International Journal of Computer Science and Security (IJCSS)

ISSN : 1985-1533

VOLUME 1, ISSUE 1

PUBLICATION FREQUENCY: 6 ISSUES PER YEAR

Editor in Chief Dr. Haralambos Mouratidis

International Journal of Computer Science and Security (IJCSS)

Book: 2007 Volume 1, Issue 1
Publishing Date: 28-02-2007
Proceedings
ISSN (Online): 1985-1553

IJCSS Journal is a part of CSC Publishers
http://www.cscjournals.org

Published in Malaysia

Typesetting: Camera-ready by author, data conversation by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers

Table of Contents

Volume 1, Issue 1, February 2007.

Analysis & Integrated Modeling of the Performance Evaluation Techniques for Evaluating Parallel Systems

Amit Chhabra chhabra_amit78@yahoo.com
Department of Computer Science & Engineering,
Guru Nanak Dev University,
Amritsar, 143001,India

Gurvinder Singh gsbawa71@yahoo.com
Department of Computer Science & Engineering,
Guru Nanak Dev University,
Amritsar, 143001,India

Abstract

Parallel computing has emerged as an environment for computing inherently parallel and computation intensive applications. Performance is always a key factor in determining the success of any system. So parallel computing systems are no exception. Evaluating and analyzing the performance of parallel systems is an important aspect of parallel computing research. Evaluating and analyzing parallel system is difficult due to the complex interaction between application characteristics and architectural features.

Experimental measurement, Theoretical/Analytical modeling and Simulation are the most widely used techniques in the performance evaluation of parallel systems. Experimental measurement uses real or synthetic workloads, usually known as benchmarks, to evaluate and analyze their performance on actual hardware. Theoretical/Analytical models try to abstract details of a parallel system. Simulation and other performance monitoring/visualization tools are extremely popular because they can capture the dynamic nature of the interaction between applications and architectures. Each of them has several types. For example, Experimental measurement has software, hardware, and hybrid. Theoretical/Analytical modeling has queueing network, Petri net, etc. and simulation has discrete event, trace/execution driven, Monte Carlo. Each of these three techniques has their own pros and cons.

The purpose of this paper is firstly to present a qualitative parametric comparative analysis of these techniques based on parameters like stage, output statistics, accuracy, cost, resource consumption, time consumption, flexibility, scalability, tools required, trustability and secondly to justify the need for an integrated model combining the advantages of all these techniques to evaluate the performance of parallel systems and thirdly to present a new integrated model for performance evaluation . This paper also discusses certain issues like selecting an appropriate metric for evaluating parallel systems.

Keywords: Integrated model, Metrics, Parallel systems, Performance, Evaluation

1.INTRODUCTION TO PARALLEL COMPUTING SYSTEMS

For the last three decades, researchers in the area of parallel processing have proclaimed that parallel computing is the wave of the future. The reason most widely cited for this wave is the rapid approach that resolves the speed limit of serial computing. This limit needs to cracked in order to solve many computing intensive applications in an acceptable amount of time. The only way to overcome the speed limit of serial computers is to harness the power of several serial computers to solve a single problem in a parallel fashion.

The advents in the today's micro-electronic technology have resulted in the availability of fast, inexpensive processors and advancement in the communication technology has resulted in the availability of cost-effective and highly efficient computer networks.

2. ROLE OF PERFORMANCE IN PARALLEL SYSTEMS

Performance is always a key factor in determining the success of parallel system. Quantitative evaluation and modelling of hardware and software components of parallel systems are critical for the delivery of high performance. Performance studies apply to initial design phases as well as to procurement, tuning, and capacity planning analyses. As performance cannot be expressed by quantities independent of the system workload, the quantitative characterization of resource demands of applications and of their behaviour is an important part of any performance evaluation study. Among the goals of parallel systems performance analysis are to assess the performance of a system or a system component or an application, to investigate the match between applications requirements and system architecture characteristics, to identify the features that have a significant impact on the application execution time, to predict the performance of a particular application on a given parallel system, to evaluate different structures of parallel applications.

3. SELECTING AN APPROPRIATE PERFORMANCE METRIC

To study the performance of systems or to compare different systems for a given purpose, we must first select some criteria. These criteria are often called metrics in performance evaluation. Different situations need different sets of metrics. Thus, selecting metrics are highly problem oriented. Different metrics may result in totally different values. Hence, selecting proper metrics to fairly evaluate the performance of a system is difficult. Another problem with selecting metrics is that in a real system different metrics may be relevant for different jobs. For example, response time may be the most suited metric for interactive jobs, while system utilization is more important for batch jobs. So issue here is to select an appropriate metric, which is suitable for every kind of systems (batch, interactive, closed and open systems).

Characteristics or properties of a good performance metric should be SMART i.e. Specific, Measurable, Acceptable, Realizable and Thorough. So far the researchers have proposed many parallel metrics. *Execution time* is the time interval between the beginning of parallel computation and the time since the last processing elements finishes execution. Another parallel metric is *Speedup. Speedup* is defined as the ratio of the time taken to execute a problem on a single processor to the time required to solve the same problem on a parallel computers with p identical processing elements. Amdahl and Gustafson [1] have proposed a Law for *Speedup* metric. The formulation for speed up using Amdahl's law are based on fixed problem size and speed up drops very rapidly with the increase in sequential fraction. So fixed load acts as deterrent in achieving the scalability in performance.

John Gustafson [1] proposed a fixed law using time concept to scale the speed up model and remove the fixed load restriction. This law states that problem size scales with the number of processors and idea is to keep all processors busy by increasing the problem size. Other variants of *speedup* are *Relative speedup* [6][7][8], *Real speedup* [3][5], *Absolute speedup* [6][7][8] and *Asymptotic relative speedup* [4].

Sun and Ni [6][8] have generalized Gustafson's scaled speedup to fixed-time relative speedup and proposed another speedup model called *memory-bounded relative speedup*. In memory constrained sealing, the problem is made to fit in the available memory.

Sun and Gustafson [7] proposed two new speedup measures; one of them is *generalized speedup* and other is *sizeup*. *Generalized speedup* is the ratio of parallel execution speed to sequential execution speed. It has been observed that when cost factor is same for all kinds of work, *Generalized speedup* equals *relative speedup*. In *sizeup* metric, the instance being solved is adjusted as for fixed-time speedup. The *sizeup* is the ratio of the serial work represented by the adjusted instance to the serial work represented by the unadjusted instance.

Another parallel metric is *efficiency*. This measure is very close to speedup. It is the ratio of speedup to the number of processors P. Depending on the variety of speedup used; one gets a different variety of efficiency. Carmona and Rice [2] define efficiency as the ratio of work accomplished (*wa*) by a parallel algorithm and the work expended (*we*) by the algorithm. Here work accomplished (*wa*) by the parallel algorithm is the work that is being done/performed by the "best" serial algorithm and work expended (*we*) is the product of the parallel execution time, the speed (*S*) of an individual parallel processor and the number P of processors. Also there is one another variant of efficiency known as *Incremental efficiency*. It is an approach to look at change of the efficiency of the implementation as the number of processors increases. The *Incremental efficiency* metric [11] looks at the ratio of the successive efficiencies for different number of processors, a quantity that tends to unity in the limit.

Another performance measure is *scalability*. *Scalability* refers to the change in the performance of the parallel system as the problem size and machine size increases. Intuitively, a parallel system is scalable if its performance continues to improve as we scale the size of the system.

Kumar, Nageshwara and Ramesh [10] proposed a scalability measure based on efficiency. In this scalability, or *isoefficiency*, of a parallel system is defined to be the rate at which workload must increase relative to the rate at which the number of processors is increased so that the efficiency remains the same. Isoefficiency is generally a function of the instance and number of processors in the parallel computer. Depending on the version of efficiency, different varieties of isoefficiency (e.g. *real isoefficiency*, *absolute isoefficiency* and *relative isoefficiency*) can be obtained. Sun and Rover [9] proposed *isospeed* that uses the average workload per processor needed to sustain a specified computational speed as a measure of scalability.

Utilization is another measure for evaluating resource utilization and may be defined as ratio of total usage time over the total available time of the resources (processors, memory). In addition to the above metrics, some other general measures such as *CPU time*, and *CPI* (Clock cycles for instruction) play important role in measuring the success of parallel implementation of problem.

4. PERFORMANCE EVALUATION TECHNIQUES

Performance evaluation can be defined as assigning quantitative values to the indices of the performance of the system under study. Evaluating and analyzing parallel system is difficult due to the complex interaction between application characteristics and architectural features. Performance evaluation techniques can be classified into three categories; Experimental measurement, Theoretical/Analytical modelling and Simulation. In this section all of these three techniques are discussed and compared with each other on the basis of some parameters.

Four major parameters for selecting a technique for performance evaluation are:

a) *Stage*-this parameter examines which performance evaluation technique should be used at what stage of system development life cycle.

b) *Output statistics*-this parameter examines the capabilities of the technique towards providing the desirable metrics.

c) *Accuracy*-this factor evaluates the validity and reliability of the results obtained from the technique.

d) *Cost/effort*-this parameter investigates the cost and effort invested in each performance evaluation strategy in context with computer and human resources.

Various other parameters for selecting a technique for performance evaluation are:

e) *Resource consumption*-this parameter examines the amount of resources consumed/required by a particular performance evaluation technique.

f) *Time consumption*- this parameter examines the amount of time consumed/required by a particular performance evaluation technique.

g) *Tools required*- this parameter examines the type of tools required for implementation of any particular performance evaluation technique.

h) *Trustability/Believability*- this parameters reveals that how much one can trust on the results of a particular performance evaluation technique.

i) *Scalability complexity*-this parameter examines the complexity involved in scaling a particular performance evaluation technique.

j) *Flexibility*-this parameter examines the flexibility of a particular performance evaluation technique towards adapting the modifications made to the model of evaluation technique and checking their effect.

4.1. Experimental Measurement

It is based on direct measurements of the system under study using a software, hardware or/and hybrid monitor. It uses real or synthetic workloads and measures their performance on actual hardware. As it uses the actual hardware and the related system software to conduct the evaluation, making it the most realistic model from the architectural point of view. A monitor is a tool, which is used to observe the activities on a system. In general, a monitor performs three tasks: data acquisition, data analysis, and result output. The data recorded by a monitor include hardware related data, e.g. processor usage throughout program run, message latency, and software data, e.g. process times, buffer usage, load balancing overhead [12]. Traditionally, monitors are classified as *software* monitors, *hardware* monitors, and *hybrid* monitors. Software monitors are made of programs that detect the states of a system or of sets of instructions, called *software probes*, capable of event detection. Hardware monitors are electronic devices to be connected to specific system points where they detect signals characterizing the phenomena to be observed [13].

A hybrid monitor is a combination of software and hardware. Many examples of software monitors can be found in the literature [14][15][16]. Examples of hardware monitors are COMTEN and SPM [17]. [18] describes the Paragon performance-monitoring environment that uses a hardware performance monitoring board. And examples of hybrid monitors are Diamond [20] and HMS [19].

Each class of monitors has its own advantages and disadvantages. Selecting an appropriate monitor involves various aspects, e.g., cost, overhead, accuracy, availability, information level, etc.. In general, hardware monitors have received less attention than software monitors [15]. This has been shown by the fact that the number of existing software monitors are far greater than that of hardware monitors. [21] describes and compares software and hardware monitors in more detail. Application suites such as the Perfect Club [22], the NAS Parallel Benchmarks [23], and the SPLASH application suite [24] have been proposed for the evaluation of parallel machines.

