Summer 5-12-2012

Proceeding of International Conference on Electrical Engineering and Computer Science ICEECS-2012

Dr. Harsh K Verma

Editorial

The mushrooming growth of the IT industry in the 21st century determines the pace of research and innovation across the globe. In a similar fashion Computer Science has acquired a path breaking trend by making a swift in a number of cross functional disciplines like Bio-Science, Health Science, Performance Engineering, Applied Behavioral Science, and Intelligence. It seems like the quest of Homo Sapience Community to integrate this world with a vision of Exchange of Knowledge and Culture is coming at the end. Apparently the quotation "Shrunken Earth, Shrinking Humanity" holds true as the connectivity and the flux of information remains on a simple command over an internet protocol address. Still there remains a substantial relativity in both the disciplines which underscores further extension of existing literature to augment the socio-economic relevancy of these two fields of study. The IT tycoon Microsoft addressing at the annual Worldwide Partner Conference in Los Angeles introduced Cloud ERP (Enterprise Resource Planning,) and updated CRM (Customer Relationship Management) software which emphasizes the ongoing research on capacity building of the Internal Business Process. It is worth mentioning here that Hewlett-Packard has been with flying colors with 4G touch pad removing comfort ability barriers with 2G and 3G. If we progress, the discussion will never limit because advancement is seamlessly flowing at the most efficient and state-of-the art universities and research labs like Laboratory for Advanced Systems Research, University of California. Unquestionably apex bodies like UNO, WTO and IBRD include these two disciplines in their millennium development agenda, realizing the aftermath of the various application projects like VSAT, POLNET, EDUSAT and many more. 'IT' has magnified the influence of knowledge management and congruently responding to social and industrial revolution.

Although the discipline like electrical engineering has narrated academic maturity in the last decades, but the limitations of the non renewable energy sources, turbulence and disturbances in the energy propagation cascades various insightfulness and stimulation in post classical electrical era. Evidence shows that there are phenomenal supplements in power generation and control after the introduction of Energy Management System (EMS) supported by Supervisory Control and Data Acquisition (SCADA). As there is increasing focus on strengthening the capacity of the power houses with the existing resources or constraints some new dimensions like FACTS, Optimal System Generation, High Voltage DC transmission system, Power Generation Control, Soft Computing, Compensation of transmission line, Protection scheme of generator, Loss calculation, economics of generation, fault analysis in power systems are emerging. Since the world is suffering with water, food, and energy crisis, energy consumption has social relevancy.

Keeping view of the ongoing energy and power issues many action research can be initiated by the research fraternity of this domain. The conference is a thought provoking outcome of all these interrelated facts.

In the quest of making this earth a better place to live we have to make a strong hold upon sustainable energy source. Sustainable energy sources include all renewable energy sources, such as hydroelectricity, solar energy, wind energy, wave power, geothermal energy, bioenergy, and tidal power. It usually also includes technologies designed to improve energy efficiency. Energy efficiency and renewable energy are said to be the twin pillars of sustainable energy. Renewable energy technologies are essential contributors to sustainable

energy as they generally contribute to world energy security, reducing dependence on fossil fuel resources, and providing opportunities for mitigating greenhouse gases.

It's my pleasure to welcome all the participants, delegates and organizer to this international conference. In the process of organizing this conference IRNet family members have shown their commitment and dedication. I sincerely thank all the authors for their invaluable contribution to this conference. I am indebted towards the reviewers and Board of Editors for their generous gifts of time, energy and effort.

Editor-in-Chief

Dr. Harsh K Verma
Associate Professor and Head of the Department,
Department of Computer Science and Engineering.
NIT, Jalandhar.

Design of a Novel Speed-Power Efficient IEEE 754 Compliant Floating Point Unit

Nija S J & T.V.J.Ganeshbabu

Francis Xavier Engineering College, Tirunelveli-627003, Tamilnadu, India
E-mail : sjnija@gmail.com

Abstract **-** IEEE 754 compliant Floating point unit for 32 bit and 64 bit computing using FPGA that consists of arithmetic modules has to be designed and developed. The Floating Point Unit acts as a co-processor that comprises of the arithmetic modules, Multiplier, Adder, subtractor and divider. The floating point data formats, operations, and special values are compared with the IEEE 754 standard with double precision format, which gives potential gains in performance. The core architecture for each module has to be selected by careful analysis based on the speed, complexity and power.This paper involves selection of suitable architecture for each module, developing FPGA program using VHDL codes and identification of the FPGA device. The Programming should be done in such a way that each module operates the mathematical data in an accurate manner. The developed FPGA program will be verified by simulation and the test cases will be developed and tested using software tools

Keywords **-** *floating point unit,IEEE 754-2008,IEEE 754-1987,rounding,binary format.*

I. INTRODUCTION

The floating point unit is a dedicated execution unit designed for performing mathematical functions on floating point numbers, typical operations are addition, subtraction, multiplication and division. Some systems can also perform various transcendental functions such as exponential or trigonometric calculations. In most modern general purpose computer architectures, one or more floating point units are integrated with the central processing unit. The IEEE 754 Standard for Floating-Point Arithmetic is the most widely-used standard for floating point computation, and is followed by many hardware like CPU, FPU and software implementations. Many computer languages allow or require that some or all arithmetic be carried out using IEEE 754 formats and operations. The current version is IEEE 754-2008, which was published in August 2008, includes nearly all of the original IEEE 754-1985, which was published in 1985 and the IEEE Standard for Radix-Independent Floating-Point Arithmetic IEEE 854-1987, which was published in 1987. . The IEEE 754-2008 standard is derived from and replaces IEEE 754-1985, the previous version, following a seven-year revision process. The standard also includes extensive recommendations for advanced exception handling, additional operations such as trigonometric functions, expression evaluation, and for achieving reproducible result.

Floating point addition is believed to be the most frequent computer arithmetic operation. Intricacies of floatingpoint number representation make floating-point addition inherently more complex than integer addition. Thus, methods for speeding up floating-point addition are of utmost importance. A floating-point number is conventionally composed of a sign bit, an exponent, and a significand. Since the ANSI/IEEE standard for binary floating-point arithmetic (IEEE 754 for short) was introduced in 1985, virtually all implementations have adhered to its representation formats, even when they do not follow the full provisions of the standard, or its revised version, IEEE 754- 2008 [3]. Implementations in this category include systems for digital signal/image processing [4] and computer graphics [5]. The short (long), also known as 32-bit or single (64-bit or double), normalized standard format incorporates a sign bit s, a biased excess-127 (excess-1023) representation e for the exponent, and a significant composed of 23 (52) bits to the right of the binary point that has a hidden 1 to its left Design of hardware floating-point units has a long history, dating back to early digital computers that were used primarily for scientific computations [13]. Initial efforts in providing high performance in floating-point units were impeded by the exorbitant cost of hardware. This forced serialization of potentially parallel steps to allow hardware reduction and sharing, even in top-of-the-line supercomputers [14]. As hardware cost decreased, an array of innovative designs began to emerge. Once the more or less straightforward

performance enhancement schemes were exhausted, replication of units and other hardware-intensive methods were employed to squeeze out incremental gains. The state of the art in floating-point adder design uses dual data paths, as depicted in Fig. 1, to separate the relatively slow alignment and post normalization shifts (Steps 2 and 6 of Algorithm 1) into different paths. This parallelization, based on the exponent difference, is credited to Farmwald [6]. Others have refined the path separation criteria. For example, Seidel and Even [15] use both the exponent difference and the actual operation for this purpose. The innovations cited above notwithstanding, there is still room for fine-tuning and improvements in speed, latency-area tradeoffs, and energy dissipation given the following challenges in high-speed floating-point adders/ subtractors:

- The prevalent sign-magnitude encoding leads to a more complex significand addition process than 1's-or 2's-complement format. Some techniques meant to speed up the addition of sign-magnitude significands (e.g., [16]) entail additional chip-area and power overheads.
- Postnormalization via counting the leading (non)-zero digits is a log-latency operation at best. However, the count of leading 0/1 digits can be obtained concurrently with addition, perhaps by deriving an approximate count quickly and finetuning the result at the end [7].
- Rounding to nearest may require an incrementation and possible exponent adjustment. With some extra hardware (e.g., the parallel-prefix adder of [15]), both the normal and an incremented result can be computed in parallel, thus allowing rapid selection of the rounded value.

Multiplying two numbers in floating point format is done by 1- adding the exponent of the two numbers then subtracting the bias from their result, 2- multiplying the significand of the two numbers, and 3- calculating the sign by XORing the sign of the two numbers. In order to represent the multiplication result as a normalized number there should be 1 in the MSB of the result (leading one).

Floating Point numbers represented in IEEE 754 format are used in most of the DSP Processors. Floating point arithmetic is useful in applications where a large dynamic range is required or in rapid prototyping applications where the required number range has not been thoroughly investigated.

Because of the increasing importance, specifications for decimal floating-point arithmetic have been added to the recently approved IEEE 754-2008 Standard for Floating-Point Arithmetic that offers a more profound specification than the former Radix-Independent Floating Point Arithmetic IEEE 754-1987.

II. PRELIMINARIES

The IEEE-754 standard was created in the early 1980s after word sizes of 32 bits (or 16 or 64) had been generally settled upon. This was based on a proposal from Intel who were designing the i8087 numerical coprocessor.[1] Among the innovations are these:

- A precisely specified encoding of the bits, so that all compliant computers would interpret bit patterns the same way. This made it possible to transfer floating-point numbers from one computer to another.
- A precisely specified behavior of the arithmetic operations. This meant that a given program, with given data, would always produce the same result on any compliant computer. This helped reduce the almost mystical reputation that floating-point computation had for seemingly nondeterministic behavior.
- The ability of exceptional conditions (overflow, divide by zero, etc.) to propagate through a computation in a benign manner and be handled by the software in a controlled way.

IEEE 754 is a widely used floating point standard.This standard has two versions: IEEE 754 - 1985 and IEEE 754 – 2008.

This standard defines a family of commercially feasible ways for new systems to perform binary floating-point arithmetic. The issues of retrofitting were not considered. Among the desiderata that guided the formulation of this standard were

1) Facilitate movement of existing programs from diverse computers to those that adhere to this standard.

2) Enhance the capabilities and safety available to programmers who, though not expert in numerical methods, may well be attempting to produce numerically sophisticated programs. However, we recognize that utility and safety are sometimes antagonists.

3) Encourage experts to develop and distribute robust and efficient numerical programs that are portable, by way of minor editing and recompilation, onto any computer that conforms to this standard and possesses adequate capacity. When restricted to a declared subset of the standard, these programs should produce identical results on all conforming systems.

4) Provide direct support for

   a) Execution-time diagnosis of anomalies

b) Smoother handling of exceptions

c) Interval arithmetic at a reasonable cost

5) Provide for development of

a) Standard elementary functions such as exp and cos

b) Very high precision (multiword) arithmetic

c) Coupling of numerical and symbolic algebraic computation

6) Enable rather than preclude further refinements and extensions.

The IEEE Standard for Floating-Point Arithmetic (IEEE 754) is a technical standard established by the Institute of Electrical and Electronics Engineers (IEEE) and the most widely used standard for floating-point computation, followed by many hardware (CPU and FPU) and software implementations. Many computer languages allow or require that some or all arithmetic be carried out using IEEE 754 formats and operations. The current version is IEEE 754-2008, which was published in August 2008; it includes nearly all of the original IEEE 754-1985 (which was published in 1985) and the IEEE Standard for Radix-Independent Floating-Point Arithmetic (IEEE 854-1987).

III. PROPOSED SYSTEM

FLOATING POINT ADDITION & SUBTRACTION

Typically, a floating-point operation takes two inputs with p bits of precision and returns a p-bit result. The ideal algorithm would compute this by first performing the operation exactly, and then rounding the result to p bit (using the current rounding mode).

Let a1 and a2 be the two numbers to be added. The notations ei and si are used for the exponent and significant of the addends ai. This means that the floating-point inputs have been unpacked and that si has an explicit leading bit. To add a1 and a2, perform these eight steps:

1. If e1 < e2, swap the operands. This ensures that the difference of the exponents satisfies d = e1–e2 ≥0. Tentatively set the exponent of the result to e1.

2. If the sign of a1 and a2 differ, replace s2 by its two's complement.

3. Place s2 in a p-bit register and shift it d = e1-e2 places to the right (shifting in 1's if the s2 was complemented in previous step). From the bits shifted out, set g to the most-significant bit, r to the next most-significant bit, and set sticky bit s to the OR of the rest.

4. Compute a preliminary significant S = s1+s2 by adding s1 to the p-bit register containing s2. If the signs of a1 and a2 are different, the most-significant bit of S is 1, and there was no carry out then S is negative. Replace S with its two's complement. This can only happen when d = 0.

5. Shift S as follows. If the signs of a1 and a2 are same and there was a carry out in step 4, shift S right by one, filling the high order position with one (the carry out). Otherwise shift it left until it is normalized. When left shifting, on the first shift fill in the low order position with the g bit. After that, shift in zeros. Adjust the exponent of the result accordingly.

6. Adjust r and s. If S was shifted right in step 5, set r: = low order bit of S before shifting and s: = g or r or s. If there was no shift, set r: = g, s: = r. If there was a single left shift, don't change r and s. If there were two or more left shifts, set r: = 0, s: = 0. (In the last case, two or more shifts can only happen when a1 and a2 have opposite signs and the same exponent, in which case the computation s1 + s2 in step 4 will be exact.)

7. Round S using following rounding rules as in Table 3.1

Table3.1Rounding

Rounding Mode	Sign of result ≥0	Sign of result <0
-∞		+1 if r ∨ s
+∞	+1 if r ∨ s	
0		
Nearest	+1 if r ∧ p_0 or r ∨ s	+1 if r ∧ p_0 or r ∧ s

If a table entry is non empty, add 1 to the low order bit of *S*. If rounding causes carry out, shift *S* right and adjust the exponent. This is the significant of the result.

8. Compute the sign of the result. If *a1* and *a2* have the same sign, this is the sign of the result.

If *a1* and *a2* have different signs, then the sign of the result depends on which of *a1*, *a2* is negative, whether there was a swap in the step 1 and whether *S* was replaced by its two's complement in step 4. Refer to table 3.2

Table 3.2 Sign of result

Swap	Complement	Sign (a_1)	Sign (a_2)	Sign (result)
Yes	Ø	+	-	-
Yes	Ø	-	+	+
No	No	+	-	+
No	No	-	+	-
No	Yes	+	-	-
No	Yes	-	+	+

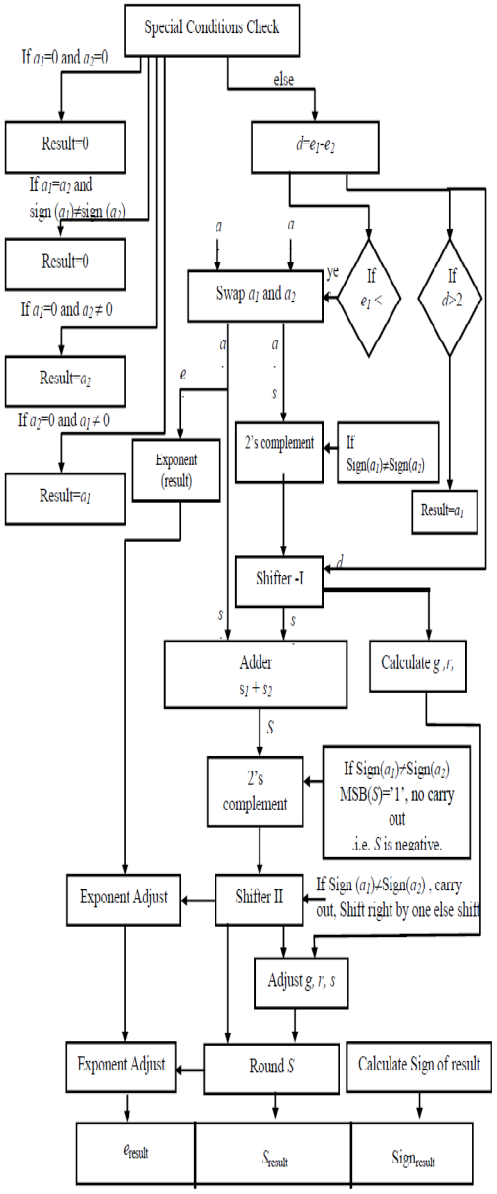
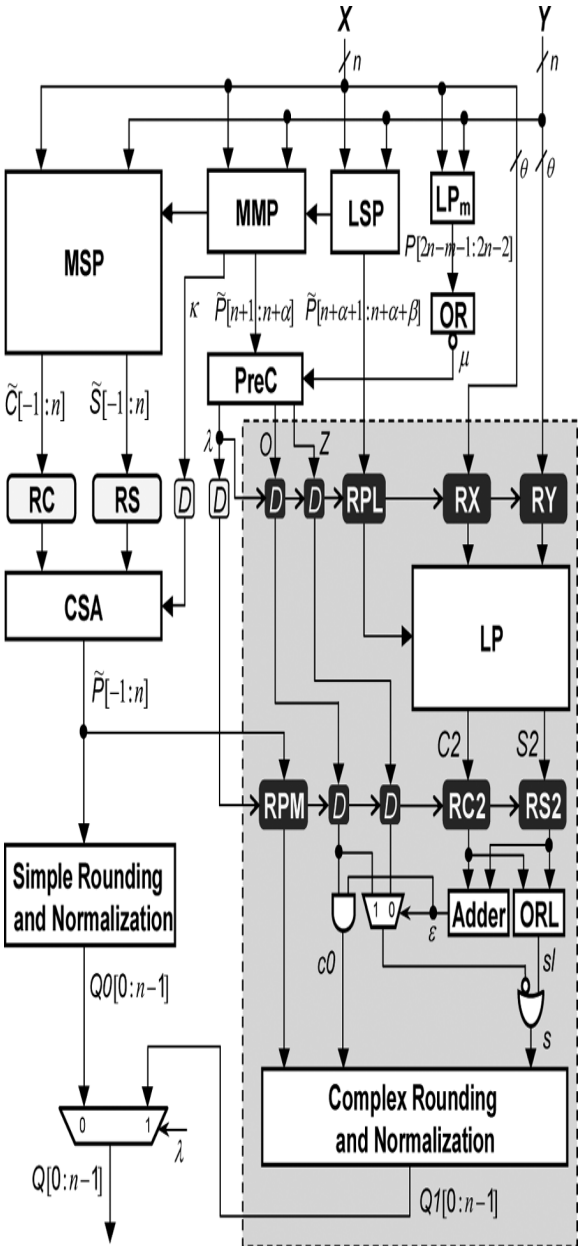


Fig 3.1 Floating point add-sub

LOW-POWER FLOATING-POINT MULTIPLICATION



Floating-point operations have been acknowledged to be useful for many real-time graphic and multimedia applications since it provides a wide dynamic range of representable real numbers. Most modern processors perform floating-point operation according to the IEEE 754-1985 Standard for binary floating-point arithmetic, which has been recently revised to include specifications for decimal floating-point arithmetic (IEEE 754R).IEEE 754-1985 floating-point multiplication is typically performed by first generating the partial products of two

n-bit (23 or 53) normalized significand in parallel, followed by reducing these partial products with an addition tree into the sum S and carry C terms. Next, a carry-propagate adder is used to sum up these two 2n-bit terms to produce a 2n-bit product P, which is then normalized and rounded to produce the final rounded n-bit product. The exponent of the product is the sum of the exponents with proper bias adjustment and increment if the prenormalized product requires a 1-bit normalization shift.

To enhance the performance, many fast floating-point multiplication algorithms compute and partition the S and C terms into upper parts and lower parts .

The featurs of this multipliers are Available for wide range of FPGA families, Pipelined, Compliance with IEEE 754 standard, Single precision real floating point support, Supports Overflow, Underflow, Invalid operation flags, Optimized for Speed and Latency, Compatible, Flexible and integrable with other module.

IV. EXPERIMENT RESULTS AND DSCUSSIONS

This design has been implemented, simulated on ModelSim and synthesized for VHDL.. Simulation based verification ensures that the design isfunctionally correct when tested with a given set of inputs. Though it is not fullycomplete, by picking a random set of inputs as well as corner cases, simulation basedverification can still yield reasonably good results.The following snapshots are taken from ModelSim after the timing simulation of the floating point multiplier and adder core.

Fig 4.1 Adder screen shot

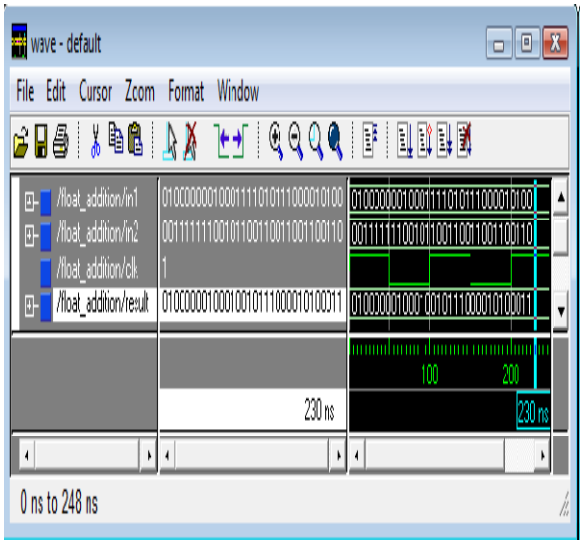


Fig 4.2 : Subtractor screen shot

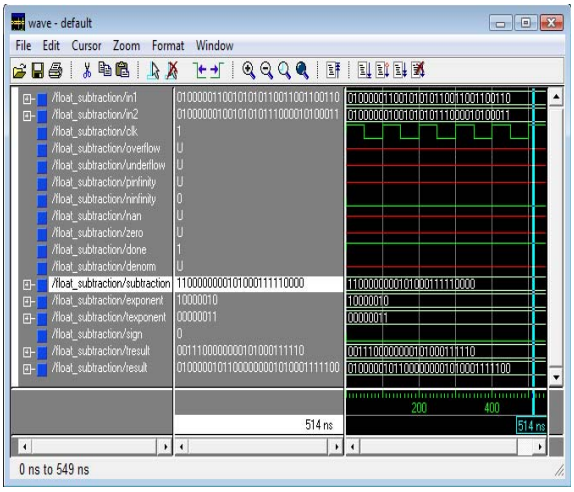
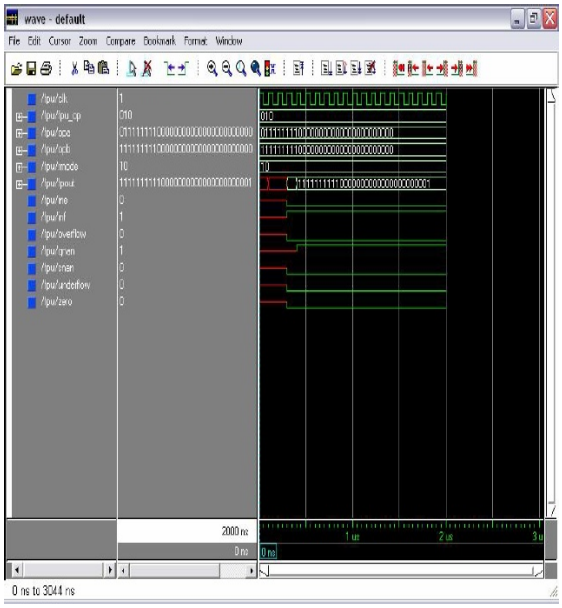


Fig 4.3 Multiplier screen shot



V. CONCLUSION AND FUTURE ENHANCEMENT

The Floating point unit is designed and developed. The design is done in such a way that the floating point unit can be effectively interfaced with any processor that is either 32 bit or 64 bit. The architectures that are used in the design are selected based on careful analysis considering the three parameters, speed, area and power consumption. In future, other arithmetic and logical modules such as trigonometric, logarithmic , square root, division, and comparison units has to be added along with this floating point unit and the floating point

unit can be extended to quadruple precision format for more advanced and scientific computations.

REFERENCES

[1] Al-Ashrafy, M.; Salem, A.; Anis, W [2011]" An efficient implementation of floating point multiplier" Communications and Photonics Conference (SIECPC), 2011

[2] Castellanos, I.D.; Stine[2006] "A 64-bit Decimal Floating-Point Comparator", International Conference on Application-specific Systems Sept. 2006

[3] C. S. Wallace, [1964] "A suggestion for parallel multipliers," IEEE Trans. Electron. Comput., ol. 13, pp. 14–17, Feb. 1964.

[4] Fahmy, H.A.H.; ElDeeb, T.; Hassan, M.Y.; Farouk, Y.; Eissa[2010] "Decimal Floating Point for future processors" IEEE International Conference on Microelectronics Dec. 2010

[5] G. Even and P.-M. Seidel, [2000]. "A comparison of three rounding algorithms for IEEE floating-point multiplication," IEEE Trans. Comput., vol. 49, no. 7, pp. 638–650, Jul. 2000

[6] Hemmert, K.S.; Underwood, K.D[2007]" Floating-Point Divider Design for FPGAs". IEEE Transactions on Very Large Scale Integration (VLSI) Systems Jan. 2007

[7] IEEE Standard for Binary Floating-Point Arithmetic, ANSI/IEEE 754- 1985, 1985.

[8] IEEE Standard for Floating-Point Arithmetic, IEEE 754™-2008, 2008.

[9] J. Y. E. Tong, D. Nagle, and R. A. Rutenbar, [2000]. "Reducing power by optimizing the necessary precision/range of floating-point arithmetic," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 8, no. 3, pp. 273–286, Jun. 2000.

[10] K. E.Wires, M. J. Schulte, and J. E. Stine, [2001] "Combined IEEE compliant and truncated floating point multipliers for reduced power dissipation," Int. Conf. Comput. Design, 2001, pp. 497–500.

[11] K. E. Wires, M. J. Schulte, and J. E. Stine, [2000] "Variable-correction truncated floating point multipliers," 34th Asilomar Conf. Signals, Syst. Comput., 2000, vol. 2, pp. 1344–1348.

[12] Kingshuk Karuri; Leupers, R.; Kedia, M.; Ascheid, G.; Meyr, H[2006]." Design and Implementation of a Modular and Portable IEEE 754 Compliant Floating-Point Unit". Design, Automation and Test in Europe, 2006.

[13] Liang-Kai Wang; Schulte, M.J [2007]" Decimal Floating-Point Adder and Multifunction Unit with Injection-Based Rounding". IEEE Symposium on Computer Arithmetic June 2007

[14] Liang-Kai Wang; Schulte, M.J.; Thompson, J.D.; Jairam, N" [2009]Hardware Designs for Decimal Floating-Point Addition and Related Operations". IEEE Transactions on Computers March 2009

[15] N. Quach, N. Takagi, and M. Flynn, [1991].On fast IEEE rounding Stanford, CA,, Tech. Rep. CSL-TR-91-459, 1991.

[16] N. T. Quach, N. Takagi, and M. J. Flynn, [2004]. "Systematic IEEE rounding method for high-speed floating-point multipliers," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 12, no. 5, pp. 511–521, May 2004.

[17] P.-M. Seidel, [1998] "How to half the latency of IEEE compliant floatingpoint multiplication," 24th Euromicro Conf., 1998, vol. 1, pp. 329–332

[18] Quinnell, E.; Swartzlander, E.E.; Lemonds, C[2008]" Bridge Floating-Point Fused Multiply-Add Design" IEEE Transactions on Very Large Scale Integration (VLSI) Systems Dec. 2008

[19] R. K. Yu and G. B. Zyner, [1995] "167 MHz radix-4 floating point multiplier," 12th Symp. Comput. Arith., 1995, pp. 149–154.

[20] Samy, R.; Fahmy, H.A.H.; Raafat, R.; Mohamed, A.; ElDeeb, T.; Farouk, Y; [2010]" A decimal floating-point fused-multiply-add unit" IEEE International Midwest Symposium on Circuits and Systems Aug. 2010

[21] Shiann-Rong Kuang, Jiun-Ping Wang and Hua-Yi Hong, [2010]. "Variable-Latency Floatingpoint Multipliers for Low power Applications" IEEE Trans.VLSI,vol.18, no.10, Oct 2010

[22] Taher, M.; Aboulwafa, M.; Abdelwahab, A.; Saad, E.M" High-Speed, [2007] " Area-Efficient FPGA-Based Floating-Point Arithmetic Modules". National Radio Science ConferenceMarch 2007

[23] Tan, D.; Lemonds, C.E.; Schulte, [2009]" Low-Power Multiple-Precision Iterative Floating-Point Multiplier with SIMD Support" IEEE Transactions on Computers Feb. 2009

[24] V. G. Oklobdzija, D. Villeger, and S. S. Liu, ,"[1996] "A method for speed optimized partial product reduction and generation of fast parallel multipliers using an algorithmic approach" IEEE Trans. Comput., vol. 45, no. 3, pp. 294–306, Mar. 1996.

◈◈◈

A Method for Inferring the Structure of Bayesian Networks in Continuous Systems Using Copulas

Hari M. Koduvely[1] & Madhu Gopinathan[2]

[1]Center for Knowledge Driven Intelligent Systems, Infosys Labs, Infosys, Bangalore, India
[2]vMobo Inc., Bangalore, India
E-mail : harimanassery_k@infosys.com[1], madhu@vmobo.com[2]

Abstract **-** We describe a new method to discover the structure of Bayesian networks from continuous data without discretizing. Our method makes use of mathematical functions called Copulas and their empirical estimation method using Bernstein Polynomials to find conditional independence between variables in the data set. Once the conditional independence between all the variables is obtained using the copula method, we make use of the Inductive Causation (IC) algorithm proposed by Judea Pearl to infer the structure of the Bayesian network. We have applied our method to a Supply Chain Performance Management System and shown that copula method along with IC algorithm produces better results compared to the use of IC algorithm on the discretized data.

Keywords **-** *component; Graphical Models; Bayesian Networks; Copula; Causal Models.*

I. INTRODUCTION

Many complex systems such as gene regulatory networks and supply chain networks contain highly interacting sub-components. One important aspect in understanding and managing of such complex systems is the cause-effect relationship between these components. For example, in a supply chain management system, one would be interested in monitoring and controlling some Key Performance Indicators or KPIs to improve the overall efficiency of the supply chain. To change the values of some high level business metrics such as total revenue one would have to typically change some operational level metrics in the system such as the number of suppliers in a given category. Therefore, it is essential to understand how the different KPIs are causally related to each other for managing a supply chain network.

To understand the cause-effect relationship between different variables, it is not enough to estimate the statistical correlations between them. Two variables could be correlated just because they both are influenced by a third variable which is acting as a common cause. It is hard to eliminate such spurious correlations using conventional statistical methods. In the last two decades, two main frameworks have been developed for measuring the cause-effect relationships. These are Bayesian Networks and Structural Equations Modeling [1].

Bayesian Networks or Belief Networks which belong to the class of Graphical Models is a powerful and intuitive method to represent causal relations between variables in a complex interacting system [1]. It is a directed acyclic graph (DAG) encoding an N-dimensional probability distribution involving N-variables, which are the nodes in the network, with the edges representing the direct dependencies between them. Bayesian Networks have been used for several practical applications such as diagnosis of diseases in medicine, root cause determination, Bio-Informatics, object tracking in computer vision, and sensor networks [2, 3].

The first step in building a Bayesian Network model is to determine the structure of the network which represents the cause-effect relationship between different variables. The next step is to determine the conditional probability distributions, which encode the strength of these cause-effect relationships from data. Once the structure of the network and conditional distributions are known, Bayesian network can be used for inference or decision making purpose. Learning the structure of the network from data is more difficult compared to learning the probability distributions or inference. It is in general a NP-Hard problem [4]. Many issues related to scalability, accuracy and discrete or continuous nature of the variables are still not solved completely. Therefore, in many cases, the structure is

constructed manually with the help of a domain expert rather than by using algorithms.

One particular issue associated with discovering the network structure is that many of the current algorithms require the variables to be discrete in nature. This is because in the case of discrete variables, there are well known and easily implementable methods to determine the conditional independency between variables. The cases of continuous variables are dealt with either discretizing or modeling them using some known parametric distributions [5]. Both discretization of continuous variables and modeling them using parametric distributions introduce approximations and hence reduces the accuracy. The error involved with discretization of continuous variables can be reduced in general by increasing the number of states. However this has two problems. Firstly, the memory requirements will go up since total number of possible states in the system increases exponentially. Secondly, more data will be required. This is because for each state of a variable some minimum amount of data would be required to accurately compute the conditional distributions or scores associated with the network. Other approach would be to use non-parametric methods. This is the approach used by Hofman and Tresp for continuous variables in their work [9].

In this paper, we describe a new method to discover the structure Bayesian networks for continuous variables, without the need to discretize them. We make use of mathematical functions called copulas and their empirical estimation methods along with the Inductive Causation algorithm of Judea Pearl for discovering the Bayesian network structure. Copulas are used in Mathematics and Statistics fields to represent multivariate distributions but its application in the Machine Learning community is very rare. Our objective of this work is to introduce the method of copulas into classical machine learning problems such as discovery of the structure of Bayesian Networks and compare its performance with other methods involving discretization of continuous data on a problem of practical interest. To our knowledge we have not seen any prior work where copulas have been used for inferring the structure of Bayesian networks from continuous data.

The rest of the paper is organized as follows. In section 2, we briefly review Bayesian Networks and Inductive Causation (IC) Algorithm which is used for discovering the network structure from data. In section 3, we introduce the concept of copulas and show how to estimate the local conditional dependencies using them. In section 4, we use the method of copulas along with IC algorithm to discover the structure of Bayesian Network in a supply chain performance management system. In section 5, we compare our method with other algorithms for discovering the Bayesian networks using their implementation in an open source machine learning software Weka.

II. BAYESIAN NETWORKS AND INDUCTIVE CAUSATION ALGORITHM

A. Bayesian Networks

A Bayesian Network consists of N variables $\{X_1, X_2,\ldots X_N\}$ which forms the set of nodes V and a set E of directed edges between these nodes representing the cause effect relationships between them. The set (V, E) forms a Directed Acyclic Graph (DAG). The Bayesian Network is a formal representation of the knowledge of conditional dependencies between the variables in V.

If $P(X_1,X_2,...,X_N)$ denotes the joint distribution of V, then the Bayesian Network represents the factorization of P into conditional probabilities given by

$$P(X_1, X_2,..., X_N)=\prod_{i=1}^{N} P(X_i \mid P_i) \qquad (1)$$

Here P_i denotes the set of parents of variable X_i (i.e. there is a directed edge from each node in P_i to X_i. Since Bayesian Networks contain directed edges, they can represent cause-effect relationship between variables rather than merely statistical correlations. An example of Bayesian Network and the factorization of the corresponding joint distribution is shown in Fig. 1.

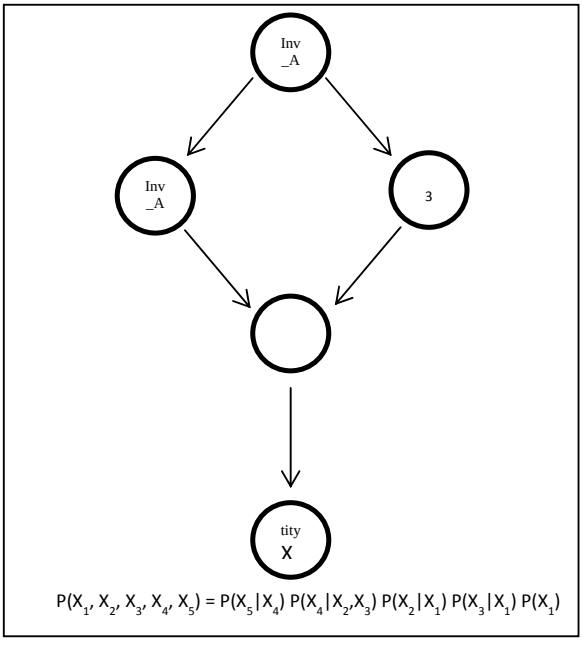


Fig. 1 : An example of Bayesian network and the factorization of corresponding joint probability distribution.

B. *Learning the Structure of Bayesian Networks from Data*

There are two approaches for learning the structure of Bayesian Networks from data. The first approach known as *Score + Search* method performs a heuristic search in the space of all possible networks and selects the network which gives the maximum score for the given data. Neither there is any guarantee that the solutions found by this method is globally optimum nor there is an accurate estimate of how close is the generated solution to the optimal one. The second is a constrained based learning approach in which the optimum structure is found by estimating the conditional dependencies between all the variables. Since Bayesian Network, by design, encodes the conditional dependencies between variables, this is a more natural approach for discovering the structure. Inductive Causation (IC) Algorithm proposed by Judea Pearl is one of the commonly used algorithms for this approach [9]. One main drawback of constrained based method is that it is very sensitive to the accuracy of estimating the conditional dependency from data.

This becomes particularly significant when the variables involved are continuous in nature. For example, consider the case of testing the dependency between two continuous variables Y and Z conditioned on X. If X is a categorical variable, and if there are sufficient observations per category it is possible to design tests for dependency between Y and Z for each category and later combine the results. On the other hand if X, Y and Z are continuous it can be shown that the hypothesis Y is independent of Z given X is not testable against any alternatives [6]. One simple explanation for this is that an *i.i.d*. sample (X, Y, Z) has, with probability one, at most one observed pair of (Y,Z) for a given X. Some methods have been suggested in the literature for testing the conditional independence of continuous variables. These are based on partial correlation coefficient and Kendall's partial tau. However these methods are applicable only when certain restrictive conditions are met [7]. Therefore in general, estimating conditional independence for continuous variables from data is a hard problem.

From the applications perspective, this is a very serious limitation. For example, consider the case of using Bayesian Networks in a Performance Management System (PMS). Typically in a PMS, there would be several KPIs (Key Performance Indicators) which are used as measures of performance of various entities. These KPIs are typically continuous variables and also aggregated measures (average over a day, week, month etc). Aggregating data would reduce their number of records. Therefore aggregated and continuous nature of KPIs poses challenges in discovering cause effect relationships between them using Bayesian Networks.

C. *Inductive Causation Algorithm*

Now we describe the Inductive Causation (IC) algorithm used for discovering the structure of Bayesian Networks from data [1]. The main steps of the IC algorithm are:

i. Start with a set of vertices V.

ii. For a pair of variables a and b in V search for a set S_{ab} so that a and b are conditionally independent on S_{ab}.

iii. Connect a and b with an undirected edge if no such set S_{ab} is found.

iv. Repeat this for all pairs of variables.

v. In the resulting undirected graph, for each pair of variables a and b connected to a common variable c through undirected edges, check if c in S_{ab}. If, then do not add any direction to the edges. If not, then add arrows from a to c and from b to c.

vi. In the partially directed graph that results, orient as many of the undirected edges as possible subjected to two conditions.

   a. The orientation should not create a new V structure $(a \rightarrow c \leftarrow b)$

   b. The orientation should not create a directed cycle.

For the last step (*vi*) Pearl suggests 4 different rules for obtaining maximally oriented graph. For details readers please refer to [1]. The most important step in the IC algorithm is step (*ii*) where the conditional dependency between two nodes a and b conditioned on a set of variables S_{ab} are found. One can use any standard statistical procedures which test the conditional dependency at this step. In the case of discrete variables, standard statistical tests such as *Chi-square* test can be applied for testing conditional dependency. However in the case of continuous variables, as mentioned earlier, testing conditional independence accurately poses a challenge.

The problem of estimating conditional independence of continuous variables from data has been studied in the domain of probability and statistics using the method of copulas [8]. These authors have used this method to discover Granger Causality between some important economic variables. However to our knowledge no one so far has used this for discovering the structure of Bayesian Networks involving many continuous variables and hierarchical relationships from data. In our work, we use the method of copulas to test conditional dependence in step 2 of the IC algorithm. In

the next section, we will give a brief review of the concept of copulas and how to test conditional independence using them.

III. ESTIMATION OF CONDITIONAL INDEPENDECE USING COPULAS

A. Introduction to Copulas

Copulas are functions used to represent the joint probability distributions of multivariate systems in a convenient way. The main advantage of copulas is that the joint distribution can be separated into two parts: the marginal distributions of each variable by itself and the copula function which combines these marginal distributions into joint a distribution. More formally, if $P(X_1, X_2,...,X_N)$ is the joint distribution of N random variables $\{X_1,X_2,...,X_N\}$ and $F_i(X_i)$ is the marginal distribution of each random variable X_i, then according to Sklar's theorem there exists the copula function C such that[10]

$$P(X_1, X_2,..., X_N)=C(F_1(X_1), F_2(X_2),..., F_N(X_N)) \quad (2)$$

Sklar's theorem also states that if the marginal distributions are continuous then the copula function is unique. Since each of the marginal distribution function is invertible we can express it as functions

$$P(X_1, X_2,..., X_N)=C(F_1^{-1}(u_1), F_2^{-1}(u_2),..., F_N^{-1}(u_N)) \quad (3)$$

or in simple notation

$$P(X_1, X_2,\ldots, X_N)=C(u_1, u_2,\ldots, u_N) \quad (4)$$

where each u_i is a continuous variable in the range [0,1].

From the copula function, which represents probability distributions, one can define the associated copula density functions as follows:

$$c(u_1, u_2,\ldots, u_N)=\frac{\partial}{\partial u_1}\frac{\partial}{\partial u_2}\ldots\frac{\partial}{\partial u_N}C(u_1, u_2,\ldots, u_N) \quad (5)$$

To estimate the copula density function from data we use a non-parametric approximation of copula functions using Bernstein's polynomial [12]. The Bernstein approximation to the copula density is given by

$$c(u_1, u_2,\ldots, u_N)= \sum_{k_1=0}^{m_1-1}\ldots\sum_{k_N}^{m_N-1} p(k_1, k_2,\ldots, k_N)\prod_{i=1}^{N} m_i\, B(m_i-1, k_i, u_i) \quad (6)$$

where $B(m,k,u)$ is the *Bernstein Polynomial* defined by

$$B(m, k, u)=\frac{m!}{(m-k)!k!}u^k(1-u)^{m-k} \qquad 0\le u\le 1 \quad (7)$$

and $p(k_1,k_2,\ldots,k_N)$ is the N dimensional empirical copula density. The empirical copula density is calculated from data by first computing the empirical copula function defined by

$$C\left(\frac{i_1}{n},\ldots,\frac{i_N}{n}\right)= \frac{\#tuples(X_1,\ldots,X_N)\mid X_1\le X(i_1),\ldots,X_N\le X(i_N)}{n} \quad (8)$$

Where $X(i_j)$ represents j^{th} order statistics of X_i and n is the number of rows of data. The empirical copula density is then found from this Empirical Formula through the differentiation operation

$$p(u_1,\ldots,u_N)=\frac{\partial}{\partial u_1}\ldots\frac{\partial}{\partial u_N}C(u_1,\ldots,u_N) \quad (9)$$

B. Testing Conditional Independence Using Copulas

The problem of testing conditional independence can be reformulated in terms of copulas as follows:

We are interested in finding from data whether two continuous random variables Y and Z are dependent conditioned on a third continuous random variable X. If Y and Z are conditionally independent then

$$P(Y\mid X, Z)=P(Y\mid X) \quad (10)$$

This means that knowing Z will not give any additional information on estimation of probability of Y given an X value. This also implies that if X is fixed there is no direct dependency between Y and Z.

The above statement can be written as a Hypothesis Test statement using conditional copula density [8].

Null Hypothesis H_0

$$P\left(\begin{matrix}c(F_X(X),F_Y(Y),F_Z(Z))=\\ c(F_X(X),F_Y(Y))\,c(F_X(X),F_Y(Z))\end{matrix}\right)=1 \quad (11)$$

Alternate Hypothesis H_1

$$P\left(\begin{matrix}c(F_X(X),F_Y(Y),F_Z(Z))=\\ c(F_X(X),F_Y(Y))\,c(F_X(X),F_Y(Z))\end{matrix}\right)<1 \quad (12)$$

where $c(F_X(X),F_Y(Y),F_Z(Z))$ is the Copula density function. To estimate similarity of two copula density functions, we use the *Hellinger Distance* defined by [11]

$$H_D=\int_{[0,1]^N}\left(1-\sqrt{\frac{c(u,v)c(u,w)}{c(u,v,w)}}\right)^2 dc(u,v,w) \quad (13)$$

Once the Hellinger distance is computed using empirical copula density estimated from data, it is relatively straight forward to find if either H_0 or H_1 holds. If Hellinger distance is close to zero (or sufficiently small) then H_0 holds, otherwise it can be rejected. For accurately estimating H_D, we have used a boot strap method. Thus using the copula density and Hellinger distance, one can do the conditional independence test on continuous variables.

IV. APPLICATION OF COPULA METHOD TO INFER BAYESIAN NETWORK IN A SUPPLY CHAIN PERFORMANCE MANAGEMENT SYSTEM

We used our copula based method for discovering the Bayesian network structure in a supply chain performance management system. This system is designed to monitor and improve the efficiency of procurement process in supply chains. It consists of several Key Performance Indicators from which we have chosen a few important ones for our study. These are Total Number of Invoices (Tot_Inv_Quantity), Total Invoice Amount (Tot_Inv_Amt), Average Price (Avg_Price), Percentage of Invoice Amount from Non-Local Suppliers (Pcnt_Nonloc_Inv_Amt), Percentage of Invoice Amount from High Cost Suppliers (Pcnt_Highco_Inv_Amt), Percentage of Defected Items (Pcnt_Defect). Though this is relatively a small system, it has the sufficient complexity for validating our method. The data set for this study was generated using a purchase order simulation tool. Daily purchase order data of two years is used for this study. From the purchase order data, the KPIs are calculated by straight forward aggregation using SQL commands. Final data set had 730 rows.

The structure of the Bayesian network obtained using our copula method is shown below. The structure captures the expected cause-effect relationships between the KPIs. For example, changing the percentage of orders from non-local suppliers or from high cost suppliers can cause changes in the average price. This in turn can cause change in the total spend. Similarly, change in the percentage of defect items (a measure of quality of supplier) can cause change in the total number of invoices because in the purchase order system we have used, defected items are re ordered by the buyer. This in turn can increase the spend amount. We experimented with different values of *m*, order of Bernstein polynomial, and found values in the range 3-5 giving best results.

Next, we used the standard implementation of IC algorithm in Weka for the generation of network structure after discretizing the data. We used 3 different discretization schemes. In the first case, we discretized each KPI into two states. In the second case, each KPI has 5 states and in the third case 10 states. We discretized by splitting the KPI value into equal size intervals and naming each interval as a state. More sophisticated methods of discretization produced qualitatively similar results. For all the cases we used IC algorithm in Weka with Bayes score function.

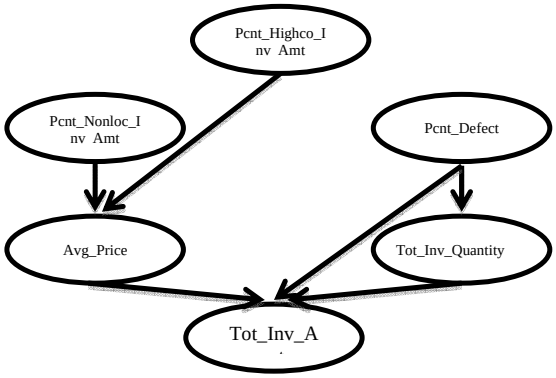


Fig. 2 : Structure of Bayesian network using Copula method

Here the first figure is the Bayesian network structure obtained with discretization to 2 states, second figure with discretization to 5 states and third figure discretization to 10 states. It is clear from the figure that keeping all other factors constant, changing the discretization scheme changes the network structure. As mentioned earlier this is due to a combination of two factors. The increase in the number of possible states and the increase in the data required for computing the conditional probability distributions accurately. The first figure captures some cause-effect relationships correctly. The arrows from percentage of non-local supplier and percentage of high cost suppliers to average price, arrow from percentage defect to total order volume are the same in the network obtained using copula method. However, the arrow between Total Number of Orders and Total Spend Amount got reversed. As the number of states increases, less number of edges are seen in the graph. This is because we need more data to compute the conditional probability distributions accurately. Therefore, the conditional independence test between some nodes is giving incorrect results leading to the removal of the edges between these nodes.

V. SUMMARY AND FUTURE WORK

In this paper, we have introduced a new method, based on copulas, for discovering the structure of Bayesian networks from data for the case of continuous variables. Copulas are commonly used in statistics for representing multivariate distributions in a more

convenient way. We have used our method in a performance management system in the context of supply chains and showed that it gives more accurate results compared to methods available in the literature which involves discretization of the variables. In the future work we will study how this method can be made more scalable in terms of computational time for systems involving large number of variables.

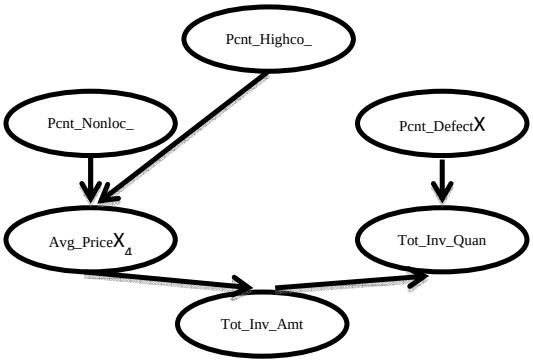


Fig. 3 : Structure of Bayesian network using discretization of the data into 2 states

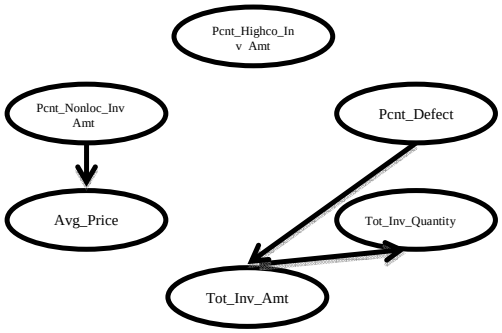


Fig. 4 : Structure of Bayesian network using discretization of the data into 5 states

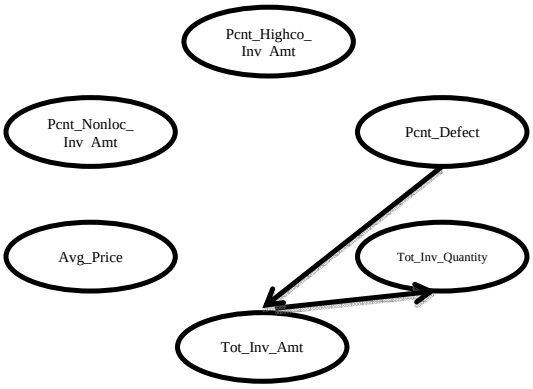


Fig. 5 : Structure of Bayesian network using discretization of the data into 10 states

ACKNOWLEDGMENT

We would like to thank Piyush Kumar Marwaha for implementing the Copula Method in .Net platform and for helping with validation.

REFERENCES

[1] Judea Peral, Causality: Models, Reasoning and Inference (Second Edition), Cambridge University Press (2009).

[2] O. Pourret, P. Naim and B. Marcot, Bayesian Networks: A Practical Guide to Applications, Wiley Publications (2008).

[3] N. Friedman, M. Linial, I. Nachman, D. Pe'er, "Using Bayesian Networks to Analyze Expression Data," Journal of Computational Biology 7 (3–4): 601–620 (2000).

[4] D.M. Chickering, C. Meek, D. Heckerman, "Large-sample learning of Bayesian networks is NP-hard," in: U. Kjaerulff, C. Meek. (Eds.), Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, Acapulco, Mexico, 2003, pp. 124–133

[5] L. D. Fu, "A comparison of state-of-the-art algorithms for learning bayesian network structure from continuous data," Masters Thesis, Vanderbilt University, 2005 (http://etd.library.vanderbilt.edu/available/etd-12022005-171510/)

[6] W.P. Bergsma, "Testing Conditional Independence for Continuous Random Variables," Unpublished

[7] E.L. Korn, "The ranges of limiting values of some partial correlations under conditional independence," The American Statistician, Vol 38, pp 61-62 (1984)

[8] T. Bouezmarnia, J. V. K. Romboutsb and A. Taamoutic, "A Nonparametric Copula Based Test for Conditional Independence with Applications to Granger Causality," Journal of Business & Economic Statistics, Dec 2011.

[9] R. Hofmann and V. Tresp, "Discovering Structure in Continuous Variables Using Bayesian Networks," Advances In Neural Information Processing System 8, MIT Press, Cambridge, MA 1996.

[10] http://en.wikipedia.org/wiki/Copula_statistics.

[11] http://en.wikipedia.org/wiki/Hellinger_distance

[12] A. Sancetta and S. Satchell, "Bernstein Approximations to the Copula Function and Portfolio Optimization," Cambridge Working Papers in Economics (2001) http://www.dspace.cam.ac.uk/handle/1810/284

◈◈◈

Path Optimization using Problem Specific Genetic Algorithm

S.V.Sastha Prashanth

Electronics and Communication Engineering, Meenakshi Sundararajan Engineering College,
Chennai, Tamil Nadu, India.
E-mail : sprashanth.msec@gmail.com

Abstract **-** This paper talks about optimization of the path taken by a mobile robot, where a knowledge based, modified genetic algorithm (GA) for path planning of a mobile robot is proposed, which uses problem-specific genetic operators for robot path planning instead of the standard GAs. The proposed knowledge based genetic algorithm contains additional Genetic operators, in addition to the standard Genetic operators, Crossover and Mutation. The Genetic Algorithm that has been proposed is a simple but effective evaluation method. The knowledge based genetic algorithm is capable of finding an optimal or near-optimal robot path in both complex static and dynamic environments. The effectiveness and efficiency of the proposed genetic algorithm is demonstrated by hardware, where IR sensors have been used to detect the obstacles and p16F877A controller has been used to implement the proposed algorithm in real time.

Keywords **-** *PSGA, Cost Function, Shortest path.*

I. INTRODUCTION

The path planning problem of mobile robots is often treated as how to find an optimum collision-free path from a start node to a goal node in an environment with obstacles. Generally, some optimization criteria (e.g., shortest distance, minimum time or lowest energy consumption) must be satisfied, and other constraints (e.g., limited velocity and acceleration of mobile robots or minimum turning radius) also should be taken into consideration. In this work, we focus the investigation on optimum path planning problem based on shortest distance criterion.

Path planning is an important issue in mobile robotics. In an environment with obstacles, path planning is to find a suitable collision-free path for a mobile robot to move from a start location to a target location. A genetic algorithm (GA) refers to an experience-based technique for problem solving, learning, and discovery. Where an exhaustive search is impractical, GAs are used to speed up the process of finding a satisfactory solution. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. Candidate solutions to the optimization problem play the role of individuals in a population, and the fitness function determines the environment within which the solutions. Evolution of the population then takes place after the repeated application of the above operators. The two major objectives of the paper are:

- The path should not collide with any obstacles
- The path should be as short as possible

II. EVALUATION OF THE PATH

Genetic Algorithm is a trial and error based algorithm, where each path taken by the robot has to be evaluated in terms of a fitness function: a function that determines the efficiency of the path taken by the robot.

The path taken by the robot in 1 run, is evaluated with the help of the ***Cost Function***, which is represented as:

$$F_{\text{cost}} = \sum_{i=1}^{N} (d_i + \beta_i C),$$

where,

N→ number of line segments.

di→ length of the line segment.

Bi→ depth of collision.

C→ constant that depends on the number of obstacles.

III. GENETIC OPERATORS

In addition to the 2 standard genetic operatiors: (a) Crossover and (b) Mutation, (c)Deletion and (d) Improvement are the additional Genetic operators used

in the paper. These Genetic operators can be easily understood from the following diagrams:

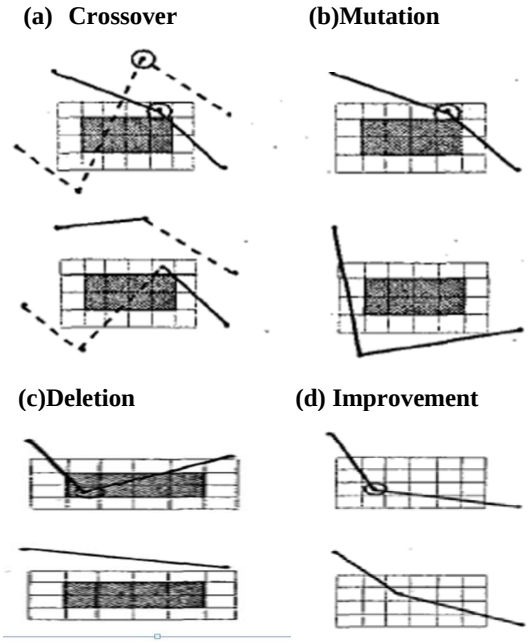


Fig.1: Representation of the genetic operators used for path optimization.

IV. CIRCUIT DIAGRAM OF THE BOT

The effectiveness of the proposed algorithm has been demonstrated by the use of hardware, where PIC microcontroller has been used to implement the proposed algorithm in real time. IR sensors have been used to detect the obstacles present on the path of navigation. A TSAL6200 Infrared transmitter module is placed at the stop point, and a TSOP1738 Infrared receiver module is placed in the robot, which makes the robot stop when the receiver detects the Infrared signal from the corresponding transmitter module.

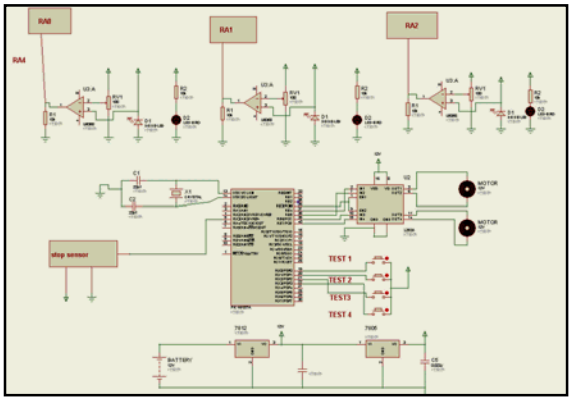


Fig. 2: Circuit Diagram

V. FLOW CHART OF ALGORITHM

A. 1^{ST} PASS:

- Initially, all the interrupts are enabled. The respective ports are declared as input/output ports.
- The controller checks whether button 1 is pressed. If no, it waits till B1 is pressed.
- Disable all the interrupts.
- The robot movement activation code is enabled, and the samples are calculated depending on the obstacles present on the path of navigation.
- The cost function of the path is calculated.
- The robot is halted if the stop sensor is activated.
- The controller now checks whether B2 has been pressed. If not, it waits till B2 is pressed.
- The same procedure is repeated again.

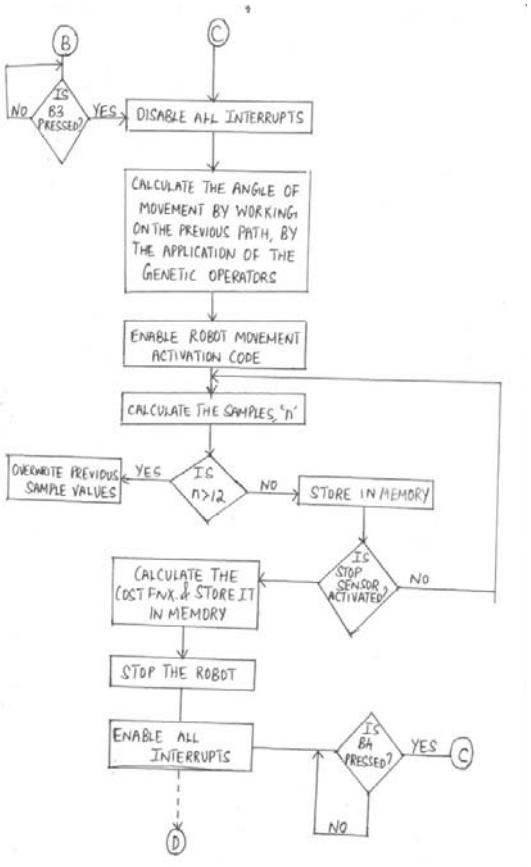


Fig. 3: Flow chart of Algorithm: ***1^{st} pass***

B. 2ND PASS

- Initially, all the interrupts are enabled. The respective ports are declared as input/output ports.
- The controller checks whether button 3 is pressed. If no, it waits till B3 is pressed.
- Disable all the interrupts.
- The controller now finds the angle of movement of the robot, by using the domain knowledge of the environment obtained during the first two runs, in order to reduce the cost function of the path.
- The robot movement activation code is enabled, and the samples are calculated depending on the obstacles present on the path of navigation.
- The cost function of the path is calculated.
- The robot is halted if the stop sensor is activated.
- The controller now checks whether B2 has been pressed. If not, it waits till B2 is pressed.
- The same procedure is repeated again.

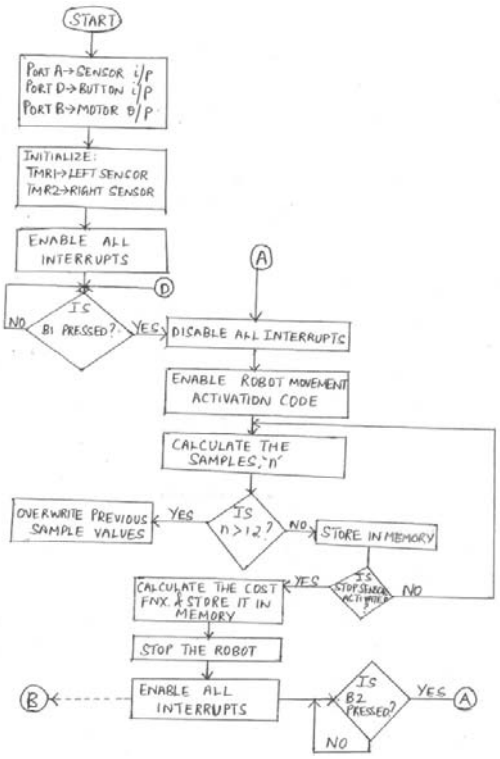


Fig. 4: Flow chart of Algorithm: ***2st pass***

VI. FUTURE ENHANCEMENTS

The efficiency of the proposed algorithm gets reduced, as the number of obstacles in the environment increases. Moreover, the number of trials required to find the shortest possible path can be reduced further, by including more genetic operators. Thus our future work would be to improve the performance of the proposed algorithm by including more genetic operators like *line repair* and *node repair*, by considering a larger, global arena, where the number of obstacles would be much more.

VII. CONCLUSION

The proposed Knowledge Based, Problem Specific GA uses a simple and unique robot path representation that combines coordinates of environment representations. The genetic algorithm incorporates the domain knowledge into its problem specific genetic operators. The developed GA also features its efficient evaluation method that is greatly beneficial for evolving good solutions from infeasible solutions. The effectiveness of the knowledge based operators is demonstrated by hardware. The results also show that the proposed genetic algorithm is effective in both complex static and dynamic environments. The efficiency of computational time makes the proposed knowledge based genetic algorithm be able to be applied to real applications.

BIBLIOGRAPHY

[1] Jaroslaw Arabas and Stanislaw Kozdrowski, "Population Initialization in the Context of a Biased, Problem - Specific Mutation" Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence.

[2] Qing Li, Wei Zhang, Yixin Yin, Zhiliang Wang Guangjun Liu, "An Improved Genetic Algorithm of Optimum Path Planning for Mobile Robots", Sixth International Conference on Intelligent Systems Design and Applications, 2006. ISDA '06.

[3] Soh Chin Yun, Veleppa Ganapathy, Lim Ooi Chong, "Improved Genetic Algorithms based Optimum Path Planning for Mobile Robot", 11th International Conference on Control Automation Robotics & Vision (ICARCV), 2010.

[4] Li, Xinhai Tong, Sijiang Xie, Yingchun Zhang, "Optimum Path Planning for Mobile Robots Based on a Hybrid Genetic Algorithm", Qing Sixth International Conference on Hybrid Intelligent Systems, 2006. HIS '06.

[5] Wang Jianguo, Zhang Yilong, Xia Linlin, "Adaptive Genetic Algorithm Enhancements for Path Planning of Mobile Robots", International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), 2010.

[6] Pu Shi, Yujie Cui, "Dynamic Path Planning for Mobile Robot Based on Genetic Algorithm in Unknown Environment", Control and Decision Conference (CCDC), 2010.

[7] Wang Haiying , Xu Rui, Bai Mingyou and Jia Jinlin, "Optimization Design of Mobile Robot Based on Genetic Algorithm" International Conference on Mechatronics and Automation, 2009. ICMA 2009.

◈◈◈

A Mobile Payment System Through Independent M-Signature Service

P.Vaishnavi & M. Ibrahim

Department Of Computer Applications, Anna University of Technology - Tiruchirappalli
E-mail : vaishmk@gmail.com, ibrahim.aut@gmail.com

Abstract **-** Nowadays a great number of applications require the use of Electronic Signature (e-signature) and non-repudiation services such as certified e-mail, contract signing, electronic payment system and etc... The authenticity and integrity generated by e-signature in electronic document is like a handwritten signature in the paper document. The mobile user would be able to use any application that requires an e-signature is called as Mobile Signature (m-signature). M-signature can be created by different ways, such as Server-based signatures, Mobile Signature Service (MSS), Mobile Signature Application Unit (MSAU). But those have several limitations. In this work, we present the implementation of mobile payment system using m- signature service to an insurance company and for a mobile shop.

Keywords **-** *Electronic signature , Mobile signature, MSAP, iMSSP.*

I. INTRODUCTION

Electronic signature (e-signature) provides interesting features such as integrity, authentication and non- repudiation. Different applications use the e-signature for non-repudiation services such as (mobile) electronic payment systems, certified e-mail, contract signing protocols, e-auctions, long-term preservation of documents, e-procurement, e-invoices, etc. The e-signature provide guarantee of authenticity and integrity for electronic documents like a paper documents. (Ruiz et al., 2007)

In developed countries almost everybody has a mobile handset (mobile phones, personal digital assistants, etc.) and almost everybody has it on them all the time. The e-signature based mobile applications could be developed for mobile devices. Any user would be able to use any application that requires an e-signature. This kind of signature is named mobile signature (m-signature)

It's proposed as a solution that can be used in any application. Furthermore, thanks to the fact that the signature is generated in a mobile device, it may be used everywhere, every time. Finally, the mobile signature is designed so that application providers do not have to develop multiple solutions for the wide range of mobile handsets, mobile operating systems and e-signature technologies that exist for mobile devices. m-signature can be used to provide another security services such as mobile authentication, service identity signing, etc.M-signature can be created by different ways, such as Server-based signatures, Mobile Signature Service (MSS), Mobile Signature Application Unit (MSAU). But those have several limitations, the server-based signatures cannot be considered legally equivalent to the handwritten signatures, The MSS proposal would require that every Mobile Network Operator supported it in order to provide a universal solution, The MSAU, it does not define an standardized interface for the invocation. But we propose an implementation for the payment system for an insurance company and for a mobile shop. The important features are being implanted are avoid intervention of MNO, the user has more control on the signature process, here all the information is exchanged in a secure way.

II. RELATED WORKS

"Electronic signature(e-signature)is fundamental in fields such as electronic commerce and government since it provides some interesting features such as integrity, authentication and non-repudiation". (Ford and Baum, 1997; Sherif, 2000). "The purpose of the e-signature is to guarantee the authenticity and integrity of electronic documents in a way equivalent to the handwritten signature in paper documents". (Rossnagel, 2004; Ruiz-Martıoinez et al.,2007). "This process MSS_Registration) is carried out by means of a MASP such as the RA. In some cases the user has to provide some information that completes registration process such as a PIN, a certificate, etc". (European Telecommunications Standards Institute (ETSI), 2003a). "The roles are similar to the ones proposed by the ETSI,

namely, Mobile Network Operators, Certification Authority, Registration Authority, Mobile Signature Service Provider and the Mobile User". (European Telecommunications Standards Institute (ETSI), 2003a)

III. PROPOSED SYSTEM

In this paper we propose a new Mobile Signature Service architecture that is MNO-independent and is not linked to any mobile handset-based specific technology to generate the electronic signature in the mobile handset.

We have improved the functionality of the MSSP and reduced its overhead because in our model the MSSP, which is named MNO-independent MSSP (iMSSP), is not responsible for checking whether the mobile handset is available to perform an m-signature.

In our proposal the mobile user, when available and able to perform signature processes, checks by means of a Web service in the iMSSP whether there are m-signature requests to be signed. Currently, it is feasible to invoke Web services from mobile devices thanks to the development of the Extensible Markup Language (XML) and Web services (WS) Application Programming Interfaces (APIs) and libraries for mobile handsets

A Web service interface to be developed by the MSSPs for mobile clients, we promote the development of our m-signature solution for a mobile handset since the m-signature application developed could be used with different MSSPs.

IV SYSTEM PROCESS FOR MPLEMENTATION OF MOBILE PAYMENT SYSTEM

The purpose of this section is to provide an overview of the different processes that will take place in our mobile signature system. The purpose of obtaining a certificate the mobile user performs a certification process with a Public Key Infrastructure (PKI) (step 0). Once the mobile user owns a certificate he performs an enrollment process (step 1) with an iMSSP. Through this process the mobile user obtains an identifier that will be used to receive mobile signature requests and a mobile signature application, if needed, to sign them. When the mobile user wants to use an m-signature-based services/application of a MASP, he has to provide her identifier to the MASP(step2).In our case of use, let us suppose that insurance application or for the mobile shop, the user chooses the options he wants to enrol in from the list of the fills in a form with his personal data. The mobile user has to sign the form to confirm his/her enrolment. For this purpose, he provides his identifier to initiate the process that will allow the application to obtain the signature of the data introduced in the form. From the identifier, the insurance/mobile application locates the WS of the iMSSP and sends an m-signature request (step 3). The iMSSP stores all the requests received from different MASPs (insurance/mobile shop) and when the mobile user is available to make m-signatures, she connects to the iMSSP to obtain m-signature requests and signs them if he agrees with them(step4).In our case of use, the mobile user, sometimes after having provided his identifier to the application, will use his m-signature application to connect to the iMSSP and receive as m-signature request the information related to the course in order to sign it. Once them-signature has been performed, the MASP can obtain it, validate it(step5)and then, the application or the service confirms the finalization of the process and performs its task or provides the information needed(step6). In our case of use, the MASP would validate the signature and would provide a receipt to his enrolment in the selected course to the mobile user.

COMPONENTS:

This system consists of the following components (shown in Figure 1):-

Mobile User:

Here the user can use m-signature based applications through network connected mobile.

iMSSP: (independent mobile signature services provider)

It is service, for creating signature for the users. The personal details of the user are gathering from applications and stored on here one database. The signatures are created by users given details based example name, organization and etc. The PKI(Public Key Infrastructure) is creates one certificate and one key for each users, that certificate is act as a m-signature. Both the key and certificate are saved on imssp database.

MASP: (Mobile Service/ Application Provider)

It's are mobile payment system based applications. Here we use two different applications though are

1. Online Mobile Shop
2. Online Insurance payment system.

These two applications have some common modules,

- **Login Form** for existing users. Here the user can give their username and password for login to their account.
- **Registration Form** for new users. Here the new user enters their personal details for

creating a new signature. It is stored on imssp database.

- **Details Form** for show/explain details about the mobile phones/insurances plans.
- **Payment Form** here we choose and pay money for our favorite mobile phone/insurance plan.
- **Report Form** after pay the money some details about products and thank you messages have been display on this form.
- **Logout Form** its final form for end the process.

Bank Service:

It is any one general bank system used for pay the money for users from their account, because who have account on bank he/she only eligible for creating the signature on imssp. Its have three modules,

- **Deposit Form** here the account holders deposit money to their account.
- **Withdrawal Form** here the user withdraw money for their use, in directly or through ATM or through our mobile payment system based.
- **Display Form** here the user seen their personal details about their account.

V. DEVELOPMENT ENVIRONMENT

We have developed based on Java language, which is based on object oriented programming concept. The IDE supports the SOAP messages and WSDL libraries for the development of two services. Those web service for the insurance service and for the mobile shop are been provided through Apache Tomcat server.

Fig 2 (a) Home Page for Insurance plan

Fig 2 (b) Home Page for Mobile shop

Fig 2 a and b shows the home page for Mobile shop and Insurance company.

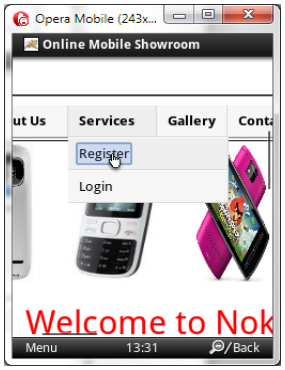


Fig 3 Register and login services for Mobile shop

Fig 3 The mobile shop have basic two forms. One is for exiting user login form and another for new user registration form. In the time of registration one key and one certificate are been created by PKI and it stored on iMSSP database.

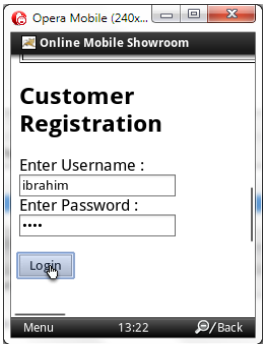


Fig 4 Login form form mobile customers

Fig 4 This is the login form for the existing users. The mobile user can enter the user name and password on here.

Fig 5 Mobile choosing form on Mobile shop

Fig 5 This form is used to select the mobile for purchase and display information about the login user details.

Fig 6 Final report details form for Insurance plan

Fig 6 This form is used for display the final report of the service for mobile shop or insurance plan. Thank you message is display on is form.

VI CONCLUSIONS

In the implementation of the mobile payment system the low cost, secure, ubiquitously accessible, auto-configurable, remotely controlled solution for automation of mobile payment system are been implemented and tested for the insurance company and the mobile shop. This process adopts the m-signature process through the web service as iMSSP. The approach discussed in the system is novel implementation for the Mobile payment system through m- signature service. The basic level of network control and remote monitoring has been implemented. In future the system will be provided with more secured features to implement m- signature for the independent mobile payment system.

REFERENCES

[1] European Telecommunications Standards Institute (ETSI). Mobile Commerce (M-COMM); mobile signatures; business and functional requirements. Technical report 102 203, May 2003a.

[2] European Telecommunications Standards Institute (ETSI). Mobile Commerce (M-COMM); mobile signatures; Web service interface. Technical report 102 204, August 2003b.

[3] Ruiz-Martinez A, Sanchez-Martinez D, Martinez-Montesinos M, Gomez-Skarmeta AF. A s urvey of electronic signature solutions in mobile devices. Journal of The oritical and Applied Electronic Commerce Research 2007.

[4] Hassinen M, Hypp¨onen K, Haataja K. An open, PKI-based mobile payment system. In: Proceedings of the emerging trends in information and communication security, 2006.p.86–100.

[5] Rossnagel H. Mobile qualified electronic signatures and certification on demand. In: Proceedings of first European PKI workshop: research and applications, EuroPKI 2004, 2004. p. 274–286.

[6] Overview of Electronic Signatures and Records Act (ESRA) Alan S. Kowlowitz Strategic Policies, Acquisitions and e-Commerce NYS Office for Technology ppt.

◈◈◈

Neural Network based Age Classification using Linear Wavelet Transforms

Nithyashri Jayaraman[1] & G.Kulanthaivel[2]

[1]Department of Computer Science & Engineering, Sathyabama University Old Mamallapuram Road, Chennai, India
[2]Electronics Engineering, National Institute of Technical Teachers Training & Research, Taramani, Chennai, India
E-mail : [1]nithyaranjith2002@yahoo.co.in, [2]gkvel@rediffmail.com

Abstract **-** The facial image analysis for classifying human age has a vital role in Image processing, Pattern recognition, Computer vision, Cognitive science and Forensic science. The various computational and mathematical models, for classifying facial age includes Principal Component Analysis (PCA) and Wavelet Transforms and Local Binary Pattern (LBP). A more sophisticated method is introduced to improve the performance of the system by decomposing the face image using 2-level linear wavelet transforms and classifying the human age group using Artificial Neural Network. This approach needs normalizing the facial image at first and then extracting the face features using linear wavelet transforms. The distance of the features is measured using Euclidean distance and given as input to Adaptive Resonance Theory (ART). The network is trained with an own dataset consisting of 70 facial images of various age group. The goal of the proposed work is to classify the human age group into four categories as Child, Adolescence, Adult and Senior Adult.

Keywords **-** *Linear Wavelet Transform, Euclidean Distance, Age group, Feature Extraction, Neural Network.*

I. INTRODUCTION

Face recognition based on Illumination , Pose, Expression and Age variations has been an active area of research and more challenging in various fields like Pattern recognition, Image processing and Computer vision. The age–separated faces has unique characteristics which makes age classification as a complex task. The classification of age has its application in finding missing individuals, identifying criminals in forensic sciences etc., This paper focuses on age classification in which the facial images are classified into four age groups: Child , Adolescence, Adult and Senior Adult. This proposed method proves that human age can be classified based on the facial features using Artificial Neural Network.

The craniofacial growth, plastic surgery, environmental changes, chronic diseases, changes the face with an increasing age. This is the first work done in age classification and was developed by Kwon and Lobo[6] . They used four categories of age group for classification : Babies, Young, Adults and Senior Adults. They used Snakelets to find wrinkles to distinguish the young and the old.

The proposed work includes normalizing the face image and this pre-processed image is said to undergo linear wavelet transform to extract the face features. The distance of different face features are measured using Euclidean distance and these features are given as inputs to the ART network to classify the various age groups. The datasets are trained using ART Network which yields better results. The ultimate goal of the proposed work is to classify the age group of the given facial image and not to recognize the faces.

II. RELATED WORK

There are various age classification methodologies that has been developed so far and they are summarized as below. The various approaches in this area includes Age prototypes[10] , Statistical models [8] and distance based techniques [9]. Moreover, age estimation techniques can be used as a basis for developing - age progression algorithms[11].

A.J.O'Toole, H. Abdi, K.A. Deffenbacher and J.CBarlett [7] used an auto associative memory technique to classify faces by gender and race. This approach consists of a set of associative memory of completely interconnected units. The Eigen vectors were extracted and the difference in co-efficients helps in classifying male/female and Japanese and Caucasian classification.

Horng, Lee and Chen [5] developed classification of age groups based on facial features. They used four categories of age group for classification : Babies, Young adults, Middle- aged adults and Senior Adults.

They used two Back Propagation Networks for age classification phase, in which, one is used to check whether the image is a baby and the other is to classify whether young adult, middle- aged adult or senior adult.

Asuman & Varif [3] developed an automatic age classification using Local Binary Pattern . In this work, the faces are divided into small regions from which LBP histograms are extracted. Later they are concatenated into a feature vector. In the classification phase, minimum distance, k- nearest neighbor classifiers are used.

Feng Gao and Haizhou [4] introduced a face age classification on consumer images using Gabour Feature and Fuzzy LDA method. To solve intrinsic age ambiguity problem, Fuzzy LDA method was used. Gabour Feature is extracted for face representation and later used in LDA classifiers.

Age Estimation based on Neural Networks using Face features was developed by Nabil Hewachi et.al [1]. They used four categories for classification : Child, Young, Youth and Old. A supervised back propagation network was used for classification. Later they tested the face images using two databases : FGNET and MORPH.

Laura , Bradley and Ken [2] developed a new set of young adult Caucasian male faces and was created with FaceGen software with which an internet based version of testing was done. They conducted a series of experiments and found that the ability to learn and recognize unfamiliar faces improves until the early 30's.

The craniofacial growth, plastic surgery, environmental changes and chronic diseases changes the face with an increasing age . It was first developed by Kwon and Lobo [6]. They used three categories of age group for classification namely Young, Adults and Senior Adults. The snakelets are used for finding wrinkles so as to distinguish young and old.

Ye Sun et al, [12] developed an Embedded Hidden Markov Model(EHMM) to recognize face and age. The nonlinear relationship between the feature points in the face are estimated and different ages of the same face are used to train EHMM , so as to estimate ageing face.

Allison C Lamont et al, [13] presented a study on recognition accuracy based on ageing effects. The face recognition accuracy decreases with young faces when compared to old aged faces.

III. PROPOSED WORK

A. Overview of Proposed Work

The face features were extracted by decomposing the input image using 2- level linear wavelet transformation. In this work linear wavelets like Haar, Symlet, and Daubechies wavelet transformations were used for decomposition.

The results of 2- level linear wavelet yields the face features which includes the eyes, nose, lips and chin etc., The distance between the face features were estimated using Euclidean distance and they are referred as Feature Point Distance(FPD). These FPD's were given as input to a Neural Network for classifying the age group. Adaptive Resonance Theory (ART) algorithm is used for classifying the human age. A own dataset including 70 face images of various age groups are trained using ART. In this work the age groups are classified into Child(0 to 12 years), Adolescence(13 to 18), Adult(19 to 55) and Senior Adult(56 & above). The entire representation of this work, can be seen in the fig 3.1.

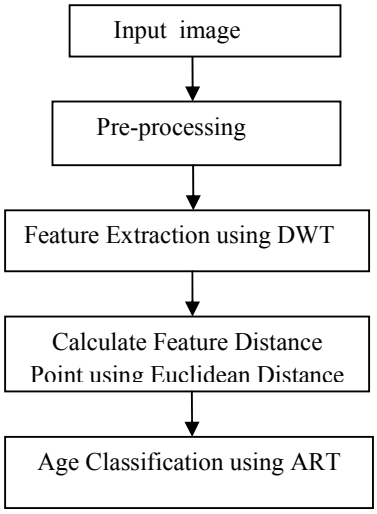


Fig. 1. Block Diagram of the Proposed Work.

B. Normalizing the image

The face images are normalized or pre-processed , which includes converting a RGB image into Gray Scale image. A RGB (Red, Blue, Green) image is an array of pixels and the number of bits used to represent a pixel value determines the bit depth of RGB image. The RGB images are converted into Gray Scale images for easier processing of facial images.

C. Feature Extraction using Linear Wavelet Transforms

The Wavelet Transform (WT) provides a time-frequency representation of the signal. Wavelet transforms are used to reduce the number of bits required to represent an image. Image Compression is done in order to get useful information from the image. A 2-D transform is done by applying 1-D transform

twice. The decomposition of can be achieved with various Linear Wavelet Transformation. It includes Haar, Symlet, and Daubechies transformations.

The gray scaled image or the normalized image, undergo 2-level Wavelet decomposition and the face image is decomposed using linear wavelets. The face features of different age groups namely Child, Adolescence, Adult and Senior Adult were obtained. The features including Eyes, Nose, Lips, Chin are extracted after two-level wavelet decomposition.

D. Finding Feature Point Distance

In this work, 4 different types of Feature Point Distance are measured namely, **FPD1, FPD2, FPD3**, **FPD4.** The mid-point of line joining the two eyes be measured as **FPD1.** The distance between the mid-point of eyes and the nose be measured as **FPD2**.The distance between the mid-point of eyes and the lips be measured as **FPD3**. The distance between the midpoint of eyes and the chin be measured as **FPD4.** These feature points are given as input to a Neural Network for classifying the age group. The distance between feature points were measured using Euclidean distance . The distance between the pixel **(i,j)** and the pixel **(k, l)** were calculated by the formula

$$D_E [(i, j), (k, l)]=[(i-k)^2 + (j-l)^2]^{1/2} \quad (1)$$

E. Age Classification using Adative Resonance Theory

Artificial Neural Network has been developed as a generalization of Mathematical models of Human Cognition or Neural Biology. Artificial Neural Network consists of many nodes and these processing units were analogous to neurons in the human brain. Adaptive Resonance Theory Network (ART) is used for classification of age and is explained in detail as below.

The basic architecture of adaptive resonance neural network consists of three types of neurons as shown in the figure 2.

1. Input Unit – F1 layer
2. Cluster Units – F2 layer
3. Reset mechanism to control the similarity of patterns placed on same cluster.

Input Processing (F1 layer) is divided into two regions namely the input region and the interface region.
(i) Input region (denoted as F1(a)).
(ii) Interface region (denoted as F1(b)).

1. Input Unit

The input region represents the input vector whereas the interface region combines the signal from the input region with the F2 layer. The interface region F1(b) layer is connected to F2 layer through bottom up weights and F2 layer is connected to F1(b) layer through top down weights.

2. Cluster Unit

The cluster unit with largest net input is selected to learn the input pattern. The activation of all other F2 units are set to zero. The intermediate units now combine the information from the input and the cluster units.

3. Reset mechanism

Based on the similarity of the input vector and the top – down weight, the cluster unit may or may not be allowed to learn the pattern. If a cluster unit is not allowed to learn , then a new cluster unit is selected.

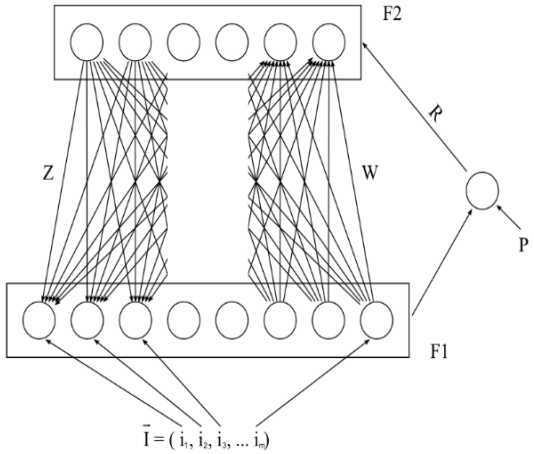


Fig. 2 : Adaptive Resonance Theory Network

The following are the steps involved in classifying the age groups.

1. The Feature point distances namely **FPD1, FPD2, FPD3**,and **FPD4** were given as input to input vector of a Adaptive Resonance Theory Network.
2. Let the threshold value be fixed as 0.6.
3. The input vector is compared with the threshold value

```
for i=1:r
  for j=1:(c-1)
    if(x(i,j)<0.6)
      ip1(i,j)=0;
    else
      ip1(i,j)=1;
    end
```

end

end

4. The ART network is said to be trained

   and the time taken and the efficiency in training the patterns are tabulated.

5. The feature point distance FPD3 (the distance between eyes and lips) consists of wrinkles near the eyes and nasal lines.

6. Therefore, pixels in this region will be greater, when compared to other regions.

7. Greater the number of pixels, more will be the age.

IV RESULTS AND DISCUSSIONS

The experimental work is done using own dataset consisting of various face age groups and totally 70 images were used. The implementation of this work is done using MATLAB. The input facial images were converted into Gray-scale images. The sample input face images and the gray scaled images are shown in Fig.3 and Fig.4 respectively.

Child Adolescence Adult

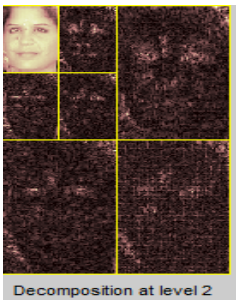
Fig.3. Sample input face images

Child Adolescence Adult

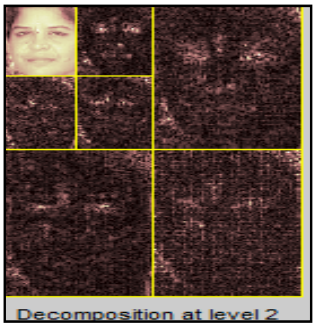
Fig.4. Gray-scaled Images

The gray scaled images were decomposed using various Linear Wavelet Transformations like Haar, Symlet (sym8), and Daubechies wavelets (db6). The figure 5, shows the decomposition of a face image using linear wavelets.

a) Haar



b) sym8.



c) db6

Fig. 5 : a) Haar b) symlet c) Daubechies wavelet transformation

The face features were extracted as a result of decomposing the gray scaled image. The features like eyes, nose, lips and chin were extracted from the facial images. The Euclidean distance for these feature points were measured and tabulated as below. The Feature Point Distance measured from 2-level Haar, sym8 and db6 are listed in the table 1, 2 and 3 respectively.

Table 1 Feature Point Distance measured from 2-level Haar wavelet transformation

FACE IMAGES	FPD1	FPD2	FPD3	FPD4
IM1	0.6756	0.8518	0.9457	1.6103
IM2	1.3841	0.9816	1.7501	2.2289
IM3	1.0279	0.9157	1.4993	2.0339
IM4	1.2828	0.9485	1.5895	2.1297
IM5	1.4159	0.9677	1.8459	2.3798
IM6	0.7443	0.8968	0.9541	1.7380
IM7	1.4763	0.9738	1.8307	2.4899
IM8	1.8086	0.9982	1.9849	2.7015
IM9	1.8960	0.9826	1.9982	2.8459
IM10	1.1774	0.9023	1.5609	2.0597

Table 2 Feature Point Distance measured from 2-level Sym 8

FACE IMAGES	FPD1	FPD2	FPD3	FPD4
IM1	0.6851	0.8520	0.9459	1.6211
IM2	1.3839	0.9811	1.7522	2.2332
IM3	1.0276	0.9160	1.4997	2.0433
IM4	1.2825	0.9488	1.5892	2.1311
IM5	1.4251	0.9679	1.8460	2.3810
IM6	0.7450	0.8965	0.9547	1.7379
IM7	1.4771	0.9733	1.8310	2.4888
IM8	1.8092	0.9984	1.9851	2.7022
IM9	1.8965	0.9816	1.9988	2.8433
IM10	1.1771	0.9027	1.5611	2.0599

Table 3 Feature Point Distance measured from 2-level db6

These feature point distances were given as input to ART neural network to classify the age group and is trained using threshold value and is listed in the table4

Table 4 Threshold ρ = 0.6

S.no	Training in %	Testing in %	Time in (ms)	Efficiency in %
1	10	90	0.0439	77.7778
2	20	80	0.0571	87.5
3	30	70	0.0718	85.7143
4	40	60	0.0666	83.3333
5	50	50	0.0911	80

The ART network is trained as shown in the table – V. The total number of pixels for the feature point FPD3 is calculated for all the images. If pixel count is lesser than 120 pixels , the classified age group is Child. If the pixel count is in between 120 and 160, then the age group is classified as Adolescence. The classified age group is Adult, if the pixel count is greater than 160 and less than 400. If the number of pixels exceeds 400, then the classified age group is Senior Adult. The results of Age classification with Haar, Sym8 and db6 are listed in the tables 5,6 and 7 respectively. The classification rate of these wavelets is compared and is shown in the figure 6.

Table 5 Results of Age classification with Haar

Face images	FPD3	pixel count	Classification
IM1	0.9457	109	Child
IM2	1.7501	344	Adult
IM3	1.4993	148	Adolescence
IM4	1.5895	155	Adolescence
IM5	1.8459	361	Adult
IM6	0.9541	115	Child
IM7	1.8307	355	Adult
IM8	1.9849	464	Senior Adult
IM9	1.9982	497	Senior Adult
IM10	1.5609	151	Adolescence

Table 6 Results of Age classification with Sym 8

Face images	FPD3	pixel count	Classification
IM1	0.9459	112	Child
IM2	1.7522	348	Adult
IM3	1.4997	149	Adolescence
IM4	1.5892	154	Adolescence
IM5	1.8460	360	Adult
IM6	0.9547	118	Child
IM7	1.8310	359	Adult
IM8	1.9851	469	Senior Adult
IM9	1.9988	499	Senior Adult
IM10	1.5611	155	Adolescence

Table 7 : Results of Age classification with db6

Face images	FPD3	pixel count	Classification
IM1	0.9455	110	Child
IM2	1.7525	349	Adult
IM3	1.4991	146	Adolescence
IM4	1.5895	157	Adolescence
IM5	1.8466	364	Adult
IM6	0.9549	118	Child
IM7	1.8315	356	Adult
IM8	1.9855	466	Senior Adult
IM9	1.9991	499	Senior Adult
IM10	1.5620	154	Adolescence

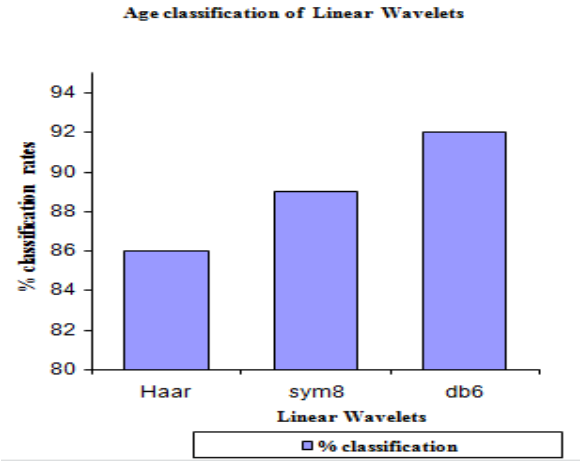


Fig. 6 : Comparison of Age classification with three Linear Wavelets

V CONCLUSION

This work proposes a Linear Wavelet Transform for extracting the face features and an Adaptive Resonance Theory for classifying age groups. The network is trained with images of different age groups and out of the three waveforms, it is observed that db6 classifies

92% of images correctly. It is proposed to improve the performance by training with more images and working with other wavelet transforms.

REFERENCES

[1] Nabil Hewahi, Aya Olwan, Nebal Tubeel, Salha EL-Asar, Zeinab Abu-Sultan," Estimation based on Neural Networks using Face Features", Journal of Emerging Trends in Computing and Information Sciences, Vol. 1, No. 2, Oct 2010.

[2] Laura T. Germine a,⇑, Bradley Duchaine b, Ken Nakayama," Where cognitive development and aging meet: Face learning ability peaks after age 30",Elsevier, International Journal of Pattern Recognition,2010,pp1-10 .

[3] Asuman Gunay , Vasif V. Nabiyev, " Automatic Age Classification with LBP", Proc. of IEEE, International conference on Computer vision and Image Processing,2008.

[4] Feng Gao, Haizhou , "Face Age Classification on Consumer Images with Gabor Features and Fuzzy LDA method",2008 .

[5] Wen – bing Horng, cheng-Ping Lee and chun-Wen chen," classification of age groups based on facial feature", Journal of Science and Engineering, vol. 4, no.3,pp.183-192, 2001.

[6] Young Ho Kwon and Niels da vitoria Lobo, " Age classification from facial images", computer vision and image understanding, vol. 74, pp.1-21,1999.

[7] A.J.O'Toole, H. Abdi, K.A. Deffenbacher and J.CBarlett, classifying faces by race using an auto-associative memory trained for recognition, in Proc. Of Annual meeting of the Cog.Sci.Soc., Chicago,1991,pp. 847-851.

[8] A Lanitis, CJ.Taylor, and T.F.Cootes . "Towards automatic simulation of aging effects on face images ". IEEE Trans Pattern Analysis .Machine Intelligence Vol.24,no,4. pp.442-455,2002.

[9] A.Lanitis, C. Dragannva, and C.Christodoulou, Comparing different classifiers for automatic age estimation. IEEE Transactions on Systems, Man and Cybernetics, vol.34(1):pp 621-628, 2004.

[10] Karl Ricanek, Tamirat Tesafaye," MORPH:A longitudinal image database of normal adult age-progression", International conference on Automatic face and Gesture Recognition",2006. pp 341-345.

[11] X.Geng,Z, H.Zhou, and K.Smith-Miles ." Automatic age estimation based on facial aging patterns", IEEE Transaction on Pattern Analysis Machine Intelligence . vol.29,pp:2234-2240,2007.

[12] Ye Sun, Jian – Ming Zhang, Liang-Min Wang,Yong-Zhao Zhan and Shun-Lin Song, " A Novel Method of Recognizing Ageing Face based on EHMM, " Proceedings of fourth International Conference on Machine Learning and Cybernetics. Pp 4599-4603, August 2005.

[13] Allison C Lamont, Steve Stewart – Williams and John Podd, "Face Recognition and ageing : Effects of Target Age and Memory Load", Journal of Memory and cognition, Vol 33 (6),pp. 1017 -1024,September 2005.

◈◈◈

Model Identification & Fuzzy Controller Implementation For the Evaporator Process in Sugar Industry

Raghul. R, Syed Mohaideen Shahidh. H, Luca Nelson. S & Ajith. B. Singh

Department of Electronics & Instrumentation Engineering, Sethu Institute of Technology, Tamil Nadu, India
E-mail : sethuraghul@gmail.com, shahidhdhubasi@gmail.com, lucanelshaa@gmail.com, ajith.b.singh@gmail.com

Abstract **-** Modeling of vapour flow is one of the most common problems in the process industry. In this paper, model identification process is attempted for evaporator process. It is a classic example of a nonlinear system. For nonlinear systems, model identification provides a better alternative to find their system transfer function. First principle modeling of vapour flow is done with so much of approximation. Model identification procedure gives almost accurate model of the process. Steam inflow rate is taken as the input data and vapour outflow rate is taken as the output for the evaporator process. Both, the inflow steam and the outflow vapour are measured using an electromagnetic flow meter. Collected data is fitted to the transfer function model set. The simulation is performed using MATLAB environment in System Identification toolbox. We have found practically the model of the evaporator process using the MATLAB environment gives the result almost identical to the ideal model of the plant. The model identification process is used here to identify the system's nature using the inputs and the outputs available. After the identification process, PID, auto tuning PID and fuzzy logic controllers are implemented.

Keywords **-** *Process control; nonlinear dynamics; nonlinear process control; System identification process; fuzzy control.*

I. INTRODUCTION

The basic control problem is to regulate the liquid flow rate in one tank by varying the speed of the circulating pump. The apparatus consist of three stage evaporator tank, inter connected by a catches, A electromagnetic flow meter is connected at the input & output of the tank variable area valve in this Channel is used to measure the inflow & outflow rate of the tanks. A pump which delivered the wet steam of 130˚C came out from the turbine to the evaporator tank. Catches are used to maintain the latent heat of the steam to the next stages. At the output side, we get the required syrup at 60 BRIX & the steam rate is around 60 m^3/hr.

Model identification involves building mathematical models of a dynamic system based on a set of measured stimulus and response data samples. Here we focus on the transfer function model design of the evaporator tank process.

Simulation is an inexpensive and fast way to practice many problems unimaginable in laboratory experiments. Simulation does not replace laboratory experiments. Analysis of plant structure and dynamics is better done with a real plant, but with the simulation experience and the already familiar control panel, students need only about 1/4 of the time to complete an experiment. This also shows that simulation experience is transferable to real-life problems.

II. EVAPORATOR PROCESS

The main function of evaporator is concentrating on clear juice containing around 83-86% of water and getting it transforms into syrup containing 35-40% of water. At the time of process gets initiated condensation by water is started and vacuum is developed. Exhaust valve is opened for vapour flow. The gas valves are opened and incondensable gases are removed in each calandria by using pumps. The juice is passed to all the calandria and the vapours from the last body are condensed in condenser through a vapour pipe. The clear juice flow is stabilized and concentrated to get syrup at 60°-65° Bx.

In this evaporator process other than the process variable like pressure and temperature are checked for their constant values. Since, the process variable flow of steam is concentrated so there is very little variation is possible it is negligible. But, the steam supply and vacuum shows no fluctuations in the process.

Fig 1: 3 Stage- Evaporator Unit

The continuous inflow of fresh reactants (and matching volumetric outflow of reactant and-product mixture) provides a thermodynamically open system, in which true steady state behavior can be sustained 'indefinitely'.

III. MODEL IDENTIFICATION

Model identification is the art and science of building mathematical model of dynamic systems from observed input-output data.

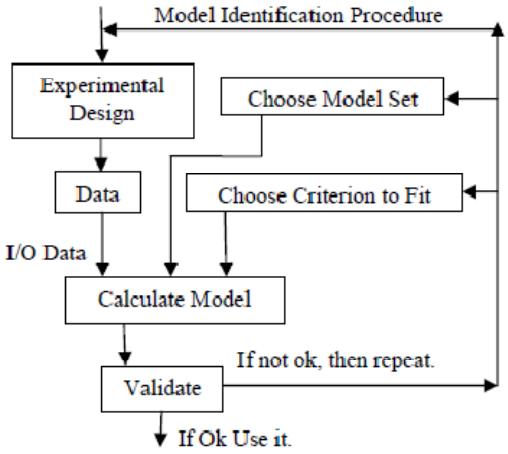


Fig. 2 : Block Diagram for Model Identification

Steps in Model Identification

The block diagram showing various steps involved in system identification are shown in fig.1.The system identification problem can be divided into a number of sub problems:

- Experimental design
- Data collection
- Model structure selection
- Model parameter Estimation
- Model validation

Experimental Planning & Data Collection

It is the basis for the identification procedure, where process experiments are designed and conducted. The purpose is to maximize the information content in the data, within the limits imposed by the process.

Model Structure Selection

The choice of appropriate model structure μ is most crucial for a successful identification application. This choice must be based both on an understanding of identification procedure and on insights and knowledge about the system to be identified.

General aspects of choice of model structure

1. *Choose the type of model set*

Selection between nonlinear and linear models, between input-output, black-box and physically parameterized state space models and so on.

2. *To choose the type of model set*

This involves issues like selecting the order of the state space model, the degree of the polynomial in a model. It also contains the problem of which variables to include in the model description. We thus have to select μ from a given, increasing chain structures

$$\mu_1 \subset \mu_2 \subset \mu_3 \ldots\ldots\ldots\ldots$$

3. *To choose the model parameterization*

When a model $\boldsymbol{\mu}^*$ has been decided on, it remains to parameterizes it, that is to find a suitable model structure μ whose range equal μ^*.

IV. MODEL STRUCTURE

The rational Transfer function model can be classified based upon linear and nonlinear regression as

FIR model	NFIR model
ARX Model	NARX model
ARMAX Model	NARAMX model
BJ model	NBJ model
Output Error method	NOE model

The structures we have discussed in this section actually may give 32 different model set based upon the five polynomials A,B,C,D, and F are used. The generalized model structure

$$A(q)y(t) = \frac{B(q)}{F(q)}u(t) + \frac{C(q)}{D(q)}e(t) \qquad (1)$$

ARX Model - The ARX model, is the simplest model incorporating the stimulus signal. The estimation of the ARX model is the most efficient of the polynomial estimation methods because it is the result of solving linear regression equations in analytic form. The following equation shows the form of the ARX model.

$$A(q)y(t) = B(q)u(t-d) + e(t) \qquad (2)$$

ARMAX Model - Model-unlike the ARX model, the ARMAX model structure includes disturbance dynamics. ARMAX models are useful when you have dominating disturbances that enter early in the process, such as at the input. The ARMAX model has more flexibility in the handling of disturbance modeling than the ARX model. The following equation shows the form of the ARMAX model.

$$A(q)y(t) = B(q)u(t-d) + C(q)e(t) \qquad (3)$$

BOX-JENKINS Model - The Box-Jenkins (BJ) structure provides a complete model with disturbance properties modeled separately from system dynamics. The following equation shows the form of the output error model.

$$A(q)y(t) = \frac{B(q)}{F(q)}u(t-d) + \frac{C(q)}{D(q)}e(t) \qquad (4)$$

The Box-Jenkins model is useful when the disturbances enter late in the process.

OUTPUT-ERROR Model-The Output-Error (OE) model structure describes the system dynamics separately. The following equation shows the form of the output error model.

$$A(q)y(t) = \frac{B(q)}{F(q)}u(t-d) + e(t) \qquad (5)$$

V. FUZZY LOGIC CONTROLLER

Fuzzy logic is a controller with decision making ability from its knowledge base. It consists of mainly three operations in it. They are Fuzzification, Knowledge base and Defuzzification.

Fuzzification performs the function that converts the input data into suitable linguistic variable. It also normalizes the measured variable within the Universe Of Discourse (UOD).Knowledge base has two components in it. They are data base and rule base. The data base is used to define linguistic variable and rule base is used to control it.

Defuzzification converts the linguistic variable into crisp solution variable. The most commonly used method of defuzzification is the 'Center of Area' method that generates the center of gravity of the final fuzzy control space. It produces a result that is sensitive to all the rules.

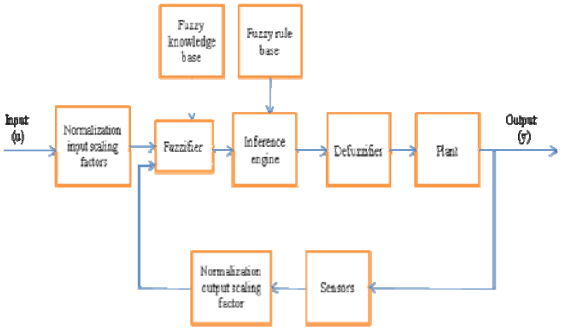


Fig 3: Fuzzy logic control System

Selection of Linguistic Variables

The fuzzy input and output variables, namely Error, Change in Error and Change in Output, are divided into seven linguistic (fuzzy) variables namely NB (negative Big), NM (Negative Medium), NS (Negative Small), ZE (Zero), PS (Positive Small), PM (Positive Medium) and PB (Positive Big).

Knowledge Base

Data Base

The membership functions for the input and output variables i.e. error, change in error and change in controller output are given.

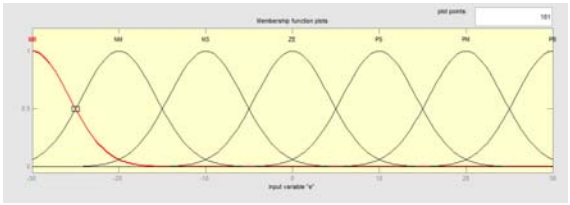


Fig. 4 : fuzzy logic Input error mf's.

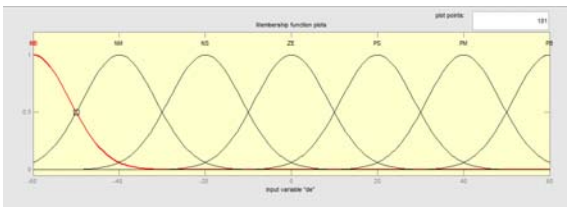


Fig. 5 : Fuzzy logic Input derivative of error mf's

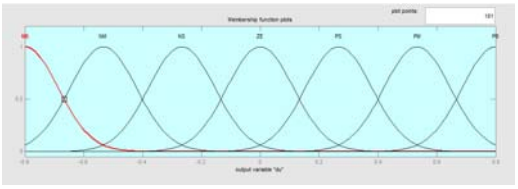


Fig. 6 : fuzzy logic output mf's

Rule Base

The rules that tie the input and output variables are given here.

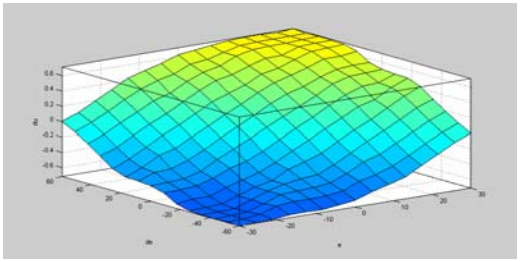


Fig 7: Surface viewer of the rules.

VI. RESULT

The steam inflow from turbine is taken as the input data and vapour outflow from calandria is taken as the output data for the evaporator process. Data collection and Identification algorithms are developed using Matlab. The fitness percentage value for each model is taken into considerations. Here the ARMAX model is selected which gives better validation fitness when compared to other models.

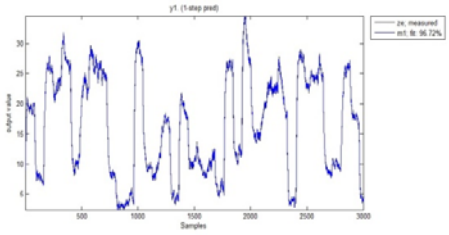


Fig 8: Estimated output

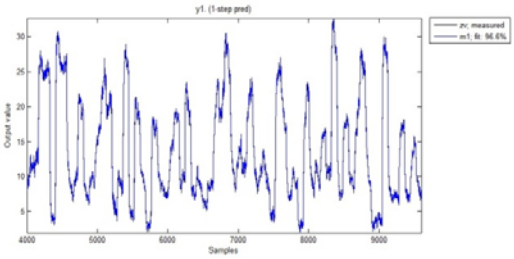


Fig 9: Validated output

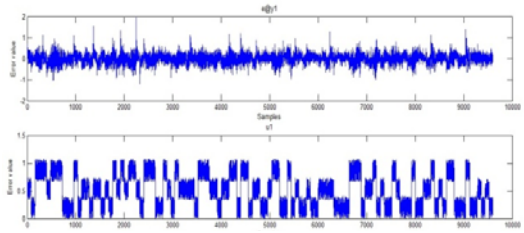


Fig 10: Error at 1-Step Prediction

The identified transfer function model of the evaporator process after various model reductions is given by:

$$Gp(s) = \frac{0.27\,S3 + 1.18S2 + 1.798S + 2.278}{S3 + 2.357S2 + 3.988S + 0.2422}$$

The PID values calculated for the process are Kp=0.14532, Ki=2.62, Kd=0.393 using zeigler Nicholas method.

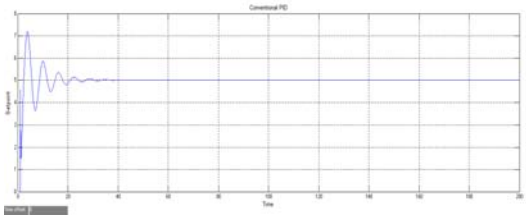


Fig 11 : Step response with PID controller.

This response shows the PID controller action for the evaporator process is shown above. It shows the evaporator process controlling action is attained at a rapid rate. This may lead to unstable condition of the evaporator process.

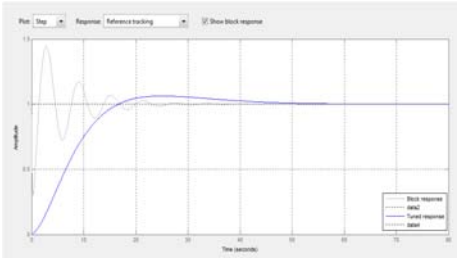


Fig 12: AUTO tuning PID Controller response.

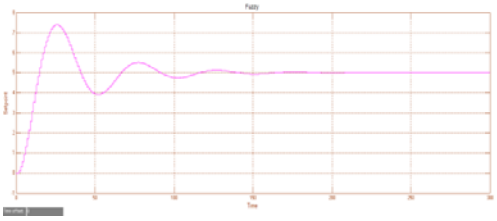


Fig 13: Fuzzy logic controller response

The fuzzy logic controller output response shows a significant improvement in the controlling action for control the process in a slow & steady rate. Thus it achieves the set point with a smooth control action which is ideal for this slow evaporator process.

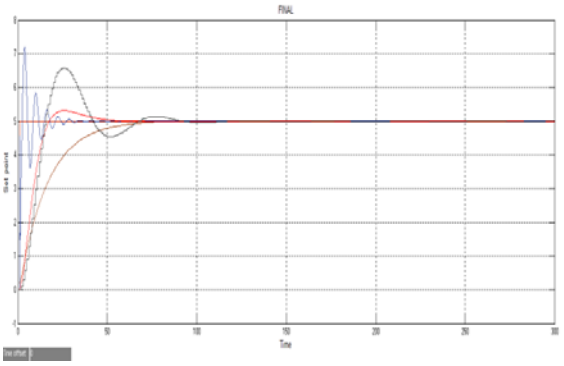


Fig 14: comparision of all controller response

VII. CONCLUSION

In this paper Model identification process is done using linear system identification models and then PID, Auto tuning PID and fuzzy logic controllers are implemented and their performance are compared. Here the evaporator process is basically a slow process it takes around 30 minutes or 1800 seconds to get the output from all the three stage of the evaporator. So in order to control this process we need to implement the controller which suits this process effectively.

The PID controller response completely settles before 45th second & Auto tuning PID controller response is also similar to the PID controller. This is not good for a slow process which takes a long time to settle, and is controlled by a controller which makes to settle at a rapid rate. Here PID controller implemented will severely affect the system performance & stability. So the fuzzy controller is implemented. The fuzzy controller takes around 200 seconds to attain the stable condition for the process. A controller can perform the control action only when the process goes beyond the set values and here the fuzzy controller action is giving a better controlling action to the evaporator system.

So an intelligent controller technique is adequate to improve the system performance and control the process in its own way by not affecting the system's stability or other parameters.

The results show an effective significance in the performance of the fuzzy logic controller. In future the same Model identification process can be done and other intelligent controller tuning methods will be implemented.

REFERENCES

[1] Model identification and model reference adaptive control implementation for a hybrid tank system.[Ajith.B.Singh ISSN: 2249-6149 Issue date: November 2011]

[2] Continuous-time model identification of a cooling system with variable delay [Tan, A.H. Cham, C.L. ISSN: 1751-8644 Issue Date: 5 May 2011]

[3] Francisco J. Arteaga, Jose R. Contreras , System Identification and Model Predictive Control for the Optimization of a Gas Sweetening Process, Escuela de Ingenieria Elecrrica University of Carabobo, Valencia, Venezuela, PDVSA, Maracaibo, Venezuela, IEEE.(Reference).

[4] Mikael Eklund a,*, Michael J. Korenberg b, P. James McLellan c, Nonlinear system identification and control of chemical processes using fast orthogonal search, Faculty of Engineering and Applied Science, University of Ontario Institute of Technology, Oshawa, Ont., Canada L1H 7K4, b Department of Electrical and Computer Engineering, Queen's University, Kingston, Ont., Canada, Journal of Process Control 17 (2007) 742–754. (Reference).

[5] G.J. Klir and B. Yuan, Fuzzy Sets and Fuzzy Logic: Theory and Application', Prentice Hall ER, New Jersey, 1995. hnp:lcsd.new~ stle.edu.aw'wnt~~simulationsl ph-siml .html.

[6] D. Driankov, H. Hellendoorn and M. Rein frank M, An infroduction to Fuzzy Control, Narcs Publishing .House, New Delhi, 1996.

[7] C.C. Lee, "Fuzzy logic in control systems: F- logic controller - Part II," IEEE Transactions on Systems, Man and Cybernetics, Vol. 20, No. 2, pp. 419435, 1990.

Books:

[1] System Identification: Theory for the user, 2nd Edition, Lennart Ljung, Prentice Hall, Inc.,1987

[2] System Identification 2nd Edition, PHP., Richard Bellman 1986

[3] Principles Of Soft Computing 2nd Edition.,Willey S.N.Sivanandam S.N.Deepa

[4] Computer Aided Process Control 3rd Edition,PHP., S.K Singh

[5] Cane Sugar Manufacture In India.,Sugar technologies association of india., D.P. Kulkarni

[6] Introduction to Fuzzy Logic using MatLab.,Springer S.N.Sivanandam S.N.Deepa

Web references

[1] www.stowann.ihe.nl/application_ ANN_Fuzzy_Logic.htm

[2] www.mathworks.com/help/toolbox/fuzzy/ fp715dup12.html

[3] www.fuzzytech.com

[4] www.sciencedirect.com

[5] www.myreaders.info/09-Hybrid_Systems.pdf

[6] www.controls.engin.umich.edu/wiki/ index.php/Image:PIDT uningClassicalRebuttaL.doc

◈◈◈

ECG Comp Decomp Using Wavelet and Intrabeat

R. Aravind Raj[1], P. Anand[2] & K. Latha[3]

[1&2]Dept of Information Technolgy, [3]Dept of Computer Science and Encgineering
Anna University of Technology, Tiruchirappalli,Trichy, India
E-mail : aravindraj25@gmail.com, anandtechnocrat@gmail.com, erklatha@gmail.com

Abstract **-** ECG is one of the most sensitive data, which we needed to compress and encode it in an efficient manner. Many telecardiology applications are mainly based on compression technique. ECG data can be obtained from patient alive monitor (a type of portable ECG monitor) and compress it with the help of proposed technique. The compressed data are sent to the hospital server and the data are reconstructed. The main purpose of this technique is to improve the compression ratio compared to the technique which is already existed. We can send the compressed data over the telecommunication networks (SMS/MMS).

Keywords **-** *ECG compression; Telecardiology; Encoding; ECG reconstruction; Compression ratio.*

I. INTRODUCTION

ECG is an electronic method of recording the heart beat of the patient. The characteristic of heart beat is determined by PQRST curve. By attaching additional leads while recording ECG signals they require higher sampling rates and finer amplitude resolution, for noncardiac signals such as blood pressure and respiration it requires huge amount of storage space. For the fast transmission of ECG signals across wireless network[1], from ambulance to hospital, telephone networks, hospital networks is not possible without compressing them. In Wireless telecardiology applications[2] ECG signal is compressed before transmission to support faster data delivery and reduce consumption of bandwidth. However, most of the ECG analysis and diagnosis algorithms are based on processing of the original ECG signal. Therefore, compressed ECG data needs to be decompressed first before the existing algorithms and tools can be applied to detect cardiovascular abnormalities.

The cost of storage devices has reduced drastically in recent times, but to store long ECG signals like 24-Hour heart rate monitoring, a very large storage capacity[1][2] is required. Again, to carry out transmission of ECG data over networks i.e., for telemedicine services the time required to transmit large files is very high. Thus, in both the cases, i.e., for the purpose of storage and for the purpose of transmission compression of ECG data is much in need. Various data compression techniques [3],[4],[8],[9] are available providing lossless and lossy compressions. The requirement of ECG signals is to apply with such compression techniques, such that, the signal should be compressed but after decompression (reconstruction), the signal must contain all the required information for the purpose of diagnosis [7] or in other words it should be in clinically acceptable form. Lossy compressors sometimes may hide such details, whereas lossless compressor generates the exactly same information as the original signal. Data compression techniques are evaluated not only in terms of compression ratio and compression and decompression bandwidths achieved but also based on their performance when the compressed data must be sent over any of the available wireless networks.

A. Heart and ECG Features

An ECG signal is the manifestation on the body surface of the myocardium electrical activity appearing as an almost periodic signal. The ECG waveform, as shown in Fig. 1, consists of three basic waves: the P, QRS, and T (Fig. 1). These waves correspond to the far field induced by specific electrical phenomena on the cardiac surface, namely, the atrial depolarization (P wave), the ventricular depolarization (QRS complex), and the ventricular repolarisation (T wave). Analyses of ECG waveforms help doctors determine many cardiac abnormalities and Arrhythmias. Accurate detection, measurement and classification of ECG complexes and their major components has always been a difficult task during computerized analysis.

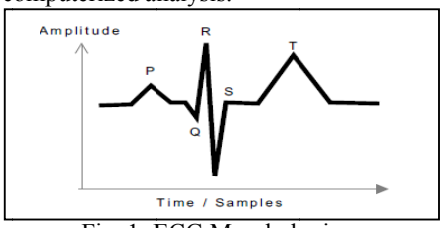


Fig. 1: ECG Morphologies

B. Large Volume of Data

The uncompressed data even from a single patient could be huge. If a patient with heart abnormality is remotely monitored, then with 12 lead ECG, 11 bit Resolution and 360 Hz Sampling Frequency the data can easily reach up to 2.77 GB in one day. When several residents are to be monitored in the same facility the amount of data generated could be tremendous. To process and transfer this large amount of data in real time is a difficult problem.

II. COMPRESSION

A. Why compression necessary

First of all, in a wireless telecardiology application, huge ECG data is transmitted from the patient to the medical server or to the doctors [2]. During this transmission of enormous ECG data, compression technology can be applied for faster transmission on limited bandwidth wireless link. At the end, faster transmission means faster treatment for the patient. Secondly, some of the messaging protocols like SMS can only allow a limited set of characters or message size. For example, SMS can only accommodate a message size of 140 bytes or 160 (140 × 8/7) characters. Since each SMS involves cost for the patient, it is imperative to transmit compressed data for economic reasons[1]. Lastly, compression algorithm also adds value to a realtime telemonitoring scenario by allowing more storage of physiological signals.

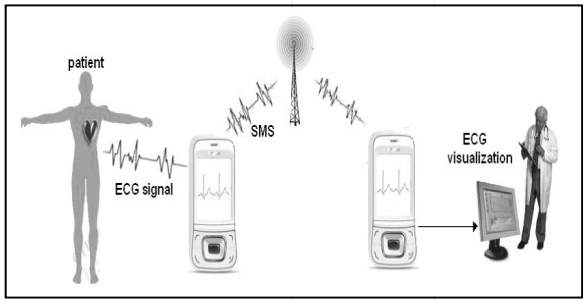


Fig. 2 : Transmission of ECG signals over the telephone network

B. Compression Techniques

1) Direct Signal Compression Techniques:

Direct methods involve the compression performed directly on the ECG signal. These are also known as time domain techniques dedicated to compression of ECG signal through the extraction of a subset of significant samples from the original sample set. The original signal is reconstructed by an inverse process, most often by drawing straight lines between the extracted samples. This category includes the FAN (Dipersio & Barr, 1985), CORTES (Abenstein & Tompkins, 1982), AZTEC (Cox et al., 1968), Turning Point (Mueller W., 1978) and TRIMS (Moody et al., 1989) algorithms. The more recent cardinality constrained shortest path technique (Haugland et al., 1997) also fits into this category. Many of the time domain techniques for ECG signal compression are based on the idea of extracting a subset of significant signal samples to represent the original signal. The key to a successful algorithm is the development of a good rule for determining the most significant samples. Decoding is based on interpolating this subset of samples. The traditional ECG time domain compression algorithms all have in common that they are based on heuristics in the sample selection process. This generally makes them fast, but they all suffer from sub-optimality.

2) Transformed Compression Methods:

Transform domain methods, as their name implies, operate by first transforming the ECG signal into another domain [4]. Signal reconstruction is achieved by an inverse transformation process. This category includes traditional transform coding techniques applied to ECG signals such as the Karhunen–Loève transform (Olmos et al., 1996), Fourier transform (Reddy & Murthy, 1986), Cosine transform (Ahmed et al., 1975), subband-techniques (Husøy&Gjerde, (1996), vector quantization (VQ) (Mammen &Ramamurthi, 1990), and more recently the wavelet transform (WT)[3] (Chen et al., 1993; (Miaou et al.., 2002). Wavelet technique is the obvious choice for ECG signal compression because of its localized and non-stationary property and the well-proven ability of wavelets to see through signals at different resolutions [3][4][11]. The wavelet decomposition splits the analyzing signal into average and detail coefficients, using finite impulse response digital filters. The main task in wavelet analysis (decomposition and reconstruction) is to find a good analyzing function (mother wavelet) to perform an optimal decomposition [11]. Wavelet-based ECG compression methods have been proved to perform well [9][10]. The discrete wavelet transform has interesting mathematics and fits in with standard signal filtering and encoding methodologies. It produces few coefficients, and the user does not have to worry about losing energy during the transform process or its inverse. While the DWT is faster and maps quickly to the sub-band coding of signals, the Continuous Wavelet Transform (CWT) allows the user to analyze the signal at various scales and translations according to the problem.

III. PROPOSED METHOD

The *wavelet transform* (WT) is a powerful tool of signal processing for its multiresolutional possibilities. The wavelet coefficients represent a measure of

similarity in the frequency content between a signal and a chosen wavelet function. These coefficients are computed as a convolution of the signal and the scaled wavelet function, which can be interpreted as a dilated band-pass filter because of its band-pass like spectrum. The *scale* is inversely proportional to radian frequency. Consequently, low frequencies correspond to high scales and a dilated wavelet function. By wavelet analysis at high scales, we extract global information from a signal called *approximations*. Whereas at low scales, we extract fine information from a signal called *details*.

In practice, the DWT is computed by passing a signal successively through a high-pass and a low pass filter. For each decomposition level, the high-pass filter h_d forming the wavelet function produces the *approximations A*. The complementary low-pass filter l_d representing the scaling function produces the *details D*. This computational algorithm is called the *subband coding*. The resolution is altered by the filtering process, and the scale is changed by either *upsampling* or *downsampling* by 2. This is described by relations.

$$D_1[n] = \sum_{k=-\infty}^{\infty} h_d[k]\, x[2n - k] \quad (1)$$

$$A_1[n] = \sum_{k=-\infty}^{\infty} l_d[k]\, x[2n - k] \quad (2)$$

Where *n* and *k* denote discrete time coefficients and *x* stands for the given signal. Half-band filters form orthonormal bases, and therefore make the reconstruction easy.

$$y[n] = \sum_{k=-\infty}^{\infty} (D_1[k]\, h_r[2k - n] + A_1[k]\, l_r[2k - n]) \quad (3)$$

The reconstruction filters l_r and h_r and identical with the decomposition filters l_d and h_d, transformed signal y(n) respectively, except the reverse time course. These filters attain to produce perfect signal reconstruction from the DWT coefficients provided that the signal is of finite energy.

Signal compression using the DWT consists of the three successive procedures, namely, signal *decomposition*, *thresholding* of the DWT coefficients, and signal *reconstruction.* We perform thresholding of the detail coefficients from level 1 to *N*. Thresholding can be either *soft* or *hard.* In hard thresholding technique, if the value of the coefficient is less than defined value of threshold, then the coefficient is scaled to zero, otherwise the value of the coefficient is maintained as it is. This process is repeated until all the amplitude values in the signal are exhausted. In soft thresholding technique, if the value of the coefficient is less than defined value of threshold, then the coefficient value is scaled to zero and otherwise the value of coefficient is reduced by the amount of defined value of threshold. This process is repeated until all the amplitude values in the signal are exhausted.

The coefficients that are received after haar wavelet transformation is taken as a input to the intra beat subtraction[6]. The coefficients of the haar wavelet transform are first normalized. During this normalization, the resultant coefficients' are multiplied with an quantization constant, *C* as in (4).

$$z(n) = y(n) \times C \quad (4)$$

The purpose of this normalization is to reduce the character size (number of digits) of each of the samples and transform the floating point values to integer values. Then, the difference between two consecutive samples is calculated using Eq. 5.

$$w(n) = z(n) - z(n - 1) \quad (5)$$

w(n) is the Differenced normalized signal. Fig 3b depicts the differenced normalized signal. Then the differenced normalized signal encoded by using the ASCII character set.

Algorithm shows encoding of differenced normalized entries. Within this algorithm *cc* holds the compressed character. The algorithm reads *q* characters at a time until the end of encoded characters. The value of *q* is the number of Differenced Normalized entries to be read by the algorithm at a time, depends on the particular character encoding. The variable *signV al* in algorithm 1 holds the decimal representation of corresponding binary value. This binary value basically represents sign of the Differenced Normalized entries, where negative is represented by 1 and positive is represented by 0.The output of the algorithm shown in Fig 3d.

Algorithm : Encoding of Data

//Variable Description:

//*cc* holds the encoded character (output of this //algorithm).

//char (*index*) returns the character for that index //value

//entry (*index*) returns the Differenced Normalized//signal of the index value.

Loop (Until the end of Differenced Normalized entries)

cc = ""

signV al = 0;

//Following codes performs the sign encoding

if(**ith entry is negative**)

signV al = 1;

cc = char (*signVal*)

//Following codes performs the value encoding

cc = *cc*+ char (entry(*i*));

endif

else

cc=*cc* + char (entry(*i*));

End Loop

In order to make the results quantitatively comparable, the most widely used numerical indexes, namely CR, SP, QS and PRD are employed in this paper. The CR [10] (Compression Ratio) is used to measure the compression efficiency, which is defined by the ratio of the bits of the original data to those of the compressed data.

$$CR = \frac{\text{original data bits size}}{\text{compressed data bits size}} \quad (6)$$

SP (Saving Percentage) is used to measure the shrinkage of the source file as a percentage. SP is shown in (7).

$$SP = \frac{sizebeforecompression - sizeaftercompression}{sizebeforecompression} \quad (7)$$

PRD (Persent root mean square difference also gives the information of the distortion rate of the reconstructed signal waveform and how the reproduced signal is compatible with the original one, PRD is shown in (8)

$$PRD = \sqrt{\frac{\sum(x(n)-x'(n))\text{^}2}{\sum x(n)\text{^}2}} \quad (8)$$

x(n) represents original signal and x'(n) represents reconstructed signal.