Parameters under consideration

a) *Stage*- As Experimental measurement uses the actual hardware, so it cannot be used at the early stage of system design. It can be only possible when the system design has been completed and a real system is being available for evaluation. So experimental measurement is only possible when actual system under study is available.

b) *Output statistics*-Experimental measurement can give the overall execution time of the application on particular hardware platform. Conducting the evaluation with different problem sizes and number of processors would thus helps to calculate metrics such as speedup, scaled speedup, sizeup, isoefficiency.

c) *Accuracy*- Experimental measurement uses real applications to conduct the evaluation on actual machines giving accurate results. But external factors like external instrumentation and monitoring intrusion can affect the accuracy of results.

d) *Cost/Effort*-Substantial costs are involved in developing the model for experimentation as experimental technique uses the actual real system to give results. This cost is directly dependent on the complexity of the application. Once the model for experimental measurement is developed, the cost for conducting the actual evaluation, that is the execution time of the given application, may be very small. Modifications to the application and hardware can thus be accommodated by the experimentation technique at a cost that is totally dependent on the type and the amount of modifications. In other words scalability can be complexier but it totally depends how much one wants to scale.

e) *Resource consumption*-This technique requires actual instruments, so resource consumption is high.

f) *Time consumption*- As experimentation requires actual system setup which consumes heavy amount of time but once the system comes into reality, time required for results may be very less.

g) *Tools required*-Actual instruments are required here to evaluate any parallel application

h) *Trustability/Believability*-As it uses actual hardware so results given by it can be highly trustworthy. Here results could be validated by atleast simulation or theoretical/analytical modelling.

i) *Scalability complexity*-Here complexity is totally dependent on the fact that how much scaling of the system is required and this scaling requires time and money.

j) *Flexibility*-It is not highly flexible because it takes lot of time to change a particular experimental setup and it takes lot of time to check the effect of this change.

4.2 Theoretical/Analytical Modelling

Performance evaluation of parallel systems is hard due to the several degrees of freedom that they exhibit. Analytical and theoretical models try to abstract details of a system, in order to limit these degrees of freedom to a tractable level. Such abstractions have been used for developing parallel algorithms and for performance analysis of parallel systems.

Abstracting machine features by theoretical models like the PRAM [25][26] has facilitated algorithm development and analysis. These models try to hide hardware details from the programmer, providing a simplified view of the machine. The utility of such models towards developing efficient algorithms for actual machines depends on the closeness of the model to the actual machine. Several machine models like Bulk Parallel Random Access Machine (BPRAM)[29], Module Parallel Computer (MPC)[30], Valiant [28] introduces the *Bulk-Synchronous Parallel (BSP)* model and LogP[27] have been proposed over the years to bridge the gap between the theoretical abstractions and the hardware.

While theoretical models attempt to simplify hardware details, analytical models abstract both the hardware and application details in a parallel system. Analytical models capture complex system features by simple mathematical formulae, parameterized by a limited number of degrees of freedom that are tractable. Such models have found more use in performance analysis than in algorithm development where theoretical models are more widely used. As with experimentation, analytical models have been used to evaluate overall system performance as well as the performance of specific system artifacts. Vrsalovic et al. [31] develop an analytical model for predicting the performance of iterative algorithms on a simple multiprocessor abstraction, and study the impact of the speed of processors, memory, and network on overall performance.

Parameters under consideration-

a) *Stage*- Theoretical/Analytical modelling can be used at the early design phase of system development life cycle, as it does not require any actual hardware for evaluation. These models try to hide hardware details from the programmer, providing a simplified view of the machine and also behaviour of these models is very close to that of actual machine.

b) *Output statistics* – Theoretical/Analytical models can directly present statistics for system overheads that are modeled. The values for the hardware and the workload parameters can be plugged into the model corresponding to the system overhead. With models available for each system overhead, overall execution time and all other metrics can be calculated. The drawback is that each system overhead needs to be modeled or ignored in calculating the execution time.

c) *Accuracy*- Theoretical/analytical models are useful in predicting system performance and scalability trends as parameterized functions. However, the accuracy of the predicted trends depends on the simplifying assumptions made about the hardware and the application details to keep the models tractable. Theoretical models can use real applications as the workload, whereas analytical models represent the workload using simple parameters and probability distributions. Thus the former has an advantage over the latter in being able to estimate metrics of interest more accurately. But even for theoretical models, a static analysis of application code, which is used to estimate the running time, can yield inaccurate results.
d) *Cost/effort*- Substantial effort is involved in the development of models for theoretical/analytical modelling. Simple modifications to the application and hardware can be easily handled with these models by changing the values for the corresponding parameters and re-calculating the results. But a significant change in the hardware and application would demand a re-design of the input models, which can be expensive.
e) *Resource consumption*-Not much resources are required to build models for theoretical/analytical modelling.
f) *Time consumption*- Good amount of time is consumed for developing models of Theoretical/Analytical modelling but once models have been developed it takes little time to get results.
g) *Tools*- It involves analyst and various analysis tools.
h) *Trustability/Believability*-This method involves assumptions that are being made while developing models. So they are not much trustworthy until there results are validated by atleast simulation or experimental measurement.

i) *Scalability complexity*-It takes time to scale these kind of models.
j) *Flexibility*-It is flexible as changes can be made to the models easily and effect of change can also be checked easily.

4.3. Simulation

Simulation is a widely used technique in performance evaluation. It provides a powerful way to predict performance before the system under study has not been implemented. It can also be used to validate analytical models, as is done in [32][34]. There are a variety of simulations presented in the literature: emulation, Monte Carlo simulation, trace-driven simulation, execution-driven simulation, and discrete-event simulation.

An example of emulation is using one available processor (host processor) to emulate the instruction set of another processor (target processor) that is not available or under design. This type of simulation is sometimes called by some authors *instruction-level simulation* [35] or *cycle-by-cycle simulation*. Programs written for the target processor are simulated by the simulator running on the host processor to study the behavior of the programs if they were executed on the target processor.

Monte Carlo simulation is a static simulation where the simulated systems do not change their characteristics with time. Computer systems are dynamic systems, and do not belong to this category.

A trace-driven simulation system consists of two components: an event generator (or trace generator) and a simulator. The event generator produces a trace of execution events, mostly addresses, which are used as input to the simulator. The simulator consumes the traced data and simulates the target architecture to estimate the time token to perform each event on the architecture under study.

Discrete-event simulation is used to simulate systems that can be described using discrete event models. Discrete-event simulation is very well suited for studying queueing systems.

Parameters under consideration-

a) *Stage*-Simulation can not be used at very early stage of system design because of the non-availability of required system details at that point but as the design process goes on and more details about the system are obtained, this technique becomes powerful tool at that point.

b) *Output statistics*-Simulation provides a convenient monitoring environment for observing details of parallel system execution, allowing the user to accumulate a range of statistics about the application, the hardware, and the interaction between the two. It can give the total execution time and all other metrics discussed earlier.

c) *Accuracy*- The accuracy of results depends purely on the accuracy of input models. Execution-driven simulation can faithfully simulate all the details of a real-world application. It is also possible to simulate all the details of the hardware, though in many circumstances a level of abstraction may be chosen to give moderately accurate results for the intended purposes. The accuracy of these abstractions may also be validated by comparing the results with those obtained from a detailed simulation of the machine or an experimental evaluation on the actual machine.
d) *Cost/Effort*- The main disadvantage associated with simulations is the cost and effort involved in simulating the details of large parallel systems. With regard to modifiability, plugging in these values into the model and re-simulating the system may handle a moderate change in hardware parameters. But such a re-simulation, as we observed, is invariably costlier than a simple re-calculation that is needed for analytical models, or experimentation on the actual machine. A significant change in the machine or application details would also demand a re-implementation of the simulation model, but the cost of re-simulation is again expected to dominate over the cost of re-implementation.
e) *Resource Consumption*-It requires small amount of resources as we are just simulating the parallel environment, no actual instrumentation is being required.
f) *Time consumption*-Good amount of time is consumed for developing the simulation model. But once model has been developed it takes little time to get results.
g) *Tools*-Simulation requires high level computer programming languages to build and develop the model.
h) *Trustability/Believability*-simulated model is just a replica of the final actual machine but results may little bit vary as this is not an actual machine.
i) *Scalability complexity*-Simulated models are very easy scalable as it artificially simulates the actual environment.

j) *Flexibility*-The main advantage of simulation is its flexibility. One can make various modifications to the simulation model and check their effect easily.

5. NEED FOR INTEGRATED MODELLING OF PERFORMANCE EVALUATION TECHNIQUES

Each performance evaluation technique has its own role to play in evaluating parallel systems. As shown in TABLE 1 each technique has its own advantages and disadvantages. So effort can be made in developing knowledge based integrated model, which combines the advantages of all the three techniques.

	Performance Evaluation Techniques		
Characteristic	**Theoretical/Analytical Modelling**	**Simulation**	**Experimental Measurement**
a) Stage	Any	Any	After prototype is available
b) Output Statistics	It can give total execution time and all other metrics discussed earlier	It can also give total execution time and all other metrics discussed earlier	It can also give total execution time and all other metrics discussed earlier
c) Accuracy	Low	Medium	High
d) Cost/ Effort	Low	Medium	High
e) Resource Consumption	Small	Medium	High
f) Time Consumption	Small	Medium	High
g) Tools	Analysts	Computer Programming Languages	Instruments
h) Trustability/ Believability	Low	Medium	High
i) Scalability Complexity	Small	Medium	High
j) Flexibility	High	High	Low

TABLE1: Showing comparison of performance evaluation techniques

This integrated model has the advantage that it benefits the realism and accuracy of experimentation in evaluating large parallel systems, the convenience and power of theoretical/analytical models in predicting the performance and scalability of the system as a function of system parameters and the accuracy of detailed statistics provided by execution-driven simulation and avoid some of their drawbacks. Also we need an integrated model that takes care of three rules of validation that says

- Do not trust the results of a Simulation model until they have been validated by at least Theoretical/analytical model or Experimental measurements.

- Do not trust the results of a Theoretical/analytical model until they have been validated by at least Simulation or Experimental measurements.
- Do not trust the results of a Experimental measurement model until they have been validated by at least Simulation or Theoretical/analytical modelling.

6. PROPOSED INTEGRATED MODEL

This integrated model shown in Fig. 1 is going to use the power of stored performance knowledge base systems. User applications are being fed to experimental measurement technique whose stored performance knowledge base system is also attached with it as indicated in Fig. 1. Based on the user application, the type of workload it is using (real or synthetic) and any other current data, a knowledge set can be extracted from this stored knowledge base system which will suggest which experimental technique (software, hardware or hybrid) is best or optimal for the current user application. Also when results are obtained after a particular technique is applied on user application, these results can in turn act as knowledge. This valuable knowledge can be updated back into the knowledge base system and this procedure of knowledge extraction and updation can be repeated. Stored Knowledge base systems can also be attached with all other evaluation techniques (Simulation and Theoretical/Analytical modelling). These stored knowledge base systems helps in choosing technique for evaluating current user application. Same procedure of knowledge extraction and updation is also true for other measurement techniques.

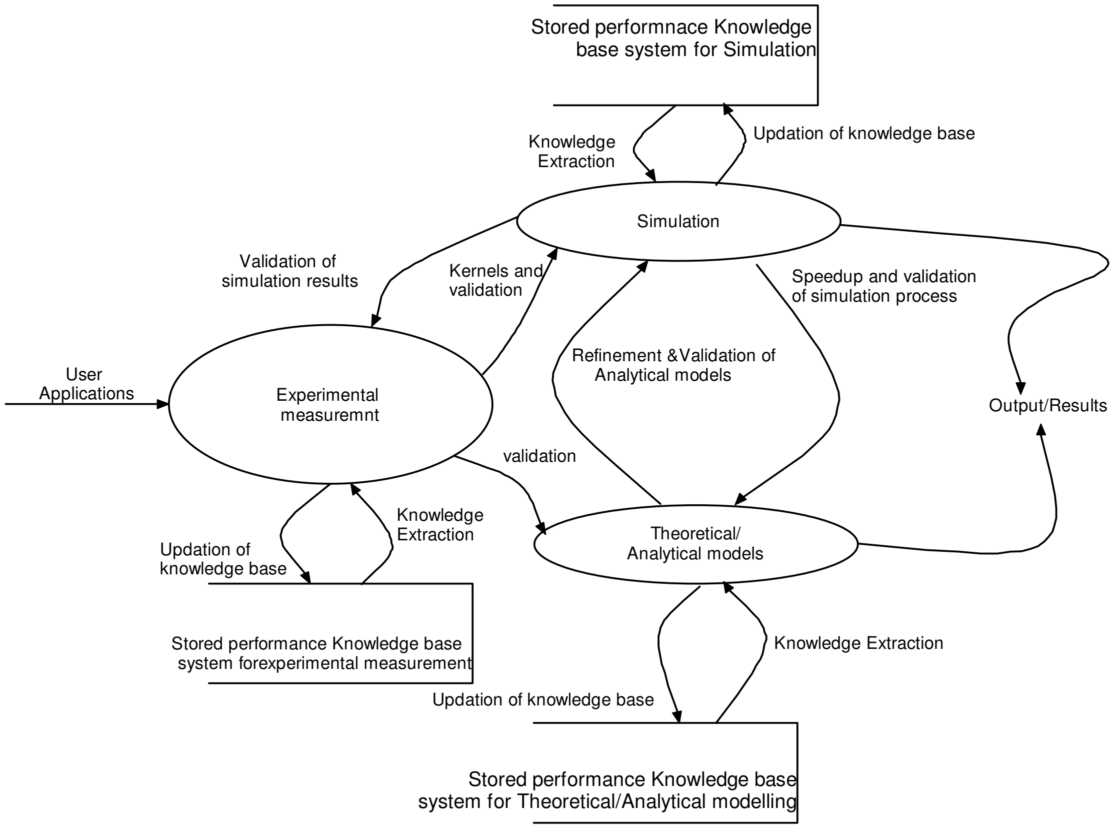

Stored performnace Knowledge base system for Simulation
Knowledge Extraction
Updation of knowledge base
Simulation
Validation of simulation results
Kernels and validation
Speedup and validation of simulation process
Refinement &Validation of Analytical models
User Applications
Experimental measuremnt
Output/Results
validation
Knowledge Extraction
Updation of knowledge base
Theoretical/ Analytical models
Stored performance Knowledge base system forexperimental measurement
Knowledge Extraction
Updation of knowledge base
Stored performance Knowledge base system for Theoretical/Analytical modelling

FIGURE 1: Diagram showing proposed integrated model for performance evaluation

Experimental measurement can be used to implement real-life applications on actual machines, to understand their behaviour and to extract interesting kernels that occur in them. These kernels are fed to an execution-driven simulator, which faithfully and successfully models the dynamics of parallel system interactions. The statistics that are drawn from simulation may be used to validate and refine existing theoretical/analytical models, and to even develop new models. The validated and refined models can help in abstracting details in the simulation model to enhance the speed of simulation.

The validity of such a simulation model can in turn be verified by comparing the simulation results with those from an experimental evaluation on the actual hardware. Such a strategy combines advantages of all three techniques, uses the power of stored knowledge base systems and avoids the shortcomings of the individual evaluation techniques.

7. CONCLUSION

Performance evaluation as a discipline has repeatedly proved to be critical for design and successful use of parallel systems. At the early stage of design, performance models can be used to project the system scalability and evaluate design alternatives. At the production stage, performance evaluation methodologies can be used to detect bottlenecks and subsequently suggest ways to alleviate them. In this paper three techniques of parallel system performance evaluation are reviewed and compared with each other with the help of four major parameters *Stage*, *Output statistics*, *Accuracy* and *Cost/Effort* involved. Other parameters involved in the selection of performance evaluation technique are *Resource consumption, Time consumption, Tools required, Trustability, Scalability Complexity and Flexibility.*Each of the three techniques has their own pros and cons. So an integrated model that uses the power of knowledge base systems and combines advantages of all the three techniques is discussed.Issue like selecting an appropriate metrics for performance evaluation is also discussed.

8.REFERENCES

[1]J.Gustafson, “*Reevaluating Amdahl’s Law*”, CACM, 31,5,532-533,1988.

[2]E.Caromona ,M.Rice, “*Modelling the serial and parallel fractions of a parallel algorithm*”,Journal of Parallel and Distributed Computing,13,286-298,1991.

[3]J.JaJa, “ *An introduction to parallel algorithms*”,Addison Wesley,1992.

[4]D.Nussbam and A.Agrawal, “ *Scalability of parallel machines*”,CACM,34,3,57-61,1991.

[5]S.Ranka, S.Sahni, “*Hypercube algorithms*”,Springer-Verlag,New York,1990.

[6]X.Sun, L.Ni, “*Another view on parallel speedup*”,Proceedings Supercomputing 90,324-333,1990.

[7]X.Sun ,J.Gustafson, “*Towards a better parallel performance metric*”,Parallel Computing,17,1093-1109,1991.

[8]X.Sun ,L.Ni, “ *Scalable problems and memory-bouneded speedup*”, Journal of Parallel and Distributed Computing,19,27-37,1993.

[9]X.Sun ,D.Rover, “*Scalability of parallel algorithm-machine combinations*”,IEEE Transactions Of Parallel and Distributed systems,5,6,599-613,1994.

[10]V.Kumar,V.Nageshwara and K.Ramesh, “*Parallel depth first search on the ring architecture*”, Proc.1988 International Conference on Parallel Processing,Penn. State Univ. Press,128-132,1988.

[11]J.Worlton,“*Toward a taxonomy of performance metrics*”, Parallel Computing 17(10-11): 1073-1092 (1991) .

[12]Jelly, I. ,Gorton, I., “*Software engineering for parallel systems*”, Information and Software Technology, vol. 36, no. 7, pp. 381-396, 1994.

[13]Ferrari, D., “*Considerations on the insularity of performance evaluation*”, Performance Evaluation Review, vol. 14, no. 2, pp. 21-32, August 1986.

[14]Plattner, B. ,Nievergelt, J., “*Monitoring program execution: A survey,*” IEEE Computer, vol. 14, pp. 76-93, November 1981.

[15]Power, L. R., “*Design and use of a program execution analyzer,*” IBM Systems Journal, vol. 22,no. 3, pp. 271-294, 1983.

[16]Malony, A. D., Reed, D. A. ,Wijshoff, H. A. G., “*Performance measurement intrusion and perturbation analysis,*” IEEE Transactions on Parallel and Distrubuted Systems, vol. 3, no. 4, pp. 443-450, July 1992.

[17]Ibbett, R., “*The hardware monitoring of a high performance processor,*” in: Benwell, N. (ed), Computer Performance Evaluation, Cranfield Institute of Technology, UK, pp. 274-292, December 1978.

[18]Ries, B., Anderson, R., Auld, W., Breazeal, D., Callaghan, K., Richards, E. and Smith, W., “*The Paragon performance monitoring environment,*” Proceedings of the conference on Supercomputing’93, pp. 850-859, 1993.

[19]Hadsell, R. W., Keinzle, M. G. and Milliken, K. R., “*The hybrid monitor system*,” Technical Report RC9339, IBM Thomas J. Watson Research Center, New York, 1983.

[20]Hughes, J. H., “*Diamond ¾ A digital analyzer and monitoring device*,” Performance Evaluation Review, vol. 9, no. 2, pp. 27-34, 1980.

[21]Jain, R., *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*, John Wiley & Sons, New York, 1991.

[22]M.Berryetal. *The Perfect Club Benchmarks: Effective Performance Evaluation of Supercomputers*. International Journal of Supercomputer Applications, 3(3):5–40, 1989.

[23]D. Bailey et al. *The NAS Parallel Benchmarks*. International Journal of Supercomputer Applications, 5(3):63–73, 1991.

[24]J. P. Singh, W-D. Weber, and A. Gupta. SPLASH: Stanford Parallel Applications for Shared-Memory. Technical Report CSL-TR-91-469, Computer Systems Laboratory, Stanford University, 1991.

[25] S. Fortune and J. Wyllie. Parallelism in random access machines. In *Proceedings of the 10th Annual Symposium on Theory of Computing*, pages 114–118, 1978.

[26]P. B. Gibbons. A More Practical PRAM Model. In *Proceedings of the First Annual ACM Symposium on Parallel Algorithms and Architectures*, pages 158–168, 1989.

[27]D. Culler et al.”*LogP: Towards a realistic model of parallel computation*”In Proceedings of the 4th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, pages 1–12, May 1993.

[28]L. G. Valiant. “*A Bridging Model for Parallel Computation*” Communications of the ACM, 33(8):103–111, August 1990.

[29]A. Aggarwal, A. K. Chandra, and M. Snir “ *On Communication Latency in PRAM Computations*” In Proceedings of the First Annual ACM Symposium on Parallel Algorithms and Architectures, pages 11–21, 1989.

[30]H. Alt, T. Hagerup, K. Mehlhorn, F. P. Preparata “ *Deterministic Simulation of Idealized Parallel Computers on More Realistic Ones*” SIAM Journal of Computing, 16(5):808–835, 1987.

[31] D. F. Vrsalovic, D. P. Siewiorek, Z. Z. Segall, and E. Gehringer “*Performance Prediction and Calibration for a Class of Multiprocessors*” IEEE Transactions on Computer Systems, 37(11):1353–1365, November 1988.

[32]Agarwal, A., *“Performance tradeoffs in multithreaded processors*,” IEEE Transactions on Parallel and distributed Systems, vol. 3, no. 5, pp. 525-539, September 1992.

[33]Menasce, D. A. ,Barroso, L. A., “*A methodology for performance evaluation of parallel applications on multiprocessors,*” Journal of Parallel and Distributed Computing, vol. 14, pp. 1-14,1992.

[35]Covington, R. G., Dwarkadas, S., Jump, J. R., Sinclair, J. B. ,Madala, S., “*The efficient simulation of parallel computer systems,*” International Journal in Computer Simulation, vol. 1, pp.31-58, 1991.

A Novel Secure Key Agreement Protocol using Trusted Third Party

Sairam Kulkarni kulkarnisairam@gmail.com
Department of Computer Science & Engg.
National Institute of Technology Rourkela
Rourkela, 769008, India

Debasish Jena debasishjena@hotmail.com
Assistant Professor
Centre for IT Education
Biju Patnaik University of Technology
Bhubaneswar, 751010, India

Sanjay Kumar Jena skjena@nitrkl.ac.in
Professor
Department of Computer Science & Engg.
National Institute of Technology Rourkela
Rourkela, 769008, India

Abstract

In the past, several key agreement protocols are proposed on password based mechanism. These protocols are vulnerable to dictionary attacks. Storing plain text version of password on server is not secure always. In this paper we utilize the service of a trusted third party, i.e., the Key Distribution server (KDS) for key agreement between the hosts. Now-a-days in large working environments two party key agreement protocols are being rarely used. In this proposed scheme, instead of storing plain text version of password we store one way hash of the password at the server. Every host and server agree upon family of commutative one-way hash functions, using which host authentication is done when a host applies for session key with KDS. Host establishes one time key with server using which server authentication is done. Due to this man-in-the middle attacks are defeated. The proposed protocol is based on Diffie-Hellman key exchange protocol.