Quality Score (QS)[12] is used to quantify the overall performance of the compression method, considering both the CR and the error rate, Qs is shown in (9).

$$QS = \frac{Compression\ Ratio}{Present\ root\ mean\ square\ difference} \quad (9)$$

IV. EMPIRICAL RESULTS

In this section we compare the result of several experiments of our method with other ECG compression methods. The proposed algorithm was tested and evaluated using actual data from MIT-BIH arrhythmia database. These ECG data were sampled at 360 Hz with the resolution of 11 bits/sample. In this paper, the performance of the algorithm is measured and compared with other methods according to its Percent Root mean square Difference (PRD), Compression Ratio (CR) saving percentage (SP) and Quality Score (QS) for each experiment. The results are obtained through simulation by MATLAB 7.0. For better comparison of our algorithm with other methods, we test proposed method using a dataset consist of 10 seconds of data from records number 100, 101, 103, 111, 114, 200 in the MIT-BIH database[5]. The performance analysis of CR and SP are depicted in the Table 1. The Proposed method got higher compression ratios compared to the other wavelet transformation techniques so the saving percentage of the source file is increased. Table 2 shows the result analysis of PRD and QS. According to the result the proposed method possesses the PRD, Which are equivalent to the haar wavelet transform. The proposed method is combined approach of lossy compression (haar wavelet) and lossless compression. Basically the haar wavelet transform holds the lower PRD as depicted in the table 2.

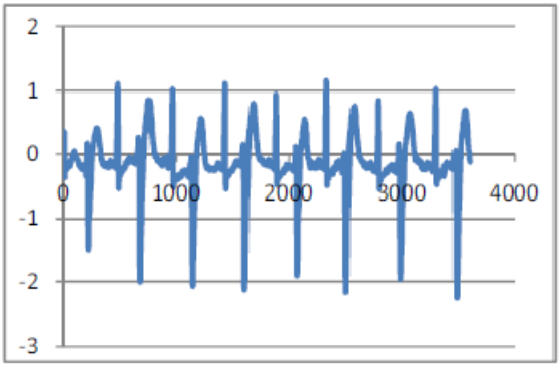


Fig. 3a : Original signal of MIT-BIH ARRHYTHMIA RECORD 200.

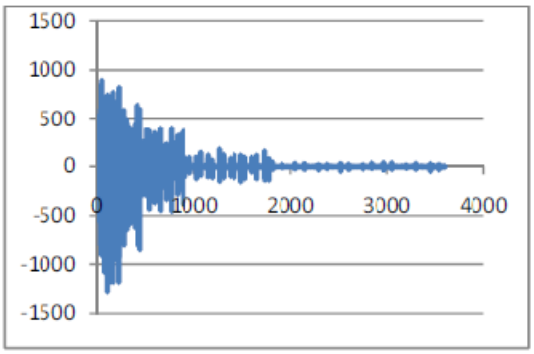


Fig. 3b : compressed coefficients of MIT BIH ARRHYTHMIA RECORD 200.

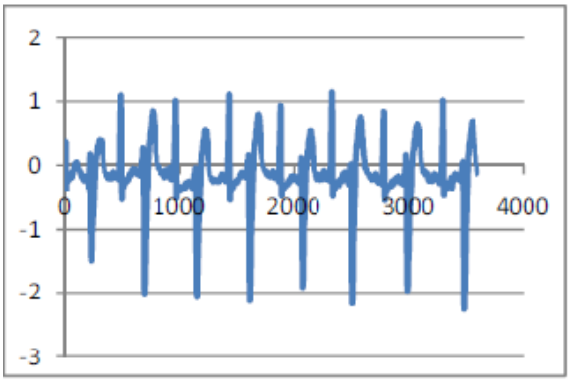


Fig. 3c : Reconstructed Signal of Proposed method (MIT BIH ARRHYTHMIA RECORD 200).

Ap.nvpr 1I Oa f i j p~Oe l06" 8"182 " 356"198 "14 "1 76 " 364 " 3084 " 1 O" nr S€O { ~bar j Oavflur ! bo:r6S 1.0;E.i; r p" t Oiht r 11.§txtnx€a!Ob€£j pa-er j PPp'pOi nT8S"1 36 " 326"244 " 2 8 "Ia ' O ~4 10" 1 7 6 " F" hJppvONr 1r pOmODt ~ xf6o c;: u nt v nO i Otvc;: r ne a pz] r r e ~ et r p Df@F" HOxnAhrt ~ €f L~22 '124" 2'.:242 " 396 "1 58"22" 250 "Om372 '40"

Fig. 3d : Encoded Signal (MIT BIH ARRHTHMIA RECORD200).

TABLE 1: PERFORMANCE COMPARISION OF COMPRESSION

MIT-BIH Arrhythmia Records (10secs)	HAAR WAVELET		DAUBECHIES-4 WAVELET		COIFLET-2 WAVELET		SYMLET-3 WAVELET		PROPOSED	
	CR	SP%	CR	SP%	CR	SP%	CR	SP%	CR	SP%
100	1.303	23.28	1.396	28.36	1.407	28.92	1.384	27.73	2.087	52.07
101	1.265	20.93	1.377	27.37	1.404	28.77	1.355	26.23	2.061	51.48
103	1.298	22.98	1.415	29.35	1.398	28.48	1.405	28.81	2.137	53.20
111	1.218	17.89	1.355	26.3	1.324	24.46	1.282	21.98	2.013	50.32
114	1.274	21.49	1.378	27.42	1.383	27.66	1.359	26.44	2.078	51.87
200	1.189	15.94	1.326	24.6	1.304	23.29	1.303	23.23	1.958	48.92

TABLE 2 : PERFORMANCE COMPARISON of DECOMPRESSION

MIT-BIH Arrhythmia Records (10secs)	HAAR Wavelet		DAUBECHIES -4 Wavelet		COIFLET-2 Wavelet		SYMLET 3 Wavelet		PROPOSED	
	PRD %	QS	PRD %	QS	PRD %	QS	PRD %	QS	PRD %	QS
100	0.666	1.957	0.732	1.907	0.737	1.909	0.731	1.893	0.666	3.134
101	0.63	2.008	0.696	1.978	0.739	1.899	0.696	1.967	0.63	3.271
103	0.675	1,923	0.704	2.009	0.693	2.017	0.693	2.027	0.675	3.166
111	1.076	1.132	1.608	1.254	1.053	1.257	1.008	1.272	1.076	1.871
114	0.864	1.478	0.915	1.506	0.897	1.542	0.884	1.537	0.864	2.405
200	0.552	2.278	0.561	2.364	0.541	2.410	0.529	2.463	0.522	3.751

V. CONCLUSION ANDFUTURE WORK

In this paper, the comparative performance of four wavelet based Algorithm with proposed technique for high-fidelity compression of ECG data were presented. The other main advantage of the proposed method is the simplicity of reconstructing the original signal. For example, in the case of the tested ECG data samples, the reconstruction error is less or equal to the haar wavelet compared to the other wavelets. The simulation results show that, proposed method provided the higher CRs.

Our future work is to apply the thresholding on detail coefficients at initial level of Haar decomposition instead of applying all the detail coefficients' after certain levels. The main motive of this approach is to reduce the error rate in reconstruction process and also incorporates the analizations of perfect threshold value which leads to increase the compression ratio.

REFERENCES

[1] F. Sufi, I. Khalil, Q. Fang, and I. Cosic, "A mobile web grid-based physiological signal monitoring system," in Proc. Int. Conf. Technol. Appl. Biomed.(ITAB), May 2008, pp. 252–255

[2] R.-G. Lee, K.-C. Chen, C.-C. Hsiao, and C.-L. Tseng, "A mobile care system with alertmechanism," IEEE Trans. Inf. Technol. Biomed., vol. 11, no. 5, pp. 507–517, Sep. 2007.

[3] B. Kim, S. Yoo, andM. Lee, "Wavelet-based low-delay ECG compression algorithm for continuous ecg transmission," IEEE Trans. Inf. Technol.Biomed., vol. 10, no. 1, pp. 77–83, Jan. 2006.

[4] R. Istepanian and A. Petrosian, "Optimal zonal wavelet-based ECG data compression for a mobile telecardiology system," IEEE Trans. Inf. Technol. Biomed., vol. 4, no. 3, pp. 200–211, Sep. 2000.

[5] Physiobank: Physiologic signal archives for biomedical research. (2009). [Online]. Available: http://www.physionet.org/physiobank/

[6] P. Hamilton and W. Tompkins. Compression of the ambulatory ecg by average beat subtraction and residual differencing. IEEE Trans. Biomed. Eng., 38(3):253–259, Mar 1991.

[7] Fahim Sufi and Ibrahim Khalil. Diagnosis of Cardiovascular Abnormalities From Compressed ECG: A Data Mining-Based Approach. IEEE Transactions on Information Technology in Biomediceine vol. 15, no.1Jan 2011.

[8] Catalina Monica Fira and Liviu Goras. An ECG Signals Compression Method and Its Validation Using NNs. IEEE Transactions on Biomedical Engineering VOl 55, NO. 4 April 2008.

[9] Bashar A. Rajoub. An Efficient Coding Algorithm for the Compression of ECG Signals Using the Wavelet Transform.IEEE Transactions on Biomedical Engineering, VOL. 49, NO. 4,April 2002.

[10] Manuel Blanco-Velascoa, Fernando Cruz-Rold´ana, J. Ignacio Godino-Llorenteb, Joaqu´ın Blanco-Velascoa, Carlos Armiens-Aparicioa, Francisco L´opez-Ferrerasa. Communication On the use of PRD and CR parameters for ECG compression. Medical Engineering & Physics 27 (2005) 798–802.

[11] Brij N. Singh , Arvind K. Tiwari Optimal selection of wavelet basis function applied to ECG signal denoising. Digital Signal Processing 16 (2006) 275–287.

[12] Hyejung Kim, Refet Firat Yazicioglu, Patrick Merken, Chris Van Hoof, and Hoi-Jun Yoo, ECG Signal Compression and Classification Algorithm With Quad Level Vector for ECG Holter System. IEEE Transactions on Information Technology in Biomedicine VOL. 14, No. 1, January 2010.

◈◈◈

An Indiscernibility Approach for Pre processing of Web Log Files

Jeeva Jose[1] & P. Sojan Lal[2]

[1]Department of Computer Applications, BPC College, Piravom, Kerala, India
[2]School of Computer Sciences, Mahatma Gandhi University, Kottayam, Kerala, India
E-mail : vijojeeva@yahoo.co.in[1], padikkakudy@gmail.com[2]

Abstract **-** World Wide Web has a spectacular growth not only in terms of the number of websites and volume of information, but also in terms of the number of visitors. Web log files contain tremendous information about the user traffic and behavior. A large amount of pre processing is required for eliminating the noise and is one of the challenging tasks in web usage mining. This paper proposes an indiscernibility approach in rough set theory for pre processing of web log files.

Keywords **-** *Web Logs, Web Usage Mining, Rough set, Preprocessing.*

I. INTRODUCTION

Web usage mining is the extraction of information from web log files generated when a user visits the web site [1][2][3].Web usage analysis includes straightforward statistics such as page access frequency as well as more sophisticated forms of analysis like clustering, classification and association rule mining. Access log files contain large amount of HTTP information. Every time a web browser downloads an HTML document from the internet, the images are also downloaded and stored in the log file [4]. A *hit* is any file from a web site that a user downloads and *accesses* are an entire page downloaded by users regardless of the number of images, sounds or movies [5]. A hit can be a text document, image, movie or a sound file. If a user downloads a web page that has 6 images on it, then the number of hits is 7(1 text page + 6 images) whereas the user has accessed just one page of the web site. Most web analysis software counts the number of *hits* a server receives rather than the number of *accesses*. In order to assure that only the data contain useful information for the users, it is absolutely necessary for the log files to be cleaned and filtered.

II. WEB LOG FILES

Web logs are maintained by web servers and contain information about users accessing the site. Logs are mostly stored as simple text files, each line corresponding to one access. The most widely used log file formats are Common Log File Format and Extended Log File Format [6] [7] [8]. Traditionally there are four types of server logs [17].

A. Transfer log

The Transfer log or Access log provides the greatest amount of server data, including the date, time, IP address and user action. The information like percentage of users accessing the site from a specific domain type (.com, .edu etc), the number of unique IP addresses accessing the site, number of accesses the server receives during specific hours and days of the week and the path a user take through a site. Such analysis enables server administrators to characterize the server load and usage pattern.

B. Agent Log

The Agent log provides data on a user's browser, browser version and operating system. This is a significant information, as the type of browser and operating system determine what a user is able to access on a web site. These data are essential for the design and development of web sites.

C. Error Log

The average web user will receive an error "Error 404 File Not Found" message several times a day. When a user encounters this message, an entry is made in the Error Log. It also contains information about stopped transmission. This occurs when a user click on the "Stop" button while downloading a file. This information can indicate patterns with large files (eg, movies and images) that users consistently stop downloading. The analysis of error log can also provide important server information such as missing files, erroneous links and aborted downloads.

D. Referrer Log

The Referrer Log indicates what other sites on the web link to a particular server. Each link made to a site generates a Referral log entry. Such referral data is critical to alleviate missing link data, path reconstruction etc.

The referrer and agent logs may or may not be turned on at the server or may be added to the transfer log file to create an extended log file format. There are three main sources to get the raw web log file [18]. They are the a) client log file, b) Proxy log file and c) Server log file. Client log files are most accurate and authentic to depict the user behavior but it is a difficult task to modify the browser for each client. In proxy servers, same IP address is used by many users and hence to identify users is difficult. Most researchers consider the web server log file as most reliable and accurate for web usage mining process. The Web log contains the following information: a) user's IP address b) user's authentication name c) the date-time stamp of the access d) the HTTP request e)the URL requested f) the response status g)the size of the requested file and optionally also h) the referrer URL i)the browser and its version and j) the operating system.

Usually there are three ways of HTTP requests namely GET, POST and HEAD. Most HTML files are served via GET method while most CGI functionality is served via POST or HEAD. The status code 200 is the successful status code. Table 1 shows various status codes of Hyper Text Transfer Protocol [9].

III. DATA CLEANING AND PREPROCESSING

The first pre processing task is data cleaning. The process of data cleaning is to remove noise or irrelevant data. Web server access logs represent the raw data source.

101	Switching Protocols	404	Not Found
200	OK	405	Method Not Allowed
201	Created	406	Not Acceptable
202	Accepted	407	Proxy Authentication Required
203	Non Authoritative Information	408	Request Time Out
204	No Content	409	Conflict
205	Reset Content	410	Gone
206	Partial Content	411	Length Required
300	Multiple Choices	412	Precondition Failed
301	Moved Permanently	413	Request Entity Too Large
302	Moved Temporarily	414	Request URL Too Large
303	See Other	415	Unsupported Media Type
304	Not Modified	500	Server Error
305	Use Proxy	501	Not Implemented
400	Bad Request	502	Bad Gateway
401	Unauthorized	503	Out of Resources
402	Payment Required	504	Gateway Time Out
403	Forbidden	505	HTTP Version not Supported

Table 1 : Status Codes of Hypertext Transfer Protocol

It is important to identify and discard the data recorded by web robots or web crawlers, the images, sounds, java scripts etc that is often redundant and irrelevant [10]. The problems identified in pre processing of log files are

- The need for a large storage space due to considerable volume of data saved on disk.
- The existence of large amount of data that is irrelevant for the web mining process containing images, sounds etc.
- The storage of requests performed by search engines and various automated scripts.
- The storage of data containing error messages such as 400(bad request).

The advantages of pre processing are

- The storage space is reduced as only the data relevant to web mining is stored.
- The search engine requests and image files are removed so that the precision of web mining is improved.

IV. INDISCERNIBILITY RELATIONS AND ROUGH SET THEORY

Indiscernibility relations in rough set theory [11] can be used for the data cleaning of web log files. Rough set is based on the assumption that with every object of the universe of discourse, some information is associated. Objects characterized by the same information are indiscernible (similar) in view of the available information about them. Any set of all

indiscernible (similar) objects is called an elementary set and forms a basic granule of knowledge about the universe. Any union of some elementary sets is referred to as crisp (precise) set otherwise the set is rough (imprecise, vague).

Let a given pair S= (U,A) of non–empty finite sets U and A, where U is the Universe of objects and A is the set consisting of attributes. The function a: U→ V_a , where V_a is the set of values of attribute a, called the domain of a. The pair S=(U,A) is called an information system. Any information system can be represented by a data table with rows labeled by objects and columns labeled by attributes. Any pair (x, a) where x □ U and a □A defines the table entry consisting of the value a(x).

Any subset B of A determines a Binary relation I(B) on U, called an indiscernibility relation defined by xI(B)y if and only if a(x)=a(y) for every a □ B, where a(x) denotes the value of attribute a for object x. I(B) is an equivalence relation. The family of all equivalence classes of I(B) will be denoted by U/I(B) or simply U/B. Equivalence classes of the relation I(B) of the partition U/B are referred to as B-elementary sets or B-elementary granules. Let U represents the set of the set of all user sessions with the path traversed. Let A be the subset of U which represents the HTTP request, URL requested and status code. An indiscernibility relation I(B) is defined for every a(x)=a(y) [12].

A rough set based feature selection for web usage mining is used in [13]. Feature selection is a pre processing step in data mining and is very effective in reducing dimensions

The indiscernibility relations in rough set theory is used for clustering in [14].

V. CONSTRUCTION OF EQUIVALENCE CLASSES FROM WEB LOGS

We consider the subset of web log file entry which includes <HTTP request, URL requested, status code> for indiscernibility relation and is denoted by B. Based on the values of these attributes, the family of all equivalence classes is generated and it is denoted as U(B) . Out of the equivalence classes generated, the indiscernibility relation having the attribute values<GET, *.html, 200> and <GET, */, 200> are the equivalence classes which is useful for further data mining to follow. The advantage of this method is that all other equivalence classes generated can be further used to perform other statistical analysis like the percentage of bad requests, mean number of redirects, percentage of reduction in the size of log file etc.

A. *Sample Code*

1. *Extract the fields from a raw log file to a spreadsheet or a data base.*
2. *Eliminate the request by bots, spiders and crawlers.*
3. *Generate Equivalence classes considering the attributes <HTTP Requested, URL Requested, Status Code>*
4. *Filter the equivalence classes with values* <GET, *.html, 200> and <GET, */, 200>
5. *Eliminate all other fields excluding IP, date, time and URL requested.*
6. *All other equivalence classes can be saved in another file for statistical analysis if required.*

We have used raw log file of a business organization called NeST ranging from January 1, 2011 to March 31, 2011. A sample output after cleaning and pre processing is shown below in Table 2.

IP	DATE	TIME	URL REQUESTED
117.196.136.242	01-Jan-11	00:45:59	/
117.196.136.242	01-Jan-11	00:46:19	/index.php?option=com_content&task=view&id=70&Itemid=72
219.64.67.45.hyd-cdma.dialup.vsnl.net.in	01-Jan-11	00:55:02	/nestit/
219.64.67.45.hyd-cdma.dialup.vsnl.net.in	01-Jan-11	01:05:05	/nestit/services.html
host-196.218.41.130-static.tedata.net	01-Jan-11	01:16:34	/
115.113.10.228.static-bangalore.vsnl.net.in	01-Jan-11	02:33:21	/index.php?option=com_content&task=view&id=6&Itemid=9
91.201.66.28	01-Jan-11	02:34:58	/
91.201.66.135	01-Jan-11	03:33:18	/
hosted-by.altushost.com	01-Jan-11	07:17:28	/index.php

Table 2 : Cleaned Web Log in Common Log File Format

B. User and Session Identification

A time out of 30 minutes is used to identify a user session [15]. Each request from the same IP before 30 minutes is considered as the same user. If the log file is an extended log file, the browser and its version, operating system etc are also considered for identifying users and sessions. After user session identification, if the task to be performed is association rule mining or frequent itemset mining, the log file is further pre processed to obtain records in the form $<T_{id}, p_1, p_2, \ldots p_n)$, where T_{id} represents the transaction id and $p_1, p_2, \ldots p_n$ represents the pages requested. Table 3 shows the statistics of the weblog data before and after preprocessing.

Total number of records	8,68,598
Number of records after removing search engine visits	7,75,738
Number of records after pre processing	60,606
Percentage in reduction	93.02%
Total number of users	25,015

Table 3 Statistics of web log data before and after pre processing

C. Prefetching and Caching

If the mining of web logs is performed for prefetching and caching of web pages in proxy servers, then the pre processing is done in a different way. The web objects like images, sounds, movies can not be eliminated. References to embedded objects are usually preceded by their HTML container[16]. If the status code is successful, then the number of bytes transferred is important for caching and prefetching purposes.

D. Sample Code

1. *Extract the fields from a raw log file to a spreadsheet or a data base.*
2. *Eliminate the request of bots, spiders and crawlers.*
3. *Generate Equivalence classes considering the attributes <HTTP Requested, URL Requested, Status Code>*
4. *Filter the equivalence classes with values* <GET, * 200>
5. *Compute the total bytes transferred including the embedded objects for each container page.*
6. *Eliminate all other fields excluding container page and total bytes transferred.*
7. *Calculate the frequency of each container page.*

VI. CONCLUSION

Rough set theory has found applications in many domains like engineering, environment, banking, medicine and others. This paper is an attempt to highlight another application of rough set theory, which is the pre processing of web log files. Pre processing of web log files is very crucial as it may bias the data mining tasks to follow. The advantage of this method is that we will get different equivalence classes which can be further used for statistical analysis including prediction. The pre processed file can be used for various web usage mining tasks.

REFERENCES

[1] M. F. Facca & P. L. Lanzi, ".Mining interesting knowledge from web logs: a survey", Data & Knowledge Engineering, 2004,Vol. 53, pp.225-241.

[2] R.Cooley, B. Mobasher and J. Srivastava," Data Preparation for Mining World Wide Web Patterns",. Knowledge and Information Systems, 1999, Vol 1(1), pp.5-32.

[3] J.Pitkow, "In Search of Reliable Usage Data on WWW",Sixth International WWW Conference,1997.

[4] R.M. Suresh & R. Padmajavall,"An Overview of Data Preprocessing in Data and Web Usage Mining",First International Conference on Digital Management, Banglore, India,IEEE, 2006,pp.193-198.

[5] M.C.Burton & B. J.Walther,"A Survey of Web Log Data and Their Application in Use-Based Design", 34th Hawaii International Conference on System Sciences, Maui, Hawaii, IEEE,2001,pp.1-10.

[6] M.Grcar,"User Profiling: Web Usage Mining",7th International Multiconference Information Society, Ljublijana,Slovenia,2004.

[7] A. H. M.Wahab,H.N.M.Mohd,F.H.Hanaf & M.F.M.Mohsin," Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm",World Academy of Science, Engineering and Technology,2008, pp.190-197.

[8] M. Spiliopoulou, "Web Usage Mining for Web Site Evaluation", Communications of the ACM, 2000, vol. 43(8), pp.127-134.

[9] Internet: Hypertext Transfer Protocol Overview,http://www.w3.org/protocols. Last retrieved October 2011.

[10] D. Mican & D. Sitar-Taut," Preprocessing and Content/ Navigational Pages Identification as Premises for an Extended Web Usage Mining Model Development", Informatica Economica, 2009,vol. 13(4),pp.168-179.

[11] Z. Pawlak,"Rough Set Theory and its Applications", Journal of Telecommunications and Information Technology, 2002,pp.7-10.

[12] Z. Pawlak & A. Skowron, "Rudiments of Rough Sets", Information Sciences, 2007, vol. 177(1), pp.3-27.

[13] H.H. Inbarani, K. Thangavel & A. Pethalakshmi,"Rough set based Feature Selection for Web Usage Mining",International Conference on Computational Intelligence and Multimedia Applications, Sivakasi, India,IEEE,2007,33-38.

[14] S. Hirano & S. Tsumoto, "An Indiscernibility Based Clustering Method", IEEE International Conference on Granular Computing, Izumo, Japan, IEEE.2005, pp. 468-473.

[15] Z. Pabarskaite, & A. Raudys, "A process of knowledge discovery from web log data: Systematization and critical review", Journal of Intelligent Information Systems, 2007, Vol.28, pp. 79-104.

[16] Qiang Yang, H. H. Zang & Tianyi Li, "Mining Web Logs for Prediction Models in WWW Caching and Prefetching", Knowledge Discovery in Databases, San Fransisco, 2001,pp. 473-478.

[17] J. C. Bertot, C.R. Mcclure, W. E. Moen & Jeffrey Rubin, "Web Usage Statistics: Measurement Issues and Analytical Techniques", Government Information Quarterly, Jai Press Inc.,1997, Vol. 14 (4), pp. 373-395.

[18] T. Hussain, S. Asghar and N. Masood, "Web Usage Mining: A Survey on Preprocessing of Web Log File", International Conference on Information and Emerging Technologies, IEEE, 2010, pp. 1-6.

◈◈◈

Speech Recognizing Powered Wheelchair for Disabled

D. K. Prabitha, Chidananda Murthy. M. V & M. Z. Kurian

Dept. of Electronics and communication, Sri Siddhartha Institute of Technology Tumkur, Karnataka, India
E-mail : dkprabitha@gmail.com, chidussit@yahoo.com, mzkurianvc@yahoo.com

Abstract **-** Freedom of mobility is a dream for every person with physical disability especially in the case of quadriplegics, those people who are paralyzed below neck and in case of multiple sclerosis patients. Though many types of mobility equipments are available for the disabled for their independent mobility, no efficient devices have yet been invented for the independent mobility of quadriplegic and multiple sclerosis patients. The quadriplegic and multiple sclerosis patients cannot drive a joystick operated wheelchair which is the only type of powered wheelchair which is commercially available. Quadriplegia and multiple sclerosis is cureless.

This paper describes the sincere devotion to design a powered wheelchair with particular features that helps the quadriplegic patients and multiple sclerosis patients move independently. This powered wheelchair is controlled by the speech of the user. This wheelchair can be easily driven to desired direction with minimum effort. The user requires minimum training to use this mobility equipment. Technically the wheelchair is integrated with a voice recognition circuit to recognize the voice, a microcontroller which can be programmed other supporting hardware components. This wheelchair named "COMRADE" will be a helping aid for quadriplegic patients and other impaired people, who have a similar grade of disability.

Keywords **-** *Quadriplegic ;Multiple Sclerosis;Voice recognition circuit; Microcontroller.*

I. INTRODUCTION

Physically challenged people have either less mobility or no mobility. Freedom of mobility is an important factor that each and every individual seeks. Quadriplegics is a kind of cureless disability and quadriplegic patients are people those who have lost the hope to a normal life. Those patients cannot drive a joystick operated powered wheelchair. As his physical condition degrades day by day, he may lose the strength and control of hands to operate a joystick operated powered wheelchair. A powered wheelchair with a speech recognition system, thus can lend a transition in the life of such disabled.

The objective of this work is to design and develop a new technology for the smooth operation of electric wheelchairs through speech recognition.

Mobility is restricted or absent at extreme of quadriplegic and multiple sclerosis patients. Multiple Sclerosis (MS) is an inflammatory disease of the Central Nervous System (CNS) - that's the brain and spinal cord. Predominantly, it is a disease of the "white matter" tissue. There is as yet no cure for MS. Many patients do well with no therapy at all, especially since many medications have serious side effects and some carry significant risks. These results indicate a need for entirely new technology for the independent mobility of such patients.

One of the best solutions to this problem is a speech recognizing powered wheelchair that the input to the equipment is voice commands. A microphone receives the voice input and the processor converts this to specific motion of the wheelchair. With this regulated motion of wheelchair most of the lost mobility can be retained.

The project aims at designing and developing a microcontroller based voice controlled electric wheelchair with a voice recognition system to drive the wheelchair using voice commands.

A wireless audio transmission unit is also integrated in this wheelchair which can be controlled wirelessly while sitting on the wheelchair or within the range of 50metres which is an additional feature of this equipment. The main circuit units that are integrated in this wheelchair are:

- Transmitter
- Receiver
- Voice Recognition circuit
- Microcontroller based Control circuit
- Isolation circuit
- Relays
- Drives on the Wheelchair

II. BACKGROUND

Several studies have shown that both children and adults benefit substantially from access to a means of independent mobility. While the needs of many individuals with disabilities can be satisfied with manual or powered wheelchairs, a segment of the disabled community finds it difficult or impossible to use wheelchairs independently. To accommodate this population, researchers have used technologies originally developed for mobile robots to create smart wheelchairs. [2]

Smart wheelchairs have been the subject of research since the early 1980s and have been developed on four continents. Several studies have shown that both children and adults benefit substantially from access to a means of independent mobility, including power wheelchairs, manual wheelchairs, scooters, and walkers. Independent mobility increases vocational and educational opportunities, reduces dependence on caregivers and family members, and promotes feelings of self-reliance. While the needs of many individuals with disabilities can be satisfied with traditional manual or powered wheelchairs, a segment of the disabled community finds it difficult or impossible to use wheelchairs independently. This population includes, but is not limited to, individuals with low vision, visual field reduction, spasticity, tremors, or cognitive deficits. These individuals often lack independent mobility and rely on a caregiver to push them in a manual wheelchair.

A. Tongue operated electric wheelchair

The Tongue Drive system was described on June 29 at the 2008 Rehabilitation Engineering and Assistive Technology Society of North America (RESNA) Annual Conference in Washington, D.C.

The novel system allows individuals with disabilities to operate a computer control a powered wheelchair and interact with their environments simply by moving their tongues.

Patients use their tongue to operate the system because unlike hands and feet, which are controlled by the brain through the spinal cord, the tongue is directly connected to the brain by a cranial nerve that generally escapes damage in severe spinal cord injuries or neuromuscular diseases. Tongue movements are also fast, accurate and do not require much thinking, concentration or effort.

The system can potentially capture a large number of tongue movements, each of which can represent a different user command. A unique set of specific tongue movements can be tailored for each individual based on the user's abilities, oral anatomy, personal preferences and lifestyle.[1]

Fig.1: Tongue operated electric wheelchair

B. Electric wheelchairs with split frame technology

World's first split frame technology was invented by Ostrich Mobility, a Bangalore based wheelchair manufacturing company. Unlike normal suspension systems, the Split Frame Suspension technology allows all the four wheels in contact with ground even when there is a level difference of six inches.

In conventional chassis electric wheelchairs, the wheels lift up when there is a level difference on the ground or floor the wheelchair is running. This can cause loss of control leading to accidents. The split frame chassis makes sure that all the four wheels touch the ground with a level difference up to 150mm or more thus giving a better drive control and better maneuverability in tough terrain conditions.[5]

Fig. 2: Picture of Split frame e-wheels

III. SYSTEM DESIGN

Designing a wheelchair is not a trivial job but designing them in a convenient manner for the disabled is a tedious job. If the features of the wheelchair happen to be inconvenient to the user it would be a disadvantage for them. The features of the wheelchair include:

A. Effortless Operation:

The user operating the electric wheelchair contributes less effort to drive the wheelchair or the wheelchair should be easy to operate. The user should

get control over the equipment so as to drive in the desired direction

B. Safe Speed

The speed range of Electric wheelchair is from 0.5 Km to 15 km/hr depending on different manufacturers. In this project since the equipment is driven by voice the speed is kept on the lower side. The decided maximum speed is 2 Km/hr so that there will be sufficient time to deviate from a potential hazard.

C. Favorable Power

The Electric wheelchair should consume less power to get a longer drive range. The electrical parts and components selected for designing the wheelchair should consume less power. Reduction in the weight of other parts like the frame, the seat etc will also help to reduce the effort of motors thereby increasing the drive range.

D. Compression

Compression of a wheelchair improves the ability to manipulate in tight indoor conditions. Trying to design a compressed wheelchair will also reduce the weight of the wheelchair thus improving the drive range and also the portability of the system. Care should be taken to have space for all the electronic units to be packed inside the wheelchair.

E. Ruggedness

The wheelchair should be capable to carry a designed load upto 100kg. therefore the frame design should be rugged enough to carry the load and run at a designed maximum speed of two kilometers per hour. The ruggedness of the frame can be brought in by using superior light weight materials .

F. Effortless service

The parts and components of the wheelchair should be easily accessible. It should also be easily removable with the help of minimum tools. Reduced number of parts and components helps serviceability of the wheelchair.

G. Less Expensive

The Optimized Design of the equipment reduces the cost. Optimization of the designing can be done by selecting the suitable material for the desired output. By avoiding unwanted features beyond design intend also reduces cost of the equipment.

H. Environmental friendly

The equipment should be designed not to pollute the environment. The parts and components of the equipment should be made of biodegradable materials or the parts should be 100 percent recyclable.

I. Ease of Transport

The weight of the equipment should be as low as possible as the compressed size of the equipment also helps to transport the equipment with ease.

J. Major parts of electric wheelchair

- Motor controller : Microprocessor based PWM controller
- Battery pack : Consists of 12 Volt 17 Ah battery
- Battery charger : 230-40 Volt Ac input and 13.6 Volt Dc out put
- Seating : Cushion type
- Human interface: Voice activated
- Foot rest , arm rest and
- Tyres

This wheelchair is driven and fully controlled by using voice commands. A wireless audio transmission unit is integrated to the wheelchair so that the patient can use it while sitting on the wheelchair or within a range of 50 meters away from the wheelchair, which is an additional and optional feature of the system. The voice recognition unit can recognize 20 different voice commands that can be set to various movements of the wheelchair. The output of voice recognition unit is the input to the motor controller unit.

The motor controller output voltage to the motors varies according to the various voice input commands. The motors that are connected to the wheels make the wheelchair move in the direction proportional to the various voice commands.

The main circuit units integrated to the wheelchair is:

- Voice Reception unit
- Voice Recognition circuit
- Microcontroller based Motor Control circuit
- Relay bank
- Drives on the Wheelchair

K. Voice reception unit

This unit consists of an audio transmitter and an audio receiver. This unit operates in a range of 50metres. This unit delivers the voice commands from the person to the voice recognition unit.

L. Voice recognition unit

This unit recognizes the voice command from the voice reception unit. IC HM2007P is the main component of this Voice Recognition circuit. This IC can recognize 20 voice different commands. This Voice Recognition circuit produces an 8-bit digital output for each voice commands.

M. Micro controller based controller unit

The 8-bit digital output obtained from the voice recognition circuit is used to drive a microcontroller based control circuit. A PIC16F877 microcontroller is used in this circuit. The microcontroller is programmed in such a way to produce the required outputs for corresponding voice commands.

N. Relays

Relays are used to switch ON and OFF the motors of the wheelchair according to the output of motor controller unit 8 relays are used.

O. Drives on the wheelchair

The drives used in the wheelchairs are PMDC motors. Two motors are used to drive the wheelchair. They are 12V, 30A brushed DC motors.

P. Power supply

Battery is the power source. The UPS can be charged from the 230V, 50Hz power supply which is available commonly. There a two 12V, 7Ah battery used to supply the power. The motors are driven directly from the battery.

Q. Battery charger

A battery charger is used to recharge the battery. The charger consists of an AC to DC circuit. The input is 230-240V AC and the output is 13.6 Volts DC at 2 ampere rating.

R. Download pattern

The download pattern process is same as the upload pattern process except that the direction of the data flow is reversed. After receiving the download command and the word number, HM2007 being to read data from external device. The first two words of the data will be treated as the pattern length and following data will be stored as pattern frame by frame. Figure shows the control flow of the downloading process

S. Reset

When Reset command is received by HM2007, the HM2007, the will clear the entire pattern in the memory.

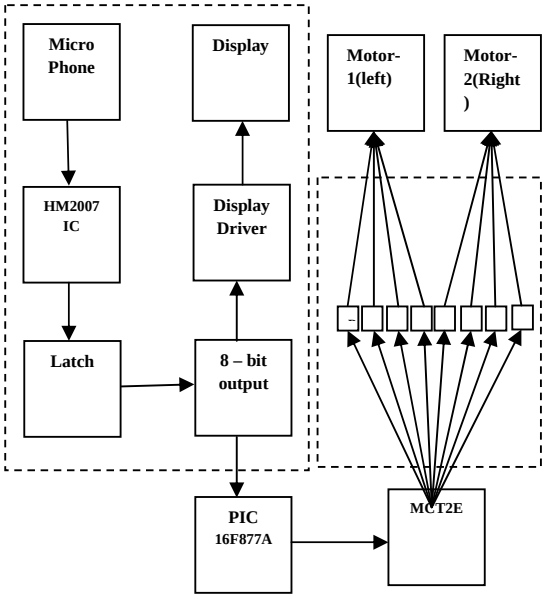


Fig 3: Block Diagram of Voice Regulated powered Wheelchair for Disabled

IV. SYSTEM IMPLEMENTATION

The required hardware has been integrated and the programmable microcontroller is programmed and tested.

A. Major parts of electric wheelchair

- Frame: The frame is made up of mild steel tubes welded to form a space frame.
- Drive motors: These are 12 Volt 30 Amp peak DC permanent magnet brushed type
- Motor controller : Microprocessor based PWM controller
- Battery pack : Consists of 12 Volt 17 Ah battery
- Battery charger : 230-40 Volt Ac input and 13.6 Volt Dc out put

Seating : Cushion type

Human interface: Voice activated

Foot rest , Arm rest and Tyres

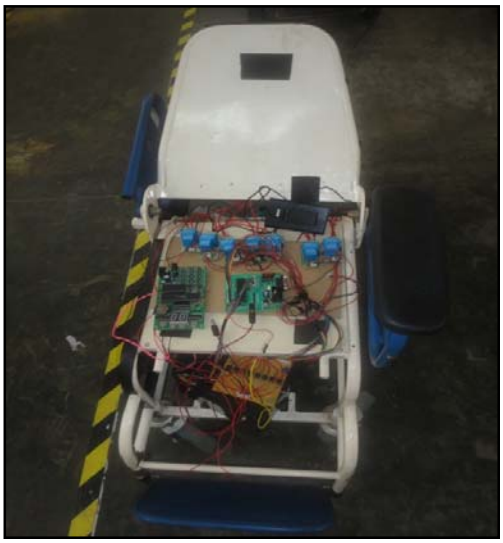


Fig 4 : Speech recognizing powered wheelchair for disabled – electronic components

B. Voice reception unit

This unit consists of an audio transmitter and an audio receiver. This unit also operates in a range of 50metres which is optional. This unit delivers the voice commands from the person to the voice recognition unit.

C. Voice recognition circuit

The voice recognition unit recognizes the voice command from the voice reception unit. The IC HM2007P is the main component of this Voice Recognition circuit. This IC can recognize 20 different voice commands. The voice commands can be in any different languages. This Voice Recognition circuit produces a 8-bit digital output for each voice commands fed to it.

D. Microcontroller based controller circuit

The 8-bit digital output obtained from the voice recognition circuit is used to drive a microcontroller based control circuit. A PIC16F877 microcontroller is used in this circuit. The microcontroller is programmed in such a way that it produces the required outputs for corresponding voice commands. The microcontroller is programmed using the C language.

E. Relay bank

Relays are used to switch ON and OFF the motors of the wheelchair according to the output of motor controller unit. 8 relays are used.

F. Drives on wheelchair

The drives used in the wheelchairs are PMDC motors. Two motors are used to drive the wheelchair. They are 12V, 30A brushed DC motors.

V. ACKNOWLEDGMENT

I am deeply indebted and I would like to express my sincere thanks to our beloved Principal Dr. K.A.Krishnamurthy, for providing me an opportunity to do this project.

My special gratitude and sincere thanks to Dr M.Z.Kurian,Head of the Department, E&C,S.S.I.T for his encouragement and wholehearted support.

My special gratitude and sincere thanks to my guide Mr Chidananda Murthy.M.V., Lecturer, Department of E & C, S.S.I.T for his guidance, constant encouragement and wholehearted support.

My sincere thanks to Mr. Hari Vasudevan, Managing Director, Ostrich Mobility Instruments Pvt Ltd, Bangalore for his guidance, constant encouragement and whole-hearted support.

I would like to express my sincere thanks to Mr. Martin Mathew, R&D – Engineer and all the employees of Ostrich Mobility for their valuable support during the project period.

VI. CONCLUSION

To achieve independence in mobility for people with physical disability, right mobility equipments has to be designed based on the severity and type of disability. This is not a trivial job just because the nature and type of disability varies from person to person. Voice recognition and the controls to drive the wheelchair in the desired direction is the most critical aspect of the overall process of Commander. By its very special features Commander become an advanced means of mobility aid for patients suffering from Multiple Sclerosis, strokes and paralysis. Commander is thus a ray of hope to freedom of mobility for many disabled.

REFERENCES

[1]. Journal of Rehabilitation research and Development, July/august 2005

[2] Smart Wheelchairs: A literature Survey, July/August 2005 Journal of Rehabilitation Research & Development

[3] www.ostrichmobility.com

[4] Multiple Sclerosis and Amyotrophic Lateral Sclerosis Reza Vosoughi, MD; Mark S. Freedman, MSc, MD, FAAN, FRCPC

[5] Freedom of mobility by Mr.Hari Vasudevan, M-Tech, Ostrichmobility, Bangalore.

◈◈◈

Deployment Analysis in Underwater Acoustic Wireless Sensor Networks

T. S. Yenganti

Department of CSE, R.T.M.Nagpur University
Tulsiramaji Gaiwkad-Patil College of Engineering and Technology, Nagpur-441108
E-mail : yengantiwar@rediffmail.com

Abstract **-** This document gives formatting instructions for authors preparing papers for publication in the Proceedings of an IEEE conference. The au In this paper, different deployment strategies for two-dimensional and three-dimensional communication architectures for Under Water Acoustic Sensor Networks (UW-ASNs) are proposed, and statistical deployment analysis for both Architectures is provided. The objectives of this paper are to determine the minimum number of sensors needed to be deployed to achieve the optimal sensing and communication coverage, which are dictated by the application, provide guidelines on how to choose the optimal deployment surface area, given a target region. study the robustness of the sensor network to node failures, and provide an estimate of the number of redundant sensors to be deployed to compensate for possible failures

Keywords **-** *Design, Performance, Reliability. Underwater Acoustic Sensor Networks, Deployment.*

I. INTRODUCTION

Underwater sensor networks are envisioned to enable applications for oceanographic data collection, ocean sampling, environmental and pollution monitoring, offshore exploration, disaster prevention, tsunami and seaquake warning, assisted navigation, distributed tactical surveillance, and mine reconnaissance. There is, in fact, significant interest in monitoring aquatic environments for scientific, environmental, commercial, safety, and military reasons. While there is a need for highly precise, real-time, fine grained spatio-temporal sampling of the ocean environment, current methods such as remote telemetry and sequential local sensing cannot satisfy many application needs, which call for wireless underwater acoustic networking. Under Water Acoustic Sensor Networks (UW-ASN) consist of sensors that are deployed to perform collaborative monitoring tasks over a given region. UW-ASN communication links are based on *acoustic wireless technology*, which poses unique challenges due to the harsh underwater environment, such as limited bandwidth capacity, high and variable propagation delays , high bit error rates, and temporary losses of connectivity caused by multipath and fading phenomena . We consider two communication architectures for UW-ASNs, i.e., the *two-dimensional architecture*, where sensors are anchored to the bottom of the ocean, and the *three-dimensional architecture*, where sensors float at different ocean depths covering the entire monitored volume region. While the former is designed for networks whose objective is to monitor the ocean bottom, the latter is more suitable to detect and observe phenomena that cannot be adequately observed by means of ocean bottom sensor nodes. We propose different deployment strategies, and provide a mathematical analysis to study deployment issues concerning both architectures, with the objectives below:

i) Determine the minimum number of sensors needed to be deployed to achieve the target sensing and communication coverage, which are dictated by the application;

ii) Provide guidelines on how to choose the optimal deployment surface area, given a target region;

iii) Study the robustness of the sensor network to node failures, and provide an estimate of the number of redundant sensors to be deployed to compensate for possible failures.

II. RELATED WORK

The problem of sensing and communication coverage for terrestrial sensor networks has been addresses in several papers. However, to the best of the authors' knowledge, this work is the first to study deployment issues for underwater sensor networks. Many previous deployment solutions and theoretical

bounds assuming spatio-temporal correlation, mobile sensors, redeployment of nodes, and particular deployment grid structures may not be feasible for the underwater environment.

In particular, methods for determining net-work connectivity and coverage given a node reliability model are discussed, and an estimate of the minimum required node-reliability for meeting a system-reliability objective is provided. An interesting result is that connectivity does not necessarily imply coverage. As the node reliability decreases', in fact, the sufficient condition for connectivity becomes weaker than the necessary condition for coverage.

Although provides useful theoretical bounds and insight into the deployment of wireless terrestrial sensor networks, the analysis is limited to grid structures. In two coordination sleep algorithms are compared, a random and a coordinated sleep scheme. It is shown that when the density of the network increases, the duty cycle of the network can be decreased for a fixed coverage. In sensor coverage is achieved by moving sensor nodes after an initial random deployment. However requires either mobile sensor nodes or redeployment of nodes, which may not be feasible for UW-ASNs. In sensing and communication coverage in a three-dimensional environment are rigorously investigated. The diameter, minimum and maximum degree of the reach ability graph that describes the network are derived as a function of the communication range, while different degrees of coverage (1- coverage and, more in general, k-coverage) for the 3D environment are characterized as a function of the sensing range. Interestingly, it is shown that the sensing range r required for 1-coverage is greater than the transmission range t that guarantees network connectivity. Since in typical applications $t \geq r$, the network is guaranteed to be connected when 1-coverage is achieved.

Although these results were derived for terrestrial networks, they can also be applied in the underwater environment. Thus, in this paper, we will focus on the sensing coverage when discussing deployment issues in 3D UW-ASNS, as in three-dimensional networks it implicitly implies the communication coverage.

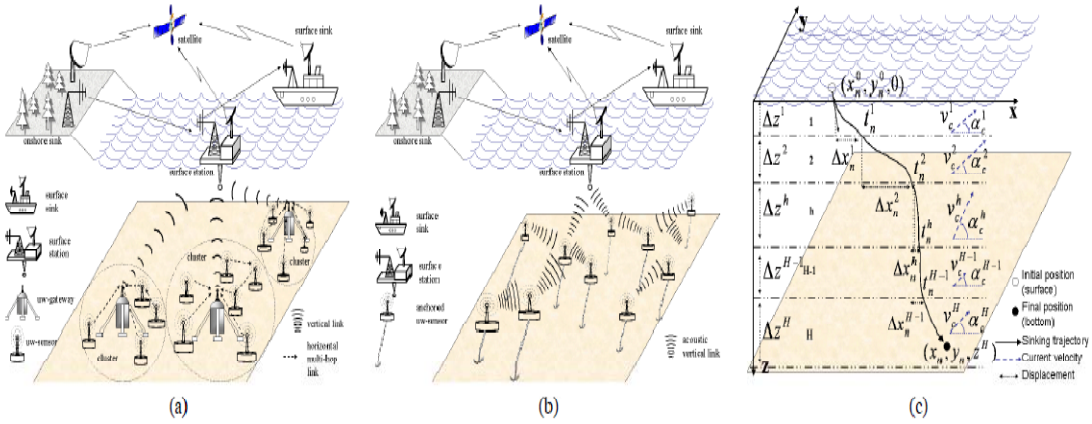


Figure 1: Architectures for two-dimensional (a) and three-dimensional (b) UW-ASNs; (c): Trajectory of a sinking object

III. COMMUNICATION ARCHITECTURE

We consider two communication architectures for underwater sensor networks, i.e., a *two-dimensional* and a *three-dimensional architecture* , and identify the relevant deployment challenges. As in terrestrial sensor networks, in UW-ASNs it is necessary to provide *communication coverage*, i.e., all sensors should be able to establish multi-hop paths to the sink, and *sensing coverage*, i.e., the monitored area should be covered by the sensors. More formally, the *sensing range r* of a sensor is the radius of the sphere that models the region monitored by the sensor (sensing sphere). A portion $A\eta$ of the monitored region A is said to be *k-covered* if every point in $A\eta$ falls within the sensing sphere of at least k sensors. The *k-coverage ratio* ηk of a monitored region A is the fraction of the volume/ area that is k-covered by a 3D/2D UW-ASN, respectively. In the following, we will consider the case of $k = 1$ both for 2D and 3D networks to obtain simple *1-cover age* $\eta 1$ of the region, since underwater sensors may be expensive

devices and spatio-temporal correlation may not be assumed.

1) Two-dimensional UW-ASNs

A reference architecture for two-dimensional underwater sensor networks is shown in Fig. 1(a), where deployed sensor nodes are anchored to the bottom of the ocean. Underwater sensors may be organized in a cluster-based architecture, and be interconnected to one or more *underwater gateways* (uw-gateways) by means of wireless acoustic links. Uw-gateways are network devices in charge of relaying data from the ocean bottom network to a surface station. They are equipped with a long-range *vertical* transceiver, which is used to relay data to a *surface station*, and with a *horizontal* transceiver, which is used to communicate with the sensor nodes to send commands and configuration data, and to collect monitored data. The surface station is equipped with an acoustic transceiver, which may be able to handle multiple parallel communications with the uw-gateways, and with a long-range radio transmitter and/or satellite transmitter, which is needed to communicate with an *onshore sink* and/or to a *surface sink*.

2) Three-dimensional UW-ASNs

Three-dimensional underwater networks are used to detect and observe phenomena that cannot be adequately observed by means of ocean bottom uw-sensor nodes, i.e., to perform cooperative sampling of the 3D ocean environment. In this architecture, sensors float at different depths to observe a given phenomenon. One possible solution would be to attach each sensor node to a surface buoy, by means of wires whose length can be regulated to adjust the depth of each sensor node. However, although this solution enables easy and quick deployment of the sensor network, multiple floating buoys may obstruct ships navigating on the surface, or they can be easily detected and deactivated by enemies in military settings. Furthermore, floating buoys are vulnerable to weather and tampering or pilfering. A different approach is to anchor winch based sensor devices to the bottom of the ocean, as depicted in Fig.1(b). Each sensor is anchored to the ocean bottom and is equipped with a floating buoy that can be inflated by a pump. The buoy pulls the sensor towards the ocean surface. The depth of the sensor can then be regulated by adjusting the length of the wire that connects the sensor to the anchor, by means of an electronically controlled engine that resides on the sensor

IV. DEPLOYMENT IN A 2D ENVIRONMENT

In this section, we provide a mathematical analysis of the graph properties of sensor devices that are deployed on the surface of the ocean, sink, and reach the ocean bottom. To achieve this, we study the trajectory of sinking devices (sensors and uw-gateways) when they are deployed on the ocean surface with known initial conditions (position and velocity). This allows us to capture both the case when sensor nodes are *randomly deployed* on the ocean surface, e.g., scattered from an airplane, or the case when sensors are *accurately positioned*, e.g., released from a vessel.

To address the deployment challenges presented in the previous section, in Section 4.1 we propose the *triangular-grid* deployment, and derive useful geo-metric properties. In Section 4.2, we study the dynamics of a sinking object and evaluate its trajectory under the presence of ocean currents. In Section 4.3, we characterize the different sinking behaviour of sensors and uw-gateways, with the objective of describing their average horizontal displacement and study the main communication properties of sensor clusters.

1) Triangular-grid Coverage Properties

In this section, we propose the *triangular-grid* deployment, and derive useful geometric properties. Let us consider the common case of sensors with same sensing range r. The optimal deployment strategy to cover a two-dimensional rectangular area using the minimum number of sensors is to centre each sensor at the vertex of a grid of equilateral triangles, as shown in Fig. 2(a). With this configuration, by adjusting the distance d among sensors, i.e., the side of the equilateral triangles, it is possible to achieve *full coverage*, i.e., $\eta = 1$. In addition, this enables to optimally control the coverage ratio η, defined as the ratio between the covered area and the target area. In particular, as it will be mathematically proven in the following, when $d = \sqrt{3}r$ the coverage ratio η is equal to1, i.e., the uncovered area ABC depicted in Figs. 2(a-b) is zero, and the overlapping areas are minimized. This allows to achieve the full coverage of a target area, but requires the highest number of sensors. Conversely, as the distance among sensors increases, i.e., the number of deployed sensors decreases, the coverage ratio decreases. Therefore, there is a trade-off between the number of deployed sensors and the achievable sensing coverage. We are interested in finding the minimum number of sensors that need to be deployed in order to guarantee a target sensing coverage η^*, which is dictated by the application requirements. To this end, we present the following theorem.

THEOREM 1. *In an equilateral grid the sensing coverage $\eta(d, r)$,i.e., the ratio of the covered area and the target area, is*

$$\eta(d,r) = \eta\left(\frac{d}{r}\right) = \begin{cases} \frac{A_{DEF} - A_{ABC}}{A_{DEF}} = 1 - \frac{A_{ABC}}{\frac{\sqrt{3}}{4}d^2} & \frac{d}{r} \in [0,2] \\ \frac{3 \cdot \frac{\pi r^2}{6}}{\frac{\sqrt{3}}{4}d^2} = \frac{2\pi}{\sqrt{3}} \cdot \left(\frac{d}{r}\right)^{-2} & \frac{d}{r} \in (2,\infty), \end{cases} \quad (1)$$

where:

$$A_{ABC} = \frac{\sqrt{3}}{4}\left(\frac{d}{2} - \sqrt{3r^2 - \frac{3}{4}d^2}\right)^2 - 3r^2 \arcsin\frac{\overline{BC}}{2r} + \frac{3}{4}\overline{BC}\sqrt{4r^2 - \overline{BC}^2}, \quad \overline{BC} = \frac{d}{2} - \sqrt{3r^2 - \frac{3}{4}d^2}. \quad (2)$$

PROOF - With reference to Fig. 2(b), which represents a zoomed portion of Fig. 2(a), $AE = r$ and $EH = d/2$, where r is the sensing range and d is the distance between sensors. Since the triangle DEF is equilateral by construction, $HO = (\sqrt{3}/6)d$. Consequently, since $AH = \sqrt{r^2 - d^2/4}$, it holds $AO = HO - AH = (\sqrt{3}/6)d - \sqrt{r^2 - d2/4}$. As triangle DEF is equilateral, triangle ABC is equilateral too. Since $AO = (\sqrt{3}/3)BC$, then $BC = d/2 - \sqrt{3r^2 - (3/4)d^2}$. Therefore, the area of triangle ABC is *A (delta) of ABC* $= (\sqrt{3}/4)BC^2$ In order to express the sensing coverage $\eta(d, r)$ as a function of d and r, we need to compute the area *delta ABC* of the *uncovered region ABC* among the circles with centres in D, E, and F, and radius r. This can be computed as *A ABC = A(delta)ABC − 3.A BTCK*, where *A BTCK* coincides with the difference of the areas of the circular sector *BTCF i.e A BTCK = A BTCF- A(delta) BCF.*

$r^2 \arcsin(\overline{BC}/2r) - (\overline{BC}/4)\sqrt{4r^2 - \overline{BC}^2}$.

Consequently, $A_{ABC} = (\sqrt{3}/4)\left(d/2 - \sqrt{3r^2 - (3/4)d^2}\right)^2 - 3r^2 \arcsin(\overline{BC}/2r) + (3/4)\overline{BC}\sqrt{4r^2 - \overline{BC}^2}$, where $\overline{BC} = d/2 - \sqrt{3r^2 - (3/4)d^2}$, which gives (1) in the non-trivial case $d/r \in [0,2]$. As far as the case $d/r \in (2,\infty)$ is concerned, no overlapping areas are formed, and the coverage η can be computed straightforward.

2) Trajectory of a Sinking Object

In this section, we study the dynamics of a sinking object and evaluate its trajectory under the presence of ocean currents. In particular, we first consider the ideal case in which the velocity of the ocean current does not change with depth; then, we extend the model to capture the more realistic case in which the velocity of the current depends on depth. According to Newton's first law of motion, the acceleration **a** describing the sinking in the water of an object with a density ρ and volume V is determined by the following vectorial motion law,

$$\vec{F}_W + \vec{F}_B + \vec{F}_R + \vec{F}_C = \rho V \cdot \vec{a}, \quad (4)$$

where:

- $\vec{F}_W = \rho V \cdot \vec{g}$ is the *weight force*, which depends on the density ρ[Kg/m^3] and volume V[m^3] of the sinking object, and on the terrestrial gravitational acceleration $g = 9.81$ m/s^2;

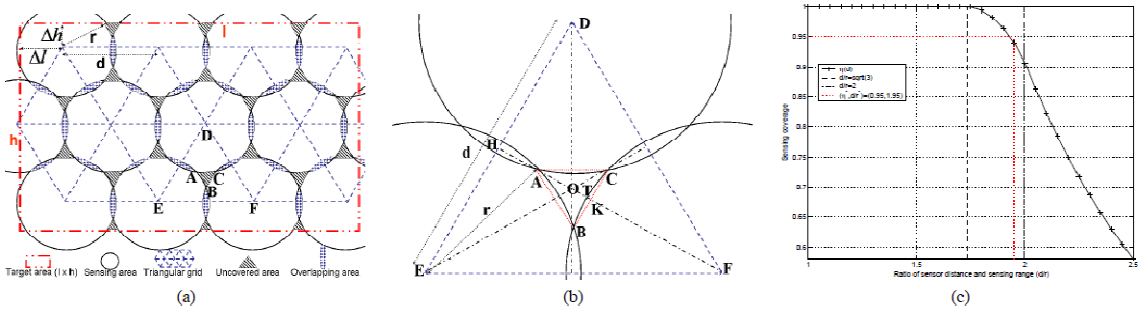


Figure 2: Triangular-grid deployment. (a): Grid structure and side margins; (b): Uncovered area; (c): Sensing coverage

We project (4) onto the x-, y-, and z- axes, which are directed as shown in Fig. 1(c), and we denote the dynamic position of the sinking object as **P** = (*x, y, z*), its velocity as _**v** = ($\dot{x}$, $\dot{y}$, $\dot{z}$), and its acceleration as **a** = ($\ddot{x}$, $\ddot{y}$, $\ddot{z}$). We then consider the velocity of the current V**c** = ($V^x c$, $V^y c$,$V^z c$), which, for the sake of clarity, is first assumed to be independent on the ocean depth (we will then relax this assumption). Under the assumption that no significant vertical movement of ocean water is observed, i.e., the considered area is neither an *upwelling* nor a *downwelling area*, the current along the z-axes can be neglected ($V^z c \approx 0$), and (4) leads to three scalar laws,

$$x: F_C^x = \rho V \ddot{x}; \quad y: F_C^y = \rho V \ddot{y}; \quad z: F_W^z + F_B^z + F_R^z = \rho V \ddot{z}. \tag{5}$$

Specifically, we obtain the following dynamic system equations,

$$\begin{cases} \ddot{x} + \frac{C\sigma A^{xy}}{\rho V}\dot{x} = \frac{C\sigma A^{xy}}{\rho V} v_c^x \\ \ddot{y} + \frac{C\sigma A^{xy}}{\rho V}\dot{y} = \frac{C\sigma A^{xy}}{\rho V} v_c^y \\ \ddot{z} + \frac{K\mu\rho_w A^z}{\rho V}\dot{z} = g\frac{\rho-\rho_w}{\rho}, \end{cases} \tag{6}$$

where *Axy* and *Az* represent the horizontal and vertical cross-sections, respectively. By solving this dynamic system, with the initial conditions of the object on the surface at time t^0 i.e its Position $\mathbf{P}(t0) = (x(t0), y(t0), 0)$ and velocity $\mathbf{v}(t0) = (x(t0), \dot{y}(t0), \dot{z}(t0))$, we obtain the solution

$$\begin{cases} x(t) = x(t^0) + v_c^x \cdot (t - t^0) + \frac{\dot{x}(t^0) - v_c^x}{C\sigma A^{xy}/\rho V} \cdot \left[1 - e^{-\frac{C\sigma A^{xy}}{\rho V}\cdot(t-t^0)}\right] \\ y(t) = y(t^0) + v_c^y \cdot (t - t^0) + \frac{\dot{y}(t^0) - v_c^y}{C\sigma A^{xy}/\rho V} \cdot \left[1 - e^{-\frac{C\sigma A^{xy}}{\rho V}\cdot(t-t^0)}\right] \\ z(t) = v_\infty^z \cdot (t - t^0) + \left[\dot{z}(t^0) - v_\infty^z\right] \cdot \left[1 - e^{-\frac{K\rho_w \mu A^z}{\rho V}\cdot(t-t^0)}\right], \end{cases} \tag{7}$$

where we denoted as $v_\infty^z = \frac{gV(\rho-\rho_w)}{K\rho_w \mu A^z}$ [m/s] the *terminal velocity* along z, which is computed by imposing in (5) the following force equilibrium, $F_W^z + F_B^z + F_R^z = 0$, i.e., $\ddot{z} = 0$ in (6).

V. DEPLOYMENT IN A 3D ENVIRONMENT

In this section, we propose three deployment strategies for three dimensional UW-ASNs to obtain a target *1-coverage* $\eta^*1 = \eta^*$ of the 3D region, i.e., the *3D-random*, the *bottom-random*, and the *bottom-grid* strategies. the sensing range *r* required for 1-coverage is greater than the transmission range *t* that guarantees network connectivity. Since in typical applications $t \geq r$, the network is guaranteed to be connected when 1-coverage is guaranteed. Thus, in the following we focus on the sensing coverage. In all these deployment strategies, winch-based sensor devices are anchored to the bottom of the ocean in such a way that they cannot drift with currents. Sensor devices are equipped with a floating buoy that can be inflated by a pump by means of an electronically controlled engine that resides on the sensor. This way, they can adjust their depth and float at different depths in order to observe a given phenomenon, as described in Section 3.2. In all the proposed deployment strategies, described hereafter, sensors are assumed to know their final positions by exploiting localization techniques.

3D-random - This is the simplest deployment strategy, and does not require any form of coordination from the surface station. Sensors are randomly deployed on the bottom of the 3D volume, where they are anchored. Then, each sensor randomly chooses its depth, and, by adjusting the length of the wire that connects it to the anchor, it floats to the selected depth. Finally, each sensor informs the surface station about its final position.

Bottom-random - As in the previous strategy, sensors are randomly deployed on the bottom, where they are anchored. Differently from the 3D-random scheme, the surface station is informed about their position on the bottom. Then, the surface station calculates the depth for each sensor in order to achieve the target 1- coverage ratio η^*. Finally, each sensor is assigned its target depth and floats to the desired position.

Bottom-grid - This deployment strategy needs to be assisted by one or multiple AUVs, which deploy the underwater sensors to predefined target locations to obtain a grid deployment on the bottom of the ocean. Each sensor is also assigned a desired depth by the AUV and accordingly floats to achieve the target coverage ratio η^*. As shown in Figs. 6(a-c), given a fixed number of sensors we achieve a better coverage ratio with increasing complexity of the deployment strategy. In fact, the coverage ratio obtained with the bottom-grid strategy is greater than the coverage ratio obtained with the bottom-random strategy, which is in turn greater than the coverage ratio of the 3D-random strategy. Moreover, given a target coverage ratio, the minimum number of sensors needed to achieve the desired coverage ratio decreases with the complexity of the deployment strategy. Figure 7 shows a comparison between the minimum normalized sensing range that guarantees coverage ratios of 1 and 0.9 with the bottom-random strategy and the theoretical bound on the minimum normalized sensing range derived in [4], where the authors investigate sensing and communication coverage in a 3D environment. According to Theorem 4 , the 3D volume is guaranteed to be *asymptotically almost surely* 1-covered iff $4\ 3\pi\ n\ V\ r3 = \ln n + \ln\ln n + \omega(n)$, with $1 << \omega(n) << \ln\ln n$,where *V* is the volume of the region to be covered, *n* the number of deployed sensors, and *r* their sensing range. Hence, to draw Fig.7 we set $\omega(n) = 1 + \ln\ln n\ 2$. This shows that the bottom-random deployment strategy very closely approximates the theoretically predicted bound, i.e., the minimum sensing range that guarantees 1-coverage with probability 1 is almost the same as that predicted by the model.

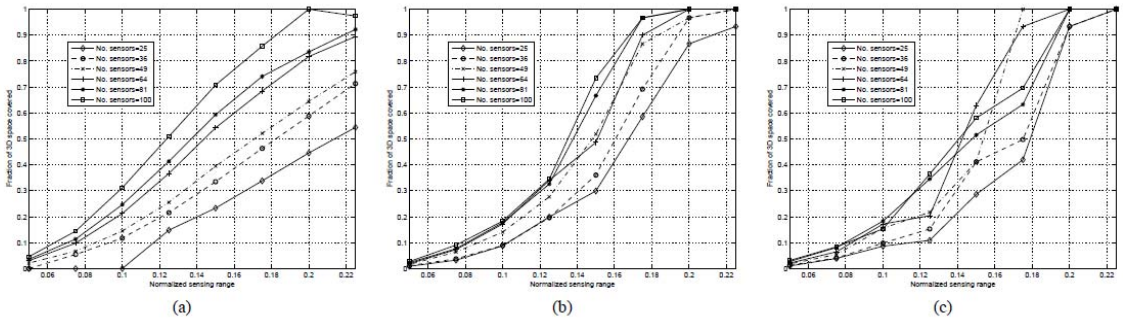


Figure 6: Three-dimensional scenario. (a): 3D coverage with a 3D random deployment; (b): Optimized 3D coverage with a 2D bottom-random deployment; (c): Optimized 3D coverage with a 2D bottom-grid deployment

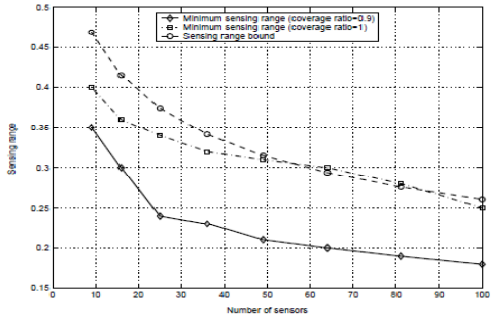


Figure 7: Theoretical and experimental sensing range

VI. CONCLUSIONS

In this paper, deployment strategies for two-dimensional and three dimensional architectures for underwater sensor networks were proposed, and deployment analysis was provided. The objectives were to determine the minimum number of sensors to be deployed to achieve the application-dependent target sensing and communication coverage; provide guidelines on how to choose the deployment surface area, given a target region; study the robustness of the sensor network to node failures, and provide an estimate of the number of required redundant sensors.

REFERENCES

[1] I. F. Akyildiz, D. Pompili, and T. Melodia. Underwater Acoustic Sensor Networks: Research Challenges. Ad Hoc Networks (Elsevier), 3(3):257–279, May 2006.

[2] C. Hsin and M. Liu. Network Coverage using Low Duty Cycled Sensors: Random and Coordinated Sleep Algorithms. In Proc. Of IEEE/ACM IPSN, pages 433–442, Berkeley, California, USA, Apr.2004.

[3] J. Proakis, E. Sozer, J. Rice, and M. Stojanovic. Shallow Water Acoustic Networks. IEEE Communications Magazine, pages 114–119, Nov. 2001.

[4] V. Ravelomanana. Extremal Properties of Three-dimensional Sensor Networks with Applications. IEEE Transactions on Mobile Computing, 3(3):246–257, July/Sept. 2004.

[5] R. A. Serway and J. W. Jewett. Physics for Scientists and Engineers. Brooks/Cole, 2004.

[6] S. Shakkottai, R. Srikant, and N. Shroff. Unreliable Sensor Grids:Coverage, Connectivity, and Diameter. In Proc. of IEEE INFOCOM, volume 2, pages 1073–1083, San Francisco, CA, USA, Apr. 2003.

[7] E. Sozer, M. Stojanovic, and J. Proakis. Underwater Acoustic Networks. IEEE Journal of Oceanic Engineering, 25(1):72–83, Jan.

◈◈◈

Outlier Detection Over Data Stream Using Cluster Based Approach And Distance Based Approach

Surekha V Peshatwar & Snehlata Dongre

Department of Computer science and Engineering,
G. H. Raisoni College of Engineering, Nagpur, Maharashtra, India
E-mail : peshatwarsurekha@yahoo.com, dongre.sneha@gmail.com

Abstract **-** Outlier detection is considered an important data mining task, aiming at the discovery of elements (also known as outliers) that show significant diversion from the expected case. Most of the existing methods are based on distance measure. But in case of data stream these method is not efficient as computational point of view. However, distance-based methods have time and space complexities that make them impractical for streaming data. So hybrid approach is used in which two technique are combined. Hybrid approach contains cluster based approach and distance based approach. In cluster based approach, clustering algorithm is used to create the fixed number of cluster and among each cluster, candidate outlier will find then distance based approach is used over the candidate outlier to find the actual outlier. These approach reduces the time and space complexity.

General Terms : *Data Stream Mining.*

Keywords : *outlier detection, data stream, cluster based, Distance based.*

I. INTRODUCTION

Outlier detection is currently very active area of research in data stream mining community. However, earlier research for the problem of outlier detection is suitable for disk resident datasets where the entire dataset is available in advance and algorithms can operate in more than single passes. But, outlier detection over data stream is a challenging task because data is continuously updated and flowing.

Finding outliers in a collection of patterns is a very well known problem in the data mining field. An outlier is a pattern which is dissimilar with respect to the rest of the patterns in the dataset. Depending upon the application domain, outliers are of particular interest. In some cases presence of outliers are adversely affect the conclusions drawn out of the analysis and hence need to be eliminated beforehand. There are varied reasons for outlier generation in the first place. For example outliers may be generated due to measurement impairments, rare normal events exhibiting entirely different characteristics, deliberate actions etc. Detecting outliers may lead to the discovery of truly unexpected behavior and help avoid wrong conclusions etc. Most of the existing work for outlier detection over the data stream only focus on detection rate of outliers while ignoring the most important issue of data stream mining like, low memory requirements and high speed algorithms to keep pace with high speed unbounded data streams.

The main objective is to find the outlier from the data stream using cluster based method and distance based method. The problem of finding outliers in data has broad applications in areas as diverse as data cleaning, fraud detection, network monitoring, invasive species monitoring, etc. While there are dozens of techniques that have been proposed to solve this problem for static data collections, very simple distance-based outlier detection methods are known to be competitive or superior to more complex methods. However, distance-based methods have time and space complexities that make them impractical for streaming data and/or resource limited sensors. So, first apply cluster based method to reduce the size of data and then apply the distance based method to find the outlier detection. So that distance based method will be practically implementable for streaming data and time and space complexity will be reduced for data stream.

The rest of the paper is organized as follows. Section 2 reviews related work in outlier detection. Section 3 provide the details of technique used in this paper. Section 4 provide system architecture.

II. REVIEW OF LITERATURE

Outlier detection (deviation detection, exception mining, novelty detection, etc.) is an important problem that has attracted wide interest and numerous solutions. These solutions can be broadly classified into several major ideas:

2.1 Density-Based [7]:

Objects in low-density regions of space are flagged.

Disadvantage: Density based models require the careful settings of several parameters. It requires quadratic time complexity. It may rule out outliers close to some non-outliers patterns that has low density.

2.2 Connectedness [5]:

In domains where objects are linked (social networks, biological networks), objects with few links are considered potential anomalies.

Disadvantage: Connectedness approaches are only defined for datasets with linkage information.

Distance-Based [2]:

Given any distance measure, objects that have distances to their nearest neighbors that exceed a specific threshold are considered potential anomalies.

In contrast to the above, distance-based methods are much more flexible and robust. They are defined for any data type for which we have a distance measure and do not require a detailed understanding of the application domain.

Cluster based approach [9]:

The clustering based techniques involve a clustering step which partitions the data into groups which contain similar objects. The assumed behavior of outliers is that they either do not belong to any cluster, or belong to very small clusters, or are forced to belong to a cluster where they are very different from other members. Clustering based outlier detection techniques have been enveloped which make use of the fact that outliers do not belong to any cluster since they are very few and different from the normal instances.

III. PROPOSED METHOD IN DETAILS

3.1 Cluster based approach

The cluster based techniques involve a clustering step which partitions the data into groups which contain similar objects. The assumed behavior of outliers is that they either do not belong to any cluster, or belong to very small clusters, or are forced to belong to a cluster where they are very different from other members. Clustering based outlier detection techniques have been enveloped which make use of the fact that outliers do not belong to any cluster since they are very few and different from the normal instances.

In this project, Cluster based approach is here act as data reduction. First, cluster based technique is used to find cluster of dataset. Once cluster are formed, centriod of each cluster are calculated. Remove the data up to certain radius as a real data. After removing the real data, remaining data are the candidate outlier. Candidate outliers are the temporary outlier.

3.2 Distance based approach

In this approach, there are n numbers of points in the original dataset so the distance of a point from its neighbors is calculated. If the neighboring points are relatively close, then the point is considered normal. If the neighboring points are far away, then the point is considered an outlier. In this project, the Euclidean distance is used as the distance function.

In this project, candidate outliers are taken from the cluster based approach and then calculate the distance of each point with the remaining point and then count the number of objects which are lie within some threshold value. If the count is less then it will declare finally as “outlier” otherwise as a “real” data.

IV. SYSTEM ARCHITECTURE

It is assumed that, the number of outliers in any dataset is expected to be extremely small as compared to the normal data. So, it is highly inefficient to apply the traditional outlier detection algorithms over the entire data set, especially in case of data stream this method can become highly expensive as well as can often led us to wrong decision in finding most outstanding outliers

Traditional methods for outlier detection can produce good results on stored static dataset. These methods cannot be applied to streaming data efficiently as these methods are suitable for the environment where the entire dataset is already available and algorithm can operate in more than one pass. A general framework for mining data streams need small constant time per record along with the minimum memory requirement, using at most one scan of data.

System can be divided in foollowing steps:

- Partition the data stream into number of chunks and each chunk contain set of data.
- Over each chunk, apply clustering method to figure out candidate outliers and safe region.
- Apply distance based outlier detection algorithm over candidate outliers and find final outliers.

In the first phase, the cluster based method is applied; safe data will be pruned out while the candidate cells will still be kept there for further processing. Safe region are pruned out from the current chunks so that next chunks can allow for storing. Then in second phase, distance based strategy is applied over the candidate to efficiently figure out the outliers while discard rest of data.

Proposed approach do not suffer from high computational cost, which is always a problem when distance based approaches are applied over high speed huge volume of data streams. Proposed algorithm efficiently prune of the safe cells and save huge number of extra calculations.

As we know that the number of outliers in a data set is very few, it is unnecessary to calculate these measures for all points. By removing the points which are probably not outliers, we can reduce the computation time.

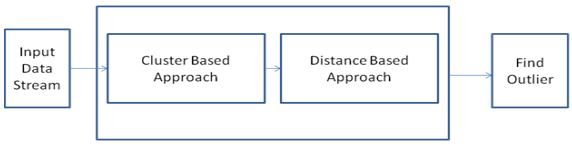


Fig 1:System Architecture

V. PROPOSED ALGORITHM

Hybride Algorithm For Outlier Detection

Require K: Number of cluster

Require X_j : { x_1,x_2,x_3……..,x_N }

Require N: Chunk size

Require T: Threshold

Step 1: Input a chunk of stream X_j : { x_1,x_2,x_3……..,x_N }

Setp 2: Cluster the chunk in fixed number number of cluster K

K-mean(X_j, K)

Step 3: Find the point having maximum distance in each cluster

Step 4: using maximum distance of each cluster separate the inliers and candidate outliers for each cluster.

Step 5: Keep the candidate outliers.

Step 6: Discard the safe region or inliers of each cluster.

Step 7: Consider candidate outlier of all cluster in each chunk . Apply distance based approach.

……

VI. RESULT AND DISCUSSION

In this section, we used stock data of Microsoft which is available at Yahoo. We were apply our hybride approach over this data. In first phase, we divide data into chunk and size of each chunk is around 252 records. Over each chunk, we apply our hybride approach. In first phase, we have to provide the number of cluster eg K=3. So 3 cluster will formed in cluster based approach. Now we assume that the element near to the centriod are the real data so we need to prune out this records. So we find maximum distance element and by using that we find threshold value for each cluster. As the size of each cluster is different so threshold value for each cluster will be different. Once the threshold has find, the data which lie within that threshold will consider as "real" data. So we will prune out that real data and the data which lie outside the threshold will considered as candidate outlier. But this candidate outlier still contain some real data and outlier value. So we need to apply Distance based approach. In this, we will find the distance of each element with the remaining point and then cont the number of objects which are lie within some threshold value. If the count is more then data object is the real data otherwise it will be the final outlier value.

In figure 2, it shows three graphs, In first graph, complete stock data is shown graphically. In second subplots, it shows the result of cluster based approach for 3 cluster. We circled candidate outlier value which we get as final outlier after applying the distance based approach which is circled within third subplot.

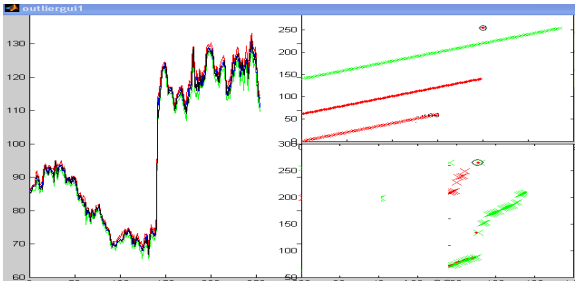


Figure2. Outlier value(circled) using proposed approach

VII.CONCLUSION

In this paper, we used hybrid approach to find the outlier in the data stream. First , we used cluster based approach to find the candidate outlier. Candidate outlier may or may not be the final outlier so we applied distance based approach to find the correct outlier.

REFERENCES

[1] K.Prasanna Lakshmi, Dr. C.R.K.Reddy,2010 " A Survey on different trends in Data Streams", International Conference on Networking and Information Technology

[2] Vit Niennattrakul, Eamonn Keogh, Chotirat Ann Ratanamahatana,2010 "Data Editing Techniques to Allow the Application of Distance-Based Outlier Detection to Streams", IEEE International Conference on Data Mining

[3] Rajendra Pamula,Jatidra kumar Deka, Sukumar Nandi, 2010 "Distance Based Fast Outlier Detection", Annual IEEE India Conference.

[4] V. Roth, "Kernel Fisher Discriminants for Outlier Detection," Neural Computation, vol. 18, pp. 942-960, April 2006.

[5] J. Tang, Z. Chen, A. W.-C. Fu, and D. W.-L. Cheung, "Enhancing Effectiveness of Outlier Detections for Low Density Patterns," In Proceedings of PAKDD'02, pp. 535-548, May 6-8 2002.

[6] Jingke Xi," Outlier Detection Algorithms in Data Mining", Second International Symposium on Intelligent Information Technology Application 2008

[7] Peng Yang; Biao Huang, "A Modified Density Based Outlier Mining Algorithm for Large Dataset" International conference on Future Information Technology and Management Engineering, 2008.

[8] Tianqing Zhu, "An Outlier Detection Model Based on Cross Datasets Comparison for Financial Surveillance" IEEE Asia-Pacific Conference on Services Computing, 2006

[9] Parneeta Dhaliwal , MPS Bhatia and Priti Bansal," A Cluster-based Approach for Outlier Detection in Dynamic Data Streams (KORM: k-median OutlieR Miner)" JOURNAL OF COMPUTING, VOLUME 2, ISSUE 2, FEBRUARY 2010, ISSN 2151-9617

[10] Peng Yang; Biao Huang;" KNN Based Outlier Detection Algorithm in Large Dataset " International Workshop on Education Technology and Training, 2008

[11] Manzoor Elahi, Xinjie Lv, Wasif Nisar, Imran Ali Khan, Ying Qiao Hongan Wang," DB-Outlier Detection Algorithm using Divide andConquer approach over Dynamic DataStream" International Conference on Computer Science and Software Engineering 2008

◈◈◈

Optimal Power-Sharing of Multiple Distributed Generators (Dgs) Considering Control Modes In A Microgrid

Chitra. N, Logeshwari. V & Senthil Kumar. A

EEE S.K.P Engineering College,Thiruvannamalai, India
E-mail : Sivakumar.poruran@gmail.com

Abstract - This paper describes the active power sharing in multiple distributed generators (DGs) in a microgrid. Microgrids have two operating modes: 1) a grid-connected mode and 2) an islanded mode. During islanded operation, one DG unit should share output generation power with other units in exact accordance with the load. For controlling the active power of DGs unit output power control (UPC) is introduced. The feasibility of the proposed power control mode is simulated by using MATLAB/Simulink.

Keywords-*Active power control; distributed generator; droop characteristics; micro grid.*

I. INTRODUCTION

Recently, interest in distributed generation systems (DGS) is rapidly increasing, particularly onsite generation. This interest is due to the facts that larger power plants are economically unfeasible in many regions due to increasing system and fuel costs, and stricter environmental regulations. In addition, recent technological advances in small generators, power electronics, and energy storage devices have provided a new opportunity for distributed energy resources at the distribution level, and especially, the incentive laws to utilize renewable energies have also encouraged a more decentralized approach to power delivery[1]-[3]. Accordingly, distributed generators (DGs) have been installed in power systems and tested for better configurations and control schemes.

Microgrids can realize a coordinated approach to facilitate the penetration of DG into the utility network [4]. The CERTS defines the microgrid as a small-scale, low-voltage system consisting of a combination of generators, loads, and energy storage elements [5]–[7]. Essentially, a microgrid is an active distribution network that can be exploited in two operating conditions. In grid-connected mode, the microgrid is connected to the distribution grid at a single point of connection, the point of common coupling (PCC). In islanded mode, the microgrid is disconnected from the main grid. A key advantage is that the microgrid appears to the power network as a single controllable unit, enabling it to deliver the cost benefits of large units. Furthermore, microgrids can enhance local reliability, reduce feeder losses, provide reactive power and local voltage support, remove transmission and distribution bottlenecks, increase efficiency through the use of waste heat and provide uninterruptible power supply functions [8], [9]. The increased amount of small-scale power sources that are not directly online requires the development of converter-based microgrids [10]. Hence, the microgrid control focuses on the control of these converters.

There are many technical issues related to microgrid operation, including interconnection schemes between microgrids and the main grid [11]; voltage-control schemes within a microgrid [12], [13], [14]; and frequency control during islanded operation [12]. Among these, this paper focuses on active power and frequency-control strategies for sound operation of a microgrid with multiple DGs.

This paper focuses on proper active power sharing of each DG. Many innovative control techniques have been used for stability of the system as well as for proper load sharing. The most common method is the use of droop characteristics for wireless load sharing. Local signals are used as feedback to control the parallel converters, since in a real system, the distance between the converters may make an inter-communication impractical. To control the active power among multiple DGs, unit output power control (UPC) is proposed [15]. During UPC, the output power of the DG is constantly controlled according to the power reference.

The remainder of this paper is divided into five sections. Section 2 presents a detailed description of the

power-control methods. In Section 3, presents a simulation model. In Section 4, presents a simulation results. In Section 5, contains concluding remarks.

II. DESCRIPTION OF THE POWER-CONTROL MODES

In this section the basic concept of UPC mode and the active power-sharing principle is analyzed.

A. Unit output Power Control (UPC) Mode

The objective of this mode is to control the power injected by a DG unit at a desired value (P_{ref}) [15]. To accomplish this, the voltage *(V)* at the interconnection point and the DG output current *(I)* are measured as shown in Fig. 1. The power injection (P_{DG}) is calculated from the measured voltage and current and fed back to the generator controller (GC).

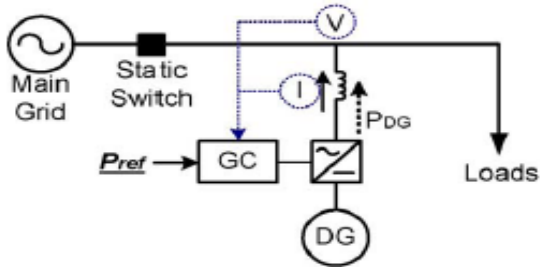


Fig. 1 : Unit output power control (UPC).

When the microgrid is connected to the main grid, the DG is able to maintain a constant output power regardless of the load variation, because the power mismatch can be compensated by the grid. However, during islanded operation, DGs must follow the load demand exactly. In numerous studies, a power versus frequency (*P–f*) droop control has been adopted for DG power-sharing methods [12], [16]–[20]. This control uses the frequency of the microgrid as a common signal among the DGs to balance the active power generation of the system [12]. *P–f* droop-based power controllers have proven to be robust and adaptive to variation in the power system operational conditions, such as frequency- and/or voltage-dependent loads and system losses [12], [20]. The relationship between the frequency (*f*) and the power output of a DG (P) can be expressed as

(1)

where K^U is the UPC droop constant, f' and P' are the frequency and DG output power at a new operating point, and f^0 and P^0 are the nominal values. When the load increases during islanded operation, the DG output power also increases, and the frequency decreases according to the droop characteristic, as given by (1).

B. DG Active Power Controller

Fig. 2 shows the active power-control block of a DG, where the inputs are local measurements of frequency (*f*) and power output (*P*), or feeder flow (*FL*), and the set points are provided by the central controller. The output is the axis current reference signal for the current controller or the angle of the desired voltage. The control block contains two additional functions: 1) frequency droop control and 2) output limit control.

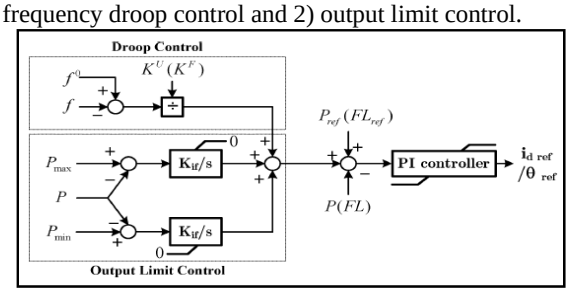


Fig. 2 : Active power-control block diagram of a DG

proposed, but the secondary load-frequency control function is not considered in this study.

The output limit control function restricts the steady-state output power of the DGs to within the limits. Since the energy sources of DGs have a finite capacity for storing or generating energy, the output limit should be enforced [15], [19], [21]. The function will be activated only when the power output violates the limits, and effectively enforces the output limits [15].

III. SIMULATION MODEL

A. Test System and Simulation Scenario

Fig. 3 shows a single-line diagram of the microgrid test system model, which is connected to a 13.8-kV, 50-Hz main grid system by a static switch. The system parameters are similar to [12] and [22], with slight modifications in the line connections and parameters. The test model contains three DGs with voltage ratings of 4.14 kV, and maximum power generation limits (arbitrarily chosen to be 2.5, 3.0, and 2.0MW, respectively) are included in the simulations. We set the UPC droop constants of the DGs to be equal to 1.2, 1.0, and 1.5 Hz/MW, respectively, which means that 0.05-p.u. frequency deviation causes a 1.0-p.u. change in the power output of each DG [23]. Three lumped balanced loads represent the sensitive loads, whose demands are arbitrarily chosen. The test system modeled with DG controllers was modeled by using the MATLAB/ Simulink.

The simulation sequence was as follows. $Load_1$ was decreased from 3.0 MW and 0.9 MVAr to 2.4 MW and 0.6 MVAr at 1.2 s to investigate power sharing in terms of load variation during grid-connected operation. At 2.0 s, the static switch was opened so that the microgrid was islanded from the grid. To demonstrate the effect of load variation during islanded operation was increased from 1.8 MW and 0.6 MVAr to 2.4 MW and 1.2 MVAr at 3.0 s.

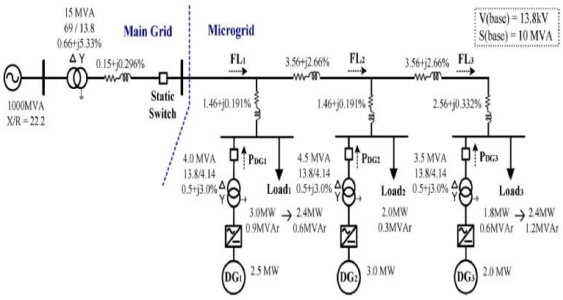


Fig. 3 : Single-line diagram of the microgrid system.

IV. SIMULATION RESULTS

Fig. 4, 5 shows the feeder flows and the DGs power outputs respectively. Since all DGs are operated in the UPC mode, the output of each DG is maintained constant at its initial reference value as -0.5, 0.98 and 0.8 MW respectively until 2 s. At 1.2 s, the main grid compensated for the variation of $Load_1$, so that the power flow from the main grid (FL_1) is reduced to 0.1 MW.

After islanding at 2 s, all DGs increased their output to match the load demands. In the new steady state, the outputs of the DGs are approximately 1.15, 1.47 and 0.99 MW respectively.

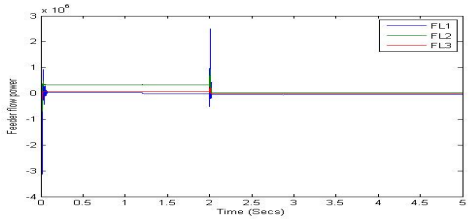


Figure-4 Power flow in the feeders

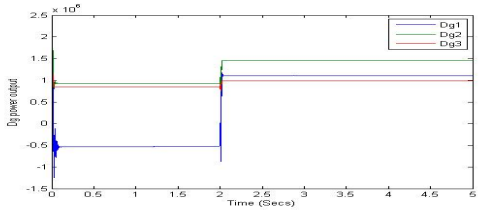


Figure – 5 : Active power output of each DG.

V. CONCLUSION

This paper has presented the active power-sharing principles of multiple DGs are examined by considering its control modes. During islanded mode of operation, the load demand has been matched by DGs alone. Thus the UPC mode is more advantageous for islanded mode operation than grid connected mode. The simulation results indicated that all DGs shared the proper amount of power especially in islanded mode.

REFERENCES

[1] B. Maurhoff and G.Wood, "Dispersed generation to reduce power costs and improve service reliability," in Proc. Rural Electric Power Conf.,2000, pp. C5/1–C5/7.

[2] J. L. Del Monaco, "The role of distributed generation in the critical electric power infrastructure," in IEEE-Power Engineering Soc. Winter Meeting, vol. 1, 2001, pp. 144–145.

[3] B. Lasseter, "Microgrids," in IEEE-Power Engineering Soc. Winter Meeting, vol. 1, 2001, pp. 146–149.

[4] A. Engler, O. Osika, M. Barnes, N. Jenkins, and A. Arulampalam, European Commission, DB1 local micro source controller strategies and algorithms, Feb. 2004 [Online]. Available: http://www.microgrids.eu/ micro2000

[5] R. H. Lasseter, A. Akhil, C. Marnay, J. Stephens, J. Dagle, R. Guttromson, A. Meliopoulous, R. Yinger, and J. Eto, Cal. Energy Commission, Office of Power Technol., The CERTS microgrid concept, White Paper on integration of distributed energy resources U.S. Dept. Energy, Apr. 2002.

[6] S. Barsali,M. Ceraolo, P. Pelacchi, and D. Poli, "Control techniques of dispersed generators to improve the continuity of electricity supply," in Proc. IEEE PES Winter Meet., 2002, pp. 789–794.

[7] J. A. Peças Lopes, C. Moreira, A. Madureira, F. Resende, P. G. Abia, X.Wu, N. Jayawarna, Y. Zhang, N. Jenkins, F. Kanellos, N. Hatziargyriou, and C. Duvauchelle, Microgrids large scale integration of microgeneration to low voltage grids: Dd1 emergency strategies and algorithms Oct. 2004 [Online]. Available: www.microgrids.eu/micro2000

[8] H. Jiayi, J. Chuanwen, and X. Rong, "A review on distributed energy resources andmicrogrid," Renewable Sustainable Energy Rev., vol. 12, no. 9, pp. 2472–2483, 2008.

[9] C. Marnay and G. Venkataramanan, "Microgrids in the evolving electricity generation and delivery infrastructure," in Proc. IEEE PES General Meet., Montreal, QC, Canada, Jun. 18–22, 2006.

[10] T. C. Green and M. Prodanovíc, "Control of inverter-based microgrids," Elect. Power Syst. Res., vol. 77, no. 9, pp. 1204–1213, Jul. 2007.

[11] Y. Li, D. M. Vilathgamuwa, and P. C. Loh, "Design, analysis, and real-time testing of a controller for multibus microgrid system," IEEE. Trans. Power Electron., vol. 19, no. 5, pp. 1195–1204, Sep. 2004.

[12] F. Katiraei and M. R. Iravani, "Power management strategies for a micro grid with multiple distributed generation units," IEEE Trans. Power Syst., vol. 21, no. 4, pp. 1821–1831, Nov. 2006.

[13] M. H. J. Bollen and A. Sannino, "Voltage control with inverter-based distributed generation," IEEE Trans. Power Del., vol. 20, no. 1, pp. 519–520, Jan. 2005.

[14] W. Freitas, J. C. M. Vieira, A. Morelato, L. C. P. da Silva, V. F. da Costa, and F. A. B. Lemos, "Comparative analysis between synchronous and induction machines for distributed generation applications," IEEE Trans. Power Syst., vol. 21, no. 1, pp. 301–311, Feb. 2006.

[15] PSERCR. H. Lasseter, "Control and design of microgrid components," Final Project Reports. [Online].Available:http://www.pserc.org/cgipserc/ getbig/publicatio/reports/2006report/lasseter_microgridcontrol_final project_report.pdf

[16] P. Piagi and R. H. Lasseter, "Autonomous control of microgrids," presented at the Power Eng. Soc. General Meeting, Montreal, QC, Canada, 2006.

[17] J. M. Guerrero, L. G. de Vicuna, J. Matas, M. Castilla, and J. Miret, "A wireless controller to enhance dynamic performance of parallel inverters in distributed generation systems," IEEE Trans. Power Electron., vol. 19, no. 5, pp. 1205–1213, Sep. 2004.

[18] J. M. Guerrero, J. Matas, L. G. de Vicuna, M. Castilla, and J. Miret, "Decentralized control for parallel operation of distributed generation inverters using resistive output impedance," IEEE Trans. Ind. Electron., vol. 54, no. 2, pp. 994–1004, Apr. 2007.

[19] C. L. Moreira, F. O. Resende, and J. A. P. Lopes, "Using low voltage microgrids for service restoration," IEEE Trans. Power Syst., vol. 22, no. 1, pp. 395–403, Feb. 2007.

[20] N. Pogaku, M. Prodanovic, and T. C. Green, "Modeling, analysis and testing of autonomous operation of an inverter-based microgrid," IEEE Trans. Power Electron., vol. 22, no. 2, pp. 613–625, Mar. 2007.

[21] J. A. P. Lopes, C. L. Moreira, and A. G. Madureira, "Defining control strategies for microgrids islanded operation," IEEE Trans. Power Syst., vol. 21, no. 2, pp. 916–924, May 2006.

[22] F. Katiraei, M. R. Iravani, and P. W. Lehn, "Micro-grid autonomous operation during and subsequent to islanding process," IEEE Trans. Power Del., vol. 20, no. 1, pp. 248–257, Jan. 2005.

[23] P. Kundur, Power System Stability and Control. New York: McGraw- Hill, 1994.

◈◈◈

Optimal Design of Microgrid in an Autonomous Mode Using Antcolony Optimization

Chitra. N, Tamizharasi.G & A. Senthilkumar

EEE, SKP Engineering College, Tiruvannamalai. INDIA
E-mail : sivakumar.poruran@gmail.com

Abstract **-** The dynamic nature of the distribution network challenges the stability and control effectiveness of the microgrid in autonomous mode. In this paper, nonlinear model of microgrid operating in autonomous mode has been presented. The controller parameters and power sharing coefficients are optimized in case of autonomous mode. The control problem has been formulated as an optimization problem where Ant colony optimization is employed to search for optimal settings of the optimized parameters. In addition, nonlinear time-domain-based objective function has been proposed to minimize the error in the measured power and to enhance the damping characteristics, respectively. Finally, the nonlinear time-domain simulation has been carried out to assess the effectiveness of the proposed controllers under different disturbances and loading conditions. The results show satisfactory performance with efficient damping characteristics of the microgrid considered in this study.

Keywords*- Autonomous mode, controller design and inverter control, distributed generator (DG), droop-control concepts, inverter-based distributed generators, microgrid stability, optimal control, optimal power sharing, Ant colony optimization (ACO), power electronic inverters****.***

I. INTRODUCTION

Recent innovations in small-scale distributed power generation systems combined with technological advancements in power electronic systems led to concepts of future network technologies such as microgrid. These small autonomous regions of power systems can offer increased reliability and efficiency and can help integrate renewable energy and other forms of distributed generation (DG) [1]. Many forms of distributed generation such as fuel-cells, photo-voltaic and micro-turbines are interfaced to the network through power electronic converters [2]–[5]. These interface devices make the sources more flexible in their operation and control compared to the conventional electrical machines. However, due to their negligible physical inertia they also make the system potentially susceptible to oscillation resulting from network disturbances.

A microgrid can be operated either in grid connected mode or in stand-alone mode. In grid connected mode, most of the system-level dynamics are dictated by the main grid due to the relatively small size of micro sources. In stand-alone mode, the system dynamics are dictated by micro sources themselves, their power regulation control and, to an unusual degree, by the network itself.

In this paper, a systematic approach to modeling an inverter-based microgrid is presented. Each DG inverter will have an outer power loop based on droop control to share the-fundamental real and reactive powers with other DGs. Inverter internal controls will include voltage and current controllers which are designed to reject high frequency disturbances and damp the output filter to avoid any resonance with the external network. The small-signal state-space model of an individual inverter is constructed by including the controllers, output filter and coupling inductor on a synchronous reference frame whose rotation frequency is set by the power controller of that inverter. An arbitrary choice is made to select one inverter frame as the common reference frame and all other inverters are translated to this common reference frame using the simple transformation techniques familiar in synchronous machine systems. It is considered that state-less impedance models of the network are inadequate for use with full-order inverter models which include high frequency modes. Instead a dynamic (state-space) model of the network is formed on the common reference frame.

Recently, computational intelligence algorithms such as genetic algorithm (GA) and ant colony optimization (ACO) have been applied to different power system problems with impressive success [6].

However, some deficiencies in GA performance such as the premature convergence have been recorded. On the other hand, ACO has been widely implemented and stamped as one of the promising optimization technique due to its simplicity, computational efficiency, and robustness.

In this paper, a single code for modeling, optimization, linearization, and nonlinear time-domain simulation has been developed. A new technique for stability enhancement of a microgrid operating in autonomous mode is proposed in this paper. Voltage source inverter (VSI), *LC* filter, coupling inductance, phase-locked loop (PLL), lines, loads, and power, current, and voltage controllers have been modeled. The design problem of different microgrid components and controllers' parameters has been formulated as an optimization problem where ACO is employed to solve this design problem . The nonlinear time-domain simulation-based objective functions are considered with the aim of autonomous microgrid stability enhancement where the controller parameters and the power sharing coefficients are optimized. The performance of the microgrid with the proposed controllers and optimal settings under different disturbances has been examined through the nonlinear time-domain simulations. The results show the effectiveness of the proposed approach to enhance the stability of the microgrid considered.

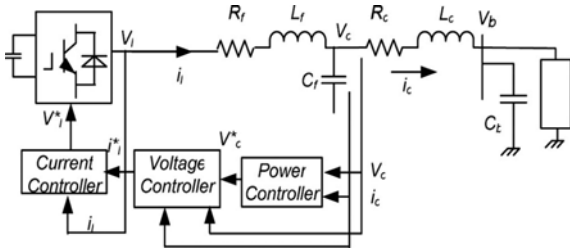


Fig. 1 : Microgrid in autonomous mode.

II. MICROGRID MODELING IN AUTONOMOUS MODE

This section provides the mathematical model of the autonomous microgrid consisting of VSI controller connected to the loads through *LC* filter and coupling inductance.

A. Modeling of VSI and Its controllerss

Power, current, and voltage controllers have been used to control the microgrid inverter in the autonomous mode, as shown in Fig. 1. First, the active and reactive powers are calculated using the measured output current and voltage of the VSI. An external power control loop sets the magnitude and frequency (and hence, phase) for the fundamental component of the inverter output voltage according to the droop characteristics set for the real and reactive powers [7]. Then, the voltage and current controllers are designed to reject high-frequency disturbances and provide sufficient damping for the *LC* filter [9].

B. Power Controller

In a conventional power system, synchronous generators share any increase in the load by decreasing the frequency according to their governor droop characteristic. In the autonomous mode, the inverter emulates the behavior of a synchronous machine. Therefore, the angle δ can be controlled by regulating P, while the output voltage is controllable through Q. Control of frequency dynamically controls the power angle,

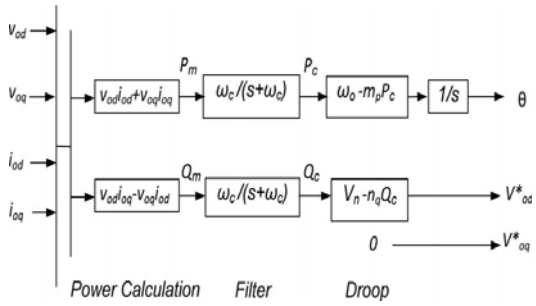


Fig. 2 : Block diagram of power controller.

and thus, the real power flow [9]. For stable operation, the real and reactive power output of the inverters should be properly controlled. First, the measured output voltage and current are used to calculate the instantaneous active and reactive power

$$\tilde{p} = v_{od} i_{od} + v_{oq} i_{oq} \tilde{q} = v_{od} i_{oq} - v_{oq} i_{od} \quad (1)$$

Second, the real and reactive powers, Pc and Qc, corresponding to the fundamental components are obtained after passing these powers through low-pass filter. Finally, the frequency ω and the output d-axis voltage magnitude reference v_{od}^* can be determined as

$$\omega = \omega_n - m_p P_c \quad , \theta = \int \omega \quad (2)$$

$$v_{od}^* = v_n - n_q Q_c, v_{oq}^* = 0 \quad (3)$$

Where m_p and n_q are the real and reactive power sharing coefficients. The different droop characteristics show that the three inverters can share the total real and reactive power.

C. Voltage Controller

The voltage controller block diagram including all feed-back and feed-forward terms. Output voltage control is achieved with a standard PI controller.

$$\dot{\emptyset}_d = v_{od}^* - v_{od} \dot{\emptyset}_q = v_{oq}^* - v_{oq}$$

Along with the algebraic equations

$$i_{ld}^* = F_{i_{od}} - \omega_n C_f v_{oq} + k_{pv}(v_{od}^* - v_{od}) + k_{iv}\emptyset_d$$

$$i_{lq}^* = F_{i_{oq}} - \omega_n C_f v_{od} + k_{pv}(v_{oq}^* - v_{oq}) + k_{iv}\emptyset_q$$

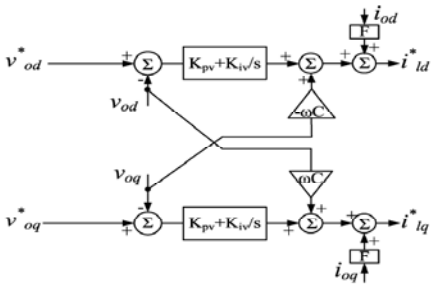


Fig. 3 : Voltage controller in autonomous mode.

D. Current Controller

The PI current controller structure is shown in Fig. The corresponding state-space model is

$$\dot{\gamma_d} = i_{ld}^* - i_{ld},\ \dot{\gamma_q} = i_{lq}^* - i_{lq} \quad (7)$$

$$v_{ld}^* = -\omega_n L_f i_{lq} + k_{pc}(i_{ld}^* - i_{ld}) + k_{ic}\gamma_d$$

$$v_{lq}^* = \omega_n L_f i_{ld} + k_{pc}(i_{lq}^* - i_{lq}) + k_{ic} \quad (8)$$

Where k_{pv} and k_{iv} are the PI current controller parameters.

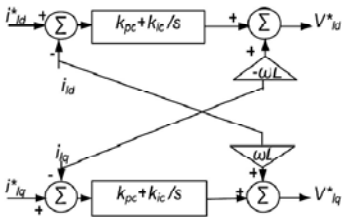


Fig. 4 : Current controller in autonomous mode.

E. LC Filter and Coupling Inductance

The LC filter and the coupling inductance model can be described with the following state equations, assuming that inverter produces the demanded voltage

$$\dot{\imath_{ld}} = -\frac{R_f}{L_f} i_{ld} + \omega i_{lq} + \frac{1}{L_f}(v_{ld} - v_{od})$$

$$\dot{\imath_{lq}} = -\frac{R_f}{L_f} i_{lq} + \omega i_{ld} + \frac{1}{L_f}(v_{lq} - v_{oq}) \quad (9)$$

$$\dot{v_{od}} = \omega v_{oq} + \frac{1}{C_f}(i_{ld} - i_{od}),\ \dot{v_{oq}} = \omega v_{od} + \frac{1}{C_f}(i_{lq} - i_{oq}) \quad (10)$$

$$\dot{i_{od}} = -\frac{R_c}{L_c} i_{od} + \omega i_{oq} + \frac{1}{L_c}(v_{od} - v_{bd})$$

$$\dot{\imath_{oq}} = -\frac{R_c}{L_c} i_{oq} - \omega i_{od} + \frac{1}{L_c}(v_{oq} - v_{bq}) \quad (11)$$

F. Complete Inverter Model:

To build the whole model of the system, the output variables of each inverter should be converted to the common reference frame using the following transformation:

$$f_{DQ} = T_i f_{dq}$$

$$T_i = \begin{bmatrix} cos(\delta_i) & -sin(\delta_i) \\ sin(\delta_i) & cos(\delta_i) \end{bmatrix} \quad (13)$$

The bus voltage that is the input signal to the inverter model should also be expressed on the common reference frame using reverse transformation.

G. Line Model:

The state equations of line current of *i*th line connected between nodes *j* and *k* can be expressed on a common reference frame as follows:

$$\dot{i_{lineDi}} = -\frac{\gamma_{linei}}{L_{linei}} i_{lineDi} + \omega i_{lineQi} + \frac{1}{L_{linei}}(v_b D_j - v_{bDk}) \quad (14)$$

$$i_{lineQi} = -\frac{\gamma_{linei}}{L_{linei}} i_{lineQi} + \omega i_{lineQi} + \frac{1}{L_{linei}}\left(v_b Q_j - v_{bQk}\right)$$

(15) Load Model:

The state equations of the RL load connected at ith node are given as follows

$$i_{lineDi} = -\frac{R_{linei}}{L_{linei}} i_{lineDi} + \omega i_{lineQi} + \frac{1}{L_{linei}}(v_{bDi})$$

.1

$$\bar{i}_{lineQi} = -\frac{R_{linei}}{L_{linei}} i_{lineQi} + \omega i_{lineDi} + \frac{1}{L_{linei}}\left(v_{bQi}\right)$$

.1

The load voltages are also given as follows:

$$\dot{v}_{bDi} = \omega v_{bQi} - \frac{1}{c_f}\left(i_{oDi} - i_{loadDi} \pm i_{lineDi,j}\right) \qquad (18)$$

$$\dot{v}_{bQi} = -\omega v_{bDi} - \frac{1}{c_f}\left(i_{oQi} - i_{loadQi} \pm i_{lineQi,j}\right) \qquad (19)$$

The sign in above equations depends on the current direction in the line.

III. ANT COLONY OPTIMIZATION

Ant Colony Optimization (ACO) is a population-based stochastic optimization technique developed by Dorigo & Stutzle [11] in 2004. ACO is a metaheuristic inspired by the foraging behavior of ant colonies. By marking the paths they have followed with pheromone trails, ants are able to communicate indirectly and find the shortest distance between their nest and a food source when foraging for food. When adapting this search metaphor of ants to solve discrete combinatorial optimization problems, artificial ants are considered to explore the search space of all possible solutions. The ACO search begins with a random solution (possibly biased by heuristic information) within the decision space of the problem. As the search progresses over discrete time intervals, ants deposit pheromone on the components of promising solutions[12]. In this way, the environment of a decision space is iteratively modified and the ACO search is gradually biased towards more desirable regions of the search space, where optimal or near-optimal solutions can be found. Due to its robustness in solving these problems, ACO has recently been applied to, and obtained some encouraging results for, real world engineering problems, such as the design of optimal water distribution systems.

As is the case with other metaheuristics, ACO can be linked with existing simulation models of power system, regardless of their complexity, when solving a power sharing problem. In addition, the unique way in which ACO problems are represented by using an equation makes ACO inherently suitable for handling power sharing and stability problems. In this section, the novel formulation that enables ACO to be applied to power sharing problems.

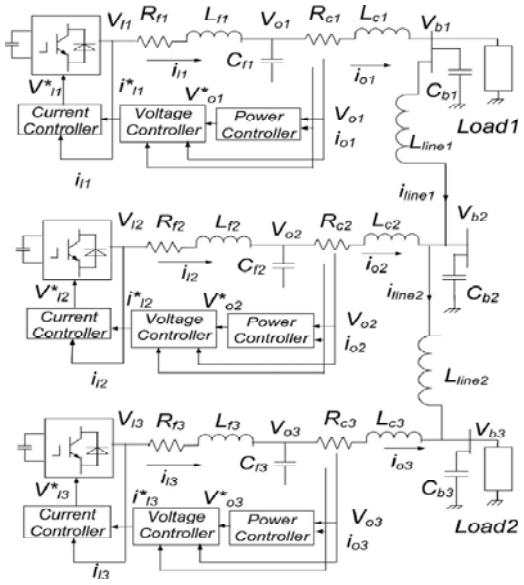


Fig. 5 : Circuit diagram of three inverter-based microgrid

IV. RESULTS AND DISCUSSION

Nonlinear model were developed using a MATLAB code to study the stability of an inverter-based microgrid when it is working in autonomous mode.

A. *Autonomous Mode*

In the autonomous mode, three inverter-based DGs (10 kVA) are connected with two loads through series *LC* filter, coupling inductance *Lc* , and lines as shown in Fig. 5. Each DG unit is represented by dc voltage source and VSI. A resistive load of 5.8 kW (25 Ω per phase) at load 1 and 7.3 kW (20 Ω per phase) at load 2 are considered as an initial operating point. The inverters are controlled to share the real and reactive powers over the lines. First, initial steady-state conditions of the system are obtained using a general power flow program. Second, nonlinear time-domain simulations have been carried out at two different disturbances to examine the effectiveness of the optimal settings of the proposed controllers and power sharing coefficients.

The first one is a step change in real power and the second is a fault disturbance at load 1. The results are given as follows. Figs.7 show the system response under fault disturbance at load 1. It can be seen that the system has satisfactory damping characteristics. Step change of 3.8 kW real power has also been applied. Figs. 6 show the system's response under this disturbance. Fig. 29 depicts the output voltage response of all the three inverters for a step load change.

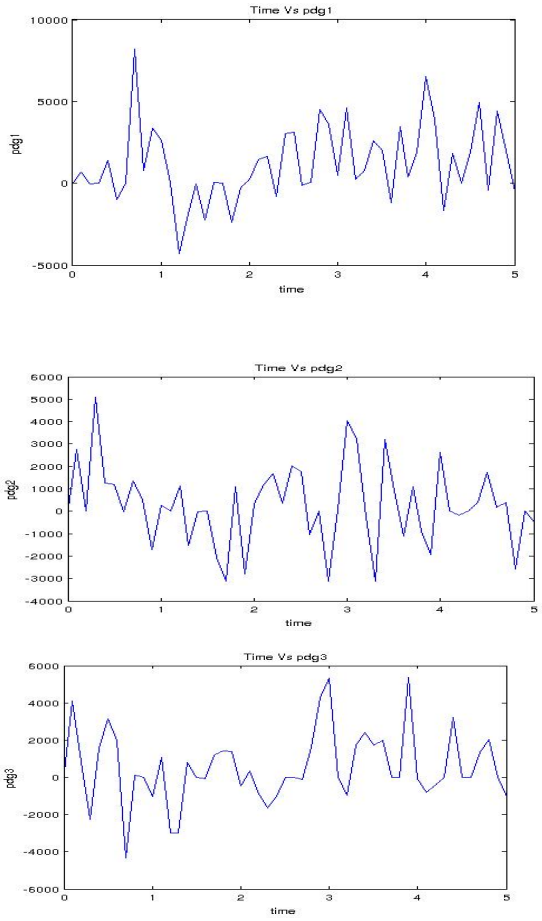
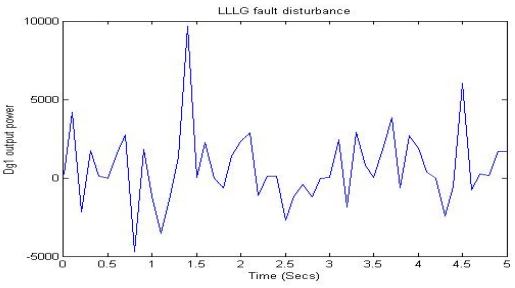


Fig. 6 : Behaviour of the system under step load change



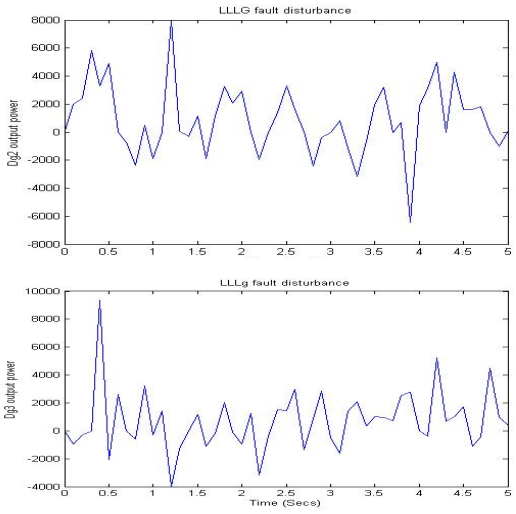


Fig. 7 : Behaviour of the system under fault disturbance at load 1

V. CONCLUSION

In this paper, nonlinear model of a microgrid operating in autonomous mode have been presented. The design of different controllers, filter, and power sharing coefficients has been formulated as an optimization problem. ACO technique is employed to search for the optimal settings of the optimized parameters. Different disturbances have been applied to demonstrate the effectiveness of the proposed design approach. System stability has been analyzed using nonlinear time-domain simulations. In an autonomous mode, step change and fault disturbances have been used to verify the system stability. The results confirm the effectiveness of the proposed ACO-based approach for optimizing the parameters of PI controllers, filter, and power sharing coefficients that achieve the satisfactory system performance under different disturbances.

REFERENCES

[1] R. Lasseter and P. Piagi, "Microgrid: A conceptual solution," presented at the Ann, Power Electron. Specialists Conf., Aachen, Germany, Jun.20–25, 2004

[2] J.Carrasco, L. Franquelo, J.Bialasiewicz, E. Galvan,R.Guisado, M. Prats,J. Leon, and N. Moreno-Alfonso, "Power electronic systems for the grid integration of renewable energy sources: A

survey," *IEEE Trans. Power Electron.*, vol. 53, no. 4, pp. 1002–1016, Aug. 2006.

[3] N. Pogaku, "Analysis, control and testing of inverter-based distributed generation in standalone and grid-connected applications," Ph.D.dissertation, Elect. Electron. Eng., Imperial College London, Univ.London, London, U.K., 2006.

[4] J. Baroudi, V. Dinavahi, and A. Knight, "A review of power converter topologies for wind generators," *Renew. Energ.*, vol. 32, no. 14, pp. 2369– 2385, Nov. 2007.

[5] R.Wai, W.Wang, andC. Lin, "High-performance stand-alone photovoltaic generation system," *IEEE Trans. Ind. Electron.*, vol. 55, no. 1, pp. 240–250, Jan. 2008.

[6] B. Panigrahi, A. Abraham, and S. Das Eds., *Computational Intelligence in Power Engineering*. Berlin, Germany: Springer-Verlag, 2010.

[7] N. Pogaku, M. Prodanovic´, and T. Green, "Modelling analysis and testing of autonomous operation of an inverter-based microgrid," *IEEE Trans.Power Electron.*, vol. 22, no. 2, pp. 613–624, Mar. 2007.

[8] E. Barklund, N. Pogaku, M. Prodanovic´, C. Hernandez-Aramburo, and T. Green, "Energy management in autonomous microgrid using stability constrained droop control of inverters," *IEEE Trans. Power Electron.*,vol. 23, no. 5, pp. 2346–2352, Sep. 2008.

[9] Y. Mohamed and E. El-Saadany, "Adaptive decentralized droop controller to preserve power sharing stability of paralleled inverters in distributed generation microgrid," *IEEE Trans. Power Electron.*, vol. 23, no. 6,pp. 2806–2816, Nov. 2008.

[10] K. De Brabandere, B. Bolsens, J. Van Den Keybus, A. Woyte, J. Driesen, and R. Belmans, "A Voltage and frequency droop control method for parallel inverters," *IEEE Trans. Power Electron*, vol. 22, no. 4, pp. 1107–1115, Jul. 2008.

[11] Dorigo M., Di Caro G., *The ant colony optimization metaheuristic*, in Corne D., Dorigo M., Glover F., *New Ideas in Optimization*, McGraw-Hill, p. 11-32, 1997.

[12] C.M. Colson, M.H. Nehrir, & C. Wang, "Ant colony optimization for microgrid multi-objective power management", *Proceedings of the 2009 IEEE/PES Power systems conference and Exposition,2009.*R.

◈◈◈

A New method of Extracting Fetal Electrocardiogram using Wavelet Transform and Genetic Algorithm

Vinolia Anandan[1] & Murugesan C[2]

Anna University of Technology Tirunelveli, Tirunelveli, India
E-mail : [1]vinojayan@gmail.com, [2]murugesan_dd@yahoo.com

Abstract **-** A new method for extracting Fetal ECG (FECG) from Abdominal ECG (AECG) using Wavelet transform and Genetic Algorithm (GA). The method used in this paper is very simple one and is found to yield good result. The performance and validity of the proposed method have been confirmed by computer simulations and experiment on real-world ElectroCardioGram (ECG) data. The result which was obtained appears to agree with the standard FECG signals.

Keywords— *FECG, AECG, ECG, GA.*

I. INTRODUCTION

A. *Problems in extracting FECG signal*

Every year about one out of 125 babies are born with some form of congenital heart defects. The Cardiac anomalies may occur due to a genetic syndrome, inherited disorder, or environmental factors such as infections and drug misuse. In any case, the regular monitoring of fetal heart and the early detection of cardiac abnormalities can help children's heart disease specialist to prescribe proper medication in time or to consider the necessary precautions during delivery or after birth [9]. The permanent appearance of mother ECG signal in which the amplitude is 5-20 times greater than that of FECG is considered to be an annoying one. The noise amplitude usually bigger than the amplitude of FECG signal is an issue. Recorded signal depends on the location of electrodes on mother abdomen. Mothers breathing are one of the permanent noise resources affecting the recorded signal. Permanent appearance of low frequency baselines noise, and also the mother's motion affect the recorded signal [5].

Fetal ECG is much weaker than the other interfering bio signals. Moreover, from the signal processing perspective, there is no specific domain (time, frequency, space) in which the fetal ECG can be totally separated from the interfering signals as shown in figure 1. Due to overlap of the fetal signals and interferences/ noises in different domains the methods that use the information in only one of these domains do not usually succeed in extracting the fetal ECG [9].

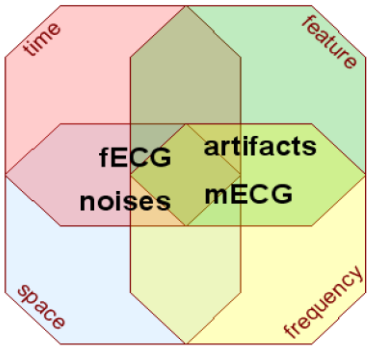


Fig. 1 The overlap of fetal ECG signals with the maternal ECG and other interferences and noise

The amplitude and frequency range of bio-signals that can interfere with fetal cardiac signal are shown in figure 2. The labels in this figure stand for the maternal electrocardiogram (mECG), electroencephalogram (mEEG), electrohistrogram (mEHG), Electrooculogram (mEOG), electromyogram (mEMG) and the fetal electrocardiogram (fECG) [4].

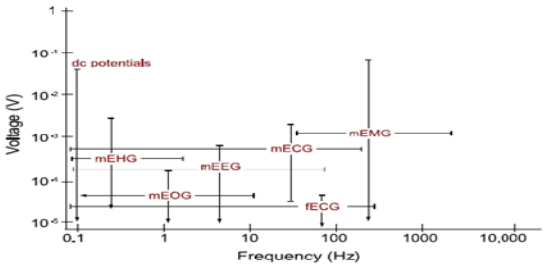


Fig. 2 Amplitude and frequency range of bio-signals

B. *The Fetal Heart*

The Heart is among the first organs developed in the fetus and undergoes a considerable amount of growth in the very early stages of pregnancy as in figure 3 . The heart is believed to begin beating by the 22nd day of life. The FECG signal can be recorded from the maternal abdomen as early as the eighteenth to twentieth week after conception. In the maternal

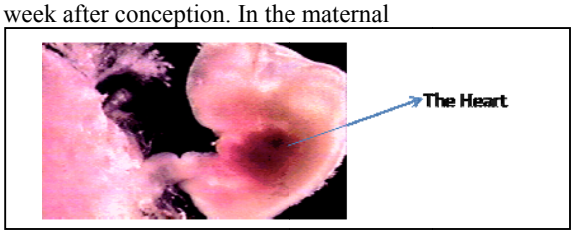


Fig. 3 Fetus and its heart in the early stages of development

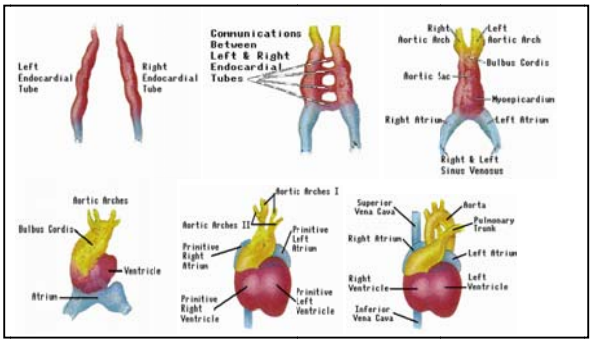


Fig. 4 Development of the fetal heart during gestation

abdomen compartments, the skin and the subcutaneous fat have a poor conductivity about ten times smaller than the muscle tissue. The above different tissues and layers form a so called volume conductor in which FECG signals propagate up to the maternal body surface. The very low conductivity vernix caseosa layer formed between the 28th and 32nd weeks of gestation. However for the normal pregnancies the layer slowly dissolves in the 37th to 38th weeks of pregnancy. The anatomy of the fully developed fetal heart is shown in figure 5.

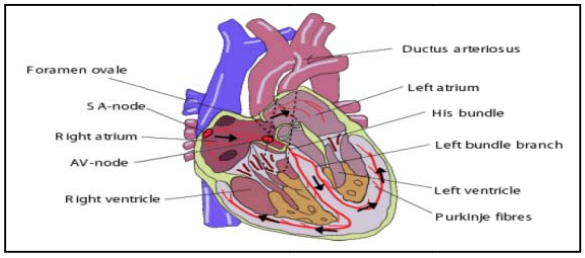


Fig. 5 The anatomy of the fetal heart

After birth the left ventricle pumps blood to the body and the right ventricles pumps the blood to the lungs for acquiring oxygen. However, for the fetus the fetal oxygen is supplied by the placenta; therefore blood is no longer pumped to the lungs for this purpose. Instead both ventricles pump the blood throughout the body [9].

Normal fetal heart rate

It is wonderful to observe how the fetus grows from conception to birth. The miraculous journey of the fetus from week 1 to week 42 is divided into three stages (Trimesters) for the sake of convenience. Depending upon the weight of the mom, position of the baby and location of the placenta, baby's heartbeats can be listen with the help of a stethoscope or fetoscope, at about 18-20 weeks, or even at about week 12. Normal fetal heart rate is quite faster than an adult's normal heart rate. Initially, in the 5th week, the fetal heart rate is almost the same as the mother's heart rate (80-85 beats per minute or bpm). Then it increases at an amazing rate, at about 3 beats per minute per day during the first month. Then again after 9th week, it settles down to about 150 bpm (average) at around 12th week. The normal fetal heart rate range after 12th week is 120-160 bpm. Table I shows the normal fetal heart rate. During the third trimester, it does not vary much. During the last (third) trimester, normal fetal heart rate slows down slightly. Once the heart beat of the fetus is heard, chances of miscarriage are very low. A fetal heart rate range of 100 – 160 is considered as normal. If baby's heart beat is not audible for a week or more, miscarriage is likely to happen. The normal fetal heart rate may vary naturally during movements. Fetal heart rate during sleeping and during other activities can be slightly different. It is normal to have different fetal heart rate at different periods of time, during a day. Gender can't be predicted with the help of fetal heart rate [19].

TABLE I

NORMAL FETAL HEART RATE CHART

Fetal Age / Size of the Fetus	Normal Fetal Heart rate
2 mm embryo and gestational sac diameter of 20 mm	75 bpm
5 mm embryo and gestational sac diameter of 30 mm	100 bpm
10 mm embryo	120 bpm
15 mm embryo	130 bpm
5 Weeks (Beginning)	80-85 bpm
5 Weeks	starts at 80 and ends at 103 bpm
6 Weeks	starts at 103 and ends at 126 bpm

7 Weeks	starts at 126 and ends at 149 bpm
8 Weeks	starts at 149 and ends at 172 bpm
9 Weeks	155-195 bpm (average175 bpm)
12 Weeks	120-180 bpm (average150 bpm)

C. *Electrical activity of the heart*

Action potential

Action potential is the electrical signal that accompanies the mechanical contraction of a single cell when stimulated by an electrical current (neural or external [8].

Resting potential

Nerve and muscle cells are encased in a semi permeable membrane that permits selected substances, while others are kept out. In their resting state, membrane of excitable cells readily permit the entry of K^+ and Cl^- ions, but effectively block the entry of Na^+ ions(the permeability for K^+ ion is 50 to 100 times that for Na^+). The outside of the cell is more positive than the inside of the cell. To balance the charge, additional K^+ ion enter the cell, causing higher K^+ concentration inside the cell than outside. A state of equilibrium is established with a potential difference, with the inside of the cell being negative with respective to the outside. A cell is in its resting state is said to be polarised [8].

Depolarization

When a cell is excited ionic currents or an external stimulus, the membrane changes its characteristics and begins to allow Na^+ ions to enter the cell. This movement of Na^+ ions constitutes an ionic current, which further reduces the membrane barrier to Na^+ ions. This leads to Avalanche effect: Na^+ ions rush in to the cell. K^+ ions try to leave the cell as they were in higher concentration inside the cell in the preceding resting state, but cannot move as fast as the Na^+ ions. The net result is that the inside of the cell becomes positive with respect to the outside due to an imbalance of K^+ ions. An excited cell displaying an action potential is said to be depolarized; the process is called Depolarization [8].