Keywords: Key Agreement, Diffie-Hellman, Online guessing attacks, Dictionary attacks.

1. INTRODUCTION

The main goal of cryptography is to enable secure communication in a hostile environment. Two parties P_i and P_j, want to safely communicate over a network occupied by an active adversary. Usually, P_i and P_j will want to ensure the privacy and authenticity of the data they send to each other. They will encrypt and authenticate their transmissions. But before P_i and P_j can use these tools they will need to have keys. Indeed, without keys, cryptography simply cannot get off the ground. Key agreement is one of the fundamental cryptographic primitive after encryption and digital signature. Such protocols allow two or more parties to exchange information among themselves over an adversarially controlled insecure network and agree upon a common session key, which may be used for later secure communication among the parties. Thus, secure key agreement protocols serve as basic building block for constructing secure, complex, higher-level protocols. Key establishment may be broadly subdivided into key transport and key agreement.

Secret communications with secret keys implies that only trusted parties should have copies of the secret key. Although secret keys can assure us of confidentiality, authentication of users, and message integrity, in a global world we must be able to securely distribute keys at a distance in a timely manner [1].

If security is to be maintained, key distribution must be as solid as the cryptographic method and must be able to ensure that only trusted parties have copies of the keys [2]. Obviously, key distribution is a significant problem. Key establishment protocols involving authentication typically require a set-up phase whereby authentic and possibly secret initial keying material is distributed. Most protocols have as an objective the creation of distinct keys on each protocol execution. In some cases, the initial keying material pre-defines fixed key which will result every time the protocol is executed by a given pair or group of users. Systems involving such static keys are insecure under known-key attacks.

Key pre-distribution schemes are key establishment protocols whereby the resulting established keys are completely determined a priori by initial keying material. In contrast, dynamic key establishment schemes are those whereby the key established by a Fixed pair (or group) of users varies on subsequent executions. Dynamic key establishment is also referred to as session key establishment. In this case the session keys are dynamic, and it is usually intended that the protocols are immune to known-key attacks. Many key establishment protocols involve a centralized or trusted party, for either or both initial system setup and on-line actions (i.e., involving real-time participation). This party is referred to by a variety of names depending on the role played, including: trusted third party, trusted server, authentication server, key distribution center (KDC), key translation center (KTC), and certification authority [3] [4].

It is generally desired that each party in a key establishment protocol be able to determine the true identity of the other(s) which could possibly gain access to the resulting key, implying preclusion of any unauthorized additional parties from deducing the same key. In this case, the technique is said (informally) to provide secure key establishment. This requires both secrecy of the key and identification of those parties with access to it [5].

In a secure system, passwords can be easily guessed if user chooses their own password in plain text [6]. Storing plain text version of password on server is not secure. This weakness exists in practically all widely used systems. The proposed protocol is secure against dictionary attacks as we use one time keys with server. This protocol is also secure against malicious insider attacks, where a host misuses the information in one protocol run to another. Proposed protocol also provides perfect forward secrecy i.e. even if one key is disclosed future session keys will not be disclosed. As we don’t use any Public Key Infrastructure (PKI), large computational power is not required. Since this is a third-party key agreement protocol every host need not share secret information with other host.

In this paper in Section 2, we review short comings of existing protocols. In section 3 we discuss our new third-party Key Agreement Protocol. Formal security analysis of proposed protocol is done in Section 4. Finally concluding remarks is done in Section 5.

2. RELATED WORK

DH-BPAKE [7] is a two party key agreement protocol based on Diffie-Hellman [8] and Encrypted key exchange protocols which were proposed by Strangio [9]. This protocol is not suitable for large networks where we cannot assume that every party shares a secret (password) with every other party. Simple Authenticated Key Agreement (SAKA) protocol proposed by Her-Tyan Yeh et al. [10] is also a two party key agreement protocol which is based on password based authentication and Diffie-Hellman key agreement. User authentication is one of the most important security services in secure communications. It is necessary to verify the identities of the communicating parties before they start a new connection. Password-based mechanism is the most widely used method for user authentication since it allows people to choose and remember their own password without any assistant device. This protocol is simple and cost effective, but is being rarely used in large networks.

STW protocol is a three party Encrypted key exchange protocol proposed by Steiner et al. [11]. Since this is a three party key agreement protocol, both the hosts share a secret key only with trusted third party. Ding et al [12] have proved that this protocol is vulnerable to undetectable online guessing attacks. According to Lin C.L. et al [13], this protocol is also vulnerable to offline guessing attacks. An attacker attempts to use a guessed password in an online transaction. Host verifies the correctness of his guess using responses from server. If his guess fails he must start a new transaction with server using another guessed password. A failed guess can not be detected and logged by server, as server is not able to depart an honest request from a malicious request. In off-line guessing attacks an attacker guesses a password and verifies his guess off-line. No participation of server is required, so server does not notice the attack. If his guess fails, the attacker tries again with another password, until he finds the proper one. Among these classes of attacks, the off-line password guessing attack is the most comfortable and promising one for an attacker. It is not noticeable and has no communication cost. Storing a plain text version of the shared password at the server is a constraint that cannot (or ought not) always be met. In particular, consider the problem of a user logging in to a computer that does not rely on a secure key server for authentication. It is inadvisable for most hosts to store passwords in either plain form or in a reversibly encrypted form.

LSH 3-PEKE protocol was proposed by Chun-Li Lin et al [13]. This protocol is secure against both the offline guessing attack and undetectable on-line guessing attacks but also satisfies the security properties of perfect forward secrecy. The most important requirement to prevent undetectable on-line guessing attacks is to provide authentication of host to server. In the STW 3-PEKE, there is no verifiable information for server to authenticate host. On the contrary, if there is any verifiable information for server combined with password will result in offline guessing attacks. LSH 3-PEKE uses server public keys for this purpose. But this is not a satisfactory solution all the times and is impractical for some environments. Communication parties have to obtain and verify the public key of the server, a task which puts a high burden on the user. In fact, key distribution services without public-keys are quite often superior in practice than PKI.

3. PROPOSED 3-PARTY KEY AGREEMENT PROTOCOL

Our proposed protocol withstands all online [12] and offline guessing attacks [13], and does not makes use of PKI. Every host and server agree upon family of commutative one-way hash functions using which host authentication is done when it applies for session key. Host establishes one time key with server using which server authentication is done. Rather than storing a plain text version of password we store one way hash of password at server. A one-way function is a function f such that for each x in the domain of f, it is easy to compute $y = f(x)$, but it is computationally infeasible to find any x given $f(x)$.

3.1 Notations

In this paper, we use the following notations

A, B	Full principal names
S	Trusted Third Party
$E_K(X)$	Encryption of plaintext block X under key K
$D_K(X)$	Decryption of plaintext block X under key K
K_{AB}	A and B share Key K
$H_{K_{AB}}(X)$	One way hash of X using key KAB
N_{AB}	Nonce generated by A and received by B
P_A	Password of A
$H(P_A)$	One way hash of password of A
g	Generator of cyclic group
P	Large prime number
$A \rightarrow B \quad M$	A sends message "M" to B

3.2 Proposed Protocol

In this subsection, we describe the steps involved in detail.

i. *A* chooses a random number ra and generates $R_A = g^{ra}(\bmod \quad p)$ then encrypts R_A with $H(P_A)$. After calculating the values sends it to server along with IDs of participating entities.

$$A \rightarrow S \qquad ID_A, ID_B, H(P_A)[R_A]$$

ii. After receiving the values sent by *A*, server *S* decrypts the packet to get R_A by previously distributed one way hash of password of *A*. server randomly chooses $rs1$ and $rs2$ and computes ephemeral key with *A* as follows

$$K_{AS} = (R_A)^{rs1}(\bmod p) = (g^{ra})^{rs1} \bmod p$$

S generates $g^{rs1}\ (\bmod p)$ and $g^{rs2}\ (\bmod p)$ and encrypts with $H(P_A)$ and $H(P_B)$ respectively. Using these quantities server establishes ephemeral keys with *A* and *B* respectively and server authentication is done. *S* sends the values to *A*

$$S \rightarrow A \qquad H(P_A)(g^{rs1} \bmod p), H(P_B)(g^{rs2} \bmod p)$$

iii. *A* decrypts this packet with $H(P_A)$ to get $g^{rs1}\ (\bmod p)$ and establishes ephemeral key with *S* as $K_{AS} = (g^{rs1})^{ra} \bmod p$.*A* calculates one way function $F_A(P_A, K_{AS})$ using which server authenticates *A*, since only *A* knows P_A it can compute this function. As this is a commutative one way hash function [14], server need not know host password to evaluate this function. Using one way hash of host password server can calculate predicate function and authenticate host. *A* sends the following values to *B*

$$A \rightarrow B \qquad F_A(P_A, K_{AS}), H(P_B)(g^{rs2} \bmod p)$$

iv. After receiving the values *B* decrypts it with $H(P_B)$ to get $(g^{rs2} \bmod p)$.*B* chooses randomly rb and generates $R_B = g^{rb}(\bmod p)$ .Then computes ephemeral key for authenticating server as $K_{BS} = (g^{rs2})^{rb} \bmod p$. *B* calculates one way function $F_B(P_B, K_{BS})$, using which server authenticates *B*. Password of *B* and ephemeral session key K_{BS} are seeds for this function. Since only *B* knows P_B it can compute this function and sends the values to *S.*

$$B \rightarrow S \qquad F_A(P_A, K_{AS}), F_B(P_B, K_{BS}), H(P_B)[R_B]$$

v. server decrypts it with $H(P_B)$ to get R_B and computes ephemeral key $K_{BS} = (g^{rb})^{rs2} \bmod p$. For authentication of *A* and *B* server evaluates one way functions $F_A(...), F_B(...)$. server need not know host passwords to evaluate these functions. Using one way hash of host password it can evaluate this function as it is a commutative one way hash function. If it results into true then it confirms that host is genuine. It defines a predicate as $T(H(P), F(P,K), K)$. This evaluates to true if and only if the genuine password *P* was used to create both $H(P)$ and $F(P,K)$. *K* can be K_{AS}, K_{BS} for A and B respectively. *S* encrypts R_B and R_A with K_{AS}, K_{BS} respectively and computes one way hash function $H_{K_{AS}}(R_A, R_B)$ using K_{AS} (one time key shared between *A* and server). Using this host *A* authenticates the server. Similarly *S* computes one way hash function $H_{K_{BS}}(R_A, R_B)$ using K_{BS} (one time key shared between *B* and server) and authenticates *B* and sends the values to *B*.

$$S \rightarrow B \qquad E_{K_{AS}}(R_B), E_{K_{BS}}(R_A), H_{K_{AS}}(R_A, R_B), H_{K_{BS}}(R_A, R_B)$$

vi. After receiving this *B* decrypts $E_{K_{BS}}(R_A)$ with K_{BS} and gets R_A. Since K_{BS} is shared between server and *B*, it ensures *B* that R_A value is from authentic source. *B* computes one way hash $H_{K_{BS}}(R_A, R_B)$ using K_{BS} as key and authenticates server. *B* computes session key with *A* as $K_{AB} = (R_A)^{rb} (\bmod p)$. *B* computes a one way hash $H_{K_{AB}}(N_{AB})$ using K_{AB} and N_{AB} as seeds, where N_{AB} is a random number. This one way hash is used for key confirmation (assures that both parties posses same session key). Since N_{AB} is transmitted in plain there is no need of decryption. One way hash suffices decryption. After computing all the values it sends to *A.*

$$B \rightarrow A \qquad E_{K_{AS}}(R_B), H_{K_{AS}}(R_A, R_B), H_{K_{AB}}(N_{AB}), N_{AB}$$

vii. *A* decrypts $E_{K_{AS}}(R_B)$ using K_{AS} to get R_B. Since K_{AS} is shared between server and *A*, it ensures *A* that R_B value is from authentic source. A computes session key with *B* as $K_{AB} = (R_B)^{ra} (\bmod p)$ .Using K_{AB} and N_{AB} *A* computes one way hash $H_{K_{AB}}(N_{AB})$ and verifies that *B* posses same key (K_{AB}) as *A*. Using K_{AB}, *A* once again calculates one way hash $H_{K_{AB}}(H_{K_{AB}}(N_{AB}))$ and sends to *B*.

$$A \rightarrow B \quad H_{K_{AB}}(H_{K_{AB}}(N_{AB}))$$

viii. Finally, after receiving this *B* computes this one way hash using K_{AB} and verifies that *A* posses same session key (K_{AB}) as *B*.

The detail is explained in Fig-1.

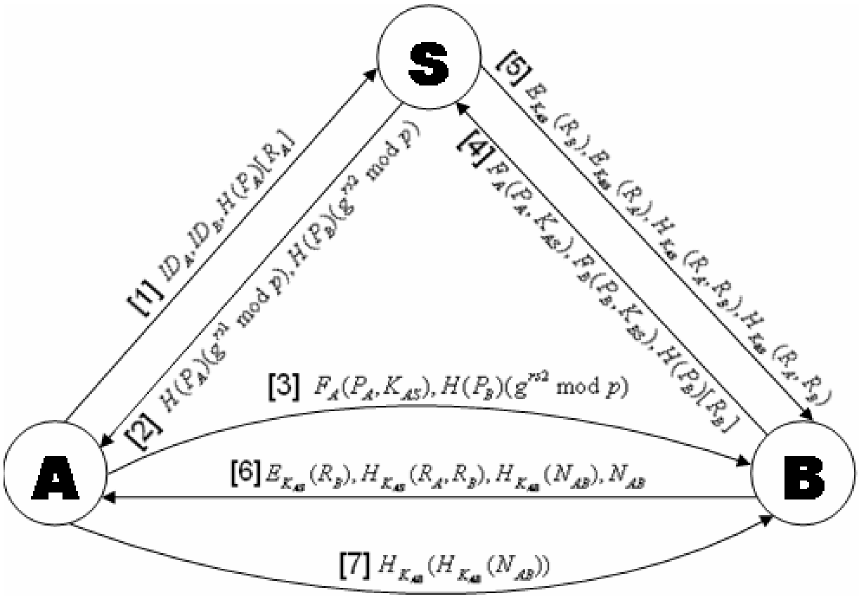

S
A
B
[1] $ID_A, ID_B, H(P_A)[R_A]$
[2] $H(P_A)(g^{rs1} \bmod p), H(P_B)(g^{rs2} \bmod p)$
[3] $F_A(P_A, K_{AS}), H(P_B)(g^{rs2} \bmod p)$
[4] $F_A(P_A, K_{AS}), F_B(P_B, K_{BS}), H(P_B)[R_B]$
[5] $E_{K_{AS}}(R_B), E_{K_{BS}}(R_A), H_{K_{AS}}(R_A, R_B), H_{K_{BS}}(R_A, R_B)$
[6] $E_{K_{AS}}(R_B), H_{K_{AS}}(R_A, R_B), H_{K_{AB}}(N_{AB}), N_{AB}$
[7] $H_{K_{AB}}(H_{K_{AB}}(N_{AB}))$

FIGURE 1: Proposed Protocol.

3.3 Commutative one way hash functions

Both host and server agree upon family of commutative one-way hash functions $\{H_0, H_1, H_2 ....H_N\}$ [14]. Let *H (P)* be defined as $H_0(P)$, a member of a family of Commutative one way hash functions. Host *A* calculates one way hash of its password as $H_0(P_A) = (P_A)^{h_0}(\bmod p)$, where h_0 is a random number. We assume that one way hash of password $H_0(P)$ of every host is distributed to server. Since one way hash is irreversible nobody can compute P from $H_0(P)$. Host *A* calculates its one way function as $F_A(P_A, K_{AS}) = H_{K_{AS}}(P_A) = (P_A^{K_{AS}})(\bmod p)$ and sends to server. Server Knows only one way hash of password P_A i.e. $H_0(P_A)$ using which it calculates predicate function of *A* as $H_{K_{AS}}(H_0(P_A)) = (P_A^{h0})^{K_{AS}}(\bmod p)$. Server computes $H_0(H_{K_{AS}}(P_A)) = (P_A^{K_{AS}})^{h_0}(\bmod p)$. Here $H_{K_{AS}}(P_A) = (P_A^{K_{AS}})(\bmod p)$ is sent by the host. Now server checks $H_{K_{AS}}(H_0(P_A))$ equals $H_0(H_{K_{AS}}(P_A))$ or not. If these two are equal it confirms server that host is genuine. Much better implementation of commutative one way hash functions can be found.

4. SECURITY ANALYSES

In this section, we provide formal security analysis of our protocol. Hosts are not forced to store plain text version of password at server as a result this protocol is not vulnerable to password file compromise attacks [14]. Though $H(P)$ is compromised there is no way to recover P from $H(P)$. Even $H(P)$ is compromised no body can mimic the host to server as only genuine host can compute one way function- $F_A(...), F_B(...)$ etc. Because only host knows password, which is seed for this function. This protocol provides host authentication and server authentication as a result man-in-the middle attacks are averted. Server authentication is done through one time keys it defeats malicious insider attacks [15]. This is a type of attack where a genuine host turns out to be hostile in subsequent protocol run and misuses the information that it has already acquired in previous protocol run.

Online guessing attacks are not possible since R_A, R_B are encrypted with one time keys. Dictionary attacks and offline guessing attacks are not possible since there is no verifiable information present in the protocol run to verify attacker's guess. This protocol also provides perfect forward key secrecy. It also provides Key non-disclosure, Key integrity, and Key confirmation. We use one way hash functions for authentication and key confirmation as conventional encryption and decryption makes protocol design messy [15]. One way hash function suffices decryption. N_{AB} in last step multiplies key space to be searched in case of brute force attack. To guard further against dictionary attacks one way function- $F_A(...), F_B(...)$ may be encrypted with K_{AS}, K_{BS} respectively. Even if *H (P)* is compromised it is equivalent to breaking Diffie-Hellman protocol [8]. Since R_A, R_B are encrypted with $H(P_A)$ and $H(P_B)$ respectively this averts identity mis-binding attacks.

5. CONCLUSION

We propose a three party protocol secure against online and dictionary attacks. It provides host and server authentication. Hosts are not forced to store plain text version of password at server. Proposed protocol does not make use of any public key infrastructure. Instead of commutative one way hash functions digital signatures can also be used for host authentication purpose.

6. ACKNOWLEDGMENT

This research was supported by Department of Communication and Information Technology, Government of India under Information Security Education and Awareness Project and being carried out at department of Computer Science and Engineering, NIT Rourkela, INDIA.

7. REFERENCES

1. Menezes A.,Oorschot P. van and Vanstone S. *"Handbook of Applied Cryptography",* CRC Press, (1996)
2. Schneier Bruce. "*Applied Cryptography: Protocols and Algorithms"*, John Wiley and Sons, (1994)
3. Stallings Williams. *"Cryptography and Network Security"*, 3rd Edition, Pearson Educa- tion, (2004)
4. Mel H.X.,Baker Doris.M. and Burnett Steve. "*Cryptography Decrypted"*, Addison- Wesley, (2004)
5. Bellare Mihir, Rogaway Phillip. *"Provably Secure Session Key Distribution-The Three Party Case"*. In Proceedings of the 27th annual ACM symposium on Theory of computing STOC '95, ACM Press, May 1995
6. L. Gong, M. A. Lomas, R. M. Needham, and J. H. Saltzer, *"Protecting Poorly Chose Secrets From Guessing Attacks".* SELECTED AREAS IN COMMUNICATIONS, vol. 11, no. 5, pp. 648–656, June 1993
7. M. Strangio, *"An Optimal Round Two-Party Password-Authenticated Key Agreement Protocol".* In The First International Conference on Availability, Reliability and Security, p. 8, April 2006
8. W. Diffie and M. Hellman, "*New Directions In Cryptography".* IEEE Transactions on Information Theory IT-11, pp. 644–654, November 1976
9. S. Bellovin and M. Merritt, *"Encrypted Key Exchange: Password Based Protocols Secure Against Dictionary Attacks".* In Proceedings IEEE Symposium on Research in Security and Privacy, pp. 72–84, 1992

10. Y. Her-Tyan and S. Hung-Min, *“Simple Authenticated Key Agreement Protocol Resistant To Password Guessing Attacks”,* ACM SIGOPS Operating Systems Review, vol. 36, no. 4, pp. 14–22, October 2002
11. M. Steiner, G. Tsudik, and M. Waidner, *“Refinement And Extension Of Encrypted Key Exchange”.* ACM Operating System Review, vol. 29, no. 3, pp. 22–30, 1995
12. Y. Ding and P. Horster, *“Undetectable On-Line Password Guessing Attacks”.* ACM Operating System Review, vol. 29, no. 4, pp. 77–86, October1995
13. C. L. Lin, H. M. Sun, and Hwang, *“Three-Party Encrypted Key Exchange: Attacks And A Solution”.* ACM Operating System Review, vol. 34, no. 4, pp. 12–20, October 2000
14. S. Bellovin and M. Merritt, *“Augmented Encrypted Key Exchange: A Password Based Protocols Secure Against Dictionary Attacks And Password File Compromise”.* In 1st ACM Conf. on Computer and Communications Security. ACM Press, pp. 244–250, December 1993
15. T. Gene and H. Van, *“On Simple and Secure Key Distribution”.* In Proceedings of the 1st ACM conference on Computer and communications security CCS 93. ACM Press, pp. 49–57, December 1993s

Active Contours without Edges and Curvature Analysis for Endoscopic Image Classification

B. V. Dhandra dhandra_b_v@yahoo.co.in
Dept. of Computer Science
Gulbarga University
Gulbarga - 585106, INDIA.

Ravindra Hegadi ravindrahegadi@rediffmail.com
Dept. of Computer Science
Gulbarga University
Gulbarga - 585106, INDIA.

Abstract

Endoscopic images do not contain sharp edges to segment using the traditional segmentation methods for obtaining edges. Therefore, the active contours or ‘snakes’ using level set method with the energy minimization algorithm is adopted here to segment these images. The results obtained from the above segmentation process will be number of segmented regions. The boundary of each region is considered as a curve for further processing. The curvature for each point of this curve is computed considering the support region of each point. The possible presence of abnormality is identified, when curvature of the contour segment between two zero crossings has the opposite curvature signs to those of such neighboring contour segments on the same edge contours. The K-nearest neighbor classifier is used to classify the images as normal or abnormal. The experiment based on the proposed method is carried out on 50 normal and 50 abnormal endoscopic images and the results are encouraging.

Keywords: Active Contours, Curvature, Endoscopy, Jacobi method, Level sets.