Repolarization

Mechanism for repolarization lies in the time dependence and voltage dependence of the membrane permeability changes for K^+ ions compared with that for Na^+ ions. However, the membrane permeability changes for Na^+ spontaneously decrease near the peak of the depolarization; where as those for K^+ ions are beginning to increase. Hence, during repolarization, the predominant membrane permeability is for K^+ ions. Because K^+ concentration is much higher inside the cell than outside, there is a net efflux of K^+ from the cell, which makes the inside more negative, thereby effecting repolarization back to the resting potential. The nerve and muscle cells repolarize rapidly, with an action potential duration of about 1 ms. Heart muscle cells repolarize slowly, with an action potential duration of 150 – 300ms [8].

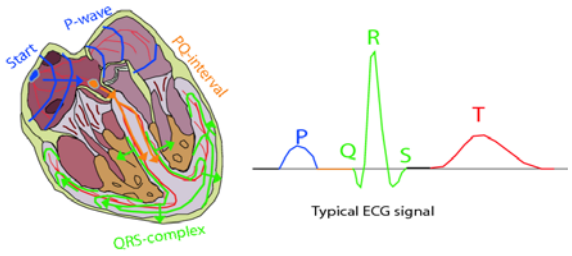


Fig. 5 The actuation cycle of the fetal heart

The SA node is the basic, natural cardiac pacemaker that triggers its own train of action potentials. The firing rate of the SA node is controlled by impulses from the autonomous Central nervous systems leading to the delivery of neuro transmitters. The action potential of a SA node propagates through the rest of the heart , causing particular pattern of excitation and contraction. The actuation cycle of the fetal heart is shown in figure 5.

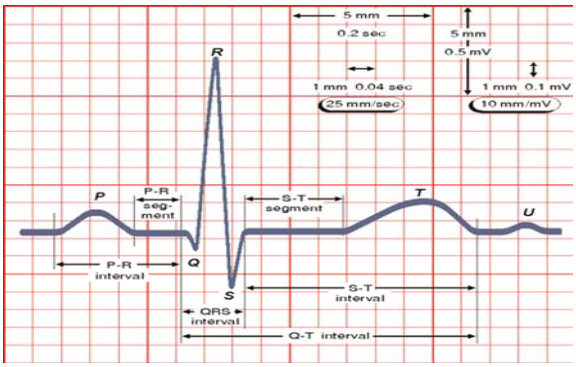


Fig. 6 Standard P,Q,R,S,T and U waveform of human heart

The sequence of events and waves in a cardiac cycle [8]

1. The SA node fires.
2. Electrical activity is propagated through the atrial musculature at comparatively low rates, causing slow- moving depolarization(contraction) of the atria. This results P-Wave in the ECG.

3. The excitation wave faces a propagation delay at the atrio-ventricular (AV) node, which results in a normally iso-electric segment after the P wave in the ECG, known as the PQ segment.
4. The bundle, the bundle branches, and the purkinje system of specialized conduction system of specialized conduction fibers propagate the stimulus to the ventricles at a high rate.
5. The wave of stimulus spreads rapidly from the apex of the heart upwards, causing rapid depolarization (contraction) of the ventricles. This results in the QRS wave of the ECG - a sharp biphasic or triphasic wave.
6. Ventricular muscle cells possess a relatively long action potential. The plateau portion of the action potential causes a normally iso-electric segment after the QRS, known as ST segment.
7. Repolarization (relaxation) of the ventricles causes the slow T wave. The standard P,Q,R,S,T and U waveform of human heart is as shown in figure 6.

TABLE II

AMPLITUDE-TIME RELATIONS IN MATERNAL AND FETAL ECG SIGNAL MEAN VALUES FROM 20 TRACES BETWEEN WEEK 36 AND 41 OF GESTATION [4]

	Mother	Fetal
QRSAmplitude (μV)	150	30
QRS width (ms)	100	50
PQ segment (ms)	70	45
P amplitude (%QRS)	20	10
P width (ms)	65	56
ST segment (ms)	110	70
T amplitude (%QRS)	30	25
T width (ms)	160	130

D. *Noise and Interferences*

High frequency noise in the ECG

The noise could be due to the instrumentation amplifiers, the recording system, pickup of ambient EM signals by the cables and so on. The signal also is corrupted by powerline interference and its harmonics, which may also be considered as a part of high frequency noise relative to the low frequency nature of the ECG signal as shown in figure 7 [8].

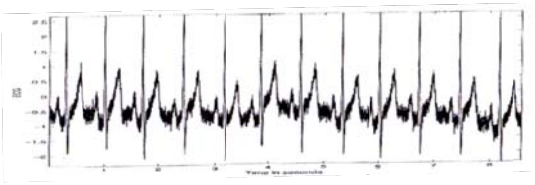


Fig. 7 ECG signal with high frequency noise

Motion artifact in the ECG

Figure 8. shows the low frequency artifacts and baseline drift may be caused in chest-lead ECG signal by coughing or breathing with large movement of the chest, or when an arm or leg is moved, in the case of limb-lead ECG acquisition. Baseline drift makes analysis iso-electricity of the ST segment difficult [8].

Fig. 8 ECG signal with low frequency artifact

Powerline interferences

The power spectrum of the signal should provide a clear indication of the presence of power line interference as an impulse or spike at 50Hz ; harmonics, if present will appear as additional spikes at integral multiples of the fundamental frequency. Figure 9 shows an segment of an ECG signal with 50 Hz interference. The regular or periodic structure of the interference, which rides on top of the ECG waves [8]

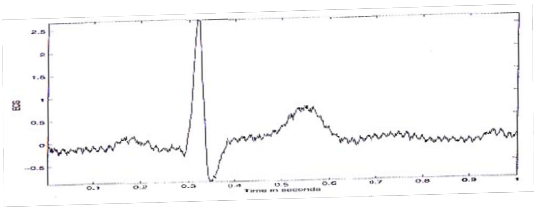


Fig. 9 ECG signal with power line interference

Maternal interference in FECG

Figure 10 shows ECG signal recorded from the abdomen of a pregnant woman , simultaneously recorded ECG from the women's chest comparing two, the abdominal ECG demonstrates multiple peaks (QRS complex) [8].

Fig. 10 Chest lead ECG(a) and Abdominal lead ECG(b)

II. STUDY AREA DESCRIPTION

A. Proposed Methodology

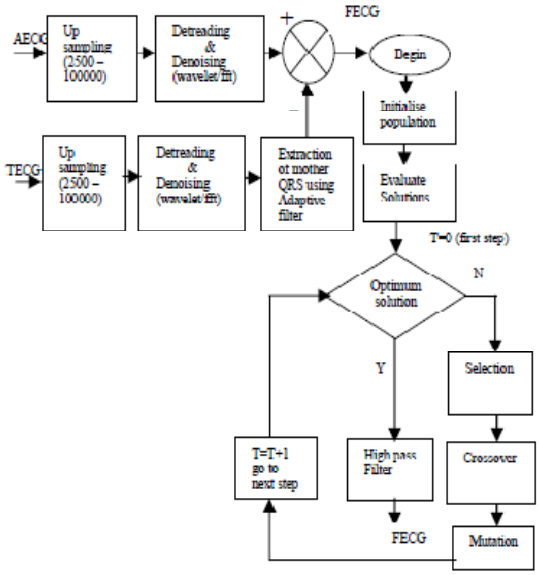


Fig. 11 Block diagram for FECG extraction using wavelet transform and genetic algorithm

The AECG signal received from the abdominal electrode upsampled (since in the consequent decomposition in wavelet transform the bandwidth reduced to half) , denoising and detreding (removal low frequency artifact mainly due to mothers breathing) are done by Fast fourier transform and Wavelet transform. Similarly Mother ECG (TECG) signal denoised and allow to filter for extracting the mother QRS. A constant K(t) is introduced in the TECG signal since mother QRS in TECG signal and mother QRS in AECG signal are not same (due to propagation delay etc). Then the FECG signal is extracted from the AECG signal. Afterwards the FECG signal is optimized by using Genetic Algorithm.The block diagram for FECG extraction using Wavelet transform and Genetic Algorithm is shown in figure 11.

B. Signal Analysis

Fourier Transform and Fast Fourier Transform

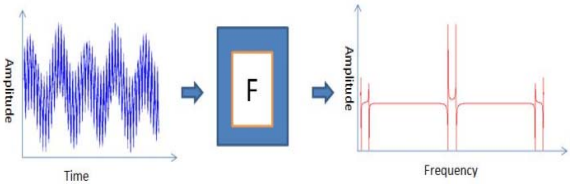


Fig. 12 Fourier Transform

Fourier analysis has a serious drawback. In transforming to the frequency domain, time information is lost [10]. But the Fast Fourier Transform(FFT) and Inverse Fast Fourier Transform (IFFT) are used to remove the noise as shown in figure 13 [12].

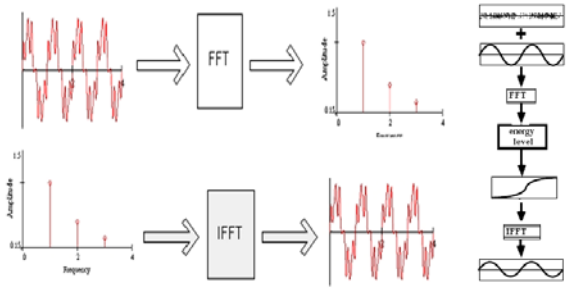


Fig. 13 Domain conversion and noise removal by using FFT and IFFT

Wavelet Transform

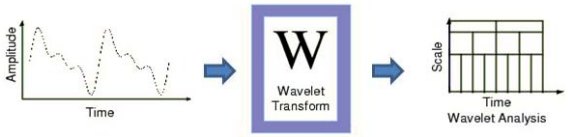


Fig. 14 Wavelet Transform

Wavelet analysis allows the use of long time intervals where more precise low frequency information and shorter regions with high frequency information is required. Wavelet analysis use a time-scale region as shown in figure 14. One major advantage afforded by wavelets is the ability to perform local analysis — that is, to analyze a localized area of a larger signal. Wavelet analysis is capable of revealing aspects of data that other signal analysis techniques miss, aspects like trends, breakdown points, discontinuities in higher derivatives, and self-similarity. Signal broken into a series of local basis functions called wavelets, which are scaled and shifted versions of the original (or Mother) wavelet. Wavelet means a small wave, the function that defines a wavelet integrates to zero. It is local in the sense that it decays to zero when sufficiently far from its centre.

Wavelet analysis is the breaking up of a signal into shifted and scaled versions of the original (or mother) wavelet as shown in figure 15. Continuous wavelet transform is the sum over all time of the signal multiplied by scaled, shifted version of the wavelet. This process produces the wavelet coefficients that are a function of scale and position.

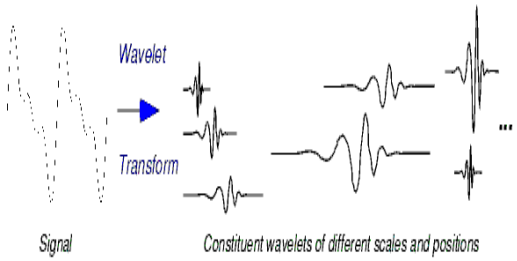


Fig. 15 Wavelets with different scales and positions

The scales and positions are chosen based on powers of two — so-called dyadic scales and positions — then the analysis will be much more efficient and just as accurate. This is obtained from the Discrete Wavelet Transform (DWT) analysis. In wavelet analysis, approximations and details are taken for consideration. The filtering process, at its most basic level is shown in figure 16. Here the high frequency signals (details) are treated as noise. The noises are removed at each level of decomposition. The low frequency signals are further allowed to consequent decomposition. To overcome the aliasing effect during multilevel decomposition the signal has to be upsampled.

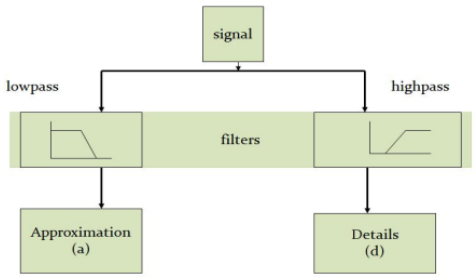


Fig. 16 Filtering process(Approximation and Details)

The wavelet expansion of a signal x(t) has the following expression [22]:

$$x(t) = \sum C_{j0k}\, \varphi_{j0k}(t) + \sum_{j=j0} \sum_{k} d_{jk}\, \psi_{jk}(t) \qquad (1)$$

Equation (1) shows that there are 2 terms. The first one is the 'approximation' and the second one is the details. The details are represented by

$$d_{jk} = \int x(t)\, \psi^{*}_{jk}(t)dt \qquad (2)$$

and ψjk (t) called the wavelet function is given by

$$\psi_{jk}(t) = \frac{1}{\sqrt{2^{j}\psi\left(\frac{t-k2^{j}}{2^{j}}\right)}} \qquad (3)$$

The approximation coefficients are given by

$$C_{jk} = \int x(t)\, \varphi^{*}_{jk}(t)dt \qquad (4)$$

$\varphi_{jk}(t)$ is called scaling function and is given by

$$\varphi_{jk}(t) = \frac{1}{\sqrt{2^{j}\varphi\left(\frac{t-k2^{j}}{2^{j}}\right)}} \qquad (5)$$

Daubechies Wavelet transform basically tries to 'smoothen' this signal by cleaning the rippling. After a number of such repeated cleaning process the 'residue ' signal left out is evidently the lowest- frequency or the slowest varying signal , the baseline.

The analysis is carried out in three steps:

Step1: In this step the abdominally recorded ECG signal is up sampled to improve the resolution at higher level of decomposition

Step2: In this step the signal vector is subjected to order of Daubechies Wavelet transform basically the wavelet transform is performed level by level, and at each level the approximation coefficients.

Step3: The signal vector is frequency analyzed each time to find the frequency range. If it is not close enough to the DC level, then step 2 is repeated increasing the level. After a certain time the vector will be found to be sufficiently close to the DC side, and it is concluded that this now plotted against the time would give the baseline successfully.

Adaptive filtering

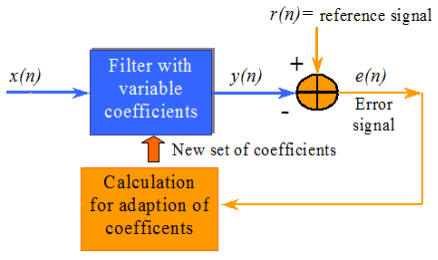


Fig. 17 Adaptive filter

An adaptive filter is a filter that self-adjusts its transfer function according to an optimization algorithm driven by an error signal. The purpose of an adaptive filter in noise cancellation is to remove the noise from a signal adaptively to improve the signal to noise ratio as shown in figure 18.

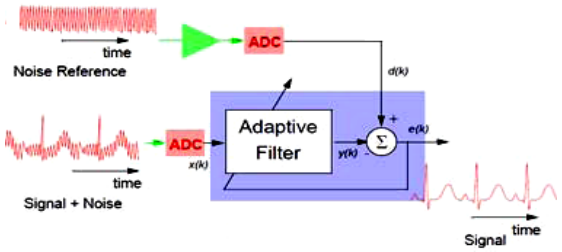


Fig. 18 Noise removal by adaptive filter

The usage of adaptive filters is one of the most popular proposed solutions to reduce the signal corruption caused by predictable and unpredictable noise [4].

Extraction of FetalECG

In this study there are two models, presented the model for signal taken by Thoracic Electrodes (TECG) represented by $'x_T(t)'$ and the model for signal taken by Abdominal Electrodes (AECG) represented by $'x_A(t)'$: [4]

$$x_A(t) = M_a(t) + F_a(t) + N_a(t) + \eta_a(t) \quad (6)$$

$$x_T(t) = M_b(t) + N_b(t) + \eta_b(t) \quad (7)$$

where:

$M_a(t)$ and $M_b(t)$ = Pure maternal ECG,

$F_a(t)$ = A pure foetal ECG

$\eta_a(t) = \eta_b(t)$ are low-rank or structured noise representing other biological sources that contaminate the ECG and the

$N_a(t)$ and $N_b(t)$ = Full-rank observation noise that always exist in physiological Measurements. As mentioned above the foetal ECG is very weak among the maternal ECG and noise, for this reason it appeared only in Eq. 6 and neglected in Eq. 7. By Some simplifications $\eta_a(t)$ and $\eta_b(t)$ can be eliminated by a high pass filter therefore (6) and (7) simplify to:

$$x_A(t) = M_a(t) + F_a(t) + N_a(t) \quad (8)$$

$$x_T(t) = M_b(t) + N_b(t) \quad (9)$$

The aim here is to extract pure Fa (t) which can be obtained by subtracting Eq. 9 from 8 in these equations $M_a(t) \neq M_b(t)$ and $N_a(t) \neq N_b(t)$, for this reasons use factor K(t) to equalize then get the equation:

$$K(t)x_T(t) = K(t)M_b(t) + K(t)N_b(t) \quad (10)$$

Then subtract Eq. 10 from 8 we get:

$$[x_A(t) - K(t)x_T(t)] = [M_a(t) - K(t)M_b(t)] + F_a(t) + [N_a(t) - K(t)N_b(t)] \quad (11)$$

Then pure FECG can be obtained from the following equation:

$$F_a(t) = [x_A(t) - K(t)x_T(t)] - [M_a(t) - K(t)M_b(t)] - [N_a(t) - K(t)N_b(t)] \quad (12)$$

After cancelling the maternal signal and noise we get:

$$F_a(t) = [x_A(t) - K(t)x_T(t)] \quad (13$$

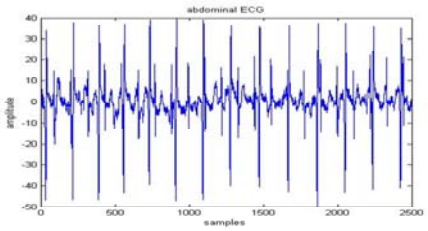


Fig. 19 Abdominal ECG

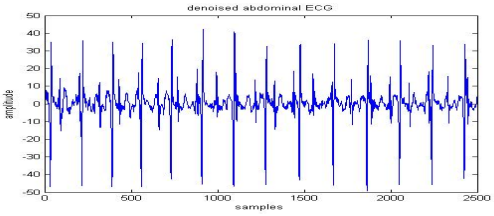


Fig. 20 Denoised abdominal ECG

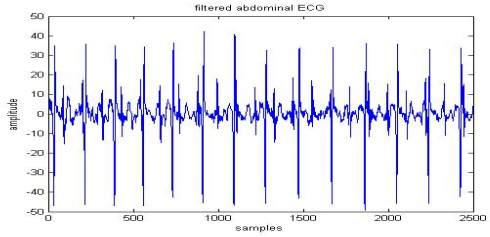


Fig. 21 Filtered abdominal ECG

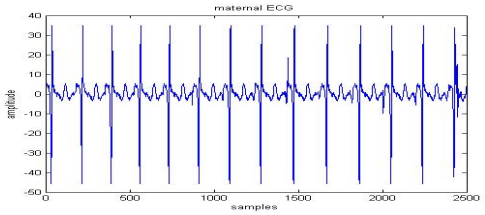


Fig. 22 Thoracic ECG

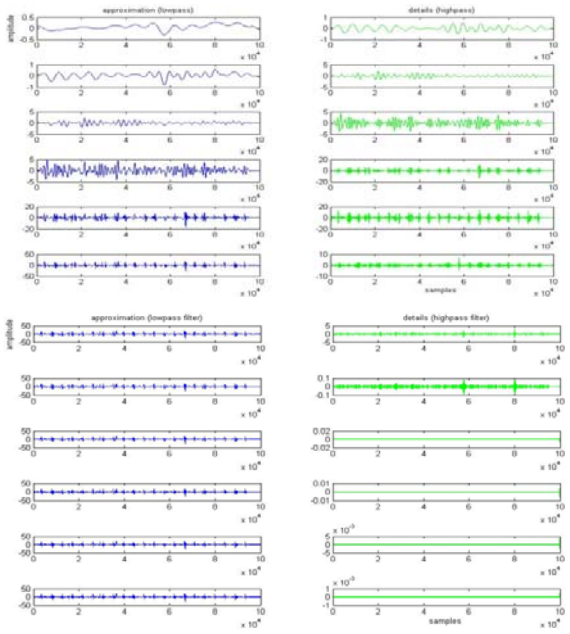


Fig. 23 12 Level wavelet decomposition

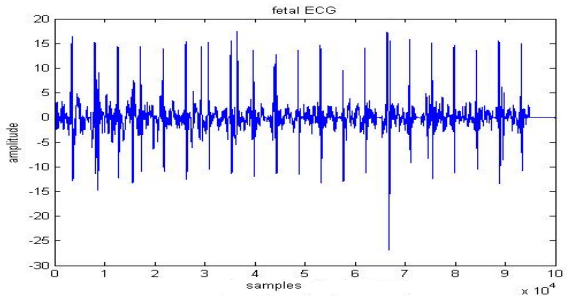


Fig. 24 Fetal ECG to Genetic Algorithm

TABLE III

COMPARISION OF DIFFERENT METHODS [9]

Method	FIR/IIR Filtering	Wavelet Denoising	Adaptive Filtering	Non-linear Filtering
Overall Performance	low	medium	medium	medium
SNR improvement	low	medium	medium	medium/ high
Computational cost	low	low	medium	high
Real time	yes	yes	yes	no
Implementation complexity	simple	medium	medium	complex
Operator interaction	not required	not required	sometimes required	required

C. Genetic Algorithm

Biological background

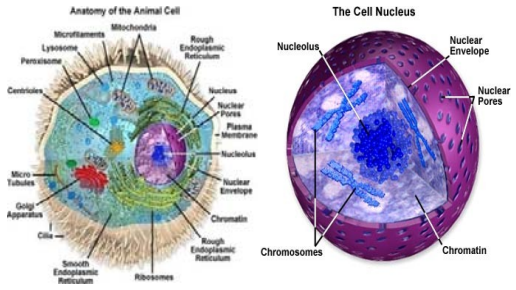


Fig. 25 Animal cell and cell nucleus

Every animal cell is a complex of many small factories working together. The center of this all is the cell nucleus shown in figure 25. The nucleus contains the genetic information. Genetic information is stored in the chromosomes and each chromosome is built of DNA shown in figure 26(a). The chromosome is divided into parts called genes, Genes code for properties. The posibilities of the genes for one property is called: allele.The entire combination of genes is called genotype, a genotype develops to a phenotype. Alleles can be either dominant or recessive. Dominant alleles will always express from the genotype to the fenotype, recessive alleles can survive in the population for many generations, without being expressed. Every gene has an unique position on the chromosome called locus [21] .

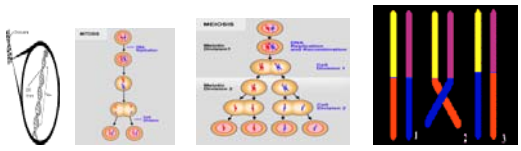


Fig. 26(a) Chromosome and genes (b) Reproduction of genetical information by Mitosis (c) Reproduction of genetical information by Meiosis (d) Cross over

Mitosis is copying the same genetic information to new offspring. Mitosis is the normal way of growing of multicell structures, like organs shown in figure 26(b). Meiosis is the basis of sexual reproduction, after meiotic division 2 gametes appear in the process shown in figure 26(c). In reproduction two gametes conjugate to a zygote wich will become the new individual. During reproduction errors occur, most important errors are Recombination (cross-over) and Mutation [21].

TABLE IV

EXPLANATION OF GENETIC ALGORITHM TERMS [14]

Genetic Algorithms	Explanation
Chromosome (String, individual)	Solution (Coding)
Genes (bits)	Part of solution
Locus	Position of Gene
Alleles	Values of Gene
Phenotype	Decoded solution
Genotype	Encoded Solution

The GA is a stochastic global search method that mimics the metaphor of natural biological evolution At each generation, a new set of approximations is created by the process of selecting individuals according to their level of fitness in the problem domain and breeding them together using operators borrowed from natural genetics. Figure 27 shows the flow diagram of genetic algorithm.

Simple Genetic Algorithm

A pseudo-code outline of the SGA is given below. The population at time 't' is represented by the time-dependent variable 'P', with the initial population of random estimates being P(0) [11].

```
procedure GA
begin
        t = 0;
        initialize P(t);
        evaluate P(t);
        while not finished do
        begin
                t = t + 1;
                select P(t) from P(t-1);
                reproduce pairs in P(t);
                        evaluate P(t);
        end
end.
```

Fitness

To help maintain diversity or differentiate between similar individuals, raw objective scores are sometimes scaled to produce the final fitness scores [18]

Rank - no scaling

Linear scaling (fitness proportionate) – normalizes based on min and max fitness in population

Sigma truncation scaling - normalizes using population mean and standard deviation, truncating low-fitness individuals

Sharing (similarity scaling) - reduces fitness for individuals that are similar to other individuals in the population.

In the canonical genetic algorithm fitness is denoted by 'fi/f' where 'fi' is the evaluation associated with string 'i' and 'f' is the average evaluation of all the strings in the population [17].

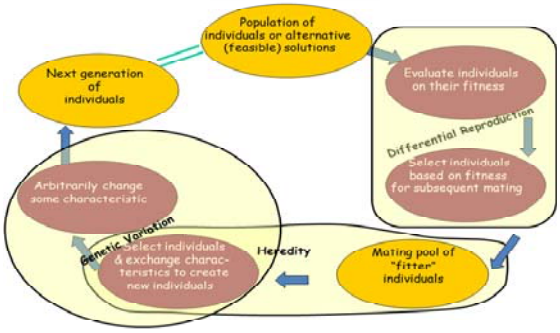


Fig. 27 Flow diagram of genetic Algorithm [20]

Performance is done through an objective function that characterises an individual's performance in the problem domain. In the natural world, this would be an individual ability to survive in the present environment. During the reproduction phase, each individual is assigned a fitness value derived from its raw performance measure given by the objective function. Highly fit individuals, relative to the whole population, have a high probability of being selected for mating whereas less fit individuals have a correspondingly low probability of being selected. Genetic operators manipulate the characters (genes) of the chromosomes directly, using the assumption that certain individual's gene codes, on average, produce fitter individuals.

Selection

The selection scheme determines how individuals are chosen for mating, based on their fitness scores [18].

Rank - pick the best individuals every time

Roulette wheel (proportionate) - probability of selection is proportional to fitness

Tournament - initial large number are selected via roulette wheel, then the best ranked are chosen

Stochastic - various methods of replenishment of less fit stock (useful) or initial selection (not useful)

Elite - in combination with other selection schemes, always keep the fittest individual around

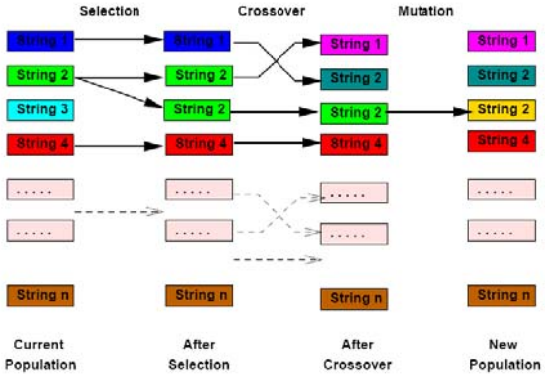


Fig. 28 Various stages of Genetic Algorithm [14]

Crossover

The recombination operator is used to exchange genetic information between pairs, or larger groups, of individuals. The simplest recombination operator is that of single point crossover as shown in Table V. Single point, Double point and Uniform crossover are as shown in figure 29.

TABLE V

SINGLE POINT CROSSOVER

Chromosome 1	11011 | 00100110110
Chromosome 2	11011 | 11000011110
Offspring 1	11011 | 11000011110
Offspring 2	11011 | 00100110110

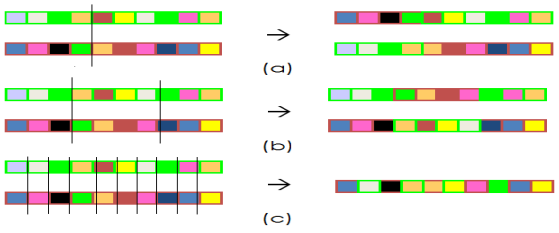


Fig. 29 (a) Single point crossover (b) Double point crossover

(c) Uniform crossover

Mutation

A further genetic operator, called mutation, is then applied to the new chromosomes, again with a set probability, 'Pm' . Mutation causes the individual genetic representation to be changed according to some probabilistic rule. In the binary string representation, mutation will cause a single bit to change its state, 0 1 or 1 0 as shown in table VI.

TABLE VI : MUTATION

Original offspring 1	**1101111000011110**
Original offspring 2	1101100100110110
Mutated offspring 1	1100111000011110
Mutated offspring 2	1101101100110100

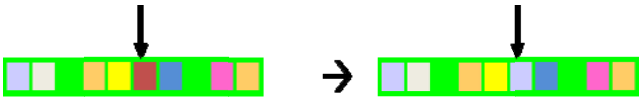


Fig. 30 Mutation

Mutation is generally considered to be a background operator that ensures that the probability of searching a particular subspace of the problem space is never zero. This has the effect of tending to inhibit the possibility of converging to a local optimum, rather than the global optimum.

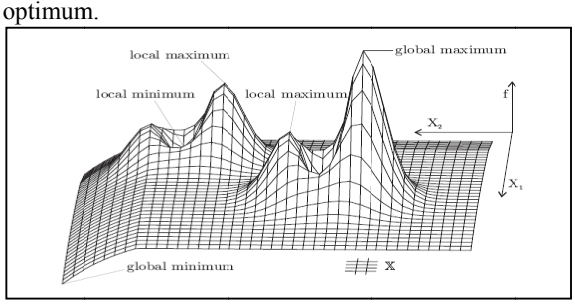


Fig. 31 Global optimum and Local optimum [15]

Replacement

Simple or generational GAs replace entire population, as per the dictates of the selection scheme, Steady state or online GAs use different replacement schemes [18]

Replace worst

Replace best

Replace parent

Replace random

Replace most similar (crowding)

In this way, the average performance of individuals in a population is expected to increase, as good individuals are preserved and bred with one another and the less fit individuals die out.

Outline of the basic Genetic Algorithm [13]

1 [Start] Generate random population of n chromosomes (suitable solutions for the problem)

2. [Fitness] Evaluate the fitness f(x) of each chromosome 'x' in the population

3. [New population] Create a new population by repeating following steps until the new population is complete

3.1[Selection] Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected)

3.2[Crossover] With a crossover probability cross over the parents to form a new offspring (children). If no crossover was performed, offspring is an exact copy of parents.

3.3[Mutation] With a mutation probability mutate new offspring at each locus (position in chromosome).

3.4[Accepting] Place new offspring in a new population

4. [Replace] Use new generated population for a further run of algorithm

5. [Test] If the end condition (for example number of populations or improvement of the best solution) is satisfied, stop, and return the best solution in current population

6. [Loop] Go to step 2

III. EVOLUTIONARY RESULT

The architecture used for this is a combination of a Wavelet transform, adaptive filter and Genetic Algorithm (GA). The process of exchanging elements among successful designs of GA also has a biological analog, which is referred to as crossover and mutation. The general framework and basic step of GA can be shown as figure 10 appeared at the right part. Number of variables - 2500, Range -200 to 200, Crossover rate – 1, Mutation rate - 1/2500, Migration rate - 0.2. The hybrid combination of wavelet transform and the GA provide the expected result.

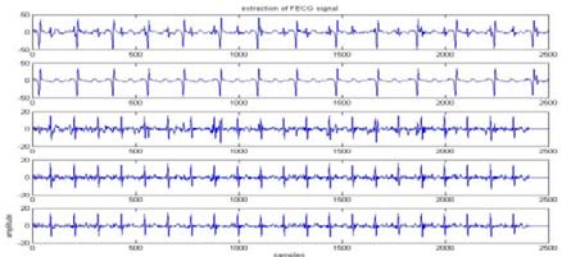


Fig. 32 AECG,TECG and FECG at various levels

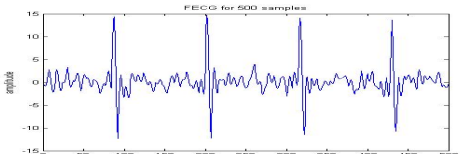


Fig. 33 FECG for 500 samples

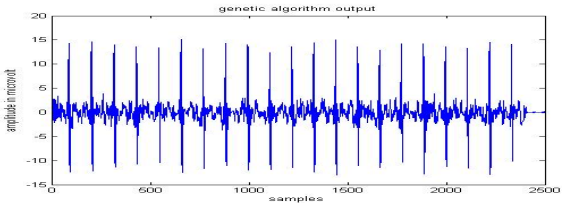


Fig. 34 Genetic Algorithm output

IV. CONCLUSION AND FUTURE SCOPE

The performance and validity of the proposed algorithm have been confirmed by computer simulations and experiment on real-world ECG data. The result which was obtained appears to agree with the standard Fetal ECG signals. In future the performance and the validation can be done by calculating the Signal to Noise Ratio (SNR) by using Synthetic ECG signal. It is also possible that the Genetic Algorithm portion of the project work can be replaced by various Soft Computing Techniques and the performance can be compared.

REFERENCE

[1] R.M. Clemente, J.L.C. Olivares, S.H. Mellado, Mar Elena and Isabel Roman (2011). "Fetal ECG Extraction from Maternal Abdominal ECG Using Neural Network" IEEE Transaction on Bio-Medical Engineering, Vol.58,No.2

[2] J.L.Camargo-Oliveres, R.M. Clemente, S.H. Mellado, M.M. Elena and I. Roman(2011) "The Maternal Abdominal ECG as Input to MICA in the fetal ECG Extraction Problem" IEEE Signal Processing letters, Vol.18, No.3

[3] M.A. Hasan, M.I. Ibrahimy and M.B.I. Reaz (2009) "Fetal ECG Extraction from Maternal Abdominal ECG Using Neural Network" Scientific Research, J.Software Engineering & Applications, 2009,2:330-334

[4] M.A.Suliman Ali and Xiao Ping Zeng (2010) "A Noval Technique for Extraction Foetal Electrocardiogram using adaptive filtering and Simple Genetic Algorithm" American Journal of Biostatistics1 (2): 75-81

[5] S.Sargolzaei, K.Faez, A.Sargolzaei (2008) 'Signal Processing Based Techniques for fetal Electrocardiogram Extraction' IEEE International Conference on BioMedical Engineering and Informatics

[6] H.Hassanpour and Amin Parsaei (2006) "Fetal ECG extraction using Wavelet Transform" International conference on computational intelligence for modeling control and automation and International conference on Intelligent Agents ,Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)

[7] R.M.Rao "Wavelet Transforms Intrduction to Theory and Application"

[8] R.M.Rangayyan 'Biomedical Signal Analysis – A Case study Approach"

[9] Reza Sameni "Extraction of fetal cardiac signals from an array of maternal abdominal recordings"

[10] Machel Misiti, Yves Misiti, Georges Oppenheim, Jean-Michel Poggi "Wavelet Tool box- For use with MATLAB"

[11] Andrew Chipperfield, Peter Fleming, Hartmut Pohlheim and Carlos Fonseca "Genetic Algorithm TOOL BOX"

[12] Charan Longton "Intutive Guide to Principles of communications"

[13] Pratibha Bajpai and Manoj Kumar "Genetic Algorithm - An approach to solve Global Optimaization Problems" Indian journel of Computer Science and Engineering Vol I NO 3 199-206

[14] Tom V. Mathew "Genetic Algorithm" IIT,Bombay

[15] Thomas Weise "Global Optimization Algorithms – Theory and Application"

[16] D.DeMoor(Ed.)daisy: database for the identification of systems, available online at:http://www.esat.kuleuvan.be/sista/.

[17] Darrell Whitely "A Genetic Algorithm Tutorial"

[18] Larry Yaeger "Tntroduction to Genetic Algorithms"

[19] Buzzle.com "Intelligent Life on the Web"

[20] Partha Chakroborthy "REACHing Optimum Designs through Processes Inspired by principles of Evolution"

[21] Peter Spijker "Basics of Genetic Algorithms and some Possibilities"

[22] Shubhajit Roy Chowdhury and Dipankar Chakrabarti (2010) "Daubechies Wavelet Decomposition based baseline wander correction of Trans-Abdominal Maternal ECG" 6th International Conference on Electrical and Computer Engineering (ICECE 2010)

◈◈◈

Diode-Clamped Based Indirect Matrix Multilevel Converter Topology

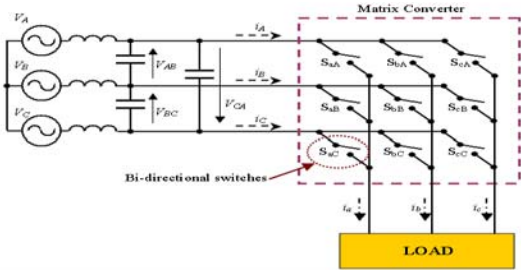
Y. Vijaya Suresh & M.Ashok Kumar,

Dept. of EEE, RGM College of Engg. & Tech. Nandyal, Andhra Pradesh, India
E-Mail: yadativijayasuresh@yahoo.com & ashok2asi@gmail.com

Abstract **-** This paper presents a novel sparse matrix converter configuration that needs no bulky energy storage capacitors or inductors, and employs only unidirectional power semiconductor switches. The source side bridge is switched at the line frequency and therefore, this allows the use of the low speed high power rating switches. This feature and the use of diode-clamped multilevel topology at the output stage make this new matrix converter suitable for medium to high power applications and improves the output voltage quality. Compared to the existing multilevel matrix converter topologies, the proposed topology requires no flying or extra capacitor, a smaller number of switches and a simpler modulation strategy. Simulation results for a 220v/60Hz input source and 20KVA load show that the proposed AC/AC topology has input unity displacement power factor, high (0.95 to 0.88) input power factor for different load Conditions and controllable output voltage at a desired frequency.

I. INTRODUCTION

Different circuit topologies have been developed to convert AC power at one frequency to AC power at another frequency. In conventional methods, the AC power is first rectified and then is converted back to AC power at a different frequency. An alternative approach is the Matrix Converter (MC) technology [1], where the input AC power is directly converted to output AC power using a matrix of bidirectional switching elements, Fig. 1. Both topologies can provide controllable sinusoidal output voltages and input currents. Compared to the conventional back to back converter, the MCs are compact, more reliable, and do not needs an extra voltage control strategy for the DC-link capacitor. Due to complicated modulation algorithms and the utilization of the bidirectional switches, conventional MCs have not been widely adopted. With respect to the early modulation techniques, a conceptually different control technique was introduced in 1983 [2] which led to the invention of the two stage direct power conversion topology. Based on this modulation technique a new class of converter topologies was invented which was simpler. These converters are also known as ”indirect MC (IMC)”, ”dual bridge MC” or ”sparse matrix converter” , [3]–[7]. These advances have made matrix converters more feasible and practical. Nevertheless, the application of matrix converters is still limited to low and medium voltage levels. Multilevel converter topologies, on the other hand, are well known solutions for medium and high power applications. Among various multilevel topologies, the most important ones are [8]: Diode-Clamped Multilevel Converter (DCMC) [9], Flying Capacitors Multilevel converters (FCMC) [10] and Cascaded Multilevel Converters (CMC). Therefore, various combinations of these topologies, known as Multilevel Matrix Converters (MMC), have been developed for high power direct AC to AC power conversion. The first attempt for developing MMCs incorporates the CMC topology, and was reported in [11]. The other approach uses the FCMC topology to form a MMC [12]. Although both techniques are innovative and interesting, they utilize excessive number of capacitors that also need to be balanced. This contradicts one of the MC objectives which was the elimination of capacitors from the converter topology. On the contrary, this paper presents a novel MMC that employs diode clamped multi-level topology in a sparse matrix converter configuration that does not utilize extra capacitors and also the cumbersome voltage balancing strategies are not necessary anymore. Similar to conventional converter topologies, the proposed topology uses unidirectional switches and consequently uses less switches than the two previously MMCs.

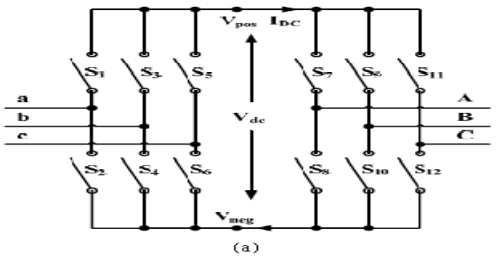


Fig/1 : Conventional indirect matrix converter topology.

A point worth mentioning would be that since the source-side bridge is switched at the source frequency, switching losses are limited to the lowest possible level. Furthermore, the modulation strategy used for this topology is simple and similar to the conventional converters and does not have the complexity of the existing MMCs. Simulation results show that the proposed AC/AC topology has input unity displacement power factor, high (0.95 to 0.88) input power factor for different load conditions and controllable output voltage at a desired frequency. The simulations are carried out for a 220v/60Hz input source and 20KVA load at 100Hz.

II. BASIC MATRIX CONVERTER TOPOLOGY AND INDIRECT MATRIX CONVERTERS (IMC)

Since the multilevel matrix converters are based on basic MC and IMC topologies, these basic topologies are reviewed in this section. The basic 3φ −3φ matrix converter topology is shown in Fig. 1. This configuration of bidirectional switches enables the connection of any of input phase a, b, or c to any of output phase A, B or C at any instant. Different modulation techniques such as triangular wave voltage command modulation [13] and space vector modulation (SVM) [14], [15] were developed for matrix converters. The indirect modulation technique was a conceptually different technique introduced in 1983 [2], which divided the MC into two parts. With the two partitions, it resembled the conventional back to back topology, but instead it uses the idea of a fictitious DC-link rather than a real rectified DC- link. The objective of the modulation strategy is to synthesize the output voltages from the input voltages and the input currents from the output currents. Thus, the matrix converter is described by an equivalent circuit combining current source rectifier and voltage source inverter connected through virtual dc link as shown in Fig. 2(a). The inverter stage has a standard 3φ - 3φ voltage source inverter topology consisting of six switches, S_7− $S_{12}$and rectifier stage has the same power topology with another six switches, S_1−S_6. Both power stages are directly connected through virtual dc-link and inherently provide bidirectional power flow capability because of its symmetrical topology.



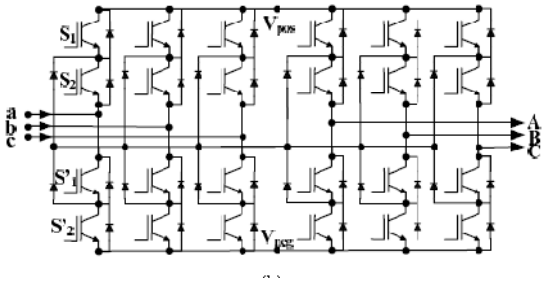
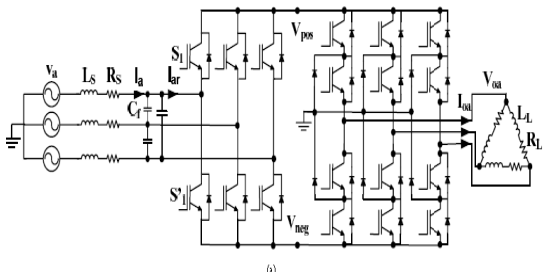


Figure 2. (a) Indirect Matrix Converter (b) One possible Multilevel matrix converter topology

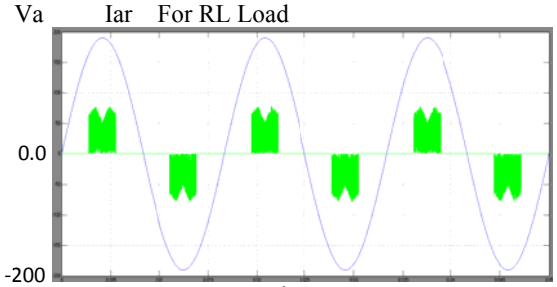
The main idea is to decouple the control of the input current and output voltage. This is done by two steps; first, by using SVM the rectifier is switched to form the input current, to keep the input power factor close to unity and to generate dc voltage at the dc bus. Secondly, based on this dc voltage the inverter is switched to generate controllable output voltages, [2], [14]. Having found the switching patterns for the equivalent circuit, it is quite straightforward to find the equivalent switching functions for the conventional matrix converter shown in Fig.1, [2]. Although the equivalent circuit shown in Fig.2(a), was originally used to find the modulation strategy for the conventional matrix converter, this concept can be physically implemented as a matrix converter topology for direct conversion. These types of converters are also known as ”indirect MC (IMC)”, ”dual bridge MC” or ”sparse matrix converter”. Different topologies are derived from this concept. For the IMC a conventional (two-quadrant switch) voltage-source type inverter is fed by a four-quadrant switch current-source type rectifier which is able to operate with positive and negative DC current for unipolar DC link voltage as required by the inverter stage. Since the dc link voltage is unipolar, the number of switches in the topology can be reduced, [4]. Another IMC topology is a hybrid of the matrix converter and the conventional back to back converter, [16]. In this configuration the bidirectional switches of the IMC topology are replaced with unidirectional ones.

III. EXISTING MULTILEVEL MATRIX CONVERTER TOPOLOGIES

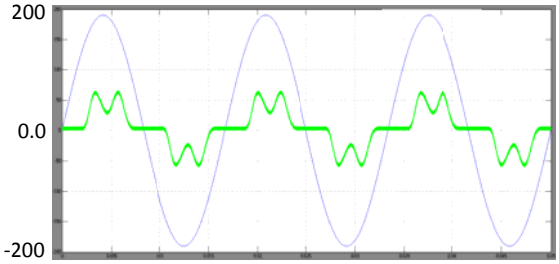
To increase the power level, achieve higher efficiency, and provide better output voltage quality, the multilevel concepts have been adapted to develop multilevel matrix converter topologies. The first attempt was inspired by CMC topology [11] which utilizes conventional matrix converter with four- quadrant switch cells. The other multilevel matrix



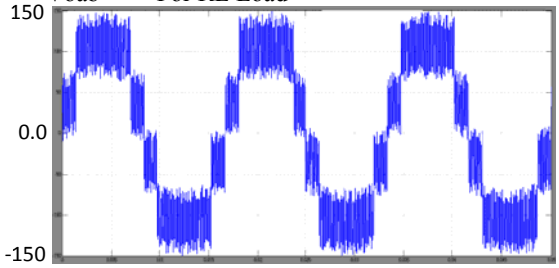
Va Iar For RL Load



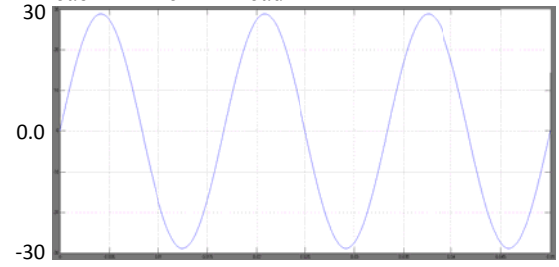
Va Ia For RL Load



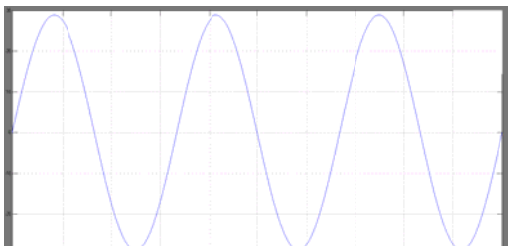
Voab For RL Load



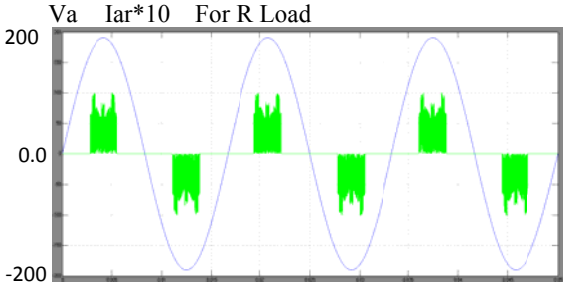
Ioab For RL Load



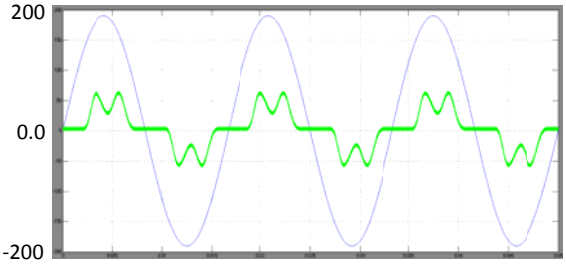
Ioab For RL Load



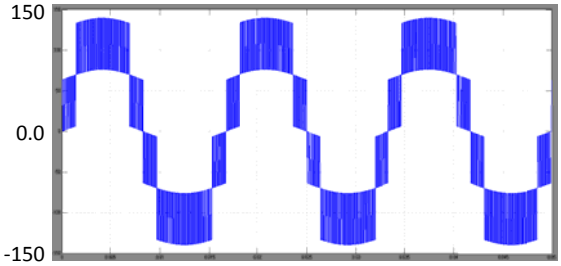
Va Iar*10 For R Load



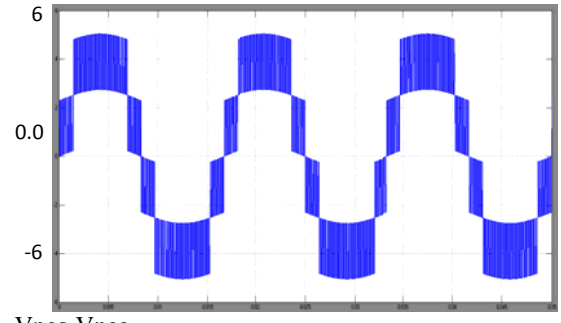
Va Ia*10 For R Load



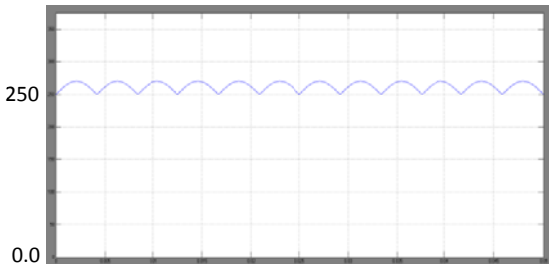
Voab For R Load



Ioab For R Load

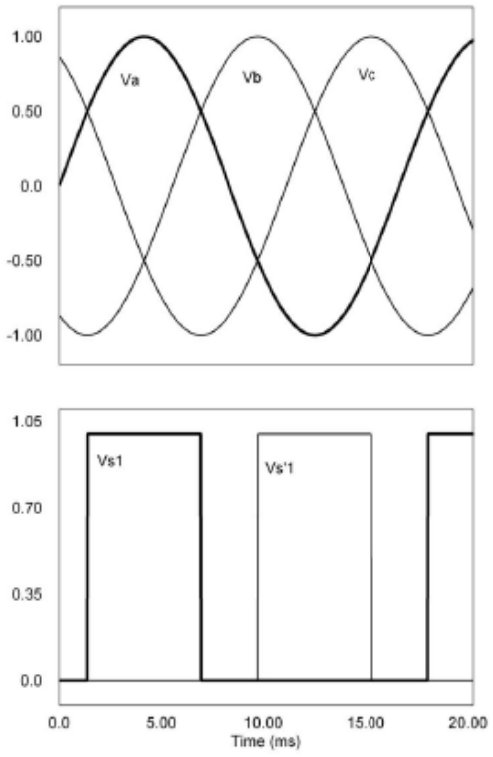


Vpos-Vneg For RL Load



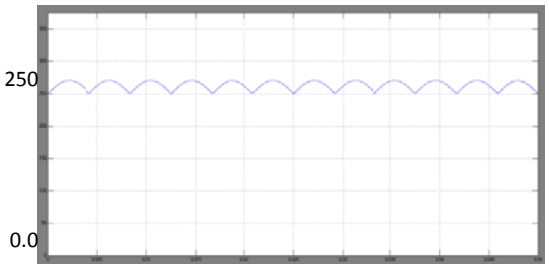
(b)

converter introduced in the literature [12] is a hybrid of the FCMC idea and the Matrix converter topology. Although both techniques are innovative and interesting, they incorporate capacitors that need to be balanced and consequently complicates the control process. On the contrary, in the conventional MCs the use of capacitors were avoided to allow a more compact form and greater longevity. Moreover, the other two existing multilevel matrix topologies, [11], [12], use 36 switches in addition to six and nine capacitors respectively, while the proposed topology in this paper uses only 24 switches.



Another shortcoming of the FCMC matrix converter is that the input current waveforms and input power factor are not controlled even thought it is one of the main objectives of the matrix converters.

Vpos-Vneg For R Load



(c)

Fig 3(a) indirect multilevel matrix converter topology (b) simulation results for R=5Ω L_L=3mH (c) simulation results for R= 50ΩL_L=0mH.

IV. DIODE-CLAMPED BASED MULTILEVEL MATRIX CONVERTERS

Considering the IMC topologies mentioned in section II, other implementations of multilevel matrix converters are possible. For this purpose, the inverter and the rectifier stages can be replaced with the multilevel converter circuit as shown in Fig. 2(b). The neutral point of the two converters are connected together and instead of two buses three current passes exist between rectifier and inverter stages. On the other hand, to avoid spikes and high current surges, the input voltage sources should not be short circuited and the output current should always find a path to flow.

But satisfying these rules for the three buses makes the modulation very complicated and it can be shown that under some circumstances for higher number of levels it is not possible to comply with the rules. However, another configuration is possible which does not have the aforementioned problems. If the input current rectifier stage remains as a two level converter, the result would be a simpler topology as shown in Fig. 3(a). In Fig. 3(b), the rectifier stage is switched at the line frequency, and because low frequency, high power and high voltage switches are available in the market with reasonable prices, there is no need to substitute this stage with multilevel converter. It is worth mentioning that the neutral point of the multilevel stage is connected to ground to keep it at zero voltage and the two positive and negative buses are supplied by the two level rectifier stage. As demonstrated in Fig. 3(b),4, the rectifier stage is synchronized with the source voltage. Therefore, the fundamental harmonics of the input currents are in phase with the corresponding source voltages. Utilizing basic SVM techniques for multiplexed inverters, the output stage can be controlled to generate controllable AC voltage wave forms with a desired frequency. To prevent a short circuit at the input stage and an open circuit at the output stage certain precaution must be

taken. To avoid an open circuit, the rectifier stage is switched at the zero vector period of the multilevel inverter stage. The short circuit problem occurs if more than two switches among upper or lower switches are turned on. However, this is prevented by using the simple switching pattern illustrated in Fig. 4, because as soon as an upper switch is turned off, another upper switch will be turned on.

V. SIMULATION RESULTS

The simulation results in Fig. 3 confirms the operation of the proposed multilevel matrix converter. Using the circuit in Fig. 3(a), computer simulation are carried out for two different sets of R-L loads. The circuit parameters used in simulation are shown in Table I.

TABLE I
SYSTEM PARAMETERS FOR SIMULATION

$Parameter$	$Values$
L_s	30μ
R_s	0.1Ω
C_f	10μ
V_{input}	$220V_{l-l}(rms)$
f_{input}	60Hz
f_{out}	100Hz
f_{sw}	10KHz

It is shown in Fig. 3 that the displacement power factor for the input stage is almost unity for the two different operating points. The input current total harmonic distortion is much better than that of the diode rectifier with dc-link capacitor, but worse than that of the PWM boost converter. However, this level of distortion is usually acceptable for most of the applications. The efficiency of the proposed topology is higher, because of the less reactive components and the lower switching frequency at the rectifier stage.

VI. CONCLUSION

This paper presents, a new AC/AC power converter that is suitable for high power applications. The proposed multi-level matrix converter topology is based on diode clamped multilevel converter at the output stage and input rectifier using unidirectional switches. Therefore, compared to other existing MMCs, it utilizes the least number of switches and minimizes switching losses. Because of the multilevel output waveforms, the output power quality is significantly improved. This topology does not require bulky DC-link capacitors, heavy input inductors or highly controlled flying capacitors and only uses small input filter capacitors. Principle of operation and characteristics of the converter were investigated and its feasibility is illustrated by simulation results for a 20KVA load. It is shown that for different loads, the output voltage is controllable at the desired frequency and the unity displacement power factor is achieved at the input stage. In conclusion it is worth mentioning that the proposed MMC topology is based on three level diode-clamped converter and further research is required for higher number of levels.

REFERENCES

[1] M. Venturini, "A new sine wave in, sine wave out, conversion technique eliminates reactive elements," *Proc. POWERCON 7*, pp. 1–15, 1980.

[2] J. Rodriguez, "A new control technique for ac-ac converters," *Proc. IFAC Control in power Electronics and Electrical Drives Conf.*, pp. 203–208, 1983.

[3] Y. Minari, K. Shinohara, and R. Ueda, "Pwm-rectifier/voltage-source inverter without dc link componentsfor induction motor drive," *IEEProceedings B [see also IEE Proceedings Electric Power Applications] Electric Power Applications*, vol. 140, no. 6, pp. 363–368, 1993.

[4] J. W. Kolar, M. Baumann, F. Schafmeister, and H. Ertl, "Novel three- phase ac-dc-ac sparse matrix converter," *Applied Power Electronics Conference and Exposition, APEC, IEEE*, vol. 2, pp. 777–791, 2002.

[5] J. Holtz and U. Boelkens, "Direct frequency convertor with sinusoidal line currents forspeed-variable ac motors," *IEEE Transactions on Industrial Electronics*, vol. 36, no. 4, pp. 475–479, Nov. 1989.

[6] C. Klumpner and F. Blaabjerg, "A new generalized two-stage direct power conversion topology to independently supply multiple ac loads from multiple power grids with adjustable power loading," *Power Electronics Specialists Conference, PESC, IEEE*, vol. 4, pp. 2855–2861, 2004.

[7] C. Klumpner, T. Wijekoon, and P. Wheeler, "A new class of hybrid ac/ac direct power converters," *Industry Applications Conference, IAS, IEEE*, vol. 4, pp. 2374–2381, 2005.

[8] J. Rodriguez, J. Lai, and F. Peng, "Multilevel inverters: A survey of topologies, controls, and applications," *IEEE Transactions on Industrial Electronics*, vol. 49, pp. 724–738, 2002.

[9] A. Nabae, I. Takahashi, and H. Akagi, "A new neutral-point clamped PWM inverter," *IEEE*

Trans. Ind. App., vol. 17, pp. 518–523, 1981.

[10] T. Meynard and H.Foch, "Multi-level conversion: High voltage choppers and voltage-source inverters," *IEEE, PESC Conf*, pp. 397–403, 1992.

[11] R. W. Erickson and O. A. Al-Naseem, "A new family of matrix converters," *Industrial Electronics Society, IECON, IEEE*, vol. 2, pp. 1515–1520, 2001.

[12] Y. Shi, X. Yang, Q. He, and Z. Wang, "Research on a novel capacitor clamped multilevel matrix converter," *IEEE Transactions on Power Electronics*, vol. 20, no. 5, pp. 1055–1065, 2005.

[13] K. Iimori, K. Shinohara, O. Tarumi, Z. Fu, and M. Muroya, "New current-controlled pwm rectifier-voltage source inverterwithout dc link components," *Proceedings of the Power Conversion Conference - Nagaoka*, vol. 2, pp. 783–786, 1997.

[14] L. Huber and D. Borojevic, "Space vector modulated three-phase to three-phase matrix converterwith input power factor correction," *IEEE Transactions on Industry Applications*, vol. 31, no. 6, pp. 1234–1246, 1995.

[15] L. Huber, D. Borojevic, and N. Burany, "Voltage space vector based pwm control of forced commutated cycloconverters," *Industrial Electronics Society, IECON, IEEE*, pp. 106–111, 1989.

[16] S. Kim, S.-K. Sul, and T. A. Lipo, "Ac/ac power conversion based on matrix converter topology with unidirectional switches," *IEEE Transactionson Industry Applications*, vol. 36, no. 1, pp. 139–145, 2000.

◈◈◈

Performance Evalution of CNTFET-Based SRAM Cell Design

Sneh Lata Murotiya[1], Aravind Matta[2] & Anu Gupta[3]

Department of Electrical and Electronics, Birla Institute of Technology and Science, Pilani,
Pilani, Rajasthan, India, 333031
E-mail : [1]snehlata@bits-pilani.ac.in, [2]aravindmatta@gmail.com, [3]anug@bits-pilani.ac.in

Abstract - Carbon Nanotube Field-Effect Transistor (CNTFET) technology with their excellent current capabilities, ballistic transport operation and superior thermal conductivities has proved to be a very promising and superior alternative to the conventional CMOS technology. A detailed analysis and simulation based assessment of circuit performance of this technology is presented here. As figures of merit speed, power consumption and stability are considered to evaluate the performance parameters of CNTFET-Based SRAM Cells with different chiral vectors for the optimum performance. A novel performance metric, presented as "SPR," is used to assess these figures of merit. This comprehensive metric includes a metric of low power delay product (PDP) for write operation and high stability in the operation of a memory cell. It is shown that an 8T SRAM cell provides 73% higher SPR than Dual-Chiral based 6T SRAM cell for CNT technology and 124% higher SPR than its CMOS counterpart, thus attaining superior performance. The CNTFET-based 8T SRAM cell demonstrates that it provides high stability, low delay and low power, which is better than CNTFET-based 6T SRAM cell as well as CMOS SRAM cell.

Keywords-*CNTFET; CNT;SNM; SINM; SPR; SRAM;PDP..*

I. INTRODUCTION

Scaling down the physical gate length (feature size) of current CMOS technology deeper in nanoscales results in various critical challenges and reliability issues, which will reduce its potential for energy-efficient applications in near future. To handle these difficulties in terms of physical phenomena and technological limitations, such as increased short-channel effects, reduced gate control, exponentially rising leakage current, larger process variations and high power dissipation, scientists and researchers are working towards new alternative technologies to replace conventional CMOS technology [1]. Nanoscaled alternatives, such as ultrathin body devices FinFETs [2], Carbon nanotube field-effect transistor (CNTFET) [3] could be the possible alternatives to bulk silicon transistors. CNTFETs are very promising and superior alternative to the conventional CMOS, due to their unique 1-D band structure that provides ballistic transport operation by suppressing backscattering, superior thermal conductivities and excellent current capabilities [4-8].

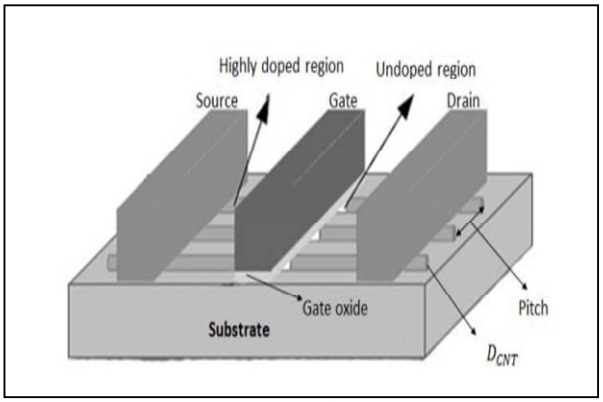
In a billion-transistor superscalar microprocessor, approximately 70% of the transistors are estimated to be used for memory arrays such as larger L2 and L3 SRAM data caches. Therefore, it is highest priority to design a fast and power-efficient memory structures to increase the performance of overall system. The unique properties of CNTFETs such as ballistic transport operation and low current under OFF condition makes them very attractive for the high performance and increased integration complexity of SRAM arrays design.

In this paper, the performance parameters of 8T Static RAM (SRAM) cell as well as 6T SRAM cell based CNTFETs are evaluated and compared in various simulation conditions and test benches by using novel efficient performance matrices. For 6T SRAM cell similar qualitative behavior has been observed as found in Ref. [16]. Hereafter, a common simulation environment has been adopted to perform a comparative analysis between 6T SRAM cell and 8T SRAM cell. At a circuit level design of CNTFET, the analysis and simulation for selection of optimum diameters for optimized threshold voltages of the two types (i.e. N or P) of CNTFETs are also performed to achieve the best overall performance in terms of power consumption, write time and stability of the CNT-based SRAM cell. Experimental results demonstrate that 8T SRAM cell outperforms 6T SRAM cell in terms of high stability, low delay and low power.

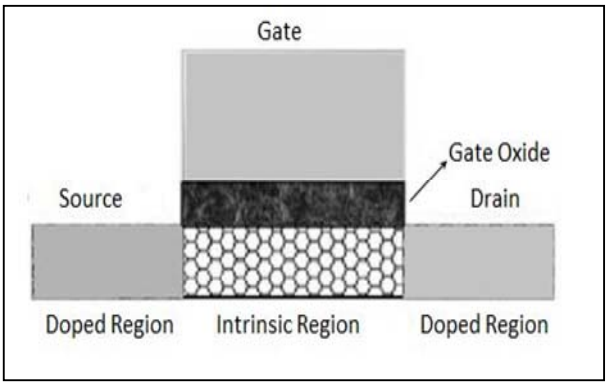
In Section II of this paper, a brief review of the characteristics and physical features of CNTFET devices is presented. The designs of 6T and 8T CNTFET SRAM cell are explained in Section III. Section IV includes the simulation result and finally, Section V concludes this paper.

II. CARBON NANO TUBE FIELD EFFECT TRANSISTOR (CNTFET)

A carbon nanotube (CNT) is a graphene sheet (with carbon atoms appearing in a hexagonal pattern) rolled up to form a hollow cylinder. CNTs have extremely low electrical resistance because electrons can travel for large distances without scattering (ballistic transport). This is partly due to their very small diameter and huge ratio of length to diameter. Also, because of their low resistance, CNTs dissipate very little energy. This will prove useful in solving the power consumption problems that are plaguing Silicon circuits. In the generic CNTFET which could be single wall (SWCNTs) or multi wall (MWCNTs), a carbon nanotube is placed between two electrodes while a separate gate electrode controls the flow of current in the channel, as presented in Figure1(a) [9-11].



(a)



(b)

Figure 1 (a).Carbon Nano Tube Field Effect Transistor, (b). Cross-section of MOSFET like CNTFET

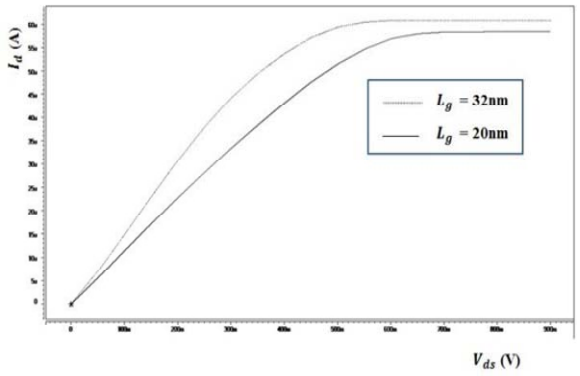


Figure 2. Current-Voltage characteristics for MOSFET like CNTFET

A typical CNTFET and the cross section of MOSFET like CNTFET is shown in Figure 1(a) and (b), respectively [9-11]. In SWCNTs, the arrangement of carbon atoms along the tube is determined by its chiral vector which is specified by (n, m) indices [9-11]. Based on a chiral vector, a SWCNT could be conducting or semiconducting. The current-voltage (I-V) characteristics of the CNTFET with different channel lengths, is shown in Figure 2, and they are similar to those of MOSFET which makes CNTFET a good candidate for ultra-low power applications. CNTFETs have the capability of setting the required threshold voltages by choosing proper diameter for the nanotubes [12], which makes them very suitable for implementing high-performance multiple-V_{th} structures. The threshold voltage of CNTFET can be approximated to the first order as the half bandgap and can be calculated by the following equation [9-11]:

$$V_{th} \approx \frac{E_g}{2e} = \frac{1}{\sqrt{3}} \frac{aV_\pi}{eD_{CNT}} \approx \frac{0.43}{D_{CNT}(nm)}$$

Where, $V\pi$ (≈ 3.033 eV) is the carbon π–π bond energy in the tight bonding model, e is the unit electron charge, parameter a (= 0.249 nm) is the carbon to carbon atom distance, and D_{CNT} is the diameter of CNTs, which is a function of chiral vector (n,m) and can be calculated by the following equation [9-11]:

$$D_{CNT} = \frac{a \times \sqrt{n^2+m^2+nm}}{\pi} \approx 0.0783 \times \sqrt{n^2+m^2+nm}$$

For example, the threshold voltage of the CNTFET having (19, 0) CNTs is 0.289 V because the D_{CNT} of (19, 0) CNT is 1.49 nm. Simulation results have corrected this threshold voltage. Figure 3 shows the threshold voltage of NCNTFET with different chiral vectors. Extensive research has been reported on manufacturing well-controlled CNTs [13-14]. In this paper, CNTFETs with

different diameters are utilized and channel length of 20nm is selected for area efficient SRAM design.

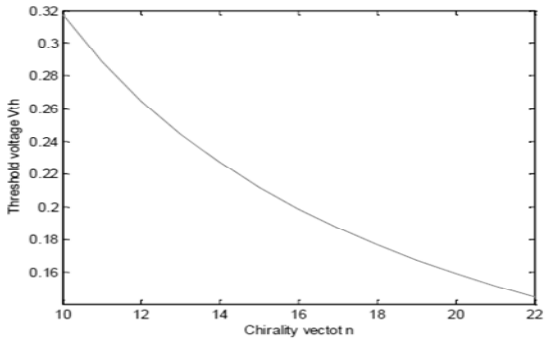


Figure 3. Threshold voltage of NCNTFET with different chiral vectors

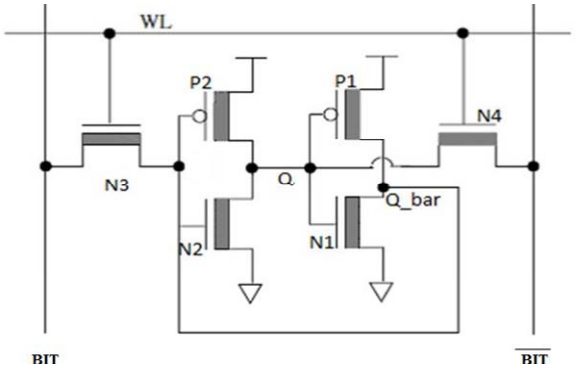


Figure 4. CNTFETs based 6T- SRAM

III. 6T & 8T SRAM CELL DESIGN

A. 6T SRAM Cell Design

The conventional [six-transistor (6T)] SRAM cell structure based on CNTFETs which is the core storage element of most register file and cache designs, is shown in Figure 4. It has four transistors P1, P2, N1 and N2 form two cross-coupled inverters for storage and two pass transistors N3 and N4 form as a combination of read/write port. With the aggressive scaling in CMOS technology, at ultra-low power supply ,the use of 6T SRAM cell leads to numerous critical problems like poor stability, high power consumption etc. In this case CNTFETs could be a good alternative with high stability and high density for high density memories.

The BIT and $\overline{\text{BIT}}$ lines in Figure 4 are pre-charged before any read operation. During read operation, the WL bit is "high" which makes both the transistor N3 and N4 to be turned ON and the stored data in SRAM cell is read. But this stored data may change due to a read-upset problem which is as follows. Suppose that the SRAM cell is currently storing a "1" so that Q is "1" and Q_bar is "0". When WL bit is high and transistor N3 and N4 are ON, then voltage level at node Q_bar will rise. In this case, an appropriate device sizing ratio between N1 and N3 is desired to limit the voltage at node Q_bar to be less than threshold voltage (V_{th}) so that the stored data will not differ during the read operation. For the appropriate selection of sizing ratio between N1 and N3, simulation has been done, which is shown in Figure 5, at the gate length of 20nm. As mentioned earlier, the threshold voltage of the (19, 0) CNTFET is 0.289V, so the sizing ratio of the N1 and N3 should not be less than 0.5. However, in this designing, sizing ratio between N1 and N3 is chosen as 1.5 for fair comparison with 8T SRAM cell as well as CMOS SRAM cell which has threshold voltage of 0.18V at gate length of 32 nm [15].

For the reliable write operation, the pull-up transistor of SRAM cell should not be very conducting. Suppose that the WL bit is high, SRAM cell is currently holding "1" and system is going to write "0" into SRAM cell. In this case the voltage level at node Q2 in Figure 4 will decrease only when the pass transistor N4 is stronger than the pull up transistor P2. Therefore, for the proper selection of sizing ratio between P2 and N4, simulations have been performed and the results are shown in Figure 6, at the gate length of 20nm. As described before, the threshold voltage of the CNTFET (19, 0) is 0.289V, so the sizing ratio between P2 and N4 should not be greater than 1.6. In this design, device sizing ratio between P2 and N4 is taken as 0.5 for proper comparison with CMOS which has threshold voltage of 0.18V at gate length of 32 nm.

The transistor width of CNTFET is defined in terms of the number of the tubes and the distance from one tube to adjacent tube is 20nm in a device [9-11]. The channel length chosen in this paper is also 20 nm. Therefore, in the presented CNTFET-based 6T SRAM cell, two NCNTFETs having device dimension of 80/20 nm, two PCNTFETs having device dimension of 40/20 nm and two NCNTFETs having device dimension of 60/20 nm are utilized. For a CMOS based.

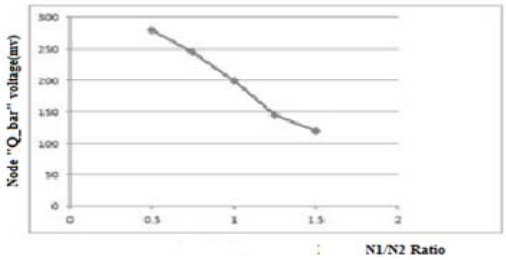


Figure 5.node "Q_bar" voltage v/s N1/N2 ratio

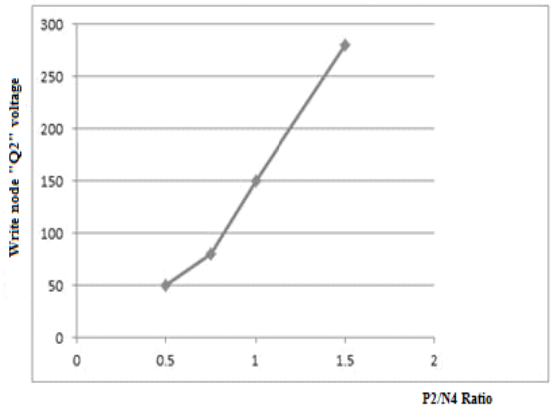


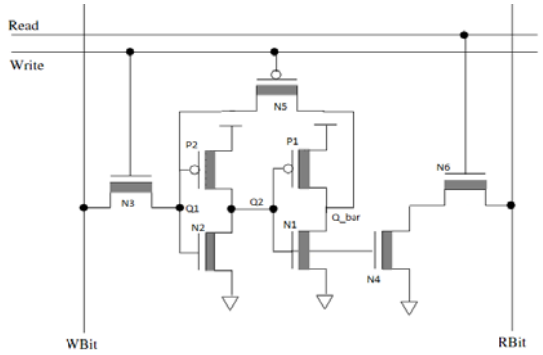
Figure 6. Write node "Q2" voltage v/s P2/N4 ratio

SRAM cell with a channel length of 32nm and similar circuit performance to the proposed CNTFET SRAM cell, the dimensions (width to length ratio) of the pull-up transistors, the pull-down transistors and the pass gate transistors are found to be 2/1, 4/1, and 3/1, respectively.

B. 8T SRAM Cell Design

The need of 8T-SRAM has originated from the fact that the 6T SRAM has more power consumption and less immunity to noise voltage during read operation as a small noise voltage is enough to flip the data. Designing an efficient cache system with 8 transistors in basic SRAM cell provides increased stability and good effective memory speed. Figure 7 shows the 8T SRAM cell configuration based on CNTFETs, where the write and read bits are separated to improve read cycle. In 8T SRAM only the Write bit is used to write for both "0" and "1" data, while BIT and $\overline{\text{BIT}}$ lines are utilized for writing data in the conventional 6T SRAM.

The writing operation starts by breaking the feedback loop of the cross-coupled inverter. During read operation, the feedback loop is maintained. The feedback loop is disconnected by setting Write bit to"1".In this case, SRAM memory cell has just two cascaded inverters. The Wbit line voltage decides the data that is going to be written into SRAM cell. The Wbit line transfers the inversion of the input data to Q2 (cell data), which drives the other cascaded inverter to get Q_bar. The Wbit line has to be pre-charged before and after each write operation. When writing "0" data, there is no. discharging at WBit line so negligible power is consumed. But writing '1' data at Q2, the dynamic power consumption is same as 6T SRAM cell because the WBit line has to be discharged to ground level. The proposed 8T SRAM cell is more power efficient in comparison with conventional ones because during write operation, the circuit does not require discharging for every write operation but discharges only when writing "1" data, and the discharging activity factor of the WBit line is less than 1.



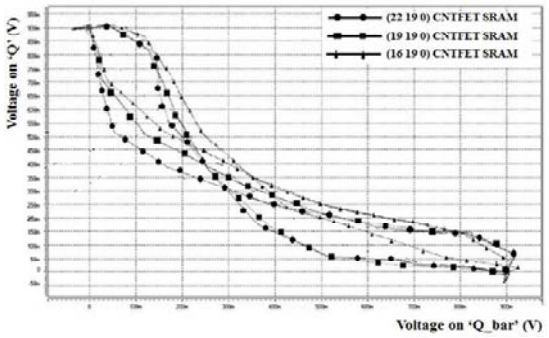
The Rbit line has to be pre-charged before any read operation. During read operation, Read bit is "high" and Write bit is low, which set turning ON condition for N5 & N6. When Q2="0", the transistor N4 is OFF maintaining the Rbit line at the pre-charged value, which shows the cell data Q2 holds "0". On the other side, when cell data Q2= "1", both the transistors N4 and N6 are ON which makes the Rbit line to be dropped at few milivolts, which is quite enough for detection in the sense amplifier. In the proposed 8T SRAM cell design, the device sizing of the read zero path with pull down transistor N4 and N6 is made 8 and 6 times larger to get a quicker discharge path to ground. In this design, two PCNTFET (P1,N5) with one tube, two NCNTFETs (N2,N3) with two tubes, one P-type CNTFET (P2) with two tube, one NCNTFET (N1) with one tube, are utilized for proper functionality and shorter delay at the minimal cost of the chip area overhead. Compared to CMOS SRAM cell with a transistor length of 32nm, a CNTFET SRAM cell offers a significant saving in chip area.

IV. SIMULATION RESULTS

In this paper, NCNTFETs and PCNTFETs with different chiral vectors are utilized for the optimized circuit performance. Therefore, a triplet index (*np*, *nn*, *m*) as shown in Ref. [16], where *np* and *nn* indicate the first chiral vector "*n*" of the PCNTFETs and NCNTFETs, respectively, and *m* is the common second chiral vector "*m*" of the two types of CNTFETs, is used. It has been shown in Ref. [16] that a comprehensive metric, SPR must be used to assess the circuit performance in terms of power dissipation, stability with respect to noise and delay. The SPR (in sec^{-1}) is defined as the ratio of product of static noise margin (SNM) and static current noise margin (SINM): (SNM X SINM) to the product of write power and write delay (write power X write delay). The SNM of SRAM cell is demonstrated as the maximum noise magnitude that does not disturb the stored bit of the SRAM cell [17]). The SINM is defined as the maximum DC noise current that can be injected in the SRAM cell

without changing its content [18].Therefore, the combined SNM and SINM is used in SPR, to define the static stability criteria for the SRAM cell [18]. To measure the SNM of the SRAM cell, simulations have been performed on CNTFET SRAM cells with index triplets of (16, 19, 0), (19, 19, 0), and (22, 19, 0) for the CNTFETs. Simulation results for SNM of the 6T SRAM cell as well as 8T SRAM cell at 0.9V power supply and room temperature are shown in Figure8 (a) and 8(b), respectively. As shown in Figure 8, the SNM of SRAM cell is increased as the chiral vector of P-type CNTFET changes from (19, 0) to (22,0). Simulation has shown that the SNM of the CMOS SRAM at 32nm channel length [15], is smaller than CNTFET SRAM cell. The PDP can be used as an important metric to compare the performance of circuits. The power dissipation of SRAM cell is high during write operation in comparison with read operation. Therefore, in this paper, write power and write delay is used to estimate PDP.

In this paper circuit simulation uses the Stanford CNTFET model [9-11] the 32nm BSIM PTM (predictive technology model) [15] for CMOS to evaluate the performance of CNTFET and CMOS SRAM cells. Table I & II show the SNM, SINM, Write Power, Write Delay



(a)

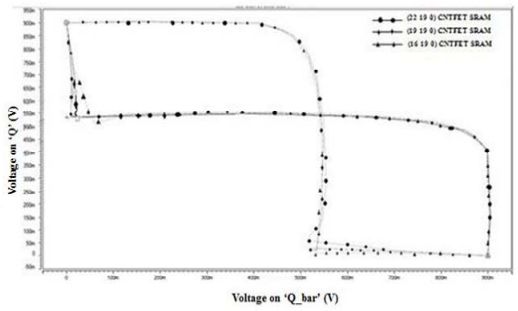


Figure.8 SNM of the (a) 6T SRAM cell (b) 8T SRAM cell

and SPR of the CNTFET SRAM cell and CMOS SRAM cell for 6T and 8T respectively, with different chiral vectors, at 0.9V power supply and room temperature. Simulation results confirm that (19,19,0) 8T SRAM cell occupies higher SPR in comparison with all specified CNTFET SRAM cells as well as CMOS SRAM cell. They also show that 8T SRAM cell provides 73% higher SPR than 6T SRAM cell in CNT technology and 124% higher SPR than 8T SRAM cell in CMOS technology.

TABLE I. ALL PERFORMANCE PARAMETERS OF THE 6T-SRAM CELL (CHANNEL LENGTH = 20 NM)

SRAM Cell (6T-SRAM)	SNM (mV)	Write time (ps)	Write Power (µW)	SINM (µA)	SPR (1/s)	SNM/ Write Delay (mV/ps)
(16,19,0)	176.00	26	24.4	70	1.94e+10	6.76
(19,19,0)	204.00	31	26.34	80	1.99e+10	6.58
(22,19,0)	225.00	35	30.10	85	1.81e+10	6.42
32nm CMOS SRAM cell	120.25	25	50.00	50	4.8e+9	4.82

TABLE II. ALL PERFORMANCE PARAMETERS OF THE 8T-SRAM CELL (CHANNEL LENGTH = 20NM)

SRAM Cell (8T-SRAM)	SNM (mV)	Write time (ps)	Write Power (µW)	SINM (µA)	SPR (1/s)	SNM/ Write Delay (mV/ps)
(16,19,0)	239.00	11	6.59	9.10	3e+10	21.70
(19,19,0)	392.00	13	7.10	7.91	3.35e+10	30.15
(22,19,0)	410.00	15	7.74	7.00	2.47e+10	27.33
32nm CMOS SRAM cell	268.56	18	20.40	28.00	2.05e+10	14.92

V. CONCLUSION

This paper has contemplated the use of CNTFETs in 8T SRAM and 6T SRAM cell designs. The extensive simulation and analysis have been done for selection of best chiral vector for the transistors used in 8T SRAM and 6T SRAM cell to attain high stability, low PDP during write operation. The performance parameters of an SRAM cell have been investigated based on a novel comprehensive metric which is known as SPR. This comprehensive metric provides versatile performance measurement of stability, access time, and power consumption in the operation of a memory cell.

It has been shown that 8T CNTFET SRAM cell has 124% higher SPR than its CMOS counterpart. In comparison with dual-chiral based 6T SRAM cell in Ref.[16], our simulation results have confirmed the superiority of the 8T SRAM cell in terms of 122% higher SNM, 50% less access time and 73% higher SPR. Thus, 8T CNTFET SRAM cell achieves superior performance in

terms of high stability, low delay and low power in CNT as well as in CMOS Technology.

REFERENCES

1) Roy, K., Mukhopadhyay, S., Meimand-Mehmoodi, H."Leakage current mechanisms and leakage reduction techniques in deep-submicron CMOS circuits," *Proc. IEEE*, vol.91,no. 2, pp. 305–327 Feb.2003.

2) 2. S. A. Tawfik, Z. Liu, and V. Kursun, "Independent-gate and tied-gate FinFET SRAM circuits: Design guidelines for reduced area and enhanced stability," in *Proc. Int. Conf. Microelectron. (ICM)*, Dec. 2007, pp. 171–174.

3) I. O'Connor, J. Liu, F. Gaffiot, F. Pregaldiny, C. Lallement, C. Maneux, J. Goguet, S. Fregonese, T. Zimmer, L. Anghel, "CNTFET Modeling and Reconfigurable Logic-Circuit Design," *IEEE Trans. Circuits Syst I Regul.* vol. 54, no.11,pp. 2365–2379, Nov. 2007.

4) A. Rahman, J. Guo, S. Datta, and M. S. Lundstrom, "Theory of ballistic nanotransistors," *IEEE Trans. Electron Device,* vol. 50, no. 10, pp.1853–1864, Sep. 2003.

5) Y. Lin, J. Appenzeller, J. Knoch, and P. Avouris, "High-performance carbon nanotube field-effect transistor with tunable polarities," *IEEE Trans. Nanotechnol.*, vol. 4, no. 5, pp. 481–489, Sep. 2005.

6) A. Akturk, G. Pennington. N. Goldsman, and A. Wickenden, "Electron transport and velocity oscillations in a carbon nanotube," *IEEE Trans. Nanotechnol.*, vol. 6, no. 4, pp. 469–474, Jul. 2007.

7) J. Appenzeller, "Carbon nanotubes for high-performance electronics—Progress and prospect," *Proc. IEEE*, vol. 96, no. 2, pp. 201–211, Feb. 2008.

8) H. Hashempour and F. Lombardi, "Device model for ballistic CNTFETs using the first conducting band," IEEE Des. Test. Comput., vol. 25, no. 2, pp. 178–186, Mar./Apr. 2008.

9) Stanford University CNTFETModelWeb site. (2008). [Online]. Available:http:// nano.stanford.edu/ model.php?id=23

10) J. Deng, H.-S.P. Wong, "A Compact SPICE Model for Carbon-Nanotube Field-Effect Transistors Including Nonidealities and Its Application—Part I: Model of the Intrinsic Channel Region," *IEEETrans. Electron Devices*, vol. 54, no. 12, pp. 3186 – 3194, Dec. 2007.

11) Jie Deng, H.-S.P. Wong, "A Compact SPICE Model for Carbon-Nanotube Field-Effect Transistors Including Nonidealities and Its Application—Part II: Full Device Model and Circuit Performance Benchmarking," *IEEE Trans. Electron Devices*, vol. 54, no.12, pp. 3195 – 3205, Dec. 2007.

12) A. Raychowdhury and K. Roy, "Carbon-nanotube-based voltage-mode multiple-valued logic design," *IEEE Trans. Nanotechnol.*, vol. 4, no. 2, pp. 168–179, Mar. 2005.

13) Y. Ohno, S. Kishimoto, T. Mizutani, T. Okazaki, and H. Shinohara, "Chiral assignment of individual single-walled carbon nanotubes in carbon nanotube field-effect transistors bymicro-photocurrent spectroscopy," *Appl. Phys. Lett.*, vol. 84, no. 8, pp. 1368–1370, Feb. 2004.

14) Y. Li, W. Kim, Y Zhang, M Rolandi, and D. Wang, "Growth of singlewalled carbon nanotubes from discrete catalytic nanoparticles of various sizes," *J. Phys. Chem.*, vol. 105, no.46, pp. 11424–11431, Oct.2001.

15) Berkeley Predictive Technology Model website [Online]. Available: http://www.eas.asu.edu/~ptm/.

16) Sheng Lin, Yong-Bin Kim,and Fabrizio Lombardi, "Design of a CNTFET-Based SRAM Cell by Dual-Chiral Selection*"; IEEE Transaction on Nanotechnology,* vol. 9, no. 1, Jan. 2010.

17) E. Seevinck, F.J. List, J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *IEEE J. Solid-State Circuits*, vol. 22, no. 5, pp. 748- 754, Oct. 1987.

18) E. Grossar, M. Stucchi, K. Maex, and W. Dehaene, "Read stability and write-ability analysis of SRAM cells for nanometer technologies," *IEEE J. Solid-State Circuits*, vol. 41, no. 11, pp. 2577–2588, Nov. 2006.

◈◈◈

Seclusion Location Based Queries In Mobile Environments

C.Vijaya Lakshmi & R. Raja Sekhar

Department of CSE, JNTUA COLLEGE OF ENGINEERING, ANANTAPUR, ANDHRAPRADESH, INDIA
E-mail : vijayaa.cs@gmail.com & drasharaj@gmail.com

Abstract - In Location based services users with location-aware mobile devices are able to make queries about their surroundings anywhere and at any time. Location cloaking is one typical approach to protecting user location privacy in location based services. Upon receiving a location-based spatial query from the user, the system cloaks the user current location into a cloaking region based on the user privacy requirements. Existing system controls the generation of cloaking regions and designed two cloaking algorithms, namely MaxAccu_Cloak and MinComm_Cloak. MaxAccu_Cloak is designed to maximize the accuracy of query results and MinComm_Cloak attempts to reduce the network communication cost. Two query processing modes, namely bulk and progressive. This paper presents extends the mobility aware location cloaking technique to t-closeness privacy metric. t-closeness requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of a sensitive attribute in the overall table. t-closeness protects against attribute disclosure but not identity disclosure.

Keywords*: Location based sevices,location privacy,query processing,mobile computing,t-closeness.*

I. INTRODUCTION

Location Based Services are emerging as a major application of mobile geospatial technologies.In LBS users are able to make queriesabout their surroundings any where at any time. Spatail range queries and k-nearest-neighbor queries are two types of the most commonly used queries in LBS.

Location cloaking is one typical approach to protecting user location in privacy in LBS. Recieving a location based spatial query from the user the system cloaks the system cloaks the users current location into cloaking region based on the users privacy requirement. The location based spatial query is transformed into a region-based spatial query befor being sent to the LBS server.The LBS server evaluates the region-based query and returns a result superset,which contains the query results for all possible location points in the cloaking region.Finally,the system refines the result superset to generate the exact results for the query location.

Existing system two query processing algorithms bulk algorithm that generates the query results all at once at the end of query evaluation and a progressive algorithm that produces the results incremently during query evaluation.To achieve identity anonymity in LBS by spatiotemporal cloaking based on k-anonymity model,that is,the cloaked location is made indistinguishablefrom the location information of at least k-1 other users.To perform spatial cloaking used Quad-tree-like algorithm.

In this paper we present t-closeness privacy metric. we propose a novel privacy notation called t-closeness that formalizes the idea of global bakground knowledge by requiring that the distribution of a sensitivesttribute in any equivalence class is close to distribution of attribute in the overall table. This effectively limits the amount of individual-specific information an observer can learn. Further,in order to incorporate distance between values of sensitive attributes,we use the Earth Mover Distance metric to measure the distance between the two distributions.

This paper is organized as follows: Section-2 k-anonymity for location privacy,Section-3 provides bulk and progressive query processing modes ,Section-4 provides t-closeness A new privacy measure,Section-5 provides Conclusion and Results.

II. K-ANONYMITY FOR LOCATION PRIVACY

In the context of location privacy, the k-anonymity metric was initially adapted to measure microscopic location-privacy by Gruteser and Grunwald [1]. In this model, each query sent to the LBS (including the user's pseudonym, her position and the query time) is equivalent to one entry in a database, and the location-time information in the query serves as the quasi-identifier. In order to protect a user's location privacy

using k-anonymity, each of her queries must be indistinguishable from that of at least k-1 other users. To this end, the pseudonyms of these k users are removed from their queries, and the location-time pair in their queries is obfuscated to the same location-area and time-window, large enough to contain the users' actual locations. The k-anonymity scheme for location privacy has become very popular, mainly due to its simplicity. A large body of research has focused on increasing the efficiency of k- anonymity schemes and reducing their cost of query obfuscation [2,3,4,5], extending the obfuscation method to protect traces [2] (i.e., location privacy at the macroscopic level), or adapting the architecture presented in [1] to different scenarios [6,7]. All of these systems can be represented by the initial model introduced in [1].

There is a set of users who access a LBS through a trusted Central Anonymity Server (CAS). Users send their LBS queries (I,q,l) to the CAS, where i is the identity of the user, q is her query, l is her precise location (expressed as a point with coordinates (x,y) in a 2-dimensional space), and t is the time at which the query is generated. In order to protect users' privacy, the CAS removes the identity i of the users. Furthermore, it obfuscates the location l = (x, y) and the time t at which the queries were generated. For this, it constructs a cloaking region R = ([x1, x2], [y1, y2], [t1, t2]) such that there are at least k users in R whose location l = (x, y) at time t satisfies that $x1 \leq x \leq x2$, $y1 \leq y \leq y2$, and $t1 \leq t \leq t2$. We note that, decentralized approaches [15] can also be represented by our model, by considering that the cloaking region R is computed by a set of entities (e.g., users them-selves) in a distributed manner (i.e., they collectively play the role of the CAS). Moreover, our model can accommodate both the systems in which users have to continuously report their location to the CAS [3], in order to build optimal regions, and the systems that rely only on user-triggered discrete queries for that purpose [2]. The k-anonymity location obfuscation technique aims at achieving two properties: query anonymity and location privacy. Achieving query anonymity implies that it is not possible for the adversary to link the identity i of a user to her query q, based on the location information (cloaking region) associated with the query. Location privacy is achieved when it is not possible for the adversary to learn the location l of a user i at time t, using queries he receives from the users and his a priori knowledge.

We consider an adversary that controls the LBS and, in addition to the received queries, has access to some background information. For example, one of the threats considered in [1] is restricted space identification". In this threat scenario the adversary knows that a given location corresponds (exclusively) to a user address, meaning that a query coming from that precise location would be linked to the user who resides at that address. Another considered threat in [1] is that the adversary (in addition to controlling the LBS) may deploy antennas in the vicinity of the users and thus knows that a given user is in location l at time t. Limitation of *k*- anonymity in situations where the values of sensitive attributes are not diverse. We propose to overcome this limitation by t-closeness privacy metric.

III. BULK AND PROGRESSIVE QUERY PROCESSING MODES

Two query processing algorithms :

A bulk algorithm that generates the query results all at once at the end of query evaluation.The bulk query processing algorithm generates the kCRNN resultsat the end of query evaluation. The server cannot start transmitting the results to the client untill end of qery evaluation. Therefore ,the query response time increased when we use bulk algorithm.

An alternative for this is progressive query processing algorithm to parallelize the query evaluation and result transmission. A progressive algorithm that produces the reults incremently during query evaluation. When compare to bulk progressive algorithm query response time is reduced. To reduce this query respone time rapidly we propose t-closeness privacy metric it reduces query response time.

IV. T-CLOSENESS:A NEW PRIVACY MEASURE

Privacy measured by the information gain of an observer. Before seeing the released table, the observer has some prior belief about the sensitive attribute value of an individual. After seeing the released table, the observer has a posterior belief. Information gain can be represented as the difference between the posterior belief and the prior belief**.**

We separate the information gain into two parts: that about the whole population in the released data and that about specific individuals.

To motivate our approach, let us perform the following thought experiment: First an observer has some prior belief $B0$ about an individual's sensitive attribute. Then, in a hypothetical step, the observer is given a completely generalized version of the data table where all attributes in a quasi-identifier are removed (or, equivalently, generalized to the most general values). The observer's belief is influenced by **Q**, the distribution of the sensitive attribute value in the whole table, and changes to $B1$. Finally, the observer is given the released table. By knowing the quasi-identifier values of the individual, the observer is able to identify the

equivalence class that the individual's record and learn the distribution **P** of sensitive attribute values in this class. The observer's belief changes to *B*2.

We choose to limit the difference between *B*1 and *B*2. In other words, we assume that **Q**, the distribution of the sensitive attribute in the overall population in the table, is public information. We do not limit the observer's information gain about the populations as a whole, but limit the extent to which the observer can learn additional information about specific individuals.

To justify our assumption that **Q** should be treated as public information, we observe that with generalizations, the most one can do is to generalize all quasi-identifier attributes to the most general value. Thus as long as a version of the data is to be released, a distribution **Q** will be released. We also argue that if one wants to release the table at all, one intends to release the distribution **Q** and this distribution is what makes data in this table useful. In other words, one wants **Q** to be public information. A large change from *B*0 to *B*1 means that the data table contains a lot of new information, e.g., the new data table corrects some widely held belief that was wrong. In some sense, the larger the difference between *B*0 and *B*1 is, the more valuable the data is. Since the knowledge gain between *B*0 and *B*1 is about the whole population, we do not limit this gain. We limit the gain from *B*1 to *B*2 by limiting the distance between **P** and **Q**. Intuitively, if **P** = **Q**, then *B*1 and *B*2 should be the same. If **P** and **Q** are close, then *B*1 and *B*2 should be close as well, even if *B*0 may be very different from both *B*1 and *B*2.

Requiring that **P** and **Q** to be close would also limit the amount of useful information that is released, as it limits information about the correlation between quasi identifier attributes and sensitive attributes. However, this is precisely what one needs to limit. If an observer gets too clear a picture of this correlation, then attribute disclosure occurs. The *t* parameter in *t-closeness* enables one to trade off between utility and privacy. Now the problem is to measure the distance between two probabilistic distributions. There are a number of ways to define the distance between them. Given two distributions

P = ($p1$, $p2$,..., pm), **Q** = ($q1$, $q2$,..., qm), two well-known distance measures are as follows:

The *variational distance* is defined as:

$$D[\mathbf{P},\mathbf{Q}] = \sum_{i=1}^{m} 1/2|pi - qi|.$$

And the Kullback-Leibler (KL) distance [10] is defined as:

$$D[\mathbf{P},\mathbf{Q}] = \sum_{i=1}^{m} pi \log pi/qi = H(\mathbf{P}) - H(\mathbf{P},\mathbf{Q})$$

Where $H(\mathbf{P}) = \sum_{i=1}^{m} pi \log pi$ is the entropy of **P** and $H(\mathbf{P},\mathbf{Q}) = \sum_{i=1}^{m} pi \log qi$ is the cross-entropy of **P** and **Q**.

This t-closeness privacy metric reduce query response time when compare to bulk and progressive query processing algorithms.

V. CONCLUSION AND RESULTS

In this paper we discussed k-anonymity for location privacy and have some limitations to overcome this we proposed new privacy metric is t-closeness.Below Table 1 shows the results of query responsed time compared to bulk and progressive query proceesing modes.

Finally t-closeness reduces the query response time and it overcome limitations of k-anonimity model.

Query processing modes	**Response Time (Seconds)**			
Bulk	57	46	35	30
Progressive	36	27	20	18
t-closeness	22	18	15	10

Table1: The comparision table of response time of bulk, progressive and t-closeness.

REFERENCES

[1] M.Gruteser and D. Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In ACM MobiSys, 2003.

[2] B.Gedik and L.Liu. Protecting location privacy with personalized k-anonymity: Architecture and algorithms. IEEE Trans. on Mobile Computing, 2008.

[3] M. F. Mokbel, C.-Y. Chow and W. G. Aref. The new casper: query processing for location

services without compromising privacy. In VLDB, 2006.

[4] K.W. Tan, Y. Lin, and K. Mouratidis. Spatial cloaking revisited: Distinguishing information leakage from anonymity. In SSTD, 2009.

[5] T.Xu and Y.Cai. Feeling-based location privacy protection for location-based services. In CCS, 2009.

[6] K.Sampigethaya,L.Huang,M.Li,R.Poovendran, K.Matsuura,and K.Sezaki. Caravan: Providing location privacy for vanet. In ESCAR, 2005.

[7] G. Zhong and U. Hengartner. A distributed k-anonymity protocol for location privacy. PerCom, 2009.

[8] P. Samarati. Protecting respondent's privacy in microdata release. *IEEE T. Knowl. Data En.*, 13(6):1010–1027, 2001.

[9] B.-C. Chen,K. LeFevre, and R. Ramakrishnan, "Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge," *Proc. of the Int'l Conf. on Very Large Data Bases (VLDB),* pp. 770–781, 2007.

[10] S.L.Kullback andR.A.Leibler.On information and sufficiency. *Ann. Math. Stat.*, 22:79–86, 1951.

◈◈◈

Performance Comparison of ANN Models For Short Term Load Forecasting

Sony Susan Varghese,S.Ashok

Dept. of Electrical Engineering, NIT Calicut, India
E-mail: sosuva@gmail.coms,ashoks@nitc.ac.in

Abstract - This paper presents a case study on short term load forecasting for a typical students apartment at NIT Calicut. Here-in the performance of three different artificial neural network models are studied for intelligent load forecasting and the results are compared. The different neural network models under study are Feed forward Back propagation model, Non linear auto-regressive with exogenous input model, and Radial Basis function model.

Keywords *: Artificial Neural Networks, Feed forward with Back propagation , Load forecasting, Non Linear auto-regressive with exogenous input , Radial Basis Function.*

I. INTRODUCTION

Load forecasting has always been important for planning and operational decision conducted by utility companies. However, with the deregulation of the energy industries, load forecasting is even more important. With supply and demand fluctuating and the changes of weather conditions and energy prices increasing by a factor of ten or more during peak situations, load forecasting is vitally important for utilities[1] . Short-term load forecasting can help to estimate load flows and to make decisions that can prevent overloading. Timely implementations of such decisions lead to the improvement of network reliability and to the reduced occurrences of equipment failures and blackouts. Load forecasting is also important for contract evaluations and evaluations of various sophisticated financial products on energy pricing offered by the market.

Load forecasting is classified into short term load forecasting and long term load forecasting depending on the period of consideration and quantum of information required according to the application. In the deregulated economy, decisions on capital expenditures based on long-term forecasting and short term load fore are also more important than in a non-deregulated economy when rate increases could be justified by capital expenditure projects. Short term load forecasting gives more precision and accurate results depending on the time frame for which it is taken.

Most forecasting methods use statistical techniques or artificial intelligence algorithms such as regression, neural networks, fuzzy logic, and expert systems. Two of the methods, so-called end-use and econometric approach are broadly used for medium- and long-term forecasting. A variety of methods, which include the so-called similar day approach, various regression models, time series, neural networks, statistical learning algorithms, fuzzy logic, and expert systems, have been developed for short-term forecasting.[2]

The end-use modeling, econometric modeling, and their combinations are the most often used methods for medium- and long-term load forecasting. Descriptions of appliances used by customers, the sizes of the houses, the age of equipment, technology changes, customer behavior, and population dynamics are usually included in the statistical and simulation models based on the so-called end-use approach. In addition, economic factors such as per capita incomes, employment levels, and electricity prices are included in econometric models. These models are often used in combination with the end-use approach. Long-term forecasts include the forecasts on the population changes, economic development, industrial construction, and technology development. End-use models. The end-use approach directly estimates energy consumption by using extensive information on end use and end users, such as appliances, the customer use, their age, sizes of houses, and so on. Statistical information about customers along with dynamics of change is the basis for the forecast.

End-use models focus on the various uses of electricity in the residential, commercial, and industrial sector. These models are based on the principle that electricity demand is derived from customer's demand for light, cooling, heating, refrigeration, etc. Thus end-use models explain energy demand as a function of the number of appliances in the market. [3]

Ideally this approach is very accurate. However, it is sensitive to the amount and quality of end-use data. For example, in this method the distribution of equipment age is important for particular types of appliances. End-use forecast requires less historical data but more information about customers and their equipment.

Econometric models**.** The econometric approach combines economic theory and statistical techniques for forecasting electricity demand. The approach estimates the relationships between energy consumption (dependent variables) and factors influencing consumption. The relationships are estimated by the least-squares method or time series methods. One of the options in this framework is to aggregate the econometric approach, when consumption in different sectors (residential, commercial, industrial, etc.) is calculated as a function of weather, economic and other variables, and then estimates are assembled using recent historical data. Integration of the econometric approach into the end-use approach introduces behavioral components into the end-use equations.

Statistical model-based learning. The end-use and econometric methods require a large amount of information relevant to appliances, customers, economics, etc. Their application is complicated and requires human participation. In addition such information is often not available regarding particular customers and a utility keeps and supports a profile of an "average" customer or average customers for different type of customers. The problem arises if the utility wants to conduct next-year forecasts for sub-areas, which are often called load pockets. In this case, the amount of the work that should be performed increases proportionally with the number of load packets. In addition, end-use profiles and econometric data for different load pockets are typically different. The characteristics for particular areas may be different from the average characteristics for the utility and may not be available.

$$L(t) = F(d(t), h(t)) \cdot f(w(t)) + R(t) \quad (1),$$

where L(t) is the actual load at time t, d(t) is the day of the week, h(t) is the hour of the day, F(d, h) is the daily and hourly component, w(t) is the weather data that include the temperature and humidity, f(w) is the weather factor, and R(t) is a random error.In fact, w(t) is a vector that consists of the current and lagged weather variables. This reflects the fact that electric load depends not only on the current weather conditions but also on the weather during the previous hours and days. In particular, the well-known effect of the so-called heat. [3,4] waves is that the use of air conditioners increases when the hot weather continues for several days.

To estimate the weather factor f(w), we used the regression model

$$f(w)=\beta o+\sum \beta_j X_j \quad (2)$$

where Xj are explanatory variables which are nonlinear functions of current and past weather parameters and βo, βj are the regression coefficients.

As mentioned earlier artificial neural networks, time series modeling approach, Fuzzy system approach can also be used load forecasting. Of the above methods artificial neural networks and Fuzzy systems are implemented in case of non-linear data and intelligent load forecasting.[5,6]

The advantage of these models are that they are adept in predicting with a good accuracy (compared to the conventional prediction methods,) considering the non linearity and the impreciseness of the data supplied. This is attributed by their adaptability, self-organisation and non-linearity property.

LOAD FORECASTING USING ARTIFICIAL NEURAL NETWORK

A. Artificial Neural Network

An artificial neural network (ANN), usually called neural network (NN), is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Modern neural networks are non-linear statistical data modelling tools. They are usually used to model complex relationships between inputs and outputs or to find patterns in data. In this paper ANN models have been used for short term load prediction over a span of a week. The data is forecasted on an hourly basis and the efficiency of different models are compared by the no. of epochs taken, no. of layers used , error in load forecasting etc.[7,8]

B. NETWORK FUNCTIONS

The word *network* in the term 'artificial neural network' refers to the inter–connections between the neurons in the different layers of each system. An example system has three layers. The first layer has input neurons, which send data via synapses to the second layer of neurons, and then via more synapses to the third layer of output neurons. More complex systems will have more layers of neurons with some having increased layers of input neurons and output neurons. The synapses store parameters called "weights" that manipulate the data in the calculations.

An ANN is typically defined by three types of parameters:

1. The interconnection pattern between different layers of neurons
2. The learning process for updating the weights of the interconnections
3. The activation function that converts a neuron's weighted input to its output activation.

Mathematically, a neuron's network function $f(x)$ is defined as a composition of other functions $g_i(x)$, which can further be defined as a composition of other functions. This can be conveniently represented as a network structure, with arrows depicting the dependencies between variables. A widely used type of composition is the *nonlinear weighted sum*,

$$f(x) = K\left(\sum_i (w_i g_i(x))\right) \quad (3)$$

where K (commonly referred to as the activation function) is some predefined function, such as the hyperbolic tangent. It will be convenient for the following to refer to a collection of functions g_i as simply a vector $g=(g_1,g_2,\ldots g_n)$.

II. NEURAL NETWORK MODELS EMPLOYED

A. FEED FORWARD BACK PROPOGATION NETWORK

Back propagation is a very common method for training multilayered feed forward networks. Back propagation can be used with any feed forward network (FFBP network) that uses a activation function that is differentiable. It is this derivative function that we will use during training. It is not necessary that you understand calculus or how to take the derivative of an equation to work with the material in this chapter. If you are using one of the common activation functions, you can simply get the activation function derivative from a chart.

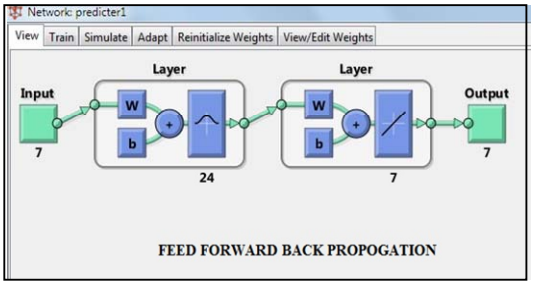


Figure 1 : FFBP network

To train the neural network, a method must be determined to calculate the error (depicted in figure 1). As the neural network is trained, the network is presented with samples from the training set. The result obtained from the neural network is then compared with the anticipated result that is part of the training set. The degree to which the output from the neural network differs from this anticipated output is the error.

To train the neural network, we must try to minimize this error. To minimize the error, the neuron connection weights and thresholds must be modified. We must define a function that will calculate the rate of error of the neural network. This error function must be mathematically differentiable. Because the network uses a differentiable activation function, the activations of the output neurons can be thought of as differentiable functions of the input, weights, and thresholds. If the error function is also a differentiable function, such as the sum of square error function, the error function itself is a differentiable function of these weights. This allows us to evaluate the derivative of the error using the weights. Then, using these derivatives, we find weights and thresholds that will minimize the error function.

B. NON LINEAR AUTOREGRESSIVE WITH EXOGENOUS INPUT MODEL

In time series modelling, a nonlinear autoregressive exogenous model (NARX) is a nonlinear autoregressive model which has exogenous inputs. This means that the model relates the current value of a time series which one would like to explain or predict to both:

- 1.past values of the same series; and
- 2.current and past values of the driving (exogenous) series — that is, of the externally determined series that influences the series of interest.

In addition, the model contains an "error" term which relates to the fact that knowledge of the other terms will not enable the current value of the time series to be predicted exactly.

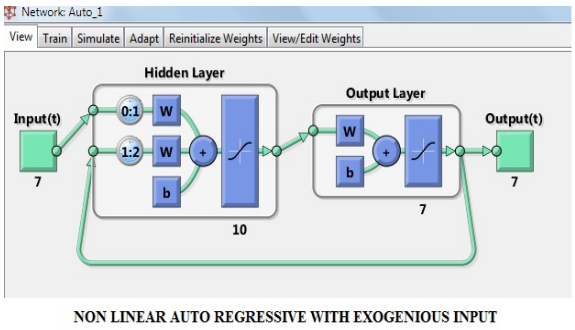


Figure 2: NARX model network

Such a model can be stated algebraically as

$$Y_t=F(Y_{t-1},Y_{t-2},..Y_{t-n},U_t,U_{t-1},U_{t-2},..)+E_t \qquad (4)$$

Here Y is the variable of interest, and U is the externally determined variable. In this scheme, information about U helps predict Y, as do previous values of y itself. Here E is the error term (sometimes called noise). For example, Y may be air temperature at noon, and U may be the day of the year (day-number within year). The above mentioned neural network simulated is illustrated in fig.2.

The function F is some nonlinear function, such as a polynomial. F can be a neural network, a wavelet network, a sigmoid network and so on.

C. *RADIAL BASIS FUNCTION*

A radial basis function (RBF) is a real-valued function whose value depends only on the distance from the origin, so that

$$\phi(x) = \| x \| \qquad (4)$$

or alternatively on the distance from some other point c, called a *center*, so that

$$\phi(x,c) = \| x - c \| \qquad (5).$$

Any function that satisfies the property $\phi(x) = \| x \|$ is a radial function.

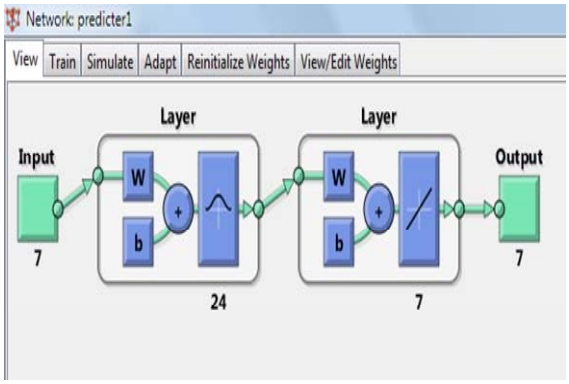


Figure 3: Radial Basis function

The norm is usually Euclidean distance, although other distance functions are also possible.

IV. COMPARITIVE STUDY

The neural networks were simulated using MATLAB NNTOOL. The following were the results were found after training and simulating these networks in MATLAB environment. In this paper the following parameters are under study and compared for each network.

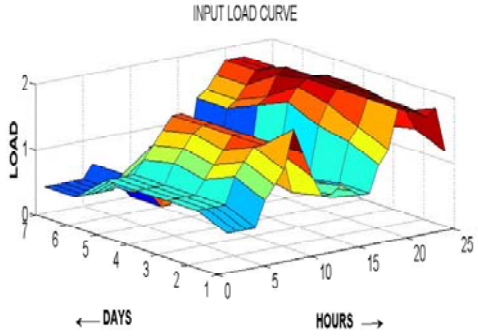


Figure 4: Input load curve (daily)

The input load curve for training the neural network is shown in figure 4.

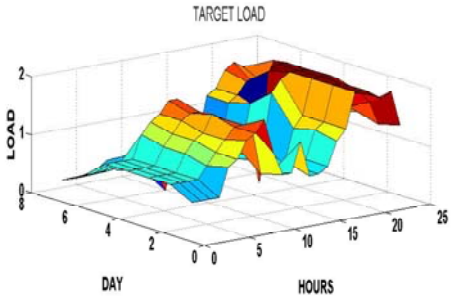


Figure 5:Target load curve(daily)

The target load curve for simulating and training the neural network is depicted in the fig.5

TABLE 1 NEURAL NETWORK PERFORMANCE

Neural Network Model	NEURAL NETWORK PARAMTERS			
	No. of layers	No. of iterations	Best validation epoch	MSE
FFBP	7	100	5	1e^-1
NARX	3	100	7	1.8
RBF	3	1	-	1e^-14

V. RESULTS AND DISCUSSION

The predicted output and error for each network , was found after conducting simulation studies using MATLAB platform. Generally load profiles are non-linear ,and can vary due to a no. of factors like temperature, climatic conditions , seasonal variations etc when considering a long term forecast model. For such models time series prediction model like non linear auto regressive model with exogenous input is the best. It is evident from the results that radial basis function gives the best predicted output with error less than 1e^-14, ideal for short term forecasting. The results for each of the network is shown below :

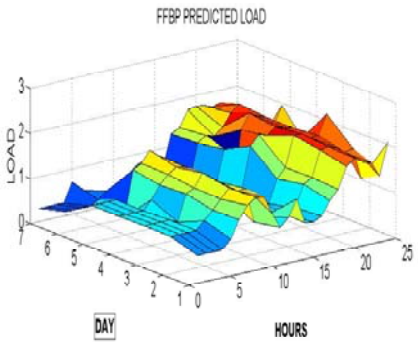


Figure 6: FFBP predicted load

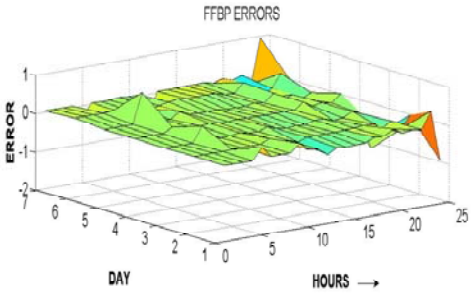


Figure 7: FFBP error

Feed Forward back propagation network predicted the output load with a maximum possible error of± 1unit as shown in fig.7.

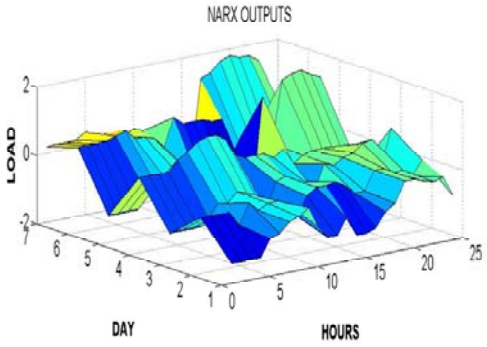


Figure 8: NARX predicted load

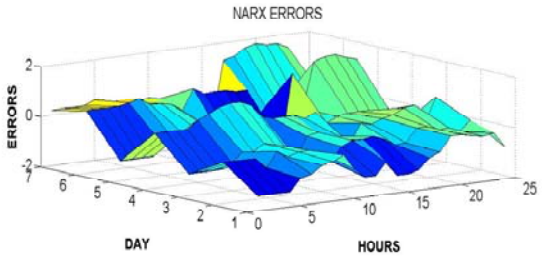


Figure 9: NARX Error

NARX model predicted the load with a relatively low accuracy, and the maximum possible error was found to be 2 units as illustrated in Fig. 9.

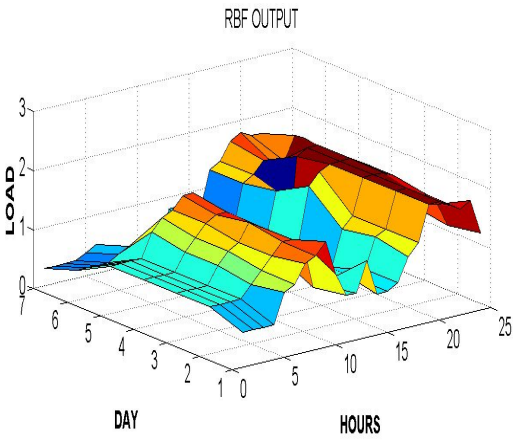


Figure 10: RBF predicted load

The RBF network predicted the load with relatively high accuracy, with minimum no. of epochs, and iterations.

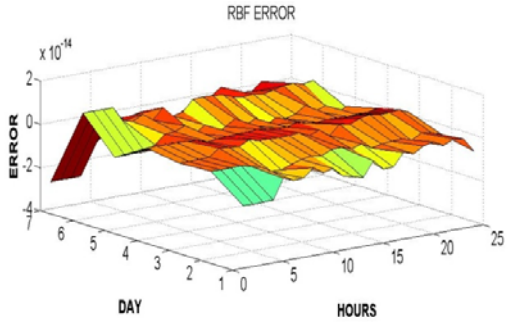


Figure 11: RBF network error

The RBF network simulated had minimum error of the order of 1e^-14

VI. CONCLUSION

The result obtained from this study clearly reveals the efficiency of the neural network employed in short term fore-casting. The neural network was able to establish the non linear relationship of the load with the historical data supplied while training and simulation phase of the network, and there in the load in the next hour. It must be ensured however that the network is not over trained, as it reduces the generalization capability of the network.

REFERENCES

[1] A.S.Dehdasthi, J.R.Tudor, "Forecasting of Hourly Load, A Deterministic approach",IEEE Transactions on Power Apparatus and Systems Vol. PAS-101,pp 3290-3294,Sep. 1982

[2] J.H.Park, "Composite Modelling for Adaptive Short-Term Load Forecastin,"; IEEE Transactions on Power Systems, Vol. 6, May1991

[3] J.Y.Fan, J.D.McDonald, "Real time implementation of Short term Load Forecasting for distribution Power Systems," IEEE Transactions on Power Systems,Vol 9, May 1994.

[4] Nima Amjady,"Short-Term Hourly Load Forecasting Using Time-Series Modeling With Peak Load Estimation Capability," IEEE Transactions on Power Systems, vol. 16,Nov. 2011.

[5] S.Rahman, R. Bhatnagar,"An Expert System Based Algorithm for short term load forecassting," IEEE Transactions on power systems, vol 3, May 1988.

[6] S.Quaiyyam, "Artificial Neural Network Based Short Term Load Forecasting of Power Systems" International Journel of Computer Applications, Vol 30, Sep.2011

[7] T.S.Dillon, "An Adaptive Neural Network approach in Load Forecasting in a Power system", Proc. IEEE ANNPS, July 1991.doi: 91TH0374-9

[8] Hiroyuki Mori and Masatarou Ohmi, "Probabilistic Short-term Load Forecasting with Gaussian Process," ISAP vol 2, July 2005.

[9] M.Peng, N.F.Hunbele, G.G.Karady, "Conceptual approach to the Application of Network for Short-term Load Forecasting,," IEEE Proc.of 1990 ISCAS, May 1990.

◈◈◈

Lung Nodule Segmentation Using Active Contour Modeling And Gabor Filtering

Shravya K N & Rajaram M Gowda

Dept. of Information Science and Engineering, M S Ramaiah Institute of Technology,
MSR Nagar, Bangalore-560054, Karnataka, India
E-mail : shravya.kn@gmail.com

Abstract **-** The total number of CT Images is more due to the increased number of patients suffering from cancer and the repeated trials. Manually examining all the CT images is practically not possible and the accuracy is also low. Automating the process of lung segmentation hence plays a very important role. Two methods are detailed here, they being active contour modeling and Gabor filtering. In the method where active contour modeling is used first, the lung is segmented by Active Contour Modeling. Then, 2D stochastic features are used to detect regions of interest. Anatomical 3D feature is then used to detect nodules. Active Contour Modeling is used to extract contours at the end. The Gabor filtering technique is used to filter the input images and segment the nodules. The nodules attached to the chest wall are not always detected in Gabor filtering. The algorithms are combined, thereby trying to improve Gabor filtering i.e. Gabor filtering with the advantages of segmenting every nodule present in the lungs, including those attached to the chest wall.

Keywords **-** *Gabor filtering, Active Contour Modeling, nodules.*

I. INTRODUCTION

Lung cancer is one of the leading causes of death in modern countries. GLOBOCAN 2008 estimates show that about 12.7 million cancer cases and 7.6 million cancer deaths are estimated to have occurred in 2008; of these, 56% of the cases and 64% of the deaths occurred in the economically developing world. Lung cancer is the leading cancer site in males, comprising 17% of the total new cancer cases and 23% of the total cancer deaths [1].

The above statistics show that Lung cancer is one of the leading causes of death. The early stages of lung cancer have no symptoms which make the presence of lung cancer obvious. Diagnosing lung cancer at an early stage becomes very important.

Nodules which are attached to the chest wall might often go unnoticed. Hence importance has to be given to methods which help in detecting nodules attached to the chest wall. The part of the proposed idea is to use active contour modeling to segment the lungs initially. The region of interests are then determined and then the nodules are detected. Finally the nodule contours are extracted using active contour modeling. One of the main papers referred here is [8].

Gabor filtering is a simple filtering technique used for edge detection. [9] is another main paper which is referred to here. This technique has been used in image processing for a long time. Gabor filtering and active contour algorithm mentioned above are implemented. The total time taken by the two to process the CT scans of lung cancer in order to detect the lung cancer nodules is measured and compared. The false positives detected by both the algorithms are also compared. This is the initial step to check if the steps to detect the nodules attached to the chest walls, in the algorithm using active contour modeling can be combined with Gabor filtering. The time taken to detect all the nodules and reduce the false positives is also taken into consideration.

II. CONCEPT AND ALGORITHM

A. Lung Segmentation by active contour modeling

The basic steps of the algorithm are:

- Lung Segmentation by active contour modeling
- Regions of interest detection
- Nodule detection
- Nodule contour extraction by active contour modeling

Lung Segmentation by active contour modeling

In the first stage the CT image is binarized using an adaptive fuzzy thresholding [2]. A hole-free lung mask is obtained in the next step by applying windows of size

5*5 and 23*23. Small and large non-isolated nodules are segmented using windows of size 50*50 and 25*25 which are rotated 45 degree in the third stage. The chain code in [7] is used to obtain small nodules attached to the lung wall.

Some pixels of bone parts and pixels near to but outside the thorax located in the mid parts of the image are removed using a simple thresholding in the fourth step. At last the lung mask is used as the initial mask of the active contour model proposed by Lankton et al. [3].

The algorithm called Yezzi energy code, which is a global region-based energy that uses mean intensities proposed by Yezzi et al. is used here [4].

Regions of Interest Detection

The variance and mean are used as stochastic 2D features. The grey level of the lung area is different from the grey level of nodules. This has been observed clinically. This helps in using average pixels in a neighbourhood as a feature. Variance is another feature which can be used, as the intensity of the pixels of the nodules vary from the intensity of pixels of the background i.e. that of the lung area. The following equation is used to calculate mean

$$M_{ij} = \frac{1}{9} \sum_{k,l=-1}^{+1} L(i+k, j+l)$$

The following is the equation to calculate variance

$$V_{ij} = \frac{1}{9} \sum_{k,l=-1}^{+1} (L(i+k, j+l) - M_{ij})^2$$

where i and j are pixel coordinates.

Nodule Detection

Nodules are detected from the candidate Region of Interests. Anatomical feature is used for detecting nodules because the pixels of bronchus and nodules have similar grey levels. The previous and next slices are used to detect nodules in the present (target) slice. Experimental results indicate that 8 slices before and next to the target slice prove to be useful, where each lung CT scan is 0.625mm thickness [8]. The binarization of the result of previous section (which is called map) is used in this section as below: (change) z is the number of slice and map (z) is concluded from the target slice.

$N1 = \text{AND}\{map(z), map(z-1), \ldots, map(z-q)\}$

$N2 = \text{AND}\{map(z), map(z+1), \ldots, map(z+q)\}$

$N = \text{OR}\{N1, N2\}$

$q = \frac{c}{T}$, T: Thickness of slices

Nodule Contour Extraction by Active Contour Modeling

The previous section causes some small deformation in the contours of detected nodules. Active contour modeling is applied to solve this problem. Tiny parts, which were detected as nodules mistakenly, will be omitted. Thus, the value of False Positives is decreased.

B. Gabor Filtering

Gabor filtering is used for automatic detection in image processing. Gabor filter can achieve function of lower bounds of uncertainty relation on frequency and time domain. Gabor filter also has the best resolution characteristic in time and frequency domain. Due to the two reasons Gabor filtering can realize signal processing considering frequency and time domain [5]. 2D-Gabor function is a function of two-dimensional Gaussian function in a given frequency and direction revised by complex sine function [6].

$$G(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left[-\frac{1}{2}\left(\frac{x'^2}{\sigma_x{}^2} + \frac{y'^2}{\sigma_y{}^2}\right)\right] exp\left[2\Pi j \frac{x'}{\lambda}\right]$$

x' and y' are got by readjusting the angle θ, formulae:

$x' = x\cos\theta + y\sin\theta$

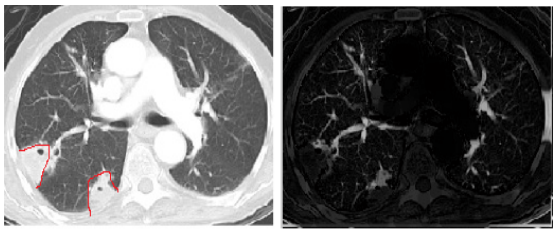
$y' = -x\sin\theta + y\cos\theta$

(x, y) is the pixel value for each point of the input image. $\frac{1}{\lambda}$ is the center frequency.

θ is the bearing. σ_x and σ_y are the spatial scale factor along x axis and y axis respectively.

III. RESULT

The active contour modeling method and Gabor filtering proposed here are automatic lung nodule segmentation methods. Thus no human intervention is required. The time taken to segment the nodules is very low. The performance is high. The number of False Positives is low [8] [9].



(a) (b)

Fig. 1: (a) Actual Image (b) Input Image

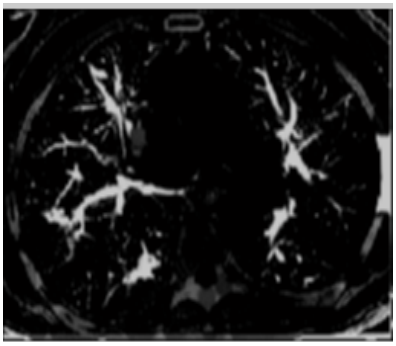


Fig. 2: Filtered image using Gabor filtering

As we can see plain Gabor filtering does not segment the nodules which area attached to the chest wall (marked in red in Figure 1.a).

Fig. 3: Result of active contour modeling, input image being the image in Figure 1.

The nodules not segmented in Gabor filtering are segmented when Active Contour Modeling is used.

IV. CONCLUSION

The results of a combining Gabor Filtering with Active Contour Modeling is better than individual algorithms, as the combined algorithm segments all the nodules, those attached to the chest wall also.

Both the methods show reduced false positives. The time taken to segment the nodules is low. Gabor filtering can leave out the nodules attached to the wall sometimes. This is overcome by the other algorithm which uses active contour modeling. This can be used in Gabor filtering, thus improving Gabor filtering.

Future work will be to try and parallelize the code, and check if there is any improvement.

REFERENCES

[1] CA: A Cancer Journal for Clinicians Volume 61, Issue 2, Article first published online: 4 FEB 2011.

[2] X. Ye, X. Lin, J. Dehmeshki, G. Slabaugh, and G. Beddoe, “Shape-based computer- aided detection of lung nodules in thoracic CT images,_ IEEE Trans. Biomedical

Engineering, vol. 56, pp. 1810-1820, Jul. 2009.

[3] S. Lankton, and A. Tannenbaum, “Localizing region-based active contours,_ IEEE Trans. Image Processing, vol. 17, pp. 2029-2039, Nov. 2008

[4] A. Yezzi, A. Tsai, and A. Willsky, “A fully global approach to image segmentation via coupled curve evolution equations, Journal of Visual Communication and Image Representation, vol. 13, pp. 195-216, Jan. 2002.

[5] Liao Kaiyang, Zhang Xuedong, Zhang Mingzhu, et al. Fast fingerprint enhancement based on Gabor filter[J]. Computer engineering and application, 2009,45 (10) :172-175.

[6] Daugman J G. Uncertainty relation for resolution in space, spatial frequency and orientation optimized by two-dimensional visual cortical filters [J].Journal of the Optical Society of America,.1985,2(7):1160-1169.

[7] X. Ye, X. Lin, J. Dehmeshki, G. Slabaugh, and G. Beddoe, “Shape-based computer-aided detection of lung nodules in thoracic CT images,_ IEEE Trans. Biomedical Engineering, vol. 56, pp. 1810-1820, Jul. 2009.

[8] Mohsen Keshani, Zohreh Azimifar and Reza Boostani, Alireza Shakibafar, “Lung Nodule Segmentation Using Active Contour Modeling” 2010 IEEE.

[9] FAN Li-nan, SUN Shen-shen, LI Dao-jing and CHANG Chao-hai, “Applied research on the automatic detection of lung nodules ROI based on Top-hat and Gabor filter”, Third International Symposium on Information Processing, 2010 IEEE.

◈◈◈

Extending and Optimizing the Visual Studio 2010 Editor for Mainframes

Abhishek Anand & Shashidhara HS

Department of Information Science Engineering, MSRIT, Bangalore, India
E-mail : aanand003@gmail.com, hs.shashidhara@gmail.com

Abstract **-** Many Companies have environment for disparate legacy systems, applications, Processes and data sources. The challenges for the industry today are to maintain these legacy systems. These legacy systems are based on legacy operating systems and programming languages. These Operating Systems show specific file structure and representation for these files. There are separate softwares to edit, Compile and run the Program. This leads to the requirement for an IDE which can provide all these functionalities at the same place. Most of the modern IDEs does not have support for these languages and for these Operating Systems. Creating an IDE from the scratch is very Complex and time Consuming task. Extending the IDE thus comes as a better solution. One of the most Important and potential candidate to be optimized in an IDE is editor as it is responsible for most of the data transfer between the host and the client system. This paper presents our approach to make the editor part efficient to perform with minimum Latency and response time.

Keywords*- Visual Studio 2010, Legacy Systems, Mainframes, Performance.*

I. INTRODUCTION

Software Systems are critical assets for companies and incorporate key knowledge acquired over lifetime of an organization. The development of these system costs huge money and time to the companies. This persuades the companies to make these systems up and running for longer span of time. These systems (Legacy systems) continue to run on old operating systems and languages. Thus the maintenance of these systems is a challenge for the new community of the developers. With the changing way of coding and with the domination of IDEs has made the companies to look forward to create IDEs to support their Programmers.

Microsoft Visual Studio 2010 was selected to be extended because of it is one of the most widely used IDE in the Programmer's Community. There are several levels of extensibility for the MS Visual Studio with increasing complexity but giving more control over Visual Studio than its predecessor which is given below.

1) Macros
2) Add-Ins
3) Packages

One type of Package is Managed VS Package. This is provided by managed Package Framework which gives us freedom to develop the extension in the managed languages like C#. It reduces our work by relieving us from working with GUIDs and cryptic Interface names.

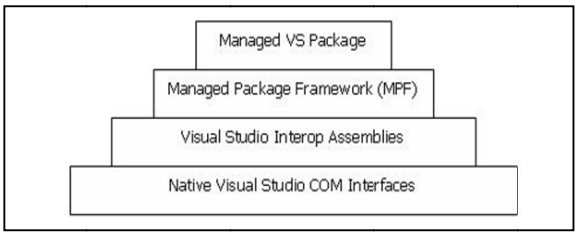


Fig. 1: Interaction between COM and MPF

Following is the generalized architecture of Visual Studio Extensibility:

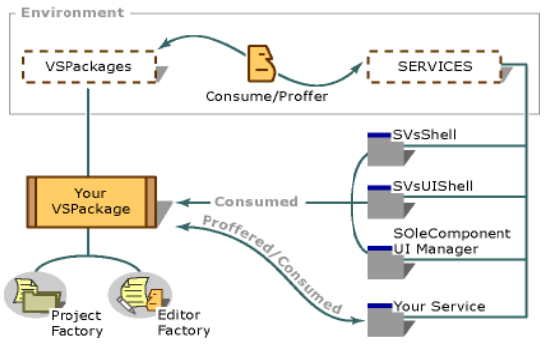


Fig. 2 : General Visual Studio Extensibility Architecture

Application functionality is provided by VSPackage, which is hosted by the IDE. Several services are offered by the VSPackage as the above image demonstrates. The VSPackage and Services show bidirectional relationship. The Project Factory extends the Project related functionalities like build, Project Node, Debug etc. The Editor Factory is responsible for the Extension of the Editor.

The mainframe which we chose for this paper is Unisys Mainframes. Unisys Mainframes are one of the most dominating in its category serving many of the most secure, reliable and scalable infrastructures. Unisys Mainframes supports many proprietary Operating Systems. We are aiming our extended Visual Studio to support these OS.

These OS, the Server and the files supported by the OS have some features which made us to optimize the editor part of the extension, e.g.

a) The Server supports only 300 records per socket call to be fetched

b) The size of the file on these mainframes are huge (thousands lines of codes per file).

c) It has different representation of the file than the Visual Studio Supports.

d) It supports file name with Special Characters (e.g. *, /)

e) The files on these OS do not have extensions.

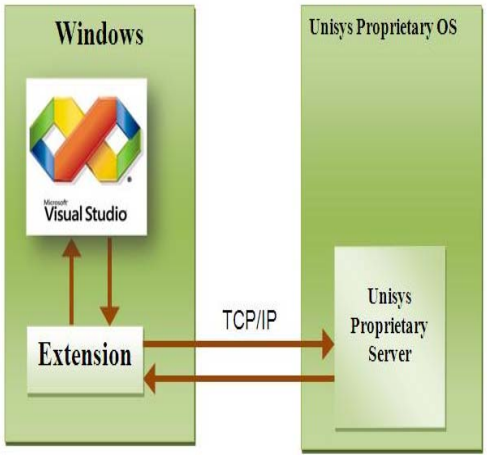


Fig. 3:High Level Architecture of the Extension

The aim of the Extension to provide a common editor supporting multiple Unisys Proprietary OS languages such as COBOL74, COBOL 85, ALGOL, DCALGOL, DMALGOL, WFL etc. Most of the operation regarding these files like build happens on the server side. We just need to send message in the prescribed format. But, in case of editor we fetch the file data to the client side and this makes the editor a valid candidate for the optimization.

We will continue with the rest of the Paper, considering the above facts. The Section II of the paper will deal about the implementation details of the Editor Extension. In section III will deal about the Optimization and results. In section IV will give the Conclusion and Section V will provide the References.

II. IMPLEMENTATION

A) Creating a Custom Editor

The Microsoft visual Studio 2010 provides two ways of creating editor extension, either we can use the Microsoft default editor which is provided to us as Managed Extension Framework (MEF) and customize it according to our requirements or can create a custom editor. The requirement was to divide the editor in three sections to display the files supported by Unisys Proprietary OS. The complexity of the above requirement and the time constraint has constrained us to use the default Visual Studio Editor. Thus the custom Editor approach is used to create and extend the Editor. For this we have used RichTextBox Control under the Forms namespace. Three different RichTextBox have been used and hooked their controls together so that they can work in synch with each other. We also have removed the scroll bars for two of the RichTextBox and left the one scroll bar to make it similar to the custom editor.

B) Extending the Visual Studio with our Custom Editor

Visual studio provides Editor Factory, which implements IVsEditorFactory interface. We need to register the file types to make the visual studio understand that the file should be opened in the custom editor. The Editor factory creates the Editor Pane object which work as a container for the editor. The Editor Pane loads the editor after checking the connection to the Server. If the connection is there it fetches the data from the server and displays on the editor. There are several things we may talk about when extending the editor. But in this paper we will talk on open and save features because these are most data intensive operations and the optimization of these may provide a better performance for the application. The sequence diagram below tries to explain how the open and save operation works.

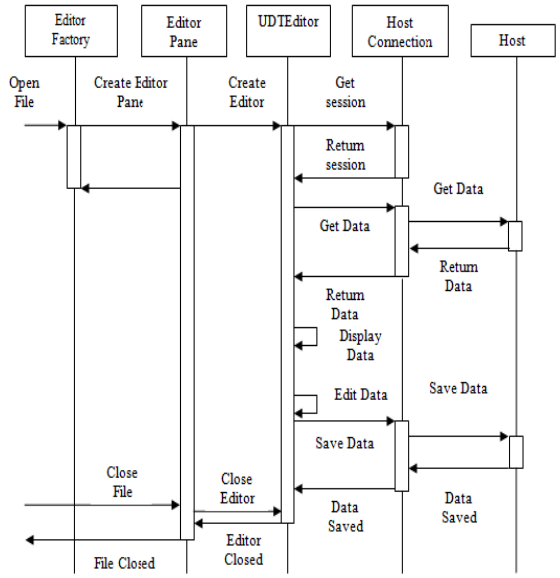


Fig. 4 : Sequence Diagram to Open and Save Functionality

C) Pseudo code to Open and Save the File without Optimization

```
START
Command=Get the user command
If Command == New File Then
      Create a new File
End If
Else If Command == Open File Then
      Connection = Get Connection Object
      If connection == null then
      Create Connection
      End If
If Editor is Empty then
      Get 300 Records from the 0th index
      Save the Records in memory
      Display the Records in the Editor
End If
If File Scrolled down is True && !End of Editor
 && !EOF then
      Get the next 300 Records from the Current position
      Append the Content to previous record in Memory
      Display the Record in the Editor
End If
If File Scrolled Up == True && !Begin of editor then
      If Editor Content Modified then
      Save the Modified data Locally.
      Display the data from saved Records
      End If
End If
End If
END
```

Fig.5 : Open File without Optimization

```
START
Command = get Command from the User
If Command == Save File then
If file is not existing then
      Create a File on the Host, newFile
      Copy Complete file data to the Host file, newFile
      Save newFile
End If
Else if file is existing then
      NewFile = Create a copy of the actual file on the Host.
      editFile = Create a File with Edited Data on Host
      Transfer the edited data and save in the editFile
      While EOF != true
      recordModified = Compare Records of editFile and newFile
      If recordModified == true then
      Replace the Records in newFile with the Modified data
      End If
      End While
      Replace actual File with newFile
End If
End If
END
```

Fig. 6 : Save File without Optimization

The above pseudo codes provide the opening and saving the File without any optimization. While Opening the File when the user requests for the data first time, the extension fetches the record from the server and saves the data in the memory and displays it on the editor. Further if the user requests for the data it again fetches the data from the host and displays it on the editor. The Up scroll is not considered here because it will display the data which has been already fetched and thus will not involve in any communication with the host. This leads to delay each time the user requests for the data.

For saving the file, we have two cases to consider. The first is to create a new file, new file save cannot be optimized because every time a new will be saved, the complete data is transferred and thus there is no scope for optimization. But, for the existing file as we can see, the content is transferred to the host system when the user presses the save button thus creating a delay based on the no of line modified. If the no of lines modified is large then this will produce larger delay. This produces a scope for improvement.

III. OPTIMIZATION AND RESULTS

A) Optimization

Performance and delay is one of the major considerations in today's software community and when it comes to the mainframes it becomes more vital issues.

The pseudo codes presented in the above section have some performance glitches which makes the user unsatisfied. To overcome the performance issue presented above, we present the modified approach which will reduce the latency to its minimum. The pseudo codes presented below demonstrate the approach.

```
START
Command=Get the user command
If Command == New File Then
        Create a new File
End If
Else If Command == Open File Then
        connection = Get Connection Object
        If connection  == null then
        Create Connection
End If
If Editor is Empty then
        Get two threads
         Thread1,
        Record1 = Get 300 records from the 0th index
        Save the Record1 in memory
        Display the record1 in the Editor
        Thread 2,
        If EOF == true End
        Record2 = null
        Return
        End If
        Record2 = get next 300 records in the memory
End If
If File Scrolled down == True && End of editor == false
         Record3! = null then
        Get two threads
         Thread1,
        If content edited before scroll
                transfer edited data to Host
        End If
        Append Record3 to Record1
        Display next records from the Record
        Thread 2,
        If EOF == true End
        Record2 = null
        Return
        End If
        Record2 = get next 300 records in the memory
        End If
        Record3 == Record2
End If
If File scrolled up = true & !Begin of Editor
        If content edited before scroll
                transfer edited data to Host
        End If
        Display the records from saved records
End If
END
```

Fig. 7 : Optimized Open

```
START
Command = get Command from the User
If Command == Save File then
If file is not existing then
      Create a File on the Host, newFile
      Copy Complete file data to the Host file, newFile
      Save newFile
End If
Else if file is existing then
      NewFile = Create a copy of the actual file on the Host.
      editFile = the file with edited data created during file open
      While EOF != true
      recordModified = Compare Records of editFile and newFile
      If recordModified == true then
      Replace the Records in newFile with the Modified data
      End If
      End While
      Replace actual File with newFile
End If
End If
END
```

Fig. 8 : Optimized save

In our approach for optimizing the open use case we have used threading concept during the scroll operation. When the fetches the data at the first one thread fetches the records to be displayed on the editor and other thread fetches the next 300 records and save the data in the memory to be displayed further. In the case of scrolling if the user reaches the end of the editor one thread fetches the next 300 records and the other will display the previously fetched records. It is also checks for the modifications in the last scrolled records and is stored on the host unlike the last pseudo code where all the modified records were copied at the end.

Saving the record during scrolling also reduces the time taken to save the file. Here, we have only shown for the down scroll but modified record save will happen even for the up scroll. Thus the only work done during save is to copy and replace the record from the modified file to the new file and to save the new file.

B) Results

The graphs below will give some insight in the improvement achieved by optimizing the approach to reduce the latency and achieve high performance.

The first two helps us to make us the difference between optimized and unoptimized open. When open is not optimized, it has a constant delay until the complete data has been opened in the editor. But, after optimizing the open after the first 300 records the delay is reduced to almost null, the small delay is introduced because of the modified data we are transferring for optimizing the save.

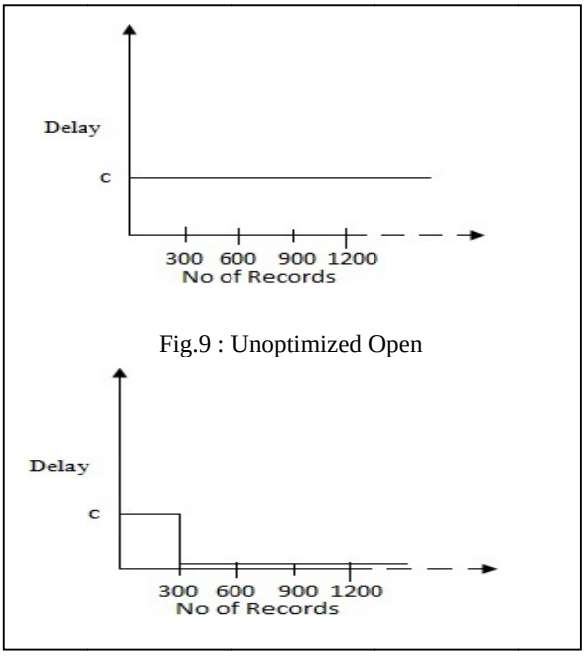


Fig.9 : Unoptimized Open

Fig. 10 : Optimized Open

The below graph shows the difference between the optimized and unoptimized save. The dotted line shows that with unoptimized save the delay or time taken keep on increasing with the no of records modified. The time taken still increase with optimized but at lower rate, this increase indicates the time taken to compare and copy the records from the modified file to new file. This performance improvement is introduced because, the modified data is transferred before the save operation is initialized.

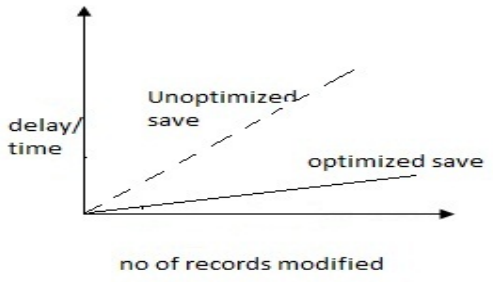


Fig. 11 : Unoptimized save Vs Optimized save

IV. CONCLUSION

The Visual Studio Editor is optimized with respect to opening and saving file to give better performance. As the graphs clearly indicate optimized open decreases the delay to a significant amount. The performance is linearly increasing with the number of records to be saved. As there is no other choice except to use the legacy system, this increase in performance has definitely helped in optimization of further application which are written on top of existing legacy systems

REFERENCES

[1] Extending Editor, http://msdn.microsoft.com /en-us/library/dd885242.aspx

[2] Software Architecture in Practice, 2nd Edition, Len Bass, Paul C. Clements, and Rick Kazman

[3] Software Engineering, Sixth Edition, Ian Somerville

[4] UML, http://www.uml.org/

[5] ClearPath Mainframe/ MCP: http://www.unisys.com/unisys/theme/index.jsp?id=16000034

[6] Burroughs's MCP, http://en.wikipedia.org/wiki/Burroughs_MCP

[7] Burroughs's large Systems, http://en.wikipedia.org/wiki/Burroughs_large_systems

◈◈◈

A Novel Face Detection and Tracking Algorithm in Real- Time Video Sequences

Nagapriya kamath K, Ashwini Holla & Subramanya Bhat

Dept. of Electronics and Communication, Canara Engineering College
Mangalore, Karnataka
E-mail : nagapriyakamath@gmail.com

Abstract **-** Face detection is a image processing technology that determines the location and size of human faces in digital images or video. This module precedes face recognition systems that plays an important role in applications such as video surveillance, human computer interaction and so on. This proposed work focuses mainly on multiple face detection technique, taking into account the variations in digital images or video such as face pose, appearances and illumination. The work is based on skin color model in YCbCr and HSV color space. First stage of this proposed method is to develop a skin color model and then applying the skin color segmentation in order to specify all skin regions in an image. Secondly, a template matching is done to assure that the segmented image does not contain any non-facial part. This algorithm works to be robust and efficient.

Keywords **-** *Skin color model, segmentation, template matching.*

I. INTRODUCTION

The goal of this proposed method is to develop multiple face detection and tracking system in real time scenario based on a fast and efficient face segmentation approach. This method consists of two image processing steps. Separate the human skin regions from non-skin regions thereby locating human faces within the skin regions. The next step is to use template matching to assure the segmented image does not contain any non-facial part. There are two types of face detection systems: face detection in still images and face detection in real time video sequences. Few approaches used in face detection are mainly based on PCA[1],neural network, edge detection based, COM and ROI based[2], feature based[3],color based [3][4] etc. Color based face detection and tracking has the following drawbacks: It gives a coarse face segmentation giving spurious results with a background cluttered with skin colored regions and found to be unsuitable when subjects in the image wear skin colored dress or their attire has patches of skin color. To overcome these drawbacks, some heuristics are developed that includes parametric skin color modeling, adaptive thresholding and dynamic template matching for face region detection and tracking. In this proposed work, a reliable skin color model is developed. Further, skin color segmentation, adaptive and optimal thresholding is done. Face detection is achieved by Euler number, heuristics and template matching. The balance of this paper is organized as follows. Methodology is discussed in section II, skin color modeling is proposed in section III, segmentation is proposed in section IV and face detection and tracking algorithm is discussed in section V. Experimental results are presented in section VI. Conclusions are made in section VII.

II. METHODOLOGY

This proposed method uses a skin color model approach in which skin color region is separated from non skin color region. To obtain skin color region, RGB color image is converted into YCbCr and HSV color space. The chromacity is fitted over a Gaussian distribution and applied over the image to obtain a skin likelihood image which is then converted into binary format by applying adaptive and optimal thresholding respectively. In order to obtain the face region from this binary image, the number of holes from the connected components in the binary image using Euler's number is determined. If the number of holes in a given connected component is greater than '1', then the centroid and orientation of skin region w.r.t. the center is determined. The height to width ratio of the oriented image is also determined and compared by developing the heuristics model. If height to width ratio is within the range of 0.8 to 1.6, then template matching is done using a template face. Template matching is performed by finding out the cross-correlation between the skin region and template

face images. If cross-correlation exceeds 0.8 then the skin region is identified as a human face.

III. SKIN COLOR MODEL

In order to segment the human skin regions from non-skin regions based on color, a reliable skin color model [1] that is adaptable to people of different skin colors and to different lighting conditions is necessary. In this work, a skin color model in chromatic color space is used for segmenting the skin region. The common RGB representation of color images is not suitable for characterizing skin-color as, the triple component represents not only color but also luminance. Luminance may vary across a person's face due to the ambient lighting and thus is not a reliable measure in separating skin from non-skin region. Luminance can be removed from the color representation in the chromatic color space. Chromatic colors in the absence of luminance, is defined by a normalization process [2] as shown in Eqn (1).

$$r = R/(R+G+B)\ ;\ g = G/(R+G+B);\quad b = B/(R+G+B) \tag{1}$$

Although skin colors of different people appear to vary over a wide range, they differ much less in color than in brightness.ie: skin colors of different people are very close, but differ mainly in intensities. Hence, YCbCr and HSV color model are often used to implement the skin color region. In the RGB domain, each component of the picture has a different brightness. However, in the YCbCr domain all information about the brightness is given by the Y- component, since the Cb (blue) and Cr (red) components are independent from the luminosity. The following conversions are used to segment the RGB image into Y, Cb and Cr components are given by the Eqn (2).

$$\begin{aligned} Y &= 0.257 * R + 0.504 * G + 0.098 * B + 16 \\ Cb &= 0.148 * R - 0.291 * G + 0.439 * B + 128 \\ Cr &= 0.439 * R - 0.368 * G - 0.071 * B + 128 \end{aligned} \tag{2}$$

The Cb and Cr components give a good indication on whether a pixel is a part of the skin or not. HSV color space represents colors in terms of Hue, Saturation and Intensity of the given pixel. Hue refers to color type, such as red, blue, or yellow. Saturation refers to the purity of the color. Value component refers to the brightness of the color. A total of 100 color images were used to determine the color distribution of human skin in chromatic color space. These samples were taken from persons of different ethnicities: Asian, American and African. As the skin samples were extracted from these color images, the skin samples were filtered using a low-pass filter to reduce the effect of noise in the samples. The impulse response of the low-pass filter is given by:

$$H(Z) = \frac{1}{9}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

The color histogram revealed that the distributions of skin color of different people are clustered in the chromatic color space and a skin color distribution can be represented by a Gaussian model N (m, C) where

$$\text{Mean: } m = \frac{1}{N}\sum_{0}^{N-1} x \quad \text{Where } x = (Cr, Cb)^T \tag{3}$$

$$\text{Covariance: } C = \frac{1}{N}\sum_{0}^{N-1}\left[(x-m)(x-m)^T\right] \tag{4}$$

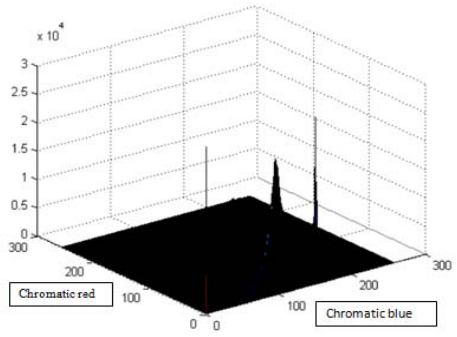


Fig 3.1: Fitting skin color into a Gaussian distribution.

With this Gaussian fitted skin color model, the likelihood of skin for any pixel of an image is obtained. Therefore, if a pixel, having transformed from RGB color space to chromatic color space has a chromatic pair value of (Cr, Cb); the likelihood of skin for this pixel is computed using Eqn (5).

$$\text{Likelihood} = P(r, b) = \exp\left[-0.5(x-m)^T C^{-1}(x-m)\right] \tag{5}$$

IV. SKIN COLOR SEGMENTATION

The first stage in segmentation [4] [5] is to transform an image into a skin-likelihood image. This involves transforming every pixel from RGB representation to chroma representation and determining the likelihood value based on the Eqn. (5). The skin-likelihood image is a gray-scale image whose gray values represent the likelihood of the pixel belonging to skin. A sample color image and its resulting skin-likelihood image are shown in Figure 4.1(a) and (b). It is seen that all skin regions like the face, hands and the arms appear brighter than the non-skin region. Since the skin regions are brighter than the other parts of the

images, the skin regions need to be segmented from the rest of the image through a thresholding process. Since people with different skins have different likelihoods, an adaptive thresholding process is used to achieve the optimal threshold value for each run.

i) ADAPTIVE AND OPTIMAL THRESHOLDING

The adaptive thresholding means stepping down the threshold value which intuitively increases the segmented region. However, the increase in segmented region will gradually decrease as percentage of skin regions detected approaches 100%, but will increase sharply when the threshold value is considerably too small that other non-skin regions get included. The threshold value at which the minimum increase in region size is observed while stepping down the threshold value gives the optimal threshold. In this proposed work, the threshold value is decremented from 0.85 to 0.05 in steps of 0.1. If the minimum increase occurs when the threshold value changes from 0.45 to 0.35, then the optimal threshold must be taken as 0.4.It is not necessary that all detected skin regions contain faces, many regions often correspond to hands, arms and other exposed parts of the body, while some correspond to objects with colors similar to those of the skin. The skin segmented image of the original color image resulting from this technique is shown in fig.4.1(c). Hence the second stage of face detection algorithm employs facial features to locate the face in all these skin regions.

(a) (b) (c)

Fig. 4.2 (a) original color image, (b) skin likelihood image, (c) skin segmented image.

To determine which regions possibly determine a human face, the number of skin regions in an image need to be determined. A skin region is defined as a closed region in the image, which has a 0, 1 or more holes in it. A skin region can be thought of as a set of connected components within an image. All holes in a binary image have pixel value of zero (black).The process of determining how many regions are present in a binary image is by labelling such regions. A label is an integer value. A 8-connected neighborhood (i.e., all the neighbors of a pixel) is often used to determine the labeling of a pixel. If any of the neighbors had a label, then label the current pixel with that label. If not, a new label is used. Finally, we count the number of labels and this gives the number of regions in the segmented image. To separate each of the regions, a new image is being created that have ones in the positions where the label occurs. The others are set to zero. After this, we iterate through each of the regions found in order to determine if the region contains a human face or not.

V. FACE DETECTION AND TRACKING

After experimenting with several images, it is found that a skin region should have at least one hole inside that region. Regions that have no holes are neglected. The number of holes inside a region is compute using the Euler number of the region given in Eqn. (6)

$$E = C - H \tag{6}$$

Where C: number of connected components and H is the number of holes in a region. Since we are considering one skin region at a time, we set the number of connected components (i.e. the skin region) to 1. Therefore, numbers of holes,

$$H = 1 - E \tag{7}$$

a) Center of the mass

The center of area in binary images is the same as the center of the mass and it is computed using Eqn (8) and Eqn (9).

$$\overline{x} = \frac{1}{A}\sum_{i=1}^{n}\sum_{j=1}^{m} jB(i, j) \tag{8}$$

$$\overline{y} = \frac{1}{A}\sum_{i=1}^{n}\sum_{j=1}^{m} iB(i, j) \tag{9}$$

Where B is the matrix of size [n x m] that represents the region and A is the area in pixels of the region.

b) Orientation

Most of the faces considered in this work are vertically oriented while some of them have a little inclination. A unique orientation is possible by elongating the object. The orientation of the axis of elongation determines the orientation of the region. The axis is computed by finding the line for which the sum of the squared distances between region points and the line is minimum. The least-squares of a line are computed to the region points in the image. Angle of inclination (theta) is given by Eqn. (10):

$$\theta = \tfrac{1}{2}\tan^{-1}\frac{b}{a - c} \tag{10}$$

Where: $a = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(x'_{ij}\right)^2 B[i, j]$

$$b = 2\sum_{i=1}^{n}\sum_{j=1}^{m} x'_{ij}\, y'_{ij}\, B[i,\ j]$$

$$c = \sum_{i=1}^{n}\sum_{j=1}^{m} \left(y'_{ij}\right)^2 B[i, j]$$

And $x' = x - \bar{x}$; $y' = y - \bar{y}$

c) Height-to-width ratio

To determine the width and height of the region, it is necessary to resize the template face so that it has the same width and height of skin region. First, we fill out the holes that the region have. Since the image is rotated by an angle theta, we need to rotate our region by theta degrees so that it becomes completely vertical. Height and width is determined by moving 4 pointers: one from the left, right, top and bottom of the image. If we find a pixel value different from 0, we stop and this marks the coordinate of a boundary. When we have the 4 values, we compute the height by subtracting the bottom and top values and the width by subtracting the right and the left values.

d) Template matching

For an image corresponding to the skin region, we first close the holes in the region and multiply this image by the original one. In template matching[4][5][6], the template face must be positioned and rotated in the same coordinates as the skin region image. Initially, the template frontal face as shown in Fig.5.1 (a) is resized according to the height and width of the region computed and the resized template face is then rotated according to angle theta so that the template face is aligned in the same direction as the skin region. The center of the rotated template face is computed as shown in the previous section. This process creates the grayscale image that will have the resized and the rotated template face model. The cross-correlation value between the parts of an image corresponding to the skin region is computed and the template face properly processed and centered. It was determined experimentally that a good threshold value for classifying a region as a face is if the resulting cross-correlation value is greater than 0.6.To plot the rectangular box, co-ordinates of the template is found that exactly determines the pattern the template face can have on the resultant matched image shown in Fig 5.1(b). With these coordinates, we draw a rectangle in the original color image shown in Fig.5.1(c) which marks the output of the face detection system.

(a) (b) (c)

Fig 5.1 (a) Original template , (b) Template matched output and (c) face detected output.

VI. RESULT AND ANALYSIS

This algorithm is found to detect the face for different illumination, orientations and poses tested across a wide range of subjects. It is also found to work well with partially occluded faces. It detects only the face region and does not detect any other skin colored objects. This algorithm works effectively for real time implementation. Fig.6.1 (a), (b) and (C) depicts real-time face detection under different illumination and position using MATLAB GUI.

Fig 6.1: (a) (b) (c)

VII. CONCLUSION

The proposed method is face detection and tracking system in real-time video sequences that captures an image sequence from a camera, segments, detects and tracks efficiently a human face. It is an optimal blend of skin color model, heuristics and template face that achieves high performance in real time. This entire algorithm for face segmentation, detection and tracking is implemented in MATLAB R2008b using MATLAB Image Processing toolbox that takes around 20secs per frame in Intel core i3 processor to segment, detect and track multiple faces. We used nearly 100 images to test the performance of this implementation and obtained 80% of accuracy. Real time implementation is done using MATLAB GUI. This proposed method can be further developed to incorporate the real-time face recognition system.

REFERENCES

[1] Liying Lang and Weiwei Gu, "The Face Detection Algorithm Combined Skin Color Segmentation and PCA", IEEE Trans 2009.

[2] J.Fritsch,S.Lang,G.A Fink, "Improving Adaptive skin color segmentation by incorporating results from face detection", Proceedings of 2002 IEEE, Workshop on robot & human interactive communication, pp 337-343.

[3] Mohamed A Berbar, "Faces and Facial Features Detection in Color Images", proceedings of Geometric modelling and imaging, IEEE Trans 2006.

[4] Chen Aiping, Pan Lian, Tong Yaobin, Ning Ning, "Face Detection Technology Based on Skin Color Segmentation and Template Matching", 2010 Second International Workshop on Education Technology and Computer Science ,IEEE trans 2010,pp 708-711.

[5] Xiaoping Li , Yinxing Li, Yongbiao Yao, Lin Zhang, "Study and Realization of Face Detection Based on Skin Segmentation and Template Matching", pp 375-378.

[6] Jizeng Wang and Hongmei Yang, "Face Detection Based on Template Matching and 2DPCA Algorithm", IEEE Trans 2008.

[7] Rafael Gonzalez and Richard E Woods, "Digital Image Processing", 2nd edition, Pearson Education, 2001.

[8] Ani.K.Jain, "Fundamentals of digital image processing", 2nd edition, Pearson Education, PHI, 2001.

◈◈◈

Analysing the Error Immunity of Nanoscale Digital Logic Circuits

R.Kavitha & P.Balasubramanian
Department of Electronics and communication Engineering,
Vel Tech Dr. RR and Dr. SR Technical University, NO 42, Avadi-Veltech road, Avadi, Chennai -600 062
E-mail : kavisumm@gmail.com, Dr.p.balasubramanian@gmail.com

Abstract - Theorems are developed which prove that certain logic functions are more robust to errors than others. These theorems are used to construct datapath circuits that give an increased immunity to error over other naive implementations. A link between probabilistic operation and ultra-low energy computing has been shown in prior work. These novel theorems and designs will be used to further improve probabilistic design of ultra-low power datapaths. This culminates in an asynchronous design for the maximum amount of energy savings per a given error rate.

I. INTRODUCTION

As digital technology marches on, ultra-low voltage operation, atomic device sizes, device mismatch, and thermal noise are becoming commonplace and so are the significant error rates that accompany them. These phenomena are causing ever increasing bit-error rates, and with billion-transistor digital chips being produced today, even a 1-in-100-million bit error rate becomes costly.

This paper will present a novel discovery of boolean logic that certain logic gates are more robust to error than others, and in fact it will be shown that some logic even *improves* the error rate just through natural computation. The paper will show how these principles translate into CMOS and other implementations, but these principles are *independent of technological implementation* since they are properties of boolean logic itself. Thus these design principles will stand the test of time.

Recently, ultra-low power computing has been achieved by lowering the supply voltage of digital circuits into near threshold or even the subthreshold region. Indeed a fundamental limit to voltage scaling technology has been proposed: the thermodynamic limit of these devices [4]. When the supply voltage becomes comparable to thermal noise levels in these types of ultra-low power designs, devices start to behave probabilistically giving an incorrect output with some nonzero probability.

II. MAIN CONTRIBUTION OF THIS PAPER

Logic gates are the fundamental building blocks of all digital technology, and it has been discovered that not all logic gates propagate error equally regardless of technological implementation. This paper analyses the following.

1. Certain gates have better error immunity when compare to other logic gates.
2. Discrete gate based implementation of the function has better error immunity than a complex gate based realization.
3. Gates that have cardinality of ON set or cardinality of OFF set is one, have better error immunity with increase in the number if inputs.
4. Gates that have cardinality of ON set is equal to cardinality of OFF set, have similar error immunity when implemented as directly or cascade.

III. ALL LOGIC GATES IS NOT CREATED EQUAL

Boolean logic functions are most simply represented by a *TruthTable* mapping each input combination to an output. When a bit-error is present at the input, or one of the input bits is flipped in other words, if this new input combination is mapped to the same output, then no error results in calculating the given logic function. In this case the logic function did

not *propagate* the error. If one assumes a small, given error rate per bit, then a single-bit error is more likely than two simultaneous bit errors which is more likely than three simultaneous errors, and so forth. A logic function that has the least input-output mappings where a single-bit error on an input causes a mapping to a different output will be the least likely to propagate a bit error. This phenomenon is shown in Figures 1(a) and 1(b) using an NAND function and an XOR function as an example.

Figure 1 illustrates the theme that not all logic gates propagate errors equally. If one calculates the average probability of error across all possible input combinations of NAND and XOR logic, an extremely interesting result emerges. Assume that a probability of error of ε = 0.2is present at the inputs of these gates. Further assume that full-adders are built such that a probability of error at the input ε= 0.2 is also present and that one of the full-adders is built with an NAND-NAND implementation and the other is built with a standard XOR implementation.

NAND w/ input error

Bit flip	In *AB*	Out	Out error?
A	10	1	
None	00	1	
B	01	1	
A	11	0	Error!
None	01	1	
B	00	1	
A	00	1	
None	10	1	
B	11	0	Error!
A	01	1	Error!
None	11	0	
B	10	1	Error!
# output errors/possible			4/8

NAND truth table

In *AB*	Out
00	1
01	1
10	1
11	0

(a)

XOR w/ input error

Bit flip	In *AB*	Out	Out error?
A	10	1	Error!
None	00	0	
B	01	1	Error!
A	11	0	Error!
None	01	1	
B	00	0	Error!
A	00	0	Error!
None	10	1	
B	11	0	Error!
A	01	1	Error!
None	11	0	
B	10	1	Error!
# output errors/possible			8/8

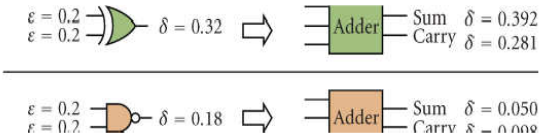
XOR truth table

In *AB*	Out
00	0
01	1
10	1
11	0

(b)

Figure 1: The effect of a single-bit error or bit flip on the inputs of an NAND function and XOR function is shown. Note that if the new input combination resulting from a bit flip maps to the same output as the input combo with no bit flip, no error is seen at the output. A NAND function is much less likely to propagate an input bit flip to the output resulting in an error than an XOR function. (a) NAND function (b) XOR function

Figure 2 shows the drastically differing output probabilities δ that result. Note that in Figure 2, the gate itself has no probability of computing erroneously associated with it.



This shows that different gates have differing abilities to mask errors regardless of input combination input. Further, δ is compared for a full-adder implemented with NAND-NAND and with a standard XOR when E = 0.2ispresent at the input. Note that not only is the probability of output error δ much lower for a NAND gate than an XOR gate, but that the error propagated from the NAND is *improved* over the input

error. In a sense, a NAND "heals" the circuit. Note that the gate itself does not compute erroneously in this example.

As shown in Figure 2, the NAND gate has a much lower output probability of error, δ, than the XOR gate when under the same input error conditions. An interesting property also emerges that a NAND actually propagates an *improved* error probability due to its logic properties through its natural computation. Obviously, a $\boldsymbol{z}$ = 0.2 is orders of magnitude higher than the actual error rates seen in digital logic, but it will later be shown that the output error probability of logic gates monotonically decreases as a function of decreasing error rates and thus this concept holds true at more realistic bit-error rates. Subsequent sections will go into more depth explaining these properties.

IV. OUTPUT ERROR PROBABILITY FOR AND GATE

Truth table

A	B	Z
0	0	0
0	1	0
1	0	0
1	1	1

Error probability for input a, ε_a

Error probability for input b, ε_b

'00' -> 11 - $\varepsilon_a\varepsilon_b$

'01' -> 11 - ε_a (1- ε_b)

'10' -> 11 - ε_b (1- ε_a)

'11' -> 00 - $\varepsilon_a\varepsilon_b$

-> 01 - ε_a (1- ε_b)

-> 10 - ε_b (1- ε_a)

Here $|Y_{ON}| = 1$

Output error probability,

$$\delta(\text{AND2}) = \frac{2\varepsilon_a\varepsilon_b + 2\varepsilon_a(1-\varepsilon_b) + 2(1-\varepsilon_a)\varepsilon_b}{4}$$

$$= \frac{\varepsilon_a + (1-\varepsilon_a)\varepsilon_b}{2}$$

simillarly

Output error probability

$$\delta(\text{AND3}) = \frac{\varepsilon_a + (1-\varepsilon_a)[\varepsilon_b + \varepsilon_c - \varepsilon_b\varepsilon_c]}{4}$$

Output error probability, δ(AND) = δ(NAND) = δ(OR) = δ(NOR)

V. OUTPUT ERROR PROBABILITY FOR XOR GATE

Truth table

A	B	Z
0	0	0
0	1	1
1	0	1
1	1	0

Error probability for input a, ε_a

Error probability for input b, ε_b

'00' -> 01 - $(1-\varepsilon_a)\varepsilon_b$

-> 10 - $\varepsilon_a(1-\varepsilon_b)$

'11' -> 01 - $\varepsilon_a(1-\varepsilon_b)$

-> 10 - $(1-\varepsilon_a)\varepsilon_b$

'01' -> 00 - $(1-\varepsilon_a)\varepsilon_b$

-> 11 - $\varepsilon_a(1-\varepsilon_b)$

'10' -> 00 - $\varepsilon_a(1-\varepsilon_b)$

-> 11 - $(1-\varepsilon_a)\varepsilon_b$

Here $|Y_{ON}| = |Y_{OFF}|$

Output error probability

$$\delta(\text{XOR2}) = \frac{4(1-\varepsilon_a)\varepsilon_b + 4\varepsilon_a(1-\varepsilon_b)}{4} - \varepsilon_a + \varepsilon_b - 2\varepsilon_a\varepsilon_b$$

simillarly

Output error probability,

$$\delta(\text{XOR3}) = \varepsilon_a[(1-\varepsilon_b)(1-\varepsilon_c) + \varepsilon_b\varepsilon_c] + (1-\varepsilon_a)[(1-\varepsilon_b)\varepsilon_c + \varepsilon_b(1-\varepsilon_c)]$$

Output error probability , δ(XOR) = δ(XNOR)

OUTPUT ERROR PROBABILITIES:

INVERTER	-
AND2	$\frac{\varepsilon_a + (1-\varepsilon_a)\varepsilon_b}{2}$
AND3	$\frac{\varepsilon_a + (1-\varepsilon_a)[\varepsilon_b + \varepsilon_c - \varepsilon_b\varepsilon_c]}{4}$
AND4	$\frac{\varepsilon_a + (1-\varepsilon_a)[\varepsilon_b + \varepsilon_c + \varepsilon_d - \varepsilon_b\varepsilon_c - \varepsilon_c\varepsilon_d - \varepsilon_d\varepsilon_b + \varepsilon_b\varepsilon_c\varepsilon_d]}{8}$
XOR2	$(1-\varepsilon_a)\varepsilon_b + \varepsilon_a(1-\varepsilon_b)$
XOR3	$\varepsilon_a[(1-\varepsilon_b)(1-\varepsilon_c) + \varepsilon_b\varepsilon_c] + (1-\varepsilon_a)[(1-\varepsilon_b)\varepsilon_c + \varepsilon_b(1-\varepsilon_c)]$
XOR4	$[(1-\varepsilon_a)\varepsilon_b + \varepsilon_a(1-\varepsilon_b)][(1-\varepsilon_c)(1-\varepsilon_d) + \varepsilon_c\varepsilon_d] + [(1-\varepsilon_c)\varepsilon_d + \varepsilon_c(1-\varepsilon_d)][(1-\varepsilon_a)(1-\varepsilon_b) + \varepsilon_a\varepsilon_b]$

consider the input error probability ε=0.2, then the output error probability for the logic gates are given below table1

Gate	Inputs		
	2	3	4
AND	0.18	0.122	0.0738
OR	0.18	0.122	0.0738
XOR	0.32	0.392	0.4352
XNOR	0.32	0.392	0.4352

Table1

V. REALIZATION METHOD

Discrete gate based implementation of the function has better error immunity than a complex gate based realization. The output error probability of various cell with input error probability ε=0.2 in discrete based realization and complex gate based realization is given in below table2.

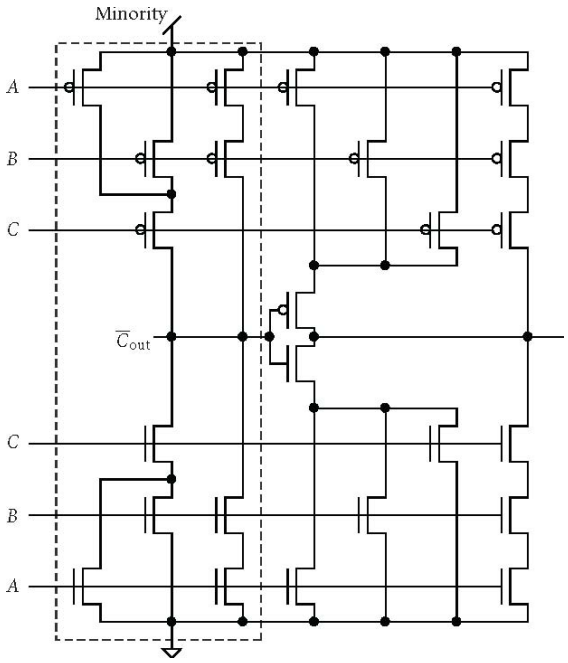
Output error probability for various cell (if $\varepsilon = 0.2$):

Cell	Function	Complex gate based realization	Discrete gate based realization
AO12	Z = ab + c	0.2220	0.1720
AO22	Z = ab + cd	0.2538	0.1638
AO31	Z = abc + d	0.2214	0.1488
AO112	Z = ab + c + d	0.1614	0.1188
AO32	Z = abc + de	0.2386	0.1400
AO212	Z = ab + cd +e	0.2152	0.1155
AO311	Z = abc + d + e	0.1770	0.1095
AO222	Z = ab + cd + ef	0.2667	0.1122
AO33	Z = abc + def	0.2061	0.1146
AO312	Z = abc + de + f	0.2263	0.1060

Case Study: The Full-Adder

The full-adder is a primary building block in digital arith-metic computation, present in nearly all digital adders and multipliers, and is one of the most well-studied circuits with a large variety of implementations. Its diverse array of circuit implementation and relative importance in creating low power datapath units make it a prime candidate for study.

Full-AdderMicroarchitecture. In a CMOS full-adder is presented that is optimized for transistor count, which is known as the 28-transistor implementation or "mirror adder"; it will be referred to as "f28". In the



authors claim the f28 is not only faster but consumes less power than the low-power carry-pass logic (CPL) implementation. The transistor level implementation of the "f28" full-adder is shown in Figure 8 It should also be noted that a majority gate is nothing more than a minority gate with an inverted output. A majority gate can also be achieved by inverting the inputs to the minority gate, which is useful should the inverted inputs be needed for some other part of the circuit.

VI. CONCLUSION

In this paper, an idea of that all gates do not propagate errors equally and that the depth of a digital circuit has a direct effect on the efficiency in terms of energy consumed for a given error rate. And also the Discrete gate based implementation of the function has better error immunity than a complex gate based realization.

REFERENCES

[1] J. George, B. Marr, B. E. S. Akgul, and K. V. Palem,"Probabilistic arithmetic and energy efficient embedded signalprocessing," in Proceedings of the International Conference onCompilers, Architecture and Synthesis for

Embedded Systems(CASES ’06), pp. 158–168, Seoul, Korea, 2006.

[2] T. Sakurai, “Perspectives on power-aware electronics,” in Proceedings of the IEEE International Solid-State Circuits Conference (ISSCC ’03), pp. 19–29, San Francisco, Calif, USA,February 2003.

[3] R. Hegde and N. R. Shanbhag, “Energy-efficient signal processing via algorithmic noise-tolerance,” in Proceedings of the International Symposium on Low Power Electronics and Design, pp. 30–35, San Diego, Calif, USA, August 1999.

[4] J. P. Hayes, I. Polian, and B. Becker, “An analysis framework for transient-error tolerance,” in Proceedings of the IEEE VLSI Test Symposium (VTS ’07), pp. 249–255, Berkeley, Calif, USA,May 2007.

◈◈◈

Reconfiguration Of Wireless Mesh Network

Sravani Achanta &.V Rama Krishna

Vignan University, vadlamudi, Guntur
E-mail : asravani99@gmail.com & bobysajja@gmail.com

Abstract - Wireless mesh networks (WMNs) have emerged as a key technology for next-generation wireless networking. Because of their advantages over other wireless networks, WMNs are undergoing rapid progress and inspiring numerous applications. In multi-hop wireless mesh networks (WMNs) experience frequent link failures caused by channel interference, dynamic obstacles and/or applications' bandwidth demands. These failures cause severe performance degradation in WMNs or require expensive, manual network management for their real-time recovery. This paper presents an Autonomous network Reconfiguration System (ARS) that enables a multi-radio WMN to autonomously recover from local link failures to preserve network performance. ARS also improves channel efficiency by more than 90% over the other recovery methods.

Keywords: *multi-radio WMNs, wireless link failures, reconfiguration network.*

I. INTRODUCTION

Wireless networks provide unprecedented freedom and mobility for a growing number of laptop and PDA users who no longer need wires to stay connected with their workplace and the Internet. Ironically, the very devices that provide wireless service to these clients need lots of wiring themselves to connect to private networks and the Internet. This white paper presents a viable alternative to all those wires - the wireless mesh network. Unlike basic Wi-Fi that simply untethers the client; the wireless mesh untethers the network itself giving IT departments, network architects and systems integrators unprecedented freedom and flexibility to build out networks in record time - with high performance and without the expensive cabling. **W**ireless mesh networks (WMNs) are being developed actively and deployed widely for a variety of applications, such as public safety environment monitoring, and citywide wireless Internet services. They have also been evolving in various forms to meet the increasing capacity demands by the above-mentioned and other emerging applications. However, due to heterogeneous and fluctuating wireless link conditions, preserving the required performance of such WMNs is still a challenging problem. For example, some links of a WMN may experience significant channel interference from other coexisting wireless networks. Some parts of networks might not be able to meet increasing bandwidth demands from new mobile users and applications. Links in a certain area (e.g., a hospital or police station) might not be able to use some frequency channels because of spectrum etiquette or regulation.

Related work: A **wireless mesh network** (**WMN**) is a communications network made up of radio nodes organized in a mesh topology. Wireless mesh networks often consist of mesh clients, mesh routers and gateways. The mesh clients are often laptops, cell phones and other wireless devices while the mesh routers forward traffic to and from

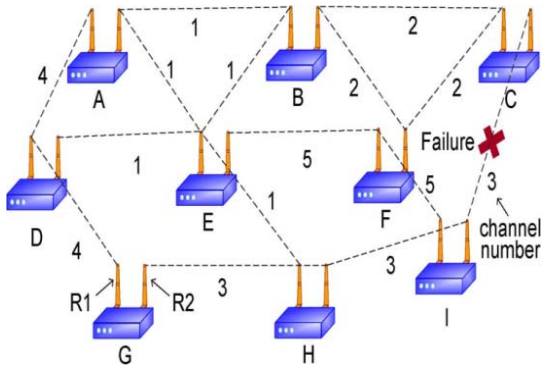


Fig: Multiradio WMN:A WMN has an initial assignment of frequency chan- nels as shown. The network often experiences wireless link failure and needs to reconfigure its settings

Configuration, and may be deployed to provide dynamic and cost effective connectivity over a certain geographic area. An ad-hoc network, on the other hand,

is formed ad hoc when wireless devices come within communication range of each other. The mesh routers may be mobile, and be moved according to specific demand as rising in the network. Wireless mesh networks can be implemented with various wireless technology including 802.11, 802.15, 802.16, cellular technologies or combinations of more than one type.

Types of Wireless Networks: Wireless networks or divided into two categories.

Infrastructure Networks: Infrastructure network have fixed network topology. Wireless nodes connect through the fixed point known as base station or access point. All of the wireless connections must pass from the base station. Whenever a node is in the range of several base stations then it connect to any one of them on the bases of some criteria.

Ad hoc Networks: Ad hoc networks also called infrastructure less networks are complex distributed the gateways which may but need not connect to the Internet. A wireless mesh network can be seen as a special type of wireless ad-hoc network. A wireless mesh network often has a more planned

TABLE I
Definition of link-change in ARS. EACH CHANGE REPRESENTS A PRIMITIVE LINK CHANGE IN CHANNEL, ASSOCIATION, OR ROUTE. MULTIPLE CHANGES CAN BE JOINTLY USED TO REPRESENT CHANGES OF MULTIPLE LINKS.

Primitive changes	Description
Channel switch $(S(A_i, B_j)_{\alpha\to\beta})$	Radios A_i and B_j of link AB switch their channel (α) to other channel (β).
Radio switch $(R(A_i, B_j)_{\alpha\to\beta})$	Radio A_i in node A re-associates with radio B_j in node B, tuned in channel (β).
Detouring $(D(A_i, B_j))$	Both radios A_i and B_j of link AB remove their associations and use a detour path, if exists.

systems consist of wireless links between the nodes and each node also works as a router to forwards the data on behalf of other nodes.

PROPOSED SYSTEM: Reconfiguration of wireless mesh network presented an autonomous network reconfiguration system (ARS) that enables a multiradio WMN to autonomously recover from wireless link failures. ARS generates an effective reconfiguration plan that requires only local network configuration changes by exploiting channel, radio, and path diversity. Furthermore, ARS effectively identifies reconfiguration plans that satisfy applications' QoS constraints, admitting up to two times more flows than static assignment, through QoS aware planning. Next, ARS's online reconfigurability allows for real-time failure detection and network reconfiguration, thus improving channel efficiency by 92%. Our experimental evaluations on a Linux-based implementation and ns2-based simulation have demonstrated the effectiveness of ARS in recovering from local link-failures and in satisfying applications' diverse QoS demands.

ANALYSIS: FEASIBILITY ANALYSIS: All projects are feasible, given unlimited resources and infinite time. Before going further in to the steps of software development, the system analyst has to analyze whether the proposed system will be feasible for the organization and must identify the feasibility studies

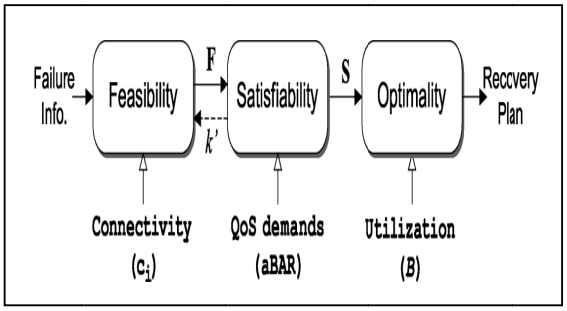


Fig: Localized reconfiguration planning in ARS. ARS generates a reconfiguration plan by breaking down the planning process into three processes with Different constraints

have to be done on any system. But there are three main feasibility tests to be performed. **OPERATIONAL FEASIBILITY**: During feasibility analysis operational feasibility study is a must. This is because; according to software engineering principles operational feasibility on the other words usability should be very high. A thorough analysis is done and found that the system is operational.

TECHNICAL FEASIBILITY: The system analyses to check the technical feasibility of the proposed system. Taking account of the hardware it is used for the system development, data storage, processing and output, makes the technical feasibility assessment. The system analyst has to check whether the company or user who is implementing the system has enough resource available for the smooth running of the application. **ECONOMICAL FEASIBILITY**: Before going further in to the development of the proposed system. The system analyst has to check the economic feasibility of the proposed system and the cost for running the system is composed with the cost benefit that can achieve by implementing the system customer needs. The main purpose of feasibility study is to determine whether the problem is worth solving. The success of a system is also lies in the amount of feasibility study done on it. Many feasibility assessment. system analyst

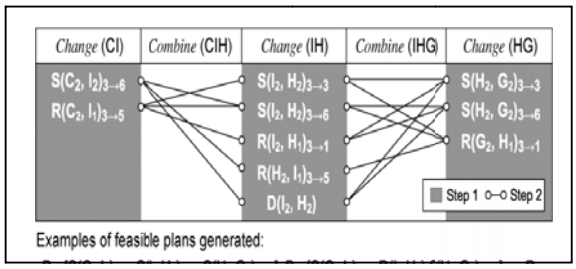


Fig: Example of network planning. ARS generates Per-link changes (gray columns) and then connects Them for feasible reconfiguration plans,

feasibility studies have to be done on any system. But there are three main feasibility tests to be performed. **OPERATIONAL FEASIBILITY**: During feasibility analysis operational feasibility study is a must. This is because; according to software engineering principles operational feasibility on the other words usability should be very high. A thorough analysis is done and found that the system is operational.

TECHNICAL FEASIBILITY: The system analyses to check the technical feasibility of the proposed system. Taking account of the hardware it is used for the system development, data storage, processing and output, makes the technical The has to check whether the company or user who is implementing the system has enough resource available for the smooth running of the application. **ECONOMICAL FEASIBILITY**: Before going further in to the development of the proposed system. The system analyst has to check the economic feasibility of the proposed system and the cost for running the system is composed with the cost benefit that can achieve by implementing the system.

Network Model and Assumptions Multi-radio WMN: A network is assumed to consist of mesh nodes, IEEE 802.11-based wireless links, and one control gateway. Each mesh node is equipped with n radios, and each radio's channel and link assignments are initially made by using global channel/link assignment algorithms multiple orthogonal channels are assumed available.

QoS support: During its operation, each mesh node periodically sends its local channel usage and the quality information for all outgoing links via management messages to the control gateway. **Link failures**: Channel-related link failures that we focus on, are due mainly to narrow-band channel failures. These failures are assumed to occur and last in the order of a few minutes to hours, and reconfiguration is triggered in the same order of failure occurrences. For short-term (lasting form milliseconds) failures, fine-grained (e.g., packet-level or in the order of milliseconds) dynamic resource allocation might be sufficient and for a long-term (lasting for weeks or months) failures, network-wide planning algorithms can be used**.**

Challenges and Requirements of Reconfiguration of Wireless Mesh Networks: The purpose of reconfiguration is in some cases Muliradio wireless mesh networks local link has to be failed. These failures are assumed to occur and last in the order of a few minutes to hours, and reconfiguration is triggered in the same order of failure occurrences. ARS architecture using the algorithm is ARS operation at mesh nodes

Localized reconfiguration: Based on multiple channels and radio associations vailable, ARS generates reconfiguration plans that allow for changes of network configurations only in the vicinity where link failures occurred while retaining configurations in areas remote from failure locations.

Planning for Localized Network Reconfiguration: The core function of ARS is to *systematically* generate localized reconfiguration plans. Reconfiguration plan is defined as a set of links' configuration necessary for a network to recover from a link(s) failure on a channel, and there are usually multiple reconfiguration plans for each link failure. Existing channel-assignment and scheduling algorithms seek "optimal" solutions by considering tight QoS constraints on all links, thus requiring a large configuration space to be searched and hence making the planning often an NP-complete problem . In addition, change in a link's requirement may lead to completely different network configurations. ARS first applies connectivity constraints to generate a *set* of feasible reconfiguration plans that enumerate feasible channel, link, and route changes around the faulty areas, given connectivity and link-failure constraints.

QoS-aware planning: ARS effectively identifies QoS-satisfiable reconfiguration plans by estimating the QoS satisfiability of generated reconfiguration plans and deriving their expected benefits in channel utilization.

Autonomous reconfiguration via link-quality monitoring:

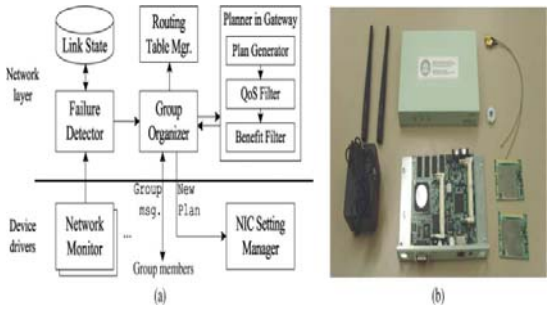
ARS accurately monitors the quality4 of links of each node in a distributed manner. Furthermore, based on the measurements and given links' QoS constraints, ARS detects local link failures and autonomously initiates network reconfiguration. **Feasible Plan Generation:** Generating feasible plans is essentially to search all legitimate changes in links' configurations and their combinations around the faulty area. Given multiple radios, channels, and routes, ARS identifies feasible changes that help avoid a local link failure but

maintain existing network connectivity as much as possible.

Maintaining network connectivity and utilization: While avoiding the use of the faulty channel, ARS needs to maintain connectivity with the full utilization of radio resources. Because each radio can associate itself with multiple neighboring nodes, a change in one link triggers other neighboring links to change their settings. **Controlling the scope of reconfiguration change**s: ARS has to limit network changes as *local* as possible, but at the same time it needs to find a locally optimal solution by considering more network changes or scope. To make this trade off, ARS uses a -hop reconfiguration parameter. Starting from a faulty link(s), ARS considers link changes within the first hops and generates feasible plans.

Per-link bandwidth estimation: ARS has to check whether each link's configuration change satisfies its bandwidth requirement, so it must estimate link bandwidth. To estimate link bandwidth, ARS accurately measures each link's capacity and its available channel time Even though numerous bandwidth-estimation techniques have been proposed, they focus on the average bandwidth of each node in a network or the end-to-end throughput to flows.

Examining per-link bandwidth satisfiability: Given measured bandwidth and bandwidth



link to check the link's QoS satisfiability. If multiple links share the airtime of one channel, ARS calculates aggregate BAR of end-radios of a link, which is defined as , where is a radio ID, a link associated with radio , and the set of directed links within and across radio 's transmission range. **Breaking a tie among multiple plans:** Multiple reconfiguration plans can have the same benefit, and ARS needs to break a tie among them. ARS uses the number of link changes that each plan requires to break a tie. Let us consider an example in Fig. 5. Assuming BAR of each directed link is 0.2 (e.g., 2Mbps10Mbps) in a tuned channel, *a BAR* of each radio tuned to channel does not exceed 1.0, satisfying each link's QoS requirement.

Cross-layer interaction: ARS actively interacts across the network and link layers for planning. This interaction enables ARS to include a rerouting for reconfiguration planning in addition to link-layer reconfiguration. ARS can also maintain connectivity during recovery period with the help of a routing protocol. Algorithm describes the operation of ARS. First, ARS in every mesh node monitors the quality of its outgoing wireless links at every (e.g., 10 s) and reports the results to a gateway via a management message. Second, once it detects a link failure(s), ARS in the detector node(s) triggers the formation of a group among local mesh routers that use a faulty channel, and one of the group members is elected as a leader using the well-known bully algorithm for coordinating the reconfiguration. Third, the leader node sends a planning-request message to a gateway. Then, the gateway synchronizes the planning requests—if there are multiple requests—and generates a reconfiguration plan for the request. Fourth, the gateway sends a reconfiguration plan to the leader node and the group members. Finally, all nodes in the group execute the corresponding configuration changes, if any, and resolve the group. We assume that during the formation and reconfiguration, all messages are reliably delivered via a routing protocol and per-hop retransmission timer. **Benefit function***:* We also studied the impact of the benefit function on the ARS's planning algorithm. We conducted the same experiment as the previous one with different values of δ in the benefit function. A high value(0.8) of δ allows ARS to keep local channel-efficiency high. By contrast, a low value (0.4) can deliver more available bandwidth (on average, 1.2 Mbps) than when the high value is used, since ARS tries to reserve more capacity. **Complexity of ARS**: ARS incurs reasonable bandwidth and computation overheads. First, the network monitoring part in the reconfiguration protocols is made highly efficient by exploiting existing data trafficandconsumes lessthan12kb/sprobing b and width Next, the computationall overhead in ARS mainly systems from the planning algorithms. Specifically, generating its possible link plans incurs complexity, where the number of available channels and the number of radios . Next, a gateway node needs to generate and evaluate feasible plans which incurs search over head in a constraint graph that consists of nodes, where the number of links that use faulty channel in the group.

The reconfiguration range*:*

We evaluated the impact of the reconfiguration range. We used the same experiment settings as the previous one and focused on reconfiguration requests at

T1. As we increase the hop count (k) from a faulty link(s), we measure the capacity improvement achieved by the reconfiguration plans. This saturation results mainly from the fixed number of radios of each node. In other words the improvement is essentially bounded by the total capacity of physical radios.

II. CONCLUSION

An autonomous network reconfiguration system (ARS) that enables a multi-radio WMN to autonomously recover from wireless link failures. ARS generates an Effective reconfiguration plan that requires only local network configuration changes by exploiting channel radio, and path diversity., ARS effectively identifies reconfiguration plans that satisfy applications' QoS constraints, admitting up to two times more flows than static assignment, through QoS aware planning. Next, ARS's online reconfigurability allows for real-time failure detection and network reconfiguration. Based on existing MAC, routing, and transport protocols, network performance is not scalable with either the number of nodes or the number of hops in the network. This problem can be alleviated by increasing the network capacity through using multiple channels/radios per node or developing wireless radios with higher transmission speed. However, these approaches do not truly enhance the scalability of WMNs, because resource utilization is not actually improved. Therefore, in order to achieve scalability, it is essential to develop new MAC, routing, and transport protocols for WMNs.

REFERENCES:

[1] Capacity of Multi-Channel Wireless Networks: Impact of Number of Channels and Interfaces. Pradeep Kyasanur, Nitin H. Vaidya.

[2] Routing in Multi-Radio, Multi-Hop Wireless Mesh Networks Richard Draves Jitendra Padhye Brian Zill.

[3] **A Survey on Wireless Mesh Networks** IAN F. AKYILDIZ, GEORGIA INSTITUTE OF TECHNOLOGYXUDONG WANG, KIYON, INC.

[4] Performance Analysis of the IEEE 802.11 Distributed Coordination Function Giuseppe Bianchi.

[5] Distributed Quality-of-Service Routing in Ad Hoc Networks Shigang Chen and Klara Nahrstedt, Member, IEEE

[6] Minimum Interference Channel Assignment in Multiradio Wireless Mesh Networks An and Prabhu Subramanian, Student Member, IEEE, Himanshu Gupta,Samir R. Das, Member, IEEE, and Jing Cao, Member, IEEE.

[7] On Accurate and Asymmetry-Aware Measurement of Link Quality in Wireless Mesh Networks Kyu-Han Kim, Student Member, IEEE, and Kang G. Shin, Fellow, IEEE

[8] Simulation of Enhanced Meshes with MASC, a MSIMD Model Johnnie W. Baker and Mingxian Jin Department of Math and Computer Science.

[9] Protocol-Independent Reconfigurable Optical Transport Networks by Leo Goyette.

[10] Self reconfigurable wireless mesh networks Kyu-Han Kim, *Member, IEEE* and Kan G G. Shin, *Fellow, IEEE*

◈◈◈

3D Depth Estimation from a Single Outdoor Image

Anju John & P. S. Sathidevi

Dept. of Electronics and Communication Engineering, National Institute of Technology, Calicut, Kerala, India-673601
E-mail: anjujohnm@gmail.com & sathi@nitc.ac.in

Abstract - Recovering depth information from images has got potential applications in computer vision. The aim of this work is to obtain 3D depth from a single still image of outdoor environments. We employ supervised learning to predict the value of depth as a function of various local and global image features and the dark channel prior information. An initial depth map is estimated from multiscale local and global image features extracted using several filter masks. This is further improved by incorporating the dark channel information and saturation of the image. Dark channel prior is a statistic derived from the fact that the minimum intensity in any colour channel in a local patch in an outdoor image is related to the haze, which in turn is a depth cue. Experimental results indicate that the proposed method is able to retrieve fairly accurate depth maps with low computational complexity compared to existing methods.

Keywords- *dark channel prior, depth estimation, markov random field, regression.*

I. INTRODUCTION

The 3D depth of an image can be defined as the distance between the corresponding scene point of the image and the camera. The three dimensional world around us is projected onto a 2D plane while imaging, causing loss of depth information. Depth perception from images is an active area of research in computer vision and can aid in object recognition, scene understanding and 3D reconstruction. Conventional methods of depth estimation employ multiple views of the same scene to estimate depth. Stereo vision techniques [1] capture two images of the same scene with two cameras separated by a baseline distance, calculate the disparity between the features in the two images and employ triangulation methods to obtain depth from known camera geometry. Depth from defocus [2] and structure from motion [3] are some other methods of depth perception from multiple view images. These algorithms consider only the geometric differences between the two images and not the global structure of the image.

Humans perceive depth of a scene with the help of both monocular and binocular cues. Depth estimation from single images is done with the help of several monocular cues like texture variations, texture gradients, known object sizes, color, haze and defocus. For example, parallel railway tracks tend to converge with increasing distance in an image. Thus the orientation of the edges is a useful monocular depth cue. Another useful observation is that distant objects appear to be hazier than nearby ones in outdoor images. We can safely conclude that many monocular cues for depth perception are contextual information, i.e., they depend on the global properties of the image.

Recently, researchers have attempted to solve for depth maps from single images by posing this as a machine learning problem. Saxena *et al.* [4] extracted local and global image features like texture variations, texture gradients and color at multiple spatial scales using appropriate filter masks. A hierarchical multiscale Markov Random Field (MRF) is then used to model depths and the relationship between depths at multiple spatial scales. After learning the parameters of the model with the help of the training set, Maximum *A Posteriori* (MAP) estimation is done to find the depth of the test image. This approach is quite successful in estimating depths of unstructured indoor and outdoor environments, but is computationally intensive.

Another related work is that of Beyang Liu *et al.* [5] where first a semantic labeling of the image is performed. The semantic classifier is trained to segment the scene so that each segment belongs to one among the following classes: sky, tree, road, grass, water, building, mountain, and foreground object. Then a separate depth estimator is learnt for each class using an MRF model.

Tien-Ying Kuo *et al.* [6] consider the problem of depth estimation from monocular outdoor images by combining the initial depth prior, the dark channel prior and multi-resolution analysis. The image is first

segmented based on edges and color similarity and an initial depth is assigned to each object based on its location in the image. Now, saturation and dark channel information are used to refine the initial depth. Finally, the depth maps obtained at different resolutions are fused together to get the final depth map.

This study draws inspiration from the works of Saxena *et al.* [4] and Tien-Ying Kuo *et al.* [6] and tries to combine these two concepts in a simple but effective manner for estimating depth maps of outdoor images. The organization of the rest of the paper is as follows. Section 2 gives an overview of MRF based depth estimation. Section 3 describes the proposed algorithm. Results and discussion are given in Section 4. Finally, Section 5 concludes the paper.

II. MRF BASED DEPTH ESTIMATION

Markov Random Field (MRF) theory is a branch of probability theory for analyzing spatial or contextual dependencies of physical phenomena. MRFs are characterized by their joint probability distribution. Here we model the relationship between the depth of an image patch and its neighboring patches using a MRF that incorporates local and global image features at multiple spatial scales [4]. First, the image is divided into uniform rectangular patches and absolute features and relative features are computed for each patch. We have to estimate a single depth value for each patch. Absolute depth features are described in detail in Section 3*A*. Relative features allow us to learn the dependencies between two neighboring patches. We compute a 10-bin histogram for all 17 filter outputs (see Section 3) leading to a 170 dimensional feature vector y_i for each image patch i. The relative depth between two patches is assumed to be related to the difference between the respective relative feature vectors.

The posterior distribution of depth at each patch is modeled by a Gaussian MRF. We have the distribution of depth d, given the image features X and parameterized by θ and σ as

$$P(d / X;\theta,\sigma) = \frac{1}{Z}\exp(-E_{\sigma,\theta}(d,X)) \qquad (1)$$

where

$$E_{\sigma,\theta}(d,X) = \sum_{i=1}^{M}\frac{(d_i - x_i^T\theta_r)^2}{2\sigma_{1r}^2} - \sum_{i=1}^{M}\sum_{j\in N_s(i)}\frac{(d_i - d_j)^2}{2\sigma_{2r}^2} \qquad (2)$$

Here, Z is a normalization constant, $E_{\sigma,\theta}(d,X)$ is a Gibbs energy function, x_i is the absolute feature vector at patch i, M is the total number of patches in the image and $N_s(i)$ are the 4 neighbors of patch i at scale s. Different rows have different parameters.

The maximum likelihood estimate of θ_r is obtained by maximizing the conditional log likelihood function over the training images. It is sufficient to solve a linear least squares problem for this purpose. The variance parameter in the first term of the exponent is modeled as $\sigma_{1r}^2 = u_r^T x_i$. The estimated value of u_r is such that σ_{1r}^2 fits to the expected value of $(d_i - x_i^T\theta_r)^2$ subject to $u_r \geq 0$ (non-negative least squares). The non-negativity constraint ensures that the estimated value of σ_{1r}^2 is non-negative. Similarly, we model the variance term σ_{2r}^2 as a linear function of the relative depth features. We put $\sigma_{2r}^2 = v_r^T |y_{ij}|$ where y_{ij} is the relative depth feature for two neighboring patches i and j given by the difference in their histograms, i.e., $y_{ij} = y_i - y_j$. Again, the parameter v_r is learnt such that σ_{2r}^2 fits to the expected value of $(d_i - d_j)^2$ subject to the condition $v_r \geq 0$.

We can express (1) in the form of a standard multivariate Gaussian distribution as:

$$P(d / X;\theta,\sigma) = \frac{1}{Z}\exp\left(-\frac{1}{2}(d - X_a\theta_r)^T\Sigma_a^{-1}(d - X_a\theta_r)\right) \qquad (3)$$

where $X_a = (\Sigma_1^{-1} + Q^T\Sigma_2^{-1}Q)^{-1}\Sigma_1^{-1}X$. Here Σ_1 and Σ_2 represent the matrices of the variances $\sigma_{1,i}^2$ and $\sigma_{2,i}^2$ in (1) respectively. Matrix Q is such that the rows of Qd give the differences of the depths in neighboring patches. From (3), we have the value of d maximizing the posterior probability as $d^* = X_a\theta_r$. Based on this, we iteratively learn the parameters θ and σ during learning, until the algorithm converges (2-3 iterations).

After learning the parameters, we are able to estimate the depth map of a new test image. This is done by maximizing the posterior distribution $\log P(d / X;\theta,\sigma)$ in terms of d. The MAP estimate is given by $d^* = X_a\theta_r$.

III. PROPOSED ALGORITHM

Color, texture gradients and texture variations are image features representative of depth in an image. We capture this information at multiple spatial scales by filtering the input images with suitable convolutional masks. Linear regression is employed to learn the dependence of depth on the image features so obtained. The initial depth map so created is further processed to take into account the relation of depth with dark channel prior and saturation. Thus, the final depth map is obtained after a two stage learning process.

A. Feature Vector

We divide the image into small rectangular patches and estimate a single depth value for each patch. The image is represented in YCbCr colour space, where Y is the intensity channel and Cb and Cr are the color channels. We apply the nine Laws' masks [7] to the intensity channel to obtain the texture energy. Haze is reflected in the low frequency information in the color channels, so we apply a local averaging filter (the first Laws' mask) to both color channels. We convolve the intensity channel with six oriented edge filters (Nevatia-Babu filters [8] at angles 0^0, 30^0, 60^0, 90^0, 120^0 and 150^0) to get the texture gradients. The filters used are shown in Figure 1.

Figure- 1 The convolutional filters used for texture energies and gradients.

We compute summary statistics of a patch i in the image $I(x, y)$. For all the 17 filter outputs (9 Laws' masks, 2 color channels and 6 texture gradients) we calculate the sum absolute energy and sum squared energy as $E_i(n) = \sum_{(x,y)\in patch(i)} |I * F_n|^k$, where F_n (n=1,2..17) are the filters and k=1,2. Thus we get an initial feature vector of dimension 34. In order to capture more contextual information, we append the initial feature vector of a patch with the features of its four neighboring patches too. We also know that objects at different depths exhibit different behaviors at different resolutions which can be captured using multiscale features. So we compute the features for a patch at a lower image resolution also (scale 1/3X). Thus the final feature vector for a patch includes features of that particular patch and its neighbors at two different spatial scales. Thus the size of the final feature vector is 34*10=340.

B. Initial Depth Assignment

The depth d_i at the ith patch in the image is assumed to be a linear function of the features. We can write

$$d_i = x_i^T \theta_r \tag{4}$$

where x_i is the feature vector for patch i and θ is the parameter of the model. We use different parameters (θ_r) for each row r in the image, because the images under consideration are taken from a horizontally mounted camera, and thus different rows of the image have different statistical properties. Linear regression is employed to find the parameters θ_r for each row. If there are N training images with M patches in each row, the parameters for the row r is obtained from

$$\theta_r = \arg\min_{\theta_r} \frac{1}{2} \sum_{n=1}^{N} \sum_{i=1}^{M} \left(d_{i,n} - x_{i,n}^T \theta_r \right)^2 \tag{5}$$

The depths are transformed to a log scale so as to emphasize multiplicative rather than additive errors in training. This is because humans can more easily distinguish small differences in depth at small distances than at larger distances.

C. Final depth map generation

The initial depth map so created has high frequency variations not present in the image. The object boundaries are also not clearly distinguished. In order to make the depth estimate closer to the ground truth, the system is also made to learn the dependence of depth on dark channel prior and saturation of the image.

Most local patches in haze-free outdoor images contain some pixels which have very low intensities in at least one color channel (dark pixels). This statistical observation is called dark channel prior [9]. Dark Channel of an image **J** is defined as:

$$J^{dark}(x) = \min_{c\in\{r,g,b\}} \left(\min_{y\in\Omega(x)} (J^c(y)) \right) \tag{6}$$

where J^c is a color channel of **J** and $\Omega(\mathbf{x})$ is a local patch centered at position **x**. At least a small amount of haze will be generally present in natural scenes. Distant objects tend to reflect more atmospheric light and hence appear to have a higher intensity in natural outdoor images. Thus the intensity of dark channel image is a qualitative descriptor of the depth of the scene. Fig. 2 shows an outdoor image and its corresponding dark channel image.

The value of dark pixels may not be indicative of depth in areas of low color saturation [6]. In Fig. 2, we see that the trees in the background have lower intensity in the dark channel image leading us to wrongly interpret it to be at a smaller distance. In order to rectify this, we also take into consideration the color saturation in a local patch while estimating the depth. The value of saturation can be directly obtained by representing the image in HSV color space.

Figure 2 : Image and its dark channel

The proposed method combines the initial depth, saturation and dark channel to get the depth of patch *i* as given below:

$$depth_i = w_1 depth_{init,i} J_{dark,i} + w_2 J_{dark,i} S_i^{(1-sigmoid(y))} + w_3 depth_{init,i} (1-S_i)^{(1-sigmoid(y))} \quad (7)$$

Here, $depth_{init,i}$ is the initial depth and $depth_i$ is the final depth. w_1, w_2 and w_3 are the weighting parameters, $J_{dark,i}$ is the dark channel value, S_i is the saturation and y is the normalized vertical coordinate (top mapped to 1 and bottom mapped to 0). The first term, $depth_{init,i}\, J_{dark,i}$ fuses the dark channel map with the initial depth map. The second term, $J_{dark,i}\, S_i^{(1-sigmoid(y))}$ suppresses the area where the dark channel prior may fail owing to low saturations. The suppressed region in the second term is compensated by the last term, $depth_{init,i}\,(1-S_i)^{(1-sigmoid(y))}$. The coefficients w_1, w_2 and w_3 are learnt through robust regression so that the performance is not degraded by outliers. The depth map so obtained is median filtered for smoothening.

IV. RESULTS AND DISCUSSION

The test inputs in the study belong to Stanford University's Make3D database [4]. This image database contains 300 outdoor images with a resolution of 1704 x 2272, and 300 corresponding ground truth depth images of resolution 86 x 107 acquired by a laser distance scanner with gray values ranging from 1-81. A subset of this database was used for training and testing the algorithm. 70 percent of the images were used for training and the rest for holdout testing. The algorithm was implemented in Matlab.

Fig. 3 shows the test results on three sample images from the dataset. The results are visually impressive. The algorithm is also tested on images downloaded from the internet (ground truth depth is unknown) to test its

Figure-3 - (a) Original image (b) Ground truth depth (c) Initial depth (d) Depth by proposed method

Since the maximum range of the sensor used to collect the ground truth measurements is 81 m, the predicted depth values are truncated to the range [0, 81]. The performance of the algorithm is also evaluated using mean absolute difference on a normalised log scale as an error metric.

Fig. 4 shows the average errors (on a normalised log scale) as a function of the distance from the camera. We see that the error in our algorithm is minimum at either end of the depth map.

TABLE I. MEAN ABSOLUTE ERRORS (NORMALISED LOG SCALE)

Linear regression	0.0899
MRF based method [4]	0.0859
Proposed method	0.0817

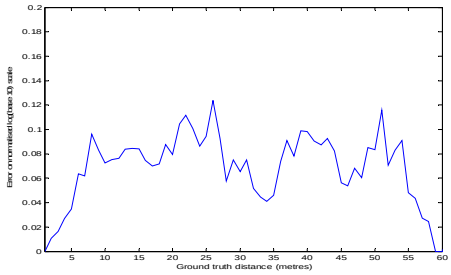


Figure- 4 - The average errors as a function of the distance from the camera

V. CONCLUSION

This work builds upon the existing methods of learning based depth estimation techniques to propose a novel way of 3D depth generation in monocular outdoor images. The image features extracted by convolutional features along with dark channel and colour saturation information are used to predict the depth from a single outdoor image. The algorithm was successfully implemented on a wide variety of natural images.

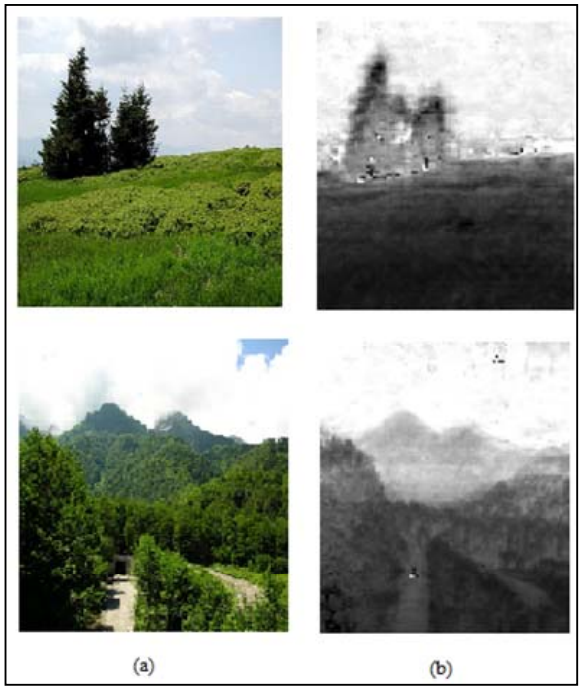


Figure- 5 : (a) Original image (b) Depth by proposed method

REFERENCES

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation o f dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, 2002.

[2] S. Das and N. Ahuja, "Performance analysis of stereo, vergence, and focus as depth cues for active vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995.

[3] D. A. Forsyth and J. Ponce, *Computer vision: a modern approach*, New York: Prentice Hall, 2003.

[4] A. Saxena, S. H. Chung, and A. Y. Ng, "3-D Depth Reconstruction from a Single Still Image," *International Journal of Computer Vision*, 2007.

[5] Beyang Liu, Stephen Gould and Daphne Koller, "Single Image Depth Estimation from Predicted Semantic Labels," *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[6] T. Y. Kuo and Y. C. Lo, "Depth Estimation from a Monocular view of the Outdoors," *IEEE Transactions on Consumer Electronics*, 2011.

[7] K. I. Laws, "Texture energy measures," *DARPA Image Understanding Workshop*, *DARPA*, 1979.

[8] R. Nevatia and K. Babu, "Linear feature extraction and Detection," *Computer Vision, Graphics and Image Processing*, 1980.

[9] K. He, J. Sun, and X. Tang, "Single Image Haze Removal Using Dark Channel Prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

◈◈◈

High Performance Generalized Punctured Convolutional Codes

Arun A V & Kashwan K. R.

Deptt. of Electronics and Communication Engineering - PG
Sona College of Tech. (Autonomous) Salem -636005, India
E-mail: arunav.aav@gmail.com & drkrkashwan@gmail.com

Abstract - Generalized Punctured Convolutional Codes (GPCC) are superior and broader in comparison with standard Punctured Convolutional Codes (PCC). The GPCC have better performance than PCC due to reduced trellis complexity. A code in this class can be represented by a module called GPCC trellis module. The conventional representation of GPCC trellis may not have minimal trellis property which means these codes can be further minimized. This research project aims at constructing a minimal GPCC trellis module which is isomorphic to minimal trellis module. The proposed minimal GPCC trellis module avoids fixed or hard trellis complexity leading to flexibility according to the applications. It supports the tradeoff between trellis complexity and performance as per priority requirements. We could reduce trellis complexity by a factor ranging from 9.30% to 10.93% for a code rate 2/3 depending upon chosen constraint length. Similarly for other rates also, there is significant reduction in trellis complexity.

Keywords-*Convolutional codes, trellis complexity, generator matrix, minimal trellis.*

I. INTRODUCTION

A convolutional encoder takes a stream of information bits and converts it into a stream of transmitted bits, using a shift register bank. Redundancy for recovering from channel errors is provided by transmitting more bits per unit time than the no of information bits per unit time. Maximum likelihood decoding can be done using Viterbi algorithm.

In this paper, the general notation for convolutional code C, is specified by three parameters; (n, k, m) where n is the number of output bits, k is the number of input bits and m will be the number of memory registers. The quantity k/n is called the code rate, is a measure of the efficiency of the code. The range selected here for n, k, m are $3<n<8$, $2<k<7$, $3<m<8$ respectively. The constraint length of the code is given by

$$L = k(m-1) \qquad (1)$$

The complexity of a Viterbi decoder increases rapidly with the code rate. Puncturing is a technique that allows the encoding and decoding of higher rate codes using standard ½ encoders and decoders. The trellis complexity of Punctured Convolutional Codes (PCC) [6]-[7] is significantly less than that of conventional convolutional codes. But PCC's has the limitation that the distance spectra of these codes are poor The research work reported in [1] defines a new set of convolutional codes that has low complexity and exhibits high performance in terms of distance spectra. The broader class of convolutional codes that has better distance spectrum than any of the PCC was defined in [2] and is called Generalized Punctured Convolutional Codes (GPCC). In this paper we are constructing a minimal trellis module for GPCC's. A canonical generator matrix G (D) for an (n, k, m) convolutional code C is a k x n matrix with polynomial entries, whose row space is C, such that the direct-form realization of an encoder for C based on G (D) uses exactly m delay elements. From a canonical generator matrix G (D), or rather from a physical encoder built using G (D) as a blueprint, it is possible to construct a conventional trellis representation for C. This trellis is in principle infinite, but it has a very regular structure, consisting (after a short initial transient) of repeated copies of what we shall call the trellis module associated with G (D). The trellis module consists of 2^m initial states and 2^m final states, with each initial state being connected by a directed edge to exactly 2^k final states. Thus the trellis module has 2^{k+m} edges. Each edge is labeled with an n-symbol binary vector, namely, the output produced by the encoder in response to the given state transition. Thus each edge has length (measured in coded symbols) n, and so the total edge length of the conventional trellis module is n^{2k+m}. Since each trellis module represents the encoder's response to *k* input bits, the conventional trellis complexity of the trellis module is defined as

$$\frac{n}{k}\frac{n}{k}2^{m+k} \; symbols\ per\ encoded\ bit \qquad (2)$$

The units of trellis complexity are also expressed as symbols per bit in short. If the code C is decoded using Viterbi's maximum-likelihood algorithm on the trellis the work factor involved in updating the metrics and survivors at each trellis module is proportional to the edge length of the trellis module, so that the trellis complexity as defined in (2) is a measure of the effort per decoded bit required by Viterbi's algorithm.

This paper is organized as follows. In section II we construct the generator matrix and in section III a method to find the trellis complexity is introduced. Section IV discusses about the minimal trellis module construction. Section V gives an account of the code search results.

II. GENERATOR POLYNOMIALS

The encoder connections are characterized by generator polynomials for describing the feedback shift register implementation of convolutional codes. There are many choices for polynomials for any m order code. They do not all result in output sequences that have good error protection properties. The generator polynomials are usually arranged in a matrix and are known as generator matrix. The inputs multiplied with the generator matrix gives the encoder of a convolutional code. The order of the generator matrix when applied in GF (2) is $k \times (m+1)\, n$. The generator matrix G is given below.

$$G = \begin{bmatrix} G_1(1)\ G_1(2).....G_1((m+1)n) \\ . \quad . \\ . \quad . \\ . \quad . \\ . \quad . \\ G_k(1).....G_k((m+1)n) \end{bmatrix} \quad (3$$

The first row shows the dependence of first input to the feedback shift register and k^{th} row indicates the dependence of k^{th} input to the shift registers used. The first column indicates the dependence of the first output to the m shift registers used. To represent an individual connection vector we need m+1 element with respect to a single output.

For a m order code 2^{m+1} values are available and they should be arranged in all possible combinations. As mentioned earlier all possible combinations of values do not generate an output. Two propositions should be followed for the generation of valid polynomials.

Proposition 1: All the columns in a row should not have the same set of values.

Proof: For a 2/3 code of m=3 the generator matrix is given by

$$G_3 = \begin{bmatrix} 1011\ 1011\ 1011 \\ 0101\ 1011\ 1111 \end{bmatrix} \quad (4)$$

This will not produce an output because all the columns in the first row has same values.

Proposition 2: All the m feedback shift registers should be used at least once.

Proof: For a 2/3 code of m=3, the generator matrix is given by

$$G_3 = \begin{bmatrix} 1011\ 0000\ 1111 \\ 0101\ 0000\ 0000 \end{bmatrix} \quad (5)$$

This will not generate an output because the second feedback shift register is never used.

Proposition 2 can be followed by bitwise ORing all the elements of the column in a particular row and checking whether the result is all '1's.

III. TRELLIS COMPLEXITY CALCULATION

Trellis complexity is in general measured in terms of state and branch complexities. These two complexities determine the storage and computation requirements of a trellis based decoding algorithm, such as Viterbi decoding algorithm. The state complexity of a trellis is measured by its state space dimension profile and the branch complexity is measured by the total number of branches (edges) in the trellis.

For the generator matrix G_m, we define

$$G' = [G_0\ G_1 .....G_m] \quad (6)$$

Now we can define G_{scalar}

$$G_{scalar} = \begin{bmatrix} G_0\ G_1 \ldots G_m & & & \\ & G_0\ G_1 \ldots G_m & & \\ & & . & \\ & & & . \\ & & & & . \\ & & & & & . \end{bmatrix} \quad (7)$$

The matrix in (7) is, except for the fact that it continues forever, the generator matrix for a binary block code (with a very regular structure), and so any of the techniques which have been developed for finding minimal trellises for block codes are useful for constructing trellis representations for convolutional codes. In the remainder of this section, we will show how to construct a trellis directly from any generator matrix for a given block code, and the minimal trellis if the generator is in "minimal-span'' form, to construct a trellis for C based on the infinite scalar genera tor matrix G_{scalar}.

From [1] G_{scalar} has LR property if it has at most one underlined and one overlined element. The first and last non zero element is called underlined and overlined element respectively. The trellis complexity can be found out by

$$TC = \frac{1}{k}\sum_{1}^{n} a^i \qquad (8)$$

where a^i is the number of active elements contained in the i^{th} column of the matrix module G_{cap}.

$$G_{cap} = \begin{bmatrix} G_m \\ G_{m-1} \\ . \\ . \\ . \\ G_0 \end{bmatrix} \qquad (9)$$

where a^i is the number of active elements contained in the i^{th} column of the matrix module G_{cap}.

For each row of this matrix we should find out the first non zero element and the last non zero element. Thus active span can be defined as the number of elements between the first non zero element and the last non zero element.

Now, for each row in G_{cap}

$$q = floor(activespan/m) \qquad (10)$$

$$r = activespan\%m \qquad (11)$$

If r is zero then trellis complexity becomes

$$TC = \sum_{i=1}^{(m+1)k} 2^m \frac{q}{k} \qquad (12)$$

If q is zero then trellis complexity becomes

$$TC = \sum_{i=1}^{(m+1)k} 2^{ri} \frac{1}{k} \qquad (13)$$

If q and r is not zero then

$$TC = \sum_{i=1}^{(m+1)k} (q_i 2^m + 2^{ri}) \frac{1}{k} \qquad (14)$$

For eg a 2/3 code with m=3

$$G_{cap} = \begin{bmatrix} 1&0&1&1&0&0&1&0&1&0&1&0 \\ 0&1&0&0&1&1&0&0&0&0&0&1 \\ 0&0&0&1&0&1&1&1&0&1&0&1 \\ 0&0&0&0&1&1&0&0&1&1&0&0 \\ 0&0&0&0&0&0&1&1&0&1&1&0 \\ 0&0&0&0&0&0&1&0&1&1&0&1 \\ 0&0&0&0&0&0&0&0&0&1&0&1 \\ 0&0&0&0&0&0&0&0&0&0&1&0 \end{bmatrix}$$

For the above G_{cap} the active span for the:

- first row is 11
- second row is 11
- third row is 9
- fourth row is 6
- fifth row is 5
- sixth row is 6
- seventh row is 3
- eighth row is 1

The trellis complexity for this code will be

$$\begin{aligned} TC' &= (3(2^3) + 2^2 + 3(2^3) + 2^2 + 3(2^3) + 2^3 + 2^2 \\ &\quad + 2(2^3) + 2(2^3) + 2^3 + 2^1) \\ &= 134 \end{aligned}$$

$$TC = \frac{TC'}{k} = \frac{134}{2} = 67$$

IV. MINIMAL TRELLIS MODULE

A trellis is said to be minimal if the total number of states in the trellis is minimum [3]-[5]. The above definition of minimality is commonly used in the literature. However, a more meaningful and useful definition of minimality of a trellis is in terms of its state space dimension trellis if the state space dimension at each time of the trellis is minimum.

If the trellis complexity mentioned in [1] is reduced it will have two advantages; (1) more number of codes can be included (2) the decoding complexity can be

reduced. In this paper the trellis complexity is reduced by applying some elementary row operations in the generator matrices.

For e.g. a 3/7 code with m=3 has the following generator matrices

$$G_3 = \begin{bmatrix} 000000100000100000000000000 \\ 000100000001000000010000000 \\ 000000000100000000001000000 \end{bmatrix}$$

$$G_2 = \begin{bmatrix} 000000000000000000000100000 \\ 000000000000000000001100000 \\ 000000000000000000001110000 \end{bmatrix}$$

$$G_1 = \begin{bmatrix} 000000000000000000001000000 \\ 000000000000000000000100000 \\ 000000000000000000000010000100 \end{bmatrix}$$

$$G_0 = \begin{bmatrix} 000000000000000000000100000 \\ 000000000000000000000010000 \\ 000000000000000000000000001 \end{bmatrix}$$

The trellis complexity obtained for this code is 42. By XORing the second and third rows of each matrix we get a new set of trellis complexities.

By varying G_0, $TC = 46$

By varying G_1, $TC = 44.6667$

By varying G_2, $TC = 39.3333$

By varying G_3, $TC = 47.3333$

From these the variation in G_2 gives the best value of trellis complexity

The weight spectra obtained from the minimal trellis and the weight spectra obtained from the trellis with n-bit branches are the same for the first few terms and are only slightly different for the other terms. In fact, for weight spectra obtained from the conventional trellis are identical to those obtained from the minimal trellis, since breaking up a zero-state-to-zero-state path of weight less than 2dfree into two zero-state-to-zero-state paths would imply that there exists a zero-state-to-zero-state path with weight less than

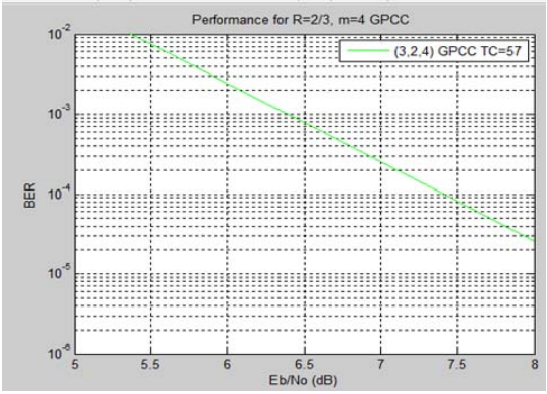


Figure1. BER simulation of a 2/3 code with m=4 over AWGN channel and QPSK modulation

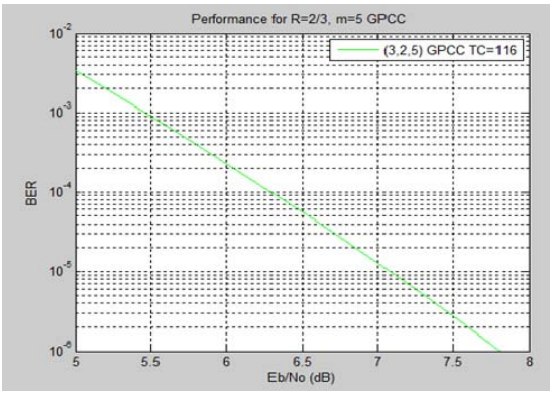


Figure 2. BER simulation of a 2/3 code with m=5 over AWGN channel and QPSK modulation

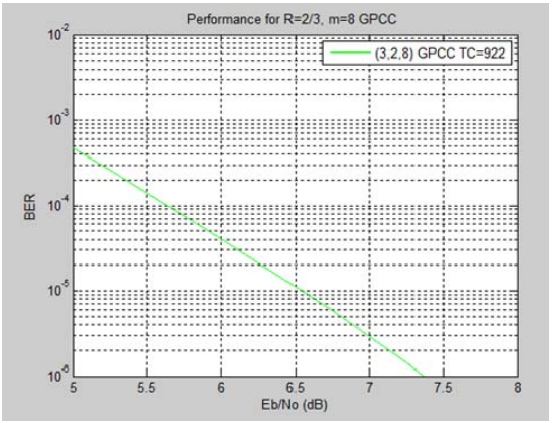


Figure 3. BER simulation of a 2/3 code with m=8 over AWGN channel and QPSK modulation

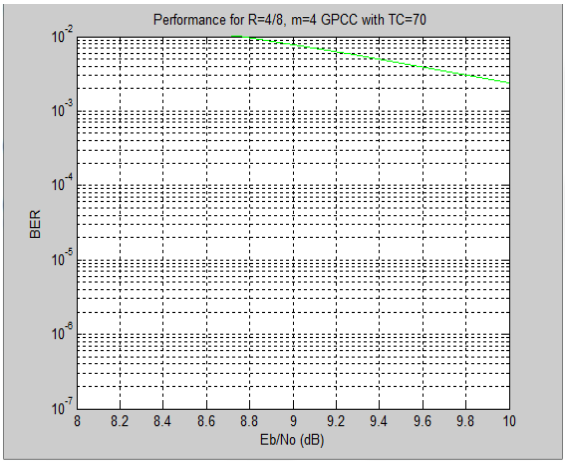


Figure 4. BER simulation of a 4/8 code with m=4 over AWGN channel and QPSK modulation

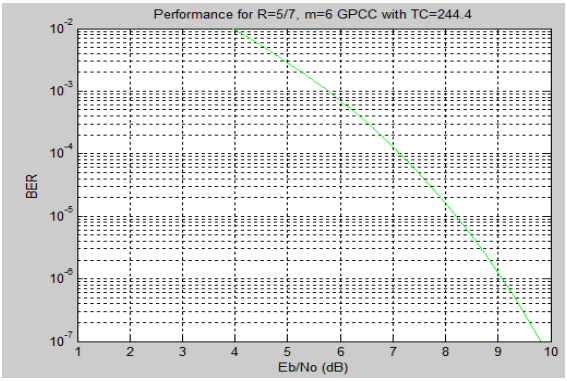


Figure 5. BER simulation of a 5/7 code with m=6 over AWGN channel and QPSK modulation

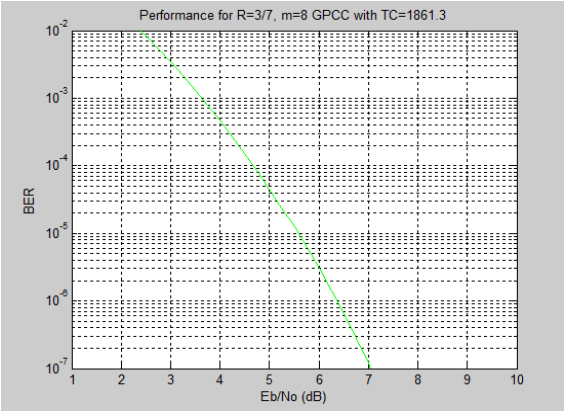


Figure 6. BER simulation of a 3/7 code with m=8 over AWGN channel and QPSK modulation

Viterbi decoding, using the conventional trellis with n-bit branches for decoding will yield the same bit error rate as that obtained from using the minimal trellis for decoding. Note that the zero states in the minimal trellis are different from the zero states in the trellis with n-bit branches. Hence, using the minimal trellis in the Viterbi decoding, some paths will be eliminated earlier as compared to using the trellis with n-bit branches. The first error events in the minimal trellis are in general different from those in the trellis with n-bit branches. Therefore, there will be a little difference between the first error event probability of decoding using the minimal trellis and that of decoding using the trellis with n-bit branches.

V. CODE SEARCH RESULTS

We conduct a refined code search, with respect to trellis complexity, within the class of GPCCs. The trellis complexity of the minimal trellis module is calculated from the "generator matrix module" of the code, following the procedure described in section III. We now describe the procedure we followed to find good GPCCs, in terms of trellis complexity. Results are obtained for different code rates and trellis complexities. In Table I, a refined list of good convolutional codes is shown for different code rates and different values of trellis complexities. The codes are defined by the polynomial generator matrix G(D) of a GPCC obtained from section II. It should be remarked that, whenever a GPCC has the same trellis complexity of the existing code, although we list the value, primacy of [1] is acknowledged. This class of convolutional codes is broad enough to contain good codes and yet has structural properties that facilitate the code search; for instance, one can easily control the span length of each row of the GPCC generator matrix, a property that allows us to search for GPCCs with fixed trellis complexity. The main contributions of this paper are as follows. First, it reveals that, when compared with respect to the given code rate, as opposed to the encoder memory size, some of the known convolutional codes in the literature present poor trellis complexities. New codes are tabulated in this paper to remedy this problem. Moreover, Table I show that, if the minimal trellis is used for the decoding of convolutional codes via the Viterbi algorithm, an error performance improvement and a reduction in the number of operations and comparisons can be achieved with the adoption of the new codes.

Simulations of some of the codes listed over AWGN channel and QPSK modulation are given in figures. It can be seen that in figure 1 trellis complexity of a 2/3 GPCC with m=4 it is reduced by 10.93% from [1] and the bit error rate from SNR from 5.3 to 8 is reduced by a factor 2.5×10^{-3}. In figure 2 for 2/3 GPCC

with m=5 the trellis complexity is reduced by 9.3% from [1] and bit error rate is reduced by a factor 0.33×10^{-3} from SNR 5 to 7.8. In figure 3 a 2/3 GPCC code with m=8 is shown. Here the trellis complexity is reduced by 9.96% from [1]. Here the bit error rate is reduced by a factor of 0.2×10^{-3} from SNR 5 to 7.3. In figure 4 for 4/8 GPCC with m=4 the trellis complexity is reduced by 2.77% from [1] and bit error rate is reduced by a factor 2.5×10^{-1} from SNR 8.6 to 10. In figure 5 a 5/7 GPCC code with m=6 is shown. Here the trellis complexity is reduced by 4.53% from [1]. Here the bit error rate is reduced by a factor of 10^{-5} from SNR 4 to 10. Figure 6 shows 3/7 GPCC with m=8. The trellis complexity is reduced by 0.82% from [1] and bit error rate is reduced by a factor 10^{-5} from SNR 2.5 to 7.

TABLE I - SEARCH RESULTS

k/n	m	TC in [1]	Proposed TC
2/3	3	32	31
2/3	4	64	57
2/3	5	128	116
2/3	6	256	241
2/3	7	512	470
2/3	8	1024	922
2/5	4	96	95
2/5	5	224	223
2/5	6	448	435
2/5	7	896	854
2/5	8	1792	1686
3/4	3	26.7	26
3/4	4	53.3	52.67
3/4	5	106.7	98.67
3/4	6	256	249.33
3/4	7	512	486.67
3/4	8	1024	959.33
3/5	3	32	30.67
3/5	4	64	62.67
3/5	5	128	127.33
3/5	6	298.7	292.67
3/5	7	512	486
3/5	8	1024	958.67
3/6	4	74.7	74.67
3/7	2	21.3	21.3
3/7	3	40	39.33
3/7	4	117.3	94.67
3/7	5	234.6	230.67
3/7	6	469.2	468.67
3/7	7	938.4	937.3
3/7	8	1876.8	1861.3
4/5	3	24	23.5
4/5	4	56	55.5
4/5	5	96	92
4/5	6	224	219.5
4/5	7	448	441
4/5	8	896	877.5
4/6	4	80	79
4/7	2	20	20
4/7	3	40	39
4/7	4	64	58
4/7	5	176	172.5
4/7	6	320	318
4/7	7	512	492
4/7	8	1152	1118
4/8	4	72	70
5/6	3	22.8	22.4
5/6	4	51.2	48.8
5/6	5	102.4	95.6
5/6	6	230.4	224.4
5/6	7	358.4	357.2
5/6	8	716.8	715.6
5/7	2	19.2	19.2
5/7	3	25.6	25.2
5/7	4	57.6	56.4
5/7	5	128	125.6
5/7	6	256	244.4
5/7	7	409.5	400
5/7	8	921.6	900.4
6/7	3	21.3	21
6/7	4	42.6	42
6/7	5	96	92.7
6/7	6	234.6	228
6/7	7	341.3	340.3
6/7	8	682.6	660.67
7/8	3	20.6	20.57
7/8	4	41.2	39.14
7/8	5	91.4	83.14
7/8	6	182.8	175.71
7/8	7	329.1	328
7/8	8	731.4	659.42

VI. CONCLUSION AND FUTURE WORK

In this paper, we have studied and analyzed results of GPCC for various code rates using MATLAB 7.12.0 (R2011a). The trellis complexities of 72 different types of convolutional codes are determined by puncturing trellis module instead of puncturing inputs to encoder, a conventional method. The trellis complexities are improved significantly over the recent works done in this field. For a particular application, if the error rate is tolerable without limiting the bandwidth of the signal, the bit error rate can be reduced considerably. The trellis complexities, determined by the proposed method are flexible so that the particular trellis module with best performance can be chosen for the required application. The reduced complexities of GPCC in proposed method will also result in reduced decoding complexity presence in the maximum likelihood Viterbi algorithm. Thus the performance will further improve. The proposed method is exhaustive and a maximum limit is reached for trellis complexity reduction. The future work will focus on improving the performance by using turbo codes, a class of forward error correcting codes. Turbo codes consist of two parallel concatenated recursive systematic Convolutional codes. Thus more flexibility in reducing the trellis complexity can be achieved provided that suitable synchronization and interleaving methods are adopted.

REFERENCES

[1] Alexandros Katsiotis, Panagiotis Rizomiliotis and Nicholas Kalouptsidis,"New Constructions of High-Performance Low-Complexity Convolutional Codes", *IEEE Trans.Communications*, vol 58, Jul 2010.

[2] B. F. Uchoa-Filho, R. D. Souza, C. Pimentel, and M. C. Lin, "Generalized punctured convolutional codes," *IEEE Commun. Lett.* vol. 9, pp. 1070–1072, Dec. 2005.

[3] R. J. McEliece and W. Lin, "The trellis complexity of convolutional codes," *IEEE Trans. Inf. Theory*, vol. 42, pp. 1855–1864, Nov. 1996.

[4] H. Tang and M. C. Lin, "On $(n, n-1)$ convolutional codes with low trellis complexity," *IEEE Trans. Commun*., vol. 50, pp. 37–47, Jan. 2002.

[5] B. F. Uchoa-Filho, R. D. Souza, C. Pimentel, and M. Jar, "Further results on convolutional codes based on a minimal trellis complexity measure," in Proc. *IEEE Intern. Telecom. Symp*., Sep. 2006, pp. 123–128.

[6] J. B. Cain, J. C. Clark, Jr, and J. M. Geist, "Punctured convolutional codes of rate $(n-1)/n$ and simplified maximum likelihood decoding," *IEEE Trans. Inf. Theory*, vol. 25, pp. 97–100, Jan. 1979.

[7] P. J. Lee, "Constructions of rate $(n-1)/n$ punctured convolutional codes with minimal required SNR criterion," *IEEE Trans. Commun*., vol. 36, pp. 1171–1173, Oct. 1988

Analysis And Control of Harmonics of Three Phase Induction Machine By Space Vector Decomposition Using Matlab Simulation

Sharda Patwa, Hemant Amhia

Jabalpur Engineering College, Electrical Engineering Department
Email.- shardapatwa@gmail.com

Abstract **-** The technique of vector space decomposition control of three phase induction machine is presented in this paper. By vector space decomposition the analytical modeling and control of machine are accomplished. The space vector decomposition technique limits the 5^{th}, 7^{th}, 17^{th}, 19^{th}….harmonic currents, which in such a system is otherwise difficult to control. This synopsis present harmonic analysis of motor current of medium and high power Variable Frequency Drive (VFD) Systems. Computer simulation of a IGBT fed induction motor based on constant voltage/frequency (V/f) operation is implemented using simulation software.

Keywords*- Induction Motor, Space vector decomposition, Harmonic Analysis, PWM Inverter, Frequency Response, IGBT inverter.*

I. INTRODUCTION

The induction motor, which is the most widely used motor type in the industry, has been used because of its good self-starting capability, simple and rugged structure, low cost and reliability etc. Along with variable frequency AC inverters, induction motors are used in many adjustable speed applications which do not require fast dynamic response. The concept of vector control has opened up a new possibility that induction motors can be controlled to achieve dynamic performance good.

Induction motor for many years has been regarded as the workhorse in industrial applications. In the last few decades, the induction motor has evolved from being a constant speed motor to a variable speed, variable torque machine. When applications required large amounts of power and torque, the induction motor became more efficient to use. With the invention of variable voltage, variable frequency drives (VVVF), the use of an induction motor has increased. Variable frequency IGBT Inverters fed space vector control method is widely used to control the harmonics of 3-phase squirrel cage induction Motors (IM) over a wide range by varying the stator frequency. In particular the IGBT fed space vector control are widely preferred in industries for individual medium to high power variable speed drive systems, driving a group of motors connected in parallel at economic costs. The "Insulated Gate Bipolar Transistor" (IGBT) is a common choice in modern VFDs. The IGBT can switch on and off several thousand times per second and precisely control the power delivered to the motor. The IGBT uses "pulse width modulation" (PWM) technique to simulate a sine wave current at the desired frequency to the motor.

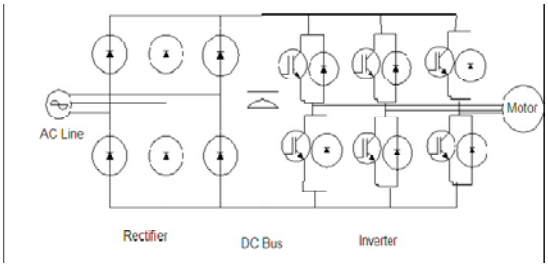


FIG 1- IGBT fed Three Phase Induction Machine.

II. METHODOLOGY OF WORK

A simulation model of such induction motor drive system is developed and its dynamic response is verified by observing harmonic analysis and control to establish acceptability of the model. Then a series of simulations are carried out for three different post fault conditions which are:

1. Open circuiting of one of the six IGBTs gate signal,

2. Blowing off one IGBT in the inverter module.

3. Space vector control method.

Computer simulation is the discipline of designing a model of an actual or theoretical physical system, executing the model on a digital computer, and analyzing the execution output. Simulation model of the motor drive system developed in Power SIM (PSIM) has been used for this study.

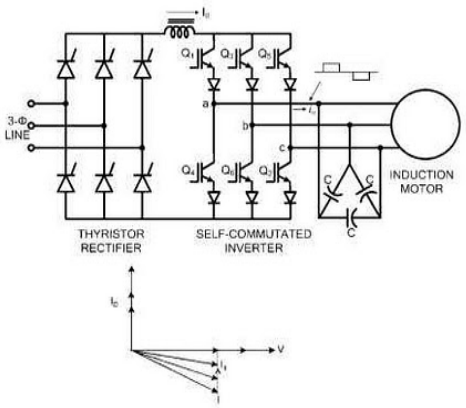


FIG 2- IGBT fed Three Phase Induction Machine.

III. RESULT AND DISCUSSION

In Figure 3, the results about harmonics present in voltage and current has been given. By this we can see the waveform of voltage and current in induction machine.

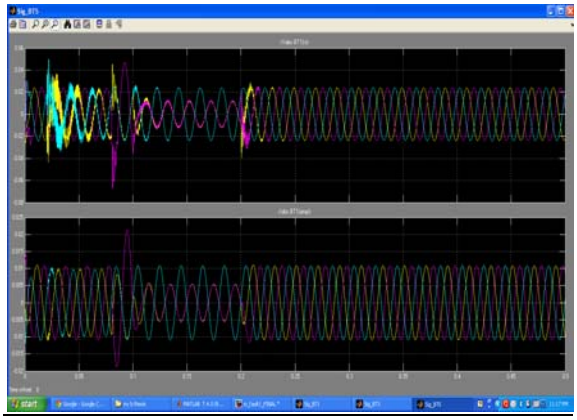


Fig.3- voltage and current waveform in induction machine with harmonics.

In Figure 4, the results of voltage and current has been given. By this we can see the waveform of voltage and current in induction machine after applying the space vector decomposition.

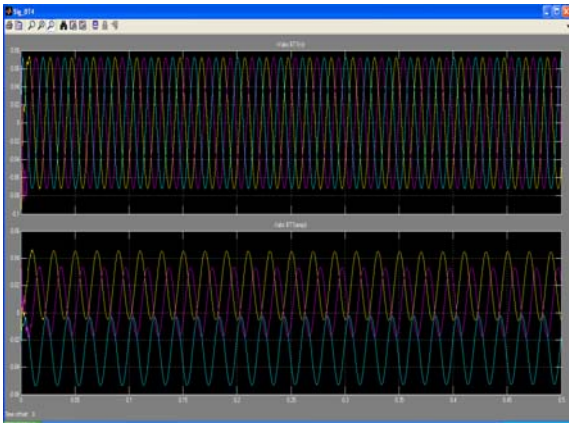


fig. 4- voltage and current waveform in induction machine without harmonics and application of space vector decomposition.

In future this controlling method can be used as-

1. The extension of work can be done on a high level for industrial purpose.
2. This process can analyses and control the harmonics of 3 phase induction motor.
3. Simulation method of harmonic analysis and control can be done on a big scale.

IV. CONCLUSION

The presented paper introduced scheme for induction motor drive harmonic control based on space vector modulation. The induction mo-tor drive controlled by the proposed harmonic control techniques was simulated using MATLAB SIMULINK package programs. Moreover, its experimental setup was build.

To validate the effectiveness of the proposed harmonic control methods, the induction motor drive was subjected to starting and reference motor speed disturbances. Simulation in terms of motor stator current, speed, and torque responses will be plotted and obtained. The results prove the powerfulness and effectiveness of the proposed control approach in sense of fast response with less settling time and less over-shoot/undershoot.

REFERENCES

1. Space vector PWM control of dual three phase induction machine using vector space decomposition. Yuan Zhao, Student Member, *IEEE*, and Thomas A. Lido, fellow, *IEEE.*
2. Wade, S., Donavan, M. W. and Williams, B.W. ‘ Modeling and simulation of induction machine

vector control and rotor resistance identification.', *IEEE Trans Power Electronics*, vol. 12, No. 3, pp. 495-505 (1997)

3. Shi, K.L., Chan, T.F. and Wong, Y.K., 'Modeling of the three phase induction motor using SIMULINK', Record of the 1997 IEEE International Electric machines and drives conference, USA, pp. WB3-6 (1997).
4. Krause, P.C., and Sendoff, S. D., Analysis of Electrical Machinery, IEEE (1995).

◈◈◈

Dynamic Job Scheduling In Grid Computing

Jani Kuntesh Ketan & Arpita Shah

Computer Science & Engineering Department CHARUSAT UNIVERSITY changa, Dist. Anand, India
E-mail : kunteshjani@gmail.com, arpitashah.ce@ecchanga.ac.in

Abstract - Grid computing is growing rapidly in the distributed heterogeneous systems for utilizing and sharing large-scale resources to solve complex scientific problems. Scheduling is the most recent topic used to achieve high performance in grid environments. It aims to find a suitable allocation of resources for each job. A typical problem which arises during this task is the decision of scheduling. It is about an effective utilization of processor to minimize tardiness time of a job, when it is being scheduled. Scheduling jobs to resources in grid computing is complicated due to the distributed and heterogeneous nature of the resources. The efficient scheduling of independent jobs in a heterogeneous computing environment is an important problem in domains such as grid computing. In general, finding optimal schedule for such an environment using the traditional sequential method is an NP-hard problem whereas heuristic approaches will provide near optimal solutions for complex problems. The Ant colony algorithm, which is one of the heuristic algorithms, suits well for the grid scheduling environment using stigmeric communication.

Keywords*-component; Grid Computing; job Scheduling; Ant Colony Algorithm.*

I. INTRODUCTION

The term Grid computing originated computer power as easy to access as an electric power grid. Grid computing is the process of applying the resources of many computers in a network to a single problem at the situations like - usually to a scientific or technical problem that requires a great number of computer processing cycles or access to large amounts of data. It involves the use of software that can divide and form out pieces of a program to as many as several thousand computers. It looks as distributed and large-scale cluster computing and as a form of network-distributed parallel processing[1].Distributed systems consist of multiple computers that communicate through computer networks. Research by [2] defined that cluster and grid computing are the most suitable ways for establishing distributed systems. Grid computing is proposed to overcome this problem where various resources from different geographic area are combined in order to develop a grid computing environment. To achieve the promising potentials of tremendous distributed resources, effective and efficient scheduling algorithms are fundamentally important. The scheduling problem is defined NP-hard problem [4] and it is not trivial. There are two types of scheduling namely static scheduling and dynamic scheduling in grid computing system. For the static scheduling, jobs are assigned to suitable resources before their execution begin. Once started, they keep running on the same resources without interruption. However, for the dynamic scheduling, reevaluation is allowed of already taken assignment decisions during job execution [3].

II. ORGANIZATION OF PAPER

The organization of the paper further is as follows. The literature review is presented in Section III, existing system is analyzed in Section IV and summary & open issues are discussed in Section V.

III. LITERATURE REVIEW

In the past few years, researchers have proposed scheduling algorithms for parallel system [5 - 9]. However, the problem of grid scheduling is still more complex than the proposed solutions. Therefore, this issue attracts the interests of the large number of researchers [10- 14]. Current systems [15] of grid resource management was surveyed and analyzed based on classification of scheduler organization, system status, scheduling and rescheduling policies. However, the characteristics and various techniques of the existing grid scheduling algorithms are still complex particularly with extra added components.

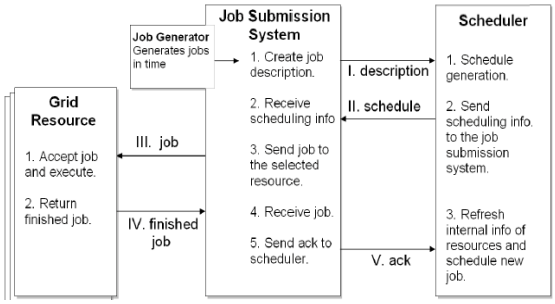


Fig. 1 : Job scheduling work flow

At the present time, job scheduling on grid computing is not only aims to find an optimal resource to improve the overall system performance but also to utilize the existing resources more efficiently. Recently, many researchers have been studied several works on job scheduling on grid environment. Some of those are the popular heuristic algorithms, which have been developed, are min-min [16], the fast greedy [16], tabu search [16] and an Ant System [17].

In 1999, the Ant Colony Optimization (ACO) metaheuristic was proposed by Dorigo, Di Caro and Gambardella, which has been successfully used to solve many NP-problem, such as TSP, job shop scheduling, etc. In the past few years, several researchers proposed solutions to solve grid scheduling problem by using ACO [20]. Several studies have been trying to apply ACO for solving grid scheduling problem. Z. Xu et al [18] proposed a simple ACO within grid simulation architecture environment and used evaluation index in response time and resource average utilization. E. Lu et al. [19] and H. Yan et al. [20] also proposed an improved Ant Colony algorithm, which could improve the performance such as job finishing ratio. However, they have never used the various evaluation indices to evaluate their algorithm. The ACO becomes very popular algorithm to apply for solving grid scheduling problem.

IV. ANALYSIS OF RELATED WORK ON ACO

Before Ant colony optimization algorithms [Dorigo 1996] are multi-agent systems, which consist of agents with the collective behavior (stigmergy) of ants for finding shortest paths. Ant colony algorithms were inspired by the observation of real ant colonies. Ants are social insects, that is, insects that live in colonies and whose behavior is directed more to the survival of the colony as a whole than to that of a single individual component of the colony. Social insects have captured the attention of many scientists because of the high structuration level their colonies can achieve, especially when compared to the relative simplicity of the colony's individuals. An important and interesting behavior of ant colonies is their foraging behavior, and, in particular, how ants can find shortest paths between food sources and their nest. While walking from food sources to the nest and vice versa, ants deposit on the ground a substance called pheromone, forming in this way a pheromone trail. When more paths are available from the nest to a food source, a colony of ants may be able to exploit the pheromone trails left by the individual ants to discover the shortest path from the nest to the food source and back. It is also interesting to note that ants can perform this specific behavior using a simple form of indirect communication mediated by pheromone laying, known as stigmergy.

```
1.  procedure ACO
2.  begin
3.  Initialize the pheromone
4.  While stopping criterion not satisfied do
5.  Position each ant in a starting node
6.  Repeat
7.  for        each ant              do
8.  Chose next node by applying the state
    transition rate
9.  end for
10. until every ant has build a solution
11. Update the pheromone
12. end while
13. end
```

III Pseudo code for Existing Ant colony Algorithm

A. Modified Ant Colony Algorithm for Grid Scheduling

The modified ant colony algorithm has changed the basic Pheromone updating rule of original ant colony algorithm. The improved pheromone updating rule is given by :

$$\tau ij(t)new=[\{(1-)/(1+\rho)\}*\tau ij(t)old]+[\{\rho/(1+\rho)\}*\Delta\tau ij(t)] \quad (1)$$

Where,

$\tau ij(t)$ =Trail intensity of the edge(i,j).

ρ = Evaporation rate.

$\Delta\tau ij(t)$ = Additional pheromone when job moves from scheduler to resource.

The proposed Ant colony algorithm as a whole is a best suited method for tracking problem with large data sets. The above approach was simulated using GRIDSim toolkit and was found to be working efficiently and effectively. Experimental test carried out for a varied range of input set to ascertain the efficiency of the algorithm. From the results it is clearly evident that the proposed Ant colony algorithm offers better optimization a very fast rate.

B. Scheduling in Computational Grid with a New Hybrid Ant Colony Optimization Algorithm

The efficiency depends on makespan and flow-time. The makespan measures the throughput of the system and flow-time measures its QOS . The main objective of the proposed algorithm is to reduce the makespan and to converge towards the optimal solution in a very faster manner. Ant algorithm for scheduling initially started with the value of ETij matrix. ETij matrix defines the expected time taken by machine j to complete task i. Here jobs are considered as independent to each other. ETij matrix consists of N×M entries where N is number of independent jobs and M is number of resources available in grid environment. Increasing the pheromone levels associated with a chosen set of good solutions makes the algorithm faster to converge to a solution. Hence on modifying the pheromone updating rule and probability matrix the solution is converging in a fast manner than the existing ACO algorithm. The probability matrix calculation uses the the same transition rule to select the job which is to be executed next in the machine. This algorithm guarantees efficient resource allocation of the machines.

C. An Improved Ant Algorithm For Job Scheduling In Grid Computing

In this approach, the general adaptive ant algorithm is used to solve the job scheduling problem in the grid simulator environment. The transition probability is trade-off between visibility, that means the power performance of resource should be chosen with high probability, and trail intensity that means if on path j there is a lot of traffic then it is high desirable. Once some resources load heavy, it comes into being a bottleneck in the grid and influences to completing of jobs. Therefore the load balancing factor is introduced in the ant algorithm to improve the load balancing capability. We define the load balancing factor λ of the resource j, which is related to the job finishing rate in the resource j. In the algorithm, the trail intensity will be changed from ΔT_J to $\Delta T_J + C\lambda_J$ (C>0 is a coefficient of the load balancing factor), the more jobs finished, the more increases the trail intensity, contrarily, the more jobs not completed, the more decreases the trail intensity. The improved algorithm practice will improve the load balance status for the resources in the grid.

D. Ant Colony Algorithm for Job Scheduling in Grid Computing

This proposed algorithm aims to minimize the computational time of each job that must be processed by available resources in grid computing system. The algorithm will select the resources based on the pheromone value on each resource. A matrix that contains the pheromone value on each resource has been used to facilitate the selection of suitable resources to process submitted jobs. The proposed algorithm has been implemented in the grid system architecture which consists of four main components namely the grid information server, grid resource broker, jobs and resources After all ants have constructed a solution, the pheromone trails are updated according to the following formula:

$$T_{Jr}\ (t+1) = (1-\rho).T_{Jr} + \rho\Delta T_{Jr}^{bs} \quad (2)$$

where $\Delta\ \tau jr\ best$ = 1/Lbest. This global pheromone update is limited to a specific upper and lower trail limit. The ant which is allowed to add pheromone may be the *iterationbest solution* or *global best solution.* If a specific resource is often used in the best solution, it will receive a larger amount of pheromone and stagnation will occur. So, lower and upper limits on the possible pheromone strengths on any resource are imposed to avoid stagnation. The resource with high pheromone value will be selected by grid resource broker. So j3 will be processed by r2. After assigning j3 to r2, the local pheromone update is performed to the second row of r2. Column 3 is no longer needed because j3 has been assigned. The scheduling process in the proposed algorithm is based on the combination of local pheromone update and trail limits. This proposed technique is different from the previous algorithm based on its initial pheromone value and the used of the PV matrix. The initial pheromone value for this algorithm considered the estimated transmission time and characteristics of each jobs and resources. The local pheromone trail update will reduce the amount of pheromone in assigned resource, to ensure the resource is less desirable for other ants while the trail limit, which is the allowed range of the pheromone strength, is limited to maximum and minimum trail strength. This is a technique to control the value of pheromone updated on each resource to ensure that already assigned optimal resources will not be chosen for newly submitted jobs. The proposed algorithm is simple to be implemented due to the existing of information of each resources and jobs. This algorithm was able to minimize the completion time of each job.