1. INTRODUCTION

Endoscopy provides images better than that of the other tests, and in many cases endoscopy is superior to the other imaging techniques such as traditional x-rays. A physician may use an endoscopy as a tool for diagnosing the possible disorders in the digestive tract. Symptoms that may indicate the need for an endoscopy include swallowing difficulties, nausea, vomiting, reflux, bleeding, indigestion, abdominal pain, chest pain and a change in bowel habits. In the conventional approach for the diagnosis of endoscopic images the visual interpretation by the physician is employed. The process of computerized visualization, interpretation and analysis of endoscopic images will assist the physician for fast identification of the abnormality in the images [1]. In this direction research works are being carried out for classifying the abnormal endoscopic images based on their properties like color, texture, structural relationships between the image pixels, etc. The method proposed by P.Wang et.al.[2] classifies the endoscopic images based on texture and neural network, where as the analysis of curvature for the edges obtained from the endoscopic images is proposed by Krishnan et.al.[3]. Hiremath et.al.[4] proposed a method to detect the possible presence of abnormality using color segmentation of the images based on 3σinterval [5] for obtaining edges followed by curvature analysis. The watershed segmentation approach for classifying abnormal endoscopic images is proposed by Dhandra et.al.[6]. In this paper the active contours using the level set method with energy minimization approach, which is also known as active contours without edges proposed by chan et.al [7] is adopted for the segmentation of the endoscopic images followed by the curvature computation of the boundary of each obtained region. The zero crossings of the curvature plot for each edge are obtained for further analysis. In the following section we shall discuss the mathematical formulation for level set method and active contours without edges. In Section 3 the curvature analysis is discussed. In Section 4 the K nearest neighborhood classification is discussed. The experimental results are analyzed in Section 5.

2. METHODS

2.1 Mathematical formulations for Level Set Method

Let Ω be a bounded open subset of R^2, with ∂Ω as its boundary. Then a two dimensional image u_0 can be defined as $u_0 : \Omega \rightarrow R$. In this case Ω is just a fixed rectangular grid. Now consider the evolving curve *C* in Ω, as the boundary of an open subset *ω* of Ω. In other words, $\omega \subseteq \Omega$, and *C* is the boundary of *ω* $(C = \partial\omega)$. The main idea is to embed this propagating curve as the zero level set of a higher dimensional function ϕ. We define the function as follows:

$$\varphi(x,y,t=0) = \pm d \qquad (1)$$

where d is the distance from *(x, y)* to *∂ω* at *t* = 0, and the plus (minus) sign is chosen if the point *(x, y)* is outside (inside) the subset *ω*.

Now, the goal is to obtain an equation for the evolution of the curve. Evolving the curve in the direction of its normal amounts to solving the partial differential equation [8]:

$$\frac{\partial \phi}{\partial t} = F\left|\nabla \phi\right|, \ \phi(x,y,t=0) = \phi_0(x,y),$$

where the set $\{(x,y), \phi_0(x,y)=0\}$ defines the initial contour, and *F* is the speed of propagation. For certain forms of the speed function, *F*, the above equation reduces to a standard Hamilton-Jacobi equation. There are several major advantages to this formulation. The first one is that $\phi(x,y,t)$ always remains a function as long as *F* is smooth. As the surface ϕ evolves, the curve *C* may break, merge, and change topology. Second advantage is that geometric properties of the curve are easily determined from a particular level set of the surface ϕ. For example, the normal vector for any point on the curve *C* is given by:

$$\vec{n} = \nabla \phi$$

and curvature *K* is obtained from the divergence of the gradient of unit normal vector to the front:

$$K = div\left(\frac{\nabla \phi}{\left|\nabla \phi\right|}\right) = \frac{\phi_{xx}\phi_y^2 - 2\phi_x\phi_y\phi_{xy} + \phi_{yy}\phi_x^2}{\left(\phi_x^2 + \phi_y^2\right)^{3/2}}$$

The third advantage is that we are able to evolve curves in more than two dimensions. The above formulae can easily be extended for higher dimensions. This is useful in propagating a curve to segment large volume of data.

2.2 Active Contours using level set method without edges

The curve *C* can be considered as the boundary of the *ω* and the region *ω* can be denoted by *inside(C)* and the region $\Omega \setminus \overline{\omega}$ by *outside(C)*. The energy minimization approach proposed by Chan et.al.[7] is adopted for segmentation is as follows:

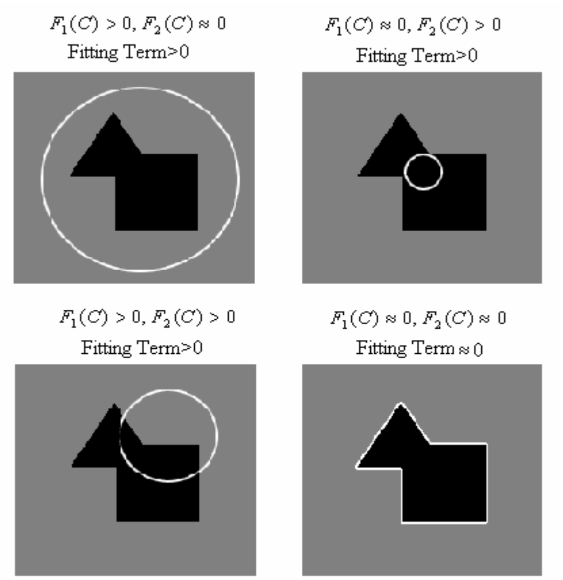

$F_1(C) > 0, F_2(C) \approx 0$
Fitting Term>0
$F_1(C) \approx 0, F_2(C) > 0$
Fitting Term>0
$F_1(C) > 0, F_2(C) > 0$
Fitting Term>0
$F_1(C) \approx 0, F_2(C) \approx 0$
Fitting Term ≈ 0

FIGURE 1: All possible cases in position of the curve

Consider a simple case where the image *u₀* is formed by two regions of piecewise constant intensity. Denote the intensity values by u_0^0 and u_0^1. Further, assume that the object to be detected has a region whose boundary is C_0 and intensity u_0^1. Then *inside(C_0)*, the intensity of u_0 is approximately u_0^1, whereas *outside(C_0)* the intensity of u_0 is approximately u_0^0. Then consider the fitting term:

$$F_1(C) + F_2(C) = \int_{inside(c)} \left| u_0(x,y) - c_1 \right|^2 dxdy + \int_{outside(c)} \left| u_0(x,y) - c_2 \right|^2 dxdy$$

where the constants c_1 and c_2 are the averages of u_0 inside and outside the curve *C* respectively.

Consider Fig. 1. If the curve *C* is outside the object, then $F_1(C) > 0, F_2(C) \approx 0$. If the curve is inside the object, then $F_1(C) \approx 0, F_2(C) > 0$. If the curve is both inside and outside the object, then $F_1(C) > 0, F_2(C) > 0$. However, if the curve *C* is exactly on our object boundary C_0, then $F_1(C) \approx 0, F_2(C) \approx 0$, and our fitting term is minimized.

Some additional regularization terms of Mumford-Shah segmentation model [9] are considered for effective energy minimization. Therefore we will also try to minimize the length of the curve and the area of the region inside the curve. So we introduce the energy function *E*:

$$E(C, c_1, c_2) = \mu \cdot length(C) + \nu \cdot Area(inside(C)) + \lambda_1 \cdot \int_{inside(c)} \left| u_0(x,y) - c_1 \right|^2 dxdy$$
$$+ \lambda_2 \cdot \int_{outside(c)} \left| u_0(x,y) - c_2 \right|^2 dxdy$$

where $\mu \geq 0, \nu \geq 0, \lambda_1 > 0, \lambda_2 > 0$ are fixed parameters. Thus, the objective is to find *C, c₁, c₂* such that *E(C, c₁, c₂)* is minimized. Mathematically, solve:

$$\inf_{C, c_1, c_2} E(C, c_1, c_2)$$

This problem can be formulated using level set method as follows. The evolving curve *C* can be represented by the zero level set of the signed distance function ϕ as in (1). So we replace the unknown variable *C* by ϕ. Now consider the Heaviside function *H* and Dirac measure *δ*:

$$H(z) = \begin{cases} 1, \ if \ \ z \geq 0 \\ 0, \ if \ \ z < 0 \end{cases}, \qquad \delta(z) = \frac{d}{dz} H(z)$$

We can rewrite the length of ϕ *= 0* and the area of the region *inside(ϕ = 0)* using these functions. The Heaviside function is positive inside our curve and zero elsewhere, so the area of the region is just the integral of the Heaviside function of ϕ. The gradient of the Heaviside function defines our curve, so integrating over this region gives the length of the curve. Mathematically:

$$Area(\phi = 0) = \int_\Omega H(\phi(x, y))dxdy$$

$$Length(\phi = 0) = \int_\Omega \left|\nabla H(\phi(x, y))\right|dxdy = \int_\Omega \delta(\phi(x, y))\left|\nabla \phi(x, y)\right|dxdy$$

Similarly, we can rewrite the previous energy equations so that they are defined over the entire domain Ω rather than separated into *inside(C) =ϕ > 0* and *outside(C) = ϕ < 0*:

$$\int_{\phi>0} \left|u_0(x, y) - c_1\right|^2 dxdy = \int_\Omega \left|u_0(x, y) - c_1\right|^2 H(\phi(x, y))dxdy$$

$$\int_{\phi<0} \left|u_0(x, y) - c_2\right|^2 dxdy = \int_\Omega \left|u_0(x, y) - c_2\right|^2 (1 - H(\phi(x, y)))dxdy$$

Therefore our energy function *E(C,c₁,ϕ)* can be written as:

$$\begin{aligned} E(C, c_1, c_2) = \mu \int_\Omega \delta(\phi(x, y))\left|\nabla \phi(x, y)\right|dxdy + \nu \int_\Omega H(\phi(x, y))dxdy \\ + \lambda_1 \int_\Omega \left|u_0(x, y) - c_1\right|^2 H(\phi(x, y))dxdy \\ + \lambda_2 \int_\Omega \left|u_0(x, y) - c_2\right|^2 (1 - H(\phi(x, y)))dxdy \end{aligned} \quad (2)$$

The constants *c₁, c₂* are the averages of *u₀* in $\phi \geq 0$ and $\phi < 0$ respectively. So they are easily computed as:

$$c_1(\phi) = \frac{\int_\Omega u_0(x, y)H(\phi(x, y))dxdy}{\int_\Omega H(\phi(x, y))dxdy} \quad (3)$$

and

$$c_2(\phi) = \frac{\int_\Omega u_0(x,y)(1-H(\phi(x,y)))dxdy}{\int_\Omega (1-H(\phi(x,y)))dxdy} \qquad (4)$$

Now we can deduce the Euler-Lagrange partial differential equation from (2). We parameterize the descent direction by $t \geq 0$, so the equation $\phi(x,y,t)$ is:

$$\frac{\partial \phi}{\partial t} = \partial(\phi)\left[\mu\, div\left(\frac{\nabla \phi}{|\nabla \phi|}\right) - \nu - \lambda_1(u_0 - c_1)^2 + \lambda_2(u_0 - c_2)^2\right] = 0 \qquad (5)$$

In order to solve this partial differential equation, we first need to regularize *H(z)* and *δ(z)*. Chan and Vese [7] propose:

$$H_\varepsilon(z) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{z}{\varepsilon}\right)$$

Implying that *δ(z)* regularizes to:

$$\delta_\varepsilon(z) = \frac{1}{\pi} \cdot \frac{\varepsilon}{\varepsilon^2 + z^2}$$