E. An Ant Colony Optimization for Dynamic Job Scheduling in Grid Environment

Here the general adaptive ACO algorithm is basically the ACS (Ant Colony System) , is applied to solve the scheduling problem in grid environment. Four main key terms of ACO algorithm are listed and defined below: The expected time complete (ETC) is defined as the amount of time that the job is processed at a machine, and it has no load to the assigned job. Where the complete time (CT) is defined as the wall clock time, which machine completes for each job. Then, *CTij = aj*

+ rj + ETCij, where *aj* is arrival time of a job *j*th, *rj* is a release time of a job *j*th and *i* is a machine *ith*. The main important of ACO algorithm is to utilize the graph representation. therefore, the design of the graph is to identify the problem to connect the arcs correspondent for each job on each machine. Let *M* be a set of machines *{m1, m2, m3,.., mm}* and let *J* be a set of jobs *{j1, j2, j3, ..., jn}*, and $n > m$. Therefore, the graph *G = (M, {CTij}mxn)*. The problem is to find the optimal resources for the jobs. It can minimize the total tardiness time. The algorithm can find an optimal processor for each machine to allocate to a job that minimizes the tardiness time of a job when the job is scheduled in the system.

V. SUMMARY AND OPEN ISSUES

Dynamic job scheduling in grid computing is an NP-hard problem which can be solved using various techniques of cooperative sub-optimal category. Ant colony system is one such type of cooperative sub-optimal solution which can meet dynamically changing needs of a heterogeneous environment like computational grids. The open issues are:

- Hybridization to produce high throughput computing
- Evaluate the different cost measures such as make-span time, Grid Efficiency and job error ratio

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] K. Yan, S. Wang, S. Wang, and C. Chang, "Towards a hybrid load balancing policy in grid computing system," *Expert Systems with Applications*, vol. 36, pp. 12054-12064, 2009.

[3] M. Chtepen, "Dynamic scheduling in grids system," Sixth Firw PhD Symposium, Faculty of Engineering, Ghent University, pp.110, 2005.

[4] D. Fernandez-Baca (1989), "Allocating Modules to Processors in a Distributed System", *IEEE Transactions on Software Engineering*. Vol.15(11): Pages 1427-1436.

[5] D.G. Feitelson. "Packing schemes for gang scheduling", In *2nd Workshop on Job Scheduling Strategies for Parallel Processing*, volume LNCS 1162, pages 89–100, 1996.

[6] D.G. Feitelson, L. Rudolph, U. Schwiegelshohn, K.C. Sevcik, and P.Wong. "Theory and practice in parallel job scheduling", In *3rd Workshop on Job Scheduling Strategies for Parallel Processing*, volume LNCS 1291, pages 1–34, 1997.

[7] J. Krallmann, U. Schwiegelshohn, and R. Yahyapour. "On the design and evaluation of job scheduling algorithms", In *5th Workshop on Job Scheduling Strategies for Parallel Processing*, volume LNCS 1659,pages 17–42, 1999.

[8] J. M. van den Akker, J. A. Hoogeveen, and J. W. van Kempen. "Parallel machine scheduling through column generation: Minimax objective functions", In Y. Azar and T. Erlebach, editors, *European Symposium on Algorithms*, volume 2996 of *Lecture Notes in Computer Science*, pages 648-659. Springer, 2004.

[9] R.D. Nelson, D.F. Towsley, and A.N. Tantawi. "Performance analysis of parallel processing systems", *IEEE Transactions on Software Engineering*, 14(4):532–540, 1988.

[10] V. Hamscher, U. Schwiegelshohn, A. Streit, R.Yahyapour, "Evaluation of job-scheduling strategies for grid computing", *Proceedings of First IEEE/ACM International Workshop on Grid Computing, Lecture Notes in Computer Science*, vol. 1971, Springer, Berlin, 2000, pp. 191–202.

[11] V. Subramani, R. Kettimuthu, S. Srinivasan, P. Sadayappan, "Distributed job scheduling on computational grids using multiple simultaneous requests", *Proceedings of the 11th International Symposium on High Performance Distributed Computing*, 2002, pp.359–366.

[12] H. Shan, L, Oliker and R. Biswas, "Job Superscheduler Architecture and Performance in Computational Grid environments", *Proceedings of the ACM/IEEE SC2003 Conference (SC03),* 2003.

[13] C. Ernemann , V. Hamscher and R. Yahyapour, "Benefits of Global Grid Computing for Job Scheduling", *Proceedings of the Fifth IEEE/ACM International Workshop on Grid Computing (GRID'04),* 2004.

[14] K. Li, "Job scheduling and processor allocation for grid computing on Metacomputers ", *Journal of Parallel and Distributed Computing*, *Elsevier,* 2005

[15] K. Krauter, R. Buyya and M. Maheswaran, "A taxonomy and survey of Grid resource management systems for distributed

computing", Software Pract. Exp. 2 (2002) 135–164.

[16] T. D. Braun, H. J. Siegel, N. Beck, L. L. Bölöni, M. Maheswaran, A. I. Reuther, J. P. Robertson, M. D. Theys, B. Yao, D. Hensgen and R. F. Freund (2001), "A Comparison of Eleven Static Heuristics for Mapping a Class of Independent Tasks onto Heterogeneous Distributed Computing Systems", *Journal of Parallel and Distributed Computing*. Vol.61(6): Pages 810-837.

[17] G. Ritchie and J. Levine, "A fast, effective local search for scheduling independent jobs in heterogeneous computing environments".

[18] Z. Xu, X. Hou and J. Sun, "Ant Algorithm-Based Task Scheduling in Grid Computing", *Electrical and Computer Engineering, IEEE CCECE 2003, Canadian Conference*, 2003.

[19] E. Lu, Z. Xu and J. Sun, "An Extendable Grid Simulation Environment Based on GridSim", *Second International Workshop, GCC 2003*, volume LNCS 3032, pages 205–208, 2004.

[20] H. Yan, X. Shen, X. Li and M. Wu, "An Improved Ant Algorithm for Job Scheduling in Grid Computing", *In Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, 18-21 August 2005.

◈◈◈

Automatic Detection of EPILEPSY EEG Using Neural Networks

Satyanarayana Vollala & Karnakar Gulla

Department of CSE & IT, Talla Padmavathi College of Engg.,Warangal-506009(AP)
Department of CSE & IT, RVRInstitute of Enng.&Technology, Ibrahimpatnam-hyd (AP)- 501510
E-mail : satya.nitw@yahoo.com, karnakarrvrcse@gmail.com

Abstract **-** The electroencephalogram (EEG) signal plays an important role in the diagnosis of epilepsy. The EEG recordings of the ambulatory recording systems generate very lengthy data and the detection of the epileptic activity requires a time-consuming analysis of the entire length of the EEG data by an expert. The traditional methods of analysis being tedious, many automated diagnostic systems for epilepsy has emerged in recent years.This paper proposes a neural-network-based automated epileptic EEG detection system that uses approximate entropy (ApEn) as the input feature. ApEn is a statistical parameter that measures the predictability of the current amplitude values of a physiological signal based on its previous amplitude values. It is known that the value of the ApEn drops sharply during an epileptic seizure and this fact is used in the proposed system.Two different types of neural networks, namely, Elman and probabilistic neural networks are considered. ApEn is used for the first time in the proposed system for the detection of epilepsy using neural networks. It is shown that the overall accuracy values as high as 100% can be achieved by using the proposed system.

Keyword: *Approximate entropy (ApEn), artificial neural network (ANN), electroencephalogram (EEG), Elman network (EN), epilepsy, probabilistic neural network (PNN), seizure.*

I. INTRODUCTION

EPILEPSY is the second most common neurological disorder, affecting 1% of world population [2]. The electroencephalogram (EEG) signal is used for the purpose of the epileptic detection as it is a condition related to the brain's electrical activity [16]. Eighty five percent of patients with epilepsy live in the developing countries [3]. Electroencephalogram (EEG) is routinely used clinically to diagnose epilepsy [4]. Long-term video-EEG monitoring can provide 90% positive diagnostic information [5] and it has become the golden standard in epilepsy diagnosis. For the purpose of this research, we define the term "the diagnosis of epilepsy" as the determination of whether a person is epileptic or non-epileptic [6]. In majority of the cases, the onset of the seizures cannot be predicted in a short period, a continuous recording of the EEG is required to detect epilepsy. The approach of using automatic seizure recognition/detection algorithms would still require the recording of clinical seizures. Therefore, very long continuous EEG recording, preferably with synchronized video for several days or weeks, are needed to capture the seizures.

II. RELATED WORK

In the previous systems, have done maximum process by manual. After that some of researchers introduced automated diagnostic systems for epilepsy have been developed using different approaches. Some of those Example approaches of existing systems and the year of proposed methods are as follows. In 1982, Gotman [14] presented a computerized system for detecting a variety of seizures. In 1991, Murro [15] *et al*. developed an automated seizure detection system based on the discriminant analysis of the EEG signal recorded from the intracranial electrodes. In 1997, Qu and Gotman [16] proposed the use of the nearest-neighbor classifier on EEG features extracted in both the time and frequency domains to detect the onset of the epileptic seizures. In 2004, Gigola *et al*[17] . Used a method based on the evolution of the accumulated energy using wavelet analysis for the prediction of the epileptic seizure onset from the intracranial epileptic EEG recordings.

Artificial Neural Network (ANN) has been used for seizure related EEG recognition. We used in this work one kind of ANN as the classifier, namely the Elmen

Neural Network (ENN), for its high speed, high accuracy and real time property in updating network structure. It is very difficult to directly use raw EEG data as the input of an ANN. Therefore, the key is to parameterize the EEG data into features prior to the input into the ANN. A time-domain feature of the EEG signal called approximate entropy (ApEn) that reflects the nonlinear dynamics of the brain activity. Hence, it is a good feature to make use of in the automated detection of epilepsy. In this paper, this feature is applied, for the first time, in the automated detection of epilepsy using neural networks.

III. EXISTING SYSTEM

The method proposed by Pradhan *et al.* uses a raw EEG signal as an input to a learning vector quantization network. In 2004, Nigam and Graupe proposed a new neural network model called LAMSTAR network, and two time-domain attributes of EEG, namely, relative spike amplitude and spike rhythmicity have been used as inputs for the purpose of the detection of epilepsy. The method proposed by Kiymik *et al.* uses a back propagation neural network with periodogram and autoregressive features as the input for the automated detection of epilepsy. Several approaches have been developed elsewhere, with varying success, in attempts to automatically detect epileptic form activity and seizures in the EEG. In most of these, the tendency has been to look for extended amplitude and frequency changes rather than aiming to capture characteristic waveforms. Because of the widely varying morphology of seizures, we also chose to incorporate measures of extended amplitude and frequency changes, as central features in our multi-stage detection algorithm. Probabilistic neural network for epileptic detection is an existing automated system for epileptic EEG detection. In probabilistic neural network, AI-based classifier is used. An AI-based classifier is essentially a mapping from the feature space to the discrete class space.

IV. PROPOSED SYSTEM

The proposed system consists of following various modules which are also shown using a flowchart in Fig 1.

Two sets of EEG data [16] of normal and epileptic subjects from 18 subjects is used. 10 subjects normal and epileptic EEG data is used for training and remaining was used for testing. The depth electrodes are placed symmetrically into the hippocampal formations and strip electrodes are placed onto the lateral and basal regions of the neocortex. The epileptic EEG segments are selected from all the recording sites exhibiting ictal activity [34].

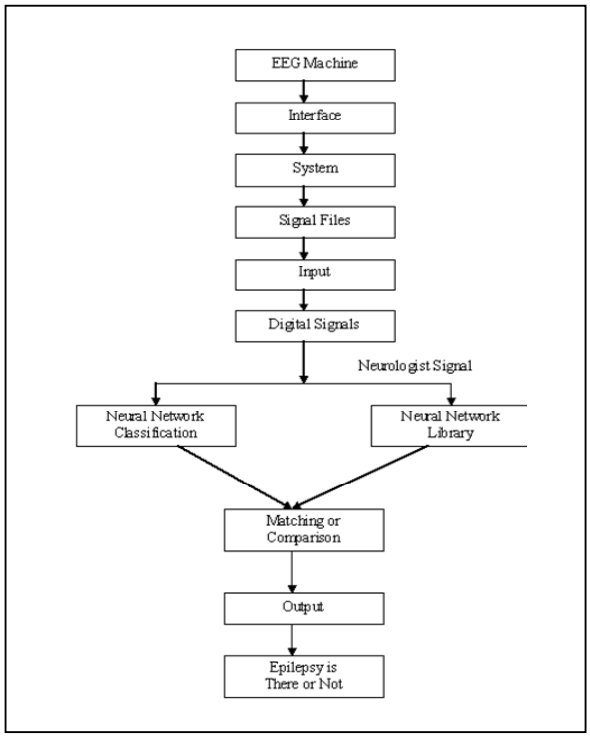


Fig. 1. Proposed System

The intracranial epileptic EEG has been chosen for this classification system as the intracranial recordings offer the most precise access to the emergence of seizure. EEG signal is recorded during the occurrence of the epileptic seizures from these intracranial electrodes. EEG signal of normal subjects are obtained using intracranial electrodes. The EEG data acquired and recorded using Neurofile NT Digital Video is greyscaled, sampled at 256 Hz, and digitized into text files of peak values of 128 channels. A line noise of 50hz is present in the data is filtered out using a Notch filter at 50hz(49hz - 51hz).An optional bandpass 0.53 to 40hz filter may be applied to increase sensitivity and remove unnecessary signal frequencies. Notch and Bandpass Filter Order is found using Trial and Error method. The order of filter used is 3*ceil (sampling rate/lower cutoff frequency). Then EEG recordings are visually inspected for artifacts.

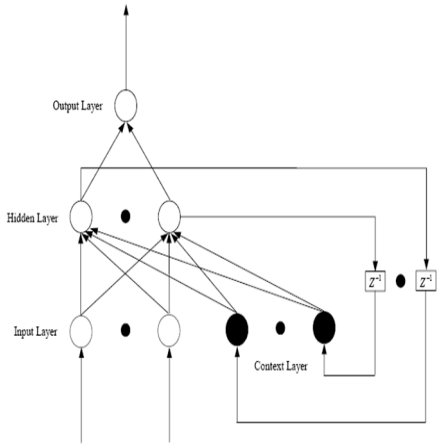
V. NEURAL NETWORK CLASSIFIER

Artificial Neural Networks are good classifiers because of their features like robustness, adaptive learning, generalization capability and self-organization. In situation where enough data are available for training the system and where the simple classification algorithm fails the ANNs (Artificial Neural Networks) are useful. To the best of our knowledge, the performance of Elman neural network has not been investigated yet. Elman

neural network has been used in this paper for detection of epilepsy. The target, threshold values and description of the configuration used for Elman neural network is given below.

Elman Network: It is a recurrent neural network with two layers the hidden layer and the output layer. It is a backpropagation network with a feedback connection from the output of hidden layer to its input. This means that the function learnt by the network can be based on the current inputs plus a record of the previous states and outputs of the network.

The spatial patterns and temporal patterns are recognized by the ENN due to its feedback connection. The inputs for the neural network are the ApEn values corresponding to the normal and epileptic EEG signals. The ApEn values of epileptic patients have a little wide spike followed by abrupt fall in ApEn value. ENN is used to learn and detect this temporal pattern of ApEn values of epileptic patients. It detects the EEG signal without that particular temporal pattern as non epileptic. The activation functions used for the two-layered ENN are tan-sigmoidal and log-sigmoidal for the hidden and output layers, respectively.



Structure of Elman Neural Networks

90 hidden nodes and one output node is used. The training of the EN is done by using Resilient Backpropogation. Train the network by using the target value 0 and 1 for normal EEG and epileptic EEG, respectively. The classification is done by using the range of values 0-0.3 for the normal EEG and 0.7-1 for epileptic EEG. The MSE Error Goal is set to 0.01, which is sufficient for accurate classification

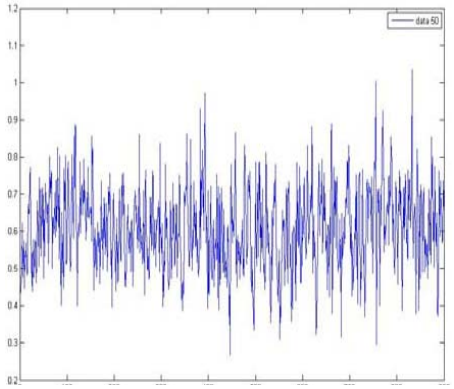


Fig.2

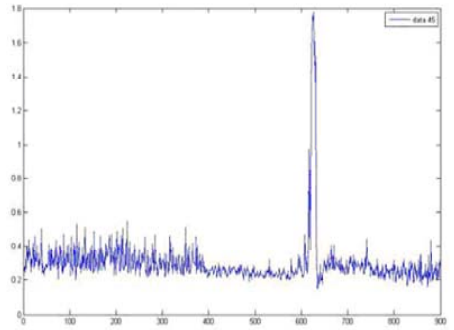


Fig.3

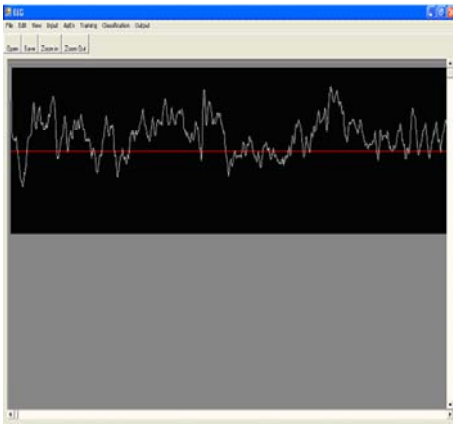


Fig.4

ApEn: The proposed system makes use of a single feature called ApEn for the epileptic detection. The ApEn is a time-domain feature that is capable of classifying complex systems [9]. ApEn is a recently formulated statistical parameter to quantify the regularity of a time series data of physiological signals [9].

It was first proposed by Pincus in 1991 [11] and has been predominantly used in the analysis of the heart rate variability [12]–[14] and endocrine hormone release pulsatility [15], estimation of regularity in epileptic seizure time series data, and in the estimation of the depth of anesthesia.

Performance Evaluation Parameters:

The performance of EN is evaluated in terms of the three parameters i.e., Sensitivity (**SE**), Specificity (**SP**) and Overall Accuracy (**OA**), which are defined which are defined in [22], [23], and [24]as:

$$\mathrm{SE}(\%) = \frac{\mathrm{TN_{CP}}}{\mathrm{TN_{AP}}} \times 100$$

Where **TNCP** depicts the total number of correctly detected positive patterns and TNAP represents the total number of actual positive patterns. A positive pattern indicates a detected seizure.

$$\mathrm{SP}(\%) = \frac{\mathrm{TN_{CN}}}{\mathrm{TN_{AN}}} \times 100$$

Where **TNCN** represents the total number of correctly detected negative patterns and **TNAN** represents the total number of actual negative patterns. A negative pattern indicates a detected nonseizure.

$$\mathrm{OA}(\%) = \frac{\mathrm{TN_{CDP}}}{\mathrm{TN_{APP}}} \times 100$$

Where **TNCDP** represents the total number of correctly detected patterns and **TNAPP** represents the total number of applied patterns. A pattern indicates both seizure and nonseizure.

Result : ApEn values are computed for selected combinations of *m*, *r*, and *N*. The values of *m*, *r*, and *N* that are used for the experiments are as follows. ApEn values are computed for both normal and epileptic EEG signals and are fed as inputs to the EN network. Among the available 100 EEG data sets, 60 data sets are used for training and the remaining are used for testing the performance of the neural networks. This choice is made arbitrarily keeping in mind that enough datasets are provided for the neural network to understand the inherent structure of the data so that it can classify the unknown datasets properly. The ApEn values are calculated for each data frame and the number of ApEn values used for training and testing the neural networks.

From 18 patients, a total of 60 datasets were selected, with 30 ictal(epileptic seizures) and 30 interictal(non epileptic) data, for testing. The Performance Evaluation Parameters were calculated and the results obtained as follows:

Overall Accuracy (OA) = 93.43%

Sensitivity (SE) = 96.87%

Specificity (SP) = 90.43%

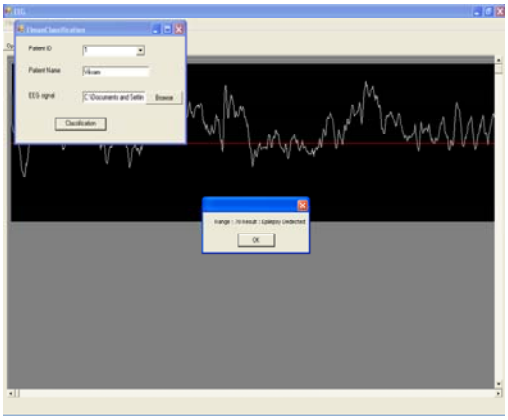


Fig.5

VI. CONCLUSION

Epilepsy is a common neurological disorder not a disease which is not contagious, fainting disorder and cause mental illness. Epileptic person has a tendency to have recurrent seizures which produces non linear dynamic system. By using Elman Neural Network and ApEn as an input feature for implementation of detection of epilepsy. Since, it is using a single input feature so that's why we having low computational burden and best suited for the real-time detection of epileptic seizures. We described one kind of ANN and conception how to implement it. Though the use of ANNs increases the computational complexity, the high overall detection accuracies expected from this system surpasses its disadvantage as in any automated seizure detection system, since the detection of the seizure with high accuracy is of primary importance. This is because it is known that ApEn possesses good characteristics such as robustness in the characterization of the epileptic patterns and low computational burden. Hence, an automated system using ApEn as the input feature is best suited for the real time detection of the epileptic seizures.

REFERENCES :

[1] V. Srinivasan, C. Eswaran, and N. Sriraam, "Approximate entropybased epileptic EEG detection using artificial neural networks," IEEE Transactions on Information Technology in Biomedicine, vol. 11, no. 3, pp. 288–295, May 2007.

[2] K. Lehnertz, F. Mormann, T. Kreuz, R. Andrzejak, C. Rieke, P. David, and C. Elger, "Seizure prediction by nonlinear EEG analysis," IEEE Engineering in Medicine and Biology Magazine, 2003.

[3] Atlas: Epilepsy Care in the World. World Health Organization, 2005. [4] R. Flink, B. Pedersen, A. B. Guekht, K. Malmgren, R. Michelucci, B. Neville, F. Pinto, U. Stephani, and C. Ozkara, "Guidelines for the use of EEG methodology in the diagnosis of epilepsy international league against epilepsy," Acta Neurologica Scandinavica, vol. 106, 2002.

[5] C. Logar, B. Walzl, and H. Lechner, "Role of long-term EEG monitoring in diagnosis and treatment of epilepsy," European Neurology, vol. 34, 1994.

[6] T. S.Walczak, P. Jayakar, and E. M. Mizrahi, Epilepsy: A Comprehensive Textbook, 2nd ed. Lippincott Williams & Wilkins, 2008, ch. 73, pp. 809 – 813.

[7] C. Bigan, "A recursive time-frequency processing method for neural networks recognition of EEG seizures," in Neural Networks and Expert Systems in Medicine and Healthcare, E. C. Ifeachor, A. Sperduti, and A. Starita, Eds. Singapore: World Scientific, 1998, pp. 67–73.

[8] J. Bruhn, H. Ropcke, and A. Hoeft, "Approximate entropy as an electroencephalographic measure of anesthetic drug effect during desflurane anesthesia," Anesthesiology, vol. 92, pp. 715–726, 2000.

[9] S. M. Pincus, I. M. Gladstone, and R. A. Ehrenkranz, "A regularity static for medical data analysis," J. Clin. Monit., vol. 7, pp. 335–345, 1991.

[10] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," Amer. J. Physiol., vol. 278, pp. H2039–H2049, 2000.

[11] S. M. Pincus, "Approximate entropy as a measure of system complexity," Proc. Natl. Acad. Sci. USA, vol. 88, pp. 2297–2301, 1991.

[12] S. M. Pincus, T. R. Cummins, and G. G. Haddad, "Heart rate control in normal and aborted SIDS infants," Amer. J. Physiol., vol. 264, no. 3, pp. R638–R646, 1993.

[13] S. M. Pincus and A. L. Goldberger, "Physiological time series analysis: what does regularity quantify?," Amer. J. Physiol., vol. 266, pp. H1643–Hl656, 1994.

[14] W. S. Kim, Y. Z. Yoon, J. H. Bae, and K. S. Soh, "Nonlinear characteristics of heart rate time series: Influence of three recumbent positions in patients with mild or severe coronary artery disease," Physiol. Meas., vol. 26, pp. 517–529, 2005.

[15] S. M. Pincus, "Older males secrete luteinizing hormone and testosterone more irregularly and joint more asynchronously, than younger males," Proc. Natl. Acad. Sci. U.S.A., vol. 93, pp. 14100–14105, 1996.

[16] N. Mc Grogan (1999). "Neural network detection of epileptic seizures in the electroencephalogram", [Online]. Available: http://www.new.ox.ac.uk/˜nmcgroga/work/transfer

[17] W. Weng and K. Khorasani, "An adaptive structure neural network with application to EEG automatic seizure detection," Neural Netw., vol. 9, pp. 1223–1240, 1996.

[18] J. Gotman and L. Wang, "State-dependent spike detection: Concepts and preliminary results," Electroencephalogr. Clin. Neurophysiol., vol. 79, pp. 11–19, 1991.

[19] N. Pradhan, P. K. Sadasivan, and G. R. Arunodaya, "Detection of seizure activity in EEG by an artificial neural network: A preliminary study," Comput. Biomed. Res., vol. 29, pp. 303–313, 1996.

[20] V. P. Nigam and D. Graupe, "A neural-network-based detection of epilepsy," Neurol. Res., vol. 26, pp. 55–60, 2004. on in EEG recording," Phys. A: Stat. Mech. Appl., vol. 273, pp. 495–505, 1999.

[21] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger, "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state," Phys. Rev. E, Stat. Nonlinear SoftMatter Phys., vol. 64, pp. 061907-1–061907-8, 2001.

◈◈◈

Analysis and Simulation Studies for the Estimation of Rotor Position in Sensorless BLDC

Neethi S.Pillai, Salitha.K & Chikku.Abraham

Rajagiri School Of Engineering & Technology Cochin, India
E-mail : neethi953@gmail.com, salitha450@gmail.com & chikkuabraham@yahoo.com

Abstract - Initial rotor position information is essential for brushless dc motor in order to ensure its stable operation. A simple method for determining the initial rotor position of a sensorless brushless dc motor at standstill is discussed in the paper. The principle behind the rotor position estimation is based on simple detection and comparison of phase voltages and dc link current responses thus relating it with stator inductances. The advantage of this method of estimation of rotor position is that it requires only three voltage pulse injection, and a resolution of 30^o is achieved. Moreover no other parameters of the machine are required. The effectiveness of the method is validated by simulation results in Matlab Simulink platform.

Keywords-*Brushless DC(BLDC),saturation, rotor position, voltage pulse injection.*

I. INTRODUCTION

. Brushless DC motors (BLDC), because of their high starting torque, high efficiency, no excitation losses, silent operation and durability is now being widely used in number of industrial applications such as compressor, in electrical vehicles, hard disc drives and in medical applications.

An inverter driven three-phase BLDC motor, as shown in Fig.1, requires the rotor position information in order to ensure a stable operation. Usually the position information is available using position sensors like Hall Effect sensors or position encoders, which increases the cost of overall system that makes the system unfavourable. Hence the concept of sensorless system is being implemented which is highly reliable and particularly for numerous low cost applications.

One of the major problems faced by sensorless BLDC is the initial start-up, since most of the sensorless techniques are based on back-EMF voltage detection which disappears at standstill. To solve this start-up problem an open-loop start-up is described in [2]. The major drawback of this method is that any temporary reverse rotation due to unknown load characteristics can result in no rotor position information. Another solution to this problem is to align the rotor to predefined position by exiting any two phases of the motor for a predefined set time [3]. This pre-alignment of rotor can fail due to large static friction. Hence it is very necessary to solve all the start-up problems and to estimate the initial rotor position of the rotor in order to have a smooth stable sensorless operation for BLDC motors.

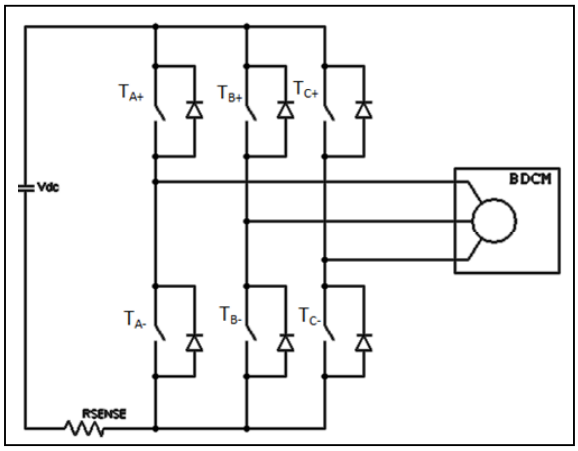


Fig.1 Inverter driven three phase BLDC Motor

Rotor position estimations based on inductance variation methods are usually done for BLDC that has iron core stator. In [4] rotor position estimation is done by ratio of differences of current responses that is obtained from applying voltage pulses. But this method proves to be inapplicable for low cost applications as it requires three current sensors for each phase.

In this method, rotor position estimation is based on simple detection and comparison of phase voltages and dc link current responses, resulting from the injection of three voltage pulses to the selected windings [1].

II. INITIAL ROTOR POSITION ESTIMATION TECHNIQUE

The principle behind the rotor position estimation is the saturation effect of the stator core, caused by the rotor magnet as described in [4]. Fig.2 shows the measured current response and equivalent inductance versus rotor position in a BLDC motor. The rotor position estimation comprises of two methods.

1. Inductance comparison process

2. Polarity determination process

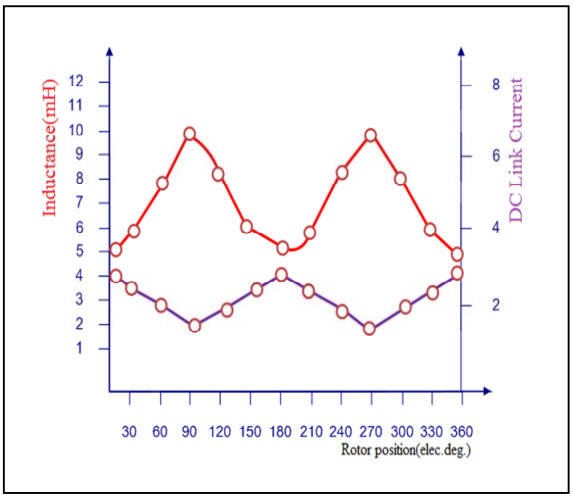


Fig.2 Current response and equivalent inductance versus rotor position of a BLDC machine (phase-A)

A . Inductance Comparison Process

The process comprises of a sequence of voltage pulses applied to selected pair of windings. Voltage pulse injection consists of two intervals: pulse injecting interval and freewheeling interval. Fig.3 shows the pulse injecting interval along with a freewheeling interval of a BLDC motor after the first voltage pulse injection. The entire process can be explained as follows.

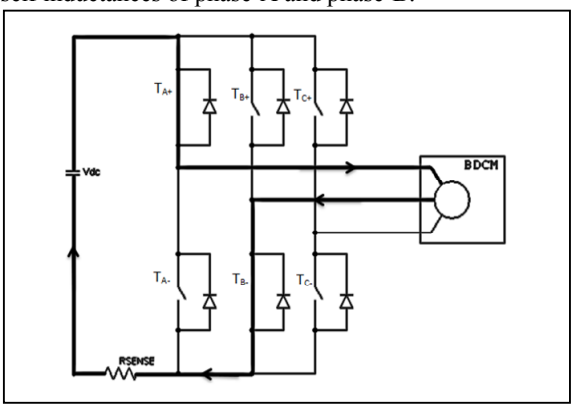
(1) First Voltage Pulse Injection:The first voltage pulse is injected to the phase-A and phase-B turning on switches T_{A+} and T_{B-}. Fig.3(a) shows the pulse injecting interval and Fig.3(b) shows freewheeling interval, while the phase voltage can be measured across the floating phase-C with respect to the negative dc bus. Taking winding resistance and the current sensing resistance negligibly small, the dc bus voltage can be approximated as

$$V_{dc} \approx [L_A(\theta_0) + L_B(\theta_0)]\frac{di_{1(on)}}{dt}$$
$$\approx L_A(\theta_0)\frac{\Delta i_{1(on)}}{T_{s1(on)}} + L_B(\theta_0)\frac{\Delta i_{1(on)}}{T_{s1(on)}} \quad (1)$$
$$= V_{AN1(on)} + V_{NB1(on)}$$

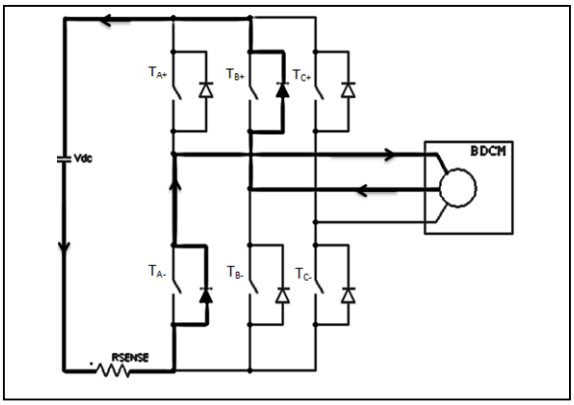
where

$$\left.\begin{aligned} L_A(\theta_0) &= L_{aa}(\theta_0) + L_{ab}(\theta_0) \\ L_B(\theta_0) &= L_{bb}(\theta_0) + L_{ba}(\theta_0) \end{aligned}\right\} \quad (2)$$

Here L_{aa} and L_{bb} are effective self inductances of phase-A and phase-B respectively. Likewise L_{ab} and L_{ba} are the mutual inductances due to currents in phase-B and phase-A. θ_0 is the initial rotor position, $\Delta i_{1(on)}$ is the current change during turn on time and $T_{s1(on)}$ is the turn on time interval.$V_{AN1(on)}$ and $V_{NB1(on)}$ are phase voltages during the pulse injecting interval. The magnitudes of mutual inductances L_{ab} and L_{ba} have same value and therefore the induced phase voltages arise mainly from self inductances of phase-A and phase-B.



(a)



(b)

Fig.3. Switching states of BLDC motor during injection of first voltage pulse. (a). Pulse-injecting interval (b). Freewheeling interval

From (1), $V_{AN1(on)}$ and $V_{NB1(on)}$ can be written in terms of V_{dc} as

$$\left.\begin{aligned} V_{AN1(on)} &= \frac{L_A}{L_A+L_B}.V_{dc} \\ V_{NB1(on)} &= \frac{L_B}{L_A+L_B}.V_{dc} \end{aligned}\right\} \quad (3)$$

After the initial voltage pulse injecting interval, switches T_{A+} and T_{B-} are turned off. Consequently the freewheeling interval occurs as shown in Fig.4(a).The DC bus voltage during freewheeling interval can be written as

$$\begin{aligned} V_{dc} &\approx -L_A(\theta_0)\frac{\Delta i_{1(off)}}{T_{s1(off)}} - L_B(\theta_0)\frac{\Delta i_{1(off)}}{T_{s1(off)}} \\ &= V_{NA1(off)} + V_{BN1(off)} \end{aligned} \quad (4)$$

Where $\Delta i_{1(off)}$ is the change in current during freewheeling interval, $T_{s1(off)}$.$V_{NA1(off)}$ and $V_{NB1(off)}$ are the voltage across phase-A and phase-B respectively during freewheeling interval. These voltages can be measured across phase-C with reference to -$V_{dc.}$.

The total voltage drop across phase-A and phase-B during the pulse injecting interval is opposite to freewheeling interval, respective change in bus voltage will reflect on the phase-A and phase-B windings. Therefore from (1) and (4) it can realized that

$$\left.\begin{aligned} &L_A(\theta_0)\frac{\Delta i_{1(on)}}{T_{s1(on)}} = -L_A(\theta_0)\frac{\Delta i_{1(off)}}{T_{s1(off)}} \text{ or} \\ &V_{AN1(on)} = V_{NA1(off)} \end{aligned}\right\}$$

$$\begin{aligned} &L_{\mathbf{B}}(\theta_0)\frac{\Delta i_{1(on)}}{T_{s1(on)}} = L_B(\theta_0)\frac{\Delta i_{1(off)}}{T_{s1(off)}} \quad \text{or} \\ &V_{NB1(on)} = V_{BN1(off)} \end{aligned} \quad (5)$$

Also phase–A and phase-B voltages during freewheeling interval can be written as

$$\left.\begin{aligned} V_{NA1(off)} &= \frac{L_A}{L_A+L_B}V_{dc} \\ V_{BN1(off)} &= \frac{L_B}{L_A+L_B}V_{dc} \end{aligned}\right\} \quad (6)$$

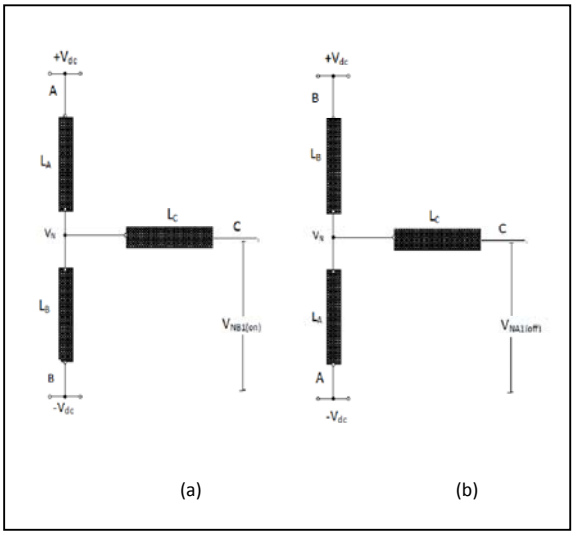
From (3) and (6)

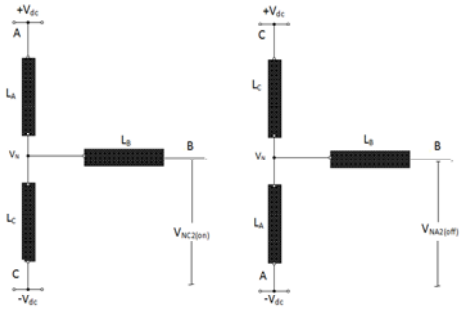
$$V_{NA1(off)} - V_{BN1(on)} = \frac{(L_A-L_B)}{(L_A+L_B)}V_{dc} \quad (7)$$

Hence from (7) it is very clear that values of L_A and L_B can be compared to each other by voltage comparison between $V_{NA1(off)}$ and $V_{BN1(on)}$, since both of them are referred to $-V_{dc}$.

(2) Second Voltage Pulse Injection: In the same way the second voltage pulse is injected to switches T_{A+} and T_{C-} of the phase-A and phase-C for the same time interval. During the pulse injecting interval phase-A is connected to $+V_{dc}$ and phase-C is connected to $-V_{dc}$. phase-B will be the floating phase, across which the phase voltages can be measured. Similar to first voltage pulse injection during freewheeling interval switches T_{A+} and T_{C-} are turned off. Hence L_A and L_C can be compared by comparing $V_{NA2(off)}$ and $V_{NC2(on)}$.Therefore it can be written as

$$V_{NA2(off)} - V_{NC2(on)} = \frac{(L_A-L_C)}{(L_A+L_C)}V_{dc} \quad (8)$$





(c) (d)

Fig.4.Terminal voltage detection. (a) Pulse injecting interval of first voltage pulse injection. (b) Freewheeling interval of first voltage pulse injection. (c) Pulse injecting interval of second voltage pulse injection (d) Freewheeling interval of second voltage pulse injection.

Where $V_{NC2(on)}$ is voltage across phase-C during pulse injecting interval and $V_{NA2(off)}$ is the phase-A voltage during freewheeling interval. These voltages can be measured across phase-B with reference to $-V_{dc}$. In the same way the relative values of L_B and L_C can be obtained from $V_{BN1(on)}$ and $V_{NC2(on)}$ as

$$V_{NB1(on)} - V_{NC2(on)} = (L_B - L_C)\frac{L_A}{(L_A + L_C)(L_A + L_B)}V_{dc} \quad (9)$$

Hence the winding inductances L_A, L_B and L_C can be compared to each other using (7) and (9). Table.1 gives information on relationship between inductance comparisons and possible initial position of the rotor (the north pole). So in order to determine the polarity of rotor magnet an additional third voltage pulse is required.

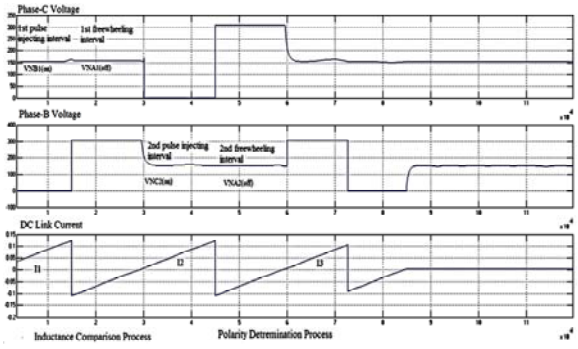


Fig.5 Actual phase voltage and DC link current for both the processes

TABLE 1
Determination of initial rotor position for BLDC

Phase Voltage Comparison	Inductance Comparison	Possible Initial Position	3rd Injection	Peak Current	Initial Rotor Position
$V_{NB1} > V_{NA1}$ $V_{NC2} \geq V_{NA2}$ $V_{NB1} > V_{NC2}$	$L_B > L_C \geq L_A$	$0^\circ < \theta^\circ < 30^\circ$ $180^\circ < \theta^\circ < 210^\circ$	T_{C+}, T_{A-}	$I_2 > I_3$ $I_3 > I_2$	$0^\circ < \theta^\circ < 30^\circ$ $180^\circ < \theta^\circ < 210^\circ$
$V_{NB1} \geq V_{NA1}$ $V_{NC2} < V_{NA2}$ $V_{NB1} > V_{NC2}$	$L_B \geq L_A > L_C$	$30^\circ < \theta^\circ < 60^\circ$ $210^\circ < \theta^\circ < 240^\circ$	T_{C+}, T_{A-}	$I_2 > I_3$ $I_3 > I_2$	$30^\circ < \theta^\circ < 60^\circ$ $210^\circ < \theta^\circ < 240^\circ$
$V_{NB1} < V_{NA1}$ $V_{NC2} < V_{NA2}$ $V_{NB1} \geq V_{NC2}$	$L_A > L_B \geq L_C$	$60^\circ < \theta^\circ < 90^\circ$ $240^\circ < \theta^\circ < 270^\circ$	T_{C+}, T_{A-}	$I_2 > I_3$ $I_3 > I_2$	$60^\circ < \theta^\circ < 90^\circ$ $240^\circ < \theta^\circ < 270^\circ$
$V_{NB1} < V_{NA1}$ $V_{NC2} \leq V_{NA2}$ $V_{NB1} < V_{NC2}$	$L_A \geq L_C > L_B$	$90^\circ < \theta^\circ < 120^\circ$ $270^\circ < \theta^\circ < 300^\circ$	T_{B+}, T_{A-}	$I_3 > I_1$ $I_1 > I_3$	$90^\circ < \theta^\circ < 120^\circ$ $270^\circ < \theta^\circ < 300^\circ$
$V_{NB1} \leq V_{NA1}$ $V_{NC2} > V_{NA2}$ $V_{NB1} < V_{NC2}$	$L_C > L_A \geq L_B$	$120^\circ < \theta^\circ < 150^\circ$ $300^\circ < \theta^\circ < 330^\circ$	T_{B+}, T_{A-}	$I_3 > I_1$ $I_1 > I_3$	$120^\circ < \theta^\circ < 150^\circ$ $300^\circ < \theta^\circ < 330^\circ$
$V_{NB1} > V_{NA1}$ $V_{NC2} > V_{NA2}$ $V_{NB1} \leq V_{NC2}$	$L_C \geq L_B > L_A$	$150^\circ < \theta^\circ < 180^\circ$ $330^\circ < \theta^\circ < 360^\circ$	T_{B+}, T_{A-}	$I_3 > I_1$ $I_1 > I_3$	$150^\circ < \theta^\circ < 180^\circ$ $330^\circ < \theta^\circ < 360^\circ$

B. Polarity Determination Process

In polarity determination process polarity of rotor current is determined by injecting an additional third voltage pulse to a selected pair of switches, As an example, for the condition $L_A > L_B \geq L_C$, the possible initial rotor positions can be either at $60^o < \theta_0 < 90^o$ or at, $240^o < \theta_0 < 270^o$. So in order to determine where the rotor magnet pole is actually located, a third voltage pulse is required to be injected The peak dc current should be noted at the end of all the three voltage pulse injecting intervals across the resistor R_{SENSE}.

The principle behind the determination of polarity of rotor magnet is clearly described in [2] that, the winding currents from the injected pulse voltages can either increase or decrease stator saturation which may result in the variation of stator inductance, as shown in Fig.2. Therefore the third voltage pulse is injected according to the following conditions.

1) If the possible initial position of the rotor is between 0^o - 90^o or 180^o - 270^o, turn on switches T_{C+} and T_{A-} and compare I_2 with I_3 .

2) If the possible initial position of the rotor is between 90^o - 180^o or 270^o - 360^o, turn on switches T_{B+} and T_{A-} and compare I_1 with I_3.

Thus by comparing the DC link currents the initial rotor position of BLDC motor can be estimated from the Table1.

III. SIMULATION RESULTS

Effectiveness of the method is validated from simulation results. Simulation was done in MATLAB Simulink platform. Fig.6 shows the simulation circuit for the estimation of rotor position of BLDC. The parameters of the motor used for simulation is given in Table.2. Voltages pulses are given to the respective inverter switches using separate signal builders for 150μs duration. The DC bus voltage is 310V. The phase voltage during the pulse injecting period is measured across the input terminals of the BLDC motor with reference to $-V_{dc}$

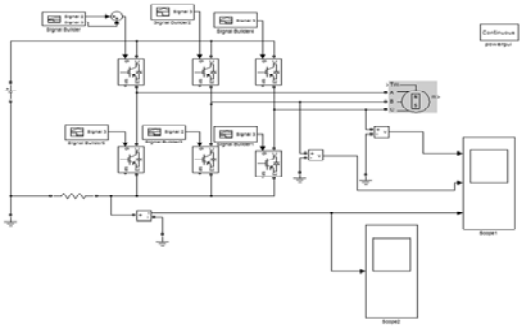


Fig.6 Simulation set up for estimation of rotor position in BLDC

TABLE
Machine Parameters

ZNo.of poles	4
DC bus voltage	310V
Stator phase inductance	0.2H
Stator phase resistance	2.8Ω

The current responses during all three pulse injecting period is measured across resistor R_{SENSE}, value of which is very negligible. So the voltage drop across this resistor is not taken into account.

The phase voltages and current responses are measured at the middle of the entire three pulse injecting interval, ie at 75μs to avoid the adverse effects due to parasitic elements. The validity of this method is verified by conducting the simulation for θ=31° by adjusting the motor model. Switching signals to inverter switches are shown in Fig.7. Initially pulse signals are given to the switches T_{A+} and T_{B-} for 150μs.After the first pulse injection voltage across phase-C is measured with reference to $-V_{dc}$, which is found to be $V_{NB1} \geq V_{NA1}$, as shown in Fig.8.

Second pulse injection is given in such a way to turn on switches T_{A+} and T_{C-}. Then the voltage is measured across phase-B with reference to $-V_{dc}$ and is found to be as $V_{NA2}<V_{NC2}$ and $V_{NB1}>V_{NC2}$. From Table1, it is clear that the choice of third pulse injection is to turn on switches T_{C+} and T_{A-} and also the current comparison is between I_2 and I_3. Accordingly the estimated initial rotor position can be between $30^o<\theta_0<60^o$ sector. Fig.9 shows the dc link current that is measured for rotor position θ=31°. It is very clear from the result currents $I_2 > I_3$.

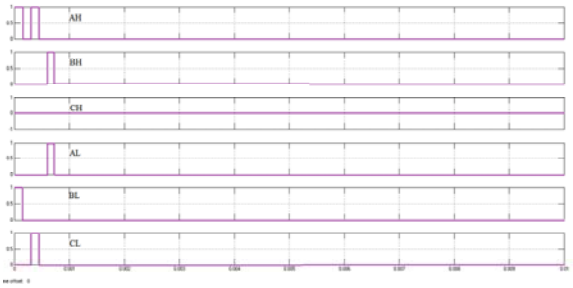


Fig.7. Switching pulses for θ=31°

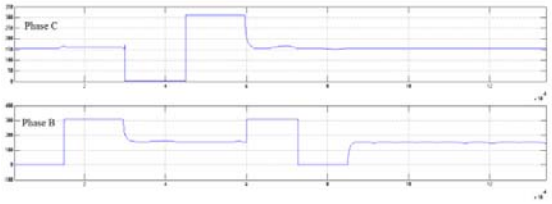


Fig.8 Actual phase voltages for both processes against corresponding gate signals at θ=31°

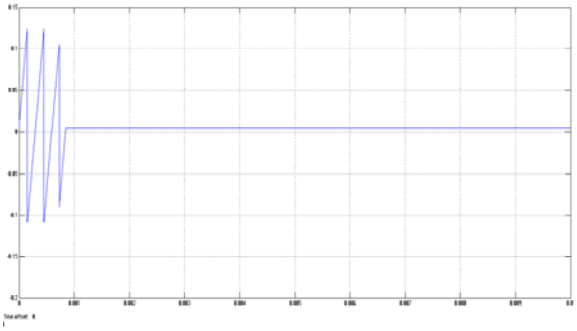


Fig.9. Actual DC link current for both processes against corresponding gate signals at θ=31°

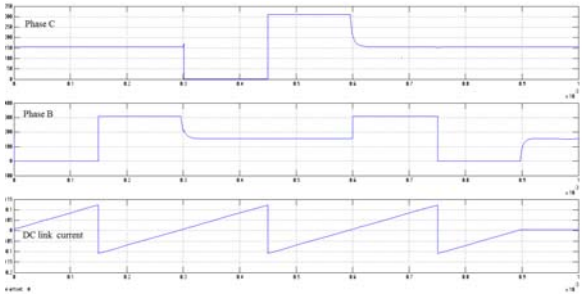


Fig.10. Actual phase voltages and dc link current for both processes against corresponding gate signals at $\theta=125^o$

Fig.10 shows the actual phase voltage and dc link current for both processes against corresponding gate signals at $\theta=125^o$. Initially pulse signals are given to the switches T_{A+} and T_{B-} for 150μs. After the first pulse injection voltage across phase-C is measured with reference to $-V_{dc}$, which is found to be $V_{NB1} \leq V_{NA1}$.Second pulse injection is given in such a way to turn on switches T_{A+} and T_{C-}, and then the voltage is measured across phase-B with reference to $-V_{dc}$ and is found to be as $V_{NA2}<V_{NC2}$ and $V_{NB1}<V_{NC2}$. From table1 it is clear the choice of third pulse injection is to turn on switches T_{B+} and T_{A-} and also the current comparison is between I_1 and I_3. Accordingly the estimated rotor position is the $120^o<\theta_0<150^o$ sector. Hence this method for rotor position estimation gives a resolution of 30^o.

IV. CONCLUSION

A simple estimation to determine the initial rotor position of sensorless BLDC has been analyzed in the paper which gives a resolution of 30^o. This paper is based on the theory that the winding inductance is varied due to the influence of saturation of stator core. It requires only three voltage pulses to be injected for the estimation of rotor position. Additionally one sensing resistor is added to a typical sensorless drive. Moreover no machine parameters are required. Hence this method proves to be very effective for the rotor position estimation in position sensorless BLDC. Once the motor has started, any other sensorless algorithms can be implemented reliably to change over the machine to running mode.

REFERENCES

[1] Prasit Champa, Pakasit Somsiri, Pongpit Wipasuramonton, Paiboon Nakmahachalasint, ”Initial rotor position estimation for sensorless brushless DC drives,” *IEEE Trans.Ind. Appl,*vol.45, no.4, pp.1318-1324, Jul/Aug.2009.

[2] ST Microelectronics, Application Note AN1276 BLDC Motor Start Routine for the ST72141 Microcontroller.

[3] S.Ogaswara and H.Agaki “An approach to position sensorless drives for brushless DC motors,” *IEEE Trans.Ind.Appl.,*vol.27, no.5,pp.928-933, Sep/Oct.1991.

[4] P.B.Schimdt, L.Gasperi,G.Ray, and A.H.Wijenayake,”Initial rotor angle detection of a non-salient pole permenant magnet synchronous machine,” in *Conf. Rec. IEEE IAS Annu.Meeting,* New Orleans, LA, 1997, pp.459-463

[5] Zezhong Xia, Wen Li, Wenjuan Sheng, Youxin Yuan, “ Design of a control system for sensorless brushless DC motor using dsPIC”, *IEEE Conf. Ind.Appl,* 2008, pp. 551-556.

[6] TI motor Compendium on motor controls, Texas Instruments 2010.

◈◈◈

Comparision of Pre, Post And Symmetrical Dispersion Compensation Scheme With 10 Gb/s NRZ link for SCM System

Ruchi Agarwal & Vivekanand Mishra

Electronics and communication Engineering ,
Sardar Vallabhbhai National Institute of Technology, Surat-395007, India
E-mail : ruchi.agarwal1987@gmail.com, ruchi.agarwal1987@gmail.com

Abstract - In this paper, the implementation of subcarrier multiplexing technique is developed and investigated with pre-, post- and symmetrical-dispersion compensation methods for 10 Gb/s non-return to zero (NRZ) links using standard and dispersion compensated fibers through computer simulations to optimize high data rate optical transmission. Mathematical analysis is done to evaluate performance in term of bit error rate. Simulation is done using Optisys Software version 10.0. Motivation to this research is to compare all three compensation methods and it is found that the symmetrical compensation method is superior to pre- and post-compensation methods. On comparing pre- and post-compensation methods, it is found that the later is superior to the former. A 10-Gb/s SCM test bed has been set up in which 4 * 2.5 Gb/s data streams are combined into one wavelength that occupies a 20-GHz optical bandwidth. Thus by using these comparisions one can get a promising system to the symmetric high capacity access network with high spectral efficiency, cost effective, good flexibility.

Keywords— *Compensating device, Electrical drivers, MZM (Mach Zender Modulator), Bit Error Rate, Quality Factor, SCM(Subcarrier multiplexing), Optisys Software version 10.0.*

I. INTRODUCTION

There are lots of technologies in optical network with which lots of work has been done previously like Time Division Multiplexing (TDM), Wavelength Division Multiplexing (WDM) by which use of the optical Bandwidth provided by the optical Fibers become more efficient. SCM follows a different approach compared to WDM. In WDM a Tera hertz optical carrier is modulated with a baseband signal of typically hundred of Mbit/s. In an SCMA infrastructure, the baseband data is first modulated on a GHz wide subcarrier that is subsequently modulated in the THz optical carrier. This way each signal occupies a different portion of the optical spectrum surrounding the centre frequency of the optical carrier. At the receiving side, as normally happens in a commercial radio service, the receiver is tuned to the correct subcarrier frequency, filtering out the other subcarriers.

Because of its simple and low-cost implementation, high-speed optical data transmission using SCM technology attracted the attention of many researchers. The most significant advantage of SCM in optical communications is its ability to place different optical carriers together closely. This is because microwave and RF devices are much more mature than optical devices: the stability of a microwave oscillator is much better than an optical oscillator (laser diode) and the frequency selectivity of a microwave filter is much better than an optical filter. Therefore, the efficiency of bandwidth utilization of SCM is expected to be much better than conventional optical WDM. SCM is existing technology that has been used in radio, cable-TV and satellite at much lower data rates. In all these applications the use of SCM technology is in CATV technology [1, 2].

It is most promising method to use dispersion compensating fiber (DCF) efficiently to upgrade installed link made of standard single mode fiber (SMF) [3].These type of passive and devices are superior for SCM systems, commercially available, easy to install and cascadable in networks, indicating that at present the DCF technique can effectively compete with any other dispersion management approach. Whenever SMF is used in any research for high bit rate transmission with low loss but dispersion is an important impairment that degrades overall system performance. As length of fiber increased or bit rate increased or number of channels increased, the dispersion-induced broadening of short pulses propagating in the fiber causes crosstalk between the adjacent time slots, leading to errors when the communication distance increases beyond the dispersion length of the fiber[4]. To combat dispersion and nonlinearities, each SCM channel can separately be pre-compensated, post-compensated or dual-compensated schemes. The use of erbium-doped

amplifiers (EDFAs) operating in the 1.55 lm region has increased the link distance as limited by fiber loss in optical communication systems and increases the gain of signals. However, these amplifiers induce nonlinear effects.

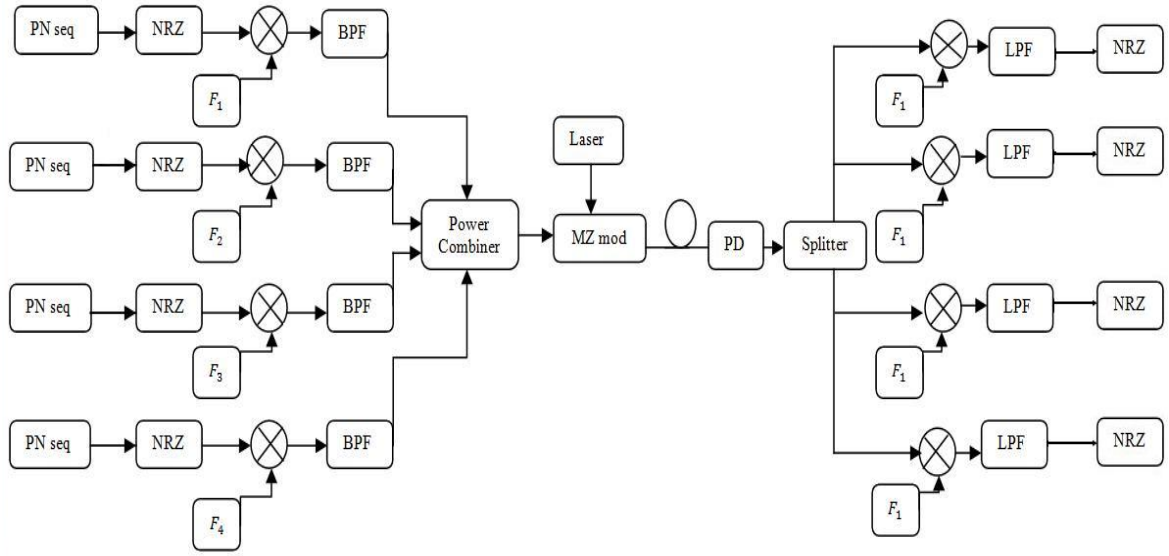
In this paper we have developed a link of SCM with NRZ (non return to zero) as line coding and Binary Phase Shift Keying (BPSK) modulation for mixing of RF signals of different frequencies with 10 Gb/s high speed digital signals. Here this paper, we compare these three dispersion compensation methods and evaluate the performance characteristics for dispersive optical communication systems. In the literature [5], post-compensation method was discussed calculated BER for all 4 channels. Here, the results of symmetrical compensation method are compared with pre- and post-compensation methods on the basis of important additional features like bit error rate, Q factor. In Section 2, the optical simulated project and parameters are defined. In Section 3 comparison results have been reported for these compensation methods and finally in Section 4, conclusions are drawn.

II. THE SYSTEM CONFIGURATION

A. Simulation Analysis

The block diagram of a basic architecture of the subcarrier multiplexing with NRZ code is shown in figure 1. The System is designed by using Optisys ver. 10 which is used in many optical fiber simulations. The bit rate in all four channels is 10 Gb/s. The transmission medium is ITU-T G.652 standard single mode fiber. The attenuation and dispersion were set at 0.25dB/km and 18ps/nm-km, respectively. The performances of the SCM with pre-, post-, pre-post method were characterized by referring BER, eye diagrams and Quality Factor (Q). The basic diagrams for all three schemes are also shown in figure 2.

.



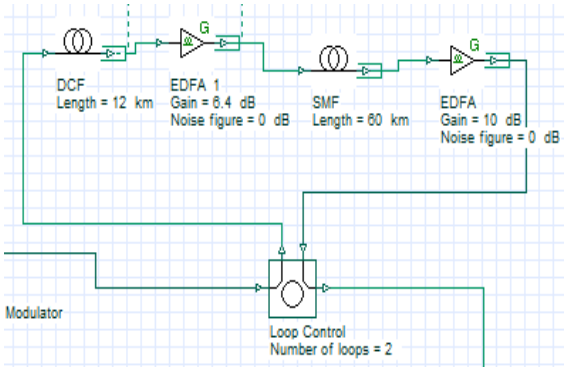
In generated link transmitter consist of NRZ coder with 10 Gb/s of data, microwave mixer, BPF, combiner and optical modulator (MZM). The receiver consists of photo detector, splitter, microwave mixer, LPF. At the transmitter, data with independent unipolar digital signal is mixed by a different microwave carrier (f_i). The subcarriers are combined and

optically modulated using an optical electrical modulator (OEM). Then n modulated code sequences are multiplexed together and transmitted through the optical fiber. Then, the decoded signal is detected by the photo detector. A splitter and an electrical Band Pass Filter (BPF) are used to split the subcarrier multiplexed signals and reject unwanted signals, respectively. In order to recover the original transmitted data, the incoming signal is electrically mixed with a microwave frequency f_i; and filtered using LPF.

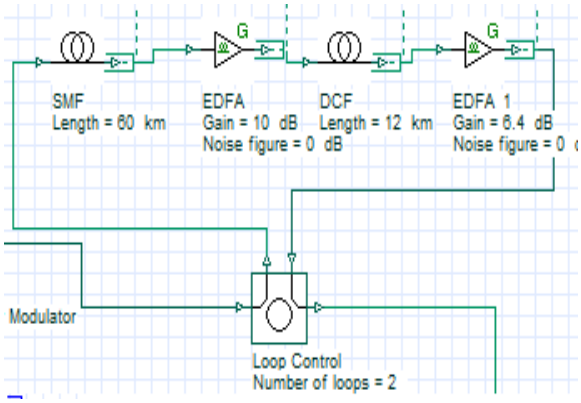
Compensation is done by three methods, pre-, post- and symmetrical compensation. In the first method, the optical communication system is pre compensated by the dispersion compensated fiber of negative dispersion against the standard fiber. In the second method, the optical communication system is post compensated by

the dispersion compensated fiber of negative dispersion against the standard fiber. In the third method, the optical communication system is symmetrically compensated by two dispersion compensated fibers of negative dispersion against the standard fiber in between. Due to the nonlinear nature of propagation, system performance depends upon power levels [6] and the position of dispersion compensated fibers [7]. Dispersion compensation fibers are specially designed fiber with negative dispersion. To compensate positive dispersion over large length of fiber high value of negative dispersion is used.

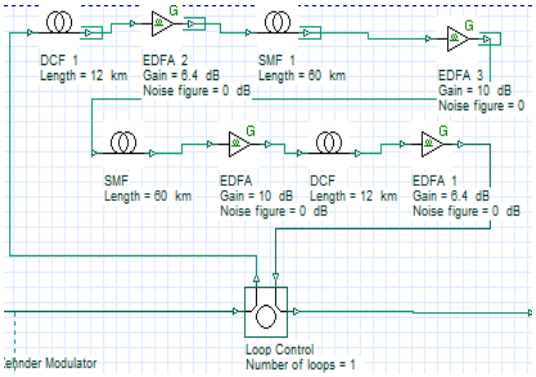
Spans made of single mode fibers and dispersion Compensated fibers are good candidates for long distance transmission as their high local dispersion is known to reduce the phase matching giving rise to four waves mixing in SCM system. The simulation setup of all three schemes is created by using software shown in figure 3.



(a)



(b)



(c)

Figure 3: Simulation setup: (a) pre-compensation with 60 km link with 2 spans (b) post-compensation with 60 km link with 2 spans (c) pre-post compensation with 120 km link with single span using single mode and dispersion compensating fibers.

B. Theoretical Analysis

In the analysis of this SCM system, we see the most important parameter is chirp v in such type of externally modulated system as it indicate the phase of the output signal modulator. Thus MZ modulators (MZM) can be designed to operate completely chirp-free, in most devices there is a small residual chirp arising from an asymmetry in the overlap of the electric fields at each electrode. A dual electrode MZM, which allows access to both electrodes, can be used to achieve a variable chirp parameter. For such a device, the chirp parameter can be related to the relative amplitude and sign of the RF drive signals to each electrode VI and V2 as [8].

$$V_1 = \frac{V_1 + V_2}{V_1 - V_2} \qquad (1)$$

The next parameter is α which is ratio of phase of amplitude modulation an defined as

$$\alpha = \frac{1}{2S}\frac{\left(\frac{d\varphi}{dt}\right)}{\left(\frac{dS}{dt}\right)} \qquad (2)$$

Where α and S are the instantaneous phase and intensity at the output of the modulator. For the MZM, α is a function of the modulator depth and de-bias point, and is related to v as [9]. If the MZM has chirp, then fiber length L [2]:

$$L = \frac{c}{2D\lambda_c f_{rfN}}\left(N - \frac{2}{\pi}\arctan(\alpha)\right) \quad (3)$$

Equation (3) indicates that the fiber-link distance in externally modulated analog systems can be increased when α is large and negative.

The signal to noise ratio SNR at the output of the PD can be expressed as:

$$SNR = \frac{(R_d P_{sc})^2}{\sigma_n^2 + \sigma_{th}^2} \quad (4)$$

Here R_d the PD sensitivity, σ_{th}^2 is the variance of receiver thermal noise and σ_n^2 is the PD shot noise variance.

The bit error rate including the effect of OBI can be expressed

$$BER = 0.5erfc\frac{(R_d P_{frf})}{\sqrt{\sigma_{sh}^2 + \sigma_{th}^2 - \sigma_{obi}^2}} \quad (5)$$

Where, σ_{obi}^2 the OBI.

Table 1 summarized the typical parameters used in BER and Q factor calculation.

TABLE I

YPICAL PARAMETERS USED IN THE PERFORMANCE

Symbol	Model Parameter	Value
R_b	Data bit rate	10 Gb/s
L_{SMF}	Length of SMF	120 km
L_{DCF}	Length of DCF	24 km
D_{SMF}	Dispersion coefficient	16ps/nm/km
D_{DCF}	Dispersion coefficient	-90ps/nm/km
G_1	Gain of $EDFA_1$	20db
G_2	Gain of $EDFA_2$	12.8db
P	Power/ channel	10dbm
α	Fiber loss	0.2dbm
A_{Eff}	Effective mode area	80μm^2
Δt	PMD coefficient	0.5ps
I_d	Dark current	10nA
2Γ	Line width	5.0 MHz

III. RESULTS AND DISCUSSION

A. Simulation Result

The Figure 4, 5 and 6 show comparisions between BER of all 4 sub-carriers at different received power. In case of multiple uplink channels, the BER of each channel will be reduced because of the accumulation of OBI noise due to the numerous interferences. It is observed that the bit error rate increases with increase in the received power. For symmetrical compensation, the bit error rate is minimum indicating the best performance. For received power up to 13 dBm, the bit error rate is 10^{-18} that is acceptable but if the power is increased from 10 to 12 dBm, it increases to 10^{-9} which is a perfect BER for high data rate optical transmission. Increasing the received power further will bring the BER higher than the defined acceptable level. For post compensation method, the bit error rate is again 10^{-13} up to 10 dBm powers by increasing the power further to 12 dBm, BER goes to10^{-10} that is good enough, when we go though pre-compensation, at 12 dBm powers BER is 10^{-19}, by increasing power it up to 20 dBm it goes to 10^{-10}. The power requirement to getting an acceptable level of BER in case of pre- compensation is more that is also a demerit. So we can easily get a conclusion that BER increases more rapidly for pre-compensation as compared to post-compensation method thereby indicating that the performance of post-compensation method is better than pre- compensation [10].

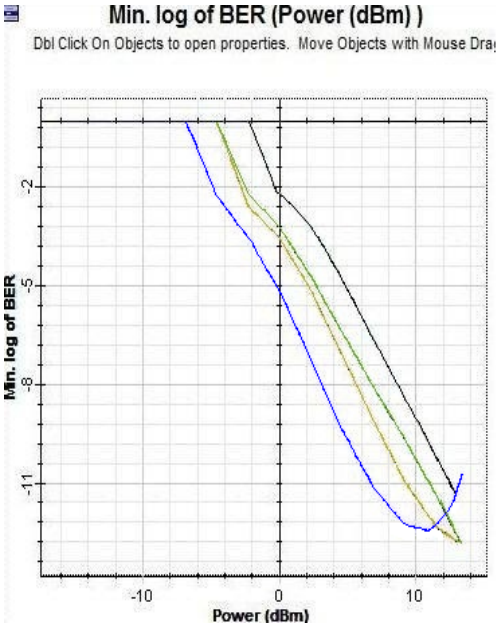


Figure 4: Bit error rate vs. Fixed received power for Post- compensation method BER= 3.30e-13

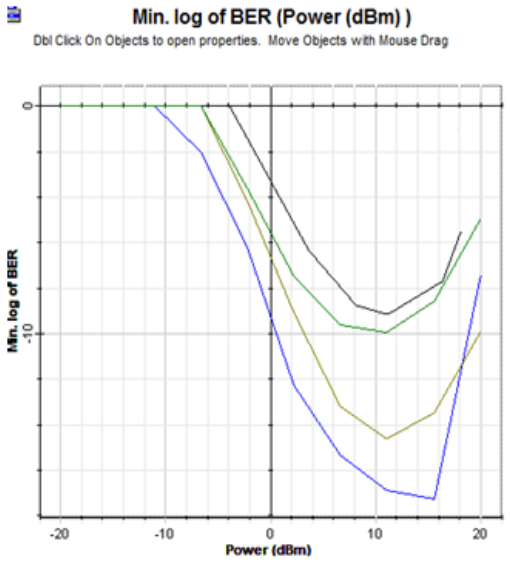


Figure 5: Bit error rate vs. Fixed received power for Pre-compensation method BER= 1.28e-18

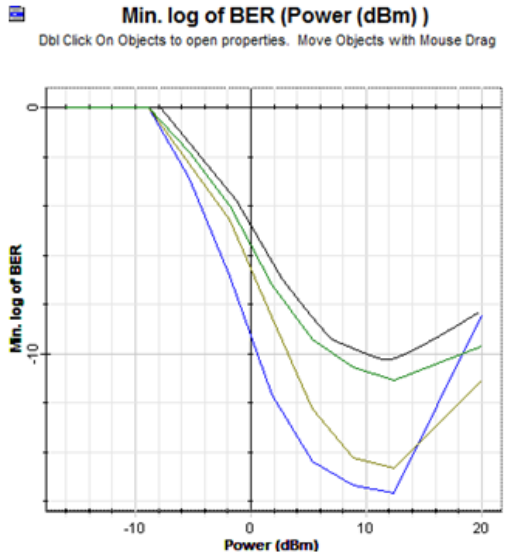


Figure 6: Bit error rate vs. Fixed received power for Pre-Post compensation method BER= 7.930e-12

In Figure 7, 8 and 9 show comparisions between Q factors of all 4 sub- carriers at different received power. As already compared BER of all three schemes Q of all three schemes goes in same manner that for post compensation Q is approximately 7 which is acceptable in range but in case of pre- compensation it goes to 9 that is not in range. In symmetrical- compensation Q factors goes to 7. So we conclude that symmetrical and post- compensation are better than pre- compensation.

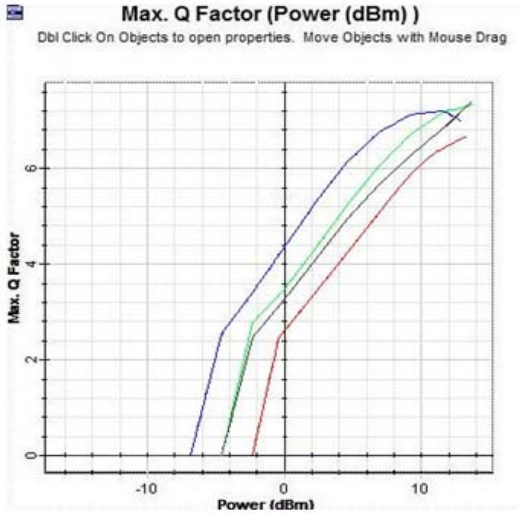


Figure 7: Q factor vs. Fixed received power for Post-compensation method Q=7.2

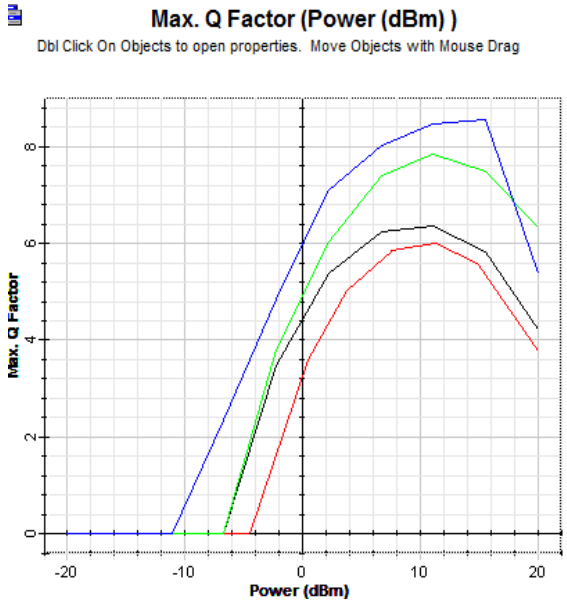


Figure 8: Q factor vs. Fixed received power for Pre-compensation method Q=9.4

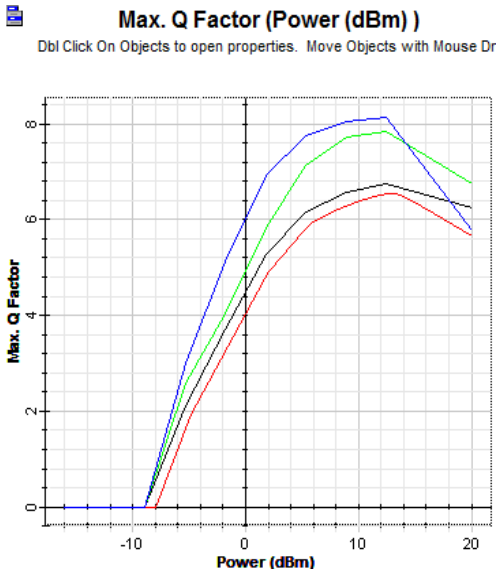


Figure 9: Q factor vs. Fixed received power for Pre-Post compensation method Q= 7

IV. CONCLUSIONS

The paper illustrate that comparisions between all three schemes pre-, post-, pre- post compensation for fiber link of 120km. and it describes that in SCM link which method is compatible. The output parameters like BER, Q factor. In simulation results, it is found that as the power increases, the bit error rate increases with non- linearity. The pre-post compensation has the best performance followed by post- and pre-compensation. As this is known that the perfect BER is 10^{-9} and Q factor is 6. So with all result we verify that in symmetrical- compensation BER minimum that is 10^{-9} but in case of both others we conclude BER is more that is not acceptable in case of SCM link with 4 RF sub-carriers. We have suggested using the same schemes for 40 Gb/s. the same technologies can be implemented with SCM/WDM system.

REFERENCES

1. R. Phillips and T. E. Darcie, "Lightwave video transmission," in *ptical Fiber Telecommunications*, I. P. Kaminon and T. L. Koch, Eds. New York: Academic, vol. IlIA , 1997.
2. P. M. Hill and R. Olshansky, "A 20-channel optical communication using subcarrier multiplexing for the transmission of digital video signals," *J. Lightwave Technol*., vol. 8, pp. 554-560, Apr. 1990.
3. [3] N. Kikuchi, S. Sasaki, and K. Sekine, "10 Gbit/s dispersion compensated transmission over 2245 km conventional fibers in a recirculating loop," *Electron. Lett.*, vol. 31, no. 5, pp. 375–377, 1995.
4. T. L. Koch, R.C. Alferness, J. *Lightwave Technol*. LT-3 (1985) 800.
5. [5] Paul D. Sargis etc. 10Gb/s subcarrier multiplexed transmission over 490km of ordinary single mode fiber without dispersion compensation. *IEEE PTL*, Dec, 1997.
6. R.M. Jopson, A.H. Gnaules, R.M. Derosier, in: *Proc.OFC'93*, paper pd3,1993.
7. A.Yariv, D.Feke, D.M.Pepper, *Opt.Lett.4* (1979) 52.
8. A. Djupsjobacka "Residual chirp in integrated optic modulators", *IEEE PhotonTech. Lett.* Vol A, pp. 41-43, Jan 1992.
9. Mohammad M. Banat, Mohsen Kavehrad. "Reduction of Optical Beat Interference in SCMlWDMA Networks Using Pseudorandom Phase Modulation." Lightwave Tech. Vol.12, No.IO, October 1994.
10. C. Peucheret, N.Hanik, R.Freund, L.Molle, P.Jeppesen, *IEEE Photon.*Technol.Lett.12 (8) (2000) 992.

◈◈◈

Sol-gel Derived Thin Films In Doped ZnO-based System and Study their V-I Characteristics

#Kiran.A, Anjan Sil & ##Kevin M. Knowles

#Department of Metallurgical and Materials Engineering,
Indian Institute of Technology Roorkee, Roorkee-247667, India
##Department of Materials Science and Metallurgy, University of Cambridge, Cambridge CB2 3QZ, UK
E-mail : kiran.a.iit@gmail.com, kmk10@cam.ac.uk

Abstract -ZnO-V_2O_5-MnO thin film has been synthesized using sol-gel process for low voltage varistor applications. The films were deposited in multi-layer depositions using spin coating technique by controlling spin coating parameters. Annealing treatment of the film was done at 700^{o}C for 1hr. The oxide films were subjected to micro-structural and XRD characterization. V-I characteristics of thin films were registered. Non-linear co-efficients (α) and breakdown voltages (V_b) were estimated from the V-I plots and values are found to be α=3.18 and 4.66 with V_b lying in the range of 22V to 25V.

Keywords*-thin film; varistor; sol-gel.*

I. INTRODUCTION

Varistors are non-ohmic resistors which are characterized by the non-linear behaviour of current and voltage. The silicon carbide bonded ceramics and zinc-oxide based ceramics are known to be the most popular varistors. However, the non-ohmic properties of silicon carbide bonded ceramics are not superior. Due to presence of excess zinc atoms in interstitial positions, zinc-oxide is non-stoichiometric which gives it semi conducting behaviour. The non-ohmic behaviour of zinc-oxide ceramics is achieved by the addition of small amount of other metal oxides such as MnO, V_2O_5, Bi_2O_3, Sb_2O_3, etc. The electrical properties can be controlled by doping or changing the grain size and the thickness of the film. Owing to rapid development of LSI and VLSI electronic circuits, there is a need to develop low voltage, lower power and highly integrated surface-mountable thin layer varistors. In this age of nano-electronics with continuously reducing size of devices and low voltage requirements of complex micro circuits, there is a huge market coming up for thin film metal oxide varistors. This is due to the fact that thin films have lower breakdown voltage with significantly high non-linear coefficient.

Various deposition techniques viz. evaporation, sputtering (RF, DC, DC Magnetron), reactive physical vapour deposition, pulsed laser deposition, chemical vapour deposition and sol-gel are available for deposition of ceramic oxide films. Uniform coating thickness can be obtained by using spin-coating technique. Due to its inexpensive nature and easy availability, spin-coating is the most widely used for the sol-gel synthesis of thin films. Moreover, the sol-gel method is suitable for realizing doped oxide film formation and films for adequate thickness.

The purpose of this work is to develop a ZnO-based ceramic thin film varistor in a system ZnO-0.5mol%V_2O_5-1.5mol%MnO by sol-gel method and study electrical characteristics viz. V-I plots for their applications as low-voltage varistors, to be applied in microelectronic devices as electrical surge arrestor.

II. EXPERIMENTAL

The flow chart showing the steps followed in the synthesis of ZnO-based ceramic thin film (ZnO: MnO : V_2O_5 = 98 : 1.5 : 0.5) is given in Fig. 1. ZnO-MnO-V_2O_5 (with molar ratio 98: 1.5 : 0.5) thin film is prepared over a gold coated ITO glass substrate using the sol-gel process. The sols are prepared by zinc acetate dehydrate (Sigma-Aldrich, >98% purity), manganous acetate tetrahydrate (Aldrich, >99% purity), ammonium metavanadate (Sigma-Aldrich, >99% purity), and the solvent 2-methoxtethanol (Aldrich, 99.8% purity). The dopants (manganous acetate tetrahydrate and ammonium metavanadate) are dissolved in 2-methoxyethanol with the addition of ethanolamine (Aldrich, >99.5% purity), by stirring the solution for 20 mins at 70-80^{o}C. Similarly the zinc acetate dihydrate is

also dissolved in 2-methoxtethanol with the addition of ethanolamine, followed by stirring. The molar ratio of ethanolamine to zinc acetate is maintained at 1.0. The resulting sol is prepared at a concentration of 1.5M/litre. The dopants solution are added to zinc acetate solution in such a way that the relative ratio of the component oxides formed after annealing becomes 98 : 1.5 : 0.5 (ZnO : MnO : V_2O_5). The solution is then stirred for 24hrs at 70-80^oC to obtain a clear, stable and homogenous sol. After this, the sol is subjected to ageing for around 24hrs for checking the stability. Ethylene glycol (Rankem, 98% purity) is the added to the final sol drop wise to increase the viscosity.

Gold-coated ITO glass is used as the substrate. ZnO-MnO-V_2O_5 ceramic films are prepared on the gold-coated ITO substrate by repeated spin coating, using spin coater (spinNXG-P1, Apex India Ltd) at room temperature, with the substrate spinning rate of 2500 rpm. After each layer deposition, the film is heated in air at 190-195°C for 20 min. After 25 layers of deposition, the film is then subjected to annealing in air at 700°C for 1hr. Voltage-Current characteristics of the thin film varistor is taken using source meter (Keithley 2400 source meter), at different locations by applying two different voltages upto 20V and 25V, and are plotted.

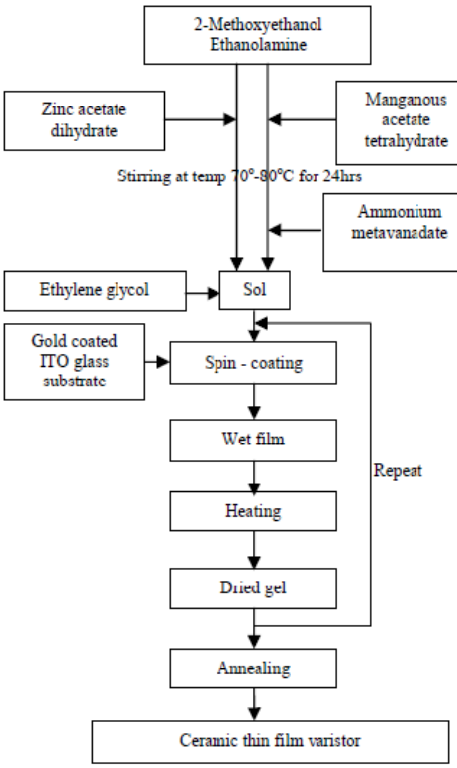


Fig. 1. Flowchart showing the sol-gel process of ZnO based thin film

The films synthesized after annealing were characterized using XRD (BRUKER D8 X-Ray Automatic Diffractometer) using CuKα radiation. Micro-structural analysis on the oxide film surface was done by FESEM using SEM (FEI Quanta 200F, Netherlands operating at 20kV). EDAX analysis at the surface was done to estimate the chemical composition on the surface.

III. RESULTS AND DISCUSSIONS

A. ZXRD pattern analysis

XRD pattern of the ZnO-MnO-V_2O_5 thin film varistor annealed at 700^oC in air for 1hr is shown in Fig. 2. The main peak is of Au (Gold) at 2θ=38.4426, which is coming out from the substrate. Because of this, other mains peaks of Zn, Mn, V and its compounds are not clearly visible, since its intensity is comparatively less. These peaks are visible in Fig. 3, where Au peak is removed.

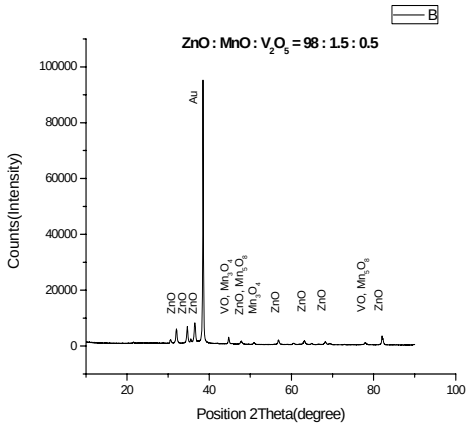


Fig. 2.XRD pattern of ZnO-MnO-V_2O_5 thin film varistor

The main purpose of editing the XRD pattern is to show the phases which are obtained, once crystallinity has taken place. From Fig. 3, it is clear that the main peaks which corresponds to that of ZnO, MnO, etc. and its compounds are coming at their standard peak positions. The strong and sharp ZnO peaks clearly shows that sufficient crystallinity has taken place. The secondary phases which are formed, like $ZnMnO_3$, are also visible from the XRD pattern (Fig. 3).

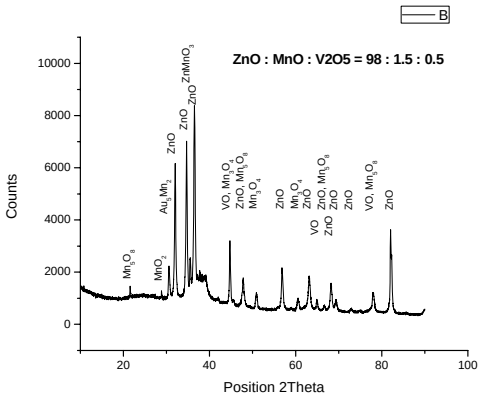


Fig. 3. XRD pattern of ZnO-MnO-V_2O_5 thin film varistor after removing the peak at 2θ=38.4426

B. Microstructure analysis using FESEM and EDAX result

The SEM micrographs of the surface of the film with different magnifications are shown in Fig .4. The EDAX analysis for the same samples is shown in Fig .5. SEM micrographs show that the film is formed uniformly over the substrate. EDAX data shows that the elemental composition of the analysed area which shows Zn as the major constituent, followed by O, V and Mn in the decreasing order. It is clear from the EDAX data that Zn is dominating the elemental composition of the oxide film surface as compared to dopants which are present. Vanadium percentage is also coming higher than Manganese.

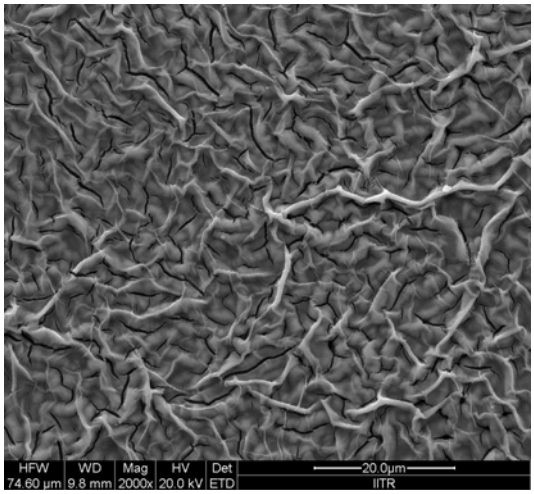


Fig .4(a) 2000X magnification

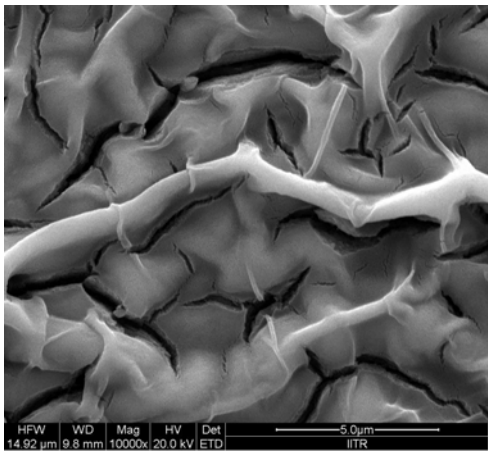
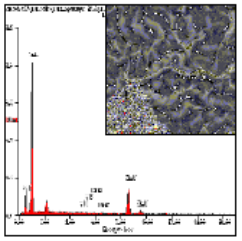
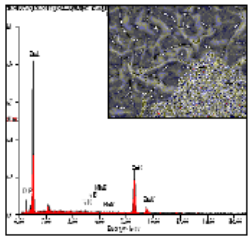


Fig .4(b) 10000X magnification



Element	Wt%	At%
OK	13.67	39.10
VK	01.56	01.40
MnK	01.30	01.08
ZnK	83.47	58.42
Matrix	Correctio n	ZAF



Element	Wt%	At%
OK	08.04	26.20
VK	01.33	01.36
MnK	01.02	00.97
ZnK	89.61	71.47
Matrix	Correctio n	ZAF

Fig .5 EDAX data of thin film varistor

V-I characteristics

Voltage and current readings have been taken from the thin film varistor with the help of source meter. These are plotted in a graph and plots which are obtained are given below (Fig .6). Two set of readings have been taken by applying two different voltages, up to 20V and 25V, that too at different locations. The non-linear V-I response which are expected is obtained.

The non linear coefficient (α) is calculated from the non-ohmic region of the plots. The relationship between voltage and current of a varistor is given as $I \propto V^{\alpha}$, where α is the co-efficient of non-linearity. The nonlinearity of a varistor is characterized by the nonlinear coefficient, α, and it is calculated as,

$\alpha = [\ln (I_2) - \ln (I_1)]/[\ln (V_2) - \ln (V_1)]$

The 'α' value is calculated at different locations. Here the graphs of two locations are shown, where we got almost same 'α' values, corresponding to both up to 20V and 25V.

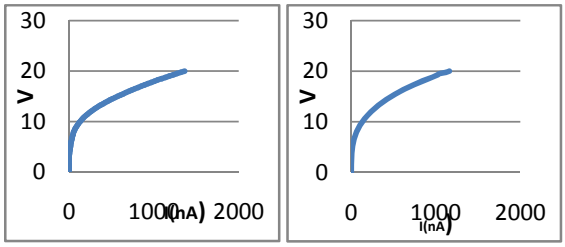


Fig .6(a). V-I plots at location 1 and 2 by applying voltage up to 20V

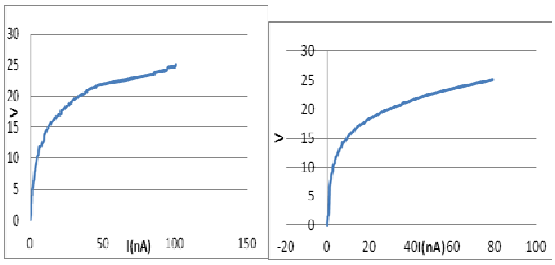


Fig .6(b). V-I plots at location 1 and 2 by applying voltage up to 25V

IV. CONCLUSION

ZnO-MnO-V_2O_5 ceramic thin film varistor has been synthesized using sol-gel technique. XRD has shown all the expected main peaks. It showed good non-linear voltage-current characteristics, which is an expected nature of varistors. Non-linearity co-efficient value obtained is α=3.1788, when voltage is given up to 20V and α=4.6617, when voltage is given up to 25V.

REFERENCES

[1] Michio Matsuoka, "Non-ohmic behaviour of ZnO ceramics", Japanese Journal of Applied Physics, 1971, 10 (6), 736-746.

[2] Hng, H. H. and Knowles, K. M., "Microstructure and current– voltage characteristics of multicomponent vanadium-doped zinc oxide varistors", Journal of American Ceramic Society, 2000, 83, 2455-2462.

[3] Yanqiu Q. Huang, Liu Meidong, Zeng Yike, Li Churong, Xia Donglin, Liu Shaobo, "Preparation and properties of ZnO-based ceramic films for low-voltage varistors by novel sol-gel process", Materials Science and Engineering, 2001, B86, 232-236.

[4] Sheng-Yuan Chu, Tser-Min Yan, Shen-Li Chen. "Analysis of ZnO varistors prepared by the sol-gel method", Ceramics International, 2000, 26, 733-737.

[5] Heriberto Pfeiffer, Kevin M. Knowles, "Effects of vanadium and manganese concentrations on the composition, structure and electrical properties of ZnO-rich MnO_2–V_2O_5–ZnO varistors", Journal of the European Ceramic Society, 2004, 24, 1199-1203.

[6] Hng, H. H. and Knowles, K. M., "Microstructure and current– voltage characteristics of multicomponent vanadium-doped zinc oxide varistors", Journal of American Ceramic Society, 2000, 83, 2455-2462.

◈◈◈

Basic Process Calculations, Simulation of Dispersed Phase of Material in Drying and Detail Discussion of Control Strategies

Shrishail S.Kattimani & S.D.Agashe

Department of Instrumentation & Control, COE, Pune, India
E-mail : shri420k@gmail.com & sda.instru@coep.ac.in

Abstract **-** Before starting with constructing and solving a specific dryer model it is recommended to classify the methods, so typical cases can easily be identified. By identifying cases and simple calculation we can easily calculate the utility requirement of our system. Also one can interpret the control of evaporation rate of material by simulation of dispersed phase. After the dryer has been designed, it must be automated to the point where it can work with minimum supervision [1].

Index Terms*- spray dryer model, dryer instrumentation.*

I. INTRODUCTION

Spray dryers have been drawing significant attention from researchers and developers in the community of process and instrumentation as far as its efficiency and control is concerned. They are expected to address various traditionally challenging problems and introduce many new applications, as spray drying has become preferred drying technology as far as powder manufacturing is concerned [2]. A spray dryer typically consists of an atomizer, a spray chamber, cyclone separator, feed tank and a heating media. When the material is atomized in atomizer small droplets of material dispersed into spray chamber so as to create the dispersed phase. Due to heating media the moisture content in small droplets in dispersed phase gets evaporated and droplet area shrinks. By removing moisture content in material powder is collected through separator. After designing dryer for its instrumentation manipulated variable is necessary, so the scope of the work is to propose the inlet heating media flow rate or feed concentration or feed flow rate can be used as manipulated variable to achieve evaporation rate at particular outlet temperature and this can be shown by simple dryer calculation and simulation using matlab and comsol multiphysics.

DRYER MODEL FORMULATION*:* When trying to derive a model of a dryer we first have to identify a volume of space that will represent a dryer. If a dryer or a whole system is composed of many such volumes, a separate sub model will have to be built for each volume and the models connected together by streams exchanged between them. Each stream entering the volume must be identified with parameters .Basically for systems under constant pressure it is enough to describe each stream by the name of the component (humid gas, wet solid, condensate, etc.), its flow rate, moisture content , and temperature. All heat and other energy fluxes must also be identified.

Balance equations: They represent Nature's laws of conservation.

Input – Output = Generation -Loss

(1) Mass Balance :

$$L_sX_1 + G_sY_1 = L_sX_2 + G_SY_2$$

(2) Enthalpy Equation :

$$L_sH_{L1} + G_sH_1 = L_sH_{L2} + G_SH_2$$

Calculation of H_{L1}, H_{L2}, H_1 for solving Enthalpy balance equation

$$H_{L1} = (C_p + X_1 * C_{pw}) * t_1$$

$$H_{L2} = (C_p + X_2 * C_{pw}) * t_2$$

$$H_1 = (C_{pp} + Y_1 * C_{pw}) + Y_1 * Le$$

Solving above equation we can calculate H_2. And by Solving Mass Balance and Enthalpy Balance simultaneously we can calculate Y_2 .

Calculation for Efficiency of dryer

Humidity of ambient air:

$$h_{amb} = Pr / (T_1 * R)$$

Specific heat of humid air:

$$C_h = C_{pp} + (h_{amb} * C_{pw})$$

Inlet heat:

$$H_{in} = (((C_{pp} * G) + (h_{amb} * G * C_{pv})) * (T_1 - T_{amb})) + ((m * X_1 * C_{pw}) + (m - (m * X_1) * C_{pm}) * (t1 - T_{amb}))$$

Dry Solid Mass:

$$m_d = m - (m * X_1)$$

Water in dry solid leaving:

$$m_w = X_2 * m_d / C_{pp} + X_2$$

Water evaporated from solid into gas:

$$S = (m * X_1) - m_w$$

Outlet heat:

$$H_{out} = [(Cpp * G) + (h_{amb} * G * Cpv)] * (T_2 - Tamb) + S * [Cpv * (T2 - T_{latent}) + Le + Cpw * (T_{latent} - T_{amb})] + [(m_d * C_{pi}) + (m_w * C_{pw})] * (t_2 - T_{amb})$$

Thermal Efficiency: $(H_{in}/H_{out}) * 100$

Disperse Phase Simulation in dryer:

The movement of droplets and particles in dryer is strongly affected by gas flow rate [4]. Droplets created by atomization move downward first and follow the main central gas flow. Near the bottom of dryer, the gas flow is reversed due to collision of material particles on dryer walls. Because of collision vortices created and dispersed particle moves at the outer edges of vortices. From comsol results "Fig-1", "Fig-2"we can identify that particle in dispersed phase follows the central gas jet and when they collide on walls creates vortices. We can clearly identify from the simulation result of dispersed phase shown in "Fig-3, how the material and heating media contacts in pressure nozzle for the atomization process. So as hot gas or air is use to evaporate the material concentration we can control the evaporation rate by means of controlling heating media flow rate at constant heater temperature.

Effects of Initial moisture content:

Vapor concentration and temperature of gas are important for both flow field and dispersed phase. Evaporation of solvent and low liquid temperature affects the gas phase properties. If temperature of heating media is lowered and solvent vapor concentration increased then location where the evaporation takes place in dryer changes. This is shown by changing feed concentration from a step 70% to 50% in matlab simulink shown in **"**Fig-4".Change in initial concentration has effect on feed flow rate, heat load requirement, product rate, outlet temperature. One can clearly identify from matlab simulink results shown in "Fig-5" that if we change initial moisture concentration then feed rate, heat load and production rate will change proportionally but outlet temperature will change inversely. For e.g. if initial moisture content decrease from 70% to 50% by means increasing solid content initially then in that case heat required to dry the solid feed is lowered and so outlet temperature will increase as we are keeping inlet temperature and final moisture content constant.

Effects of Feed rate:

For a constant evaporation rate by varying feed flow rate we can maitain outlet temperature constant.In "Fig-6" we varied feed flow rate from 5 kg/hr to 10 kg/hr at constant evaporation rate of 5 kg/hr without varying initial moisture content. From matlab simulink result shown in "Fig-7" one can easily predict that product rate changes in prapotion with feed rate and outlet temperature will decrease,provided evaporation rate will be constant. Feed flow rate can also have effect on evaporation rate.This can be achived by varying feed flow rate and keeping inlet temperature,outlet temperature and initial and final moisture content constant.In the open loop system shown in "Fig-8"we changed feed rate from 5 kg/hr to 10 kg/hr by keeping inlet temperaure,outlet temeprature ,initial and final moisture content constant.Variation in feed rate cause change in evaporation rate,product rate,heat load and mass of air as well. We can clearly identify from matlab simulink result shown in "Fig-9"how evaporation rate can also vary with feed flow rate if inlet temperature,oulet temperature,intial and final moisture content kept constant.

III. RESULTS AND DISCUSSION:

All results in both comsol and matlab simulink are obtained with respect to time in open loop system. We can interpret following sentences from the results, these are as follows:

1) Evaporation rate can be controlled by maintaining heating media temperature. This can be achieved by heater control, but from comsol results we can clearly identify that evaporation rate can be achieved by regulating heating media also by maintaining heater at constant temperature.

2) Change in initial moisture content also have effect on feed flow rate, product rate, outlet temperature, heat load provided that evaporation rate, final moisture content, initial temperature should be constant.

3) Increase in feed rate while keeping initial solid content, inlet temperature ,outlet temperature and

final moisture content keeping constant will cause increase in evaporation rate, mass of air required for evaporation increases

The effects of inlet air temperature, feed flow rate initial concentration and outlet air temperature are important for controlling dryer operation. From investigation, the effect of inlet air temperature on moisture content and outlet air temperature was greater than initial concentration [5]. Thus, inlet air temperature and feed flow rate selected as manipulated variable in many cases. From discussion it is clear that by controlling inlet air temperature we can control the evaporation rate. This can be done either by controlling the heater which is used to heat the media which might cause continuous on –off the heater or by the simulation results and basic calculation we can say that if we control the inlet heating media flow rate then humidity of heating air is controlled and hence the evaporation rate. For controlling outlet temperature we use feed flow rate as manipulated variable. This can be easily predicted from "Fig-11'' which we got from matlab simulink result of a spray dryer system shown in "Fig-10".This can be obtained by simple feedback loop using PID control.

IV. NOMENCLATURE

G_s Heating media flow rate Kg/hr.

$L_{s,}$ Solid flow rate Kg/hr.

$X_{1,}$ Moisture continent in the entering solid Kg/Kg

X_2 Moisture continent in the exit solid Kg/Kg

Y_1 Humidity of entering air Kg/Kg

Y_1 Humidity of entering air Kg/Kg

Y_2 Humidity of outlet air Kg/Kg

T_1 Temperature of inlet air

T_2 Temperature of outlet air

t_1 Temperature of inlet solid

t_2 Temperature of exist solid

R Gas Constant

Pr Inlet pressure of heating media.

m Mass of material

C_p Specific heat of solid J/(kg K)

C_{pw} Specific heat of water J/(kg K)

C_{pv} Specific heat of water vapor J/(kg K)

C_{pp} Specific heat of air J/(kg K)

H_{L1} Enthalpy of entering solid KJ/Kg

H_{L2} Enthalpy of exit solid KJ/Kg

H_1 Enthalpy of entering air KJ/Kg

H_2 Enthalpy of outlet air KJ/Kg

Le Latent heat of vaporization KJ/Kg

T_{amb} Wet bulb temperature as per psychometric chart

E Evaporation rate

P Product rate

Q Heat load

V. CONCLUSION:

By Simulation we can conclude that by controlling the inlet pressure of heating media, that is by heating media flow rate the evaporation rate and the location inside the dryer it take place can be changed. According to that we can choose inlet heating media as manipulated variable for dryer evaporation control and instrumentation. For controlling outlet temperature we can use feed flow rate as manipulated variable provided at constant evaporation rate.

REFERENCES:

[1] Aiken, Fred A, "Industrial dryers, Theory and Operation,'' in *American Control Conference*, pp.1669-1676, June 18-20, 1986.

[2] Keith Masters, "Applying Spray drying to customized powder manufacturing," *Chinese Journal of Chemical engineering*, issue 6, pp.744-749,2006.

[3] Alapati Surynarayna, *Mass Transfer operations.* New Delhi: New Age Publications, pp.571-576, 2002.

[4] EvangelosTsotsas, Arun S. Mujumdar, *Modern Drying technology-computational tools at different scales*, vol.1.Wiley-VCH, pp.192-196, 2007. vol. 6, no. 1, pp. 100–114, Jan. 2007.

[5] Lee Woun Tan, Taip, F.S.;Ibrahim, M.N.; Kamil,R.; "Empirical modeling and control for spray drying of orange juice powder'' in *4th international conference on Modeling, simulation and applied optimization (ICMSAO)*,pp.1-6,April 19-21,2011.

◈◈◈

Power-Delay Aware Resource Allocation in Cloud Data Centers

Kanchan Chowdhary

Stdt.Electronics and Computer Engineering Department ,IIT Roorkee, Roorkee,India
E-mail : Kanchan.besuit@gmail.com

Abstract - Rapid growth in demand for computational power has led to the creation of large scale data-centers which consume enormous amount of electrical power resulting in high operational cost and emission of carbon dioxide. Virtualization is a promising approach to reduce this power consumption by consolidating multiple virtual servers onto a smaller number of computing resources. But modern Cloud computing environments have to provide high Quality of Service for their customers resulting in the necessity to deal with power-performance trade-off. In this paper an efficient resource management policy for virtualized Cloud data centers is proposed. Power consumption by data-centers can be reduced by leveraging live migration of VMs and switch off idle nodes. Again dynamic reallocation incurs network propagation delay that can affect the operation cost. The objective is to optimize the power consumption as well as network delay.

Keywords *: Cloud Computing, Virtual Machines (VM), Quality of Service, Utilization, Dynamic Reallocation, Propagation Delay*

I. INTRODUCTION

Cloud Computing is defined by a large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet[1]. Cloud computing is a model that enables on demand access to a shared pool of configurable computing resources [2]. Clouds aim to power the next generation data centers as the enabling platform for dynamic and flexible application provisioning.

In recent years, demand for computational power created by modern compute-intensive business and scientific applications have driven IT infrastructures to grow rapidly. This lead to the creation of computing infrastructures such as large-scale data centers. However, a large scale computing infrastructure consumes enormous amount of electrical power. This result in operational costs exceeds the cost of the infrastructure in few years. For example, in 2006 the cost of electricity consumed by IT infrastructures in US was estimated as 4.5 billion dollars and tends to double by 2011[3]. This high power consumption results in overheating of the hardware devices which in turn reduces the system reliability and device lifetime along with increasing the operational cost. Another problem is significant emission of carbon dioxide, which contribute to the greenhouse effect.

There are number of ways to reduce power consumption by data centers, such as improvement of application algorithm, energy efficient hardware, terminal servers and thin client, Dynamic Voltage and Frequency Scaling [4], and virtualization of computer resources [5]. Virtualization technology allows one to consolidate several servers to one physical node as Virtual Machines (VMs). Results in reducing the amount of the hardware in use and thus improve utilization of resources. Recently emerged Cloud computing paradigm uses virtualization for providing scalable, on-demand resource provisioning over the Internet on pay-as-you-go basis [6].

Again one important requirement of the Cloud computing environment is to provide reliable quality of service (QoS), which is defined in terms of Service Level Agreement (SLA) e.g. throughput, response time. Although modern virtualization technology is capable of isolating the different workload of VMs, but since VMs are sharing the physical host some VM may not get required amount of resource when requested. So cloud providers have to deal with power-performance trade-off as aggressive consolidation lead to performance loss.

In this work reduction in power consumption is achieved by leveraging live migration of VM. Different heuristics are used for dynamic reallocation of VMs

according to current resource requirement while ensuring reliable QoS. The objective of reallocation is to minimize the number physical serving the current work, whereas idle nodes are switched off in order to decrease power consumption. Since dynamic reallocation of VM needs to transmit VMs from one host to another. This incurs network delay overhead which in turn increase the operation cost as physical nodes may be located geographically apart. Goal of this work is to minimize both power consumption as well as network delay ensuring reliable QoS. Doing a comprehensive study on this overall problem in the real world will be extremely difficult, so the best approach to study such a dynamic and massively distributed environment is through simulation. Even in simulation environment there are some challenges that need to be addressed.

1) Optimal solution for trade-off between energy saving and performance.
2) How to determine when and which VMs and where to migrate to minimize power consumption while minimizing migration overhead and ensuring SLA.
3) Development of an efficient and scalable algorithm for resources allocation.

II. SYSTEM ARCHITECTURE

In this work the underlying infrastructure is represented by a large-scale Cloud data center comprising *n* heterogeneous physical nodes. Each node has a CPU, which can be multi-core, with performance defined in Millions Instructions Per Second (MIPS). Besides that, a node is characterized by the amount of RAM and network bandwidth. Users submit requests for provisioning of m heterogeneous VMs with resource requirements defined in MIPS, amount of RAM and network bandwidth. SLA violation occurs when a VM cannot get the requested amount of resource, which may happen due to VM consolidation.

The software system architecture is layered. It consists of a dispatcher, global and local managers. The local managers reside on each physical node. They observe current utilization of node's resources. The local managers choose VMs that have to be migrated to another node under following cases.

- The utilization of some resources is close to threshold that creates a risk of SLA violation.
- The utilization is low, therefore all VMs should be reallocated to another node and idle node should be turned off.

The local manager sends to the global manager the information about the utilization of resources and VM chosen to migrate. Beside that they issue command for VM resizing, application of DVFS, turn on/off idle nodes. Each global manager is attached to a set of nodes and processes data from their local managers. The global manager continuously apply distributed version of a heuristic for semi-online multidimensional bin-packing, where bin represents physical nodes and items are VMs that have to be allocated. The decentralization removes a Single Point of Failure and improves scalability. Each dimension of an item represents the utilization of a particular resource. After obtaining allocation decision, the global mangers issue commands for live migration of VM.

As shown in figure 1, the system operation consists of the following steps:

1) New requests for VM provisioning. Users submit requests for provisioning of VMs.
2) Dispatching requests for VM provisioning. The dispatcher distributes requests among global managers.
3) Intercommunication between global managers. The global managers exchange information about utilization of resources and VMs that have to be allocated.
4) Data about utilization of resources and VMs chosen to migrate. The local managers propagate information about resource utilization and VMs chosen to migrate to the global managers.
5) Migration commands. The global managers issue VM migration commands in order to optimize current allocation.
6) Commands for VM resizing and adjusting of power states. The local managers monitor their host nodes and issue commands for VM resizing and changes in power states of nodes.
7) VM resizing, scheduling and migration actions. According to the received commands, VMM performs actual resizing and migration of VMs as well as resource scheduling.

2.1 Power-Delay Model

As proposed in this work, minimization of power consumption by data centers is accomplished by live migration and dynamic reallocation of VM from one host to another. Since hosts are connected through network this reallocation incurs some delay/cost to transmit VMs. if we try only to optimize power it may increase the transmission cost. So in this work we propose to integrate power and delay in a single equation and optimize this value instead of optimizing power and delay individually. For integrating these two dissimilar quantities we propose to use following equation (1).

$$\text{power-delay} = w \cdot P + (1 - w) \cdot D, \quad (1)$$

P is the power required to run the VM on new host and D is the network delay/cost required to transmit the VM from previously allocated host to the new host and w (<=1) is a weight factor. Now one can optimize this value to optimize power and delay together by providing the weight factor according to its requirement.

III. REALLOCATION HEURISTICS

The problem of VM allocation can be divided in two: the first part is the admission of new requests for VM provisioning and placing the VMs on hosts, whereas the second part is the optimization of the current VM allocation.

3.1. VM Placement

The first part can be seen as a bin packing problem with variable bin sizes and prices. To solve it we apply a modification of the Best Fit Decreasing (BFD) algorithm that is shown to use no more than $^{11}/_{9} \cdot OPT + 1$ bins (where OPT is the number of bins given by the optimal solution) [7]. In our modification, the Modified Best Fit Decreasing (MBFD) algorithms, we sort all VMs in

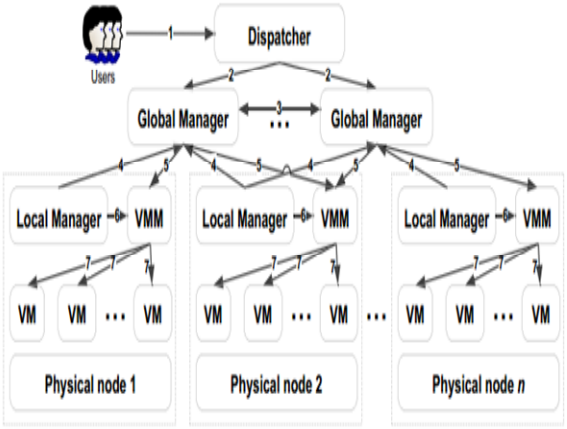


Figure 1. The System Architecture

allocate each VM to a host that provides the least increase of power-delay (PD) factor due to this allocation. This allows leveraging the heterogeneity of resources by choosing the most power-efficient nodes first. The pseudo-code for the algorithm is presented in Algorithm 1. The complexity of the allocation part is n • m, where n is the number of VM that have to be allocated and m is the number of hosts.

Algorithm 1: Modified Best Fit Decreasing (MBFD)

```
1  Input: hostList, vmList  Output: allocation of VM
2  vmList.sortDecreasingUtilization()
3  for each vm in vmList do
4      minPD ← MAX
5      allocatedHost ← NULL
6      for each host in hostList do
7          if host has enough resource for vm then
8              power ← estimatedPower(host,vm)
9              delay ← estimatedDelay(host, preHost)
10             PD ← w*power + (1-w)*delay
11             if PD < minPD then
12                 allocatedHost ← host
13                 minPD ← PD
14     if allocatedHost ≠ NULL then
15         allocate vm to allocatedHost
16 return allocation
```

3.2. VM Selection 9

The optimization of the current VM allocation is carried out in two steps: at the first step we select VMs that need to be migrated, at the second step the chosen VMs are placed on the hosts using the MBFD algorithm. To determine when and which VMs should be migrated, we introduce three double-threshold VM selection policies. The basic idea is to set upper and lower utilization thresholds for hosts and keep the total utilization of the CPU by all the VMs allocated to the host between these thresholds. If the CPU utilization of a host falls below the lower threshold, all VMs have to be migrated from this host and the host has to be switched to the sleep mode in order to eliminate the idle power consumption. If the utilization exceeds the upper threshold, some VMs have to be migrated from the host to reduce the utilization. The aim is to preserve free resources in order to prevent SLA violations due to the consolidation in cases when the utilization by VMs increases. The difference between the old and new placements forms a set of VMs that have to be reallocated. The new placement is achieved using live migration of VMs [8]. In the following sections we discuss the proposed VM selection policies.

3.2.1 The Minimization of Migrations Policy

The Minimization of Migrations (MM) policy selects the minimum number of VMs needed to migrate from a host to lower the CPU utilization below the upper utilization threshold if the upper threshold is violated.

The pseudo-code for the MM algorithm for the over- utilization case is presented in Algorithm 2. The algorithm sorts the list of VMs in the decreasing order of the CPU utilization. Then, it repeatedly looks

through the list of VMs and finds a VM that is the best to migrate from the host. The best VM is the one that satisfies two conditions. First, the VM should have the utilization higher than the difference between the host's overall utilization and the upper utilization threshold. Second, if the VM is migrated from the host, the difference between the upper threshold and the new utilization is the minimum across the values provided by all the VMs. If there is no such a VM, the algorithm selects the VM with the highest utilization, removes it from the list of VMs, and proceeds to a new iteration. The algorithm stops when the new utilization of the host is below the upper utilization threshold. The complexity of the algorithm is proportional to the product of the number of over-utilized hosts and the number of VMs allocated to these hosts.

3.2.2. The Highest Potential Growth Policy

When the upper threshold is violated, the Highest Potential Growth (HPG) policy migrates VMs that have the lowest usage of the CPU relatively to the CPU capacity defined by the VM parameters in order to minimize the potential increase of the host's utilization and prevent an SLA violation. The pseudo-code for HPG for over-utilization is presented in Algorithm 3. The algorithm sorts the list of VM in increasing order of their relative utilization. Here relative utilization is ratio of current utilization to the maximum utilization vm has requested for. Then migrate VMs from this sorted list until utilization of host comes below the upper threshold.

Algorithm 2 : Minimization of Migration (MM)

```
1  Input: hostList     Output: migrationList
2  for each host h in hostList do
3      vmList← h.getvmList()
4      vmList.sortIncreasingUtilization()
5      hUtil ← h.getUtil()
6      bestFitUtil← MAX
7      while hUtil > THRESH_UP do
8          for each vm in vmList do
9              if vm.getUtil() > hUtil – THRESH_UP then
10                 hUtil ← hUtil – vm.getUtil()
11                 migrationList.add(vm)
12                 vmList.remove(vm)
13                 falg← 1
14             else
15                 tempVm ← vm
16         if flag == 0 then
17             hUtil ← hUtil – tempVm.Util
18             migrationList.add(tempVm)
19             vmList.remove(tempVm)
20     if hUtil < THRESH_LOW then
21         migrationList.add(h.getVmList())
22         vmList.remove(h.getVmList())
23 return migrationList
```

Algorithm 3 : Highest Potential Growth (HPG)

```
1  Input: hostList     Output: migrationList
2  for each host h in hostList do
3      vmList← h.getvmList()
4      hUtil ← h.getUtil()
5      vmList.sortIncreasingRelativeUtilization()
6      if hUtil > THREAH_UP then
7          for each vm in vmList do
8              hUtil← hUtil – vm.getUtil()
9              migrationList.add(vm)
10             vmList.remove(vm)
11             if hUtil ≤ THRESH_UP then
12                 break
13     if hUtil < THRESH_LOW then
14         migrationList.add(h.getvmList())
15         vmList.remove(h.getvmList())
16 return migrationList
```

3.2.3. The Random Choice Policy

The Random Choice (RC) policy relies on a random selection of a number of VMs needed to decrease the CPU utilization by a host below the upper utilization threshold.

We are not providing pseudo-code for RC because it is similar to HPG.

IV. CONCLUSION

In this work an efficient resource management policy for virtualized Cloud data centers is proposed. The proposed technique reduces power consumption by data-centers by leveraging live migration of and dynamic reallocation of VMs from one host to another. Since dynamic reallocation incurs network propagation delay that can affect the operation cost, the proposed power-delay model takes into consideration both power required to run VM on new host and network delay overhead in transferring VM from old host to new host. The objective is to optimize the power consumption as well as network delay (with relative importance determined by weighing factors). Future work includes integrating the proposed VM reallocation policy in simulated Cloud environment and its performance comparison with commonly used allocation policies – Minimization of Migration, Highest Potential Growth, and Random Choice.

REFRENCES

[1] Foster I., Zhao Y., Raicu I., Lu S., "Cloud Computing and Grid Computing 360-Degree Compared", *Grid Computing Environments Workshop*, GCE'08,12-16 Nov.2008

[2] A nist notional definition of cloud computing, http://www.csrc.nist.gov/groups/SNS/cloudcomputing.

[3] R. Brown et al., "Report to congress on server and data center energy efficiency: Public law 109-431," Lawrence Berkeley National Laboratory,2008.

[4] Semeraro G., Magklis G., Balasubramonian R., Al-bonesi D., Dwarkadas S., and Scott M., "Energy-efficient processor design using multiple clock domains with dynamic voltage and frequency scaling", *in Proceedings of the 8th International Symposium on High-Performance Computer Architecture*, 2002, pp. 29–42.

[5] Barham P., Dragovic B., Fraser K., Hand S., Harris T., Ho A., Neugebauer R., Pratt I., and Warfield A., "Xen and the art of virtualization," *in Proceedings of the 19th ACM symposium on Operating systems principles*, 2003, p. 177.

[6] Buyya R., Yeo C., and Venugopal S., "Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities," *in Proceedings of HPCS. IEEE CS Press, Los Alamitos*, CA, USA, 2008.

[7] Yue M., A simple proof of the inequality FFD (L)< 11/9 OPT (L)+ 1,for all l for the FFD bin-packing algorithm, Acta Mathematicae Applicatae Sinica (English Series) 7 (4) (1991) 321–331.

[8] Clark C., Fraser K., Hand S., Hansen J., Jul E., Limpach C., Pratt I., Warfield A., Live migration of virtual machines, in: *Proceedings of the 2nd Symposium on Networked Systems Design and Implementation (NSDI 2005)*, USENIX, Boston, MA, USA, 2005.

◈◈◈

Power Synchronization Control of VSC-HVDC Transmission For Weak AC System

Seena. K. R & Sindhu.T.K

Dept. of Electrical Engineering, National Institute Of Technology Calicut, Kerala – 673 601
E-mail : krseena@gmail.com & tk_sindhu@nitc.ac.in

Abstract **-** In this paper voltage source converter based HVDC transmission system is used for connecting two ac systems. The control method used is power synchronization control. This method is different from other control methods and it uses the internal synchronization mechanism in ac systems. It is applied for all grid connected VSC's specially for HVDC application. This control method gives strong voltage support to a weak ac system. It shows that the proposed control allows 0.86 p.u power to be transferred from a system with short circuit ratio of 1.2 to a system with an SCR of 1.The result is compared with the vector current control for the same ac system where it can transfer only 0.4 p.u. The simulations in MATLAB/Simulink are done to demonstrate the system and observe the system behavior under three phase AC faults.

Keywords: *HVDC,* ***Voltage source converter, power synchronization control, Short circuit ratio.***

I. INTRODUCTION

High voltage DC transmission is a high power electronics technology used in electric power systems. It is an efficient, economic and flexible method to transmit large amounts of electrical power over long distances by overhead transmission lines or underground/submarine cables. Factors such as improved transient stability, dynamic damping of electrical system oscillations and possibility to interconnect two systems at different frequencies influence the selection of dc transmission over ac transmission. The HVDC transmission based on voltage source converters (VSC) is a comparatively new technology, where the valves are built by IGBTs (Insulated Gate Bipolar Transistors) and PWM (Pulse Width Modulation) is used to create the desired voltage waveform Compared to conventional line commutated HVDC systems. The principal characteristics of VSC transmission are it needs no external voltage source for commutation, it can independently control the reactive power flow at each ac network and reactive power control is independent of active power control. These features make VSC transmission technology very attractive for connecting weak ac systems, island networks, and renewable sources into a main grid.

Two existing methods for control of VSC-HVDC are power angle control and vector-current control. Power angle control is simple and straight forward to implement. In power angle control the active power is controlled by the phase angle shift between the VSC and the ac system, while the reactive power is controlled by varying the VSC voltage magnitude. The major disadvantage of power angle control is that the control system does not have the capability to limit the current flowing in to the converter.[1]

The vector current technology is the dominant control system for grid connected VSCs and it limits the current flowing in to the converter. The basic principle is to control the active power and the reactive power independently through an inner-current control loop. By using a dq decomposition technique with the grid voltage as phase reference, the inner current control loop decouples the current into d and q components. The d component of current is used to control active power or direct voltage, and the q component to control reactive power or alternating voltage. The problems with vector control are the low-frequency resonance. This can interfere with the fast inner current control loop, thereby limiting the VSC control performance. All VSCs connected to the ac system uses a PLL to obtain an accurate synchronization. But the PLL dynamics might have a negative impact on the performance of VSC-HVDC in weak ac-system connections. The power synchronization control is the combination of power angle control and vector current control. In this no PLL is needed under normal system conditions. The problems such as the resonant peak at grid frequency and converter over-current limitation are properly treated in the proposed control.[1]

II. SYSTEM OUTLINE AND CONTROL

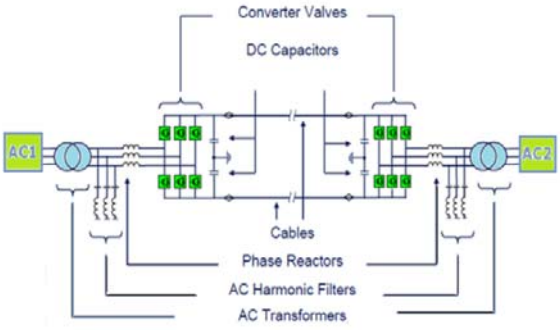


Fig.1 VSC-HVDC main circuit diagram

The schematic diagram of a VSC-HVDC transmission system connected between two ac systems is shown in fig.1.The ac source is a stiff constant frequency voltage source. The mathematical model of the VSC-HVDC can be written as shown below.[2]

$$L_c \frac{di_c}{dt} = v - u_f - R_c i_c - j\omega_1 L_c i_c \quad (1)$$

$$C_f \frac{du_f}{dt} = i_c - i_g - j\omega_1 C_f u_f \quad (2)$$

$$L_g \frac{di_g}{dt} = u_f - E - R_g i_g - j\omega_1 L_g i_g \quad (3)$$

Where L_c = The inductance of the phase reactor

Rc =Resistance of the phase reactor

E=Voltage vector of ac source

u_f=Voltage vector of filter bus

v=Voltage vector across VSC

Lg=Inductance of ac system

Rg=Resistance of ac system

P=Active power

Q=Reactive power

i_c =Current vector of phase reactor

i_g=Current vector to the ac source

ω_1 =Angular velocity of the ac system

The short circuit ratio (SCR) is defined as the ratio of the short circuit capacity of the ac system at the filter bus and the rated power of the HVDC link.

$$SCR = \frac{S_{ac}}{P_{dN}} \quad (4)$$

Where S_{ac} is the short circuit capacity of the ac system at the filter bus and P_{dN} is the rated power of the HVDC link.

Weak ac systems are those having SCR between 3.0 and 2.0, whereas very weak systems have a value lower than 2.0. The SCR can be regarded as an important reference for the design and operation of the HVDC system. It may cause the following problems in the HVDC system when connected to weak ac system: high temporary overvoltage, voltage instability, risk for harmonic oscillation and harmonic instability. [2]

III. CONCEPT OF POWER SYNCHRONIZATION CONTROL

The basic principle of power-synchronization control is similar to the operation of a synchronous machine. In an ac system the synchronous machines maintain synchronism by transient power transfer. This power transfer involves a current which is unknown and is determined by the interconnecting network. Here the VSCs synchronize with the ac system through the active-power control instead of using a phase-locked loop (PLL) [1] (1)

The active power output from the VSC is directly controlled by the power-synchronization loop and the reactive power (or alternating voltage) is controlled by adjusting the magnitude of the voltage. The power synchronization control law for VSCs is now proposed as[1]

$$\frac{d\Delta\theta}{dt} = K_p (P_{ref} - P) \quad (5)$$

Where

P_{ref} is the reference for the active power

P is the measured active power output,

$K_p$is the controller gain and $\Delta\theta$ is the output of the controller.

Here $\Delta\theta$ directly provides synchronization for the VSC. During normal operation an additional PLL is not necessary. The dynamic process of a VSC using power-synchronization control is very similar to that of interconnected synchronous machine. The transmitted

power is increased or decreased by shifting the output voltage phasors of the VSC forwards or backwards. A simplified schematic diagram of the GSVSC control using power synchronization control system is shown in Fig.2.

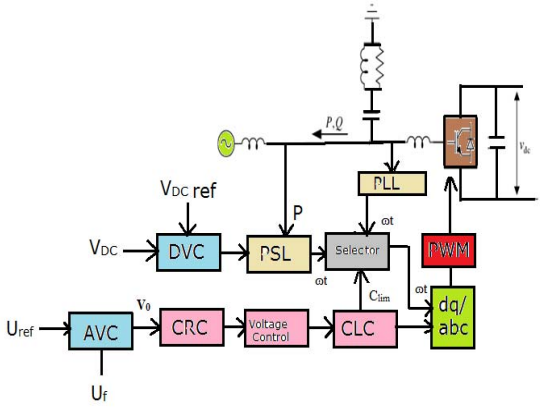


Fig.2 The schematic diagram of the GSVSC control using power synchronization control

A. Power Synchronization Control Loop(PSL)

PSL maintains synchronism between the VSC and the ac system. It is the active power control loop. The control law is given by equation 5. The power control error is converted to a frequency deviation, which is then integrated to an angle increment. The output of the PSL is used to transform the voltage reference from the converter dq frame to the stationary frame. The block diagram of PSL is shown in fig.3[1]

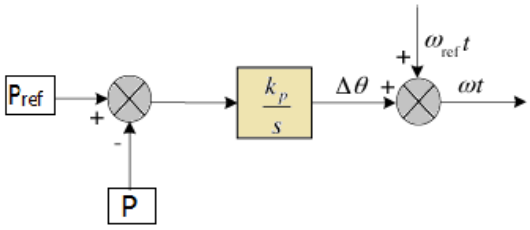


Fig.3. Power Synchronization Control Loop

B. Direct Voltage Control (DVC)

GSVSC has to keep the DC voltage constant to keep the balanced active power flow between the two sides. The control law is given as

$$P_{ref} = \left(K_{pd} + \frac{K_{id}}{s} \right) \frac{V^{ref}{}_{dc} - V_{dc}{}^2}{2} \qquad (6)$$

C. Alternating-voltage control (AVC)

The function of alternating-voltage control mode is to keep the point of common coupling (PCC) voltage constant and control the active power to/from the PCC.

The control law is given by

$$\Delta V = \frac{K_u}{s}(U_{ref} - U_f) \qquad (7)$$

D. Current Limitation Controller(CLC)

The current limitation controller is designed in such a way that under normal condition the voltage control law is

$$V^c_{ref} = V_0 - H_{HP}(s) i^c \qquad (8)$$

Where $H_{HP}(s)$ is a high pass filter

$$H_{HP}(s) = \frac{K_v S}{s + \alpha_v} \qquad (9)$$

During severe ac system fault if converter current exceeds the maximum current rating of the converter the control system has to be switched to vector current control given by

$$V^c_{ref} = \alpha_c L_c (i^c_{ref} - i^c) + j\omega_1 L_c i^c + u^c_f \qquad (10)$$

And the reference value of the converter current I^c_{ref} is calculated by

$$i^c_{ref} = \frac{1}{\alpha_c L_c} [V_0 - u^c_f - j\omega_1 L_c i^c - H_{HP}(s) i^c] + i \qquad (11)$$

During severe ac-system faults the VSC is blocked. The control system has to limit the current flowing into the converter valve. So the control system has to switch to a back up PLL during fault condition and after the fault is cleared, it has to switch back to PSL.[1]

IV. SIMULATION STUDIES

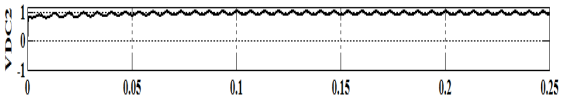
The VSC-HVDC transmission system model along with its controls is simulated in MATLAB/Simulink. The simulation set up contains a 400 kV ac system with a short circuit ratio of 1.0 at the receiving end (inverter station) while the sending end is having an SCR of 1.5.

A. Steady State Condition of VSC-HVDC system

Under normal condition the VSC-HVDC system transferring power between two ac system using power

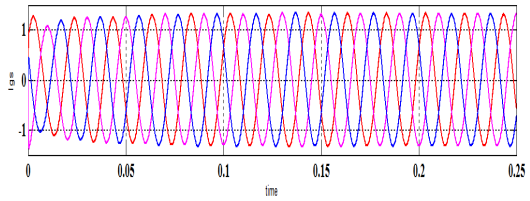
synchronization control is shown below. The DC voltage is set at 1 p.u and the active power is set at 0.8 p.u. The Grid side (Inverter side) Ac voltage, current, active power and the reference value of converter current are shown in fig.4.

(a)Grid side DC Voltage in P.U

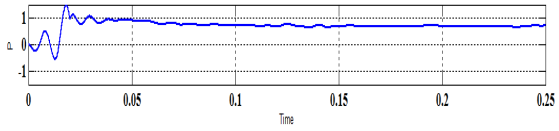


(b).Grid side AC voltage in P.U

(c). Grid side AC Current in p.u



(d). Active Power in p.u at the receiving end



(e). Converter reference current

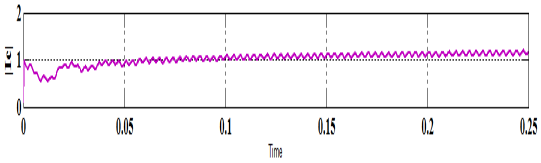
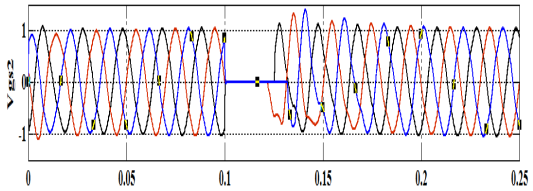


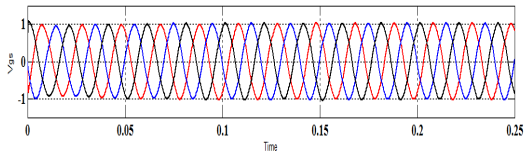
Fig.4. (a)DC Voltage. (b) AC voltage. (c)Current (d)Active Power and (e) converter reference current at the Inverter side (Grid side) during normal operation of the system using power synchronization control.

B. Three-phase ac system fault at the Grid side

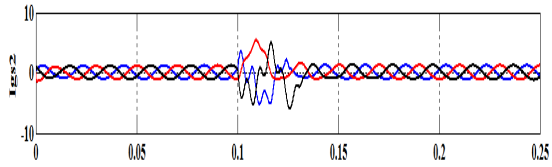
The VSC-HVDC link is supposed to ride through ac-system faults without relying on communication between the two converter stations. In fig.5 a three phase ac fault is occurs at 0.1 s for a duration of 0.02s at the grid side. From the fig.5 it can be seen that the Dc voltage is increased to 1.4 and it is going to the normal value soon. Also during fault time the converter current Ic is increased to high value but the controller reduces the converter current to 1 p.u.

(a). AC voltage in p.u at the grid side during fault

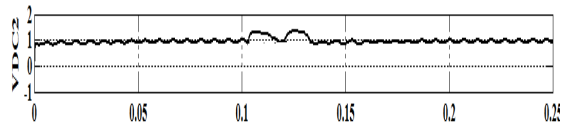




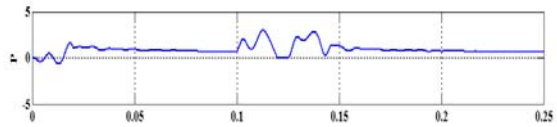
(b). AC Current in p.u at the grid side during fault



(c). DC voltage at the grid side in p.u during fault



(d).Active power in p.u at the grid side during fault



(d). Converter reference current at the grid side during fault

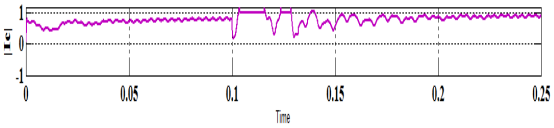
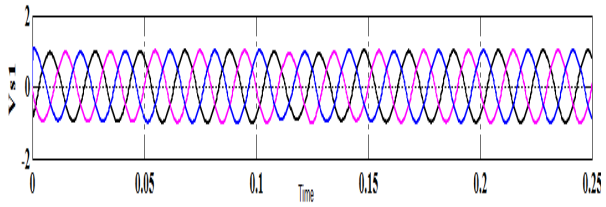


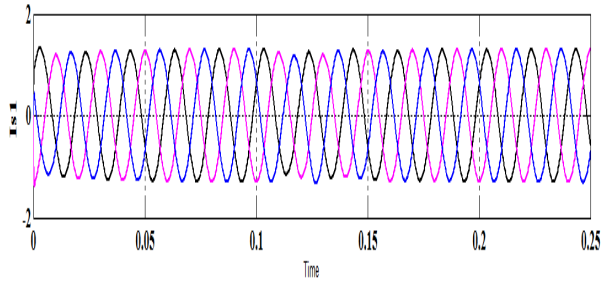
Fig.5(a)AC voltage.(b)Current(c)Dc voltage(d)Active Power and (e)converter reference current at the Inverter side (Grid side) during fault time.

The effect of fault at the sending end is shown in fig.6.The effect of fault at the grid side is less at the sending end.

(a). AC voltage in p.u at sending end during fault



(b)AC Current in p.u at sending end during fault



(c)Active Power at the rectifier side during fault

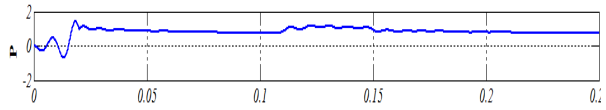
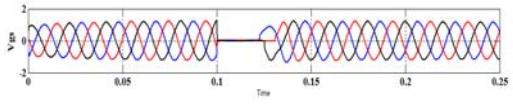


Fig.6(a)AC voltage.(b)Current(c)Active Power at the rectifier side (sending end side) when a three phase fault occurs at the inverter side while using power synchronization control.

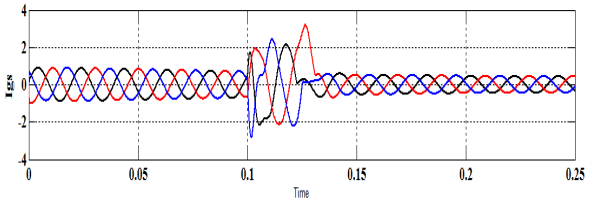
C. Comparison with the vector current control

To make comparison with the vector current control the fault analysis is shown in fig.7.Here 0.86 p.u power is transferred from a system with short circuit ratio of 1.2 to a system with an SCR of 1.The result is compared with the vector current control for the same ac system where it can transfer only 0.4 p.u.

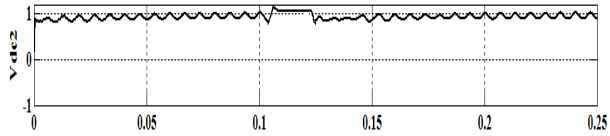
(a). AC voltage in p.u at the grid side during fault



(c). AC current in p.u at the grid side during fault



(c). DC voltage in p.u at the grid side during fault



(d). Active power in p.u at the grid side during fault

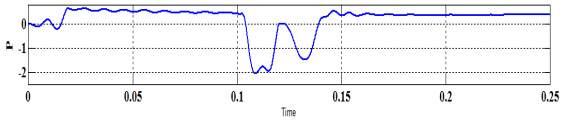


Fig.7 (a)AC voltage.(b)Current(c)Dc voltage and(d)Active Power (Grid side) during fault time for vector current control.

In vector current control also the DC voltage is set at 1 p.u. fault occurs at 0.1 p.u for 0.02 second. Here the active power transfer is only o.4 p.u approximately. During fault time the variation of active power Dc voltage, ac voltage and current is shown in fig.7.

IV. CONCLUSION

In this paper power synchronization control of VSC-HVDC is done for connecting two weak ac systems. From the simulation results it can be concluded that the system response is fast. It shows a very stiff DC link control. The control strategy is a combination of Vector current control and Power angle control. The control strategy is working well for keeping the grid side voltage constant. It can be seen that the very important property of power synchronization control is current limiting capability during ac system fault.

REFFERENCES

[1.] Lidong zhang,Lennart Harnefors,Hans peter," Power synchronization control of grid connected voltage source converters"IEEE trans.Power Syst.vol.25,no.2,pp.809 820,May 2010.

[2.] Lidong zhang,Lennart Harnefors,Hans peter",Interconnection of Two very weak Ac system by VSC-HVDC Links using Power-

Synchronization control" IEEE trans,Power systems,voi.26,no.1,pp.976-984,Feb 2011.

[3]. .Lidong zhang,Lennart Harnefors,Hans peter" Modeling and control of VSC-HVDC links connected to island system"IEEE trans.Power Syst.vol.26,no.2,pp.783-791,May 2011.

[4]. Lidong zhang,Lennart Harnefors,Hans peter,"Analysis of stability Limitations of a VSC-HVDC link using Power-Synchronization Control", IEEE trans.Power Syst.vol.26,no.3,August 2011.

[5]. Yi wang,Xiaorong Zhu,Lie xu," Contribution of VSC-HVDC connected Wind farms to grid frequency Regulation and Power damping" 36 th Annual conference on IEEE industrial electronics society,IECON 2010 .

[6.] B. R. Andersen, L. Xu, P. Horton, and P. Cartwright,"Topologies for VSC transmission," *Inst. Elect. Eng. Power Eng. J.*, vol. 16, no. 3, pp. 142–150, Jun. 2002

◈◈◈

H_∞BasedObserver for Disturbance Compensation in Decoupled TRMS Using LMI Optimization

Lekshmi S & Jeevamma Jacob

Dept. of Electrical Engineering , National Institute of Technology, Calicut
E-mail : lekshmi.eee@gmail.com & jeeva@nitc.ac.in

Abstract - Twin Rotor MIMO System is a laboratory model of helicopter. In this paper, the problem of disturbance rejection in TRMS is dealt with. Using disturbance observers,without any additional sensors is an attractive method to attenuate theeffects of disturbancesas they are highly cost effective. This method uses a simple form of DOBs, which doesnot need to solve the plant model inverse, and uses H_∞control method using LMIs to design the Q-filter in the DOB. The estimation capability of DOB is verified using simulation results in frequency domain as well as in time domain.

Keywords- *Twin Rotor; TRMS; Disturbance Ohserver; Disturbance Rejection; linear matrix inequalities;* H_∞ *control.*

I. INTRODUCTION

paper presents an application of the H_∞ control framework using linear matrix inequalities (LMIs) to a Twin Rotor MIMO System (TRMS), which is a laboratory sized helicopter model, for achieving disturbance rejection using observers.

The helicopter is easily subjected to external disturbances in the form of unexpected air currents, vibration etc. The structural vibrations occurring due to the presence of rotor load at the end of the cantilever beam and motor torque, induces a bending movement, while in operation. These vibrations induce oscillations with long settling timein the system response. Thus it is important to reject these incoming disturbances to ensure smooth operation of the TRMS. However, there haven't been sufficient researches in the area of disturbance rejection in TRMS. This paper is oriented in this direction.

The idea of using disturbance observer (DOB) to improve the performance of servomechanism was introduced in 1970s [1][2] where with the disturbances assumed to be generated by a linear time-invariant dynamic system and then estimated from the system measurements using Luenberger observer. The effects of the disturbances were compensated by feeding back the disturbance estimates into the system. However, the disturbance model is not always available in linear time-invariant form and the identification of disturbance model is not always easy. Subsequently, a new type of DOB that does not need the assumption that the disturbance model is linear time invariant nor require the full information of the disturbance model have been introduced [3]. However, the model of the controlled plant needs to be known accurately and must be invertible. Various ways to design, implement and represent disturbance observers were discussed in [4]. In [5] a simple disturbance observer is introduced that does not need to solve the plant model inverse. The DOB consists of two transfer functions which can be derived directly from the plant model and a Q-filter. As majority of the disturbances are of low frequency, the Q-filter is designed as a low pass filter. An H_∞ control-based method is applied to the design of Q-filter.

H_∞ methods are used in control theory to synthesize controllers achieving robust performance or stabilization. In the standard H∞ control problem the stabilizing controllers are obtained by the solution of a set of Riccatti equations [6]. But the Riccatti approaches have some inherent restrictions tending to limit the scope and performance of control systems. To get around such problems, we give an alternative to the parameterization of ***H***∞ controllers based on the LMI approach. The LMI approaches yield not only existence conditions valid for singular as well as regular problems but also characterizations of H∞ controllers leading to a convex or quasiconvex optimization problem. These resulting optimization problems can be solved numerically very efficiently using recently developed interior-point methods. [7][8].

In this paper, the design procedure of the DOB in [5] is extended to the TRMS set up. The TRMS nonlinear model obtained from the Feedback systems operation manual [9] and [10] is linearized and

decoupled using classical decoupling techniques given in [11]. In order to reduce the complexities related with controller design,the simple PID controller is used to ensure that the outputs follow the reference commands [12]. The design of DOB is done using the H∞ control method with the help of LMI algorithms in MATLAB using LMI toolbox provided with it [13].

The remainder of the paper is organized as follows: a description of the system along with the mathematical model is exposed in Section II, followed by an introduction of the disturbance rejection using DOB and the model of the DOB in section III. In Section IV an application of the DOB framework is provided for the TRMS, and simulation resultsare shown and discussed in Section V. Finally, the majorare conclusions drawn are given in Section VI.

II. SYSTEM AND MODELLING

In order to simplify the mechanical design of the system, the mathematical model employed in TRMS is designed slightly differently. Thus TRMS could be seen as a special purpose helicopter where the blades of the rotor have a fixed angle of attack and control is achieved by controlling rotor speeds. As a consequence of this, TRMS presents higher coupling between dynamics of the rigid body and dynamicsof the rotors as compared to a conventional helicopter, andyields a highly nonlinear, strongly cross coupled dynamics. The schematic diagram of TRMS is shown in Fig.1.

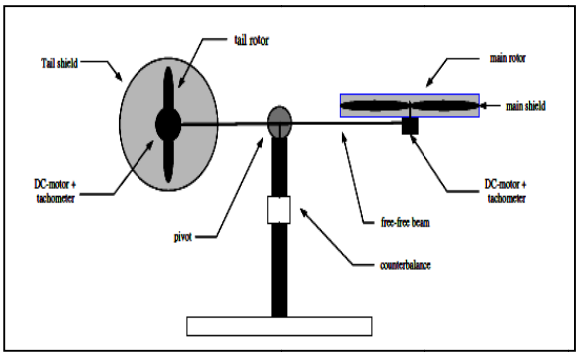


Figure 1: TRMS set up

The TRMS mechanical unit consists of a beam pivoted on its base in such a way that it can rotate freely both in its horizontal and vertical planes. At both end of a beam, there are two propellers driven by DC motors. The aerodynamic force is controlled by varying the speed of the motors. Therefore, the control inputs are the supply voltages of the DC motors. The TRMS system has main and tail rotors for generating vertical and horizontal propeller thrust. The main rotor produces a lifting force allowing the beam to rise vertically making a rotation around the pitch axis. While, the tail rotor is used to make the beam turn left or right around the yaw axis. There is a counter weight fixed to the beam and it determines a stable equilibrium position. This TRMS system has two degrees of freedom (2-DOF), the pitch and the yaw and the only mode of flight is hovering. Either the horizontal or the vertical degree of freedom can be restricted to 1 degree of freedom using the screws.

Apart from the mechanical unit, the electrical unit placed under the unit pays an important role for TRMS control. It allows for the measured signal transfer to PC and the control signal application via an I/O card. The measured signals are: position of beam in space, i.e. two position angles. The controls of the system are the motor supply voltages. A change in the voltage value results in a change of rotation speed of the propeller which results in a corresponding change in position of beam. Thus there are two inputs–horizontal and vertical motor voltages and two output – pitch and yaw angles which make TRMS a MIMO system.

TRMS MODEL

Various forces acting on the system are represented in Fig:2 [9]. According to the electrical-mechanical figure the nonlinear model of the system can be developed:

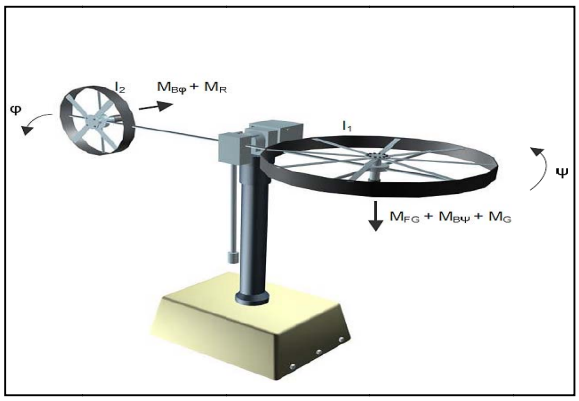


Figure 2: TRMS phenomenological model

As far as the mechanical unit is concerned, the following momentum equations can be derived for the vertical movement, where ψ represents the pitch angle and ⍰ the yaw angle.

(1)

where.

$$M_{B\psi} = B_{1\psi}.\dot{\psi} + B_{2\psi}.sign(\dot{\psi}) - frictional\ forces\ momentum$$

$$M_G = K_{gy}.M_1.\dot{\phi}.cos\psi - gyroscopic\ momentum$$

The motor and the electrical circuit are approximated by a first order transfer function. Thus in Laplace domain, the motor momentum is given by,

$$\tau_1 = \frac{k_1}{T_{11}s+T_{10}}.u_1 \quad (2)$$

Similar equations refer to the horizontal plane motion:

$$I_2.\ddot{\phi} = M_2 - M_{B\psi} - M_R \quad (3)$$

where.

$$M_2 = a_2.\tau_2^2 + b_2\tau_2 - nonlinear\ static\ characteristic$$

$$M_{B\psi} = B_{1\psi}.\dot{\psi} + B_{2\psi}.sign(\dot{\psi}) - frictional\ forces\ momentum$$

The cross reaction momentum is approximated by,

$$M_R = \frac{k_1(T_0 s+1)}{(T_p s+1)}.\tau_1$$

Again, the DC motor with electrical circuit is given by,

$$\tau_2 = \frac{k_2}{T_{21}s+T_{20}}.u_2 \quad (4)$$

Table 1shows the experimentally obtained values of the various parameters of the TRMS set up available at the Lab. Using the values the nonlinear model and later the linear model have been obtained and are given in section IV.

PARAMETERS	VALUE
I_1 – moment of inertia of vertical motor	6.8 x 10^{-2} kg m^2
I_2 – moment of inertia of horizontal motor	2 x 10^{-2} kg m^2
a_1 – static characteristic parameter	0.0135
b_1 – static characteristic parameter	0.0924
a_2 – static characteristic parameter	0.02
b_2 – static characteristic parameter	0.09
M_g – gravity momentum	0.32 N m
$B_{1\psi}$ – friction momentum function parameter	6 x 10^{-3} N m s/rad
$B_{2\psi}$ – friction momentum function parameter	1 x 10^{-3} N m s^2/rad
$B_{1\square}$ – friction momentum function parameter	1 x 10^{-1} N m s/rad
$B_{2\square}$ – friction momentum function parameter	1 x 10^{-2} N m s^2/rad
K_{gv} – gyroscopic momentum parameter	0.05 s / rad
k_1 – motor 1 gain	1.1
k_2 – motor 1 gain	0.8
T_{11} – motor 1 denominator parameter	1.1
T_{10} – motor 2 denominator parameter	1
T_{21} – motor 1 denominator parameter	1
T_{20} – motor 2 denominator parameter	1
T_p – cross reaction momentum parameter	2
T_0 – cross reaction momentum parameter	3.5
k_0 – cross reaction momentum gain	-0.2

Table 1 :TRMS Model parameters

III. DISTURBANCE REJECTION

Using disturbance observers, without any additional sensors is an attractive method to attenuate the effects of disturbances as they are highly cost effective. Disturbance Observers (DOB) are used to estimate the disturbance and disturbance estimate is injected into the designed control law to improve the disturbance rejection performance.

Several techniques exist to incorporate disturbance rejection requirements in a linear controller design. Contrary to, for example the H-infinity controller design technique where only one degree of freedom is available to obtain both disturbance rejection and performance, a DOB adds a degree of freedom, thereby enabling a separate design of the disturbance rejection and the performance. A simple form of DOB which does not

require the calculation of plant inverse is discussed in the following section [5].

A. DOB Model

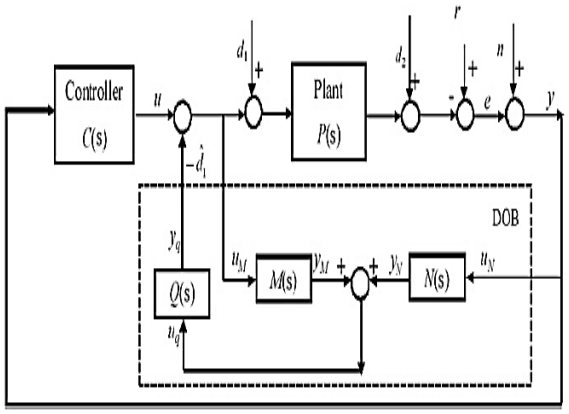


Figure3: DOB model

Fig. 3 shows the block diagram of the DOB structure [5], where C(s) is the feedback controller, P(s) is the model of the plant, d_1 is the input disturbance, d_2 output disturbance, n measurement noise. The DOB aims to compensation d_1 using its estimate$\widehat{d_1}$.

If the model of the plant P be expressed as:

$$P(s) = \frac{B(s)}{A(s)}$$

with,

$$B(s) = b_m s^m + b_{m-1} s^{m-1} + \cdots + b_0$$

$$A(s) = s^n + a_{n-1} s^{n-1} + \cdots + a_0$$

The components of the DOB M(s) and N(s) can be obtained as,

$$M(s) = \frac{B(s)}{s^m}; \qquad N(s) = \frac{A(s)}{s^n} \qquad (5)$$

Thus, the design of DOB reduces to the design of the Q-filter Q(s). The H∞ optimization method will be applied to design Q(s).

IV. APPLICATION TO TRMS

Since, the DOB model discussed in [5] is designed for a linear SISO plant, the TRMS model needs to be linearized and decoupled into two different SISO plants. PID controller is used as the controller C(s)and is tuned to ensure that the pitch and yaw are following the desired reference commands. Here the DOB is designed to compensate external disturbance,$d_2$using its estimate$\widehat{d_2}$, as its effect was found to be significantly affecting the system performance.

A. Linearization and Decoupling of TRMS

From the nonlinear differential equations given in (1) to (4), the state equations were derived after Jacobian linearization, the states are pitch angle ψ, yaw angle , pitch rate $\dot{\psi}$, yaw rate$\dot{\phi}$, main rotor torque τ_1, tail rotor torque τ_2 and cross reaction M_R. Two output variables can be measured and feedback for control purpose and the two control inputs to two DC motors are two manipulated signals. Using dynamic equations given by operation manual for TRMS [9], the state space model of linearized plant is derived as below.

$$\begin{bmatrix} \dot{\psi} \\ \dot{\phi} \\ \ddot{\psi} \\ \ddot{\phi} \\ \dot{\tau 1} \\ \dot{\tau 2} \\ \dot{MR} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ -4..7 & 0 & -0.088 & 0 & 1.35 & 0 & 0 \\ 0 & 0 & 0 & -5 & -10 & 4.5 & -50 \\ 0 & 0 & 0 & 0 & -0.909 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -0.1 & 0 & -0.5 \end{bmatrix} \begin{bmatrix} \psi \\ \phi \\ \dot{\psi} \\ \dot{\phi} \\ \tau 1 \\ \tau 2 \\ MR \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0.8 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u1 \\ u2 \end{bmatrix}$$

$$\begin{bmatrix} \psi \\ \phi \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \psi \\ \phi \\ \dot{\psi} \\ \dot{\phi} \\ \tau 1 \\ \tau 2 \\ M_R \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u1 \\ u2 \end{bmatrix} \qquad (6)$$

From the state model, the transfer matrix of TRMS can be obtained. The system can now be decoupled into two SISO subsystems using classical decoupling techniques[11]. The decoupled system has been obtained as below:

$$\mathrm{D(s)G(s)} = \begin{bmatrix} \frac{1.359}{s\hat{}3 + 0.9091\, s\hat{}2 + 4.706\, s + 4.278} & 0 \\ 0 & \frac{3.6}{s\hat{}3 + s\hat{}2} \end{bmatrix} \qquad (7)$$

whereD(s) represents the decoupler transfer matrix and G(s), the TRMS transfer matrix.

B. GeneralH_∞ control configuration

For the application of H∞ control method for the design of DOB it is required that the system is represented in the general control configuration [6]**.** Each of the SISO sub-systems can be represented as shown in the Fig4.

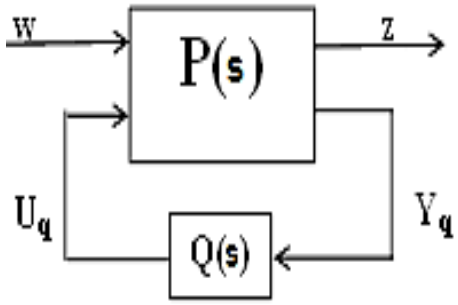


Figure: 4 : General H_∞control configuration

The exogenous input vector ***w*** includes the inputs to system such as reference signal (r), input disturbance (d_1), output disturbance (d_2) and noise (n). The error signal z represents the error signal ***e*** between the reference signal and the output.The formulation can be obtained by considering one of the SISO sub-system components alone and later the same can be extended for the other. *(The suffix 1 and 2 can be used to represents the SISO sub-systems 1 and 2 separately.)*

The state space representations of each component of the block diagram in Fig 3 can be represented as:

P(s) : (A_p, B_p, C_p, D_p)

C(s) : (A_c, B_c,C_c, D_c)

M(s) : (A_m, B_m,C_m,D_m)

N(s) : (A_n, B_n, C_n, D_n)

The state vector can be taken as:

$$x_p(t) = \begin{bmatrix} x_p(t) \\ x_c(t) \\ x_m(t) \\ x_n(t) \end{bmatrix}$$

where, $x_p(t)$, $x_c(t)$, $x_m(t)$ and $x_n(t)$ represents the corresponding states of P(s), C(s) M(s) and N(s).The augmented plant can be represented in the form of general plant configuration as:

$$\begin{bmatrix} \dot{x_p}(t) \\ e(t) \\ u_q(t) \end{bmatrix} = \begin{bmatrix} A1 & B11 & B21 \\ Cz1 & Dzw1 & Dzu1 \\ Cy1 & Dyw1 & Dyu1 \end{bmatrix} \begin{bmatrix} x_p(t) \\ w(t) \\ y_q(t) \end{bmatrix} \quad (8)$$

The component matricesof (8)can be obtained by the manipulation of the state equations of the components of block diagram.

$$A = \begin{bmatrix} Ap - BpCpDc & BpCc & 0 & 0 \\ -BcCp & Ac & 0 & 0 \\ -BmDcCp & BmCc & Am & 0 \\ -BnCp & 0 & 0 & An \end{bmatrix}$$

$$B1 = \begin{bmatrix} BpDc & Bp & -BpDc & BpDc \\ Bc & 0 & -Bc & Bc \\ BmDc & Bm & -BmDc & BmDc \\ Bn & 0 & -Bn & Bn \end{bmatrix}$$

$$B2 = \begin{bmatrix} BpDc \\ Bc \\ BmDc \\ Bn \end{bmatrix}$$

$$Cy = [-DmDcCp - DnCp \quad DmCp \quad Cm \quad Cn]$$

$$Cz = [-Cp \quad 0 \quad 0 \quad 0]$$

$$Duw = [DmDc + Dn \quad Dm \quad -(DmDc + Dn) \quad (DmDc + Dn)]$$

$$Duy = [(DmDc + Dn)] \quad Dzw = [1\ 0\ -1\ 0] \quad Dzy = -1$$

Obtaining DOB parameters

Now, the components of DOB - M(s), N(s) and Q(s) in the block diagram of Fig 3 are to be designed. The componentsM(s) and N(s) can be obtained from the plant transfer function a given in (5). The transfer functions thus obtained for the two SISO subsystems of TRMS are given below:

$$M_1(s) = \frac{8.389e{-}015\, s^2 + 3.356e{-}014\, s + 8.389e{-}015}{s^2} \quad (9)$$

$$M_2(s) = \frac{2.222e{-}014\, s^2 + 8.889e{-}014\, s + 2.222e{-}014}{s^2} \quad (10)$$

$$N_1(s) = \frac{s^3 - 3\, s^2 + 3\, s - 1}{s^3} \quad (11)$$

$$N_2(s) = \frac{s^3 - 3\, s^2 + 3\, s - 1}{s^3} \quad (12)$$

Now the design of DOB reduces to the design of the Q-filters - $Q_1(s)$ and $Q_2(s)$. They are designed based on$H\infty$ control method using LMI algorithms. Denote the transfer function from $\boldsymbol{w}$ to $\boldsymbol{e}$ as

$$T_{ew}= [T_{ed1}\ T_{ed2}\ T_{en}]$$

where T_{ed1},T_{ed2} and T_{en}representsthe transfer functions from the disturbance d_1, $d_2$and noise n to the error e.

The H∞ optimization method is applied to design Q(s) to minimize the H_∞ norm

$$\|[Ted1\ Ted2\ Ten]^T\|_\infty$$

The objective of the proposed simple DOB design can then be stated as: Given a positive scalar γ design a stable Q(s): (A_Q, B_Q, C_Q, D_Q)such that:

$$\|[Ted1\ Ted2\ Ten]^T\| < \gamma$$

The H∞ control design problem can be solved via theLMIapproach as stated in [7], [8].

V. SIMULATIONS AND RESULTS

The designed DOB was implemented along with decoupled TRMS and the simulation was done on MATLAB. The LMI algorithm was realized in MATLAB using LMI Tool Box [11] provided with it. The Q- filter transfer functions as obtained from MATLAB are as follows:

1.325e009 s^10 - 3.322e012 s^9 - 6.302e012 s^8 - 1.425e013 s^7 - 7.336e014 s^6 - 8.951e014 s^5 - 9.892e013 s^4 - 9.055e011 s^3 + 5.465e009 s^2 + 0.0003766 s+ 4.371e-019

$Q_1(s)$ = ---------- ----------

s^11 + 4.085e004 s^10 - 9.19e006 s^9 - 4.405e008 s^8 - 6.149e009 s^7 - 1.893e010 s^6 - 1.452e010 s^5 - 1.315e009 s^4 - 5.081e006 s^3 + 2.919e004 s^2 - 2.661e-009 s - 1.791e-023

s^11 + 1.841e009 s^10 + 1.095e014 s^9 + 1.18e014 s^8 + 4.441e014 s^7 + 3.964e016 s^6 + 4.518e016 s^5 + 1.552e015 s^4 + 3.045e013 s^3 - 1.687e011 s^2 + 0.2803 s + 2.261e-015

$Q_2(s)$ = ----------

s^11 - 5.812e005 s^10 + 1.08e008 s^9 - 7.923e009 s^8 - 6.385e010 s^7 + 3.977e011 s^6 + 4.696e011 s^5 + 1.642e010 s^4 + 4.424e007 s^3 - 2.125e005 s^2 + 1.139e-006 s + 9.173e-020

where $Q_1(s)$ and $Q_2(s)$ corresponds to the filter for pitch subsystem and yaw subsystem respectively. The frequency responses of the filters are given in Fig 5(a) and 5(b).The responses show that $Q_1(s)$ and $Q_2(s)$ are low pass filters as required.

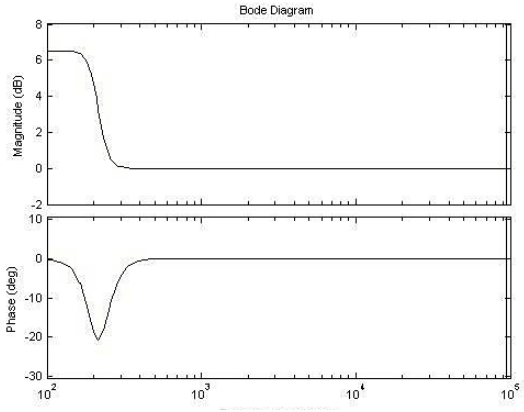


Figure 5(a) frequency response of $Q_1(s)$

The sensitivity functions corresponding to the two designed filters were obtained as given in Fig. 6(a) and 6(b). It is clear from the sensitivity function plot that the designed DOB is able to suppress the disturbances with frequency less than 30Hz. It is known that the possible incoming disturbances will be of the order of 3-5Hz. Thus it is clear that the DOB designed is successful to suppress possible incoming disturbances into the TRMS.

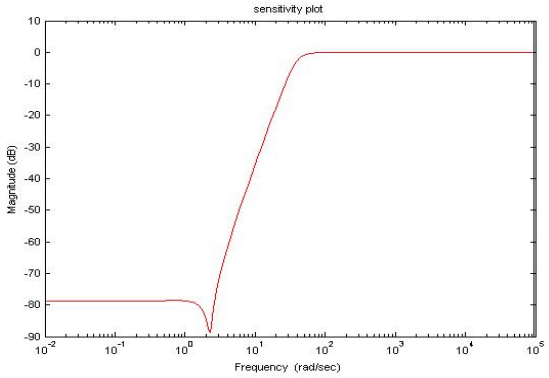


Figure 6(a) sensitivity function of $Q_1(s)$

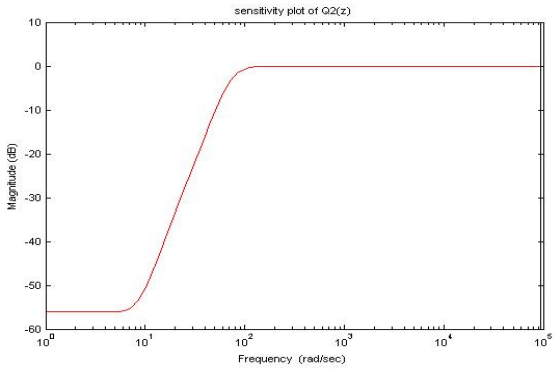


Figure6(b) sensitivity function of $Q_2(s)$

After obtaining acceptable results in frequency domain analysis, the time domain analysis of the DOB performance has been done. The designed DOB is implemented along with the TRMS model and the disturbance estimating capability was analyzed with a Gaussian white noise signal given as disturbance. The estimate obtained has been compared with the original incoming disturbance and the accuracy of the estimator was analyzed. The results obtained are shown in Fig7(a)and 7(b).

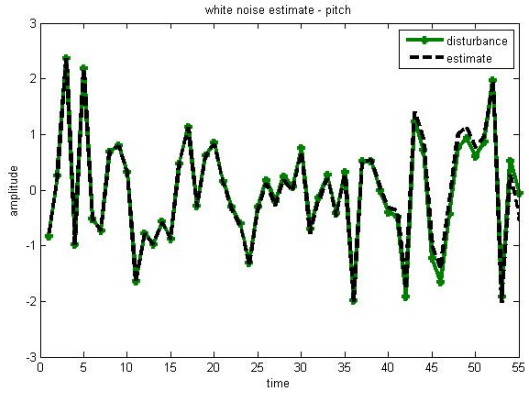


Figure7(a) disturbance estimate of the pitch

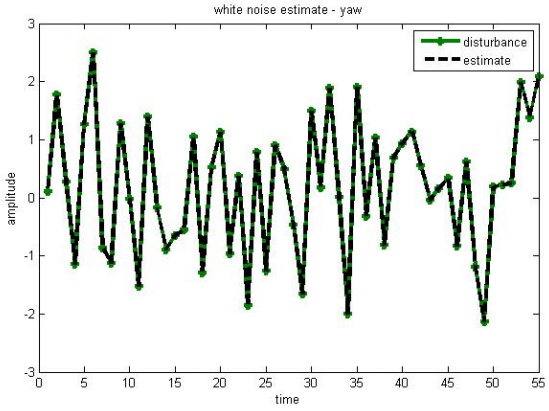


Figure7(b) disturbance estimate of the yaw

It is seen that the estimate closely follows the disturbance in both cases. Thus the designed DOB is successfully estimating the incoming disturbances which will cancel off the effects of incoming disturbance and ensure stabilized performance.

VI. CONCLUSION

The problem of disturbance rejection on TRMS has been addressed in the paper. The method of using Observers for disturbance rejection has been utilized here. A simple form of DOB has been discussed and designedbased on the H∞ control method using LMI to achieve desireddisturbance/noise rejection. The DOB does not need tosolve the plant model inverse and this benefit is of great significance especiallythus its design issimplified and this benefit is of great significance especially for non-minimum phase plant. Thesimulationresults show that theDOB designed in this paper is able to effectively improve the attenuation ofthe disturbance in frequency lower than 30Hz, and will not sacrifice the stability and performance of the nominal feedback control loop.

REFERENCES

[1] JOHNSON C.: 'Accommodation of external disturbances inlinear regulator and servomechanism problems', IEEE Trans.Autom. Control, 1971, 163, (6), pp. 635–644

[2] LEVIN V. JR., KREINDLER E.: 'Use of disturbance estimator for disturbance suppression', IEEE Trans. Autom. Control, 1976, 21, (5), pp. 776–778

[3] DAVISON E.: 'The robust decentralized control of a general servomechanism problem', IEEE Trans. Autom. Control, 1976, 21, (1), pp. 14–24

[4] Erwin Schrijvert , Johannes van Dijkt and Henk Nijmeijert: 'Equivalence of Disturbance Observer Structures for LinearSystems',IEEE Conference on Decision and Control Sydney, Australia December, 2000

[5] C. Du, H. Li C, K. Thum, F.L. Lewis, Y. Wang : ' Simple disturbance observer for disturbance compensation', IET Control Theory Appl., 2010, Vol. 4, Iss. 9

[6] Skogestad and Postlethwaite, 'Multivariable Feedback Control- Analysis and Design', John Wiley and Son,Ltd

[7] PASCAL GAHINET: ' A Linear Matrix Inequality approach to *H*∞ Control', International Journal of Robust and Nonlinear Control, vol 4, 421-448

[8] PASCAL GAHINET:'Explicit *H*∞ Controller Formulas for LMI-based Synthesis', *Auromorico, Vol.* 32, No. 7, pp. lOU7-1014, 19%

[9] Twin rotor MIMO system control experiments-33949s, TRMS manual, Feedback instruments, UK.

[10] TING-KAI LIU AND JIH-GAU JUANG: 'A Single Neuron PID Control for Twin Rotor MIMO System', IEEE/ASME International

Conference on Advanced Intelligent Mechatronics, July 14-17, 2009

[11] M.Tham : 'Multivariable control : An introduction to Decoupling Control', MTT, July 1999.

[12] Karl Johan Astrom , 'PID Controller', Control system design, 2002

[13] Pascal Gahinet, Arkudii Nemirovskii, Aim J. Laub Muhmoud Chiluli: 'The LMI Control Toolbox', Proceedings 33rd Conference *on* Decision and control Lake Buena, vista F December 1894.

◈◈◈

Image Encryption by Using Pixel Value Rotation

Honnaraju B, Manoj Kumar M,Shiva Sumanth Reddy

Abstract **-** With the fast progression of data exchange in electronic way, information security is becoming more important in data storage and transmission. Because of widely use of images in industrial process, it is important to protect the confidential image data from unauthorized access. So Encryption is used to securely transmit data in open networks. Each type of data has its own features; therefore different techniques should be used to protect confidential image data from unauthorized access. Most of the available encryption algorithms are mainly used for textual data and may not be suitable for multimedia data such as images. In thispaper, wehave introduced a block-based transformation algorithm based on the pixel value rotation of image. The number of pixel rotation is based on the random number. Result shows the some out of the encryption and the decryption.

Key words: *Encryption, Decryption, RSA, Rotation, Block-based.*

I. INTRODUCTION

THE field of encryption is becoming very important in the present era in which information security is of utmost concern. Security is an important issue in communication and storage of images, and encryption is one of the ways to ensure security. Image encryption has applications in internet communication, multimedia systems, medical imaging, telemedicine, military communication, etc.

Images are different from text. Although we may use the traditional cryptosystems to encrypt images directly, it is not a good idea for two reasons. One is that the image size is almost always much greater than that of text. Therefore, the traditional cryptosystems need much time to directly encrypt the image data. The other problem is that the decrypted text must be equal to the original text. However, this requirementis not necessary for image data. Due to the characteristic of human perception, a decrypted image containing small distortion is usually acceptable.

II. BACKGROUND

Encryption is the process of transforming the information to insure its security. With the huge growth of computer networks and the latest advances in digital technologies, a huge amount of digital data is being exchanged over various types of networks. It is often true that a large part of this information is either confidential or private. As a result, different security techniques have been used to provide the required protection. The security of digital images has become more and more important due to the rapid evolution of the Internet in the digital world today. The security of digital images has attracted more attention recently, and many different imageencryption methods have been proposed to enhance the security of these images. Image encryption techniques try to convert an image to another one that is hard to understand. On the other hand, image decryption retrieves the original image from the encrypted one. There are various image encryption systems to encrypt and decrypt data, and there is no single encryption algorithm satisfies the different image types.

Most of the algorithms specifically designed to encrypt digital images are proposed in the mid-1990s. There are two major groups of image encryption algorithms: (a) non-chaos selective methods and (b) Chaos-based selective or non-selective methods.

Most of these algorithms are designed for a specific image format compressed or uncompressed, and some of them are even format compliant. There are methods that offer light encryption (degradation), while others offer strong form of encryption. Some of the algorithms are scalable and have different modes ranging from degradation to strong encryption.

The proposed algorithm generates a random numbers in between 1 to 8 and rotates the pixel value based on the random number. The random number for each pixel is different. At present we have worked on gray scale images. Another proposed method is to divides the image into random number of blocks with predefined maximum and minimum number of pixels, and apply the above mention method to each block independently.

III. THE PROPOSED TECHNIQUE

A. Description of the Transformation Approach

The transformation technique works as follows: the

*original*image is divided into a random number of blocks. And generate the random number for each block and rotate the each pixel value by using the random number.

Random numbers are stored in the file and then encrypted and send to the receiver.

The block diagram of the proposed method is shown below:

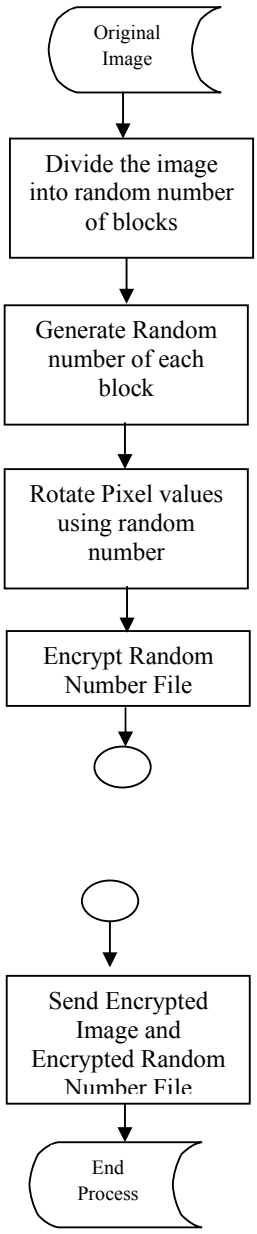


Figure 1: Flow chart Encryption Process

The function of the receiver is to decrypt the encrypted random file. The decrypted file is stored in the text file. This random file is used for the image decryption. The flow chart of the image decryption is shown in the below figure.

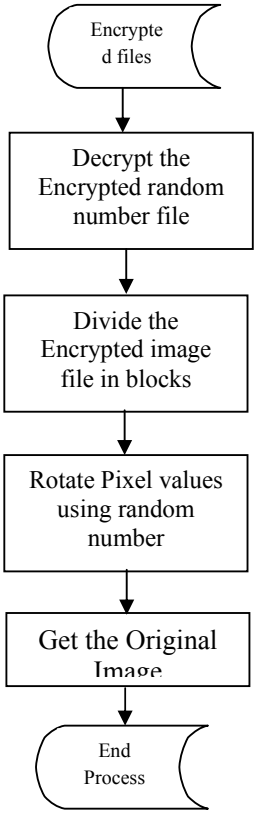


Figure 2: Flow chart for decryption process

Algorithm:

EncryptImageUsingPixelRotation()

Input: Original Image I

Output: Encrypted Image, and Random Number File

Steps:

1. Read Image file.
2. Divide the image file into number blocks.
3. Generate random number for each block
4. Rotate each pixel using the random number.

   (val>>n) | (val<<(8-n))

   Where val is the actual value of the image pixel, and n represent the number bit rotation.
5. Encrypt the random number file using RSA algorithm.

Algorithm:

DecryptImageUsingPixelRotation()

Input: Encrypted Image, and Encrypted Random number file.

Output: Decrypted Image File

Steps:

1. Read Encrypted Image File and Encrypted Random file.
2. Decrypt the random number file.
3. Decrypt the image by rotating pixel valuesby using the decrypted random number file.

```
if(Rnum == 1)
Rnum = 7;
else if(Rnum == 2)
Rnum = 6;
else if(Rnum == 3)
Rnum = 5;
else if(Rnum == 4)
Rnum = 4;
else if(Rnum == 5)
Rnum = 3;
else if(Rnum == 6)
Rnum = 2;
else if(Rnum == 7)
Rnum = 1;
elseRnum = 0;
(val>>Rnum) | (val<<(8-Rnum));
```

4. Output decrypted image.

IV. RSA SCHEME

The RSA scheme is a block cipher. Each plain text is an integer 0 to n-1 for some n, which leads to a block size <log2(n). The typical size for n is 1024 bits. The details of the RSA algorithm are described as follows

- Key Generation

1. Pick two large prime number p & q, p≠ q

2. Calculate n = p x q.

3. Calculate $\phi(n) = (p-1)(q-1)$;

4. Pick e, so that gcd(e, $\phi(n)) = 1$ $1 < e < \phi(n)$

5. Calculate d, so that $d \cdot c \; mod \; \phi(n) = 1$

i.e d is the multiplicative inverse of e in mod $\phi(n)$

6. Get public key as $k_U = \{e, n\}$;

7. Get Private key as $k_R = \{d, n\}$;

- Encryption

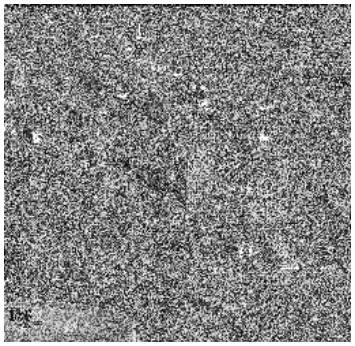
For plain text P<n, its cipher text $C = P^e mod \; n$

- Decryption

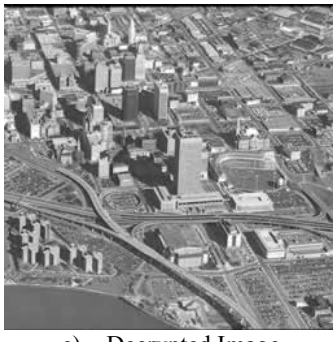
For plain text C, its plain text $P = C^d mod \; n$

V. RESULTS

a) Input Image



b) Encrypted Image



c) Decrypted Image

Figure 3: Encryption and Decryption output

a) Input Image

b) Encrypted Image

c) Decrypted Image

Figure 4: Encryption and Decryption output

VI. CONCLUSION

We have introduced Lossless Encryption for Gray scale Images using a Binary rotation of image pixel values. We divide the image into random number of blocks. Each block pixel is rotated using the single random number. We RSA encryption algorithm to encrypt the random number sequence. This algorithm is easy to implement in hardware because they operate at the binary levels. They are also suitable for multimedia applications and real time application such as mobile phone services and wireless networks.

REFERENCES

[1] National Institute of Standards and Technology, "Advanced Encryption Standards (AES), "http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf, 2001.

[2] J. Cheng; J.I. GUI, "A new chaotic key-based design for image Encryption and Decryption," The 2000 IEEE International Symposium on Circuits and Systems, 2000.Proceedings. ISCAS Geneva, vol.4, pp. 49-52, May. 2000.

[3] S. Li, C. Li, G. Chen, N. G. BourBakis, and K.-T. Lo, "A General quantitative cryptanalysis of Permutation-only Multimedia ciphers against plaintext attacks," Signal Processing: Image Communication, vol. 23, no. 3, pp.212-223, 2008.

[4] K. C. lyer and A. Subramanya, "Image Encryption by Pixel Property Separation," http://eprint.iacr.org/2009/043.pdf, Cryptology ePrint Archive, 2009.

[5] M. Ashtiyani, P. M. Birgani, and H. M. Hosseini, "Chaos- Based Medical Image Encryption Using SymmetricCryptography," in Information and Communication.

[6] M. Yang, N. Bourbakis, and S. Li, "Data-image-video Encryption," potential, IEEE, vol. 23, no. 3, pp. 28-34, 2004

[7] Y. Zhou, S. Again, V. M. Joyner, and K. Panetta, "Two Fibonacci p-code based image scrambling algorithms," in Image Processing: Algorithms and Systems VI, San Jose, CA, USA, 2008, pp. 681215-12.

◈◈◈

Biotelemetry - Wireless Transmission Of Biomedical Data

[1]Anjali Krishanlal Arora, [2]Priyanka M. Dubal[*], [3]Sayali D. Kulkarni[*], [4]Pranali P. Janbandhu[*] & [5]Jyoti Gupta[*]

Dept. of Electronics and Tele-communication, Dr. Babasaheb Ambedkar Technological University, "Vidyavihar", Lonere-402103, Tal. Mangaon, Dist. Raigad (M.S.)

Abstract **-** In hospitals numbers of patients are admitted. Some of the patients are in ICU, some in general wards, some are in the special rooms etc. Monitoring the different parameters of these patients such as temp, BP, ECG/heart rate etc. is done after particular time intervals. They go for rounds in different wards and rooms and take the different parameters and record them manually. To make a system, one slave unit will be mounted near the every patient's bed. This unit will consist of sensors like temperature sensor, Heart rate sensor etc. These values are displayed on LCD at slave unit and send to the central system. The master unit will receive the data coming from slave unit and display it on the LCD.RF transmitter receiver pair will be used for the data transfer between master card and slave card. The Buzzer indication is provided at both ends to indicate the arrival of new data or any panic condition etc. Thus it is possible to monitor the various parameters of different patients without physically going to their bed from the consulting room.

Keywords- *Microcontroller, RF Transmitter-Receiver, Biotelemetry, ECG, Sensors.*

I. INTRODUCTION

The monitoring, recording, and measuring of a living organism's basic physiological functions is known as Biotelemetry. The recording and measuring of certain vital phenomena of living organisms that are situated at a distance from the measuring device comes under Biotelemetry. The transmission of physiological data, such as electrocardiographic (ECG) and electroencephalographic (EEG) recordings, heart rate, and body temperature by radio or telephone systems. Transmission of such data uses sophisticated electronic devices developed for the study of the effects of space travel on animals and humans; it has progressed to the use of communication satellites for relaying such data from one part of the world to another part.

The most common usage for biotelemetry is in dedicated cardiac care telemetry units or step-down units in hospitals. Although virtually any physiological signal could be transmitted, application is typically limited to ECG, heart rate and body temperature.

A. History Of Biotelemetry

Historically, Einthoven, the originator of ECG i.e., Electrocardiogram, as a means of analysis of electrical activity of the Heart, transmitted ECGs from a hospital to his laboratory which was many miles away in 1903. The transmission was done via Telephone Lines. The telephone lines in this instance were merely used as Conductors for the current produced by the biopotentials.

Some of the first uses of biotelemetry systems date to the early space race, where physiological signals obtained from animals or human passengers were transmitted back to Earth for analysis. The device was manufactured in 1958 by Spacelabs Healthcare.

B. Current Trends in Biotelemetry

Because of crowding of the radio spectrum, the Federal Communications Commission (FCC) as well as similar agencies elsewhere have recently begun to allocate dedicated frequency bands for exclusive biotelemetry usage, for example, the Wireless Medical Telemetry Service (WMTS). In addition, there are many products that utilize commonly available standard radio devices such as Bluetooth.Mostly, different bands of Radio Frequencies are used, called as "Radio Telemetry".

C. *Basic Operation Principle*

The simplest application of the principle of biotelemetry is the Stethoscope. In a stethoscope, the heart beats are amplified acoustically and transmitted through a hollow tube system to be picked up by the ear of a physician for interpretation.The media of transmission changes in a biotelemetry system as the

physical parameter to be measured is changed. As described, RF bands are used.

II. COMPONENTS OF A BIOTELEMETRY SYSTEM

Mainly, two units are involved viz, Transmitting unit and Receiving Unit.

A. Biotelemetric Transmitter

The most common circuitry required for biomedical data transmitter is shown in fig below. The stages are broken in typical functional blocks.

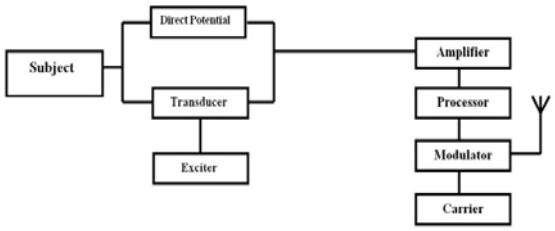


Figure 1: Block diagram for Biotelemetric Transmitter

Physiological signals are obtained from the subject by means of appropriate transducers. The signals are then passed through a stage of amplification and processing circuits. These include generation of a subcarrier and a modulation stage of transmission.

The transducers, processor circuit change according to the subject and the parameters.

A. Biotelemetric Reciever

The receiver consists of a tuner to select a transmitting frequency. A demodulator is provided to separate a signal from the carrier wave. Any device can be provided for display unit or for recording purpose. The block diagram for receiver is shown below.

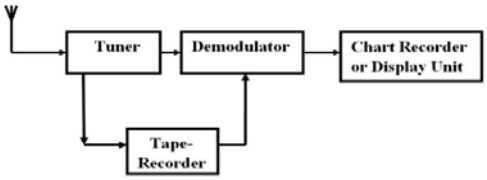


Figure 2: Block diagram of Biotelemetric Receiver and Display Unit

The received signal can also be stored in the modulated state by the use of a tape recorder, as shown in figure 2.

C. Some Terminologies Related to Biotelemetric Transmitter and Receiver

1) *Subject:-* A living organism whose physical parameter is to be measured.
2) *A Radio- Frequency Carrier Wave:-* A high frequency sinusoidal signal, propagated in the form of electromagnetic wave.
3) *Range:-* The distance in which transmitted signal can be received.
4) *Modulation:-* Information to be transmitted is impressed upon the carrier.
5) *Transmitter:-* Circuit which generates the carrier wave and modulates it.
6) *Receiver:-* Equipment capable of receiving the transmitted signal and demodulating it.
7) *Tuner:-* Selects the appropriate frequency.
8) *Display and Storage unit:-* used for storage and interpretation of received information.

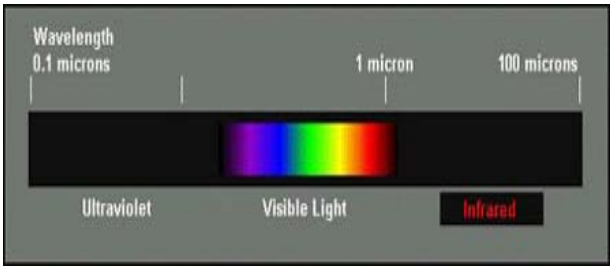
III. RFID AND MICROCONTROLLER

Biotelemetry involves storage and transmission of the data. For this purpose, it is dependent on RFID and Microcontroller. They are the basic parameters used in biotelemetry.

A. RFID?

RFID is a tracking technology used to identify and authenticate tags that are applied to any product, individual or animal. Radio frequency Identification and Detection is a general term used for technologies that make use of radio waves in order to identify objects and people.

The IR i.e. Infrared Frequency also come under RF i.e., Radio Frequency. The IR band can be shown as



Introduction:-

Purpose of Radio frequency Identification and Detection system is to facilitate data transmission through the portable device known as tag that is read with the help of **RFID** reader; and process it as per the needs of an application. Information transmitted with the help of tag offers location or identification along with other specifics of product tagged – purchase date,

color, and price. Typical **RFID** tag includes microchip with radio antenna, mounted on substrate.

The RFID tags are configured to respond and receive signals from an RFID transceiver. This allows tags to be read from a distance, unlike other forms of authentication technology. The RFID system has gained wide acceptance in businesses, and is gradually replacing the barcode system.

1) History and Key Developments:-

RFID has been around since II World War but was viewed as too limited and expensive in functionality for most of commercial use. With advancement in technology, cost of system components has reduced and capabilities have increased, making RFID more popular.

Léon Theremin invented a surveillance tool for Soviet Union in the year 1945. This tool retransmitted the incident radio waves along with audio information.

Sound waves vibrated diaphragm that altered the shape of resonator, modulating reflected sound frequencies. This tool was not identification tag but a secret listening device. But it is still considered as predecessor of the RFID technology due to it being energized, passive and stimulated by outside electromagnetic waves. Similar technology as transponder was invented in UK in the year 1915 and was regularly used by allies in the II World War for identifying aircrafts as foes or friends. The transponders are used for by powered aircrafts till date.

Invented in 1973, device by Mario Cardullo is known to be a true ancestor of the modern RFID. Initially the device was passive and was powered by interrogating signals and had transponder 16 bit memory for application as toll device. The basic patent by Cardullo covers application of RF, light and sounds as the transmission media.

Early exhibition of the reflected power RFID tags, semi passive and passive was presented by Robert Freyman, Steven Depp and Alfred Koelle. This portable system used around 12 bit tags and worked at 915 MHz. And the first patent associated with abbreviation of RFID was approved to Mr. Charles Walton in the year 1983.

4) How RFID and RF Works?

Antenna emits the radio signals to activate tag and to read as well as write information to it. Reader emits the radio waves, ranging from one to 100 inches, on the basis of used radio frequency and power output. While passing through electronic magnetic zone, RFID tag detects activation signals of readers.

Most RFID tags contain at least two parts: one is an integrated circuit for storing and processing information, modulating and demodulating a radio-frequency (RF) signal, and other specialized functions; the other is an antenna for receiving and transmitting the signal

Powered by its internal battery or by the reader signals, the tag sends radio waves back to the reader. Reader receives these waves and identifies the frequency to generate a unique ID. Reader then decodes data encoded in integrated circuit of tags and transmits it to the computers for use.

5) *Types of RFID:-*

Active and passive RFID are different technologies but are usually evaluated together. Even though both of them use the radio frequency for communication between tag and reader, means of providing power to tags is different. Active RFID makes use of battery within tag for providing continuous power to tag and radio frequency power circuitry. Passive RFID on the other hand, relies on energy of radio frequency transferred from reader to tag for powering it.

Passive RFID needs strong signals from reader but signal strength bounced from tag is at low levels. Active RFID receives low level signals by tag but it can create higher level signals to readers. This type of RFID is constantly powered, whether in or out of the reader's field. Active tags consist of external sensors for checking humidity, temperature, motion as well as other conditions.

6) RFID frequencies:-

Just like you can tune a radio in various frequencies for listening to different channels, RFID readers and tags need to be tuned in to a same frequency for communication. RFID system uses various frequencies but most common and popularly used frequency is low, high and ultra high frequency. Low frequency is around 125 KHz, high is around 13.56 MHz and ultra high varies between 860-960 MegaHz. Some applications also make use of microwave frequency of 2.45 GHz. It is imperative to choose right frequency for an application as radio waves work different at various frequencies.

7) RFID Applications:-

The role of RFID is not just confined to Aircraft identification anymore; it is also lending a hand in various commercial uses. Asset tracking is one of the most popular uses of RFID. Companies are using RFID tags on the products that might get stolen or misplaced. Almost each type of Radio frequency Identification and Detection system can be used for the purpose of asset management.

Manufacturing plants have also been using RFID from a long time now. These systems are used for

tracking parts and working in process for reduction of defects, managing production of various versions and increasing output. The technology has also been useful in the closed looped supply chains for years. More and more companies are turning to this technology for tracking shipments among the supply chain allies. Not just manufacturers but retailers also are using this RFID technology for proper placement of their products and improvements in the supply chain. RFID also plays an important role in the access and security control. The newly introduced 13.56 MHz RFID systems provide long range readings to the users. The best part is that RFID is convenient to handle and requires low maintenance at the same time. It is being used in Biotelemetry for data transmissions. As the distance changes, frequency range changes. For ex., the biomedical data of Astronauts is sent to the Earth for monitoring which uses 2.45 GHz.

B. MICROCONTROLLER

The Microcontroller is a device which is used in the biotelemetry for Storage purposes. Due to low power consumption and high performance, microcontrollers are used for memory purposes.

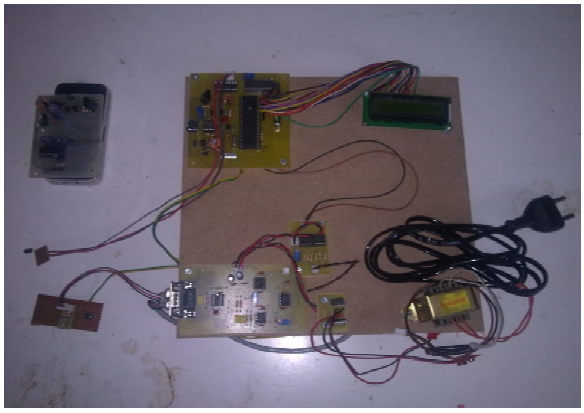
A microcontroller basically contains one or more following components:

- Central processing unit(CPU)
- Random Access Memory)(RAM)
- Read Only Memory(ROM)
- Input/output ports
- Timers and Counters
- Interrupt Controls
- Analog to digital converters
- Digital analog converters
- Serial interfacing ports
- Oscillatory circuits

The microcontroller AT89c51 is used for data storage. AT89C51 is an 8-bit microcontroller and belongs to Atmel's 8051 family. AT89C51 has 4KB of Flash programmable and erasable read only memory (PEROM) and 128 bytes of RAM. It can be erased and program to a maximum of 1000 times. As the memory is erasable, it can be used for an array also. For storage of data of various patients, the microcontroller is used. Port P_0 and P_2 are also used to provide low byte and high byte addresses, respectively, when connected to an external memory. Port 3 has multiplexed pins for special functions like serial communication, hardware interrupts, timer inputs and read/write operation from external memory. AT89C51 has an inbuilt UART for serial communication. It can be programmed to operate at different baud rates. Including two timers & hardware interrupts, it has a total of six interrupts. Multiplexed pins are provided for address lines. 8 registers are provided for memory storage, called as Register bank. Some SFRs are also in-built on-chip which performs special functions.

IV. DATA INTERPRETATION, TRANSMISSION AND RECEPTION

The hardware unit involved in Biotelemetry system may be given as follows:



The biomedical data is interpreted in following steps:

A. Data Collection:-

This is done with the use of Biosensors. The biosensors are attached to the subject. The sensors have the electrodes. They may be internal or external. They sense the potential or the related data and send to the ADC i.e., Analog to Digital Converter. This data is temporarily stored in microcontroller.

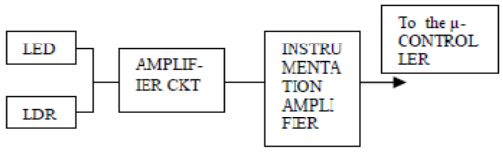


Figure 3: Pulse Rate Sensor Block Diagram

The above fig shows the block diagram for rate sensor as a mean of data collection

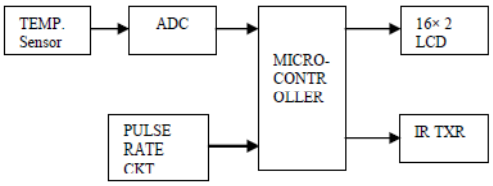


Figure 4: Slave unit for DATA COLLECTION & TRANSMISSION

B. Data Transmission:-

This step is followed using modulation and RF or IR Transmitter and Receiver. Antenna emits the radio signals to activate tag and to read as well as write information to it. Reader emits the radio waves, ranging from one to 100 inches, on the basis of used radio frequency and power output. While passing through electronic magnetic zone, RF detects activation signals of readers.

C. Data Reception & Processing:-

This is done by Master unit. The information is received by an antenna. This data is then processed within microcontroller and then given to the display and recorder.

IR RECEIVER TO THE PC FOR

Figure 5: Master unit for DATA RECEPTION AND DISPLAY

D. Data Display Unit:-

This unit is used for data display, record and management. In a hospital, the data of various patients is to be stored. Hence, DBMS, Data Base Management System is used.The LED or Buzzer arrangement may be provided for an emergency in a hospital. PC will display the appropriate value that will be measured by sensor.

V. RESULT AND DISCUSSION

The output of the Biotelemetry system can be taken on the PC in the form of the windows as follows. The following windows has the Temperature and Heart Rate of the different persons taken via the system.

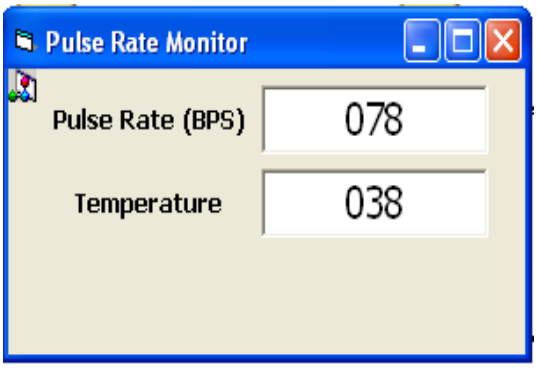


Figure 6: Output of Biotelemetry System for Person 1

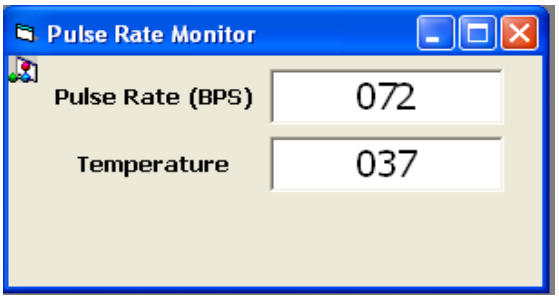


Figure 7: Output of Biotelemetry System for Person 2

The temperature of ambience is continuously measured via the temperature sensor ckt which is a temperature dependent resistor. The heart rate is measured is using LDR and LED. When the finger is placed, then heart rate is displayed. The values are measured for different persons as shown above.

VI. APPLICATIONS AND ADVANTAGES

The Biotelemetry is a very useful branch relating to Engineering and Medical. Both of these technologies, on combining, gives following advantages.

1) Radio Frequency transmissions for monitoring astronauts in the space.
2) Patient monitoring in an ambulance or in a remote area away from hospital.
3) Collection of medical data from home or office.
4) Research on rare animals in their habitat.
5) Patient monitoring where freedom of movement is desirable (Ex. ECG while exercising)
6) Continuous patient monitoring is possible.

VII. CONCLUSION

The Biotelemetry is a fast growing branch of science. Using both medical and engineering technologies, it is increasing an ease to the doctors all over the world.

It can be concluded that the measurements of different biological parameters depends on the respective sensors. This telemetric system can also be employed to other fields by changing the sensors and the RF ranges.

ACKNOWLEDGEMENT

We are heartly thankful to our guide Prof. Sanjay V. Khobragade, Ms.Punam More and Mr. Mahesh M. Kumbhar for giving us the guidance required at each

step. We are also thankful to the persons who directly or indirectly helped us in the task.

REFERENCES

[1] Webster, John G. (1998). *Medical Instrumentation: Application and Design*.

[2] IEEE- Microcontroller-based wireless recorder forbiomedical signals

[3] New Jersey: John Wiley & Sons, Inc.

[4] Lucas, Bill (1991). .An Evaluation System for Direct Interface of the MPX5100 Pressure Sensor with a Microprocessor,. Freescale Application Note AN1305

[5] Leslie Cromwell, Fred Weibell and Erich A Pfeiffer, "Biomedical Instrumentation and Measurement", PHI. http//www.atmelcorporation.com

◈◈◈

Machine Learning on DNA Sequences of Hepatitis B Virus

Anusia. S, Arifa.A & K. Latha

Department of Computer Science and Engineering, Anna University of Technology
Tiruchirappalli., India
E-mail : anusiacse@gmail.com, arianish@gmail.com, erklatha@gmail.com

Abstract - Extraction of meaningful information from large data sets is a key element in bioinformatics research. One of the challenges is to identify genomic markers in Hepatitis B virus(HBV)that are associated with HCC(Hepato Cellular Carcinoma) development by comparing the complete genomic sequences of HBV among patients with HCC and those without HCC(liver cancer). In this study, a data mining framework, which includes feature selection, classifier learning, and classification, is introduced. Our group has collected HBV DNA sequences, of genotype C from over 300 patients specifically for this project. In the feature selection process, potential markers are selected based on the combination of ReliefF and CFS(Correlation Based Feature Selection) for further classifier learning. The features ranked using ReliefF are passed as input to CFS for feature selection. This combination of methods increases the performance of the classification. Based on features chosen by these methods, error rates of several classification algorithms were obtained for analysis. Our results demonstrate the importance of feature selection in accurately classifying new samples. The results shows that the classification methods have more than 70 percent sensitivity for most data sets, which are considered high as an initial scanning method for liver cancer diagnosis.

Keywords- *HBV, Deoxyribonucleic Acid,* Attribute Relation File Format, HCC, Genetic Markers.

I. INTRODUCTION

Data mining, also known as knowledge discovery in databases, has been recognized a new area for database research. The area can be defined as efficiently discovering interesting rules from large collections of data. It is the process of analyzing data from different perspectives and summarizing it into useful information. It allows users to analyse data from many different dimensions, categorize it and summarize the relationship identified.

A. *Biological Data mining*

The past decade has seen an explosive growth in genomics, proteomics, functional genomics, and biomedical research. Examples range from the identification and comparative analysis of the genomes of human and other species (by discovering sequencing patterns, gene functions, and evolution paths) to the investigation of genetic networks and protein pathways, and the development of new pharmaceuticals and advances in cancer therapies. Biological data mining has become an essential part of a new research field called bioinformatics.DNA sequences form the foundation of the genetic codes of all living organisms. All DNA sequences are comprised of four basic building blocks, called nucleotides: adenine (A), cytosine(C), guanine (G), thymine (T). These four nucleotides (or bases) are combined to form long sequences or chains that resemble a twisted ladder. Chronic hepatitis B patients recruited since 1997 were prospectively followed up for the development of HCC. HCC was diagnosed by a combination of alpha fetoprotein, imaging, and histology. Liver cirrhosis was defined as ultrasonic features of cirrhosis together with hypersplenism, ascites, varices, and/or encephalopathy. It is reported that a high proportion of the Patients with genotype C have been shown to be positive for hepatitis B e antigen (HBeAg). It is shown from the figure.

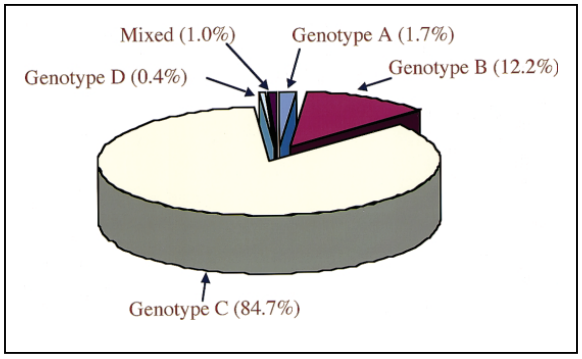


Figure 1. HBV genotype in Japan. Of 720 patients, genotype A was found in 12 patients (1.7%), genotype B in 88 (12.2%), genotype C in 610 (84.7%), genotype D

in 3 (0.4%), and the mixed genotype in 7(1.0%). In Japan, genotypes C and B are prevalent. In this study, we have DNA sequences from 150 Control (normal) patients and 150 HCC (cancer) patients. The aim of this study is to develop a data mining framework with a classification model for HBV DNA and clinical data. This Classification model should have high sensitivity for HCC diagnosis and prediction. The basic units of virus DNA are nucleotides. There are four different types of nucleotides found in DNA. The four nucleotides are given one-letter abbreviations as shorthand for the four bases, which are A, G, C and T respectively. An example of DNA strand is:

ATGGGCTAATCCTCTAATCTGG

There are 3215 units in each HBV virus DNA strand. Each unit is named as a particular site which is ordered from sites 1 to 3215.

II. STATE OF ART

The focus of this study is to identify genetic marker(s) for liver cancer (HCC) from HBV DNA sequences. One of the past researches is an HIV genomic study [1]. The researchers align each DNA sequences with a reference sequence, and then select the genes using their expert knowledge, and then use Decision Tree and Support vector Machine (SVM) for analysis. The other researchers focussed on the identification of HBV DNA sequences that are predictive of response to one therapy[2]. Some sites in sequences were observed to have caused the effect. Next, they studied the risk factors in HBV sequences with respect to medicine [3], [4].There are similar medical research reports, but all of them are focussed on the specific gene positions, proteins or part of a virus genome. This is the limitation of the existing system. Our project is the first study on the complete viral genome. The aim of this study is to develop a data mining framework which contains an appropriate classifier for liver cancer based on HBV DNA and clinical data. We identified the important mutation sites (markers) in the HBV sequences that could have caused or been related to liver cancer. We use the combination of ReliefF and CFS for finding genetic markers of HCC in the HBV genome data.

III. METHODOLOGIES

A. Feature Selection Algorithm

The main purpose of feature selection [5], [6], [7], [8] is to reduce the number of features used in classification while maintaining acceptable classification accuracy. We use a simpler algorithm based on information gain to select initial features. In our approach, information gain criterion [9] is used to find the useful features to distinguish between the Control (normal) and the HCC (cancer) groups of HBV carriers. Information gain is a common criterion for feature selection. The information gain of a feature (attribute) is the uncertainty (entropy) that can be reduced if the attribute is used for classification. Hence, the information gain should be aimed at the higher the better. Equation (1) is the entropy E of an attribute X with m values, X1, X2,. . .,Xm, and P(Xi) is the probability of the value Xi

$$E(X) = \sum - P(|X_i|) \log_2 P(X_i) \qquad (1)$$

Specific to a typical DNA classification problem, we assume that the data have K classes, C = c1, c2, ..., ck. For each aligned site position, it has m possible nucleotides V1,V2,. . . Vm. We define |ck|, k = 1, 2, . . . , K, as the number of sequences in class ck.|cki|is the number of sequences in Class ck whose character at the aligned site is Vi, which can be A, T, G, or C in our case. The Remainder of X, R(X), is defined as follow

$$R(X) = \sum_{i=0}^{m} \frac{\sum_{i=0} |C_{Ki}|}{\sum_{k=1} |C_K|} E(P(c_{1i}).....P(c_{ki})) \quad (2)$$

Information Gain IGj of the aligned site j is the difference between the original information content E(C) of the data set and the amount of information needed to classify all the unclassified data left in the data set after applying site j for classification

$$IGj = E(C) - R(j) \qquad (3)$$

The features are ranked by the information gains, and then the top-ranked features are chosen as the potential attributes used in the classifier. A site with higher information gain will contribute more in the classification and be able to distinguish more examples (cases).

B. ReliefF

A key idea of the original Relief algorithm (Kira and Rendell, 1992b), is to estimate the quality of attributes according to how well their values distinguish between instances that are near to each other. For that purpose, given a randomly selected instance Ri (line 3), Relief searches for its two nearest neighbors: one from the same class, called nearest hit H, and the other from the different class, called nearest miss M (line 4). It updates the quality estimation W[A] for all attributes A depending on their values for Ri, M, and H (lines 5 and 6). If instances Ri and H have different values of the attribute A then the attribute A separates two instances with the same class which is not desirable so we decrease the quality estimation W[A]. On the other hand if instances Ri and M have different values of the

attribute A then the attribute A separates two instances with different class values which is desirable so we increase the quality estimation W[A]. The whole process is repeated for m times, where m is a user-defined parameter.

C. *Correlation Based Feature selection*

CFS is a simple filter algorithm that ranks feature subsets according to a correlation based heuristic evaluation function. The bias of the evaluation function is toward subsets that contain features that are highly correlated with the class and uncorrelated with each other. Irrelevant features should be ignored because they will have low correlation with the class.

Redundant features should be screened out as they will be highly correlated with one or more of the remaining features. The acceptance of a feature will depend on the extent to which it predicts classes in areas of the instance space not already predicted by other features. CFS's feature subset evaluation function (Equation 4.2) is repeated here (with slightly modified notation) for ease of reference:

$$M_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

Where MS is the heuristic "merit" of a feature subset S containing k features, rcf is the mean feature-class correlation (f 2 S), and rff is the average feature-feature intercorrelation. The numerator of Equation can be thought of as providing an indication of how predictive of the class a set of features are; the denominator of how much redundancy there is among the features.

D. *Rule Learning*

Rule learning tries to learn rules from a set of training data (examples). It can be modeled as a search problem of finding the best rules that classify the training examples with minimum error. However, the search space can be very large; a robust search algorithm is required. Here, Generic Genetic Programming (GGP), which is a type of the Evolutionary Algorithms (EA), is adopted as our search and optimization algorithm. First, a population is initialized by generating individuals (a set of rules) randomly. A fitness function is used to evaluate how good an individual is, that is, how many cases it can classify correctly. Then, some individuals are selected to evolve (generate) new individuals with the genetic operators. Individuals become better and better through the evolution process until the termination criterion is met. The input is the training data set, and the output is a rule set, which can classify the training data with higher accuracy. We use a simple example to illustrate the rules deduced by the Rule Learning. For HBV data set B, which will be introduced in the following section, we have learned the rules for diagnosing liver cancer (HCC) and non liver cancer (CONTROL) cases. The rules are given as follows:

IF A1762 and G1764 and C53, then HCC,

IF T1762 and A1764 and CG2712, then HCC,

IF T1762 and A1764 and T2712 and C2525, then HCC,

ELSE CONTROL

E. *Decision Tree*

A decision tree is a tree-structured classifier. The Decision Tree method learns decision tree using a recursive tree growing process. Each test corresponding to an attribute is evaluated on the training data using a test criteria function. The test criteria function assigns each test a score based on how well it partitions the data set. The test with the highest score is selected and placed at the root of the tree. The subtrees of each node are then grown recursively by applying the same algorithm to the examples in each leaf. The algorithm terminates when the current node contains either all positive or all negative examples.[10]

F. *Support Vector Machine*

SVMs are a set of related supervised learning methods used for classification and regression. Viewing input data as two sets of vectors in an n-dimensional space, an SVM will construct a separating hyperplane in that space, which maximizes the margin between the two data sets. To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are "pushed up against" the two data sets. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the neighbouring data points of both classes, since, in general, the larger the margin the smaller the generalization error of the classifier. The original optimal hyperplane algorithm proposed by Vladimir Vapnik in 1963 was a linear classifier. The classification model produced by SVM (as described above) only depends on a subset of the training data, because the Cost function for building the model does not care about training points that lie beyond the margin.[11]

G. *Naive Bayes [12]*

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent Feature model." Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of

maximum likelihood; in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian method. In spite of their oversimplified assumptions, Naive Bayes classifiers often work much better in many complex real world situations than one might expect. Recently, careful analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficiency of Naive Bayes classifiers [13]. An advantage of the Naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

IV. EVALUATION METHODOLOGY

In classifying an unknown case, depending on the class predicted by the classifier and the true class of the patient (Control or HCC), four possible types of results can be observed for the prediction as follows:

	actual class (expectation)	
predicted class (observation)	**tp** (true positive) Correct result	**fp** (false positive) Unexpected result
	fn (false negative) Missing result	**tn** (true negative) Correct absence of result

Figure 2: Confusion Matrix

1. True positive—the result of the patient has been predicted as positive (Cancer) and the patient has cancer.
2. False positive—the result of the patient has been predicted as positive (Cancer) but the patient does not have cancer.
3. True negative—the result of the patient has been predicted as negative (Control), and indeed, the patient does not have cancer.
4. False negative—the result of the patient has been predicted as negative (Control) but the patient has cancer.

Let TP, FP, TN, and FN respectively, denote the number of true positives, false positives, true negatives, and false negatives. For each learning and evaluation experiment, Sensitivity, Specificity, Precision, Recall and F-Measure defined below are used as the fitness or performance indicators of the classification:

Sensitivity = TP/TP + FN;

Specificity = TN/TN +FN;

Precision = TP/TP +FP;

Recall = TP/TP +FN;

F-Measure=2(Precision* Recall) / (Precision +Recall)

.

Performance / K-Fold Cross Validation	True Positive Rate (Sensitivity)	False Positive Rate (Specificity)	Precision	Recall	F-Measure
2-Fold	0.975	0.211	0.951	0.975	0.963
3-Fold	0.962	0.150	0.962	0.962	0.962
4-Fold	0.925	0.158	0.961	0.925	0.943
5-Fold	0.949	0.294	0.938	0.949	0.943
6-Fold	0.925	0.150	0.961	0.925	0.943
7-Fold	0.962	0.100	0.974	0.962	0.968
8-Fold	0.950	0.100	0.974	0.950	0.962
9-Fold	0.937	0.100	0.974	0.937	0.955
10-Fold	0.937	0.118	0.974	0.937	0.955

.For screening tests, medical professionals usually will prefer to have higher sensitivity, i.e., lower specificity is an acceptable trade-off for high sensitivity as long as the specificity are reasonable. It means that we rather send more people for Confirmation tests than miss any true cancer patients. In the data sets, all attributes are categorical attributes. There are four symbolic values A, C, G, and T for each attribute. We adopt K-fold cross-validation method to make sure that the whole data set can be used as testing data in turn and overtraining (over fitting) can be avoided.

We must ensure that there is at least one positive case for each class in the testing data set. If the number (K) of splits is too large, the size of the testing set will be too small, and it may not even have a positive case. On the other hand, if the numbers of splits are too small, it will result in small training sets which may not contain sufficient information for training. So, we need to find a balance between the sizes of training and testing sets in order to reduce the probability of overtraining (overfitting) and undertesting (i.e., not enough positive and negative examples for testing). We have tried several feasible K values for Rule Learning.We can see that the testing sensitivity and Specificity are best by taking 10-fold. Consequently, we have chosen a 10-fold method for our experiments.

V. EXPERIMENTAL RESULTS

In this section, we first present the results of classifiers to classify the HBV DNA data into liver cancer (HCC) and normal (CON, control) classes, and then compare them with several traditional classification methods which include See5.0 (Decision Tree) [10], SVM [11], and Naive Bayes [12]. The biochemists and doctors can see explicitly and clearly the influences of the mutated sites or markers and their potential interactions toward the formation of liver cancer. For reducing computational complexity, we reduce the number of attributes by including the feature selection method.

TABLE 2

COMPARISON RESULTS WITH CLASSICAL METHODS USING INFORMATION GAIN

Performance	Rule Learning	Decision Tree	Naive Bayes	Lib Support Vector Machine
True Positive	0.920	1.000	1.000	1.000
False Positive	0.000	0.333	0.333	0.667
Precision	1.000	0.933	0.933	0.875
Recall	0.920	1.000	1.000	1.000
F-Measure	0.958	0.966	0.966	0.933

It is clear from the Table 2 and Table 3,that after applying the combination of ReliefF and CFS the performance of the classification methods are improved..In our case, sensitivity(True positive rate),Precision, Recall, F-Measure should be high and Specificity should be low. It is also clear from the following graph

TABLE 3

COMPARISON OF CLASSIFICATION RESULTS AFTER FEATURE SELECTION USING RELIEFF AND CFS FOR ALL DATASETS

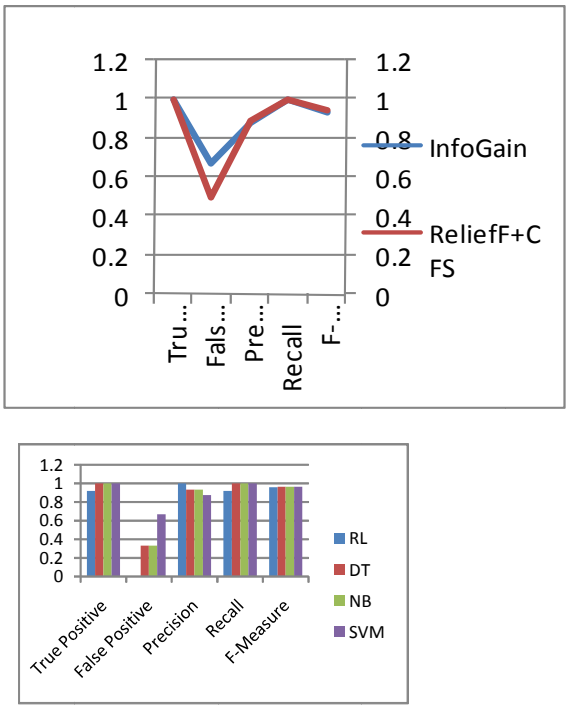


Figure 4. Comparison of all methods using Information Gain

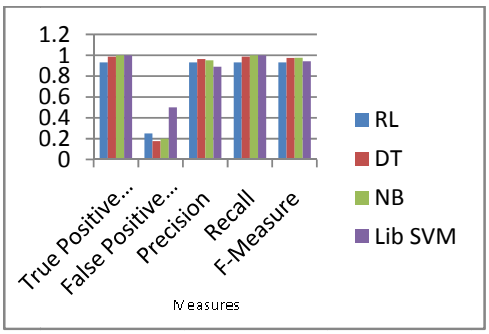


Figure 5. Comparison of all methods using ReliefF and CFS

It is clear from the Figure 4 and Figure 5, the combination of ReliefF and CFS algorithm reduces the False Positive Rate which in turn increases the performance of the classification methods.

VI. CONCLUSION

In this paper, a data mining framework for DNA sequence biological data sets has been presented. It has been applied to the Hepatitis B Virus DNA data sets which are among the largest in the world and have been collected from gene bank for this project. We have developed a framework for markers discovery. This framework has incorporated four algorithms, Rule Learning, Naïve Bayes, Decision Tree, and Support Vector Machine. These Classifiers can explicitly give the importance of the markers and their interactions and have shown good performance in cancer prediction. The experimental results have shown that the combination of ReliefF and Correlation Based Feature Selection Methods increases the performance of the classifiers especially in terms of sensitivity, specificity, precision. Among the classifiers considered Naïve Bayes classifies efficiently compared to others. Hence, it is very useful for preliminary diagnosis and screening test of liver cancer caused by HBV.

VII.FURURE WORK

The final confirmation experiments, like many other bioinformatics problems, need to be carried out by biochemists to identify and study the true markers. We believe that this comparative study can be helpful to do further research for biochemists. We hypothesize that the genomic makeup of HBV affects the carcinogenic potential of the virus. In this case-control study, we have demonstrated that some genotype-specific mutations are more commonly found among HCC patients than their age and gender-matched controls. These markers can therefore be used as biomarkers to stratify the cancer risk of chronic hepatitis B patients. Our findings have been validated by independent data sets in the validation process.

REFERENCES

[1] R.B. Potter and S. Draghici, "A Soft Approach to Predicting HIV Drug Resistance,"Proc. Pacific Symp. Biocomputing (PSB '02), 2002.

[2] A. Ciancio, A. Smedile, and M. Rizzetto, "Identification of HBV DNA Sequences that Are Predictive of Response to Lamivudine Therapy," Hepatology, vol. 39, pp. 64-73, 2004.

[3] H.L.Y. Chan et al., "Genotype C Hepatitis B Virus Infection Is Associated with an Increased Risk of Hepatocellular Carcinoma,"Gut, vol. 53, pp. 1494-1498, 2004.

[4] H.L.Y. Chan, C.H. Tse, E.Y.T. Ng, K.S. Leung, K.H. Lee, K.W. Tsui, and J.J.Y. Sung, "Phylogenetic, Virological and Clinical Characteristics of Genotype C Hepatitis B Virus with Tcc at Codon 15 of the Precore Region," J. Clinical Microbiology, vol. 44, no. 3, pp. 681-687,2006.

[5] H. Almuallim and T. Dietterich, "Learning Boolean Concepts in the Presence of Many Irrelevant Features," Artificial Intelligence, vol. 69, nos. 1/2, pp. 179-305, 1994.

[6] M. Dash and H. Liu, "Feature Selection for Classification," Intelligent Data Analysis, vol. 1, no. 3, pp. 131-156, 1997.

[7] G. John, R. Kohavi, and K. Pdlwfwe, "Irrelevant Features and the Subset Selection Problem," Proc. 11th Int'l Conf. Machine Learning, pp. 121-129, 1994.

[8] P. Langley, "Selection of Relevant Features in Machine Learning," Proc. AAAI Fall Symp. Relevance, pp. 1-5, 1994.

[9] T.M. Mitchell, Machine Learning. The Mc-Graw-Hill Companies, Inc., 1997.

[10] Data Mining ToolSee5 and C5.0, Software, http://www.rulequest.com/see5- info.html, May 2006.

[11] C.C Chang and C.J. Lin, "LIBSVM: A Library for Support Vector Machines," Software,http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[12] C. Borgelt, Bayes Classifier Induction, Software, http://fuzzy.cs.uni-magdeburg.de/~borgelt/bayes.html, 2009.

[13] H. Zhang,"The Optimality of Naive Bayes," Proc. 17th Int'l Florida Alliance of Informationand Referral Services (FLAIRS) Conf., 2004.

◈◈◈

Design of a Communication System Based on IEEE 804.15.4 for Telemedical Applications

G C Madhu, T.Reshma, N. Suvarna, K.Nomitha Raj & Y.I.Swetha

Sree Vidyanikethan Engineering College, Department of Electronics and Control Engineering
E-mail : msnaidu417@gmail.com, reshma.krc@gmail.com

Abstract - This paper deals with the design and implementation of the prototype of a Digital communication system that transmits all the information from one point to another using an Xbee controller. To test this system we have chosen to transmit the values of Blood Pressure, body Temperature and position of a patient from home place to the doctor room.

Keywords- *Sensors, Xbee module, LabVIEW, Digital Communication*.

I. INTRODUCTION

The present technological evolution is leading to many new forms of revolutionary needs that meet the needs of patients. In order to improve elderly life conditions at home and to reduce the costs of long hospitalizations, the medical world is more and more interested in telemonitoring techniques. These techniques will allow elderly people to stay safely at home, to benefit from an automated medical supervision and will delay their entrance in nursing homes. Medical professionals believe that one of the best ways to detect emerging physical and mental health problems, before it becomes critical (particularly for the elderly), is analyzing the human behaviour and looking for changes in the activities of daily living (ADLs).

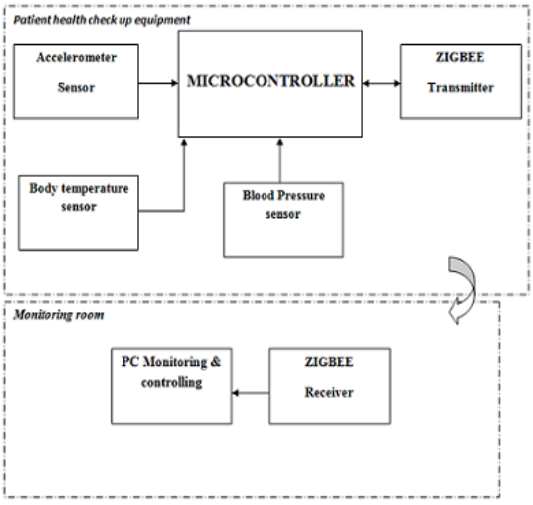


Fig. 1 : Block Diagram of the System

II. PROTOTYPE

This system consists of two Zigbee nodes with each node can be controlled by an intelligent ZigBee Embedded Microcontroller. To build this application 8-bit Microcontroller is used as the central core of the project. Here one ZigBee transceivers act as End Device can be placed at patient home (or this system attached for patient belt), another one ZigBee act as a Coordinator can be placed to doctor's room for monitoring the heart beat and fall status.

Here blood pressure sensor, body temperature sensor and accelerometer sensor can be connected to the Microcontroller. Each collection of data's are transferred through the Zigbee controller end device sent the information in control room (coordinator). Blood pressure sensor measure systolic & diastolic & pulse rate can be measured using this sensor. Any body movements quickly inform the Accelerometer sensor. Body temperature sensor is used to measuring the human body temperature. Then the system realizes the real time remote supervisory for the patient information. The command window which is the output window used here is the LabVIEW.

2.1 MICROCONTROLLER

The microcontroller used here is P89C51RD2BN. It contains a non-volatile 64KB Flash program memory that is both parallel programmable and serial In-System and In-Application Programmable. The device supports 6-clock/12-clock mode selection by programming a Flash bit using parallel programming or In-System Programming. This device is a Single-Chip 8-Bit Microcontroller manufactured in an advanced CMOS process and is a derivative of the 80C51 microcontroller

family. The device also has four 8-bit I/O ports, three 16-bit timer/event counters, a multi-source, and four-priority-level, nested interrupt structure, an enhanced UART and on-chip oscillator and timing circuits. The added features of the P89C51RD2BN make it a powerful microcontroller for applications that require pulse width modulation, high-speed I/O and up/down counting capabilities such as motor control.

2.2 SENSORS

Here we used three sensors to find the three parameters of a patient. A pressure sensor to find the blood pressure using a digital blood pressure meter is used. The temperature of the body is measured using a temperature sensor called LM35. For the position and movements of the body, an accelerometer sensor ADXL335 is used. These sensors are interfaced with the microcontroller.

2.2.1 BLOOD PRESSURE SENSOR

The method used in electronics systems for measuring blood pressure is the oscillometric method. The pressure sensor is connected directly to the cuff, which is inflated or deflated via a motor and valve.

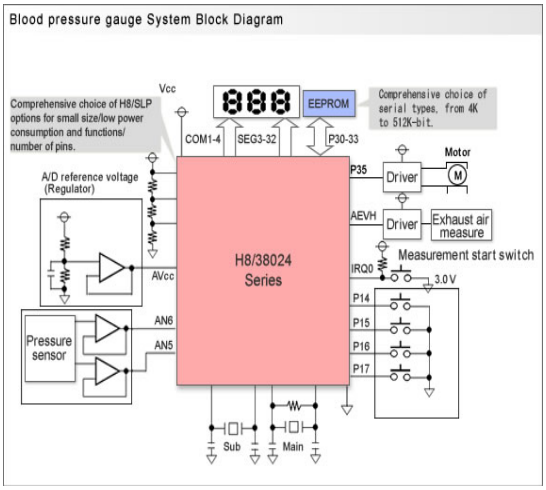


Fig. 2 : Digital Blood Pressure meter

Digital Blood Pressure Meter uses an integrated pressure sensor, analog signal-conditioning circuitry, microcontroller hardware/software and a liquid crystal display. The sensing system reads the cuff pressure (CP) and extracts the pulses for analysis and determination of systolic and diastolic pressure. This design uses a 50 kPa integrated pressure yielding a pressure range of 0 mm Hg to 300 mm Hg.

2.2.2 TEMPERATURE SENSOR

The LM35 series sensors used here are precision integrated-circuit temperature sensors, whose output voltage is linearly proportional to the Celsius (Centigrade) temperature. The LM35 thus has an advantage over linear temperature sensors calibrated in ° Kelvin, as the user is not required to subtract a large constant voltage from its output to obtain convenient Centigrade scaling.

Fig. 3: Temperature Sensor LM35

The LM35 does not require any external calibration or trimming to provide typical accuracies of ±1⁄4°C at room temperature and ±3⁄4°C over a full −55 to +150°C temperature range. Low cost is assured by trimming and calibration at the wafer level. The LM35's low output impedance, linear output, and precise inherent calibration make interfacing to readout or control circuitry especially easy.

2.2.3 ACCERELOMETER SENSOR

The ADXL335 is a small, thin, low power, complete 3-axis accelerometer with Signal conditioned voltage outputs. The product measures acceleration with a minimum full-scale range of ±3 *g*. It can measure the static acceleration of gravity in tilt-sensing applications, as well as dynamic acceleration resulting from motion, shock, or vibration.

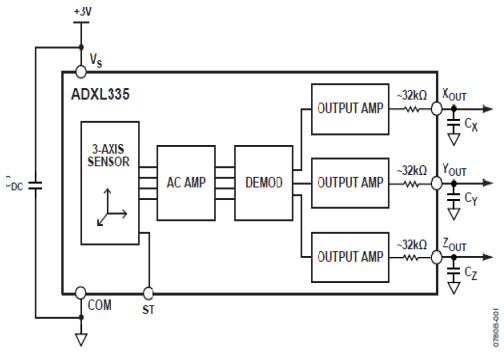


Fig.4 : Accelerometer Sensor ADXL335

The user selects the bandwidth of the accelerometer using the CX, CY, and CZ capacitors at the XOUT, YOUT, and ZOUT pins. Bandwidths can be selected to suit the application, with a range of 0.5 Hz to 1600 Hz for the X and Y axes, and a range of 0.5 Hz to 550 Hz for the Z axis.

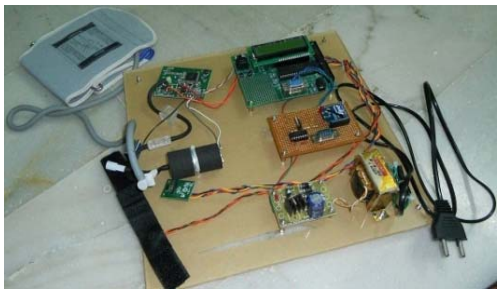


Fig. 5. Transmitter part of the overall system

Fig. 6. Receiver part of the overall system

2.3 COMMUNICATION

The data from the patient needs to be transmitted from the patient's room to the doctor's room. In this project wireless communication is used through the Xbee module. It consists of a transmitter and receiver both based on Zigbee Protocol. Zigbee is specifically designed to provide a cost-effective, standard-based and flexible wireless network, which supports low power consumption, reliability, interoperability and security for control and monitoring applications with low to moderate data rates.

The Xbee module can transmit low data rate through a distance of about 50m. The communication card is essentially based on the Xbee module, the microcontroller and the USB port. It can provide the communication between PC-Xbee, the communication between PIC-Xbee and the communication between PIC-PC. In addition, it allows users to make Analog to Digital conversion to one of the five analog inputs. The presence of the microcontroller on the card is mandatory as it is considered as the brain of the card. Indeed, it manages all communication and determines the mode of communications and run the Analog to Digital Conversion.

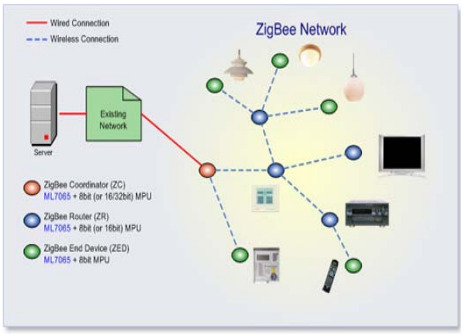


Fig. 6. A Zigbee Network

Zigbee is a wireless communications standard that provides a short-range cost effective networking capability. It has been developed with the emphasis on low-cost, battery-powered applications, such as building automation, industrial and wireless, personal healthcare and advanced tagging.

III. LABVIEW

LabVIEW (Laboratory Virtual Instrument Engineering Workbench) is a graphical programming language that uses icons instead of lines of text to create applications. In contrast to text-based programming languages, where instructions determine the order of program execution, LabVIEW uses dataflow programming. VIs or virtual instruments, is LabVIEW programs that imitate physical instruments.

LabVIEW is used to communicate with hardware such as data acquisition, vision, and motion control devices, and industrial automation on a variety of platforms including Microsoft Windows, various versions of UNIX, Linux. LabVIEW contains a comprehensive set of tools for acquiring, analyzing, displaying, and storing data.

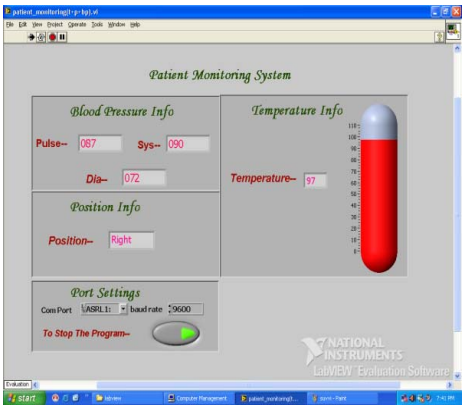


Fig. 7. LabVIEW output at the doctor's screen

IV. EXPERIMENTAL RESULTS

To test this prototype many experiments have been done with many patients to evaluate the quality of transmission from different distances. Initially when it is tested with one patient at a distance less than 10m, all the parameters that were measured are accurately transmitted through the Xbee module to the receiver section without any error.

Next it was repeated for gradually increasing distances. Every time there was an accurate transmission of the data from the transmitter to the receiver.

V. CONCLUSION

In this prototype the biological data from the patient to the doctor's screen is transmitted via a wireless digital communication. As this prototype is used for only a distance of about 50m using the Xbee module, in future this can be extended to a more farther distances efficiently.

REFERENCES

[1] Nadia Ghariani, Mondher Chaoui, Hamadi Ghariani, Mongi Lahiani "Design Of A Digital Communication System Based On a XBee Module For Biomedical Applications", in 8th International Multi-Conference on Systems, Signals & Devices.

[2] S. Borromeo, C. Rodriguez-Sanchez, F. Machado, J.A. Hernandez-Tamames, R. de la Prieta, "A Reconfigurable, Wearable, Wireless ECG System", in 29th Annual International Conference of the IEEE EMBS, France, August 23-26, 2007, pp.1659-1662

[3] C. Ken, L. Xiaoying, "A Zigbee Based Mesh Network for ECG Monitoring System", in 4th International Conference of Bioinformatics and Biomedical Engineering (iCBBE), 2010, pp. 1-4

[4] Yu-Cheng, Su Huan, Chen Ching-Lun, Hung Shuenn-Yuh Lee, "Wireless ECG detection system with low-power analog front-end circuit and bio-processing ZigBee firmware", in Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS), 2010, p 1216 – 1219

[5] http://www.laboratoiresdelaretine.com/ dmla/wpcontent/up loads/AQDM-telemedicine.pdf\

[6] http://www.digi.com/products/wireless/zigbee-mesh/xbeedigimesh-2-4.jsp#overview

[7] V. Auteri, C. Lamberti, L. Roffia, and T. S. Cinotti, "ZigBee-based wireless ECG monitor", in Computers in Cardiology 2007 (CINC07), 2007, pp. 133-136.

Improved Genetic Algorithm For Optimal Power Dispatch In Multinode Deregulated Electricity Market

S.Padmini, Teresa George & Dinesh Pillai

Dept of Electrical Engineering, Srm University, Kattankulathur
E-mail : padminisp81@gmail.com , treesag77@gmail.com & dineshpillai98@yahoo.com

Abstract - The electricity industries in number of countries have recently been deregulated to introduce competition. In a competitive electricity market, generation resources are, scheduled based on offers and bids of the suppliers and consumers. This paper proposes the application of improved genetic algorithm (IGA) to solve the optimal power (OPD) in electricity market. Improved Genetic Algorithms (IGA) explore the possibility to work with populations in order to reduce the processing time. The objective of the algorithm is to maximise the participants benefit in a system. The proposed algorithm using IGA for OPD of multi-node electricity market is simple to implement and can incorporate additional constraints. IGAs are more efficient to solve this kind of problem as they are faster and give nearly global optimal solutions. The developed algorithms are tested on a 17-node, 26-line system.

Keywords*-multinode auction market,optimal power dispatch, Genetic algorithm(GA), Independent system operator(ISO).*

I. INTRODUCTION

The electric power industry has over the years been dominated by large utilities that had an overall authority over all activities in generation, transmission and distribution of power. Such utilities have often been referred to as vertically integrated utilities. The electric power industry has been undergoing a process of transition and restructuring since the nineties decade or so. The restructuring process of power industry has been the separation of the transmission activities from the electricity generation activities. The subsequent step was to introduce competition.

One of the competitive electricity market models is the auction market model, in which participants place their bids to sell or buy electricity as explained in literature [4].. In an electricity auction market, the two main participants are distribution companies and generation companies. These participants will submit their bids to an independent system operation (ISO) company. A supply bid is given as a cost per MW and a quantity in MW which a generation company is willing to generate in a particular period. Each generation company may place several bids. A demand bid is given as a cost per MW and a quantity in MW which a distribution company is willing to consume in a particular period. Several demand bids may be placed by each distribution company.

The optimal power dispatches have been proposed by several researchers [2] and have the objective to maximize the total benefit to the participants in the auction market. In [1] genetic algorithm is applied for optimal power dispatch. This paper demonstrates the application of improved genetic algorithm to solve the optimal power dispatch problem for a multi-node auction market. The model used in this thesis, like most of the models available in literature, does not directly consider the reactive power market. The advantage of the proposed genetic algorithm is the simplicity of handling non-linear constraints in less processing time. In addition, the algorithm is easy to implement and additional features such as security constraints can be easily incorporated in the algorithm**.**

A new model using improved genetic algorithm is developed to solve the optimal power dispatch problem for a multi-node auction market. The methods are tested on 17-node 26-line system and compared with other methods to demonstrate their performance**.**

II. PROBLEM DESCRIPTION FOR SINGLE NODE ELECTRICITY MARKET

The spot price at a single node is the price which matches the supply and demand bids, i.e., the point at which the supply and demand curves intersect each other. The supply curve is plotted in increasing order of price and demand curve in decreasing order of price. At the spot price, the benefit of participants is maximized. For a single node auction market, the supply and demand curves at each node can be illustrated as shown

in fig. 1. The maximisation of participants benefit is shown by shaded area.

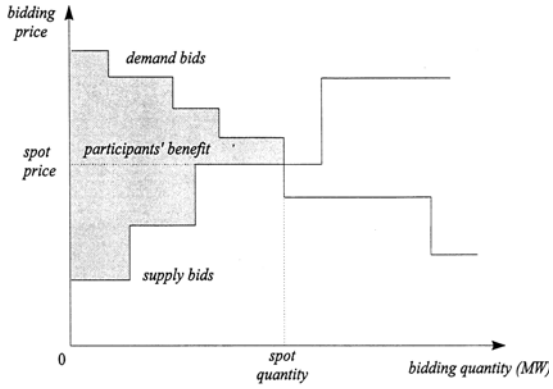


Fig2.. An example of supply and demand curves

Assuming that there are M_k supply bids and N_k demand bids at the k^{th} node. Let S_{ik} be the ith supply bid at node k and is given by $S_{ik} = \{x^s_{ik}, p^s_{ik}\}$, where x^s_{ik} is the selling price and p^s_{ik} is the selling quantity. In addition, let B_{ik} be the i^{th} demand bid at node k and is given by $B_{ik} = \{x^d_{ik}, p^d_{ik}\}$, where x^d_{ik} is the buying price and p^d_{ik} is the buying quantity. If $\hat{x}_k$ denotes the spot price and $\hat{p}_k$ denotes the spot quantity, then the maximum participants' benefit, which is the sum of suppliers' benefit and consumers' benefit, can be given as

$$B_k = \sum_{i \in M_k^s} (\hat{x}_k - x^s_{ik})\, \tilde{p}^s_{ik} + \sum_{j \in N_k^d} (x^d_{jk} - \hat{x}_k)\, \tilde{p}^d_{jk} \quad (1)$$

where $\tilde{p}^d_{ik}$ and $\tilde{p}^s_{ik}$ are consumer's and supplier's dispatched quantity, respectively, M^s_k and N^d_k are the sets of all dispatched suppliers and dispatched consumers, respectively.

III. PROBLEM DESCRIPTION FOR MULTINODE ELECTRICITY MARKET

For a multi-node electricity auction market, there are transmission lines connected between bidding nodes. The connections result in real power P_k and reactive power q_k injection to the network at each node. The real power injection to a node can be modeled as an additional demand bid (or a supply bid if the real power injection is negative) by the network for the quantity p_k at the selling (or buying) price $\hat{x}_k$, which is equal to spot price. As an example, Fig. 3 illustrates the dispatch of the bids when the real power injection is considered. In Fig. 2(a), the injection of Pk to the node is supplied by the partly dispatched generator bid. The spot quantity has increased and the price has not changed. If the injected power is greater than the undispatched amount of the partly dispatched supply bid then the additional amount cannot be supplied at the same price. Therefore, the spot price will increase as shown in Fig. 2(b). This will result in displacing some consumers as shown by *dc* in Fig. 2(b). It can be seen in Fig. 2 that the spot price and spot quantity may be changed due to the effect of the real power injection. This may result in changing the sets Bik and Sik of all dispatched suppliers and dispatched consumers.

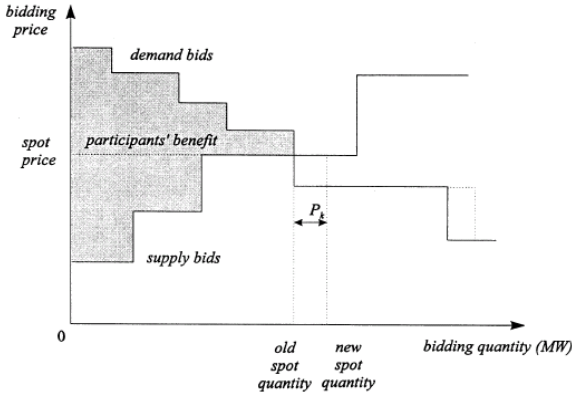


Fig 3(a)

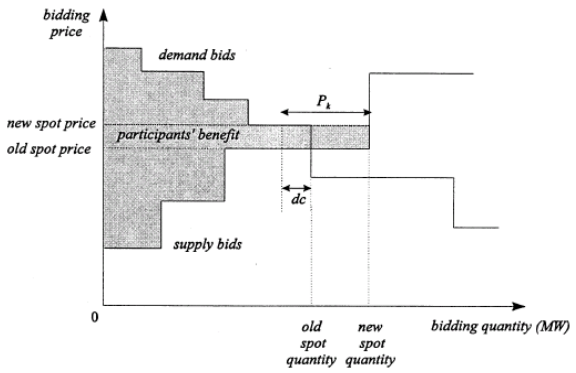


Fig 3(b)

Fig 3. Examples of the network effects.

Consequently, the participants' benefit at node k is now given by

$$B'_k = \sum_{i \in M_k^s} (\dot{x}_k - x^s_{ik})\, \tilde{p}^s_{ik} + \sum_{j \in N_k^d} (x^d_{jk} - \dot{x}_k)\, \tilde{p}^d_{jk} - \dot{x}_k\, p_k \quad (2)$$

where $\dot{M}_k^s$ and $\dot{N}_k^d$ are the new sets of all dispatched suppliers and dispatched consumers respectively, $\dot{x}_k$ is the new spot price and the last term is the amount paid by the transmission line. In addition, the total participants' benefit at all nodes can be expressed as

$$B_k^{'} = \sum_{k=1}^{K} \left\{ \sum_{i \in \dot{M}_k^s} (\dot{x}_k - x_{ik}^s)\tilde{p}_{ik}^s + \sum_{j \in \dot{N}_k^d} (x_{jk}^d - \dot{x}_k)\tilde{p}_{jk}^d - \dot{x}_k p_k \right\} \quad (3)$$

where *K* is the number of nodes.

The participants' benefit at each node ($B_k^{'}$) is a function of the real power injection. Therefore, the optimization problem of the total participants' benefit at all nodes is similar to the conventional optimal load flow problem, with the exception that the objective is to maximize the participants' benefit, rather than minimize the generation cost. This optimization problem can be described as

Maximize

$$\sum_{k=1}^{K} \left\{ \sum_{i \in M_k^s} (\dot{x}_k - x_{ik}^s)\tilde{p}_{ik}^s + \sum_{j \in N_k^d} (x_{jk}^d - \dot{x}_k)\tilde{p}_{jk}^d - \dot{x}_k p_k \right\} \quad (4)$$

Subject to the following constraints:

The capacity constraints which provide the limits on real power (p_k) and reactive power (q_k) injection to the network by any node, i.e.

$$\underline{p}_k \le p_k \le \overline{p}_k \quad (5)$$

$$\underline{q}_k \le q_k \le \overline{q}_k \quad (6)$$

Where $\underline{p}_k$, $\overline{p}_k$ are the minimum and maximum real power injection limit associates with node *k* and $\underline{q}_k$, $\overline{q}_k$ are the minimum and maximum reactive power output limits of generators associate with node *k*.

Constraints on the limit of power flow along lines which are given by

$$\left| p_{kl} \right| \le \overline{p}_{kl} \quad (7)$$

where $\overline{p}_{kl}$ is the maximum limit of a power flow in a line connecting node *k* and node *l*.

In addition, the real and reactive power injection at each node can be determined as a summation of the real and reactive power flows along lines which are connected to that node. These are given by

$$p_k = \sum_{\substack{l=1 \\ l \ne k}}^{K} p_{kl} \quad (8)$$

$$q_k = \sum_{\substack{l=1 \\ l \ne k}}^{K} q_{kl} \quad (9)$$

where p_{kl} and q_{kl} are the real power and reactive power flow along the transmission line connecting node *k* and node *l*, respectively. Furthermore, the real power and reactive power flow are given by the following equations

$$p_{kl} = G_{kl}(v_k^2 - v_k v_l \cos(\theta_k - \theta_l)) - B_{kl}(v_k v_l \sin(\theta_k - \theta_l)) \quad (10)$$

$$q_{kl} = -B_{kl}(v_k^2 - v_k v_l \cos(\theta_k - \theta_l)) + G_{kl}(v_k v_l \sin(\theta_k - \theta_l)) \quad (11)$$

Where G_{kl} and B_{kl} are real and imaginary component of the admittance of the line connecting node *k* and node *l*, θ_k and θ_l are angles at node *k* and *l* and v_k and v_l are voltages at node *k* and node *l*.

This optimization problem has non-linear constraints which is difficult to solve using the linear programming technique. An improved genetic algorithm is proposed in the following section to solve the above problem. The genetic algorithms are simple to implement and it is easy to incorporate additional constraints into the problem.

IV. APPLICATION OF IMPROVED GENETIC ALGORITHM FOR OPTIMAL POWER DISPATCH

Genetic algorithms have been successfully applied to many areas in science and engineering applications. Genetic algorithm starts with a population of randomly generated candidates. These are called chromosomes. The population is a set of solutions to a problem. Each chromosome is made up of a sequence of genes. Chromosomes can be represented as binary digits, floating point numbers, integers, real values, matrices, etc. Generally natural representations are more efficient and will produce better solutions. Real-coded

representation is more efficient in terms of CPU time and offers more consistent results.

In this thesis real coded genetic algorithm is used. Each chromosome will have a fitness value which indicates the quality of the solution. The genetic algorithm iteratively produces a new population from the old population by means of genetic recombination processes like selection, mutation and crossover. When this genetic recombination process is iterated for many generations, the overall fitness of the population generally improves.

In this study the objective function of optimisation problem is used as the fitness function. The objective is to maximise the fitness function.

Maximise

$$F = \sum_{k=1}^{K} \left\{ \sum_{i \in M_k^s} (\dot{x}_k - x_{ik}^s)\tilde{p}_{ik}^s + \sum_{j \in N_k^d} (x_{jk}^d - \dot{x}_k)\tilde{p}_{jk}^d - \dot{x}_k p_k \right\} \quad (12)$$

Initially many individuals are randomly generated to form an initial population covering the entire range of possible solutions (search space). Each individual will be having its own fitness value. Selection process is done by selecting a proportion of the existing population to bread a new breed of generation. Several selection methods are there like roulette wheel selection and its extensions, scaling techniques, tournament, normal geometric, elitist models and ranking methods. In this algorithm tournament selection is used to prevent loss of diversity in population. First, the individuals are selected (only once) for a tournament where couples are randomly formed to compete between themselves. The most adapted individual of each couple wins. Then, the tournament is repeated and the selected individuals form couples to begin crossover. In this way, not only do the most developed individuals have an opportunity to participate in the reproduction but all of them do.

The new individuals are created from existing population by mutation and crossover processes. A crossover process combines the feature of two parents to form two similar offspring which inherit features from both parents. In this thesis uniform crossover is done. Mutation process injects new information in the offspring by altering a small part of a chromosome. A solution is found when it satisfies the minimum criteria or when the highest ranking solution's fitness is reaching or has reached.

In addition to performing the fitness evaluation, tournament selection, and uniform crossover, IGA uses the elitism mechanism. Elitism, on the other hand, guarantees that the best string individual survives until the last generation. More specifically, if the best offspring individual is worse than the best parent individual, the best parent individual will randomly replace any offspring individual. Elitism significantly improves performance.

IGA produces better and consistent results compared to basic genetic algorithm. As it uses the elitism mechanism and tournament selection best off springs are retained. Mutation is not needed as best individual survives until the last generation.

ALGORITHM FOR MULTINODE ELECTRICITY MARKET USING IGA

1. Read the system data and GA data.
2. Read the suppliers bidding data and consumers bidding data (quantity of power injection, price).
3. Randomly generate a *p* initial population and go to step6.
4. Run Newton Raphson Load Flow.
5. Compute power flows through all the lines, real and reactive power injections of slack bus.
6. Evaluate fitness value (total benefit) using
7. Select individuals for reproduction using tournament selection and elitism.
8. Apply crossover and mutation.
9. Calculate the fitness of the new chromosomes using (12).
10. Repeat steps 10, 11 until the population reaches the homogeneous degree previously chosen.
11. Find the best chromosome, keep it, and discard the others.
12. Repeat steps 5 to 12, until the best individual is identified after a maximum number of generations.
13. Calculate spot price, participants benefit, lines benefit and total benefit at all the nodes.
14. end

V. RESULTS AND DISCUSSIONS

The developed improved genetic algorithm was tested on17 node 26 line system.The objective is to maximize the total participants' benefit at all nodes in the system, which in turn depends on the real power injection to the system. In addition, the real power and reactive power injection at the reference node can be obtained from the load flow solution which is given in Table1. The total participants' benefit is given as a chromosome's fitness and it has been determined by solving the load flow problem. The real power injection at a given node is maximum when all selling bids are dispatched. Therefore the maximum possible injection is

equal to the total amount of power offered by suppliers at that node. Similarly, the minimum power injection (i.e. maximum negative injection) is when no selling bids are dispatched and all buying bids are dispatched. It is equal to the total amount of power bid by the consumers. The evaluated spot price, participants benefit and total benefit for single node is given in table2 and the results for multinode are given in table3 and table 4.

TABLE 1
POWER INJECTION, VOLTAGE AND PHASE ANGLES

Node	P(MW)	Q(MVAR)	Voltage(V)	Angle(degree)
1	162.00	139.51	1.05	7.00
2	281.34	11.72	1.05	8.99
3	-51.16	-38.19	1.049	6.92
4	-123.55	-39.37	1.05	6.90
5	46.40	-36.98	1.05	5.18
6	15.23	30.54	1.05	3.47
7	234.84	81.92	1.05	2.50
8	-156.52	-85.05	1.05	1.72
9	-35.0	-2.17	1.042	2.42
10	-3.26	12.99	1.05	1.67
11	-234.14	-160.73	1.05	1.32
12	231.23	363.40	1.05	1.82
13	-113.96	-60.97	1.05	1.76
14	-30.10	-5.26	1.041	0.751
15	-437.96	-158.88	1.05	0.58
16	443.20	126.14	1.05	4.24
17	-213.00	-310.80	1.0	0.0

TABLE 2. SPOT PRICE AND PARTICIPANTS BENEFIT FOR SINGLE NODE ELECTRICITY MARKET

Node	Spot price($/MW)	Participant's benefit($)
1	1.140000	14.440000
2	0.700000	376.500000
3	-----	-----
4	1.400000	2.700000
5	1.100000	8.300000
6	1.000000	2.000000
7	0.010000	30.760000
8	-----	-----
9	-----	-----
10	1.000000	12.900000
11	1.400000	34.200000
12	1.000000	4.800000
13	1.300000	35.500000
14	-----	-----
15	1.000000	65.500000
16	1.000000	63.800000
17	-----	-----
Total		651.56

TABLE 3 .PARTICIPANTS BENEFIT USING GA

Node	**With Network Effect**			
	Spot price($/MW)	**Total benefit($)**	**Participant's benefit($)**	**Line's benefit($)**
1	1.130811	-168.824	14.367	-183.191
2	1.003244	113.929	396.181	-282.253
3	1.020072	75.635	23.448	52.187
4	1.133647	158.998	18.936	140.062
5	1.107560	-43.807	7.583	-51.391
6	1.013704	-11.984	3.455	-15.439
7	1.000366	3.249	238.175	-234.926
8	1.000366	213.114	8.197	204.917
9	1.005474	42.409	7.217	35.192
10	1.107513	18.913	15.303	3.610
11	1.294779	344.687	41.528	303.159
12	1.006862	-226.414	6.402	-232.817
13	1.303282	187.455	38.933	148.522
14	1.008177	43.990	13.644	30.346
15	1.210129	644.099	114.111	529.988
16	1.000839	-384.672	58.900	-443.572
17	1.006831	327.310	112.855	214.455
Total		**1338.086**	**1119.235**	**218.851**

TABLE 4.PARTICIPANTS BENEFIT USING IGA

Node	**With Network Effect**			
	Spot price($/MW)	**Total benefit($)**	**Participants benefit($)**	**Line's benefit($)**
1	1.116666	-165.389	15.511	-180.900
2	1.024761	113.753	402.060	-288.306
3	1.023278	75.531	23.180	52.351
4	1.162419	159.465	15.848	143.617
5	1.114978	-43.994	7.741	-51.735
6	1.047990	-11.036	4.925	-15.961
7	1.033771	2.400	245.171	-242.771
8	1.329962	215.399	7.233	208.166
9	1.065120	43.062	5.783	37.279
10	1.107378	18.909	15.299	3.610
11	1.342315	357.975	43.685	314.290
12	1.005983	-226.287	6.327	-232.614
13	1.217671	178.612	39.846	138.766
14	1.077561	42.751	10.316	32.435
15	1.205641	643.799	115.777	528.022
16	1.057066	-383.687	84.804	-468.492
17	1.037417	322.168	101.198	220.970
Total		**1343.431**	**1144.704**	**198.727**

A. COMPARISON BETWEEN GA AND IGA

The participants benefit for single node market without network effect for this system is 651.56$. The spot prices and participants benefit for multinode electricity market using genetic algorithm are given in table 3. The improved genetic algorithm results are given in table 4. Table 5 shows the comparison of total benefit using genetic algorithm and improved genetic algorithm. It shows IGA produces better and consistent results.

TABLE 5. COMPARISON OF GA AND IGA

ALGORITHM	TOTAL BENEFIT
GA	1338.086
IGA	1343.431

VI. CONCLUSION

This paper presented IGA approaches for solving the optimal power dispatch in a multi-node electricity market. The objective of the algorithm is to maximize the total participants' benefit at all nodes in the system. GAs have proved themselves to be effective solutions to optimization problems. But their slow convergence is a disadvantage when applied to a real time system studies. Any improvement in the convergence of such algorithms is appropriate as far as the real time studies are concerned. IGAs are more efficient to solve this kind of problems, as they are faster and converge to better optimal solutions.

REFERENCES

[1] K..S Pandhya, S.K Joshi , " A survey of optimal power flow methods '', Journal of Theoretical and Applied information technology 2005-2008

[2] T. Numnonda, U.D. Annakkage, "Optimal power dispatch in multinode electricity market using genetic algorithm", Elect. Power Syst. Res. 49(1999) pp 211–220.

[3] R.W. Ferrero, S.M. Shahidehpour, "Optimality conditions in power transactions in deregulated power pools",Elect.Power Syst. Res. 42 (1997) pp 209–214.

[4] D.L. Post, S.S. Coppinger, G.B. Sheble, "Application of auctions as a pricing mechanism for the interchange of electric power", IEEE Trans.Power Syst. 10 (1995) pp 1580–1584

[5] R.A.S.K. Ranatunga, U.D. Annakkage, N.C. pahalawaththa, C.S. Kumble, "A Case Study of Network Characteristics Based Optimal Pricing and Dispatch", IEEE Trans. Power Syst. (1999) pp 190–196.

[6] R.A.S.K. Ranatunga, U.D. Annakkage, N.C. pahalawaththa, C.S. Kumble, "A Case Study of Network Characteristics Based Optimal Pricing and Dispatch", IEEE Trans. Power Syst. (1999) pp 190–196.

[7] T.T. Maifeld, G.B. Sheble, "Genetic-based unit commitment algorithm" IEEE Trans. Power Syst. 11 (1996) pp 1359–1370.

[8] S.O. Orero, M.R. Irving, "Economic dispatch of generators with prohibited operating zones:: A genetic algorithm approach", IEE Proc. Gener. Trans. Distrib. 143 (1996) pp 529–534.

[9] D.E. Goldberg, Genetic Algorithms in Search, Optimisation and Machine Learning, Addison-Wesley, Reading, MA, 1989.

[10] Kankar Bhattacharya, Math H.J. Bollen, Jaap E. Daadler, Operation of restructured power systems, Kluwer Academic Publishers, 2001.