It is easy to see that as $\varepsilon \to 0$, $H_\varepsilon(z)$ converges to *H(z)* and $\delta_\varepsilon(z)$ converges to $\delta(z)$. Authors claim that with these regularizations, the algorithm has the tendency to compute a global minimizer. Chan and Vese [7] give the following discretization and linearization of (5):

$$\begin{aligned}\frac{\phi_{i,j}^{n+1} - \phi_{i,j}^{n}}{\Delta t} = \delta_\varepsilon\left(\phi_{i,j}^{n}\right)\Bigg[& \frac{\mu}{h^2}\Delta_-^x \cdot \left(\frac{\Delta_+^x \phi_{i,j}^{n+1}}{\sqrt{\left(\Delta_+^x \phi_{i,j}^{n}\right)^2 /(h^{2)} + \left(\phi_{i,j+1}^{n} - \phi_{i,j-1}^{n}\right)^2 /(2h)^2}}\right) \\ & + \frac{\mu}{h^2}\Delta_-^y \cdot \left(\frac{\Delta_+^y \phi_{i,j}^{n+1}}{\sqrt{\left(\phi_{i+1,j}^{n} - \phi_{i-1,j}^{n}\right)^2 /(2h)^2 + \left(\Delta_+^y \phi_{i,j}^{n}\right)^2 /(h^2)}}\right) \\ & - \nu - \lambda_1\left(u_{0,i,j} - c_1(\phi^n)\right)^2 + \lambda_2\left(u_{0,i,j} - c_2(\phi^n)\right)^2 \Bigg]\end{aligned} \qquad (6)$$

where the forward differences of $\phi_{i,j}^{n}$ are calculated as:

$$\Delta_-^x \phi_{i,j} = \phi_{i,j} - \phi_{i-1,j}, \quad \Delta_+^x \phi_{i,j} = \phi_{i+1,j} - \phi_{i,j},$$
$$\Delta_-^y \phi_{i,j} = \phi_{i,j} - \phi_{i,j-1}, \quad \Delta_+^y \phi_{i,j} = \phi_{i,j-1} - \phi_{i,j}$$

This linear system also depends on the forward differences of $\phi_{i,j}^{n+1}$, which is an unknown. However these can be solved using the Jacobi method [10]. In practice, the number of iterations until convergence was found to be small. The segmentation algorithm is then given by:

```
1. Initialize $\phi^0$ by $\phi_0$, n=0
2. For fixed number of iterations do
   2.1 Compute $c_1(\phi^n)$ and $c_2(\phi^n)$ using (3) and (4)
   2.2 Estimate forward differences of $\phi^{n+1}$ using Jacobi method
   2.3 Compute $\phi^{n+1}$ using (6)
3. End.
```

ALGORITHM 1: Energy minimization with Jacobi method.

3. CURVATURE COMPUTATION

The result of the method proposed in Section 2 will generate a number of regions. The boundary of each region is considered for the curvature computation. Due to the discrete boundary representation and quantization errors, false local concavities and convexities along a contour are formed. This noisy nature of binary contours must be taken into account to obtain reliable estimates of contour curvature. Hence, a Gaussian filter is used to smooth the contour points to reduce the noise effect [11]. However, the width of Gaussian filter, *w*, that controls the degree of smoothing has to be chosen suitably. A large value of *w* will remove all small details of the contour curvature, while a small value will permit false concavities and convexities to remain in the contour, thus enforcing an appropriate choice of *w*. To overcome this problem a support region is employed which will dynamically determine the parameter of the Gaussian filter.

3.1 Determination of Support Region

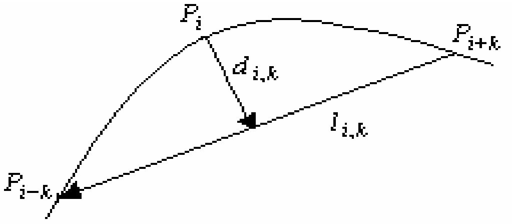


FIGURE 2: Representation of Support Region.

The support region concept can be explained using the Fig. 2. The support region for each point on the curve is the number of points obtained from the implementation of the Algorithm 2:

1. *Determine the length of the chord joining the points* $P_{i-k,}P_{i+k}$ *as*

$$l_{i,k} = \left|\overline{P_{i-k}\,P_{i+k}}\right|$$

2. *Let $d_{i,k}$ be the perpendicular distance from $P_{i,}$ to the line joining $\overline{P_{i-k}P_{i+k}}$, start with k=1, compute $l_{i,k}$ and $d_{i,k}$ until one of the following conditions hold:*

   a. $l_{i,k} \ge l_{i,k+1}$

   b. $\frac{\mathrm{d}_{\mathrm{i,k}}}{\mathrm{l}_{\mathrm{i,k}}} \ge \frac{d_{i,k+1}}{l_{i,k+1}}$ *for* $d_{i,k} \ge 0$

3. *Now, the region of support of P_i is the set of points satisfying either condition (a) or condition (b), that is,*

$$D(P_i) = \left\{P_{i-k}, \ldots\ldots, P_{i-1}, P_{i,} P_{i+1,} \ldots\ldots, P_{i+k} \mid condition\ (a)\ or\ (b)\right\}$$

ALGORITHM 2: Determination of support region

3.2 Gaussian Smoothing

A planar curve can be defined in parametric form as *(x(t), y(t))* ∈ *R^2*, where t is the path length along the curve. Smoothing is performed by the convolution of *x(t)* and *y(t)* with the Gaussian filter. A one dimensional Gaussian filter is defined as

$$\eta(t,w) = \frac{1}{\sqrt{2\pi w^2}} e^{\left(-t^2 / 2w^2\right)}$$

where *w* is the width of the filter, which needs to be determined. The smoothed curve is denoted by set of points *(X(t,w), Y(t,w))*, where,

$$x(t,w) = x(t) \otimes \eta(t,w), \quad y(t,w) = y(t) \otimes \eta(t,w)$$

where $\otimes$ denotes the convolution.

The measurement of the curvature of the point is based on the local properties within its region of support, and the length of Gaussian smooth filter is proportional to the region of support [12]. This implies that the neighboring points closer to the point of interest should have higher weights than those points further away. This method is less sensitive to noise. The Gaussian filter applied here will have the following window length and width [3]:

Window Len =2xSupp. Region D(P_i)+1,
Width w =Support Region D(P_i) / 3

Further, the curvature for each point on the curve is calculated in the following way.

3.3 Curvature Computation

For the continuous curve *C*, expressed by *{x(s), y(s)}*, where s is the arc length of the edge point, the curvature can be expressed as:

$$k(s) = \frac{\dot{x}\ddot{y} - \ddot{x}\dot{y}}{\left(\dot{x}^2 + \dot{y}^2\right)^{3/2}} \tag{7}$$

where $\dot{x} = dx/ds$, $\ddot{x} = d^2x/ds^2$, $\dot{y} = dy/ds$, $\ddot{y} = d^2y/ds^2$.

For digital implementation, the coordinate functions *x(s)* and *y(s)* of the curvature are represented by a set of equally spaced Cartesian grid samples. The derivatives in the equation (7) are calculated by finite differences as:

$$\begin{aligned} \dot{x}_i &= x_i - x_{i-1}, \ \ddot{x}_i = x_{i-1} - 2x_i + x_{i+1}, \\ \dot{y}_i &= y_i - y_{i-1}, \ \ddot{y}_i = y_{i-1} - 2y_i + y_{i+1} \end{aligned} \tag{8}$$

The algorithm for curvature computation is presented below.

1. *Determine edge contours in the input image using active contours without edges.*
2. *Select a large edge contour for further processing.*
3. *Determine the support region for each contour point.*
4. *Smooth the contour by a Gaussian filter with the width proportional to the support region.*

```
5. Compute the curvature for each point on the Gaussian smoothed
   curve using equation (7) and (8).
```

ALGORITHM 3: Curvature computation

In the process of curvature computation we come across with two special conditions for which the alternate solutions need to be given. They are:

1. When the edge point is on a straight line, the curvature for that point is assigned to zero.
2. When the support region for an edge point is 1, this point will not be smoothed. So, the smoothing on this point is performed using the following equation:

$$(\hat{x}_{i,} \hat{y}_i) = \frac{1}{2}(x_i, y_i) + \frac{1}{4}[(x_{i-1}, y_{i-1}) + (x_{i+1}, y_{i+1})]$$

where, $(\hat{x}_{i,} \hat{y}_i)$ is the smoothed point of (x_i, y_i).

4. K-NEAREST NEIGHBOR CLASSIFIER

The nearest neighbor and K-nearest neighbor classifiers are applied on two parameters namely the number of regions obtained from the segmentation process and the total number of zero crossings of curvature plot of every region edge to classify the images as either normal or abnormal. Basically this classifier finds the closest training point to the unknown point and predicts the category of that training point for this unknown point. The experiment is carried out by varying the number of neighbors (K=3, 5, 7). Performance of the algorithm is optimal when K=3.

5. EXPERIMENTAL RESULTS

For the experimentation 50 normal and 50 abnormal endoscopic color images of size 128X128 pixels are chosen. The parameters are set to the values as: $\lambda_1 = \lambda_2 = 1$, *h =1*, and $\Delta t = 0.1$. Varying the value of ν generally had little or no effect on the segmentation results. The ε parameter for the regularized Heaviside and Dirac functions was set to 1 as suggested by Chan and Vese [7]. The length parameter μ is fixed to a small value (i.e. $\mu \approx 0.0001 \cdot 255^2$) .The detection of possible presence of abnormality is performed by analyzing the curvature change along boundary of each region, which is considered as a edge contour. The curvature of each edge point on the edge contour is computed using the Algorithm 3. Two thresholds, c_{th} and $n_{th,}$ are used in the analysis. c_{th} is the curvature threshold value, and n_{th} is number of edge points in a segment. Along the edge contour, if the absolute curvature of the point is bigger than c_{th}, the point counting starts until the absolute curvature value of the point is less than c_{th}. If the point count is bigger than n_{th}, an edge segment is formed. The possible presence of abnormality in the image is detected when the curvature of a segment has opposite sign to those of such neighboring segments on the same edge contour. Also such a segment is bounded by two significant zero crossings. The stages in the proposed method for the abnormal images and normal images are shown in the Fig. 3 and 4 respectively. In these figures image [A] is abnormal endoscopic color image, [B] shows regions obtained after the segmentation of images using active contours without edges. Its binary image is shown in image [C]. A largest edge contour obtained is shown in image [D]. Its curvature profile is shown in image [E]. In the abnormal image shown in Fig. 3, the number of regions formed is 12, and the number of zero crossings obtained for a large edge contour is 137, where as for the normal image shown in Fig. 4, the number of regions formed is 5, and the number of zero crossings obtained for a large edge contour is 24. Fig. 5 and Fig. 6 show the endoscopic images with their curvature plot for their largest edges. The original abnormal and normal endoscopic images are shown in the first column and the corresponding curvature profile for the largest edge segment is shown in the second column. From these figures, considerably less number of zero crossings for the normal image edge curvatures can be observed as opposed to the number of zero crossings of edge curvatures in the abnormal images.