◈◈◈

Implementation of RSA Key Generation based on RNS using Verilog

Vishak M & N.Shankaraiah

Dept. of Electronics and Communication Sri Jayachamarajendra College of Engineering, Mysore, INDIA
E-mail : Vishak.msp@gmail.com & shankarsjce@gmail.com

Abstract - RSA key generation is of great concern for implementation of RSA cryptosystem on embedded system due to its long processing latency. In this paper, a novel architecture is presented to provide high processing speed to RSA key generation for embedded platform with limited processing capacity. In order to exploit more data level parallelism, Residue Number System (RNS) is introduced to accelerate RSA key pair generation, in which these independent elements can be processed simultaneously. A cipher processor based on Transport Triggered Architecture (TTA) is proposed to realize the parallelism at the architecture level. In the meantime, division is avoided in the proposed architecture, which reduces the expense of hardware implementation remarkably. The proposed design is implemented by Verilog HDL and verified in matlab. A rate of 3 pairs per second can be achieved for 1024-bit RSA key generation at the frequency of 100 MHz.

I. INTRODUCTION

Public key cryptography has gained extreme popularity in many applications such as smart cards, digital certificate and so on. One of the most widely used public key algorithms is RSA .However, the process of RSA key generation is very complex and time consuming. Some implementations try to generate the RSA key pairs on a desktop and upload the pair, or only the private key, into a smart card. Take communication security and high-performance processing into consideration, the entire procedure of RSA key generation is preferably performed totally inside cipher chip in order to guarantee the efficiency and the security of the applications. Nevertheless, the limited computational power of embedded systems cannot afford high speed RSA key generation. In this paper, the issue about how to provide high processing speed to RSA key generation for embedded platform with limited processing capacity is discussed. An on-card implementation is presented in which takes up to 6.82 seconds to create a key pair. A scalable hardware architecture is proposed in they have presented the architecture applied for the multiple word Radix-2 Montgomery multiplication algorithm and a processing element (PE) is designed. We propose an efficient solution to accelerate RSA key pair generation from both data and architecture level parallelism, in which RNS and TTA are combined closely. The advantage of RNS Montgomery multiplication algorithm is that large number multiplications can be divided into small elements and each independent element can be processed simultaneously. The advantages of Transport Triggered Architecture (TTA) is that it can be used as an application specific processor, especially as a coprocessor for different DSP applications in SoC. Comparing with the traditional ASIC, TTA is more flexible in application and consumes less silicon area. And comparing with traditional DSP processor, it has more efficiency and better performance. In this paper, function unit (FU)"MMAC" is elaborately designed, which meets the highly parallelism of RNS Montgomery multiplication and reduces redundant data transmission. So an optimized architecture, called TTA-like, is proposed to meet the desire for high-efficient performance. In order to reduce the expense of hardware implementation, division is discarded in the process due to an appropriate selection of algorithms. This improvement makes it possible to fulfill RSA key pair generation with pure hardware design, which is quite different from prior methods that use software programming to do the division. The rest of paper is organized as follows. Section 2 describes the RSA key pair generation and details how to reduce timings related to prime finding algorithm which is the kernel of the whole process. Section 3 presents our architecture for it. The implementation results in this paper are compared finally, in section 4, the conclusions and future work are discussed.

II. DESIGN FLOW AND ALGORITHM ANALYSIS

Common RSA key pair generator generally includes a stage of trial division in the sieve function procedure. We investigate in this section a way of how

to avoid this by utilizing invertible numbers which is quite suitable for the hardware implementation. Some algebraic techniques are introduced to speed up modular exponentiation, so the primality test, which is the most time-consuming section practically, can be optimized greatly to achieve a satisfactory implementation; and the private key generation algorithm is improved as well for the sake of limited computation resources in our embedded processor.

Figure 1 gives out a common procedure of the RSA key pair generation.

2.1. Sieve Function

The purpose of the sieve function is to reduce the times of the primality test which is the most time-consuming part of RSA key pair generation. Unlike ordinary sieve functions which use small primes to divide the candidate number, trial divisions is replaced by one-time modular exponentiation as shown in Algorithm 1. According to the architecture presented in this paper, modular exponentiation can be done very fast and no extra hardware consuming is needed. So the processing time of the sieve function can be trivial and circuit area will be saved as well.

2.2. Primality Test

The candidate must be tested for primality in order to be useful for the generation of a RSA key pair. In this section, Miller-Rabin's method is used for our primality test. Since Miller-Rabin test is dominant in the processing time of RSA key pair generation, it is important to improve this part for the sake of high performance. Take notice of $b \leftarrow au \bmod n$ which is the most timing-cost computation in Algorithm 2, we focus on this part and make use of RNS.

2.3. Calculation of Private Key

The private key is the modular inverse of the public key .The particular method chosen for computing modular inverse avoids trial division to make it easier to be implemented in hardware.

2.4. RNS Montgomery Modular Multiplication

Modular multiplication is the kernel operation of RSA key pair generation which is called in quantity and takes up the most time of the whole procedure, it is of great importance to analyze this part and bring out the optimal algorithm for

The hardware we present in this paper.

Algorithm 1 Invertible Number Generation

Input: randomly selected $k\_ = \Pi 74$

i=1 $pi;\ \_(\_) =$

$lcm(p1 - 1; p2 - 1; :::; p74 - 1); pi = 3; 5; 7; :::; 379$

$L = 8 \times \_; M = 3 \times$ pi

Output: q co-prime with the smallest 74 odd primes

if $k\_(\_) \bmod \_ = 1$ then

$q = k + L$

else $k = k + 1$, go to 1

end if

if q is even then

$q = q + M$

end if

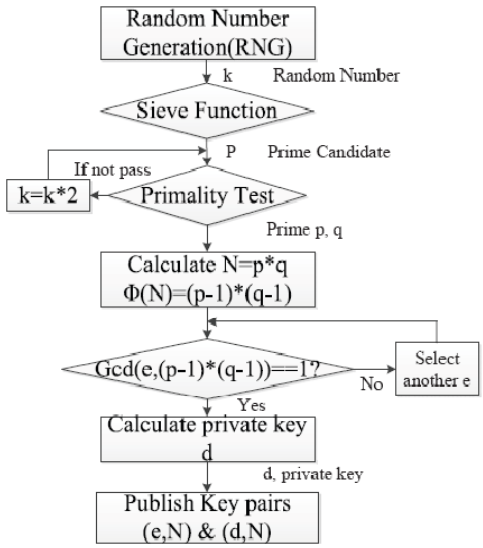


Figure 1 A common flow of RSA key pair generation.

Algorithm 2 Miller-Rabin Test

Input: Odd integer $n \geq 3$

Output: Confirm the primality of n

Find u and k so that $n - 1 = u * 2k$

Let a be randomly chosen from $2; 3; :::; n - 1$

$b \leftarrow a\ u \bmod n$

if $b \in 1; n - 1$ then

print "n is prime"

end if

for $i = 1$ to $k - 1$ do

$b \leftarrow b2 \bmod n$

if $b = n - 1$ then

print "n is prime"

end if

if $b = 1$ then

print "n is composite"

end if

end for

print "n is composite"

Algorithm 3 Improved Stein's method for modular inverse

Input: e(public key),$\_(n)$($\_(n) = (p - 1) \times (q - 1)$,$p; q$ are two primes)

Output: $d = e$ 1 mod $\_(n)$

$(X1;X2;X3) = (1; 0; e)$, $(Y1; Y2; Y3) = (0; 1; (n))$

while $X3$ □= 1 or $Y3$ □= 1 do

if $X3$ is even then

if $X1;X2$ are even then

$(X1;X2;X3) = (X1=2;X2=2;X3=2)$

else

$X1 = X1 + \_(n), X2 = X2 - e$

end if

else if $Y3$ is even then

if $Y1; Y2$ are even then

$(Y1; Y2; Y3) = (Y1=2; Y2=2; Y3=2)$

else

$Y1 = Y1 + \_(n), Y2 = Y2 - e$

end if

else if $X3 > Y3$ then

$(X1;X2;X3) = (X1 - Y1;X2 - Y2;X3 - Y3)$

else

$(Y1; Y2; Y3) = (Y1 - X1; Y2 - X2; Y3 - X3)$

end if

if $X3 = 1$ then

return $X1 = e$□1 mod $\_(n)$

else

return $Y1 = e$□1 mod $\_(n)$

Algorithm 4 RNS Montgomery Modular Multiplication

Input: $|X|a[b; |Y|a[b; (X; Y < 2N)$

Output: $[r]a[b; (r = XY A$□1)mod$N, r < 2N$

for i=1 to k do

step1: $zi = (xi \times yi)$ modai

step2: $qi = (zi \times | - N$□1$|ai)$modai

end for

step3: $qj = BT(qi; 0)$

for j=1 to k do

step4: $zj = (xj \times yj)$ mod bj

step5: $wj = (zj + qj \times Nj)$ mod bj

step6: $rj = (wi \times |A$□1

$i |bj) (i = j)$

end for

step7: $ri = BT(rj ; 0x80000000)$

Algorithm 5 Base Extension *qj = BT (qi; _)*

Input: $[q]a$

Output: $[q]b; qr$

for i=1 to k do

$li = qi \times |A$□1

$i |mi$

end for

$w = (\_ +$

Σk

i=1 $li) >> 32$

for j=1 to k do

$qj = |$

Σk

n=1 $|An|ai \times li|mi$

end for

return $[q]b$

III. IMPLEMENTATION FOR RSA KEY PAIR GENERATION

In this section, a TTA-like architecture is presented to illustrate the fast RSA key pair generation in our way. The implementation results are compared with in section 3.

4.1. Architecture Design

Transport Triggered Architecture (TTA) is statically programmed ILP modular architecture which

is similar to VLIW architecture. Instead of specifying operation typing and controlling the FUs directly, TTAs specify the required data transports. These transports may trigger operations as side effect implicitly. Figure 2 shows the difference of instruction format between VLIW and TTA. The proposed architecture in this design is shown in Figure 3. It mainly consists of five parts: FUs, RFs, control logic, transport network and on-chip RAM/ROM. Similar to common processors, the control logic composes of instruction fetch, instruction decoder and PC control units. In this design, JMP unit can affect the PC value to realize jump, branch and loop operations.

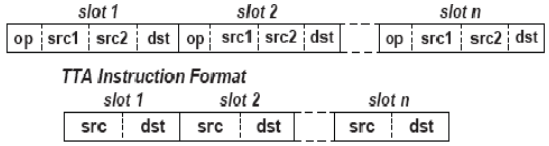


Figure 2.The difference of instruction format between VLIW and TTA

The FUs are the key factors which decide the performance of this processor. According to different applications, various of FUs can be designed and attached to the transport network. To implement RNS Montgomery multiplication algorithm, modular multiplication and modular multiplication and-accumulation are the key operations in n-bit level, where n is the base size. So, the MMAC units are designed to accelerate the execution speed of these key operations. MMAC units can only do 32-bit× 32-bit modular multiplication, But combing them together with RNS Montgomery multiplication algorithm, long bit modular multiplication can be done very fast. In this work, four MMAC units are designed according to the maximum number of function units used in the processing of executing the operations in Figure 4 before TIME E which refers to step 1 from step 3 in algorithm 4;

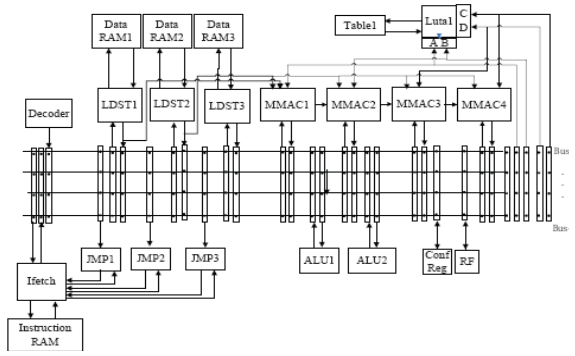


Figure 3. The proposed architecture design

The remaining operations after TIME E also depend on these four MMAC units in order to increase their usability of the architecture and save the valuable hardware resources. Two ALU units are included to implement the modular addition and modular subtraction operations. And some controlling operations are needed such as logical right-shift, arithmetic right-shift and case-select. They are included in ALU units to help finish the extra operation in the process of RSA key pair generation as well. There are three Load-store units (LDST), one Look-up Table unit (LUT). LDSTs are connected with the independent Data RAM. LUT are used to store the pre-computed data. There are direct data paths between LDSTs and MMAC units and also between LUT and the group of MMAC, which are used to reduce redundant data transmission. The transport network is used to transport data from source register to destination register. Four buses are adopted in the transport network to fully exploit the computation capacity of four MMAC units in parallel. The width of each bus is 32-bit which is decided by the selected base size.Figure 5 illustrates the operational process of RSA key pair generation from function unit level. The whole process has three stages:

1) Sieve Function

2) Primality Test

3) Private Key Calculation

According to the algorithm analysis in section 2, these function units are arranged to fulfil the RSA key pair generation: MMAC units do the modular exponentiation; ALU units complete the operation of addition and subtraction To compare the performance of our implementation with other works, the processing time consumed in primality test, prime finding and RSA key pair generation is analyzed. Because of the architecture specified for modular multiplication, the consuming time in primarily test is reduced greatly and numbers of times in prime finding are 36 which is a satisfactory compromise in both circuit area and processing time. As shown in Table 1, the proposed work requires less clock cycles than the other works.

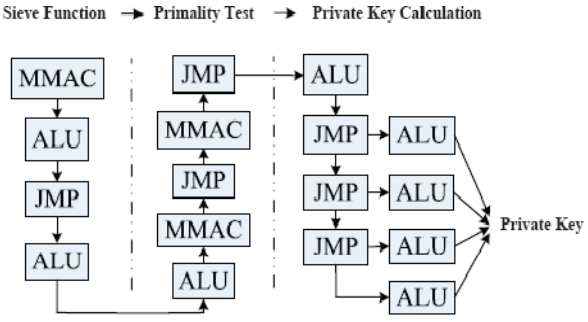


Figure 4. The operation process of RSA key pair generation from functional unit level

	Ref [3]	Ref [4]	This work
Frequency (MHz)	32	10.52	100
Primality test (ms)	84.2	25.29	4.1
prime finding (ms)	3100	/	151
RSA key pair generation (ms)	6820	/	306

Table1. RSA key pair comparison

IV. RESULT

The design was implemented by Verilog HDL and the processing time of 1024-bit RSA key pair generation is 306 ms in average, and the logic area is 131k gates. To compare the performance of our implementation with other works, the processing time consumed in primality test, prime finding and RSA key pair generation is analyzed. Because of the architecture specified for modular multiplication, the consuming time in primarily test is reduced greatly and numbers of times in prime finding are 36 which is a satisfactory compromise in both circuit area and processing time.

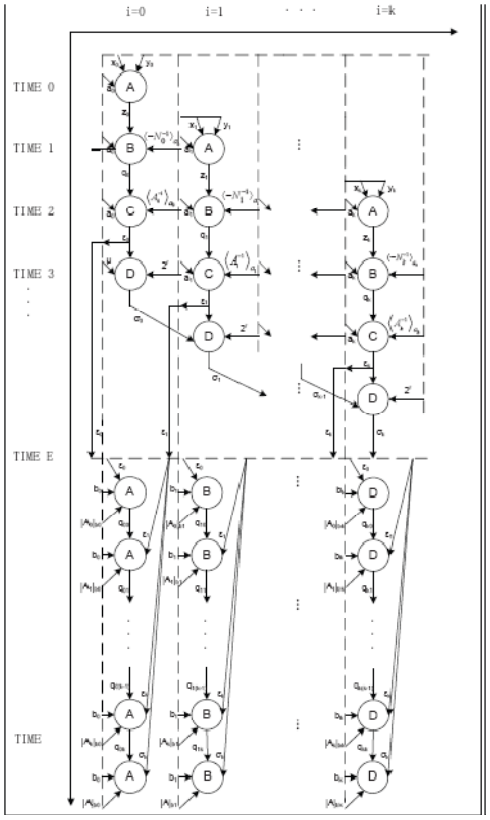


Figure 5. The operation process of MMAC unit

V. CONCLUSION AND FUTURE WORK

This paper presents a novel RSA key pair generation hardware implementation based on TTA, and Montgomery modular multiplication based on RNS is adopted in both sieve function and primality test to improve the performance significantly. FUs suiting the algorithms is designed on TTA,and direct data paths are used to reduce redundant data transmission. Above all, pipeline and parallel technology to improve the computing speed are introduced. At the frequency of 100 MHz, 1024-bit RSA key pair generation needs 306 ms in average, the logic area of the proposed architecture consists of 131k gates. This result shows that our proposed work can achieve high performance and small area for RSA key pair generation.

On-going and future developments include: (1) Preparation for some pre-computed data especially in RNS Montgomery multiplication can be optimized which affect the rate of RSA key pair generation significantly. (2) The conceptof scalable and reconfigurable architecture is introduced, in which not only 1024-bit RSA key pair but also 2048, 4096bit can be implemented in this platform.

REFERENCES

[1] R. L. Rivest, A. Shamir, L. Adleman, ”A Method for Obtaining Digital Signatures and Public-key Cryptosystems,” Communications of the ACM, vol. 21, pp. 120-126, 1978

[2] Chenghuai Lu, Andre L. M. dos Santos, Francisco R. Pimentel, ”Implementation of Fast RSA Key Generation on Smart Cards,” Proceedings of the ACM Symposium on Applied Computing, pp. 214-220, 2002

[3] Bahadori M, Mali M. R, Sarbishei O, Atarodi M, Sharifkhani M, ”A Novel Approach for Secure and Fast Generation of RSA Public and Private Keys on SmartCard,” 2010 8th IEEE International NEWCAS Conference (NEWCAS 2010),pp. 265-268, 2010

[4] Cheung Ray C. C, Brown Ashley, Luk Wayne, Cheung Peter Y K, ”A Scalable Hardware Architecture for Prime Number Validation,” Proceedings - 2004 IEEE International Conference on Field-Programmable Technology, FPT ’04, pp. 177-184, 2004

[5] Laurent Imbert, Jean-Claude Bajard, ”A Full RNS Implementation of RSA,” IEEE Transactions on Computers, vol. 53,2004, pp. 769-774.

[6] Wei Guo, Jizeng Wei, Yongbin Yao et al, "Design of a configurable and extensible Tcore processor based on Transport Triggered Architecture," World Congress on Computer Science and Information Engineering, pp. 536-540, 2009

[7] Hanae Nozaki, Masahiko Motoyama, Atsushi Shimbo et al,"Implementation of RSA Algorithm Based on RNS Montgomery Multiplication," Cryptographic Hardware and Embedded Systems(CHES), pp. 364-376, 2001

[8] Kenneth R, Castleman, "Digital image processing[M]," NewJersey:Prentice-Hall, 1996

[9] Joye M, Paillier P, Vaudenay S, "Efficient generation of prime numbers," Cryptographic Hardware and Embedded Systems-CHES 2000. Second International Workshop. Proceedings(Lecture Notes in Computer Science Vol. 1965), pp. 340-54,2000

[10] Maurer U. M, "Fast generation of prime numbers and secure public-key cryptographic parameters," Journal of Cryptology,pp. 123-55, 1995

◈◈◈

Novel QR Decomposition Methods for LTE Standard

***K.Kalyani, **N.Yoga & ***S.Rajaram**

Department of Electronics and Communication Engineering
Thiagarajar College of Engineering, Madurai – 625015, Tamil Nadu, India
E-mail:k_kalyani@tce.edu,yoga@tce.edu, rajaram_siva@tce.edu

Abstract - Multiple input multiple output (MIMO) transmission is an emerging technique targeted at 3G long term evolution (LTE) systems. One vital baseband function in MIMO receivers is QR decomposition of the channel matrix. In this paper, QR decomposition using Modified Gram-Schmidt and household transformation are presented. The multiplications and divisions used in these techniques are replaced by repeated addition and repeated subtraction .Thus the proposed design has minimal hardware and computational complexity to meet the requirements of LTE standard. The VHDL coding for these methods are downloaded onto Xilinx xc3s1000-4fg456 and results are verified.

Keywords- *Field Programmable Gate Array (FPGA), 3rd Generation Partnership Project (3GPP), Long Term Evolution (LTE), Multiple-Input Multiple-Output (MIMO) - Orthogonal Frequency-Division Multiplexing (OFDM), QR Decomposition (QRD), Modified Gram-Schmidt (MGS).*

I. INTRODUCTION

Wireless communications has developed into a key element of modern society. Mobile wireless communication devices have expanded dramatically from their inception as mobile telephones. The LTE is used to provide an extremely high performance radio-access technology that offers full vehicular speed mobility and that can readily coexist with HSPA and earlier networks. Because of scalable bandwidth, operators will be able to easily migrate their networks and users from HSPA to LTE over time.

The 3GPP LTE receiver systems with multiple transmit and receive antennas (MIMO) present a better performance on two different angles, the diversity and the multiplexing [1]. MIMO exploits higher transmission rates, higher spectral efficiencies, greater coverage, improved link robustness, without increasing total transmission power or bandwidth [1]. But, the 3GPP LTE receiver also increases the computational and the hardware complexities greatly. It is a challenge to realize the receiver with the MIMO OFDM system with minimal hardware complexity and power consumption—especially the computational complexity—in VLSI implementation. The QR Decomposition is one of the computational complexity modules in the baseband function of 3GPP LTE receiver. It transforms a complex-valued channel matrix **H** to a decomposition of an orthogonal **Q** and an upper triangular **R** matrices. The matrices are required by a list sphere detection (LSD) algorithm, which detects the received, complex-valued symbols. There are several ways to obtain QR decomposition. In this work, in order to realize 3GPP LTE receiver with the MIMO-OFDM with minimal hardware and computational complexity, we have proposed novel multiplier less Modified Gram-schmidt and household algorithms to perform QRD.

II. DESIGN ISSUE IN QR DECOMPOSITION FOR LTE RECEIVER

A high-level description of the targeted 2-antenna MIMO OFDM receiver is presented in Figure 1. The input ports are connected to radio-frequency functions of the receiver. The upcoming 3GPP long term evolution (LTE) standard receiver will support data rates up to 100 Mbps [8]. Such a high data rate will be achieved in 20MHz bandwidth by using transmission techniques like orthogonal frequency division multiplexing (OFDM) , multiple-input multiple-output (MIMO) , that is, the use of multiple antennas, and an efficient forward error correction method, the turbo coding [8].

The 3GPP LTE standard based on the MIMO-OFDM system provides very high data throughput rate because the technique of the MIMO can increase the data rate by extending an OFDM based system. However, the 3GPP LTE standard also increases the computational and the hardware complexities greatly, compared with the current WLAN standards. Symbol estimation at the receiver is intractable with multiple antennas and large constellations. The symbol estimation can be simplified by using QR decomposition of channel matrix. With this practice, the computational

complexity is lowered. By using very sophisticated algorithm, area utilization of QRD can be reduced.

There are several ways to obtain QR decomposition. It can be computed, e.g, using Householder transformations, Givens rotations, or Gram-Schmidt process. However, quite often some variations of the basic algorithms are used. The inherent regularity of matrix operations can be utilized with systolic structures [2]. Elementary operations can be alleviated with a CORDIC algorithm [3] which lends itself to low-complexity hardware realization. Such an approach was followed, e.g., in [4, 5, 6]. Another way to alleviate hardware complexity is to carry out computations in logarithmic domain. This practice is used in [7]. The MIMO receiver requires relatively small matrix size and low processing speed for QR decomposition. Therefore, extensively parallel solutions like systolic array processors in [2, 6] or processor arrays with reduced dimensions [4] are not justified to be used for such systems. Givens rotations require a more sophisticated degree of understanding in VLSI design if CORDIC arithmetic is employed. So the household and Modified Gram-schmidt algorithms are considered here. Gram-Schmidt and household prevail in software programs that use floating-point arithmetic. Furthermore, Multiplications and divisions used in household and Modified Gram-schmidt algorithm incur high area overhead. In this work, in order to reduce these overheads, and to realize 3GPP LTE receiver with the MIMO-OFDM with minimal hardware and computational complexity, we have proposed novel multiplier less Modified Gram-schmidt and household algorithms to perform QRD.

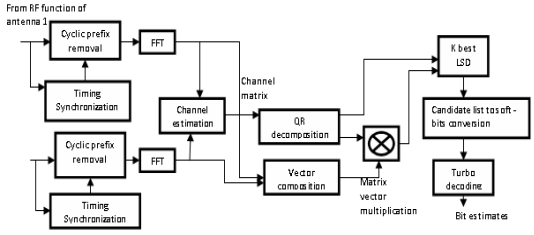


Fig.1 A Simplified Diagram of Base Band Processing of a Two Antenna MIMO OFDM Receiver.

2.1 QR Decomposition in MIMO System modeling and Detection

In principle, an MIMO system [8] with a complex-valued channel matrix, H, noise vector, n, transmitted symbol, s, and received symbol, y, can be described with

$$y = Hs + n \qquad (1)$$

The number of receive and transmit antennas equals the numbers of rows and columns of **H**, respectively. The transmitted symbol **s** can be estimated by ML detection by solving $s = \arg\min\left\|y - Hs\right\|^2$ (2) which gives the optimal result. However, solving (2) is intractable with multiple antennas and large constellations. Instead of solving (2), the symbol estimation can be simplified by using QR decomposition of **H**. With this practice, the computational complexity is lowered. Instead of ML detection, a substitute

$$s' = \arg\min\left\|y - Rs\right\|^2 \qquad (3)$$

where $y' = Q^H y$

R is in upper triangular form, approximation of **s** is computationally simpler with the aid of (3). In this paper, QR decomposition using Modified Gram-Schmidt and household transformation are presented.

2,2 QR decomposition using Modified Gram-Schmidt algorithm

The modified Gram-Schmidt algorithm is used for the QR decomposition. The modified Gram-Schmidt algorithm has better numerical properties than the classical Gram-Schmidt algorithm. In principle, the algorithm orthogonalizes a set of vectors. The modified Gram-Schmidt algorithm [9] is given for square matrix $\mathbf{H}_{n\times n}$.

for $k = 1:n$

$R_{k,k} = \left\|H_{1:n,k}\right\|$

$Q_{1:n,k} = H_{1:n,k} / R_{k,k}$

for $j = k+1:n$

$R_{k,j} = Q_{1:n,k}^H H_{1:n,j}$

$H_{1:n,j} = H_{1:n,j} - Q_{1:n,k} R_{k,j}$

end

end

2.3 QR decomposition using Household transformation

A Householder reflection (or *Householder transformation*) is a transformation that takes a vector and reflects it about some plane or hyperplane. This operation is used to calculate the QR factorization of an m-by-n matrix *A* with m ≥ n. *Q* can be used to reflect a vector in such a way that all coordinates but one disappear. The algorithm for household transformation [10] is given below.

Matrix A

$R_0 = A$

$Q_0 = I$

for $i = 1$ *to* n

for $j = i$ *to* n

$a_j <= Rj,i$

end

$g = \sqrt{\sum_{j=i}^{n} a_j^2}$

$u_i = a_j + sign(a_i)ge$

$t = \sqrt{\sum_{i=0}^{n} u_i^2}$

$v_i = \frac{u_i}{t}$

$H_i = I - 2v_i v_i^T$

$R_i = H_i R_{i-1}$

$Q_i = H_i Q_0$

end

$Q_n = (Q_{n-1} \ldots . Q_2 \, Q_1)^T$

A straightforward implementation of these algorithms would require square root function for distance computations, multiplication and division operation, but they are demanding to implement on hardware. In this paper, multiplications are replaced by repeated additions and divisions are replaced by repeated subtraction to reduce the area.

III. NOVEL ARCHITECTURES FOR QR DECOMPOSITION WITHOUT USING MULTIPLICATION AND DIVISION

The Modified Gram-schmidt and Household transformations based QR decomposition are proposed without use of multiplications and divisions. In this work, the CORDIC based Given's rotation QR decomposition is proposed.

3.1 Proposed Modified Gram-Schmidt based QR decomposition without using multiplication and division

Gram–Schmidt algorithm obtains the orthogonal basis spanning the column space of the matrix by the orthogonality principle. Through a series of projection, subtraction, norm and division, the column vector of the unitary matrix containing the orthogonal basis can be acquired one by one. The upper triangular matrix is also generated as a by-product [11]–[13] which incur high area overhead. Thus the MGS based QRD is proposed as shown in Figure 1.

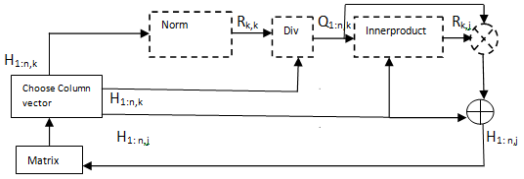


Fig2.Modified Gram-Schmidt based QR Decomposition without using Multiplication and division.

In Fig 2, the dotted line boxes where the multiplications and divisions used in MGS are replaced by repeated addition and repeated subtraction which occupy less number of slices and 4 input LUTs compared to conventional MGS for QR Decomposition.

Proposed Household transformation based QR decomposition without using multiplication and division

Householder transformation (HT) tries to zero out the most elements of each column vector at a stroke by reflection operations. The upper triangular matrix is derived after each transformation matrix being applied to every column vector sequentially. The unitary matrix involves the multiplications of these Householder transformation matrices and thus the complexity is much higher [14], [15]. To compensate for this, in the propose work, multiplications are replaced by repeated additions and divisions are replaced by repeated subtraction as given in fig 3.

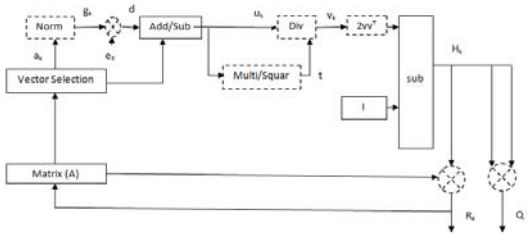


Fig 3.Household transformation based QR Decomposition without using Multiplication and division.

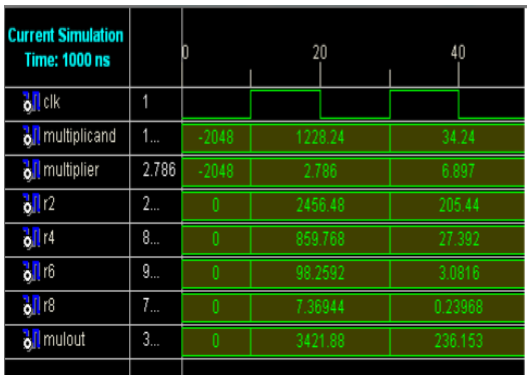
From the Fig 3, the dotted line boxes where the multiplications and divisions used in Household transformation are also replaced by repeated addition and repeated subtraction which occupy less number of slices and 4 input LUTs compared to conventional Household transformation for QR Decomposition.

IV. SIMULATION AND IMPLEMENTATION

The VHDL coding for QR decomposition for LTE is simulated and downloaded onto Xilinx xc3s1000-4fg456. The simulated results and comparison result were given below.

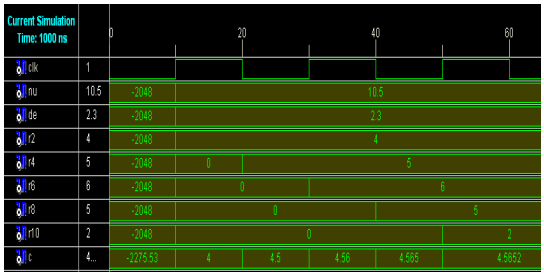
4.1 Simulation Results

4.1.1 Simulation Result for multiplication using repeated addition



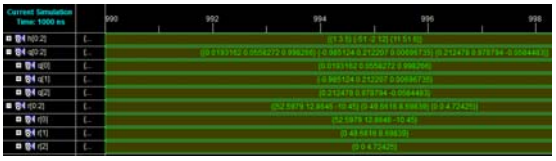
The multiplication using repeated addition is simulated by, taking 'multiplicand' as multiplicand, 'multiplier' indicates how many number of times multiplicand will be added, r2,r4,r6,r8 are intermediate results. To verify this simulation result with manual, the applied input values are, multiplicand=1228.24, multiplier=2.786 then intermediate results are obtained by, adding multiplicand 1228.24 two times r2=2456.48, adding 122.824 seven times r4= 859.768, adding 12.2824 eight times r6=98.2592 and adding 1.22824 seven times r8=7.36944 then final output, mulout is obtained by addition of r2+r4+r6+r8=3421.88 which is same as manual multiplication result.

4.1.2 Simulation result for division using repeated subtraction



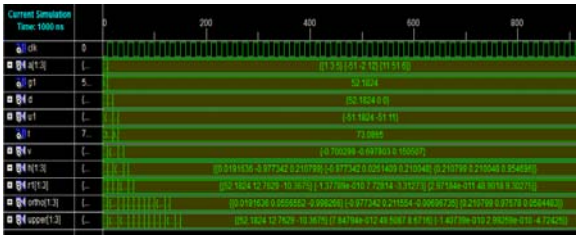
The division using repeated subtraction is simulated by, taking 'nu' as numerator, 'de' as denominator , r2,r4,r6,r8,r10 are intermediate results and 'c' as final output. To verify this simulation result with manual, the applied input values are, nu=10.5, de=2.3 then intermediate results are obtained by, repeatedly subtracting 'de' from 'nu' until the subtracted value becomes less than 'de' then r2=5 indicates that how many number of times 'de' subtracted from 'nu', the remaining value is again subtracted from 'nu' to find r4=6, this process continue to find r6=6, r8=5 and r10=2. Then final output is 'c'=4.5652 which is same as manual division result.

4.1.3 Simulation result for Modified Gram-Schmidt based QR decomposition without using multiplication and division



The QR decomposition using Modified Gram-Schmidt algorithm is simulated by, taking 'h' as input channel matrix, 'q' as output orthogonal matrix and 'r' as upper triangular matrix. To verify this simulation result with manual calculation, the applied input 'h' = [{1 3 5} {-51 -2 12} {11 51 6}] then the output obtained from this simulation are 'r' =[{52.5979 12.8646 -10.45} {0 49.6616 8.69839} {0 0 4.72425}] and 'q'=[{0.0193162 0.0558272 0.998266}{-0.985124 0.212207 0.00696735}{0.212478 0.978794 -0.0584483}]. This result is same as that of result obtained from manual calculation.

4.1.4 Simulation result for Household transformation based QR decomposition without using multiplication and division



The QR decomposition using household transformation is simulated by, taking 'a' as input channel matrix, 'ortho' as output orthogonal matrix and 'upper' as upper triangular matrix. To verify this simulation result with manual calculation, the applied input 'a' = [{1 3 5} {-51 -2 12} {11 51 6}] then 'g1' = 52.1824 is the norm of first column vector of given matrix, 'u1'= {-51.1824 -51 11} is subtraction of 'd' from first column vector of 'a', 't'= 73.0865 is norm of u1,'v'=(-0.700299 -0.697803 0.150507} is division of u1 by t, then final output obtained from this simulation are 'upper'=[{52.1824 12.7629 -10.3675} {0 49.5087 8.6716} {0 0 -4.72425}] and 'ortho'= [{0.0191636 -0.977342 0.210799}{0.0556552 0.211554 0.97578}{-

0.998266 -0.00696735 0.0584483}. This result is same as that of result obtained from manual calculation.

4.2 Implementation on FPGA

The VHDL coding for conventional Modified Gram-Schmidt, Modified Gram-Schmidt without using multiplications and divisions, conventional household transformation and household transformation without use of multiplication and division are downloaded onto Spartan xc3s1000-4fg456 and the comparison results are given below.

4.2.1 Comparison of synthesis report for QR Decomposition using Modified Gram-Schmidt algorithm

Table I: Comparison of conventional Modified Gram-schmidt QRD with Modified Gram-Schmidt without using multiplications and divisions.

S.No	Logic Utilization	Available	Conventional MGS	MGS without using multiplier and division
1. 2.	Slices 4 input LUTs	7680 15360	64% 60%	28% 26%

From the report, Modified Gram-Schmidt QR decomposition have improvement in terms of area when the multiplications are replaced by repeated additions and and divisions are replaced by repeated subtractions. Thus the proposed architecture reduces the hardware complexity.

4.2.2 Comparison of synthesis report for QR Decomposition using Household transformation

S.No	Logic Utilization	Available	Conventional household transformation	Household without using multiplier and division
1. 2.	Slices 4 input LUTs	7680 15360	45% 42%	19% 17%

Table2. Comparison of conventional Household QRD with Household QRD without using multiplications and divisions.

From the report, Household QR decomposition have improvement in terms of area when the multiplications are replaced by repeated additions and divisions are replaced by repeated subtractions. Thus the proposed architecture reduces the hardware complexity.

V. CONCLUSION

Thus the proposed hybrid architecture reduces the overhead of hardware and computational complexity of 3GPP LTE receiver. Thus it is evident from the comparison result that the number of slices and 4 input LUTs required in FPGA implementation of QR Decomposition is tremendously reduced thereby making the proposed design cost effective and computationally efficient. By reduction of complexity of the highest computational module the 3GPP LTE standard will become efficient. The proposed QR Decomposition using MGS and Household can meet the specifications of most OFDM communication systems, including VDSL, 802.16, DAB and DVB. The proposed QR decomposition methods are synthesized and implemented on FPGA. In future, this work will be extended to implement FFT, K-best LSD and turbo decoding of 3GPP LTE receiver on FPGA.

VI. ACKNOWLEDGMENTS

The authors would like to thank TIFAC (Technology Information Forecasting and Assessment Council) and Thiagarajar college of Engineering, Madurai for supporting this research. In particular, we acknowledge the permission provided by the organizing committee of Integration journal for giving an opportunity to submit and review our paper .

REFERENCES

[1] Lee, K.F. and Williams, D.B.: A space-frequency transmitter diversity technique for *OFDM* systems. In Proc. Global Telecommunications Conf., San Francisco, CA, pp. 1473-1477. (Nov. 2000)

[2] S. Y. Kung, *VLSI Array Processors*. Upper Saddle River, NJ, USA: Prentice-Hall, 1987.

[3] R. Andraka, "A survey of cordic algorithms for fpga based computers," in *Sixth ACM/SIGDA Int. Symp. on Field- Programmable Gate Arrays*, Monterey, CA, USA, Feb. 1998, pp. 191–200.

[4] Z. Liu, K. Dickson, and J. V. McCanny, "Application-specific instruction set processor for SoC implementation of modern signal processing algorithms," *IEEE Tran. on Circuits and Systems*, vol. 52, no. 4, pp. 755–765, Apr. 2006.

[5] "Implementation of CORDIC-based QRD-RLS algorithm on Altera Stratix FPGA with embedded Nios soft processor technology," Altera Corporation, San Jose, CA, USA, white paper, Mar. 2004.

[6] A. Maltsev, V. Pestretsov, R. Maslennikov, and A. Khoryaev, "Triangular systolic array with reduced latency for QRdecomposition of complex matrices," in *IEEE Inter. Symp. On Circuits and Systems*, Kos, Greece, May 2006, pp. 385–388.

[7] C. K. Singh, S. H. Prasad, and P. T. Balsara, "VLSI architecture for matrix inversion using

modified Gram-Schmidt based QR decomposition," in *Int. Conf. on VLSI Design*, Bangalore, India, Jan. 2007, pp. 836–841.

[8] Perttu Salmela, Juho Antikainen,Teemu Pitk anen,Olli Siv'en,and Jarmo Takala.: 3G Long Term Evolution Baseband Processing with Application-Specific Processors. International Journal of Digital Multimedia Broadcasting Volume (2009), Article ID 503130, 13 pages

[9] *Perttu Salmela, Adrian Burian, Harri Sorokin, and Jarmo Takala,* "Complex valued QR Decomposition implementation for MIMO receivers",IEEE ICASSP 2008.

[10] Semih Aslan, Erdal Oruklu, and Jafar Saniie "*Realization* of *area efficient QR factorization using unified division*, *square root*, and *inverse square root hardware*", IEEE Electro/Information Technology, pp. 245-250, June 2009.

[11] C. K. Singh, S. H. Prasad, and P. T. Balsara, "VLSI architecture for matrix inversion using modified Gram–Schmidt based QR decomposition," in *Proc. Int. Conf. VLSI Design*, 2007, pp. 836–841.

[12] .R.C.-H. Chang, C.H.Lin,K. H. Lin, C.L. Huang, and F.C.Chen, "Iterative decomposition architecture using the modified Gram–Schmidt algorithm for MIMO systems," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 5, pp. 1095–1102, May 2010.

[13] P. Luethi, C. Studer, S. Duetsch, E. Zgraggen, H. Kaeslin, N. Felber, andW. Fichtner, "Gram–Schmidt-based QR decomposition for MIMO detection: VLSI implementation and comparison," in *Proc. IEEE Asia Pacific Conf. Circuits and Systems*, 2008, pp. 830–833

[14] C. F. T. Tang, K. J. R. Liu, and S. A. Tretter, "On systolic arrays for recursive complex Householder transformations with applications to array processing," in *Proc. Int. Conf. Acoustics, Speech, and SignalProcess.*, 1991, pp. 1033–1036.

[15] K.-L. Chung and W.-M. Yan, "The complex Householder transform," *IEEE Trans. Signal Process.*, vol. 45, no. 9, pp. 2374–2376, Sep. 1997.

◈◈◈

Investigation on THD Mitigation by using Shunt Active Filter

Y. Vijaya Suresh, U. Shantha Kumar

Dept. of EEE, RGM College of Engg. & Tech., Nandyal, Andhra Pradesh, India
E-Mail ID: yadativijayasuresh@yahoo.com & santha243@gmail.com

Abstract **-** This paper presents a new topology for multilevel Current source converter. The new converter uses parallel connections of full-bridge cells. Also by Adding or removing the full-bridge cells, modularized circuit layout and packaging is possible, where the number of output current levels can also be easily adjusted. Using adequate levels, the multilevel current converter generates approximately sinusoidal output current with very low harmonic distortion. Based on this converter a shunt active filter has been modeled. The simulation results of the lacking shunt active filter and through shunt active filter in diode bridge rectifier shows that the THD alleviation by means of shunt active filter.

Key words**:** *Multilevel Converter, Shunt Active Power Filter, Power Quality.*

I. INTRODUCTION

Recently multilevel power conversion technology has been a very rapidly growing area of power electronics with good potential for further developments. The most attractive applications of this technology are in the medium to high-voltage range[1] Multilevel converters work more like amplitude modulation rather than pulse modulation, and as a result:

- Each device in a multilevel converter has a much lower dv/dt
- The outputs of the converter have almost perfect currents with very good voltage waveforms because the undesirable harmonics can be removed easily,
- The bridges of each converter work at a very low switching frequency and low speed semiconductors can be used and
- Switching losses are very low [2].

The general function of the multilevel converter is to synthesize a desired output voltage from several levels of DC voltages as inputs. The DC voltage sources are available from batteries, capacitors, or fuel cells. There are three types of multilevel converters:

• Diode-Clamped Multilevel Converter

• Flying-Capacitor Multilevel Converter

• Cascaded-Converters with Separated DC Sources

The first practical multilevel topology is the diode-clamped multilevel converter topology and first introduced by Nabae in 1980 [3]. The converter uses capacitors in series to divide the DC bus voltage into a set of voltage levels. To produce N levels of the phase voltage, an N −level diode-clamp converter needs N-1 capacitors on the DC bus. The flying capacitor multilevel converter proposed by Meynard and Foch in 1992 [4], [5]. The converter uses a ladder structure of the DC side capacitors where the voltage on each capacitor differs from that of the next capacitor. To generate N −level staircase output voltage, N − 1 capacitors in the DC bus are needed. Each phase-leg has an identical structure. The size of the voltage increment between two capacitors determines the size of the voltage levels in the output waveform. The last structure introduced in the paper is a multilevel converter, which uses cascade converters with separate DC sources and first used for plasma stabilization [6], it was then extended for three-phase applications [7]. The multilevel converter using cascaded-converter with separate DC sources synthesizes a desired voltage from several independent sources of DC voltage. A primary advantage of this topology is that it provides the flexibility to increase the number of levels without introducing complexity into the power stage. Also, this topology requires the same number of primary switches as the diode-clamped topology, but does not require the clamping diode. However, this configuration uses multiple dedicated DC-busses and often a complicated and expensive line transformer, which makes this a rather expensive solution. In addition, bidirectional operation is somewhat difficult (although not impossible) to achieve [8]. Modularized circuit layout and packaging is possible because each level has the same structure, and there are no extra clamping diodes or voltage balancing capacitor. The number of output voltage levels can be adjusted by adding or removing the full-bridge cells

The converters that were focused upon were voltage source converters, with multilevel voltage waveforms. These converters divide the total input voltage among a number switches, and allow a reduction of the voltage harmonics. As mentioned, these are the most commonly used and best-understood multilevel converters. The most multilevel converters discussed in the literature are multilevel voltage source converters [9]. However, in many current applications, such as shunt active filters, active power line conditioners, VAR compensations etc., we need to use multilevel current converters. This paper presents a new multilevel current converter. Then the proposed multilevel current source converter is the core of a shunt active filter, which is obtained based on this converter. The proposed new multilevel current converter consists of a set of par- allel single-phase full- bridge converter units. The AC current output of each levels full-bridge converter is connected in parallel such that the synthesized current waveform is the sum of the converter outputs. In other words, for high current applications many switches can be placed in parallel, with their current summed by inductors.

II. THE PROPOSED MULTILEVEL CURRENT CONVERTER

II.1.The Proposed Topology

The full-bridge topology is used to synthesize a three-level square-wave output current waveform. The full-bridge configuration of the single-phase current source converter is shown in Fig1.

In as single-phase full-bridge configuration, four switches are needed. In full-bridge configuration, by

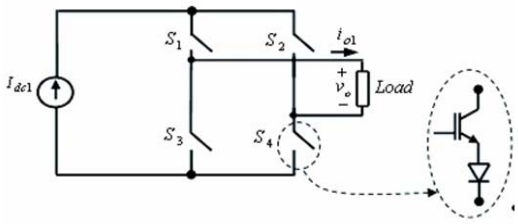


Fig.1. A multi level current converter

Turning the switches S_1 and S_4 on and S_2 and S_3 off a current of I_{dc1} is available at output i_{o1}, while reversing the ing the operation we get current of i_{dc1}. To generate zero level of a full-bridge converter, the switches $S_1$and S_3 are turned on while S_2 and S_4 are turned off or vice versa. The typical output waveform of full bridge of single-phase multilevel shown in Fig.1 is shown in Fig. 2

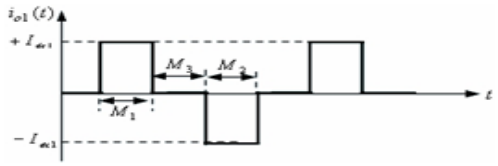


Fig.2. Typical output wave form of three level configuration

The three possible levels with respect to above discussion are shown in Table 1. Note that S_1 and S_2 should not be open at the same time, nor should S_3 and S_4. Otherwise, an open circuit would exist across the DC current source.

Table 1: output current with corresponding conditions

MODES	CONDUCTING SWITCHES	OUTPUT CURRNT
1	S1,S4	$+I_{dc1}$
2	S2,S3	$-I_{dc1}$
3	S1,S3 or S2,S4	0

Fig.3 shows equivalent circuits of the proposed topology at different modes.

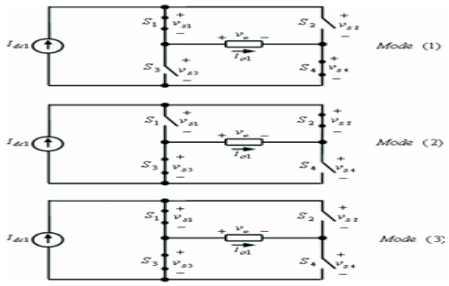


Fig.3. The Equivalent Circuits of the Proposed Topology at Different Modes

From Fig. 3, the instantaneous switches voltages of each module are given in Table2

Table 2: Instantaneous switches voltages

Mode	Vs1	Vs2	Vs3	Vs4
1	0	Vo(t)	Vo(t)	0
2	-Vo(t)	0	0	-Vo(t)
3	0	Vo(t)	0	-Vo(t)

Using parallel connections of many converters like the one shown in Fig. 1, we can synthesize multi-level current converter. The general function of this multilevel current source converter is to synthesize a desired current from several independent sources of DC currents. Fig. 3 shows a single-phase structure of a parallel converter with a separate DC current source. By different combinations of the four switches, S_1–S_4, each full-bridge converter can generate three different current outputs, $+I_{dc1}$, $-I_{dc1}$ and zero current. The AC outputs of each of the different level of full-bridge converters are connected in parallel such that the synthesized current waveform is the sum of the converter outputs. An output phase current waveform is obtained by summing the output currents of the converter bridges:

$$\text{Ion}\,(t) = \text{io1}\,(t) + i_{o2}\,(t) + \cdots + i_{on}\,(t) \qquad (1)$$

Where N is the number of parallel bridges bridges[10].

In the following we propose a new method for determining the levels of different DC current sources, which are used in the proposed multilevel converter.

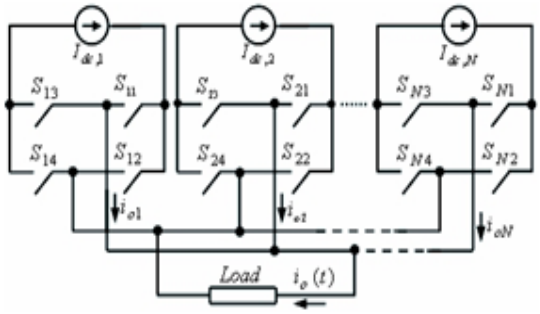


Fig.3. Single phase parllel multilevel current source converter

II.2. DETERMINING THE LEVELS

If all DC current sources in Fig. 3 are equal to I_{dc} the converter is then known as symmetric multilevel current source converter. With having a number of full-bridge converter units, this technique results in an output current of the converter that is almost sinusoidal.

The maximum output current of the N paralleled multilevel current source converter is

$$I_{MAX} = N * I_{dc} \qquad (2)$$

In this topology; the number of overall output current(S) is given by:

$$S = 1 + 2N \qquad (3)$$

For example, a 13-level multilevel current source converter using the technique can be implemented as shown in Fig. 4. In Fig. 4, i_{o1} to i_{o6} are DC current supplies, which are from either regulated inductors or separated DC sources.

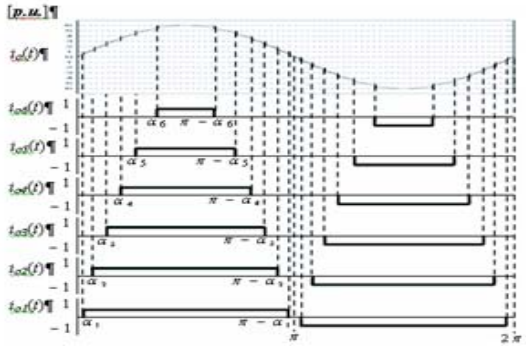


Fig.4 The 13 level converter output

III. THE SHUNT ACTIVE FILTER BASED ON MULTILEVEL CURRENT SOURCE CONVER-TER

III. 1.SHUNT ACTIVE FILTER PRINCIPLE

In recent years, the usage of modern electronic equipment has been increasing rapidly. These electronic equipments impose nonlinear loads to the AC main that draw reactive and harmonic current in addition to active current [11]. In order to overcome these problems, different kinds of active power filters, based on force-commutated devices, have been developed. Particularly, shunt active power filters, using different control strategies, have been widely investigated. These filters operate as current sources, connected in parallel with the nonlinear load generating the current and the current harmonic components required by the load. However, shunt active filters present the disadvantages that are difficult to implement in large scale where the control is also complicated. To reduce the drawbacks, the proposed solution in this paper is to use a multilevel current source converter. A shunt active filter consists of a controllable voltage or current source. This topology is shown in fig.5 it consisting of DC link capacitor C, power electronic switch and inductor Lf.

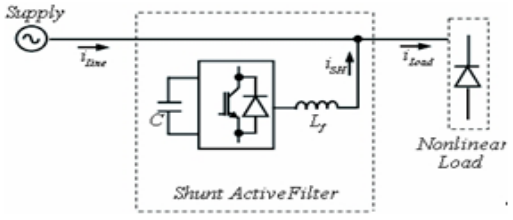


Fig.5 configuration of voltage source converter based on shunt active filter

III.2. SUGGESTED SHUNT ACTIVE FILTER

Fig.6 shows the schematic of the suggested shunt active power filter consisting of the new multilevel current source inverter with a control unit, to solve the power quality problems. The operation of the shunt current source multilevel inverters is based on the injection of current harmonic, i_{SH} , which is in phase with the load current, $i_{Load,}$ thus eliminating the harmonic current of the line(supply) current iLine. Now, suppose that the load current can be written as the sum of the fundamental and harmonic current as in equation (4)

$$i_{Load} = i_{Load,Fund} + i_{Load,Harmonics} \quad (4)$$

Then the injected current by shunt inverter should be:

$$i_{SH} = i_{Load,Harmonics} \quad (5)$$

With resulting the line current

$$i_{Line} = i_{Load} - i_{SH} \quad (6)$$

$$i_{Line} = i_{Load,Found} \quad (7)$$

As it is seen, the equation (7) only contains the fundamental component of the load current and thus free from the harmonics.

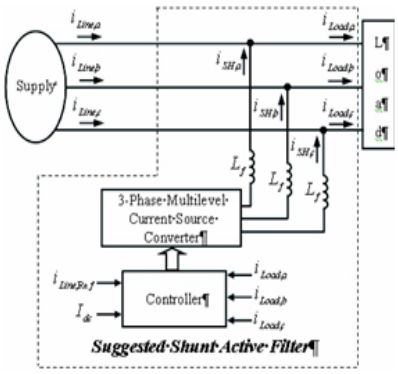


Fig.6. Suggested Shunt Active filter configuration

III. WITHOUT SHUNT ACTIVE FILTER

The industrial loads usually have complex nonlinear dynamics. In connecting nonlinearities to a power network, they induce some undesirable distortions to the sinusoidal signal of the network. For showing this effect, a three phase diode rectifier is used as a nonlinear load connected to grid Fig.7shows the circuit of a three-phase diode rectifier. The input phase voltages can be written as:

$$v_a = v_m sinw_i t \quad (8)$$

$$v_b = v_m \sin\left(w_i t - {}^{2\pi}/_{3} \right) \quad (9)$$

$$v_c = v_m \sin\left(w_i t + {}^{2\pi}/_{3} \right) \quad (10)$$

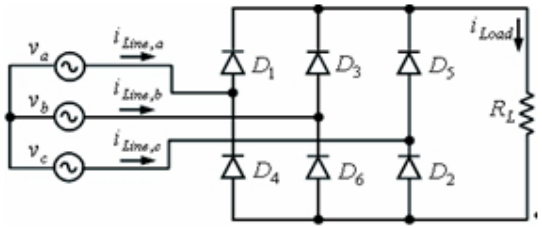


Fig.7. Three Phase Diode Rectifier as a Nonlinear Load

If the load is assumed a pure resistance, the output current Peak is:

$$I_{MAX} = \sqrt{3}V_m / R_L \quad (11)$$

In this study, the parameters of the system are as $V_m = 110\sqrt{2}V, w_i = 100\pi$ *and* $R_L = 40$ohms Fig.8 shows waveforms of input line voltages, load current and line currents. As the Fig.8 shows, nonlinear loads may pollute power lines seriously with their high levels harmonic current and reduction in power factor. Fig.9 depicts the Fast Fourier Transform of ac utility line current Harmonics up to the 20th have been considered. The THD of the input current of the rectifier also the ac utility line current is 30.87%.

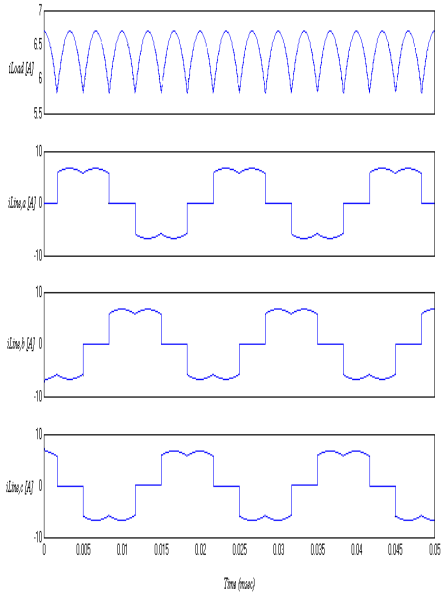


Fig.8. The Outputs of Diode Rectifier

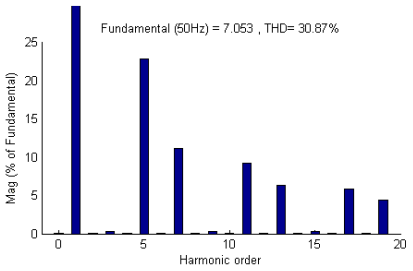


Fig.9. FFT analysis of the line current of Diode Rectifier

IV. WITH SHUNT ACTIVE FILTER

The ability of shunt active filters to suppress these problems has attracted a great deal of attention to these systems. This paper proposed a new structure for shunt active filter based on multilevel current source converter. For showing the capability of the proposed shunt active filter, a 13 level multilevel current source converter is simulated. Fig.10. Shows a single phase structure of the multilevel converter. The converter consists of seven full-bridges with all current sources are equal to *Idc*. Fig. 11 shows the load, line and shunt active power filter output currents. The shunt active power filter with multilevel current converter is able to successfully compensates reactive power and mitigate current harmonics distortions with excellent transient performance.

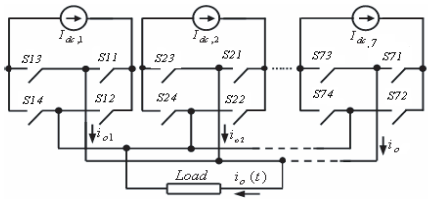


Fig.10. Single-Phase 13-Level Multilevel Current Converter Used in the Shunt Active Filter System

Fig.12. depicts the Fast Fourier Transform of ac utility line current Harmonics with shunt active filter up to the 20th have been considered. The THD of the input current of the rectifier also the ac utility line current is 10.33%.

V.CONCLUSION

In this paper, a new topology for multilevel current source converters has been presented. The most important feature of the system is being convenient for expanding and increasing the Number of output levels. The proposed strategies generate a current with minimum error with respect to the sinusoidal reference. Therefore, it generates very low harmonic distortion.

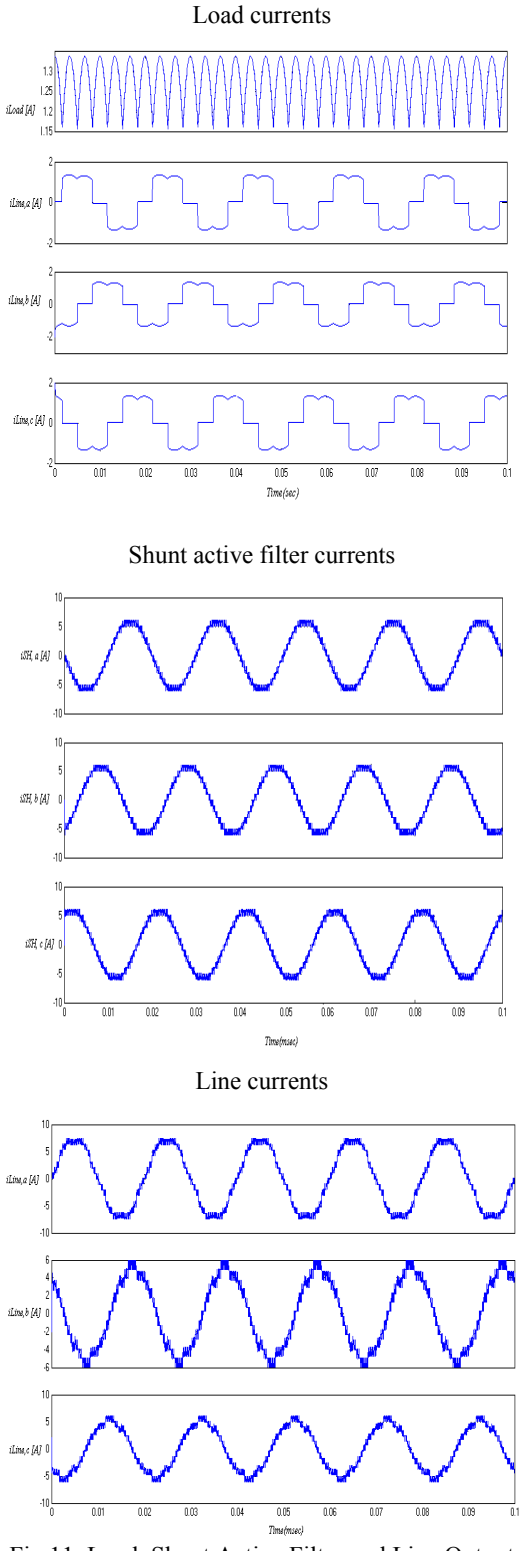


Fig.11. Load, Shunt Active Filter and Line Output Currents

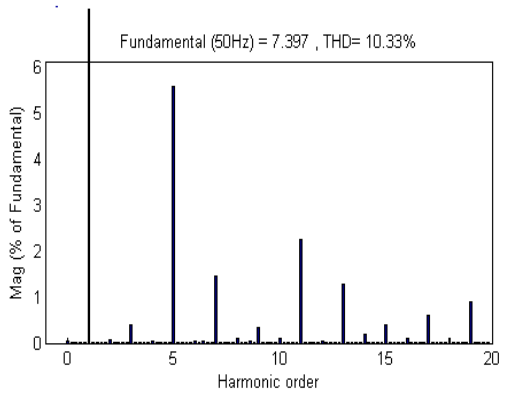


Fig.12. FFT analysis of the line current $i_{A,}$ at ac utility connected through shunt active filter

REFERENCES

[1] Nikola Celanovic, "Space Vector Modulation and Control of Multilevel Converters," Ph.D. Thesis, Blacksburg, Virginia, Sep. 2000.

[2] Giri Venkataramanan, and Ashish Bendre," Reciprocity-Transposition-Based Sinusoidal Pulse width Modulation for Diode-Clamped Multilevel Converters," IEEE Transaction on Industrial Electronics, Vol. 49, No. 5, pp. 1035-1047, Oct. 2002.

[3] Nabae, I. Takahashi and H. Akagi, "A New Neutral-Point Clamped PWM Inverter," IEEE Transactions on Industry Applications, Vol. IA-17, No.5, pp. 518-523, September=October 1981.

[4] T. Meynard and H. Foch, "Multi-Level Conversion: High Voltage Choppers and Voltage Source Inverters," IEEE PESC92, pp. 397403, 1992.

[5] T. Meynard and H. Foch"Imbricated Cells Multi-Level Voltage Source Inverters for High Voltage Applications," European Power Electronics Journal, Vol. 3, No. 2, pp. 99106, June 1993.

[6] M. Marchesoni, M. Mazzucchelli and S. Tenconi, "A Non Conventional Power Converter for Plasma Stabilization," IEEE Transactions on Power Electronics, Vol. 5, No. 2, April 1990.

[7] F. Z. Peng, J. S. Lai, J. McKeever and J. Van Coevering, "A Multilevel Voltage-Source Inverter with Separate DC Sources for Static Var Generation," IEEE-IAS Conference Record, pp. 2541-2548, 1995.

[8] J. Rodriguez, L. Moran, A, Gonzales and C.Silva, "High Voltage Converter with Regeneration Capability," IEEE-PESC Conference Proceedings, Vol. 2, pp. 1077-1082, 1999.

[9] Geoffrey R. Walker B.E. (Hons), "Modulation of Multilevel Converters," Ph.D. Thesis, Queensland, Nov. 1999.

[10] Ebrahim Babaei, Seyed Hossein Hosseini and Mahrdad Tarafdar Haque, "A Novel Approach to Multilevel Current Source Converters," Proceedings of the 2005 Electrical Engineering/Electronics, Telecommunications and Information Technology International Conference (ECTI-CON2005), Vol.I. 177-180, 12-13 May 2005, pattava, choburi, Thailand.

[11] Fang Zheng Peng, John W. McKeever, and Donald J. Adams, "A Power Line Conditioner Using Cascade Multilevel Converters for Distribution Systems," IEEE Transaction on Industry Applications, Vol. 34, No. 6, pp. 1293-1298, Nov./Dec.1998.

◈◈◈

Interleaved Phase Sequence matrices for PAPR reduction in OFDM systems

S.Bhoopalan[1], M. Palanivelan[2], Dr. Sheila Anand[3],

[1]Student, Dept. of ECE, [2] Associate Professor, [3] Dean (Research) Computer Studies,
Rajalakshmi Engineering College, Chennai, India.
E-mail : [1]bhoopalanjp@gmail.com [2]velan.research@gmail.com [3]sheila.anand@gmail.com

Abstract - The Orthogonal Frequency Division Multiplexing (OFDM) uses as a high-speed information transmission technology, with excellent anti-interference capabilities. A major drawback of OFDM signals is their high Peak-to-Average Power Ratio (PAPR), which causes serious degradation in system performance when a linear Power Amplifier (PA) is driven into nonlinear region. In this paper, we propose an Interleaved Phase Sequence (IPS) method using special matrices such as Riemann and Hilbert for reduction of PAPR in OFDM system. OFDM system is implemented in several Broadband Communication Systems like Wireless Local Area Network (WLAN), Worldwide interoperability for Microwave Access (WiMax), Digital Video Broadcasting (DVB) and Digital Audio Broadcasting (DAB). Our evolutionary computation technique can be used to reduce PAPR with commonly used PTS and SLM conventional schemes. Also proposed technique offers low computational and phase search complexity. Side Informations (SI) need not be transmitted to receiver for original data recovery.

Keywords: -*OFDM, PAPR, PTS, SLM, IPS, CCDF, OBO*

I. INTRODUCTION

A. OFDM System

Orthogonal Frequency Division Multiplexing (OFDM) [1] is considered as a root technology for broadband mobile communications [2] due to its high spectral efficiency. OFDM can be easily implemented by the Inverse Fast Fourier Transform (IFFT) and Fast Fourier Transform (FFT) process in digital domain and has the property of high-speed Broadband transmission [3]. The channel bandwidth is divided into many channels so that in a multi-user environment each channel is allocated to a single user. However the difference lies in the fact that the carriers chosen in OFDM are much more closely spaced than in FDMA. The orthogonality principle [5] [10] [20] essentially implies that each carrier has a null at the centre frequency of each of the other carriers in the system while also maintaining an integer number of cycles over a symbol period. OFDM signal has high Peak to Average Power Ratio(PAPR) [5] because of the overlapping of multi-carrier signals with large number of sub- carriers.PAPR reduction is one of the most important research interests for the OFDM systems.

To transmit high PAPR signal without distortion requires more expensive power amplifier with high linear characteristics. OFDM has been adopted for various wireless communication systems [6][7][12] such as wireless local area networks (WLANs) wireless metropolitan area networks (WMANs) digital audio broadcasting (DAB) and digital video broadcasting (DVB).OFDM is an very attractive technique for achieving high data rate in the wireless communication systems.

B. PAPR FORMULATION IN OFDM SYSTEMS

In the discrete time domain, an OFDM signal x_n of *N*subcarriers can be expressed in (1)

$$x_n = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X_k e^{j2\pi kn/N}, 0 \le n \le N-1 \qquad (1)$$

where x_k, $k = 0, 1,\ldots, N\text{-}1$, are input symbols modulated by QAM and is the discrete time index [13]. The PAPR of an OFDM signal is defined as the ratio of the maximum to the average power of the signal, as given in (2)

$$\text{PAPR}(x) = 10\log_{10} \frac{\max_{0 \le n \le N-1} \{|x_n|^2\}}{E\{|x|^2\}} (\text{dB}) \qquad (2)$$

where E {.} denotes the expected value operation and $X = [X_0, X_1,..., X_{N-1}]^T$.

It should be noted that the PAPR of a continuous-time OFDM signal cannot be precisely described by the use of *N* samples per signal period. Therefore, some of signal peaks may be missed and PAPR reduction performance estimates are unduly optimistic. To avoid this problem, the oversampling is usually employed, which can be obtained by *LN* – point IFFT of data sequence with (*L - 1*) *N* zero-padding. It is shown in

[10] that an oversampling factor $L = 4$ is sufficient to approximate the real PAPR results.

Moreover, the distribution of PAPR bears stochastic characteristics in a practical OFDM system, usually being expressed in terms of Complementary Cumulative Distribution Function (CCDF) [3]. The CCDF can be also used to evaluate and compare the performance of any PAPR reduction schemes, and the CCDF of discrete-time PAPR is given in (3)

$$\text{CCDF}(N, \text{PAPR}_0) = \Pr\{\text{PAPR} > \text{PAPR}_0\}$$
$$= 1 - (1 - e^{-\text{PAPR}0})^N \qquad (3)$$

where N is the number of subcarriers in an OFDM system and PAPR_0 is a certain value of PAPR.

The rest of the paper is organized as follows. In section 2, we discuss the existing techniques for PAPR reduction. In section 3, the proposed system is discussed and section 4 shows the simulation results supporting the ideas presented. Finally, the results are summarized in section 5.

II. RELATED WORK

Partial Transmits Sequence (PTS)

In PTS, shown in Fig. 1, the input data sequence of an OFDM system with N subcarriers is firstly partitioned into V disjoint sub blocks X_i, i = 1,2, . . . , V, where all the subcarriers which are occupied by the other subblocks are set to zero[9],[11]. The frequency domain input sequence is given as

By applying a phase weighting factor $b_i = \exp(j\varphi_i)$, $\varphi_i \in [0,2\pi]$ to the subblocks $X_i = [X_{i,1}, X_{i,2}, \ldots, X_{i,N}]^T$, i = 1, 2, . . . , V, alternative frequency signal sequence is given as

After being transformed to time domain by IFFT, the time domain signal sequence becomes

wherex' denotes the candidate sequence.

In the practical application of PTS, a set of phase weighting factors is usually selected for generating phase weighting sequences. Assume that there are W allowed phase weighting factors in this set. Without any loss of performance, we can set phase weighting factor for the first subblock to one and observe that there are $(V-1)$ subblocks to be optimized. To match the optimal phase weighting sequence for each input data sequence W^{V-1} subblocks to be optimized. To match the optimal phase weighting sequence for each input data sequence for transmitting.

For optimal PTS (O-PTS), the optimum PAPR performance can be found after searching W^{V-1} alternative combinations if the number of subblocks is V and the number of allowed phase weighting factors is W. In the process of phase weighting combination, large numbers of complex multiplications are needed, and the computational complexity is very large.

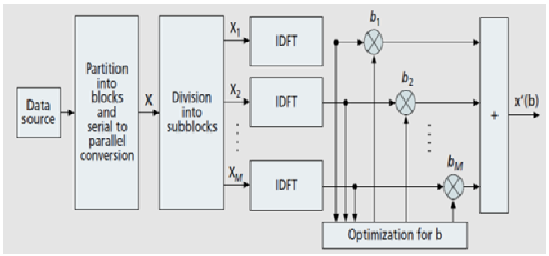


Fig 1: Block diagram of PTS scheme.

At the side of receiver, in order to recover the received signals successfully, the side information about phase weighting factors for PTS scheme is required. This information must be transmitted accompanying with the transmitted signal, and $[log_2 W^{V-1}]$ bits are required to represent this side information, where [.] rounds the elements to the nearest integers toward zero.

Selected Mapping (SLM)

In a Selected mapping (SLM) shown in Fig.2 [14] is a specific scheme for PAPR reduction. The SLM takes advantage of the fact that the PAPR of an OFDM signal is very sensitive to phase shifts in the frequency-domain data. PAPR reduction is achieved by multiplying independent phase sequences to the original data sequence and determining the PAPR of each phase sequence combination. The combination with the lowest PAPR is transmitted. In other words, the data sequence X is element-wise phased by D N-length phase sequences [15] [16].

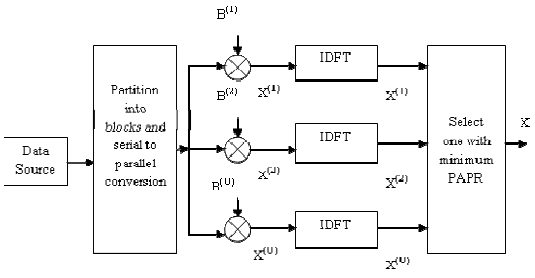


Fig 2: Block diagram of SLM Scheme

PAPR reduction is achieved by multiplying independent phase sequences to the original data and determining the PAPR of each phase sequence combination. The combination with the lowest PAPR is transmitted.

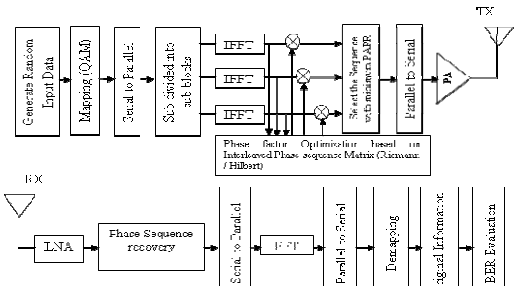


Fig 3: Proposed OFDM system with IPS

The CCDF of the PAPR in SLM OFDM symbol is

$$\Pr[\mathrm{PAR}\{x[n]\}>\gamma] = 1-(1-e^{-\gamma})^{N} \quad (7)$$

III. PROPOSED SYSTEM

The block diagram of the proposed OFDM system with Interleaved Phase Sequences (IPS) technique for PAPR reduction is shown in figure 3. IPS is formed based on Riemann and Hilbert matrix [3], [5]. The input data stream is generated from random source. Data is mapped on to Quadrature Amplitude Modulation (QAM) [3]-[8] constellation. These serially mapped symbols are converted into parallel, and then symbols are partitioned into sub blocks based on PTS [14]. Each sub-block is multiplied with phase sequence generated using IPS technique.

After phase sequence multiplication the sequence which yields minimum PAPR will be selected and transmitted. With the help of special structure of the phase sequence matrix, the receiver can able to recover the data without any Side Information. In this model, Channel noise is assumed to be Gaussian. At the Receiver phase sequence recovery is done and signals are passed through FFT. Demapping of symbols is carried out to recover the original data. It is seen that this method greatly reduces the phase search complexity with comparable computational complexity [4].

The proposed IPS technique to reduce PAPR described as a flow diagram given in figure 4.

Matrix Based Interleaved Phase Sequence (IPS) Method

I. CONVENTIONAL PTS (C-PTS) IN MATRIX FORM

Both b_i and x_i from equation (5) can be shown in matrix form as follows:

It should be noted that all the elements of each row of matrix *b* are of the same values and this is in accordance with the C-PTS method. It should be noted that in order to have exact PAPR calculation, at least 4 times oversampling is necessary [10] [18]. As the oversampling of x, add zeros to the vector, hence the number of phase sequence to multiply to matrix x will remain the same. Now, the process is performed by choosing theoptimization parameter with the following condition:

After finding the optimum then the optimum signal is transmitted to the next block. For finding the optimum , we should perform exhaustive search for (*V-1*) phase factors since one phase factor can remain fixed, b_1=1. Hence to find the optimum phase factor, W^{V-1} iteration should be performed, where *W* is the number of allowed phase factors.

Flow chart

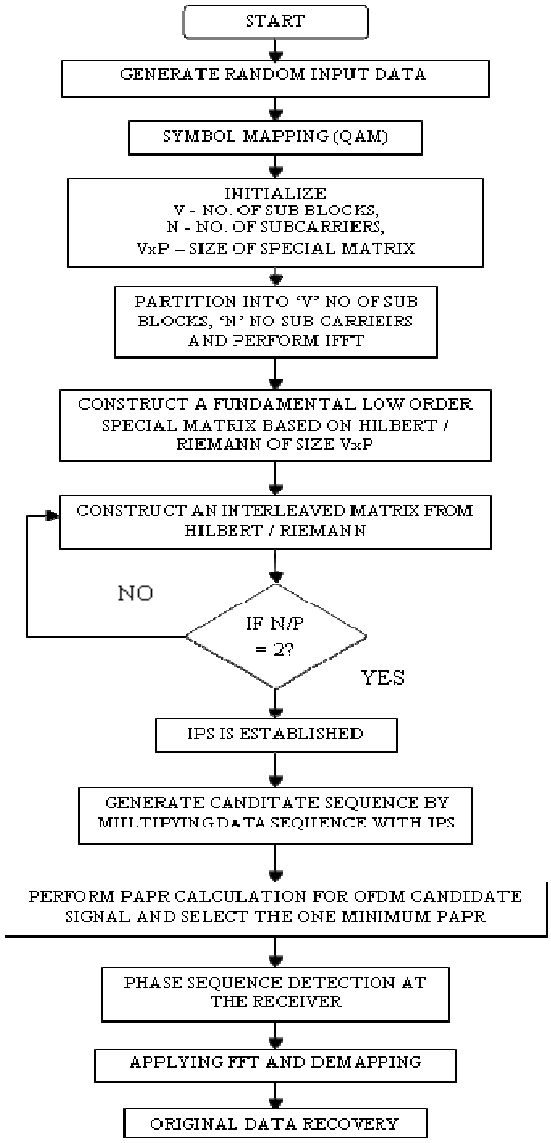


Fig 4: Flow diagram of IPS based OFDM System

II. CONSTRUCTION OF CONTINUOUS PERIODIC ORTHOGONAL SEQUENCES

Based on the DFT matrix and the interleaving technique [6], we can obtain a sequence set$b = \{b_i,\ 0 \le i \le N/2 - 1\}$, where $b_i$can be constructed as

$$b_i = \begin{bmatrix} j_i \lozenge j_i \\ j_i \lozenge j_{i+1} \\ j_i \lozenge j_{i+2} \\ j_i \lozenge j_{i+3} \end{bmatrix} \quad (11)$$

The generated sequence set $\boldsymbol{b}$ comprises N / 2 groups and each group contains 4 sequences with length 2N. Hence, there are 2N sequences in total.

If the number of allowed phase factor is constant then the value of b [5] depends on the number of subblock V .If the system requires more number of subcarriers essentially we need to go for more number of subblocks for better PAPR reduction. Increase in subblocks simultaneously increases the number of phase sequences to optimise. This will lead to greater phase search [15] complexity. In our proposed IPS method, there are N numbers of phase factors which are formed from Riemann and Hilbert matrix. In this system, even if the number of subblocks increases the phase search complexity is reduced with good PAPR reduction performance. Most of the PAPR reduction techniques use random phase sequences for optimization. This forces transmitter to send side information to the receiver. The optimum phase sequence requires N^8 iterations if the N different random phase factors are used, which is not practical. In IPS, we apply the same iteration as applied in C-PTS. If the number of sub-carriers is high for example IEEE 802.16, LTE the number of used sub carries are large, therefore the phase search complexity for such systems are very high.IPS matrix based on Riemann and Hilbert are discussed below.

IPS BASED ON RIEMANN MATRIX

Riemann matrix is formed from the equation,

$$A(n,m) = \begin{cases} \text{n} - 1 \text{ if n divides m} \\ -1 \text{ otherwise} \end{cases} \quad (12)$$

Riemann matrix of order 8x8 is generated from equation (12) as given in (13)

$$R = \begin{bmatrix} 1 & -1 & -1 & -1 \\ -1 & 2 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 4 \end{bmatrix} \quad (13)$$

Riemann matrix in IPS form for b_0 and b_2

$$b_1 = \begin{bmatrix} 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 2 & -1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 & -1 & 3 & -1 & -1 \\ 1 & -1 & -1 & -1 & -1 & -1 & -1 & 4 \end{bmatrix} \quad (14)$$

IPS BASED ON HILBERT MATRIX

Hilbert matrix is special kind of Cauchy matrix [8] .Cauchy matrix of dimension m x n is of the form as in

$$a_{ij} = \frac{1}{x_i - y_j}; \quad 1 \le i \le m, \quad 1 \le j \le n; \quad (15)$$

wherex_i and y_i are elements of field F and x_i - $y_i \neq 0$. Hilbert matrix is calculated from the Cauchy matrix using x_i - y_i = i+ j -1 Hilbert matrix (HL) with dimensions m=n=8 is given asHL matrix with order (8 x 8) is generated using equation (15) as given in equation (16)

$$HL = \begin{bmatrix} 1 & 1/2 & 1/3 & 1/4 \\ 1/2 & 1/3 & 1/4 & 1/5 \\ 1/3 & 1/4 & 1/5 & 1/6 \\ 1/4 & 1/5 & 1/6 & 1/7 \end{bmatrix} \quad (16)$$

Hilbert matrix in IPS form for b_0 and b_2

$$b_1 = \begin{bmatrix} 1 & 1 & 1/2 & 1/2 & 1/3 & 1/3 & 1/4 & 1/4 \\ 1 & 1/2 & 1/2 & 1/3 & 1/3 & 1/4 & 1/4 & 1/5 \\ 1 & 1/3 & 1/2 & 1/4 & 1/3 & 1/5 & 1/4 & 1/6 \\ 1 & 1/4 & 1/2 & 1/5 & 1/3 & 1/6 & 1/4 & 1/7 \end{bmatrix} \quad (17)$$

IV. SIMULATION RESULTS

To compare and evaluate the PAPR reduction performance, extensive simulations have been performed based on Interleaved Phase Sequences (IPS) technique using MATLAB. In simulations, an OFDM system has been considered with N=8 and 16, oversampling factor L=4 and Quadrature Amplitude Modulation (QAM) is implemented. The simulation parameters used are given in Table 1. It is seen that, highest PAPR results when all the bits in a data block are same and orthogonal [19]. It can be seen from the results that OFDM with IPS works effectively gives reduced PAPR for all types of data blocks.

In the simulation we used 16-QAM baseband modulation scheme. Each modulated symbol is transmitted through N=16 sub carriers by 64-point IFFT and L=4 oversampling is employed to estimate PAPR precisely. To analyze PAPR reduction and power amplifier efficiency, we consider class A power amplifier which is the mostlinear with power efficiency.

Table 1: Simulation Parameters

Parameter	**Specifications**
Modulation	QAM
Number of data subcarriers M	8, 16
Number of FFT/IFFT points(N)	64
Number of data symbols	16
Over sampling factor	L=4
Bandwidth, BW	1 MHz
Sampling Frequency, (BW x L)	4 MHz
Number of Guard Interval Samples	32
Channel Model	Gaussian

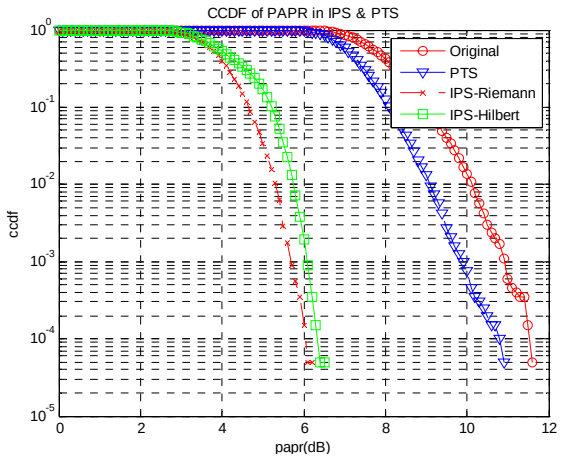


Fig 5:CCDF of PAPR in IPS& PTS

Figure 5 shows the CCDF of PAPR in the IPS technique with QAM Modulation scheme. It is easy to observe that at 0.0001% of CCDF the PAPR value for the PTS offers 1 dB reduction when compared with actual value of original data blocks. Our proposed technique based on IPS with Special Matrices is also shown in this figure 5. From this we found that Riemann Matrix performedbetter in reduction compared to other schemes also it offers 6 dB reduction from the original values.

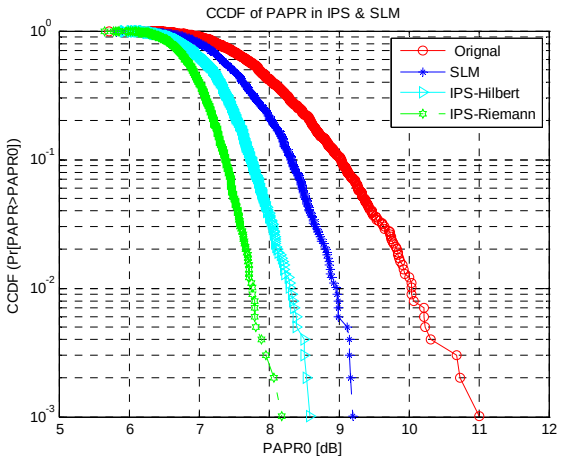


Fig 6:CCDF of PAPR in IPS & SLM

Figure 6 shows the CCDF of PAPR in the IPS technique with QAM Modulation scheme. It is easy to observe that at 0.001% of CCDF the PAPR value for the SLM offers 1 dB reduction when compared with actual value of original data blocks. Our proposed technique based on IPS with Special Matrices is also shown in this figure 6. From this we found that Riemann Matrix performedbetter in reduction compared to other schemes also it offers 2.3 dB reduction from the original values.

CONCLUSION

In this paper, the performance of OFDM transmission systems in relation to Peak to average power ratio is evaluated using Interleaved Phase Sequences technique using Special matrices. Here we constructed, Interleaved PhaseSequences using Riemann and Hilbert matrix. From the simulation results, it is seen that the proposed technique offers better PAPR reduction using Riemann matrix with reduced phase search complexity when compared with conventional schemes such as SLM and PTS. Also the proposed technique does not require the side information to be sent to the receiver for recovery. Further the effect of Out of Band (OBO) distortion can be obtained for the proposed scheme to ensure its required BER performance.

REFERENCES

[1] Seung Hee Han, Jae Hong Lee, "An Overview Of Peak-To-Average Power Ratio Reduction Techniques For Multicarrier Transmission" IEEE Wireless Communications April 2005.

[2] L.J.Cimini, Jr, and N.R.Sollenberger, "Peakto-Average power ratio reduction of an OFDM signal using PTS," IEEE Communication Letters... vol. 4, no. 3, pp.86-88.Mar. 2000.

[3] Pooria Varahram, Borhanuddin Mohd Ali, "Partial Transmit Sequence Scheme with New Phase Sequence for PAPR Reduction in OFDM Systems" IEEE Transactions on Consumer Electronics, Vol 57, No 2, May2011.

[4] Zhenyu Zhang, Fanxin Zeng, Guixin Xuan, Wei He, Jun Gu, "Periodic Odd-Shift Orthogonal Sequences Based on Interleaved DFT Matrix" IEEE September 2009

[5] M.Palanivelan, Dr.Sheela Anand, M.Gunasekaran, "Matrix Based Low Complexity PAPR Reduction In OFDM Systems" IJECT Volume 2, Issue 2, Pg no 158-162 June 2011.

[6] C.G. Han, T. Hashimoto, and N. Suehiro, "Poly phase zero-correlation zone sequences based on complete complementary codes and DFT matrix," in Proc. IWSDA07, Chendu, China, pp. 172-175, Sept. 2007.

[7] N.T Hieu., S.W. Kim and H.G Ryu, "PAPR Reduction of the Low Complexity Phase Weighting Method in OFDM Communication System", *IEEE Trans. on Consumer Electronics*, vol. 51, no. 3, pp. 776-782, 2005.

[8] Cauchy matrix, Cauchy determinants, Generalization. Available: http://en.wikipedia.org/wiki/Cauchy_matrix.

[9] H. Park and H. K. Song, "A new PAPR reduction technique of OFDM system with nonlinear high power amplifier," *IEEE Trans. Consumer Electr.,* vol. 53, no. 2, pp. 327-332, May 2007.

[10] Lingyin Wang and Ju Liu, "PAPR Reduction of OFDM Signals by PTS With Grouping and Recursive Phase Weighting Methods" IEEE TRANSACTIONS ON BROADCASTING, VOL. 57, NO. 2, JUNE 2011

[11] J. Hou, J. Ge, D. Zhai, and J. Li, "Peak-to-average power ratio reduction of OFDM signals with nonlinear companding scheme," *IEEE trans. Broadcast.*, vol. 56, no. 2, pp. 258–262, Jun. 2010.

[12] J. Hou, J. Ge, and J. Li, "Peak-to-Average Power Ratio Reduction of OFDM Signals Using PTS Scheme With Low Computational Complexity," IEEE TRANSACTIONS ON BROADCASTING, VOL. 57, NO. 1, MARCH 2011.

[13] V. Kumbasar, Oğuz Kucur, "Better wavelet packet tree based OFDM for multipath powerline channel," Gebze Institute of Technology, Electronics Engineering Department, April 2009.

[14] R. J. Baxley and T Zhou., "Comparing selected mapping and partial transmit sequence for PAPR reduction", *IEEE Trans. Broadcast*., vol. 53, no. 4, pp.797-803, December 2007.

[15] H. G Ryu., and K. J Youn, "A new PAPR reduction scheme: SPW (subblock phase weighting)," *IEEE Trans. on Consumer Electronics*, vol. 48, no. 1, pp. 81 – 89, 2002.

[16] B. S. Krongold and D. L. Jones, "PAR reduction in OFDM via activeconstellation extension," *IEEE Trans. Broadcast.*, vol. 49, no. 3, pp.258–268, Sep. 2002.

[17] R. W. Bäuml, R. F. H. Fischer, and J. B. Huber, "Reducing the peak-toaveragepower ratio of multicarrier modulation by selected mapping,"*Electron. Lett.*, vol. 32, no. 22, pp. 2056–2057, Oct. 1996.

[18] A. D. S. Jayalath and C. Tellambura, "SLM and PTS peak-power reductionof OFDM signals without side information," *IEEE Trans. WirelessCommun.*, vol. 4, no. 5, pp. 2006–2013, Sep. 2005.

[19] E. Alsusa and L. Yang, "Redundancy-free and BER-maintained selectivemapping with partial phase-randomising sequences for peak-to-averagepower ratio reduction in OFDM systems," *IET Commun.*, vol. 2,no. 1, pp. 66–74, Jan. 2008.

[20] C. Zhao, R. Baxley, and G. Zhou, "Peak-to-average power ratio andpower efficiency considerations in MIMO-OFDM systems," *IEEE CommunicationsLetters*, vol. 12, no. 4, pp. 268 –270, Apr. 2008.

◈◈◈

Advanced Cryptographic Steganography Using Multimedia Files

Shery Elizabeth Thomas, Sumod Tom Philip, Sumaya Nazar, Ashams Mathew & Niya Joseph

Department of Computer Science & Engineering, Amal Jyothi College of Engineering, Kottayam, India
E-mail : Sumodtomphilip@gmail.com,sherythomas91@gmail.com,ashamsmat@gmail.com,sumize246@gmail.com

Abstract **-** Nowadays, the extensive usage of internet for communication has increased the threat level to the users. Various kinds of threats such as malware, malicious intruders.etc can tap the secret information and use it for illegal purpose. A combination of cryptography using AES and steganography is used here for security. To increase the security level , the key is hashed using SHA-1. This method proposes an advanced concept of combining cryptography and steganography in image, audio and video. Here the data encrypted using advanced encryption algorithm will be hidden into a multimedia image, audio and video file according to the user's choice. Steganography is implemented by means of Least Significant Bit insertion technique.

Keywords **-** *SHA-1;AES;Cryptography,Steganography*

I. INTRODUCTION

The widespread use of internet for communication has increased the attacks to users. The security of information is an important issue related to privacy and safety during storage and communication. Cryptography and Steganography are two popular ways of sending vital information in a secret way . Cryptography is the method of converting plaintext into cipher text. The messages are converted into an encrypted format using a key and then this cipher text is hidden into an image, audio or video file according to the user's choice. The encryption is done using Advanced Encryption Algorithm and the key is hashed using Secure Hash Algorithm-1.

Cryptography protects information by transforming it into unreadable format . Only those who possess a secret key can decipher the cipher text into plain text. There are many cryptographic techniques and it is mainly classified into symmetric and asymmetric cryptography. Among them, AES is one of the most powerful techniques and it uses symmetric key cryptography. To enhance the security level, the secret key is hashed and this hashed key is used for encryption and decryption. The key is hashed using SHA-1. SHA stands for "Secure Hash Algorithm". SHA-1 is a common hash function which is as fast as MD5 and generates a 160-bit hash. Steganography is most often associated with embedding data in some form of electronic media. Data from a source file is hidden by altering insignificant bits of information in a host file. The information hiding process in a steganographic system starts by identifying a cover medium's redundant bits. The embedding process creates a stego medium by replacing these redundant bits with data from the hidden message. Steganography can be classified into image, audio and video steganography depending on the cover medium.

II. ADVANCED ENCRYPTION ALGORITHM

AES is a specification for the encryption of electronic data. The algorithm used by AES is a symmetric algorithm, meaning the same key is used for both encryption and decryption. AES is based on a principle known as substitution - permutation network. It is fast in both software and hardware. It does not use a fiestel network. AES has a fixed block size of 128 bits and a key size of 128,192 or 256 bits. The block size has a maximum of 256 bits and key size has no theoretical maximum.

III. SECURE HASH ALGORITHM-1

In cryptography, SHA-1 is a cryptographic hash function.SHA stands for "secure hash algorithm". It is an iterative, one-way hash function that can process a message to produce a condensed representation called a message digest. This algorithm enable the determination of a message's integrity: any change to the message wills, with a very high probability, result in a different message digests. This property is useful in the generation and verification of digital signatures and

message authentication codes, and in the generation of random numbers (bits).The algorithm can be described in two stages: preprocessing and hash computation. Preprocessing involves padding a message, parsing the padded message into m-bit blocks, and setting initialization values to be used in the hash computation. The hash computation generates a message schedule from the padded message and uses that schedule, along with functions, constants, and word operations to iteratively generate a series of hash values. The final hash value generated by the hash computation is used to determine the message digest. SHA-1 is the most widely used of the existing SHA hash functions, and is employed in several widely used applications and protocols. SHA-1 appears to provide greater resistance to attacks. SHA-1 is another common hash function. It is as fast as MD5 and generates a 160-bit hash.

IV. PROPOSED SYSTEM

Fig 1 shows the proposed architecture. The sender provides the plaintext to the encryption module and the key provided by the user is first hashed using SHA-1 algorithm to prevent attacks and this hashed key is given to the encryption module. The encryption is done using AES algorithm. The cipher text thus obtained is given to the steganography module. The users can hide the cipher text in an image, audio or video file. The cipher text is hidden to the appropriate cover medium in the Least Significant Bits as it will not cause much variation in the original file. Here, LSB Steganography is used. Then this hidden file is obtained at the receiver and is given to an extractor. As a result, hidden information is extracted from the medium. The receiver should provide the same key after hashing it with SHA-1 for decryption. For decryption, AES algorithm is used. After decryption, the plain text is retrieved.

The proposed architecture is used to enhance the security of data communications. It mainly consists of three modules:

- Encryption Module
- Steganography Module
- Decryption Module

The Encryption and Decryption Modules make use of Advanced Encryption Standard. The symmetric key used is hashed using SHA-1 algorithm. The Steganography Module is done using Least Significant Bit Insertion Technique. Users can select the cover media according to their choice. The cover media can be an image, an audio file or a video file.

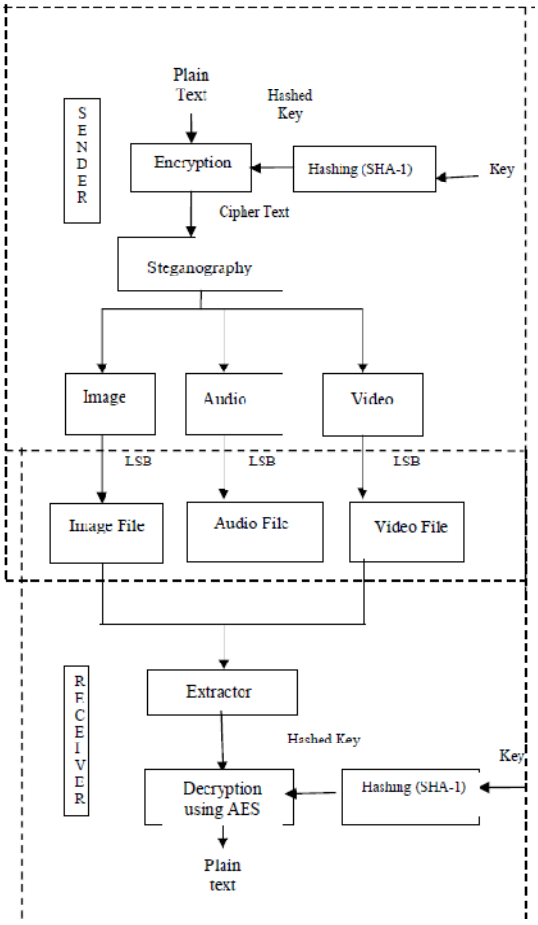


Fig. 1 : Advanced Cryptographic Steganography Architecture

A. *Encryption Module*

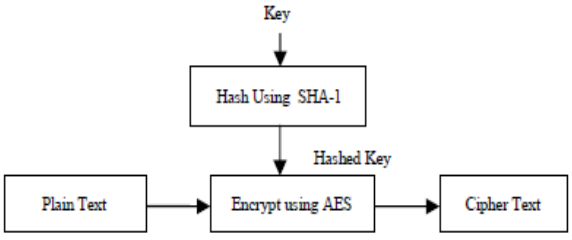


Fig. 2: Encryption Module

The encryption is done using Advanced Encryption Algorithm. First the key is converted to byte code using UTF-8 format and it is hashed using SHA-1 algorithm. This byte array can be converted into appropriate key by using AES algorithm. Similarly the plaintext is converted into byte format and it is encrypted using AES algorithm.

B. Steganography Module

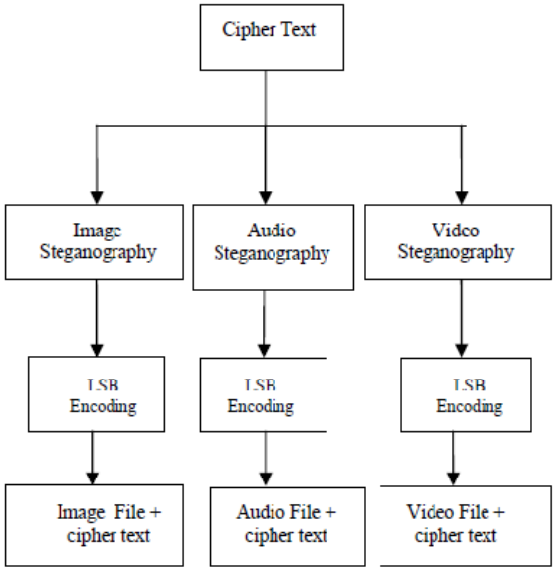


Fig. 3 : Steganography Module

Steganography can be performed using image, audio or video as cover signal. Based on cover media,there are three types of steganography:-

- Image Steganography
- Audio Steganography
- Video Steganography

1) Image Steganography: Image Steganography[1] is implemented by means of Least Significant Bit insertion.LSB is always the last bit on the right-hand side of any binary number. Changing this bit causes the least possible effect to the original value.In a 24-bit image, there are 3 bytes of data to represent RGB values for every pixel in that image. This implies that we can store/hide 3 bits in every pixel. For example, if the image has the following bits:

10010101	00001101	11001001
10010110	00001111	11001010
10011111	00010000	11001011

To store **101101101**. we replace with the original LSBs like this:

1001010**1**	0000110**0**	1100100**1**
1001011**1**	0000111**0**	1100101**1**
1001111**1**	0001000**0**	1100101**1**

To reveal the stored message, the LSBs are extracted alone from the Stego Medium and combined together.

2) Audio Steganography: In Time Domain, a message can be stored in the LSBs, something similar to images. To maintain CD quality, it is important to encode at 16 bits per sample at a rate of 44.1kHz.It is possible to record at 8 bits per sample using the high significant bits and save the other 4 LSBs to hide our message without making any perceptible change to the audio quality. Time domain technique is used here.

Sampling technique followed by Quantization converts analog audio signal to digital binary sequence.In this technique LSB of binary sequence of each sample of digitized audio file is replaced with binary equivalent of secret message. For example if we want to hide the letter 'A' (binary equivalent **01100101**) to an digitized audio file where each sample is represented with 16 bits, then LSB of 8 consecutive samples (each of 16 bit size) is replaced with each bit of binary equivalent of the letter 'A'.

TABLE I : AUDIO STEGANOGRAPHY USING LSB

Sampled Audio Stream (16 bit)	**'A' in binary**	**Audio stream with encoded message**
1001 1000 0011 1100	**0**	1001 1000 0011 110**0**
1101 1011 0011 1000	**1**	1101 1011 0011 100**1**
1011 1100 0011 1101	**1**	1011 1100 0011 110**1**
1011 1111 0011 1100	**0**	1011 1111 0011 110**0**
1011 1010 0111 1111	**0**	1011 1010 0111 111**0**
1111 1000 0011 1100	**1**	1111 1000 0011 110**1**
1101 1100 0111 1000	**0**	1101 1100 0111 100**0**
1000 1000 0001 1111	**1**	1000 1000 0001 111**1**

3) Video Steganography: To conceal the data in the video file[3], LSB modification algorithm is used. The main advantage is a very high watermark channel bit rate and a low computational complexity. In this method modification is done to the least significant bits of the carrier file's individual pixels, thereby encoding hidden data. Each pixel stores 3 bits of secret information, one in each RGB values. As a simple example of LSB substitution, imagine "hiding" the character 'A' across the following eight bytes of a carrier file:

(00100111 11101001 11001000)

(00100111 11001000 11101001)

(11001000 00100111 11101001)

Letter 'A' is represented in ASCII format as the binary string 10000011.

These eight bits can be "written" to the LSB of each of the eight carrier bytes as follows (the LSBs are italicized and bolded):

(00100111 11101000 11001000)

(00100110 11001000 11101000)

(11001001 00100111 11101001).

With such a small variation in the colors of the video image it would be very difficult for the human eye to discern the difference thus providing high robustness to the system .

C. Decryption Module

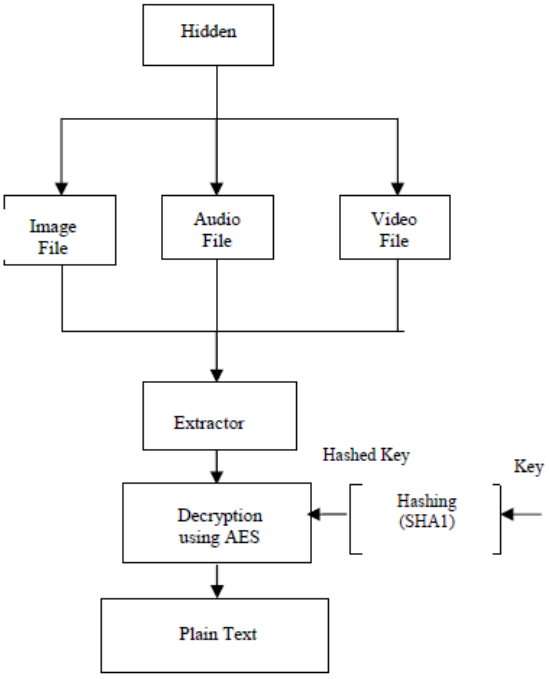


Fig. 4 : Decryption Module

The hidden cipher text is extracted from the cover media before decryption. The decryption is done using Advanced Encryption Algorithm. First the key is converted to byte code using UTF-8 format and it is hashed using SHA-1 algorithm. This byte array can be converted into appropriate key by using AES algorithm. The cipher text is then decrypted using AES algorithm.

V. CONCLUSION

In this method, AES algorithm is used for encryption and decryption as it is more secure and is significantly faster than the Triple DES algorithm and it can be implemented in hardware. AES requires only one pass to encrypt data while DES requires two passes to do the same. In addition to security, AES provides superior performance and better use of resources.

Here, the key is hashed using SHA-1 algorithm which produces a condensed form of the message .It provides additional protection. Steganography combined with cryptography allows data transfer without an eavesdropper knowing about it. The data can be hidden in either image, audio or video file. As a result, more security is achieved.

REFERENCES

[1] Piyush Marwaha, Paresh Marwaha, " Visual Cryptographic Steganography In Images", 2010 Second International conference on Computing, Communication and Networking Technologies.

[2] Soumyendu Das, Subhendu Das, Bijoy Bandyopadhyay and Sugata Sanyal **,"**Steganography and Steganalysis: Different Approaches".

[3] Mritha Ramalingam, "Stego Machine – Video Steganography using Modified LSB Algorithm", World Academy of Science, Engineering and Technology 74 2011.

[4] K. Geetha and P.Vanitha Muthu ,"Implementation of ETAS (Embedding Text in Audio Signal) Model to Ensure Secrecy", (IJCSE) International Journal on Computer Science and EngineeringVol. 02, No. 04, 2010, 1308-1313

[5] Nedeljko Cvejic and Tapio SeppÄanen, "Increasing Robustness of LSB Audio Steganography by Reduced Distortion LSB Coding"

[6] Kaliappan Gopalan,"Audio Steganography Using Bit Modification".

[7] Domenico Bloisi and Luca Iocchi,"Image Based Steganography And Cryptography".

[8] Niels Provos And Peter Honeyman, University Of Michigan ,"Hide and Seek: An Introduction to Steganography", Published By The IEEE Computer Society,1540-7993/03/$17.00 © 2003 IEEE, IEEE Security & Privacy.

◈◈◈

Authentication of Digital Documents Using Secret Key Biometric Watermarking

V.Anitha & Dr. R.Leela Velusamy

Department of Computer Science and engineering, National Institute of Technology, Tiruchirappalli, India
E-mail : anitha.v2000@gmail.com

Abstract - Digital documents play a major role in modern era. They are easy to generate, modify and manage. The easy modifiable property of digital document makes it more vulnerable to forgery. It can be easily tampered or forged. So the challenge is to produce digital documents that are highly resistant to forgery and reliably confirms the real owner of the document. This can be resolved by biometric watermarking which make a direct relation between the document and its owner. A new biometric watermarking technique with secret key is proposed to digitize the authoritative documents issued by government / other organizations as a part of UID / Aadhar card project of India using biometric watermarking. Biometric code is generated from the biometric data collected from the owner of the document. The biometric code is watermarked in the document with a secret key to generate a biometric watermarked document that authenticates the real owner. De-watermarking the document with the same key yields the biometric code that can be used for authentication of the document. If the document is tampered in any way it will be indicated in the extracted watermark. Experimental results show that 100% accuracy is obtained in authenticating the genuine documents.

Keywords- *secret key watermarking, biometric watermarking, tamper detection, digital document authentication, ownership verification, biometric authentication.*

I. INTRODUCTION

In the modern era, digital document concept is becoming increasingly important as more of the world's information is stored as readily transferable bits.

Digital documents have so much of advantages over print documents. Digital documents are less expensive and easy to store, transport and search compared to traditional print documents. But it has its own limitations too. A simple image editor can be used to modify and make a forged document. Digital documents can be tampered easily. In order to utilize the whole benefits of digital document, these limitations have to be overcomed.

Digital watermarking technique has been used in order to overcome these limitations by embedding some text/logo/pseudorandom sequence that identifies the owner of the document. Some applications of watermark are:

- Establishing ownership by embedding identifying data
- Tracking the movement of authorized copies by embedding a unique serial number in each copy
- Attaching meta-data that pertains to the image such as a time, date, and location stamp

These digital watermarks suffer from certain limitations. Watermarks are least correlative to the owner of the digital document. Watermark doesn't directly link the digital document to the owner of it. And the related characteristics of the watermark with the digital document may change over time. So the traditional watermarking method does not convincingly validate the claimed identification of the person. Using such kind of watermark may lead to easy forgery or tampering.

Recently biometrics is merged into watermarking technology to enhance the credibility of the conventional watermarking technique. Biometric watermarking is a special case of digital watermarking in which the watermark content is biometric data. Access control or authenticity verification has been addressed by both digital watermarking as well as by biometric authentication. By embedding biometrics in the host, it formulates a reliable individual identification as biometrics possesses exclusive characteristics that can be hardly counterfeited. Hence, the conflicts related to the intellectual property rights protection can be

potentially discouraged [1]. Consequently, it has been decided by governmental institutions in Europe and the U.S. to include digital biometric data in future ID documents. In India, biometric based UID scheme, AADHAR is started with the goal of issuing a unique identification number to all the Indian citizens.

Adhaar card project of India is a major motivation beyond this project. It involves the collection of biometric data from all Indian citizens. Aadhaar is a 12-digit unique identity number which is to be issued by Unique Identification Authority of India (UIDAI) to all residents in India. The number will be stored in a centralized database and linked to the basic demographics and biometric information – photograph, ten fingerprints and iris – of each individual. UID is easily verifiable in an online, cost-effective way. And also, it is unique and robust enough to eliminate the large number of duplicate and fake identities in government and private databases. The random number generated will be devoid of any classification based on caste, creed, religion and geography [2].

The goal of this project is to propose a technique for the creation of robust, forgery resistant digital documents which replace/used along with the paper documents possessed by an individual. It makes use of the issued UID as well as the biometric data collected from the Indian citizens for the generation of biometric watermarked document which uniquely identify the owner of the document. This invokes an additional layer of authentication to the underlying system.

The rest of the paper is organized as follows. In the next section, a brief discussion on the related work is given. Section III explains the proposed system in detail. The next section discusses the experiments and results and finally the concluding remarks and future work are mentioned in the last section.

II. RELATED WORK

Digital watermarking is a method that has received a lot of attention in the past few years. A digital watermark can be described as a visible or preferably invisible identification code that is permanently embedded in the data. Most of the watermarking model is used to embed a logo or private information into the document/other multimedia data for the purpose of copyright protection. By adding watermark, we add a certain degree of protection to the document/ image (or to the information that it contains). The goal is to embed some information in the image without affecting its visual content. In the copyright protection context, watermarking is used to add a key in the multimedia data that authenticates the legal copyright holder and that cannot be manipulated or removed without impairing the data.

F.Liu et.al [3] proposed a watermarking scheme for multiple cover images and multiple owners. It makes use of the visual cryptography technique, transform domain technique, chaos technique, noise reduction technique and error correcting code technique where the visual cryptography technique provides the capability to protect the copyright of multiple cover images for multiple owners, and the rest of the techniques are applied to enhance the robustness of the scheme. A new technique of color image digital watermarking based on visual cryptography is proposed by S.Kandar et,al [4]. In a digital image watermarking model proposed by H.Nyeem et.al. [5], a secret key is employed in both watermark encoding and decoding by which the model gains a resistant against many attacks. A very simple and effective watermarking technique is proposed by P.W.Wong et.al.[6]. Block by block the image is being watermarked with the help of secret key/ public key which provides additional protection to the watermarked image. S.Kandar et al. [7] proposed Visual Cryptographic scheme for color images where the divided shares are enveloped in other images using invisible digital watermarking. The shares are generated using Random Number.

Merging biometrics with the digital watermarking technique is called as biometric watermarking. A few proposals are made in biometric watermarking too. V.S.Inamdar et,al.[1] proposed a biometric watermarking scheme to watermark handwritten signature for signature authentication. Another one biometric watermarking technique is proposed by A.E.Hassanien [8] to hide iris data in digital images for protecting those iris data from tampering and stealing. A multiple watermarking technique based on wavelets and visual cryptography is proposed by S.Radharani et.al. [9]. Among all the reviewed watermarking techniques, the secret key based block by block watermarking technique proposed by P.W.Wong et.al.[6] is chosen to be the best and effective technique and is followed as the watermarking technique in this paper.

And for the generation of biometric code, a literature review is done on various biometric recognition methods. Among various existing biometrics, iris is considered to have more uniqueness and hence yield a high accuracy. Extensive research is been done in the field of iris recognition. Avila et al. [10] proposed iris recognition for biometric identification using dyadic wavelet transform zero-crossing and achieved a recognition rate of 98%. Li Ma et al. [11] proposed an iris recognition which uses circular symmetric filters for feature extraction and nearest feature line approach is used for iris matching. Tisse et al. [12] proposed iris based personal authentication technique based on gradient decomposed hough transform or integro-differential operators combination for iris localization

and analytic image concept to extract information from iris texture. . Libor Masek [13] developed an iris identification system which employs hough transform for iris segmentation and 1D Log-Gabor filters for feature extraction. Simple and effective techniques such as majority voting and haar transform are applied to the template / iriscode generated by iris algorithms to reduce the storage space requirement and improve the accuracy in a iris recognition system proposed in [14]. It employs Libor Masek process to generate the iris template and it achieves a high accuracy rate of 99.82%. Naveen singh et al. [15] designed a iris recognition system using a Canny Edge Detection scheme and a Circular Hough transform, to detect the iris boundaries in the eye's digital image Among all reviewed iris biometric recognition techniques, iris recognition system based on haar transform and majority voting technique [14] is chosen to generate the biometric code for biometric watermarking, since it yield a good accuracy with less requirement of storage space.

III. PROPOSED SYSTEM

The proposed solution has the following four modules listed below:

- Biometric code generation
- Biometric watermark Generation
- Secret key biometric watermarking of the digital document
- Authentication of biometric watermarked digital document

A. Biometric code generation

The first module collects the biometric data from the user and generates the biometric code. Any biometric data such as face, fingerprint, palmprint, etc., or a combination of them can be used as the source to generate the biometric code. In this paper, the biometric code is generated using iris since iris is more popular among all biometrics because of its greater uniqueness and accuracy.

It is wise to choose a biometric model that need less storage space and high accuracy. Storage space of the biometric templates in the database is of major concern because applications that employ biometric recognition system deal with a huge set of data. So the template which has to be created from the biometric data for storage should occupy less space. So iris recognition system based on haar transform and majority voting technique [14] is used for the iriscode generation. The technique applies multilevel haar transform on biometric templates in order to reduce the storage space and to increase the accuracy by combining the important features of the templates after the haar transform. The generated biometric code is efficient in terms of accuracy and storage space.

Iris image is collected from the Aadhaar database. .Libor Masek iris recognition model [13] is used to generate the iris template from the iris image. Fig.1 explains the steps involved in generating the iris template.

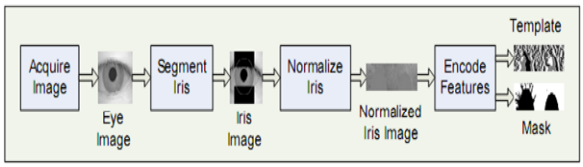


Figure 1: Iris template generation using Libor Masek process

A method based on multilevel haar transform is applied to the generated iris template and the important features of it are combined to generate a feature vector. Then the feature vector is encoded to give biometric code in binary bits. MVIC45 (Majority voting Iris Code 45) method described in [11], is used to generate a 225 iriscode. MVIC45 method is chosen because of it has a good tradeoff between low storage space and high accuracy. This 225 bit iriscode is considered as the biometric code that is to be used in the watermark generation. Fig.2 shows the generated biometric code.

Figure 2: Generated biometric code

The generated biometric code is stored in the database for the authentication purpose. The UID (bio_uid) allotted by the government to each person can be used as an index to store the biometric code in the database. This enables a faster retrieval of the biometric code from the database when many documents of a particular user have to be authenticated.

B. Biometric watermark generation

A watermark has to be generated such that it covers the entire size of the digital document. An authoritative document 'X' of size $M_X \times N_X$ is considered as the cover image for the watermarking procedure. Let 'A' is considered as the binary biometric code image which is of dimension $M_A \times N_A$. 'A' is the source for the generation of the invisible watermark which is to be inserted 'X'. Note that always the size of 'A' should be less than or equal to the size of 'X'.

If 'A' is less than 'X', from 'A' of size $M_A \times N_A$, another binary image 'B' of size $M_X \times N_X$ (same as X) is

formed by tiling process i.e., periodically replicate the image A to generate an image B which is of the same size as 'X'. Else if 'A' is of the same size as that of 'X', 'A' is considered to be 'B'. Fig.3 (a)(b) shows the document 'X' (degree certificate) and the generated biometric watermark 'B' for that document.

Figure 3: (a) Authoritative document X (b) Generated biometric watermark of the document B

C. Secret key biometric watermarking of the digital document

A secret key watermarking scheme based on cryptographic hash function [6] is used. The document can be a grayscale image or a RGB image. Watermarking procedure for a single plane of image is explained below. If it is a grayscale image, procedure can be applied directly in the single plane. If the document is a RGB image, the watermarking has to be applied independently to each of the 3 planes to get the final watermarked RGB image. Fig 4 explains how the generated watermark is inserted into the cover image/document.

The document 'X' and the generated watermark 'B' is divided into blocks of certain size say i×j for the watermarking procedure. Each block in the image is indexed with an integer number 'r'. X_r, B_r represents the r^{th} block of X and B respectively. For each block of X, say X_r, a block X_r is generated by setting the LSB of all the pixels of block X_r to 0. Fig.4 explains how the r^{th} block of generated watermark is inserted into the r^{th} block of cover image / document.

A cryptographic hash function H is used in the watermarking insertion and extraction process of each image block. A cryptographic hash function is a hash function that can be defined as a deterministic procedure that takes an arbitrary block of data and returns a fixed-size bit string, the (cryptographic) hash value, such that an accidental or intentional change to the data will change the hash value.

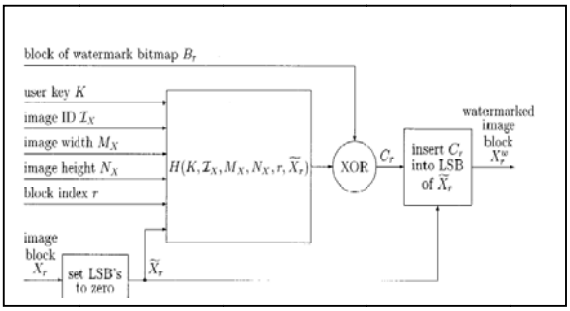


Figure 4: Biometric watermark insertion into the document

The ideal cryptographic hash function has four main or significant properties: 1) it is easy to compute the hash value for any given message 2) it is infeasible to generate a message that has a given hash 3) it is infeasible to modify a message without changing the hash 4) it is infeasible to find two different messages with the same hash.

A cryptographic hash function H(S) is used to generate a message digest of a fixed size p from the string S. String S is a concatenated string which is composed from the hash inputs secret key k, Unique id of the image I_x, width of the document M_x, height of the document N_x, block index r. String S is used as the input to the hash function to generate the digest D of p bits as the hash output. Hash function is applied to each and every block of the image.

$$H(k, I_x, M_x, N_x, r, X_r) = H(S) = (d_1, d_2, d_3, \ldots, d_p)$$

where d_i is the i^{th} bit of the message digest and p is the number of bits in the message digest.

The secret key k is a character array of arbitrary length. Secret key is known to the issuer as well as the verifier. Issuer and the verifier of the document is same in most of the cases. For image id I_x, a table containing a list of documents type and its corresponding unique id say doc_uid is maintained. Each person in the database will be assigned with a unique id say bio_uid. It is a 12 digit number of Aadhar card that uniquely identifies each person with his/her biometrics. The unique image id I_x, is a combination of the UID of the person and the UID of the document i.e. the concatenation of bio_uid and doc_uid. Table 1 lists some of the common authoritative documents and its assigned unique id.

Hash digest D contains a p bit array. If p<i×j, the p bits are replicated to form a bit array of dimension i×j. If p>i×j, then the exceeding bits of p can be truncated to form a bit array of dimension i×j. The resultant array of size i×j is said to be $d_{i\times j}$. The watermark block B_r of size i×j is combined with $d_{i\times j}$ by XOR operation to form C_r. Then C_r is inserted into the least significant bit of X_r to generate the watermarked block of image

X_r^w. Repeat this procedure to generate the watermarked block for all the blocks. All the output watermarked blocks of the image are merged together to form the watermarked document Y. The watermarked document Y is issued to the user and also stored in the database.

The biometric watermarked document 'Y' is stored in the central database using the image id Ix as index i.e. 'Y' is stored in the bio_uid directory (a folder which belongs to a specific person with his/her bio_uid as the folder name) with doc_uid as its document name. This makes the database to be more efficient and responsive while searching a specific document of a specific person (or) all the documents of the same person. Storing of the issued document and the generated watermark in the database is optional as the biometric code required for verification is already been stored in the Aadhaar database.

TABLE I. DOCUMENT TYPES AND ITS CORRESPONDING UID

Document type	Document UID
Birth certificate	001
Death certificate	002
PAN card	003
Driving license	004
Degree certificate	005
10th certificate	006
12th certificate	007
Land registration	008
Passport	009

D. Authentication of biometric watermarked document

Authentication of biometric watermarked document includes watermark extraction from the biometric watermarked document. Watermark extraction is the exact reverse procedure of watermark insertion procedure which is shown in Fig.5. The watermarked document 'Y' is divided into blocks of fixed size i×j with block index r. The LSB bit of Y_r is extracted to generate G_r before generating $\tilde{Y}_r$ by setting the LSB of Y_r to 0. The same hash function used for insertion is used in the extraction model. $\tilde{Y}_r$ is used instead of $\tilde{X}_r$, all the other inputs are same. The hash output is replicated using the same procedure mentioned in insertion, to generate $\tilde{d}_{i\times j,}$ which is XOR ed with G_r to give Y_r^0 which is the watermark block. All the Y_r^0 blocks of the image Y are combined to give the extracted watermark W_e. The same secret key should be used in order to reconstruct the correct watermark.

In the authentication phase, the biometric code of the person is retrieved from the database using the bio_uid as index. 'A' watermark W_g is generated from the retrieved biometric code by the same tiling procedure mentioned above to generate the watermark for insertion. A comparison is made between the generated watermark 'W_g' and the extracted watermark 'W_e'. The biometric watermarked document is authenticated if a hamming distance of 0 is achieved between them. Otherwise the document is not authenticated.

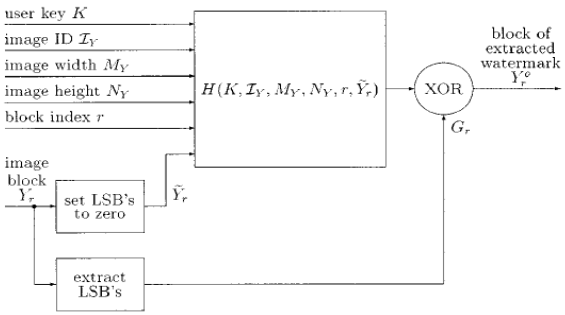


Figure 5: Watermark extraction from the watermarked document

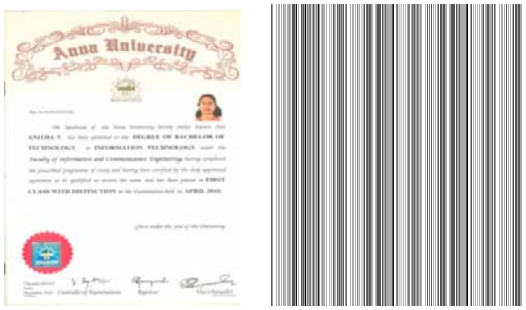
IV. EXPERIMENTS AND RESULTS

A total of 25 documents belonging to 5 different persons (each person has 5 documents) are chosen for the experiment. Biometric watermark for those 5 different persons are generated from their iris. Each biometric watermark is embedded in all the 25 documents i.e., a person's biometric watermark is embedded in all his/her documents as well as all the other person's documents too. And this watermarking is done with 5 different secret keys. Totally there are 25 documents, 5 biometric code and 5 keys. Various combinations of it generate 625 different documents among which only 25 documents are genuine with the watermark of its owner and with its own secret key. The 25 biometric watermarked documents having the original owner's biometric code embedded with the correct secret key alone got authenticated among the 625 different biometric watermarked documents. This shows 100% accuracy in authenticating the real owner of the biometric watermarked document.

Experiments with original document and forged document show how the changes made to the original document reflected in the extracted watermark. Fig.6 (a)(b) shows the original document and the extracted watermark pattern. The extracted watermark shown in Fig.6 (b) has no distortion in the extracted pattern which means the document is not tampered and it can be authenticated. Fig.6 (c) shows the forged document in which the photograph of the original document has been modified. Fig.6 (d) shows the extracted watermark from the forged document. It has a heavy noise in the modified portion and mild noise in the neighborhood

portion around it. From the unmodified portion/pattern, the original owner of the forged document can be found. Original owner can be identified by performing an identification search in the database with the extracted biometric code.

If the biometric watermarked document is extracted with the wrong key, then the extracted watermark will be an image full of noise without any pattern as shown in Fig7. Full noise image is also produced when the watermarked document is cropped/rescaled/when an extraction is made from an un-watermarked document.



(a) (b)

(c) (d)

Figure 6: (a)Original document (b) Forged document (c) Extracted watermark from original document (d) Extracted watermark from the forged document

Any change in the hash inputs during the watermark extraction process will make a huge distortion in the hash output. According to the avalanche effect property of hash algorithm, a change of a single bit in the hash input will lead to a major change in the hash output. In addition to the input document image block X_r, the other inputs to the hash algorithm are (k, I_x, M_x, N_x r). Since the dimension parameters (M_x, N_x) are included in the hash input, any change in the dimension i.e. rescaling/cropping of the watermarked document will result a noisy image during the extraction pattern. The same result will occur in case of wrong secret key (k) /wrong image id (I_x). The block index r is also added to the hash input such that it adds more resistant against attacks. Resistance to Holliman-Memon attack [16] is increased by adding the block index r. If a attacker attempts to reveal the biometric code from a block of image, all the blocks of the other similarly biometric watermarked documents can't be used for the attack. Only the same block from other similar watermarked documents can be used since we are using the block index r. Hence, the size of the code book is small, i.e., one entry in the codebook per watermarked image in the collection. So the chance of the attacker to find the watermark from the document is less.

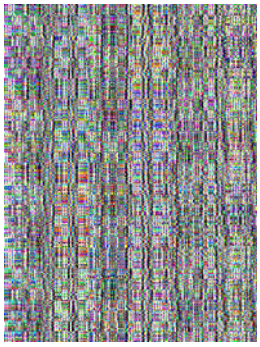


Figure 7: Noisy extracted watermark

V. CONCLUSION AND FUTURE WORK

A simple and efficient technique for digitizing the authoritative documents of an individual based on secret key biometric watermarking key is proposed. It reliably authenticates the real owners of the authoritative documents. Only the authorized person (government/ any organization issuing the document) who owns the secret key, can extract the biometric watermark from the biometric watermarked document for authenticating the document. Hence a third person can't de-watermark and forge the document, since he doesn't possess a valid secret key. If a document is rescaled/ cropped/modified, the changes are directly reflected in the extracted watermark showing where it modified. The original document which is used for forgery can also be figured out. Experiments show that the proposed technique show100% accuracy in authenticating the genuine documents and is resistant against attacks. The future work is to make a study on the performance analysis of the proposed model under attacks.

REFERENCES

[1] V. Inamdar, P. Rege, M. Arya,"Offline handwritten signature based blind biometric watermarking and authetication technique using biorthogonal wavelet transform," International Journal of Computer

Applications (0975 – 8887), vol. 11, Dec. 2010, pp. 19-27

[2] Wikipedia, the free encyclopedia, "Unique identification authority of India", http://en.wikipedia.org/wiki/Unique_Identification_ Authority of_India. Retrieved on : 10 th April, 2012.

[3] F. Liu and K. Wu, "Robust visual cryptography-based watermarking scheme for multiple cover images and multiple owners," IET Information Security, June 2010, doi: 10.1049/iet-ifs.2009.0183, ISSN 1751-8709

[4] S. Kandar, A. Maiti, C. Dhara, "Visual cryptography scheme for color image using random number with enveloping by digital watermarking," International Journal of Computer Science Issues, May 2011 ,Vol. 8, pp.543-549, ISSN : 1694-0814.

[5] H.Nyeem, W. Boles, C. Boyd, " Developing a Digital Image Watermarking Model," Proc. IEEE Symp. International Conference on Digital Image Computing: Techniques and Applications, 2011, pp.468-473,doi: 10.1109/DICTA.2011.85.

[6] W. Wong and N. Memon, "Secret and Public Key Image Watermarking Schemes for Image Authentication and Ownership Verification," IEEE transactions on image processing, vol. 10, no. 10, Oct. 2001.

[7] S. Kandar, A. Maiti, C. Dhara**,"**Visual Cryptography Scheme for Color Image Using Random Number with Enveloping by Digital Watermarking," International Journal of Computer science Issues, Vol. 8, Issue 3, No. 1, May 2011,ISSN :1694-0814.

[8] E.Hassanien,"Hiding iris data for authentication of digital images using wavelet theory," GVIP 05 Conference, Egypt, Dec 2005.

[9] S. Radharani and L. Valarmathi,"Multiple watermarking scheme for image authentication and copyright protection using wavelet based texture properties and visual cryptography," International Journal of Computer Applications (0975 – 8887), Vol. 23, June 2011,pp.29-36

[10] S. C. Avila, S. R. Reillo R, I. D. Martin, "Iris recognition for biometric identification using dyadic wavelet transform zero-crossing," Proc of the IEEE 35th International. Camahan Conference on Security Technology, 2001, pp. 272 -277

[11] L. Ma, Y. Wang, T. Tan, " Iris Recognition Using Circular Symmetric Filters ," Proc. of the 16th International Conference on Pattern Recognition, IEEE press, Vol. 2, 2002, pp. 414 -417.

[12] C. L. Tisse, L. Torres and M. Robert, "Person Identification Technique Using Human Iris," Proc. of the 15th International Recognition conference on Vision Interface, 2002

[13] L. Masek, "Recognition of human iris patterns for bio-metric identification," Tech. Rep., The School of Computer Science and Software Engineering, The university of Western Australia, 2003.

[14] V. Anitha, L. Velusamy, "Iris recognition systems with reduced storage and high accuracy using Majority Voting and Haar Transform," Springer Proc. Advances in Intelligent and Soft Computing book Series, International Conference on communications security and information assurance, May 2012.,in press.

[15] N. Singh, D. Gandhi and K. P. Singh, "Iris recognition system using a canny edge detection and a circular hough transform," International Journal of Advances in Engineering & Technology, vol 1, May 2011, pp.221-228.

[16] M. Holliman and N. Memon, "Counterfeiting attacks on oblivious block-wise independent invisible watermarking schemes," IEEE Trans.Image Processing, vol. 9, pp. 432–441, Mar. 2000.

◈◈◈

Performance Comparison with Accessibility Prediction and Link Breakage prediction in MANETs

Tejal Arvind Sonawale & Shikha Nema

EXTC, V.E.S.I.T, Chembur, Mumbai, India
E-mail : tejal10sonawale@gmail.com, aishuup@gmail.com

Abstract **-** Ad Hoc Networks face a lot of problems due to issues like mobility, power level, load of the network, bandwidth constraints, dynamic topology which lead to link breaks, node break down and increase in overhead. As nodes are changing their position consistently, routes are rapidly being disturbed, thereby generating route errors and new route discoveries. The need for mobility awareness is widely proclaimed. In our dissertation we present a scheme AOMDV-APLP that makes AOMDV aware of accessibility of neighbor nodes in the network. Nodes acquire the accessibility information of other nodes through routine routing operations and keep in their routing table. Based on this information route discovery is restricted to only "accessible" and "start" nodes. Further route with the strongest signal strength is selected from multiple routes using Link life value predicted by Link Breakage prediction technique. Simulation result shows that using accessibility and link life knowledge in route discovery process MAC overhead, routing overhead and average delay is reduced **3** times, and improve the Packet delivery ratio to a large extent than standard AOMDV which reflects effective use of network resources.

Keywords*- Ad hoc networks; Routing protocols; ; QoS,link breakage, accessibility prediction.*

I. INTRODUCTION

Wireless technologies are unequivocally among the most rapidly progressing technology sectors. There is a vast range of wireless technologies, applications and devices, which are either already a substantial part of our daily life or could play this role in future. Wireless ad hoc networking is one of these applications, which can potentially enhance our abilities to solve real life challenges.

Wireless ad hoc networking or Infrastructure-less networking can be considered as an extension to the autonomy that was anticipated with the introduction of wireless networking. Wireless ad hoc networking makes those real life scenarios possible where there is a need of instantaneous and prompt communication. There is a widespread range of scenarios, from conventions or meetings with people quickly sharing information to the emergency search-and-rescue operations, where such networks are well suited. A wireless ad hoc network is a random collection of devices with radio transceivers that accompany each other without any prior infrastructure in a temporary manner to collaboratively accomplish a task.

- The participants i.e. the devices or the nodes can be stationary, mobile, or both, and they can join or leave the network as per their requirement. Similarly, wireless ad hoc networks have technically no geographical limitations on their size; a wireless ad hoc network can be as large as possible provided that all the nodes are able to communicate with each other, though the commonly available range is restricted from the body area to the local area. The concept of wireless ad hoc networking has numerous real life applications as it provides a simple, flexible, effortless, and instant approach to communicate in a cooperative scenario.

- Efficient communication
- Technological limitations
- Resource limitations
- Security
- Quality of service

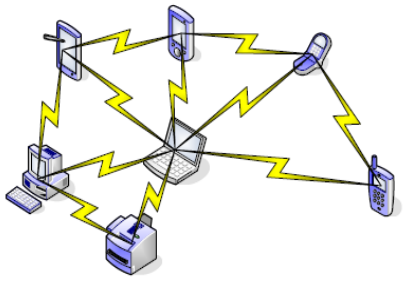


Fig 1-1: A typical wireless ad hoc networks

II. ROUTING PROTOCOL

AODV [11] is an improvement on DSDV. AODV makes use of the on-demand approach for finding routes. A route is established only when it is required by a source node for transmitting data packets and it maintains these routes as long as they are needed by the sources. AODV performs hop-by-hop routing by maintaining routing table entries at intermediate nodes. A node updates its route information only if the destination sequence number of the current received packet is greater than the destination sequence number stored at the node. This indicates freshness of the route and prevents multiple broadcast of the same packet. AODV makes use of the broadcast identifier number that ensures loop freedom since intermediate nodes only forward the first copy of the same packet and discard the duplicate copies. There are three phases of the AODV Routing Protocol. First is the Route Request, Route Reply and Route Maintenance phase. The Figure 2.1 displays a Wireless Ad Hoc scenario, which consists of 9 mobile nodes where the route has to be set from source (S) to destination (D).

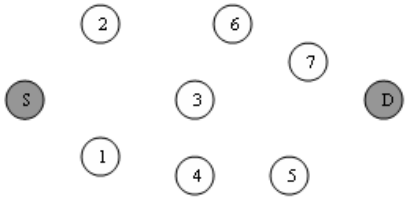


Fig. 2-2 : Wireless Ad Hoc Network Scenario

- **Route Request Phase:**

The route discovery process is initiated when a source needs route to a destination and it does not have a route in its routing table it floods the network with RREQ packet specifying the destination for which the route is requested.

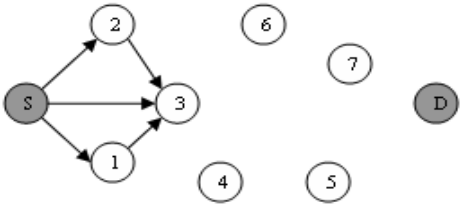


Fig. 2-3 : Route Request Broadcast

The figure 2.2 shows the broadcast of Route Request to the neighboring nodes. Here if node 3 has already received request from node S then it will discard the request that will come from node 1 and node 2. The nodes 1 and 2 will further broadcast it to their neighboring nodes 6 and 4 and if all the intermediate nodes do not have a route to the destination then the request is further broadcast to node 7 and 5 and thus it reaches the destination node D.

- **Route Reply Phase:**

The second phase is the Route Reply phase if the neighboring nodes have route to the destination then the node generates a RREP and sends back to the source along the reverse path and if it does not have the route then the request is forwarded to other nodes. Once the source node receives the RREP it can start using the route to send data packets. The source node rebroadcasts the RREQ if it does not receive a RREP before the timer expires, it attempts discovery up to maximum number of attempts or else aborts the session. It also makes a reverse route entry in its routing table and then forwards the packet. S starts sending the data from whichever route it receives the RREP and then changes the route if it receives the route with a less hop count.

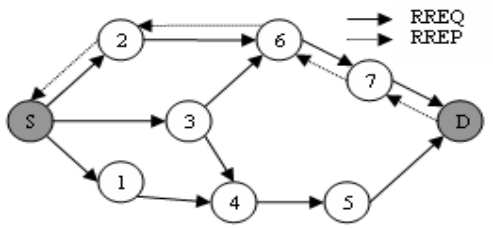


Fig. 2-4 : Route Reply Phase

The table 2.1 shows the routing table of AODV maintained by each node. The table consists of 5 fields the address of the destination node, sequence number, hop count, next hop and expiration time out. As each node just contains a single route to the destination if this route fails then a new route discovery has to be run by the source node. The Destination entry of the routing table specifies the node D where the source has destined the packet. The Sequence Number helps in maintains the freshness of the route. The Hop Count specifies the number of hops required by the source to reach the destination. The Next Hop specifies the next hop taken by the data to reach the destination D.

Table 2-1 Routing Table of AODV

Destination	Sequence Number	Hop Count	Next Hop	Expiration Timeout
D	1234	4	2	…..

- **Route Maintenance phase:**

If one of the intermediate nodes changes its position or fails then the neighboring node realizes the link failure and sends a link failure notification to its upstream neighbors. After the link failure notification

has reached the source it will reinitiate a route discovery if needed. The HELLO messages are sent at regular intervals by the intermediate nodes to find the correct information of the neighboring node. Here if the link between node 2 and node 6 goes down then a new route discovery is run and a path is set up between node S and node D.

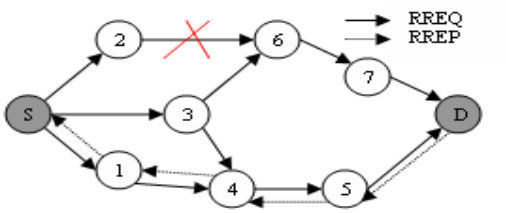


Fig. 2-5 : Route Maintenance Phase

The routing table2.2 displays the modified table of AODV protocol after the link failure.

Table 2-2 Modified Routing Table of AODV

Destination	Sequence Number	Hop Count	Next Hop	Expiration Timeout
D	1234	4	1	…..

The biggest drawback of AODV is with respect to its route maintenance. If a node detects a broken link while attempting to forward the packet to the next hop then it generates a RERR packet that is sent to all sources using the broken link. The source runs a new route discovery after receiving RERR packet. The frequent route breaks cause intermediate nodes to drop packets because no alternate path to destination is available. This reduces overall throughput, packet delivery ratio and increases average end-to-end delay if there is high mobility. The other drawback is that multiple RREP packets are received in response to a single RREQ packet and can lead to heavy control overhead. The HELLO message leads to unnecessary bandwidth consumption. Let us have a look at the already existing AODV protocol.

AOMDV

Adhoc On-demand Multi-path Distance Vector (AOMDV) [12] is an extension to the AODV. The main difference lies in the number of routes found in each route discovery. A little additional overhead is required for the computation of multiple paths. This protocol does not require any special type of control packets but makes use of AODV control packets with a few extra fields in the packet headers. The AOMDV protocol computes multiple loop-free and link-disjoint paths. There are three phases of the AOMDV protocol. The first phase is the Route Request, second is the Route Reply and the third phase is the Route Maintenance phase.

- **Route Request:**

The protocol propagates RREQ from source towards the destination. The figure 2.5 will show the working of AOMDV, which allows multiple RREQ to propagate. The node S as shown in Figure 2.5 has to set a path to the destination node D. So node S as in AODV broadcasts multiple requests to its neighboring nodes 1 and 2. This means that request with same sequence numbers are sent to the destination node. They further broadcast the request to the other neighboring nodes, which are further sent to the destination node D.

- **Route Reply:**

The protocol establishes multiple reverse paths both at intermediate nodes as well as destination. Multiple RREPs traverse these reverse paths back to form multiple forward paths to the destination at the source and intermediate nodes. If the intermediate nodes have the route defined for the destination then they send the RREP to the source node S. The protocol is designed to keep track of multiple routes where the routing entries for each destination contain a list of next hops together with the corresponding hop counts. All the hop counts have the same sequence number then the path with the minimum hop count is selected and all the other paths are discarded. The protocol computes multiple loop-free and link-disjoint paths. Loop-freedom is guaranteed by using a notion of “advertised hop count”. Each duplicate route advertisement received by a node defines an alternative path to the destination. To ensure loop freedom, a node only accepts an alternative path to the destination if it has a lower hop count than the advertised hop count for that destination. The advertised hop count is generally the maximum hop count value possible for a node S to reach a node D. If any value that is received by the source S is greater than the advertised hop count value then a loop is formed so this RREP is discarded. The multiple RREPs are received by the source via multiple paths and a minimum hop count route is selected, the other routes carrying a higher hop count value are discarded.

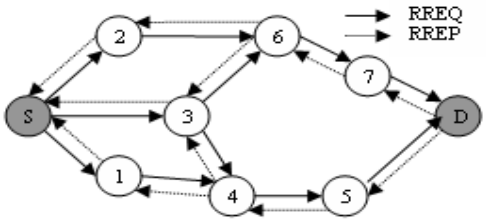


Fig. 2-6 : Working of AOMDV

Destination is the node where the packet is destined to, the sequence number to maintain the freshness of the routes, the advertised hop count that avoids the formation of loops. The route list consists of Hop Count required to reach a particular destination, Next Hop is the next hop the packet is supposed to take to reach the required destination, Last Hop is the last hop taken to reach the destination. If the packet is following the same path then this value is same as the Next Hop or else it changes and Expiration Timeout is the time for which the path will exist. There are multiple entries for a single destination but the routes that contain the lowest hop count are only recorded in the routing table and the other routes are discarded.

- **Route Maintenance Phase:**

The third phase is the Route Maintenance Phase. This phase works in exactly same as AODV. If the intermediate nodes are not able to receive a response of the HELLO message then they broadcast a Route Error message. After receiving this message all the nodes that use the particular route to reach the destination make this particular route as infinity and inform the source node to run a fresh route discovery. The routing table after a link break will appear as follows:

As node 3 has gone down the modified routing table of S will appear as above. When node 7 or node 5 goes down and there are no routes left in the routing table of S then the route discovery will be run. So it surely provides an improvement over AODV.

The above mechanism establishes loop free paths at every node but these paths have to be made disjoint. There are two types of disjoint paths, one is the node disjoint and the other is the link disjoint. Node-disjoint paths do not have any nodes in common, except the source and destination. The link disjoint paths do not have any common link.

An AODV protocol is been developed which develops route on-demand. The biggest drawback of AODV is with respect to its route maintenance. If a node detects a broken link while attempting to forward the packet to the next hop then it generates a RERR packet that is sent to all sources using the broken link. The source runs a new route discovery after receiving RERR packet. The frequent route breaks cause intermediate nodes to drop packets because no alternate path to destination is available. This reduces overall throughput, packet delivery ratio and increases average end-to-end delay if there is high mobility. The other drawback is that multiple RREP packets are received in response to a single RREQ packet and can lead to heavy control overhead. The HELLO message leads to unnecessary bandwidth consumption.

The AOMDV is an extension to the AODV protocol for computing multiple loop-free and link-disjoint paths. The protocol computes multiple loop-free and link-disjoint paths. Loop-freedom is guaranteed by using a notion of “advertised hop count”. Each duplicate route advertisement received by a node defines an alternative path to the destination. To ensure loop freedom, a node only accepts an alternative path to the destination if it has a lower hop count than the advertised hop count for that destination. With multiple redundant paths available, the protocol switches routes to a different path when an earlier path fails. Thus a new route discovery is avoided. Route discovery is initiated only when all paths to a specific destination fail. For efficiency, only link disjoint paths are computed so that the paths fail independently of each other.

In AOMDV RREQs reaching the node may not be from disjoint paths, if RREQ is from one common node one of the RREQ is discarded, this messages implicitly provide knowledge about the mobility and accessibility of their sender and originator. for example, if node A is constantly receiving messages initiated by another node B, this implies that node B is relatively stationary to node A. furthermore a valid route from node A to node B is available either directly or through other nodes. Instead of discarding repeated RREQs messages node can perform additional computation on available routing data and predict accessibility of other nodes. In terms of cost, AOMDV-AP has two additional characteristics. Firstly, repeated RREQs are used for routing table maintenance. Certainly, the additional overhead of performing this action is negligible because this RREQ is already available to the routing agent and all it has to do is to update one or two entries in the routing table. Secondly, routing entries remain permanently in the routing table. As a result, routing tables have more entries (and they also have an additional field in every entry). Use of repeated RREQs further stimulates this issue by adding entries, which were usually discarded. However, in our view, for an ad hoc network with a fair number of nodes such a situation will not cause serious problems. Larger routing tables have a positive role too. During the route discovery process, intermediate nodes can generate RREPs if they have a valid route to the destination; thereby, flooding of RREQ is obstructed. Undoubtedly, flooding has the worst effects on the performance of an ad hoc network.

Now AOMDV[8] routing make use of pre-computed routes determined during route discovery. These solutions, however, suffer during high mobility because the alternate paths are not actively maintained. Hence, precisely when needed, the routes are often broken. To overcome this problem, we will go for link breakage prediction. Prediction will be done only for multiple paths that are formed during the route

discovery process. All the paths are maintained by means of periodic update packets unicast along each path. These update packets are MAC frames which gives the transmitted and received power from which distance can be measured. This distance can be used to predict whether the node is moving inward or outward relative to the previous distance value that is it give the signal strength. At any point of time, only the path with the strongest signal strength is used for data transmission.

III. AOMDV WITH ACCESSIBILITY PREDICTION

In AOMDV repeated RREQs are not discarded. All duplicate RREQs arriving at the node are examined but not propagated further as each duplicate defines an alternate route. Thus AOMDV allows for multiple routes to same destination sequence no. With multiple redundant paths available, the protocol switches routes to a different path when an earlier path fails. Thus a new route discovery is avoided. Route discovery is initiated only when all paths to a specific destination fail. Routing table entry has one common expiration timeout regardless of no of paths to the destination. If none of the paths are used until the timeout expires, then all the paths are invalidated and the advertised hop count is reinitialized. While doing all this, routing information such as RREQs,RREP and REER packets collected can be used to predict the accessibility of nodes. This prediction is used to reduce routing overhead, MAC overhead and to enhance packet delivery ratio and connection success ratio.

3.1. Accessibility Prediction algorithm

a) If a node A receives a routing packet from another node B, node B is in A's neighborhood and is accessible to A.

b) If a node A receives a routing packet originated by a node B, node B is accessible to node A and there exists a valid route from node A to node B.

c) If a node A receives a RERR from a node B, all the unreachable nodes mentioned in this RERR are no more accessible to node A through node D.

Routing entries will never be deleted a new field "Accessible" is added to each routing table entry depicts the predicted accessibility information

Possible values

Start = No information

Accessible = A valid route to node exists or would be possible

Inaccessible = A valid route to node would not be possible

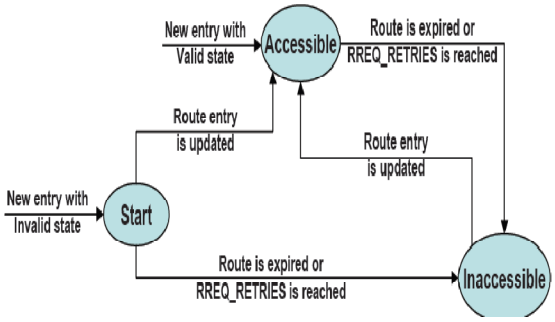


Fig. 3.1: State diagram of AOMDV with accessibility Prediction

Table 3-1 : Routing Table of AOMDV-APLP

Destination
Sequence number
Advertised_hopcount
Expiration _timeout
Route list {(nexthop1,hopcount1),(nexthop2,hopcount2),…
Accessibility

Cost

No extra messaging is required.

Additional computation due to "Accessibility" field is negligible.

Computation cost of using repeated RREQs is negligible.

Routing table entries are never deleted.

Size of routing table might not be a problem in a reasonable size network.
Relative stationary nodes are good candidate to be included in route

3.2 Modified route discovery with accessibility prediction

There is no route discovery for "Inaccessible" nodes, which reduces overhead. The value of the accessibility field is just a prediction. It is likely that this information gets stale. To assume an "Inaccessible" node "Accessible" is not an issue as in such a situation usual AOMDV procedures will be followed. However, the converse could have serious consequences. For example, nodes can conserve plenty of resources by not performing route discoveries for "Inaccessible" nodes, provided the prediction is correct. However, if this prediction is incorrect, this resource conservation will cost them in the form of connectivity loss and consequently throughput loss. Thus, in such a situation

there is a trade-off between overhead reduction (or resource conservation) and connectivity (or throughput).

IV. MODIFIED AOMDV WITH LINK BREAKAGE PREDICTION

Now AOMDV with accessibility prediction routing protocol make use of pre-computed routes determined during route discovery. These solutions, however, suffer during high mobility because the alternate paths are not actively maintained. Hence, precisely when needed, the routes are often broken. To overcome this problem, we will go for link breakage prediction. Prediction will be done only for multiple paths that are formed during the route discovery process. All the paths are maintained by means of periodic update packets unicast along each path. These update packets are MAC frames which gives the transmitted and received power from which distance can be measured, this distance can be used to predict whether the node is moving inward or outward relative to the previous distance value that is it give the signal strength. At any point of time, only the path with the strongest signal strength is used for data transmission. Following is the method to calculate link lifetime.

From two ray ground model we get Transmitted power PT and Received power PR using which we can calculate distance 'd' by given formula.

P_r = **k** $\frac{P_t}{d^4}$ **where k =** $G_{t.}.G_r.(h_t.h_r)^2$ **is a constant**

A link breakage algorithm is used to predict the value of $\mathbf{t_{break}}$ using **'d'.**

4.1 Link Breakage Algorithm

Now $\mathbf{t_{break}}$ can be calculated by the following algorithm

Always assume nodes moving radially outward. Initially

$$V = V_{prev} = V_{\max}\ m/s, d_{prev} = 0.0m$$

$$v = \left| \frac{d - d_{prev}}{t - t_{prev}} \right|$$

$$V = (w)^* v + (1 - w)^* V_{prev}\ \text{k}$$

w based on ratio of time since last sample (!t = t – t prev) and average sample interval T

Time dependency of w ensures quick adaptation to change

$$t_{break} = \left[\frac{d_{\max} - d}{V} \right]$$

$$V_{prev} = V; d_{prev} = d$$

Thus Accessibility and Link Breakage Prediction (APLP) techniques are implemented in AOMDV protocol the proposed protocol has produced good results. The proposed AOMDV-APLP protocol has reduced MAC overhead, Routing overhead and end-to-end delay. On account of which packet delivery ratio is increased a lot as compared to AODV-AP (accessibility prediction) and standard AOMDV.

V. PERFORMANCE METRICS

- **MAC overhead** – the total number of all kinds of MAC packets generated during the simulation time. The retransmission of data frames are also included in it.
- **Routing overhead** – it includes all kinds of AOMDV packets generated as well as forwarded during simulation.
- **Average Delay** – The average end–to-end delay is defined packets traveling from the source to the destination node. The packets generally sometimes get delayed due to transmission, processing, collision and queuing.
- **Packet Delivery Ratio** – The ratio of total number of data packets successfully received by all the destinations to the total number of data packets generated by all the sources.

VI. CONCLUSION

AODV came up with the advantage of the routes being discovered a single route on-demand but this caused a lot of packet delay, Routing and MAC overhead on node failure as a new route discovery had to be run by the source and RREQs are send to all the nodes. AODV-APLP came up with the solution of above problem but the number of routes to the destination is one. AOMDV came up with the advantage of multiple routes being discovered and the route carrying the minimum hop count value is selected but it suffers from large Routing, MAC overhead and Packet delay on node failure, because RREQs are send to all the nodes neighboring nodes. We proposed and implemented AOMDV-APLP where RREQs or route discovery is initiated only for "Accessible" and 'start" nodes which reduces the MAC overhead, Routing overhead, Packet Delay.

Results show that, our proposed protocol, reduces packet delay by 70%, and increases packet delivery ratio

considerably as compared to standard AOMDV protocol. Our protocol also gives stable connectivity as route with the strongest signal strength is selected with the help of Link lifetime.

REFERENCES

[1] www.ietf.org/html.charters/manet-charter.html

[2] Murthy, S. and J.J. Garcia-Luna-Aceves, An Efficient Routing Protocol for Wireless Networks, ACM Mobile Networks and App. J., Special Issue on Routing in Mobile Communication Networks, Oct. 1996, pp. 183-97.

[3] C. E. Perkins and P. Bhagwat, "Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers", in Proceedings of ACM SIGCOMM 1994, August 1994, pp. 234-244.

[4] G. Pei, M. Gerla and T.-W. Chen, Fisheye State Routing in Mobile Ad Hoc Networks. In Proceedings of the 2000 ICDCS Workshops, Taipei, Taiwan, Apr. 2000, pp. D71-D78

[5] B. Johnson, D.A. Maltz, Y.-C. Hu, "The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks", Internet-Draft, draft-ietf-manet-dsr-10.txt, July 2004,

[6] V. Park, and S. Corson, Temporally Ordered Routing Algorithm (TORA) Version 1 Functional Specification. IETF Internet draft, 1997.

[7] Y. B. Ko and N. H. Vaidya. Location Aid Routing (LAR) in mobile ad hoc networks. In Proc. ACM/IEEE MOBICOM, Oct. 1998.

[8] Z. J. Haas and M.R Pearlman, "The Zone Routing Protocol (ZRP) for ad hoc networks", IETF Internet draft , August 1998.

[9] Navid Nikaein, Christian Bonnet and Neda Nikaein, "HARP - Hybrid Ad Hoc Routing Protocol", in proceeding of IST 2001: International Symposium on Telecommunications, Iran/Tehran 2001.

[10] M. Joa-Ng and I-Tai Lu, "A peer-to-peer zone-based two-level link state routing for mobile ad hoc net-works", IEEE on Selected Areas in Communications, vol. 17, no. 8, pp. 1415 1425, 1999.

[11] C. E. Perkins and E. M. Royer, "Ad Hoc On-Demand Distance Vector Routing", Proceedings of IEEE Workshop on Mobile Computing Systems and Applications 1999, February 1999, pp. 90-100.

[12] Mahesh K. Marina and Samir R. Das,"On-Demand Multipath Distance Vector Routing in Ad Hoc Networks", in IEEE Proceedings 2001.

[13] C. E. Perkins, E. M. Royer, S. R. Das, and M. K. Marina,"Performance comparison of two on-demand routing protocols for ad hoc networks",. IEEE Personal Commun. Mag., vol. 8, pp. 16.28, Feb. 2001.

[14] J. Broch, D. A. Maltz, D. B. Johnson, Y.-C. Hu, and J. Jetcheva, "A performance comparison of multi-hop wireless ad hoc network

[15] routing protocols", in Proceeding. Int. Conf. on Mobile Computing and Networking, Oct. 1998.

[16] C. E. Perkins, ed., Ad Hoc Networking, ch. 3. Addison Wesley, 2001.

[17] Rohit Dube, Cynthia D. Rais, Kuang-Yeh Wang, and Satish K. Tripathi, "Signal stability based adaptive routing (SSA) for adhoc mobile networks", in IEEE Personal Communication, February 1997, vol. 4.

[18] Ian D. Chakeres and Elizabeth M. Belding-Royer, "The utility of hello messages for determining link connectivity", in Proceedings of the 5th International Symposium on Wireless Personal Multimedia Communications (WPMC), Honolulu, Hawaii, October 2002.

[19] T. Camp, J. Boleng, V. Davies, "A survey of mobility for ad hoc network research", in Proc. WCMC, 2002, pp.483-502.

[20] Michael Bahr, "Proposed Routing for IEEE 802.11s WLAN Mesh Networks", In Proc. 2nd Annual International Wireless Internet Conference, 2006.

[21] P Sambasivam, A Murthy, EM Belding," Dynamically Adaptive Multipath Routing based on AODV", In Proc. 3rd Annual, 2004- Citeseer

[22] Rehman, H. Wolf, L. "Performance enhancement in AODV with accessibility prediction", in mobile adhoc and sensor system, MASS 2007 IEEE International conference,pp 1-6

[23] http://www.isi.edu/nsnam/ns-build.html#allinone

◈◈◈

Price Demand Model For A Cloud Cache

Shaik Mohammed Gouse[1] & **Dr.G. Prakash Babu**[2]

Computer Science and Engineering, Intell Engineering College,Anatapur, Andhra Pradesh, India.
E-mail : smilessame@gmail.com & gpbabu27@gmail.com

Abstract - Cloud applications that offer data management services are emerging. Such clouds support caching of data in order to provide quality query services. The users can query the cloud data, paying the price for the infrastructure they use. Cloud management necessitates an economy that manages the service of multiple users in an efficient, but also, resource economic way that allows for cloud profit. Naturally, the maximization of cloud profit given some guarantees for user satisfaction presumes an appropriate price-demand model that enables optimal pricing of query services. The model should be plausible in that it reflects the correlation of cache structures involved in the queries. Optimal pricing is achieved based on a dynamic pricing scheme that adapts to time changes. This paper proposes a novel price-demand model designed for a cloud cache and a dynamic pricing scheme for queries executed in the cloud cache. The pricing solution employs a novel method that estimates the correlations of the cache services in an time-efficient manner. The experimental study shows the efficiency of the solution.

Keywords *: Computing, Optimal, Infrastructure, Centralizing, Adaptivity, Service.*

I. INTRODUCTION

The name cloud computing was inspired by the cloud symbol that's often used to represent the Internet in flowcharts and diagrams. Cloud computing is a technology that uses the internet and central remote servers to maintain data and applications. Cloud computing allows consumers and businesses to use applications without installation and access their personal files at any computer with internet access. This technology allows for much more efficient computing by centralizing storage, memory, processing and bandwidth.

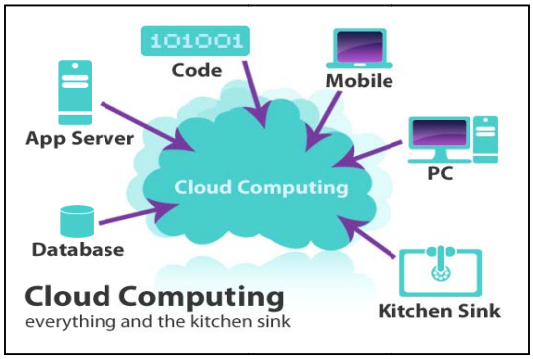
A simple example of cloud computing is Yahoo email or Gmail etc. You don't need software or a server to use them. All a consumer would need is just an internet connection and you can start sending emails. The server and email management software is all on the cloud (internet) and is totally managed by the cloud service provider Yahoo, Google etc. The consumer gets to use the software alone and enjoy the benefits. The analogy is, '***If you only need milk, would you buy a cow?'*** All the users or consumers need is to get the benefits of using the software or hardware of the computer like sending emails etc. Just to get this benefit (milk) why should a consumer buy a (cow) software /hardware? For instance, without a cloud, you might run a web server that is a single computer. Maybe that computer is powerful enough to serve 1000 pages per minute. If your website suddenly becomes more popular, and the audience demands 2000 pages, it will take two minutes for all of them to get their pages displayed on their web browsers. If it gets even more popular, your server will slow to a crawl, and your audience might start losing interest. During other times, your server might only handle a few pages per minute, and the rest of the time it'll sit there waiting, instead of serving pages. If you moved your server to a cloud, however, you might rent computer power from a service provider who might have thousands of servers, all connected together so they can share work between each other. You would share those servers with perhaps thousands of other websites, some big, some small, like an apartment building for websites. If your website suddenly becomes more popular, the cloud can automatically direct more individual computers to work to serve pages for your site, as you pay more rent for all the extra usage. If it spends months being unpopular,

however, the rent you pay may dwindle down to a trickle. Meanwhile, some other, more popular site can use the computers that you are not using. In this way, computing power becomes more fluid.

A. How Does It Work?

Cloud computing provides the ability to store, access, manipulate, and share information without ever having to save the data to a local computer. Instead, the data rests on servers located off-site. There are three basic forms of cloud computing services, each offering the user a certain amount of control.

- Infrastructure as a Service (IaaS) - A service in which a client is provided with an entire infrastructure, including software and hardware, and pay a fee based upon usage. An example of this is Amazon's S3.
- Platform as a Service (PaaS) - Services like Google App Engine allow a client to develop their own applications, but host them upon the proprietor's infrastructure for a fee.
- Software as a Service (SaaS**)** - This is where a user is presented with the software only, sometimes for free or by subscription.



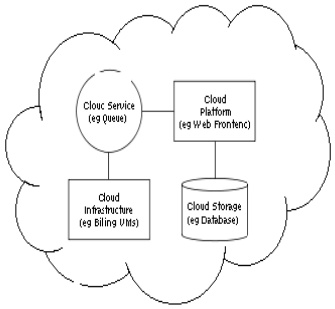
There was a time when Cloud Hosting was been used only by the governments but not it is making its way towards business, both large and small. As mentioned in Cloud Hosting "is dynamically scalable and provides virtualized resources as a service over the internet." Think about the multitude of servers that are connected via networks to create a cloud where companies can store their data. Basically this cloud acts as an outsourcing agent for server and storage requirements. As we all know that cloud hosting has become a new buzzword, adoption may or may not be the good choice for your business or company. Just go through these advantages and disadvantages to learn more about your options with cloud computing.

III. ARCHITECTURE

Cloud architecture, the systems architecture of the software systems involved in the delivery of cloud computing, typically involves multiple *cloud components* communicating with each other over loose coupling mechanism such as messaging queue.

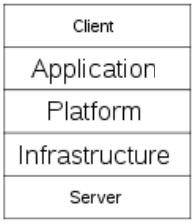
The two most significant components of cloud computing architecture are known as the front end and the back end. The front end is the part seen by the client, i.e., the computer user. This includes the client's network (or computer) and the applications used to access the cloud via a user interface such as a web browser.

The back end of the cloud computing architecture is the *cloud* itself, comprising various computers, servers and data storage devices.



Layers

Once an Internet Protocol connection is established among several computers, it is possible to share services within any one of the following layers



.

Provider

A *cloud provider* is the Company responsible for providing the cloud service.

Client:

cloud client consists of computer hardware and ccvc/or computer software that relies on cloud computing for application delivery, or that is specifically designed for delivery of cloud services and that, in either case, is in essence useless without it.

Examples include some computers, phones and other devices, operating systems, and browsers. Cloud Desktop as a Service or Hosted Desktop, is a term often used to refer to a container of a collection of virtual objects, software, hardware, configurations etc., residing on the cloud, used by a client to interact with remote services and perform computer related tasks.

Application

Cloud application services or "Software as a Service (SaaS)" deliver software as a service over the Internet, eliminating the need to install and run the application on the customer's own computers and simplifying maintenance and support. People tend to use the terms "SaaS" and "cloud" interchangeably, when in fact they are two different things.

Key characteristics include

- Network-based access to, and management of, commercially available (i.e., not custom) software
- Activities that are managed from central locations rather than at each customer's site, enabling customers to access applications remotely via the Web
- Application delivery that typically is closer to a one-to-many model (single instance, multi-tenant architecture) than to a one-to-one model, including architecture, pricing, partnering, and management characteristics

Centralized feature updating, which obviates the need for downloadable patches and upgrades

Platform

Cloud platform services, also known as Platform as a Service (PaaS), deliver a computing platform and/or solution stack as a service, often consuming cloud infrastructure and sustaining cloud applications. It facilitates deployment of applications without the cost and complexity of buying and managing the underlying hardware and software layers.

Infrastructure

Cloud infrastructure services, also known as Infrastructure as a Service (IaaS), deliver computer infrastructure – typically a platform virtualization environment – as a service. Rather than purchasing servers, software, data-center space or network equipment, clients instead buy those resources as a fully outsourced service. Suppliers typically bill such services on a utility computing basis; the amount of resources consumed (and therefore the cost) will typically reflect the level of activity. IaaS evolved from virtual private server offerings.

Cloud infrastructure often takes the form of a tier 3 data center with many tier 4 attributes, assembled from hundreds of virtual machines.

Server

The servers layer consists of computer hardware and/or computer software products that are specifically designed for the delivery of cloud services, including multi-core processors, cloud-specific operating systems and combined offerings.

II. EXISTING SYSTEM

Existing clouds focus on the provision of web services targeted to developers, such as Amazon Elastic Compute Cloud (EC2), or the deployment of servers, such as Go Grid. There are two major challenges when trying to define an optimal pricing scheme for the cloud caching service. The first is to define a simplified enough model of the price demand dependency, to achieve a feasible pricing solution, but not oversimplified model that is not representative.

A static pricing scheme cannot be optimal if the demand for services has deterministic seasonal fluctuations. The second challenge is to define a pricing scheme that is adaptable to

(i) Modeling errors,

(ii) Time-dependent model changes.

(iii) Stochastic behavior of the application. The demand for services, for instance, may depend in a nonpredictable way on factors that are external to the cloud application, such as socioeconomic situations.

Static pricing cannot guarantee cloud profit maximization. In fact, as we show in our experimental study, static pricing results in an unpredictable and, therefore, uncontrollable behavior of profit. Closely related to cloud computing is research on accounting in wide-area networks that offer distributed services. Mariposa discusses an economy for querying in distributed databases. This economy is limited to offering budget options to the users, and does not propose any pricing scheme. Other solutions for similar frameworks focus on job scheduling and bid negotiation, issues orthogonal to optimal pricing.

III. OUR CONSTRUCTION

The cloud caching service can maximize its profit using an optimal pricing scheme. Optimal pricing necessitates an appropriately simplified price-demand model that incorporates the correlations of structures in the cache services. The pricing scheme should be adaptable to time changes.

Price adaptivity to time changes:

Profit maximization is pursued in a finite long-term horizon. The horizon includes sequential non-overlapping intervals that allow for scheduling structure availability. At the beginning of each interval, the cloud redefines availability by taking offline

Some of the currently available structures and taking online some of the unavailable ones. Pricing optimization proceeds in iterations on a sliding time-window that allows online corrections on the predicted demand, via re-injection of the real demand values at each sliding instant. Also, the iterative optimization allows for re-definition of the parameters in the price-demand model, if the demand deviates substantially from the predicted.

Modeling structure correlations:

Our approach models the correlation of cache structures as a dependency of the demand for each structure on the price of every available one. Pairs of structures are characterized as competitive, if they tend to exclude each other, or collaborating, if they coexist in query plans. Competitive pairs induce negative, whereas collaborating pairs induce positive correlation. Otherwise correlation is set to zero. The index-index, index column, and column-column correlations are estimated based on proposed measures that can estimate all three types of correlation. We propose a method for the efficient computation of structure correlation by extending a cache based query cost estimation module and a template-based workload compression technique.

Query Execution:

The cloud cache is a full-fledged DBMS along with a cache of data that reside permanently in back-end databases. The goal of the cloud cache is to offer cheap efficient multi-user querying on the back-end data, while keeping the cloud provider profitable. Service of queries is performed by executing them either in the cloud cache or in the back-end database. Query performance is measured in terms of execution time. The faster the execution, the more data structures it employs, and therefore, the more expensive the service. We assume that the cloud infrastructure provides sufficient amount of storage space for a large number of cache structures. Each cache structure has a building and a maintenance cost.

Optimal Pricing:

We assume that each structure is built from scratch in the cloud cache, as the cloud may not have administration rights on existing back-end structures. Nevertheless, cheap computing and parallelism on cloud infrastructure may benefit the performance of structure creation. For a column, the building cost is the cost of transferring it from the backend and combining it with the currently cached columns. This cost may contain the cost of nte grating the column in the existing cache table. For indexes, the building cost involves fetching the data across the Internet and then building the index in the cache.

Since sorting is the most important step in building an index, the cost of building an index is approximated to the cost of sorting the indexed columns. In case of multiple cloud databases, the cost of data movement is incorporated in the building cost. The maintenance cost of a column or an index is just the cost of using disk space in the cloud. Hence, building a column or an index in the cache has a one-time static cost, whereas their maintenance yields a storage cost that is linear with time.

IV. CONCLUSION

This work proposes a novel pricing scheme designed for a cloud cache that offers querying services and aims at the maximization of the cloud profit. We define an appropriate Price-demand model and we formulate the optimal pricing problem. The proposed solution allows: on one hand, long-term profit maximization, and, on the other, dynamic calibration to the actual behavior of the cloud application, while the optimization process is in progress. We discuss qualitative aspects of the solution and a variation of the problem that allows the consideration of user satisfaction together with profit maximization. The viability of the pricing solution is ensured with the proposal of a method that estimates the correlations of the cache services in an time-efficient manner.

REFERENCES

1. Verena Kantere , Debabrata Dash , Gregory Francois , Sofia Kyriakopoulou , Anastasia Ailamaki Ecole Polytechnique F´ed´erale de Lausanne, *“Optimal service pricing for a cloud cache”*, vol. 23, no. 9, pp. 1345-1358, 2011.
2. G.R. Bitran and R. Caldentey, "An Overview of Pricing Models for Revenue Management," Manufacturing and Service Operations Management, vol. 5, no. 3, pp. 203-209, 2003.
3. N. Bruno and S. Chaudhuri, "An Online Approach to Physical Design Tuning," Proc. Int'l Conf. Data Eng. (ICDE '07), 2007.
4. X.-R. Cao, H.-X. Shen, R. Milito, and P. Wirth, "Internet Pricing with a Game Theoretical Approach: Concepts and Examples," IEEE/ACM Trans. Networking, vol. 10, no. 2, pp. 208-216, Apr. 2002.

5. C. Chen, M. Maheswaran, and M. Toulouse, "Supporting Co-Allocation In an Auctioning-Based Resource Allocator for Grid Systems," Proc. 16th Int'l Parallel and Distributed Processing Symp. (IPDPS '02), 2002.

6. S. Choenni, H.M. Blanken, and T. Chang, "On the Selection of Secondary Indices In Relational Databases," Data and Knowledge Eng., vol. 11, no. 3, pp. 207-233, 1993.

7. D. Dash, Y. Alagiannis, C. Maier, and A. Ailamaki, "Caching All Plans with One Call to the Optimizer," Proc. Self-Managing Database Systems (SMDB), 2010.

8. D. Dash, V. Kantere, and A. Ailamaki, "An Economic Model for Self-Tuned Cloud Caching," Proc. IEEE Int'l Conf. Data Eng. (ICDE '09), 2009.

◈◈◈

Implementation of Data Mining Technique for BPR

***Saju Mathew & **Salu George. T**

*Department of Information Technology, Troy University, USA, Sharjah Campus, Sharjah, U.A.E
**College of Information Technology and Design, Department of Information Technology,
University of Jazeera, Dubai, U.A.E
E-Mail : papers4conference@gmail.com & salugeorge@gmail.com

Abstract **- Business Process Re-engineering (BPR)** is one of the fundamental concerns for radically redesign and rethinking of business process to obtain a steady growth and sustained improvement in terns of quality, cost, services provided, time and the innovation the business is carrying out. It provides a great deal of opportunity to the engineers to redesign and reengineer the whole process and to reduce radically the number of process or activities carried out in the business with the help of advanced tools of information technology. The most fundamental decision for a business intelligence implementer is to check whether the new data evolved should be copied into the central data warehouse or accessed where it is. To implement such an approach effectively the business intelligence enterprise should understand what information technology can do to make the architecture flexible and take a good decision. One such development and technique used in information technology is "**Data Mining (DM)".**
This paper focuses on the use of data mining as a technique to support the process of re-engineering the business by extracting the much needed information and hidden knowledge from huge volume of data stored and maintained in the data warehouse by the organization.

Keywords*-Business process re-engineering, information technology, business intelligence, data mining, knowledge, data warehouse.*

I. INTRODUCTION

One of the crucial factor on which the business process design and its re-engineering depends is the procedure of manufacturing and business services provided to the customer to achieve the business goals and its objective. Very few techniques, methods or formula are exists to support this kind of reasoning as an analytical task. Due to which the individual do their own design based on which the decisions are taken which are very hard to relate to the business perspective and business objective. It without the proper understanding and knowledge of BPR is carried out, then the most likely result will be unsatisfactory and only the outdated process will be automated. This kind of practice misses opportunities for innovation and rationalization. This modeling and analysis of business processes along with business strategies and organizational structures are essential to study before the implementation of BPR.

BPR helps in rethinking a process in order to enhance its performance. Many business practitioners have been developing different methodologies to support the development of BPR principles. But one development in the field of information technology that has helped and recently attracted the attention of many researchers and organizations is "data mining".

"Data mining is a process of discovering actionable and meaningful patterns, profiles and trends by sniffing through your data using patterns recognition technologies such as neural networks, machine learning and genetic algorithms " [10]. In other words, data mining solutions serve to find hidden new data (trends, segmentation, behaviors, patterns, tariff simulators etc.) not visible with ordinary analytical tool. Data mining brings true value added to business.

Data mining tools can answer different business questions that traditionally were too time-consuming to resolve. They search huge database for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. So before the implementation of data mining a proper research should be carried out and few questions should be investigated. They are,

a. Educational background of DM users.
b. Which organization uses DM techniques?
c. Is there any relationship existing between DM and BPR?
d. From where BPR data are obtained?
e. How relevant they are?

f. What categories of knowledge managers will use DM techniques for BPR?

g. What are their likely consequences?

II. KNOWLEDGE MANAGEMENT (KM)

One of the important factors for data mining is to have the proper knowledge. Knowledge which is a complex and fluid concepts is the most expensive article of trade which if managed properly and correctly can prove to be a major asset for the company and business. In nature knowledge can be either explicit or tacit. Explicit knowledge can be easily expressed and transferred to others, whereas tacit knowledge is the personal knowledge residing in the individuals head. It is very difficult to express and communicated to others.

Knowledge management (KM) has achieved a high level of popularity among the various firms worldwide. As such KM has no standard or unique definition. But it can be defined as "systematic process of finding, selecting, organizing, distilling and presenting knowledge in a way that improves the organization interest". The main objective of knowledge management is to ensure that the right knowledge is available at the right time in a manner that enables the timely decision making [9].

There are five different knowledge management strategies. They are;

a. Knowledge management as a business strategy.

b. Intellectual asset business strategy

c. Personal knowledge asset responsibility strategy.

d. Knowledge creation strategy

e. Knowledge transfer strategy

III. DATA MINING TECHNIQUES

Data mining which is an extraction of hidden predictive information from large database is a powerful new technology with great potential to help companies focus on the most important information in their data warehouse. Data mining tools predict future trends and behaviors allowing businesses to make proactive, knowledge driven decisions. Most companies have already collected a large volume of data and refined them. The recent developments and the areas where data mining can be incorporated in medical treatment (disease symptoms identification), retail industry, telecommunication, DNA sequence, natural disaster, web mining, music track selection, content based email processing, analyzing data from a specific experiments which is conducted over time, nations census database etc.

Data mining techniques cab be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources and can be integrated with new products and systems as they are brought online. The most common used techniques in data mining are [10];

1. Artificial Neural Network (ANN): this is a non linear predictive model that learns through training and resembles biological neural network in structure.

2. Decision Trees: the set of decisions are represented in tree form. These decisions generate rules for the classification of a dataset.

3. Genetic Algorithm: they are optimization techniques that use process such as genetics combination, mutation and natural selection in a design based on concepts of evolution. It tries to mimic the way nature works. It is an adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetics.

4. Rule Induction: the extraction of useful if-then rule from data based on statistical significance.

5. Regression Methods; this tries to identify the best linear pattern in order to predict the value of one characteristic studied in relation to another.

A. Data Mining Tasks

Following are some of thee tasks that are solved by data mining;

a. Prediction: the tasks of learning a pattern from examples and using the developed model to predict future values of the target variable.

b. Classification: a task of finding a function that maps records into one of several discrete classes.

c. Detection of relation: a task of searching for the most influential independent variables for a selected target variable.

d. Explicit modeling: a task of finding explicit formula describing dependencies between various variables.

e. Clustering a task of identifying groups of records that are similar between themselves but different from the rest of the data.

f. Market based analysis: processing transactional data in order to find those groups of products that are sold together.

g. Deviation detection: a task of determining the most significant changes in some key measures of data from previous or expected values.

B. Benefits of Data Mining Techniques

If an organization is using data mining techniques then it can enjoy number of benefits such as understanding customer's behavior, making judgment on the effectiveness of the company's website if any, benchmarking marketing campaigns [7][10]. These benefits can be summarized as follows;

a. Establishing the probability of customers coming back to the company or their website.

b. Calculating the number of new customers coming to the company or their website.

c. Identify patterns relating either to navigate the routes that customers follow or to what they buy.

d. Discover who buys what and look for any cross-relationships between the clients.

e. Developing a better layout of the company's website.

f. Identify popular and non-popular areas of the website.

g. Personalizing online advertisement.

IV. BUSINESS PROCESS RE-ENGINEERING (BPR)

As new business functions are added to the existing process it becomes inevitable to think of alternative for having better processes. It always not advisable and practical to replace an existing system, but it may be possible to enhance the existing system or redesign the system. Before changing the process it is necessary to know all the conditions which affect the process and accordingly incorporate a change mechanism. The methodologies which are adopted to bring changes are called as re-engineering [19].

When the existing system in the organization is totally reexamined and radically modified for incorporating the latest technology. This process of change for the betterment of the organization is called as business process re-engineering (BPR). With business process being reengineered, the organization has to change the workflow and business procedures for efficiency in the organization.

When business process re-engineering is used properly and carefully, it can take the organization into new phase of competitiveness with good effectiveness. But the redesign of the individual processes will always have a limited impact unless it is implemented as a part of wider view of the organization.

This paper tries to focus on the data mining techniques that would allow the business practitioners, senior managers and decision makers in the organization to extract the useful, relevant and previously hidden knowledge from organization's database which after careful management of this knowledge leads to business process re-engineering.

C. Importance of BPR

BPR is a major innovation changing the way organization conduct their business. Such changes are often necessary for making the profit and even for survival. BPR is employed when the major IT projects such as ERP are undertaken and reengineering will totally change the organizational structure, culture and processes. The support system such as Expert system (ES), Decision support system (DSS), artificial intelligence (AI) plays a very vital role in BPR. Reengineering is basically done to achieve cost reduction, increase in quality, improvement in speed and services and makes the company more competitive in the market. Following are the factors that are common to all the BPR initiatives [2][19];

a. The need for IT solutions adapted to fit the business.

b. The focus on processes.

c. The intent to use a pilot project approach

d. The need for top management commitment

e. The need for the communication plans

B. The BPR Structure

The idea behind the BPR structure is to help the practitioner's to identify the related topics [1]. The BPR is derived as a synthesis of Work-Centered Analysis (WCA) framework, the MOBILE workflow model, the CIMOSA enterprise modeling views and the process description classes [1][3][13].

C. Methodology

A survey was carried out with some 200 students of computer science covering some academic institutions. Participations in the data gathering exercise was voluntary after the researches had explained the importance of the study to the students. The participants returned 190 useable responses of which 64.9 % were male and 35.1 % were females and 16 questionnaires were either not filled or contain missing data making the overall responses rate as 92.27 %. All the participants are in full time employment with university and research institutes, banks, insurance, ministries if science and technology, accounting and business consultancy firms. A total of 32 work organizations were represented with 9 participants from each organization type. The significance of this heterogeneous sample is that the

respondents are not uniformly influence by the contextual constraints of any single organization [8]. The mean age of participants was 39.39 years of which 74.3 % were married 68 % of respondents use primary data as their source of BPR data while 32 % rely on secondary data.

The scale design phase of the questionnaires used focuses on conduct validity and reliability, operational issues investigating whether the scales chose are true constructs describing the events or merely artifacts of the methodology itself [4][5]. The process started by arranging the selected items in a questionnaires format in preparation for data collection. The items were arranged in random order to reduce the bias. The response options for some of the questions items was graded on five-point scale ranging from strongly disagree to strongly agree.

D. Results

Due to the scope of this research work all the respondents either have theoretical or practical knowledge of data mining techniques. Some criteria are used to determine the manager's knowledge about the relevant BPR data, theoretical or practical knowledge of data mining. The manger's ability to describe any of these criteria gives him a score of 1 mark, manager who explains both the criteria scores 2 marks, while one who cannot explain any of the criteria has zero (0) score.

Following table shows percentage distribution of managers who has the knowledge of data mining techniques.

Knowledge Score	Percentage ratio	Interpretation
0	28	Very Poor
1	41	Average
2	31	Good

Table 1: Percentage distribution of managers who has the knowledge of data mining techniques.

The above table 1, shows that 72 % of responses have the knowledge of data mining techniques since most of the respondents are working in IT department of their organization while 28 % are below average and does not have any knowledge about data mining techniques.

Following table shows respondents view on the use of data mining and BPR.

Questions	Strongly Disagree	Strongly Agree	Total
Anticipating a relationship that may exist between DM and BPR	1	136 (71.6 %)	190
There would be consequence if DM actually succeeds in making BPR with KM	5	107 (57.4 %)	190
DM can affect the power within the organization and the power of organization	5	113 (59.5 %)	190

Table 2: Respondents view on the use of DM and BPR

The above table 2 clearly shows that nearly;

- 71.6% of respondents strongly anticipate that there exists a relationship between DM and BPR. Due to this percentage shows that much research is need in this area.
- 57.4% respondents strongly suspect consequence when data mining actually succeeds in making BPR with KM.
- 59.5% respondents strongly agree that effective use of DM can affect the power within the organization and the power of the organization.

V. CONCLUSION

The process of extracting knowledge hidden from large volume of data has proved very successful in solving many business and scientific problems to achieve competitive advantage. As suggested in DM/BPR framework, the DM model can be deployed on the massive data collected from the past business processes of the organization which helps in finding the previous unknown knowledge and trends needed by the top manages or decision makers in the organization for effective business process redesigning.

Due to recent developments in Internet and telecommunication has made it possible to transmit large amount of data over distance in short period of time, resulting in the accumulation of data on the Internet. This data are stored in files specially created for this purpose called log files, generated by the servers showing list of actions that occurred.

There are many data mining tools available that can convert the raw data which is in log files to useful information and knowledge. A customized program can be written to achieve better results. If these potentials are fully and properly utilized, then decision makers in the organization would be able to answer many unanswered question which was difficult in the past. As

a result a re-engineering can be done to the old business and a new system in the business can be developed.

REFERENCES

[1] Alter,S.," Information Systems: A Management Perspective", Addison Wesley, Amsterdam, 1999.

[2] Ascari.A, Rock. M, and Dutta.S, " Reengineering Organizational Change European Management Journal, Vol.13, No.1, pp 1-30, 2005.

[3] Berrot. G, Vemadat.F, "Enterprise modeling with CIMOSA: Functional and Organizational aspect", production planning and control, Vol.12, No.2, pp 128-136, 2001.

[4] Campbell D.T, Fiske D.W, "Convergent and discremenant validation by the multitrait-multimethod matrix", Psychological Bulletin 56(2), pp 81-105, 2003..

[5] Cronbach L.J, "Test Validation in educational measurement", American Council on Education, Washington, DC, pp.443-507, 1971.

[6] Crow T.J, Fong P.M, Bauman T.A, Zayas-Castro J.L, "Quantitative risk level estimation of business process re-engineering efforts", Business Process Management Journal, Vol.8, No. 5, 22 MCB University Press, pp. 490-511, 16-oct-2002.

[7] http:// www.dmreview.com/ editorial/ dmdirect/dmdirectarticles.efm?

[8] Gupta. A, McDaniel J, "Creating competitive advantage by effectively managing knowledge management,", Journal of Knowledge Management Practice, Vol.3, No. 2, pp. 40-49, 2002

[9] Hariharan.A, "Knowledge Management: A Strategic Tool", Journal of Knowledge Management practice, Vol.#, pp. 50-59, 2002.

[10] Mena.J, "Data Mining your website", dpDigital Press, ISBN: 1-55558-222-2, The United Stated of America, 1999.

[11] Neal. A, Griffin M.A, " Developing a model of individual performance for human resource management", Asia Pacific Journal of Human Resource, 37, pp. 44-59, 1999.

[12] http:// www.tjhsst.edu/dhyatt/perl/ex3.html

[13] Seidmann.A, Sundararajan. A, "The effects of task and information asymmetry on business process redesign", International Journal of Production Economics, Vol. 50, No.213, pp. 117-128, 2007

[14] Selma. L.M, Farhi. M, Hago A.R, " Case-Based reasoning as a technique for knowledge management in business process redesign", Electronic Journal of Knowledge Management, Vol. 1, issue 2, pp. 113-124, 2003

[15] Smith.S, " Rules of engagement", Computer Weekly, pp. 36-37, 14-march-1996.

[16] Strassmann. P, " The Politics of Information Management", Information Economics Press, 2003.

[17] Wendy. R, " Strategic Management and information system: an integrated approach", second edition, Financial Times Professional Limited, 2005.

[18] Wiig. K.M, " Knowledge management: an introduction and perspective", The Journal of Knowledge Manahement, Vol.1, No. 1, pp.6-14

[19] http:// www.websukat.com/smu-books/ management -Information-system/4-bpr.pdf

◈◈◈

Overcoming Limitations of Good-Old Asynchronous Serial Communication by Extending into its Next Generation

Yadynesha Behere & N. Srivastava

Department of Electronics, Bharati Vidyapeeth University College of Engineering, Pune, India-411043
E-mail : yadnyeshab@yahoo.co.in, nksrivastava@bvucoep.edu.in

Abstract **-** Universal Asynchronous Receiver Transmitter (UART) is also known as RS232 standard which is almost 50 years old, so there's a lot of crust in it. It asynchronous interface always has an advantage that there is no need of clock between transmitting and receiving ends. It is still used in many embedded & VLSI applications because it is relatively simple & does not need complicated driver support. Based on these benefits of being asynchronous & simple, we propose an architecture that extends the standard UART format. The extension is in order to support transmission & reception of relatively larger amount data, bring about configurability in extended baud rates and yet enable it's compatibility to an existing standard of RS232 upon configuration. The proposed architecture is an attempt to extend and conFigure for serial communication with 8/16/32 bits transmit & receive over range of standard baud rates.

Keywords **-** *FPGA, RS 232, Serial Communication, UART*

I. INTRODUCTION

Serial communications (RS232) are in use over decades. These days they are just indispensible in any embedded and VLSI applications. That is because of key important role they serve in debugging, testing, configuration etc. UART are low speed in comparison with all other advanced serial communication interfaces. Yet they are very important because of their simplicity and ability to be used with many scalable embedded and VLSI applications.

RS232 is a standard developed by Electronics Industry Association (EIA). This is one of the oldest and most widely used communication interfaces. In embedded system, normally an ASIC/FPGA evaluation board or a microcontroller based board is interfaced to a host machine. These are DCE (Data Communications Equipment) and DTE (Data Terminal Equipment). Data Communications Equipments are devices such as your modem, TA adapter, plotter etc while Data Terminal Equipment is your Computer or Terminal. A Null Modem is used to connect two DTE's together. This is commonly used as a cheap way to network games or to transfer files between computers using Zmodem Protocol, Xmodem Protocol etc. This can also be used with many Microprocessor Development Systems. With asynchronous communication, the present UART supports a data size of 5, 6, 7 and 8 bits of data along with one start bit, one or two stop bits and parity bit as optional. Having a maximum data length support of 8 bits and minimum overheads there is considerable use of bandwidth. A rough estimation of bandwidth occupancy for data and overhead can be given as below.

PACKEAT = OVERHEAD/ + DATA.

= (8 + 2) bits.

This conveys that for every one byte there is ¼ bytes of additional overhead associated. For large amount of data (in hundreds of KBs and MBs) this is very significant. If we want to transmit about 10 Kilo Byte of data between a host and target emulation unit, there is over head of 20 Kilo-Bits (=2.570 kilo bits).With this there is lot of wastage of bandwidth in available channel.

"In order to overcome above limitations of inefficiency in bandwidth wastage, we propose an "Extension" in existing standard of UART/RS232 and entitle it as XUART as a VLSI core".

II. UART FUNCTIONALITY

The Universal Asynchronous Receiver / Transmitter (UART) controller is the key component of the serial communications subsystem of a computer. The UART takes bytes of data and transmits the individual bits in a sequential fashion. At the destination, a second UART re-assembles the bits into complete bytes. Serial transmission is commonly used

with modems and for non-networked communication between computers, terminals and other devices. There are two primary forms of serial transmission: Synchronous and Asynchronous. Depending on the modes that are supported by the hardware, the name of the communication sub-system will usually include a if it supports Asynchronous communications and a S if it supports Synchronous communications. The UART consists of three main blocks [4]

- A serial transmit block.
- A serial receive block.
- A CPU Interface (I/F) block.

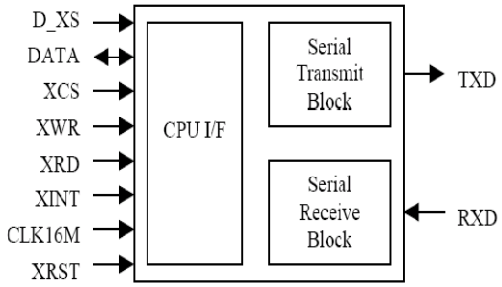


Fig. 1: Basic UART Block Diagram.

Pari indicates parity bit.

Fig. 2 : Serial Data Format.

As Shown in Figure 2 for transmission or reception of 8 bit from data communication equipment we require One START bit and One STOP bit along with Parity bit as optional. For sending large amount of data (in hundreds of KBs and MBs) this is very significant. If we want to transmit about 10 Kilo Byte of data between a host and target emulation unit, there is over head of 20 Kilo-Bits(=2.57 kilo bits). This can be explained as, If we have to transmit or receive 8 Kilo byte data then for transmitting 8000 byte data (8 Kilo byte =8000 Byte) we require….

8000 Byte * 8 Bits +
8000 Byte * 1 START BIT +
8000 Byte * 1 STOP BIT

Which is equals to = 64000 BITS + 16000 BITS

= 80,000 BITS

So for transmitting 8 Kilo Bytes we have to transmit 2 Kilo Bytes extra, total 10 KILO BYTES we have to transmit in existing UART. This conveys that for every byte there is ¼ bytes of additional overhead.

By using XUART we can transmit one start bit & one stop bit after the 16 bit data or 32 bit data, it can reduce the bandwidth wastage problem it is explained as follows: For transmitting 8 KILO BYTE DATA from 16 bit UART we require……

8000 Byte * 8 Bits +
4000 Byte * 1 START BIT +
4000 Byte * 1 STOP BIT

Which is equals to = 64000 BITS + 8000 BITS

= 72,000 BITS

So for transmitting 8 Kilo Byte Data in XUART we require only 1 Kilo Byte extra, half of the existing UART, it will reduce the bandwidth wastage problem. Same can be explained for 32 Bit Data……..

8000 Byte * 8 Bits +
2000 Byte * 1 START BIT +
2000 Byte * 1 STOP BIT

Which is equals to = 64000 BITS + 4000 BITS

= 68,000 BITS

For transmitting 8 Kilo Byte data in 32 bit XUART we require only 500 Bits extra it will reduce the bandwidth requirement at very great extent. It can save 1.5 Kilo Bytes.

III. IMPLEMENTATION

A. Software Implementation of FPGA Design:

The synthesis output files required by Xilinx ISE software are in EDIF (Electronic Design Interface File) and UCF (User Constraints File) formats, which represent the optimally synthesized netlist of integrated design, and timing constraints and FPGA pin assignment respectively to implement the synthesized design into Spartan-3E development board, Xilinx ISE performs the steps below:

1) Translate:

Convert netlist of integrated design in EDIF format to NGD (Native Generic Database) file that contains logical description of hierarchical components and Xilinx primitives by using NGD Build program.

*2) Map***:**

Perform logical DRC (Design Rule Check) on the NGD file, and then map the design logic to slices and I/O cells in Spartan-3E to create Native Circuit Description (NCD) file. The area constraints can be

sized properly using PACE (Pin-out Area Constraints Editor); then re-run Map.

3) Bit Generation:

Translate the fully routed NCD file to configuration bit-stream (.BIT) file using Bit Gen program.

4) Place and Route (PAR):

Place and route the design in mapped NCD file into Spartan-3E based on timing constraints using Timing Analysis tools. The output is fully routed NCD file.

5) Program Download:

Download the bit-stream file into Spartan-3E via JTAG cable using iMPACT program.

B. Hardware Implementation:

Extended Universal Asynchronous Receiver Transmitter is the serial communication interface. RxD is the received serial data signal and TxD is the transmitted data signal.

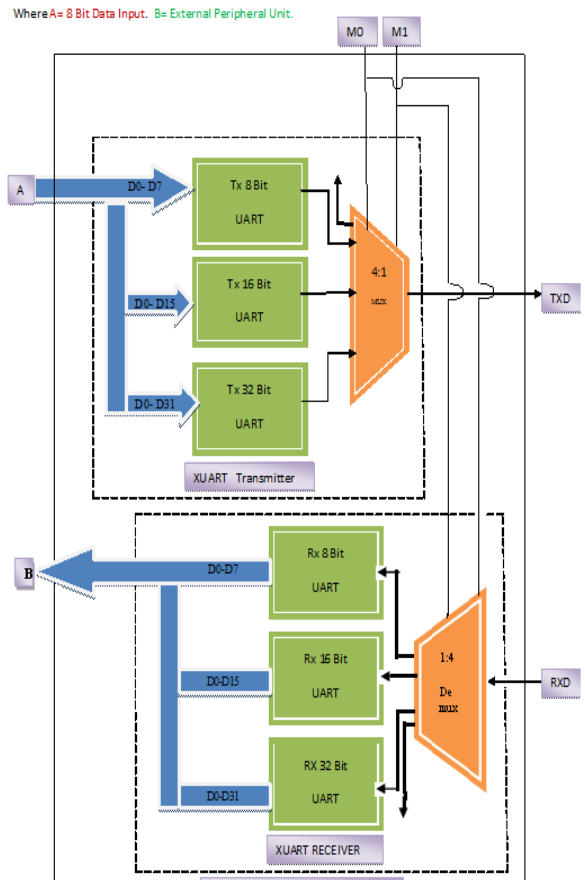


Fig. 3 : Architecture Of Extend UART.

The data is transmitted asynchronously, one byte at a time. When there is no data to transmit, Data D remains high. At the start of transmission, Data D goes low for one bit time, which is referred to as the start bit. Then 8/16/32 bit data bits are transmitted, LSB first. $8^{th}/16^{th}/32^{th}$ can be used as a parity bit. After 8/16/32 bits are transmitted, D must go high, for at least one bit time, which is called as the stop bit. By using counter as a clock divider the available baud rate in XUART is 9600Hz, 19200Hz and 115200Hz. As shown in architecture the XUART consist of two main components- XUART Transmitter & XUART Receiver.

1) XUART Transmitter:

As shown in architecture serial data transmission for XUART takes place on TxD pin. During transmission of data, the XUART takes 8 bit parallel data. The same data bus is connected to the 8 bit UART, 16 bit UART, and 32 bit UART. The first byte is connected to the 8 bit, 16 bit & 32 bit UART then 2^{nd} incoming byte is connected to the 16 bit UART & 32 bit UART only, similarly the 3^{rd} incoming byte is connected to the 32 bit UART. Thus 8,16,32 bit UART takes the 8,16,32 bits of parallel data respectively and converts the data into a serial bit stream which consists of the data format as a START bit (‘0’), 8/16/32 data bits (LSB first), and STOP bit (‘1’). All serial data coming from 8, 16, 32 bit UART is connected to the 4:1 MUX as shown in Figure 3 and two select lines are provided to the user. As per the user requirements the 8/16/32 bit UART is selected. If user wants to transmit 8 bit data frame then 8 bit UART is selected by this select lines. If 16 bit data frame has to transfer then 16 bit UART is selected and if 32 bit data frame has to transfer then 32 bit UART is selected by these select lines.

2) XUART Receiver:

The RxD pin receives 8/16/32 data frames serially. XUART receiver has 1:4 Demultiplexer, 8 bit UART, 16 bit UART and 32 bit UART receiver. The same select line is connected to XUART Mux and Demux. The 8/16/32 bit data frames are available at the Demux input line it can be connected to the 8 bit receiver UART, 16 bit receiver UART, 32 bit receiver UART according to the select lines. The XUART receiver converts 8/16/32 bit serial data frames into 8/16/32 bit parallel data. Then it can be connected to the peripheral unit.

For demonstrating our project we have use two FPGA boards for transmitting & receiving 16 & 32 bits as host PC supports only 8 bits.

C. Requirement Analysis:

TABLE I: MODE SELECTION.

MODE SELECT		
MO	**M1**	**UART SELECTED**
0	0	IDLE
0	1	MODE 8 BIT UART
1	0	MODE 16 BIT UART
1	1	MODE 32 BIT UART

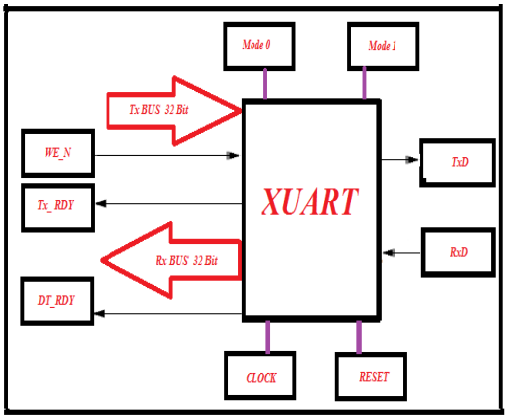


Fig. 4 : Specification Diagram Of UART

IV. EXPERIMENTAL RESULTS

A Universal Asynchronous Receiver/Transmitter is a type of "asynchronous receiver/transmitter", a piece of computer hardware that translates data between parallel and serial forms. UARTs are commonly used in conjunction with communication standards such as EIA, RS-232, RS-422 or RS-485. The universal designation indicates that the data format and transmission speeds are configurable, by using counters as a clock divider 9600Hz, 19200Hz, 115200Hz baud rates are available in extended UART. Transmitting and receiving UARTs must be set for the same bit speed, character length, parity, and stop bits for proper operation. Typical serial ports used with personal computers connected to modems use eight data bits, no parity, and one stop bit; for this configuration the number of ASCII characters per second equals the bit rate divided by 10. The existing UART's use one start bit and one stop bit for transmitting 8 bit data. When large amount of data has to be transferred then there is lot of wastage of bandwidth in available channel. For transmitting 8 Kilo Byte data we require 2 Kilo Byte extra total 10Kb has to be transfer. We transmit only one start bit and one stop bit with 8/16/32 bit data according to the user requirement. For transmitting 8KiloByte data through XUART we require only 500 Bits extra. This can be explained with the help of Figure 5 and Figure 6. For example if (A1B2C3D4)H, 32 bit data has to be transferred by XUART. First WE_N(tx_en signal in Figure 5 & 6) signal should be activated which means XUART Transmitter is ready to receive the 8/16/32 bit parallel data, then Tx_RDY is activated, now 8/16/32 bit parallel data is available on Tx bus 32 bit parallel (data_tx in Figure 5 & 6).The 8/16/32 bit transmitter UART converts this parallel data into 8/16/32 bit serial data frame respectively and which is available on TxD(_demo/txd signal in Figure 5 & 6) bus. On the receiver side Rx_Rdy signal is activated first which means receiver is ready to receive the data. The 8/16/32 bit serial bit stream data is available on RxD pin (-demo/rxd signal in Figure 5 & 6), the receiver UART converts 8/16/32 bit serial data stream into 32 bit parallel data.

The complete simulation window for 16bit XUART, 32bit XUART with synthesis report is shown in Figure 5 and Figure 6 respectively.

A. Simulation Result:

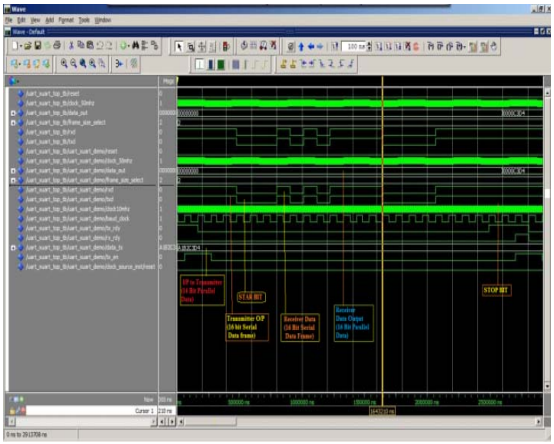


Fig. 5 : Simulation of 16 Bit UART.

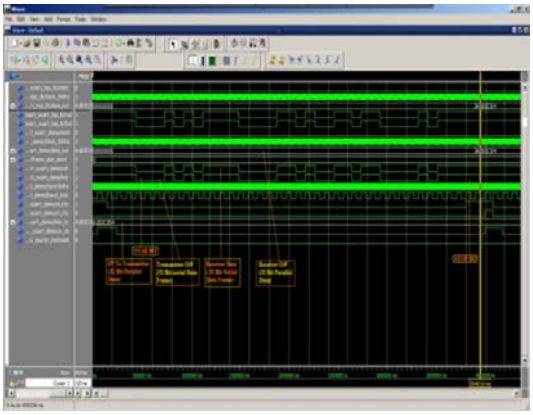


Fig. 6 : Simulation of 32 Bit UART.

TABLE II SYNTHESIS REPORT

Synthesis Report			
Target Device:	xc3s100e-4tq144	• **Warnings:**	52 Warnings

TABLE III DEVICE UTILIZATION SUMMARY.

Device Utilization Summary				[-]
Logic Utilization	***Used***	***Available***	***Utilization***	***Note(s)***
Number of Slice Flip Flops	231	1,920	12%	
Number of 4 input LUTs	485	1,920	25%	
Number of occupied Slices	358	960	37%	
Number of Slices containing only related logic	358	358	100%	
Number of Slices containing unrelated logic	0	358	0%	
Total Number of 4 input LUTs	541	1,920	28%	
Number used as logic	482			
Number used as a route-thru	56			
Number used as Shift registers	3			
Number of bonded IOBs	21	108	19%	
Number of BUFGMUXs	3	24	12%	
Average Fanout of Non-Clock Nets	4.08			

Minimum period: 7.100ns (Maximum Frequency: 140.845MHz).
Minimum input arrival time before clock: 13.382ns.
Maximum output required time after clock: 10.809ns.
Maximum combinational path delay: 9.480ns.

V. CONCLUSION

This work has presented a detailed guideline to design and implement "Overcoming limitations of good-old asynchronous serial communication by extending into its next generation". Many software and design tools have been used to verify the design output in terms of behaviour, functionality, synthesis, timing, and area constraints. The similarity of empirical real-time and simulated results shows the success of FPGA and Extended UART implementations. By this design we can successfully overcome the bandwidth limitation of existing UART by extending it up to 16 bit, 32 bit UART. Also provide standard baud rates of 9600Hz /19200Hz /115200Hz which is supported by hyper terminal. On the other hand, timing issues such as sample rate, constraints, and matching should be concerned in-depth if the input bit is coming from external source which is different from the presented design.

VI. ACKNOWLEDGMENT

Special thanks to Prof. Mrs. A. A. Shinde. (Bharati Vidyapeeth Deemed University College of Engineering Pune) giving precious guidance and information related VLSI, UART & Xilinx FPGA.

REFERENCES

[1] H. Dincer, "The design of high speed UART" Applied Electromagnetics, 2005. APACE 2005. Asia-Pacific Conference on 20 March 2006.

[2] H. Shouqian Yu Lili Yi Weihai Chen Zhaojin Wen, "Implementation of a Multi-channel UART Controller Based on FIFO Technique and FPGA" Industrial Electronics and Applications, 2007. ICIEA 2007. 2nd IEEE Conference on 24 September 2007.

[3] Idris, M.Y.I.,Yaacob, M., "A VHDL implementation of BIST technique in UART design" TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region, 15 March 2004.

[4] Dr. James P. Davis "Universal Asynchronous Receiver Transmitter" CSCE 612 – Project #2 Specification, v.04, 7 March 2003.

[5] Xilinx ISE 12.1 Software Manuals: Constraints Guide, and Development System Reference Guide, Xilinx Inc.,

[6] Spartan-3E Development Board User Guide.

◈◈◈

Edge Histogram for Sketch Based Image Retrieval Using Contourlet Edge Detection

Annie Julie Joseph & V.K.Govindan

Department of Computer Science and Engineering, National Institute of Technology, Calicut, India
E-mail: anniejulieshaji@gmail.com, vkg@nitc.ac.in

Abstract - This paper presents a novel approach for sketch based image retrieval (SBIR) using Contourlet edge detection. The performance of the proposed edge detection based approach is analyzed and compared with the existing approaches based on Edge Histogram Descriptor and improved Edge Histogram Descriptor. The experimental result demonstrates that the proposed approach is more efficient in sketch based image retrieval due to the inherent multi scale and flexible directional decomposition properties of Contourlets. The improved retrieval performance, compactness of feature vector, simplicity of implementation, and accurate feature extraction and improved matching time are the major advantages of this approach. Test results carried out on a database of 1240 images provide the precision rate as 0.8 and the recall rate as 0.5.

Keywords*-component; Sketch Based Image Retrieval (SBIR); Contourlet; Edge Histogram Descriptor (EHD); Modulus Maxima Detection; Feature Vector; Histogram Similarity.*

I. INTRODUCTION

Content Based Image Retrieval is the process of querying and retrieving images from a large collection of image database by specifying image's visual contents. If the image is unavailable for querying, user will draw main feature lines or rough sketch to retrieve matching images. The Sketch Based Image Retrieval (SBIR) deals with mapping of a user sketch to a similar image in the image database.

Edge detection is a fundamental problem to be solved in the case of SBIR. The efficiency of the edge detection algorithm significantly influences the performance of SBIR system. Recently wavelet or other advanced wavelet like Contourlet based edge detection approaches are proposed to acquire better performances. In this work, we developed a model to capture directional edges using Contourlet. The proposed approach extracted significant directional edges of the image and thus exhibits better image retrieval accuracy for a given sketch.

The remaining part of the document is organized as follows. Section 2 briefly presents the related work in the area of sketch based image retrieval. Section 3 describes an overview of Contourlet transform. Section 4 discusses about Modulus Maxima edge detection in Contourlet domain. Section 5 presents the proposed approach- the sketch based image retrieval using contourlet edge histogram. Section 6 outlines the conducted experiments and analyses the achieved results. Finally, section 7 presents the conclusion.

II. LITERATURE SURVEY

In this section, we review the existing techniques to solve the SBIR problem. SBIR problem is first introduced in Query by Visual Example (QVE) [1]. The novelty of this work is due to the transformation of images and sketches to an edge representation. The similarity of edges in sketch and images are then computed. QVE needs more computational power.

In order to reduce the computation expense, [2] presents a similarity matching based on signature file. As there is variations in details of the user sketch, edge signature file is created under multiple scales in this work.

The application of Edge Histogram Descriptor (EHD) was first proposed in the visual standard MPEG 7[3]. The edge information derived from each of the 16 non overlapping blocks of an image is categorized in to horizontal, vertical, diagonal and anti-diagonal and non-directional histogram bins .A total of 80 edge histogram bins are used in this method.

An improvement of EHD was proposed in [4]. Since [3] deals with only local distribution of edges. The performance of image retrieval [3] can be improved by deriving semi global and global edge histogram from the local edge histogram. This uses 150 histogram bins. The

edge histogram approach in [3, 4] is scale and translation invariant but not rotation invariant.

Another edge histogram based approach which is made rotation invariant is proposed in [5]. This Histogram of Edge Local Orientations (HELO) use polar coordinates and its application in sketch based retrieval is also specified. The limitation of the approach is that the query sketches need to be drawn in continuous strokes.

The major disadvantage of any histogram based image retrieval [3, 4 and 5] approaches is that it is based on global properties of the image and not on topological properties like shape.

Recently, in [6] a new approach based on oriented gradients computed on each edge point is presented. They provided a benchmark for evaluating the performance of large scale sketch-based image retrieval systems. The retrieval rate of this descriptor is dependent on database content and it is not rotation invariant.

Edge detection accuracy leads to better SBIR retrieval rate. In [7], Ren demonstrated that multi scale improves boundary detection. In this, local boundary cues including contrast, localization and relative contrast to train a classifier to integrate them across scales is used. Multi scale detection varies the scale of window and combines signals from multiple scales. In this method, probability of boundary operator (Pb) is used for contour detection. They computed Pb operator at 6 different scales. At small scale, precision is lower but more details are recovered. At large scale, precision is higher in the beginning showing salient boundaries being more detected. Multi scale approach combines the strengths of both small and large scale.

In [8], a high performance contour detector using global probability of boundary is used. It is an improvisation of the technique in [7]. The multiple cues are combined to a globalization framework based on spatial clustering. This globalized Pb detector uses a spectral Pb detector to improve the precision of the local multi scale Pb signal.

Numerous techniques for sketch based image retrieval have been proposed in literature. We found some application of wavelet and Contourlet in Content Based Image Retrieval (CBIR) [9, 10]. Use of Walshlet in CBIR can be found in [16]. But the application of wavelet or Contourlet for solving SBIR has not been found. Contourlet transform provides flexible multi resolution and directional decomposition of image. Mallat and Zhong [13] proposed the wavelet based Modulus Maxima edge detection and proved the method is efficient in edge determination. Ma et al. [14] extended this edge detection method in Contourlet domain.

In this paper we propose a new and efficient edge histogram for sketch based image retrieval using Contourlet edge detection. In this method, we have used Contourlet transform because of its efficiency in identifying edges in different directions..

III. CONTOURLET TRANSFORM

Contourlet transform was first proposed by Do and Vetterli [11] as an extension of wavelets with directionality property. They introduced a double filter bank structure by combining the Laplacian pyramid with a directional filter bank. The main feature of Contourlet transform is that it allows flexible number of directions at each scale.

The Contourlet transform block diagram is given in Figure 1.

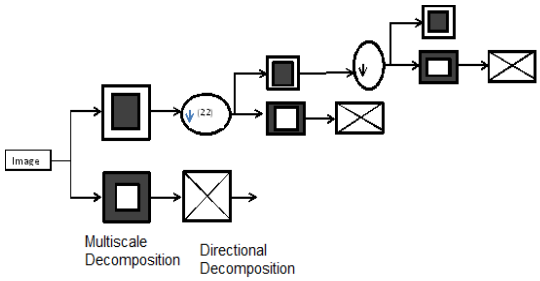


Figure 1.The Contourlet Transform Block diagram. It uses an iterated combination of the Laplacian pyramid and the directional filter bank

The Contourlet transform of an image can be obtained by applying two successive decomposition states. First step is the multiscale decomposition in to a Laplacian Pyramid (LP). The second step is decomposition of each LP scale level into directional bands by a Directional Filter Bank(DFB).

In Contourlet, at each scale level by LP decomposition an image is decomposed into one low frequency sub band and another high frequency sub band. DFB is employed to capture the directional information from the high frequency sub band.

The significant Contourlet coefficient values represent edges in the direction. By chaining the significant coefficients at each level of Contourlet decomposition, we can get a directional edge representation of the image.

IV. EDGE DETECTION USING CONTOURLET TRANSFORM MODULUS MAXIMA EDGE DETECTION

Mallat and Zhong [13] proposed an edge detection method using Wavelet. Its principle is based on finding local maxima of wavelet coefficients that will represent edges in the image. This method of edge detection extended in Contourlet domain is proposed in [14] is used in our sketch based image retrieval system.

Each directional decomposition sub band in Contourlet domain can be considered as equivalent gradient angle direction. Each coefficient's modulus value in the directional sub band is compared with the modulus of the adjacent point. Consider the example where modulus maxima in horizontal direction.

$$M=\begin{cases}1 & \text{If } MOD[C_{s,k}(x,y)]>MOD[C_{s,k}(x-1,y)] \text{ and } MOD[C_{s,k}(x,y)]>MOD[C_{s,k}(x+1,y)] \\ 0 & \text{Otherwise}\end{cases} \quad (1)$$

Cs, k(x, y) in the above representation is the Contourlet coefficient in location(x, y) at scale s and direction k.

M=1 means it is a local maximum point in the horizontal direction. If modulus maximum is 1 for multiple scales it means there is an edge in that direction. The significance of edge can be determined by the span of modulus maxima value 1 in multiple scales.

The modulus maxima of all directions are determined in the same way and edges are determined by the presence of modulus maxima value 1 in corresponding location in multiple scales.

V. SBIR ARCHITECTURE

The two main steps in building a SBIR system are feature extraction and similarity matching. In our proposed system, the histogram of the edges of each direction is used as the feature vector. The edge histogram of the sketch and image is compared using histogram similarity measure. The proposed SBIR architecture is given in Figure 2.

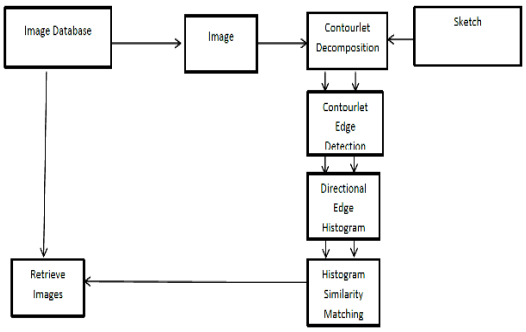


Figure 2. Proposed SBIR System Architecture

VI. EDGE HISTOGRAM FOR SKETCH BASED IMAGE RETRIEVAL USING CONTOURLET EDGE DETECTION

A. *FEATURE EXTRACTION*

The following steps are involved in the generation of feature vectors:

1. Decompose each image in the database by a 4 level Contourlet transform. We get directional decomposition of image at multiple scales.
2. Modulus Maxima Detection: Set the modulus maxima of all directional sub bands at each scale to 1 and set non maxima coefficients to 0. The resultant matrix is called as Modulus Maxima Matrix for directional sub bands at each scale.
3. Edge Detection: For each directional sub band if value 1 exists in the Modulus Maxima Matrix of multiple scales in the corresponding location identify it as an edge.
4. Count the number of edges in each directional sub band and generate a directional edge histogram ordered by edge significance.

B. *HISTOGRAM SIMILARITY*

Once the directional edge histogram of image and sketch is ready, the next step is to find a good histogram similarity measure. In paper[17], the following similarity measure is proposed.

$$S(H_1,H_2) = \frac{1}{1+\sum_{i=1}^{n} |h1(i)-h2(i)| \Big/ (1+h1(i)-h2(i))} \quad (2)$$

This similarity measure is used in our method for comparing the directional edge histogram of images and sketches.

Steps involved in Histogram Similarity Matching are as follows:

1. Compute the similarity measure of corresponding directional edge histogram of sketch and images in the database using S(H1,H2) in equation 2.
2. Retrieve all images based on minimum histogram similarity distance measure.
3. The most common evaluation measures, precision and recall are compared.

$$\text{Precision} = \frac{No.of\ relevant\ images\ retrieved}{Total\ No.images\ retrieved} \quad (3)$$

$$\text{Recall} = \frac{No.of\ relevant\ images\ retrieved}{Total\ No.of\ relevant\ images\ in\ the\ database} \quad (4)$$

VII.EXPERIMENTAL RESULTS

We tested the performance of the directional edge histogram using Contourlet using the benchmark database containing sketches and classes of images for SBIR system given with [13]. The database contains 31 sketches and each sketch contains 40 associated images yielding a total of 1240 images. Each image of size 256x256 in the database is decomposed by 4 level Contourlet transform. The decomposition parameters used is same for simplicity of implementation. The decomposition parameter used is (3, 3, 3, 3) and (4, 4, 4, 4); that is, eight and sixteen directional sub bands decomposition corresponding to each scale level. The filter used for Laplacian pyramid and directional filter bank is '9-7' filter. In the experimental results, the sketch is displayed first and retrieved images based on similarity measure goes from left to right as shown in Figure 3.

Figure 3. Results of sketch based retrievals. Shown on the left most of each row is the sketch, and the rest are the retrieved images arranged in the descending order of matches.

The retrieval rates obtained based on average precision and recall rates for a collection of sketches in the database for Edge Histogram Descriptor, Improved Edge Histogram Descriptor and Multi Directional Edge Histogram Using Contourlet are shown in Table 2 for comparative performance study. This result is depicted graphically as shown in Figure 4.

TABLE 1 : AVERAGE PRECISION AND RECALL RATES

Type of Histogram	Feature Vector Size	Precision Rate	Recall Rate
Edge Histogram Descriptor[3]	80	0.5	0.3272
Improved Edge Histogram Descriptor[4]	150	0.6	0.373
Contourlet Statistical Parameter (Kurtosis& Absolute Mean Energy)[17]	60	0.65	0.45
8 Directional Edge Histogram Using Contourlet	128	0.7	0.48
16 Directional Edge Histogram Using Contourlet	256	0.8	0.52

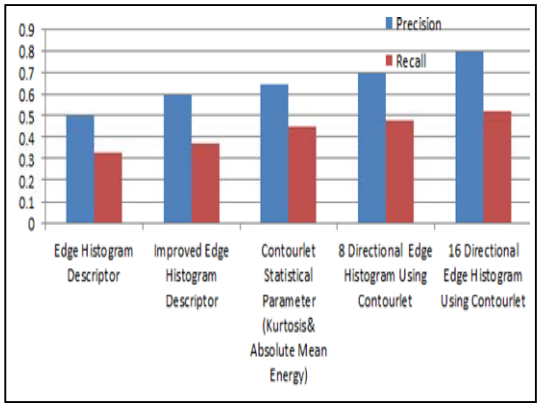


Figure 4. Average Precision and Recall Rate

We find that the retrieval results of directional edge histogram using Contourlet is superior to all other approaches. The reason is that Contourlet features are efficient at capturing directional information and smooth contours. The flexibility to specify any two's power of directions and multi-level scales in Contourlet enable us to identify more significant directional edges. The performance accuracy is higher in the case of sketches and images having sharp edges.

VIII. CONCLUSION

Content Based Image Retrieval is the process of querying and retrieving images from a large collection of image database by specifying the visual content of the

image. The Sketch Based Image Retrieval deals with mapping of a user sketch to a similar image in the image database. In this paper, we have proposed a Contoutlet based directional edge histogram approach for sketch based image retrieval. The effectiveness of Contourlet transform for capturing the directional information is demonstrated. The performance of the proposed approach is compared with three other approaches in the literature. It is observed that SBIR using Contoutlet directional edge histogram is superior to all the other approaches. Also, the result of 16 directional edge histogram is better than 8 directional edge histogram cases. By using Contourlet, multiple scale and directional decomposition are easily possible due to its inbuilt directional filter design properties. Hence, the Contourlet edge detection is efficient in significant directional edge detection, and thus preferable for sketch based image retrieval systems.

REFERENCES

[1] Kato T., Kurita T., Otsu N.andHirata K."A sketch retrieval method for full color image-Query by Visual Example," In Proceedings of the 11th IAPR International Conf. on Computer Vision and Applications, Conf. A: Pattern Recognition, pages 530–533, April 1992.

[2] Raj Kumar Rajendran and Shih-Fu Chang."Image retrievalwith sketches and compositions," In: IEEE International Conference on Multimedia and Expo, Vol. 2, pages717–720, May 2000.

[3] T. Sikora."The MPEG-7 visual standard for content description-an Overview," IEEE Transactions on Circuits and Systems for Video Technology 11 (6), pages 696–702, May 2001.

[4] Chee Sun Won, Dong Kwon Park and Soo Jun Park. "Efficient use of MPEG-7 edge histogram descriptor," In Etri Journal 24 (1), pages23–30, May 2002.

[5] Jose M Saavedra and Benjamin Bustos."An improved histogram of edge local orientations for sketch based image retrieval," In Proceedings of the 32nd conference on pattern recognition pages 432-441, 2010.

[6] Eitz M., Hildebrand K., Boubekeur, T. andAlexa M. "Sketch Based Image Retrieval: Benchmark and Bag of Features Descriptors," In IEEE Transactions on Visualization and Computer Graphics, Volume 17, No.11 November 2011.

[7] XiaofengRen "Multi-Scale Improves Boundary Detection in Natural Images", InEuropean Conference on Computer Vision 2008

[8] Pablo Arbelaez, Member, Michael Maire, CharlesFowlkes and Jitendra Malik .Contour Detection and Hierarchial Image Segmentation" EECS Department, University of California, Berkeley 2010.

[9] Ch.SrinivasaRao, S. Srinivas Kumar andB.N.Chatterji. "Content Based Image Retrieval System Using Contourlet Transform," ICGST-GVIP Journal, Volume 7, 2007.

[10] F.Pulido, O.Starostenko, D Flores Quechol, J.I Rodrigues Flores, IngripKirrschning and J.A.Chavez Aragon "Content Based Image Retrieval Using Wavelets," 2nd WSEAs International Conference on Computer Engineering and Applications Acapulco, Mexico, pages 25-27, January 2008.

[11] Min N Do. and Martin Vetterli. "Contourlets" In the Text "Beyond Wavelets" Academic Press 2001

[12] Duncan D., Y. Po and MinN.Do. "Directional Multiscale Modeling of Images Using the Contourlet Transform", IEEE transactions on image processing Vol.15,No.6 June 2006

[13] S. Mallat, S. Zhong, "Characterization of signals from multi scaleedges", IEEE Trans. Pattern Anal. Machine Intell. 1992, pp. 710–732.

[14] Shun-feng Ma, Geng-fengZheng, Long-xu Jin, Shuang-li Han, Ran-feng Zhang "Directional Multiscale Edge Detection Using the Contourlet Transform" In IEEE 2010.

[15] V. Megalooikonomou, J. Zhang, D. Kontos, P.R. Bakic,"Analysis of texture patterns in medical imageswith an application to breast imaging", Proceedings of theSPIEConference on Medical Imaging, 2007

[16] Rakhee M and V K Govindan, "A Walshlet Based Approach for Color and Texture Extraction in CBIR", Proceedings on International Conference on Futuristic Trends in Computer Science and Engineering & Information Technology (ICCT2012), 28th and 29th January 2012, pp 87-91, 2012

[17] Annie Julie Joseph and V.K Govidan "Sketch Based Image Retrieval Using Contourlet Features," In the Proceedings of First International Conference on Futuristic Trends in Computer Science and Engineering & Information Technology Vol:2 Pages:301-305,January 2012.

◈◈◈

2-D Implementation of Digital Gabor Filter Design Using VERILOG

Sushanth K J & N.Shankaraiah

Dept. of Electronics and Communication Sri Jayachamarajendra College of Engineering, Mysore, INDIA
E-mail : Sushanthkj@gmail.com & shankarsjce@gmail.com

Abstract **-** Fingerprint or Face Image enhancement using Gabor filter is one of highly computational complexity in fingerprint verification process. Gabor filter has a complex valued convolution kernel and a data format with complex values is used. So implementing Gabor filter is very significant in fingerprint verification process. Designing Gabor filter will help enhancing the quality of fingerprint image. In fingerprint recognition, Gabor filter optimally capture both local orientation and frequency information from a fingerprint image. By tuning a Gabor filter to specific frequency and direction, the local frequency and orientation information can be obtained. Thus, it is suited for extracting texture information from images. This paper presents the implementation of 2-D Gabor Filter design using Verilog HDL. This paper details important enhancement made to the 2D -Digital Gabor filter to minimize the sizing problem and the coding style that synthesizable. The intention is to study, analyze, simplify and improvise the design synthesis efficiency and accuracy while maintaining the same functionality. The result provides area efficiency architecture for the effective design.

Keywords **-** Digital filter, digital design, face recognition, Gabor filter, MAC, verilog HDL, Xilinx

I. INTRODUCTION

The Gabor filter (Gabor Wavelet) represents a band-pass linear filter whose impulse response is defined by a harmonic function multiplied by a Gaussian function. Thus, a bidimensional Gabor filter constitutes a complex sinusoidal plane of particular frequency and orientation modulated by a Gaussian envelope. It achieves an optimal resolution in both spatial and frequency domains.

Even though the design might compromise the speed, but the area consumption was reduced. The speed of serial design can be overcome by operate at a higher frequency.

In this article we are represent the 2-D digital gabor filter in MAT LAB, and design and implementation in verilog using Xilinx 10.1

1. Representation of 2-D gabor filter in MAT LAB.

A. Digital Gabor Filter

Gabor Filter was designed by transforming the designinto verilog using *xilinx 10.1*. The target device is Spartan 3Afamily. The figure shown below is the summary of the synthesized design. It can be seen that the utilization of the resource of the device exceeded 100%. This particular point was where the improvement needed to be done to achieve an effective and efficient design.

- **Texture analysis :** Texture is a fundamental property of natural images, and thus it is of much interest in the computer vision and computer graphics. Regular repetition of an elements or pattern of surface.

- **Frequency domain:** Gabor filter can be represented in frequency domain by using, the 2-D Gaussian functions.

Formally, a 2D Gabor filter in the spatial domain is defined by the following expression:

$$\psi_{u,v}(x,y) = \frac{f_u^2}{\pi\gamma\eta} e^{-\left(\frac{f^2}{\gamma^2}x'^2+\frac{f^2}{\eta^2}y'^2\right)} e^{j2\pi f_u x'},$$

where x′ = x cos Θv +y sin Θv, y′ = Θx sin Θv +ycosΘv, and the parameters fu and Θv are defined as fu = fmax/2(u/2) and Θv = vΘ/8. As we can see, Gabor filters represent Gaussian kernel functions modulated by a complex plane wave whose center frequency and orientation are defined by fu and Θv, respectively. The parameters and Θ(theta) determine the ratio between the center frequency and the size of the Gaussian envelope.

Mainly 2-D gabor filter designed by using two aspects

- Filters in frequrncy domain:

- Filters in space domain:

Fig 1: Design summary

Basically there were 3 major parts in the filter: CLU, ALU and MEMORY [4]. The 'convolution' signal indicates the operation of the filter. If the signal is high then the convolution process takes place. If it is low then the filter receives image input and stores it to the memory based on the input location. The data enters the filter pixel by pixel. The 'PIXEL_X' and 'PIXEL_Y' signal gave the address of the memory location.

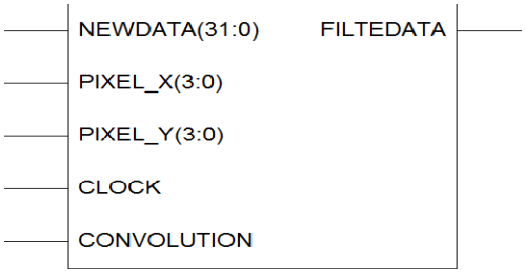


Fig 2: Top level

B. Arithmetic Unit

The main part of the digital gabor filter that is doing the convolution process. This is where the Gabor coefficient is stored[4]. It consists of 3 parts: ROM, DECODER and MAC. The ROM has 16 address locations but only 9 of it are used to store the coefficient. The MAC divided into 2 parts: multiplier and adder. The multiplier has 9 parallel multipliers. So the multiplication will be done in the same time as to speed up the convolution process. The adder consists of 8 adder connected in sequence. The adder is to sum up all the 9 multiplier outputs. Since the design uses 9 parallel multipliers and 8 adders, the design is significantly large. Both multiplier and adder use Xilinx IP cogen floating_point V3.0. This IP Cogen is generated from the Xilinx library.

II. METHODOLOGY

The focus of this work is not to design a new digital Gabor filter but to improve the design so it can be implemented on the device. As an ASIC designer, there are three major factors needed to be considered, maximization of speed, minimization of area and power consumption. In this work, minimization of area consumption will be the main priority.

A. Design

The design of the new multiplication-accumulation unit must be done precisely[6]. This is due to the sensitivity of the transition in a single data path. Below is the design flow of the filter.

Firstly, when the convolution signal is '0' the input data which is in pixel format will enter the filter and stored in the memory. The size of the memory depends on the pixel size. If the pixel is 16x16 then the memory size will be 32x32 too. It means that every memory location will be stored for value for 1 image pixel[4][10].

When the convolution signal is triggered to '1' the convolution process starts. The controller will read the image that is stored in the memory and send the data to the arithmetic unit. The controller will call the data from the determined memory location. In arithmetic unit there is also a ROM which will permanently store the coefficient kernel value. The value of kernel will also be called by the control into the convolution circuit. When both data has entered the convolution circuit the process of multiplication and accumulation will take place.

Only one series of data will be convoluted at a time. The counter will count for 9 convolution operation before giving out the result of filtered image.

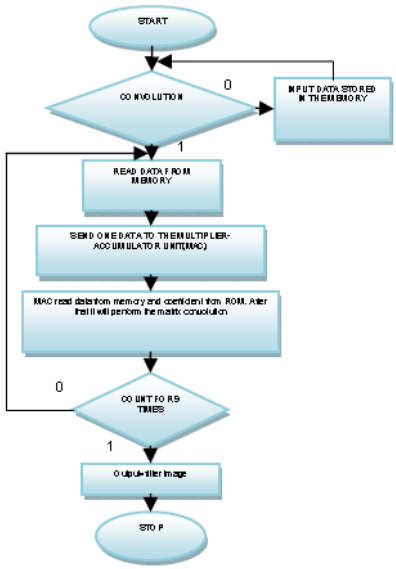


Fig 3 : Gabor filter flow

The reason why count for 9 consecutive cycle stands for the 9 coefficient kernel value. This will also be the result of the filter.

III. RESULT AND DISCUSSION

MAT LAB RESULTS:

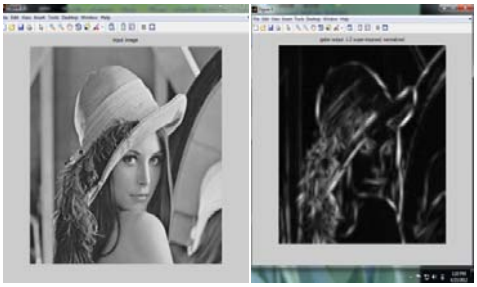


Fig 3: Test image Fig 4: Gabor image

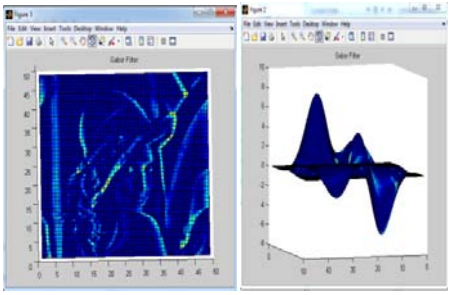


Fig 5: Gabor filter with 3-D view

The above MAT LAB results are just design and representation of digital gabor filter.

1. Design and implementation of 2-D gabor filter in verilog.

Redesigning the gabor filter in verilog using Xilinx 10.1 software, the code was then synthesized. The summary of the design was shown in figure 6. From the summary, the numbers of warnings were reduced from 92 to 21 warnings. The warnings generated are related to the incomplete if and else statement which a latch might be generated. In this summary, the target device Spartan3-S200 was used. This device contains large resources suitable for a design such as this. The numbers of Slices, Slice Flip Flops and LUTs were reduced.

RESULTS IN VERILOG:

- Below table values and waveform window from the Xilinx are the expected results when implemented in verilog

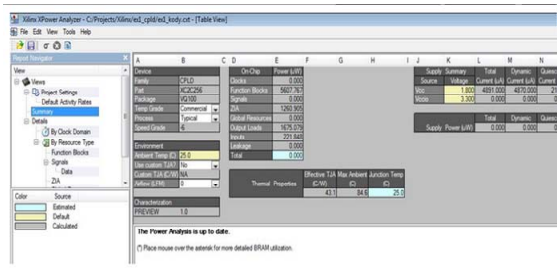


FIG 6: Design Summary

A. *Top level*

The above figure 7 shows the schematic view of the top level filter. There were 6 input pins and one output pin on the top level. New data stands for an unfiltered 32-bits image data. Pixel-X and Y hold the position of the memory when the write memory occurred. Clock and reset pins indicates the generated clock with 40ns period and reset button for the filter. The 'convolution' signal is to indicate the operation of the filter. If the signal is high then the convolution process takes place. If it is low then the filter receives image input and stores it to the memory based on the input location.

Fig 7: Toplevel

Output result for the top level filter is real convoluted data of filter is 0.006764772(3BDDAB06) but the expected result in figure 8 was 0.006764705(3BDDAA75). The difference was 0.00000068. The error was only 0.001%. This new design verifies that even though the multiplication and accumulation were design in serial, it can still give and maintain the same result from previous parallel design. It took 222 cycles to finish the convolution process in serial design.

IMAGE	COEFF	IMAGE*COEFF		SUM OF IMAGE*COEFF	
1	0.006737943	0.006737943	3BDCC9F6		
2	1.29E-05	2.57E-05	37D7E06A	0.006763677	3BDDA1D5
3	-4.00E-08	-1.21E-07	B402751D	0.006763556	3BDDA0D1
4	2.36E-07	9.43E-07	357D0CF5	0.006764499	3BDDA8BA
5	4.41E-08	2.07E-07	345E42E6	0.006764706	3BDDAA77
6	1.45E-12	8.69E-12	2D18D7D0	0.006764706	3BDDAA77
7	-1.36E-11	-9.51E-11	AED12154	0.006764705	3BDDAA75
8	2.65E-14	2.12E-13	2A6EE220	0.006764705	3BDDAA75
9	8.53E-17	7.68E-16	265D5A7A	0.006764705	3BDDAA75
		0.006764705	3BDDAA75		

Fig 8: Real convolution data

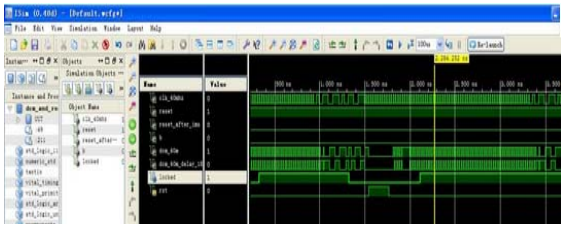


Fig 9: Verification of top level filter

B. Controller (CLU)

The control logic unit functions as controller for the data flow in the filter. It gives instruction to the other blocks to do their job. Basically, it gives the memory address to read data to the MEMORY and give address of coefficient to the ALU.

This CLU will only generate the address location when the 'START' signal is high. This signal indicats the convolution process that has taken place but if the signal is low, it indicates that the writing of image data into the memory takes place.

This CLU contains only 2 different blocks. One is the counter for the coefficient and memory address, and the other one is the counter decoder. The design of the counter gives the relationship between the coefficient and memory address. When the coefficient address was counted up until 9, the memory address for Y- direction will count a plus one. And the X-direction address must wait until Y-direction counts until 16 then it counts a plus one.

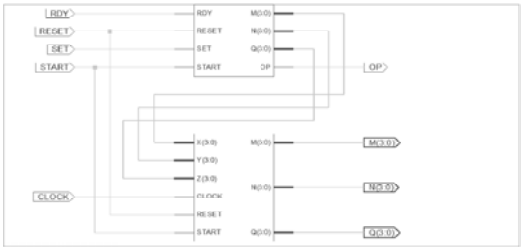


Fig 10: Controller

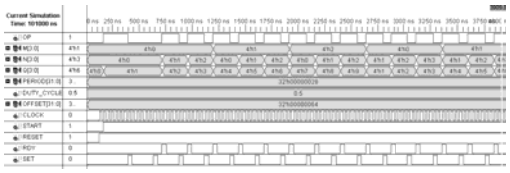


Fig 9: CLU Verification

C. Memory

The memory block is used to store the image pixel. The decoder only decodes address for Y-direction only. The clock was removed from the decoder so the decoded Y-direction can arrived at the same clock cycle with the X-direction. The adress for X-direction is supplied directly from the CLU or from the filter input. The image input is also connected directly from the filter input. The writenable signal indicates whether the operation is a write data or read data.

From the figure 10, first the 'WRITENABLE' signal is high to indicate the writing process is taking place. Then the signal goes low to read the data in the memory. The memory will give the output on the same clock cycle as the address location enters.

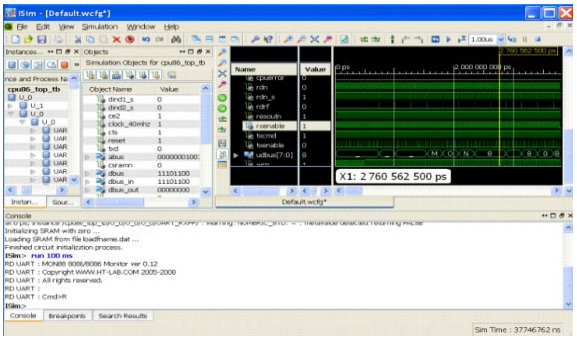


Fig 11: Verification memory unit

D. Arithmetic (ALU)

This is the main part of this work, arithmetic unit. This is where the convolution process takes place. It consists of two parts: the ROM and the MAC. The ROM is used to store the 9 coefficient values that are needed to convolute with the image while MAC consists of a buffer, a multiplier, an adder and a counter. The crucial part of this design was to make sure that the convolution process happened align with the correct image data and coefficient. The 'CONVO' signal plays important role to ensure there was no mismatch of data read.

From the figure 11 below, the 'CLOCK' and the 'CONVO' both were connected to the ROM and MAC. When the 'CONVO' went from low to high, the convolution process starts. The feedback 'READY' and 'SET' were sent to the CLU indicates convolution process completed. The CLU then will push the 'CONVO' signal from high to low before the next convolution takes place. These processes take 9 complete convolutions before sending the convoluted data out.

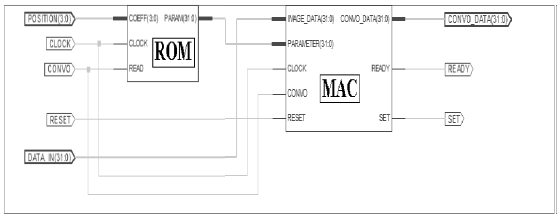


Fig 12: Arithmetic

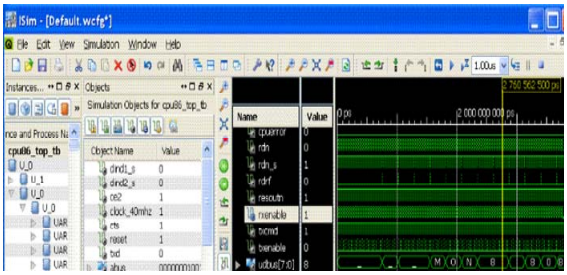


Fig 13: Verification of ALU

The buffer was used to hold the 'CONVO' operation for 1 cycle before the multiplier. The intention was to wait for the correct data sent from the memory for the convolution process. The Multiplier and the adder are connected in series. The design was done in such a way to lessen the area consumption of the filter. After 9 consecutive multiplications and additions, Since the design is a single data path also known as pipeline, the multiplication and the addition will take a longer period of time. The total cycles required for convolution for this design is 222 clock cycles with a time period of 40ns per cycle.

IV. CONCLUSION

The design and implementation of digital Gabor Filter has successfully reached. The area of the design has been significantly reduced while the function of the filter is perfectly maintained.

The numbers of slices used from previous design reduce from 5759 slices to 1625 slices. This significant change is due to the reduction of multiplier and adder used in the multiplication and accumulation unit. The enhancement made in the multiplication-accumulation unit has been proven effectively reliable and functional.

By adjusting the memory and the controller unit, the functionality of a complete and correct digital Gabor Filter is obtained. Even though, the precision of this Gabor Filter is 0.001% away from the calculated data. By minimizing the area, The result provides area efficiency architecture for the effective design.

REFERENCES

[1] A.D. Jepson and D.J. Fleet, "Image Segmentation" , Internet Draft 2007.

[2] Dennis Dunn, William E. Higgins, and Joseph Wakeley, Texture Segmentation Using 2-D Gabor Elementary Functions, IEEE TRANSACTIONS ON PAITERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 16. NO. 2, FEBRUARY 1994 : pp. 130-149.

[3] Razak, A.H.A. Taharim, R.H. "Implementing Gabor Filter for Fingerprint Recognition using verilog HDL," IEEE explorer , March 2009.

[4] P. H. W. L. Ocean Y. H. Cheung, Eric K.C. Tsang,Bertam E.SHi, "Implementing Of Gabor-type Filters on Field Programmable Gate Arrays," 2005.

[5] Heikkilä M, Pietikäinen M & Schmid C (2009) Description of interest regions with local binary patterns. Pattern Recognition 42(3): 425.436.

[6] Heisele B, Ho P, Wu J & Poggio T (2008) Face recognition: component-based versus global approaches. Compter Vision and Image Understanding 91(1.2): 6.21.

◈◈◈

Design of CGI in Embedded Dynamic Web Technology

Manisha P. Kodag[1] & M. M. Jadhav[2]

[1&2]Sinhgad College of Engineering, Pune-41
[1]Modern College of Engineering, Pune-05
E-mail : mpkodag@gmail.com, Makj123@yahoo.com

Abstract - In this paper a new idea of dynamic Web server based on LM3S6965 ARM processor is given, which is used to manage and control the embedded devices by browser connected to the Web server. With the development of embedded system, the combination of embedded technology and network technology will enhance the intelligence and flexibility of embedded devices greatly. Web browser become a standard interface to manage embedded devices by accessing to embedded Web server after embedded device connecting to the Internet. Traditional Web server simply provides a fixed page. The modern Web server can update pages in time, or create pages in accordance with each requirement. Dynamic Web can be realized by Common Gateway Interface (CGI). CGI is choosed due to the small space of embedded system. CGI is a standard of connecting external application to server. It is a software module in Web server, and it can execute the script stored in the server. The embedded system is resource restricted; while any of the TCP, IP or HTTP protocol is too large to be implemented. The embedded system does not have enough storage resource to implement a completed protocol stack. So Lightweight IP (LwIP) is used. LwIP stack is an open-source implementation of the TCP/IP stack developed specifically to reduce resource usage while maintaining a full- scale TCP/IP stack. For embedded systems, LwIP makes it possible to connect the system to a local intranet or the Internet. This paper describes the hardware and software constructs and role of CGI in embedded dynamic web Technology.

Keywords - *embedded Web server, Common Gateway Interface (CGI), dynamic Web server, Lightweight IP(LwIP).*

I. INTRODUCTION

Use of Internet has been expanding exponentially; it is now extensively used as both connectivity and reference tool for commercial, personal and educational purposes. The Internet revolves around the client-server architecture. Statically composed HTML pages lack the flexibility to accommodate fast-paced content changes and the feedback capability essential to support on-line interaction. As a result, dynamically created HTML pages, which are results of program execution, emerge as the mechanism of choice in commercial Web sites.

The rest of this paper is organized as follows. Section 2 reviews previous client-server architecture and solutions to this architecture. Section 3 introduces the role of CGI. Section 4 describes the need of LwIP & its stack. Section 5 describes the design of Embedded Dynamic Web. Section 6 describes application design based on CGI .Section 7 concludes this paper.

II. CLIENT-SERVER ARCHITECTURE:

Your computer runs software called the client and it interacts with software known as server located at remote computer. The client is usually a browser such as internet explorer. Browser interacts with server using a set of instructions called protocols. The client requests for an HTML file stored on server. The server locates this file and passes it to the client as shown in Fig.1.

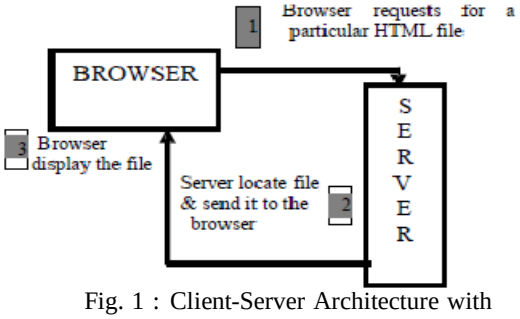


Fig. 1 : Client-Server Architecture with static HTML page

In this case, the HTML page is static. Static page do not change until the developer modifies them. This drawback is removed by using CGI program, which is executed in real time so that it can output dynamic information. This idea is given in Fig.2

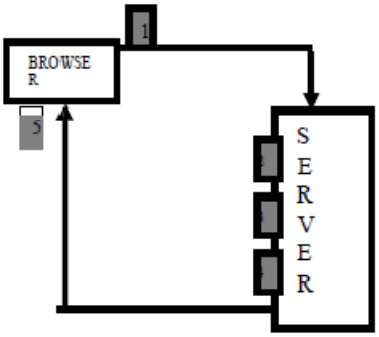


Fig.2. Client-server architecture with dynamic HTML page

1. Browser send request to the server
2. Server locates CGI program and passes the request information
3. CGI program processes the equest & sends data to the server
4. Server sends data to the browser
5. Browser display the data

A CGI program is stored on web server and executed on web server in response to a request from a user [9]. Thus CGI program generates a dynamic HTML page.

III. ROLE OF COMMON GATEWAY INTERFACE (CGI):

CGI is program used for communication of client with server. It is a standard service invocation mechanism that web server support either to provide "dynamic content", HTML pages that created dynamically to respond to user queries/requests [1]. Working of CGI is shown in Fig.3.

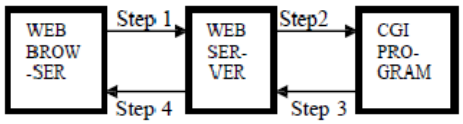
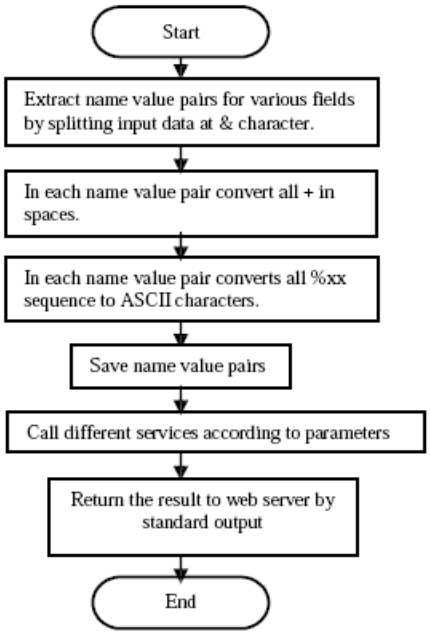


Fig. 3 The realization of CGI

The server (step 1), waiting for a request from the client (web browser). When a request comes, the server will create a sub-process for connecting services of the user. Web server then starts the CGI program and delivers the received data in the form of parameters (step 2). The output of CGI program is guided to the web server (step 3). Then web server sends them all as response to the web browser (step 4).The script file is placed on the same machine on which web server runs and not on your local machine [2].

Flow-chart of CGI:



IV. NEED OF LwIP & Its Stack:

As we know in order to access a Web page, the HTTP client sends an HTTP request to the HTTP server. To send an HTTP request, the HTTP client program (the Web browser) must establish a TCP connection to the HTTP server. Here TCP/IP protocol stack comes in point. Some embedded system such as Embedded Linux, VxWorks and WinCE had already provided TCP/IP protocol stack support; they are not suitable for applying in most of the low cost area due to their relative high cost on hardware and software. So how to utilize the limited resource of the embedded system to design a reduced and application oriented protocol stack is the key technology of implementing an embedded web server [10]. There are significant differences of developing web server on embedded system from common machines. The embedded system is resource restricted; while any of the TCP, IP or HTTP protocol

is too large to be implemented. The embedded system does not have enough storage resource to implement a completed protocol stack. So we must evaluate the protocols in standard TCP/IP protocol stack carefully to decide which parts of the protocol are necessary in the stack and which parts of the protocol can be saved. The Lightweight IP (LwIP) stack is an open-source implementation of the TCP/IP stack developed specifically to reduce resource usage while maintaining a full-scale TCP/IP stack. For embedded systems, LwIP makes it possible to connect the system to a local intranet or the Internet[8].

The LwIP stack was originally developed by Adam Dunkels of the Networked Embedded Systems group at the Swedish Institute of Computer Science. It is now being actively developed by a world-wide team headed by Leon Woestenberg and is being used today in many commercial products. Written in the C programming language, LwIP is an open-source TCP/IP stack developed for embedded systems with a focus on reducing resource usage. LwIP can run with or without an underlying operating system. Typical code size is on the order of 25 to 40 kilobytes while RAM requirements are approximately 15 to a few tens of kilobytes. LwIP TCP/IP protocol suite is shown in Fig.4

LwIP features:

- IP (Internet Protocol) including packet forwarding over multiple network interfaces
- ICMP (Internet Control Message Protocol) for network maintenance and debugging
- UDP (User Datagram Protocol) including experimental UDP-lite extensions
- TCP (Transmission Control Protocol) with congestion control.
- DHCP (Dynamic Host Configuration Protocol)
- PPP (Point-to-Point Protocol)
- ARP (Address Resolution Protocol) for Ethernet

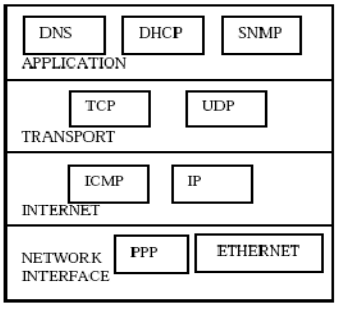


Fig. 4 LwIP TCP/IP protocol stack

V. THE DESIGN OF EMBEDDED DYNAMIC WEB:

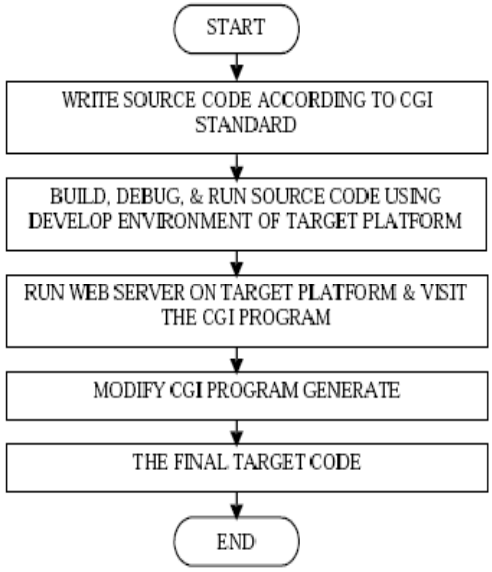
Embedded systems usually consist of embedded processor, embedded peripheral equipment, embedded operating system and application software etc [6]. This design uses LM3S6965 of Stellaris ARM7 chip as the processor, windows as OS.

5.1 Hardware:

TheStellaris LM3S6965 Evaluation Board is a compactand versatile evaluation platform for the Stellaris LM3S6965 ARM. The evaluation kit uses the LM3S6965 microcontroller's fully integrated 10/100 Ethernet controller to demonstrate an embedded web server [4]. 32-bit RISC performance using ARM architecture with 50 MHZ in operation, 256 KB Flash, 64 KB SRAM, four 10-bit channels (inputs) when used as single-ended inputs, two independent integrated analog comparators, two I2C modules, three PWM generator blocks, 0 to 42 GPIOs, depending on user configuration.

5.2 Software:

Embedded web server as a data carrier can transmit information and data of embedded equipment to the users through the network [3]. General CGI program developing flow under ARM is shown below.



VI. APPLICATION DESIGN BASED ON CGI:

I implement this idea for controlling speed of DC FAN & for changing status of DC Lamp [5]. System architecture is shown in Fig.5

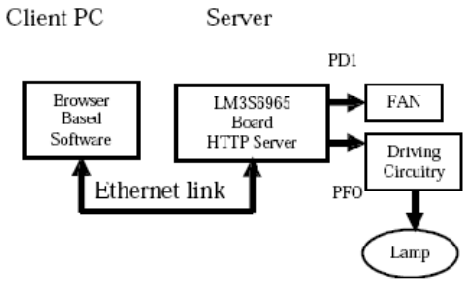


Fig. 5 : System Architecture

Here I used Code Composer Studio 4.1 software for developing the program. After designing CGI for specific application, build, debug & run that program. IP address of boardis displayed on OLED display. Start web browser, and enter the address shown on OLED display and press Enter. A web page served from the LM3S6965 should appear. It is shown in Fig.6

Fig.6.HTML page served by server

After clicking on CGI Demo, I get another page which is as shown in Fig.7. This page is specially designed for my application.

Fig.7. Web page for application.

As I tick on LED State & PWM State through checkbox with giving duty cycle for PWM then accordingly LAMP will be ON/OFF and with duty cycle speed of FAN will be changed.

VII. CONCLUSION:

The dynamic web technology is one of the core contents of embedded technology. Exploiting the dynamic web of CGI is becoming more efficient and easier. Based on this methodology control of a FAN &ON/OFF of LAMP is developed. By the simply function extending, this design method can be used widely for the intelligent family appliance, remote monitoring of the automated device, and some other related fields.

REFERENCES:

[1] Ganesh Venkitachalam, Tzi-cker Chiueh, "High Performance Common Gateway Interface Invocation," IEEE Trans.Comput.-Aided Design Integr.Circuits Syst., vol.20, no.2, pp.281-288, Feb.2001.

[2] Dubrovnik, Croatia, "Embedded Web Server Technology for Remote Online Labs", IEEE ISIE 2005, 2005, pp.120-145.

[3] Michael Sweeney, "BUS: Browser based User interface Service for web based application," Foreign Electronic Measurement Technology. Vol.25, No 11, 2006.

[4] Dan Li, Chongquan Zhong, Xiaofeng Yuan and Li Zhang, "Study on Embedded Equipment Web Monitoring and Control Based on Industrial Ethernet ," 6th World Congress on Intelligent Control and Automation, Dalian, China ., vol.20, no.5, pp.281-284,2006

[5] Jagadeesh Chandra A.P, "A Client-Friendly Near Real-Time Multimedia Technology for Web-Browser Based Remote Experimentation," IEEE Transactions on Education, Vol.48, No.3, pp.520-525, 2007.

[6] Yan Shi, "Development of Direct Current Motor Control System Based on Embedded Linux System," Proceedings IEEE/RSJ international conference on intelligent robots and system, 2007, pp. 226-230.

[7] Zhan mei-qiong, Ji chang-peng, " Research and Implementation of Embedded Web Server," International Conference on MultiMedia and Information Technology, 2008, pp. 305-308.

[8] Wei chen, Shu-bo qiu, Ying-chun zhang, "The Porting and Implementation of Light- Weight TCP/IP for Embedded Web Server ," Microcontrollers & Embedded Systems, vol 7, no. 3, 2008, pp. 80-84

[9] Wang ZhenXing, Ren XianYi, "A Study on Cgi of Embedded Web server," International Symposium on ComputerScience and Computational Technology, vol 4, No. 2 2008, pp. 180-184.

[10] Zhan mei-qiong, "Research and Implementation of Embedded Web Server," International Conference on Multimedia and Information Technology,2008,pp.670-674.

◈◈◈