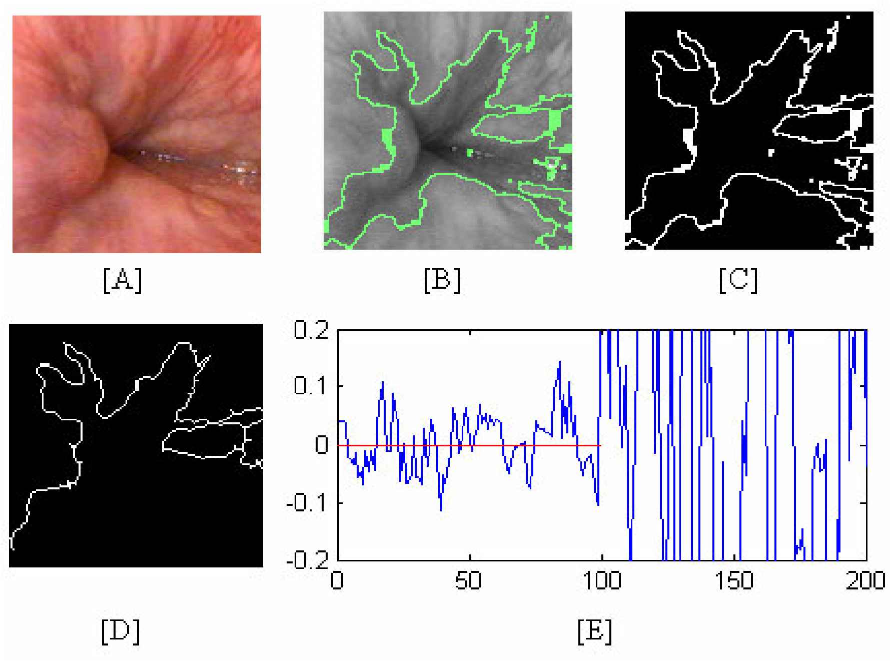


FIGURE 3: The proposed method for Abnormal Image.

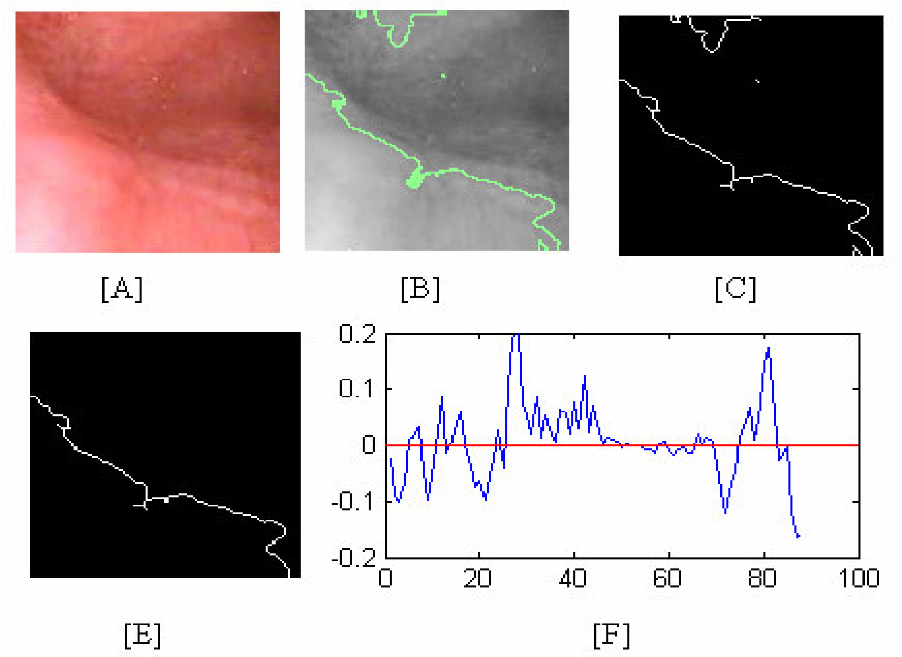


FIGURE 4: The proposed method for Normal Images.

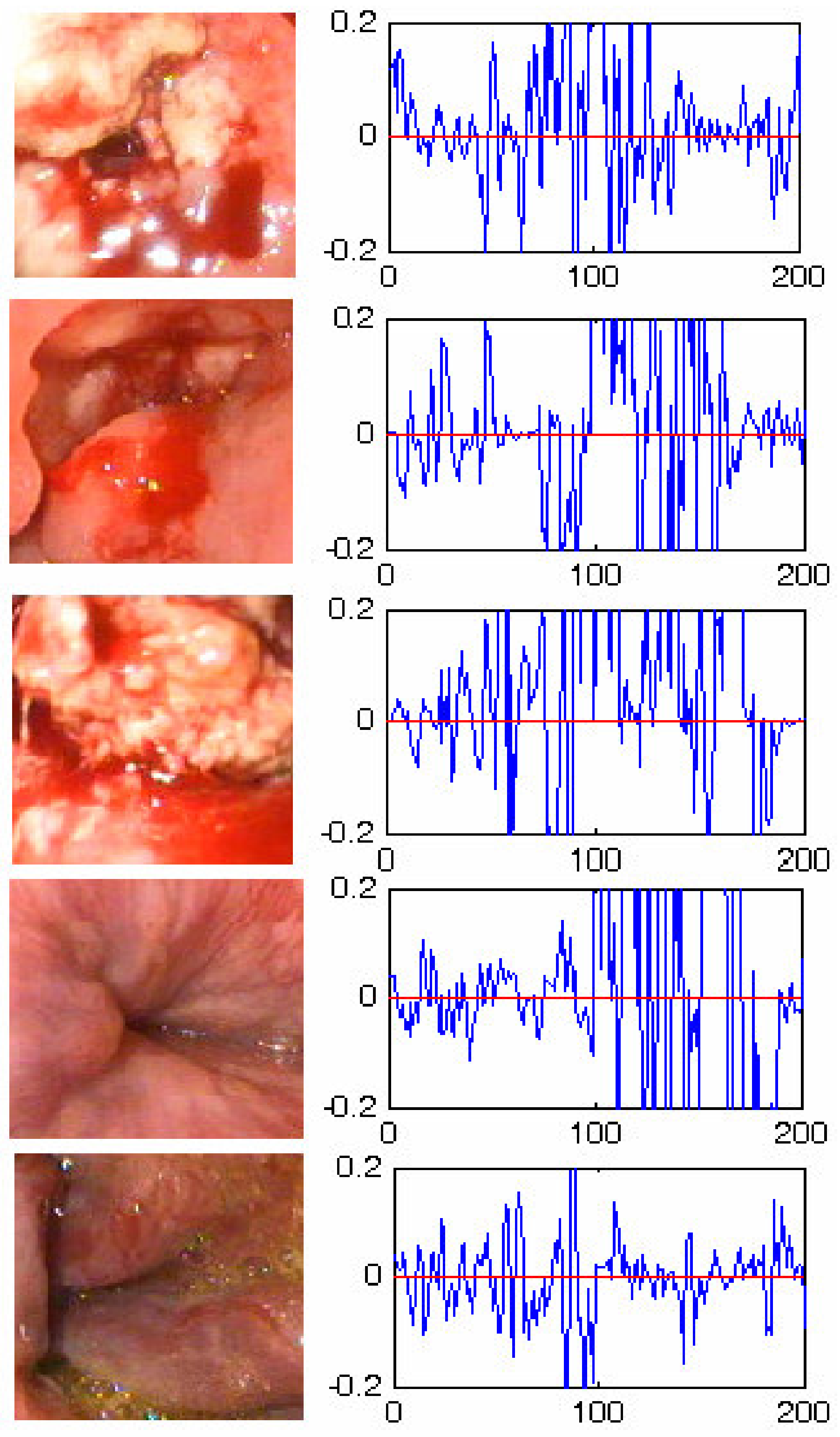


FIGURE 5: Curvature profile for Abnormal Images.

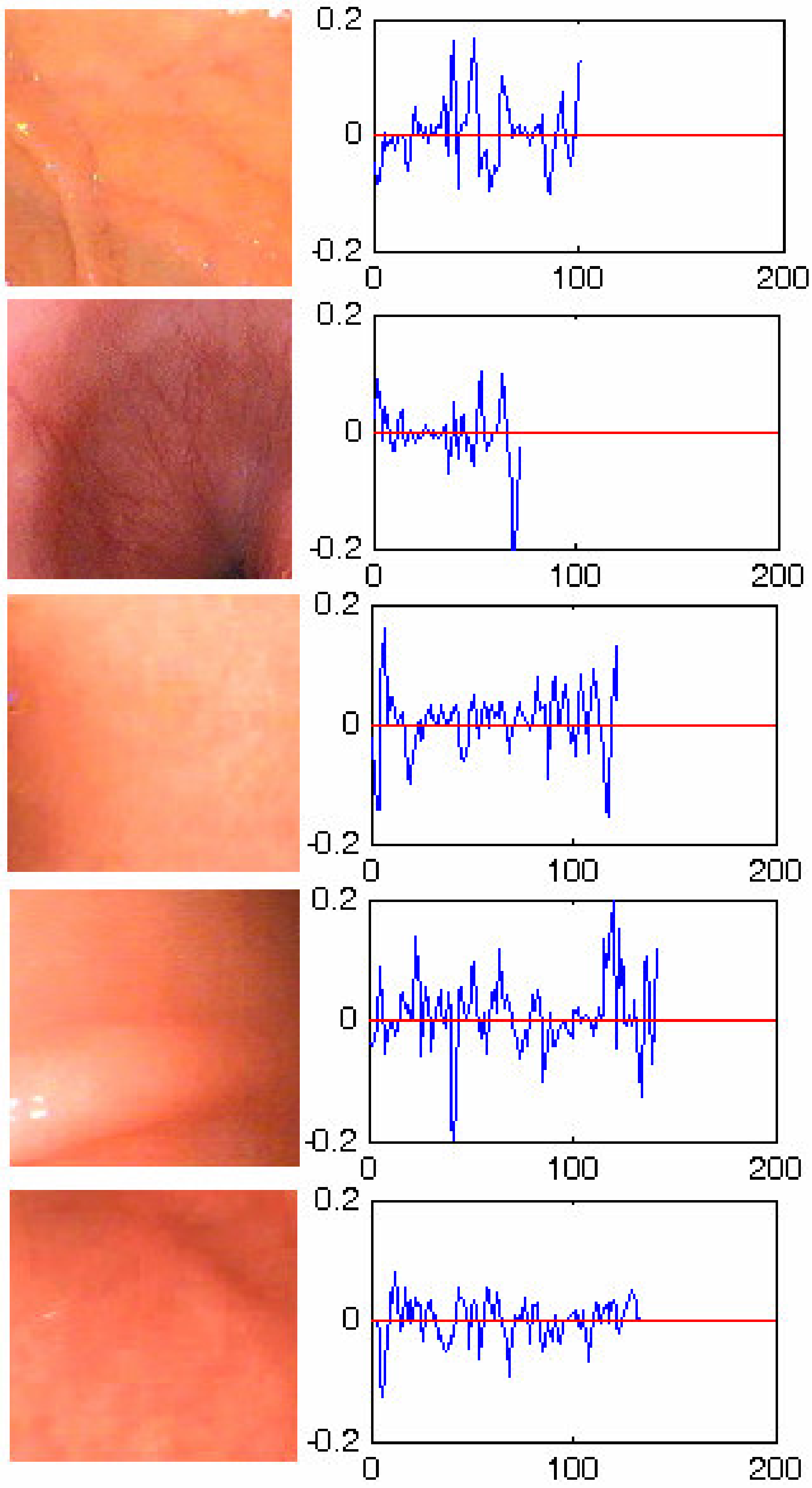


FIGURE 6: Curvature profile for Normal Images.

Abnormal Images			Normal Images		
Image ID	Regions	Zero Crossings	Image ID	Regions	Zero Crossings
Abn01	31	164	Nor01	5	89
Abn02	29	105	Nor02	9	29
Abn03	14	106	Nor03	4	69
Abn04	18	160	Nor04	9	81
Abn05	8	122	Nor05	2	53
Abn06	29	155	Nor06	4	78
Abn07	25	156	Nor07	9	48
Abn08	26	157	Nor08	5	53
Abn09	14	111	Nor09	5	70
Abn10	10	115	Nor10	2	69
Abn11	9	91	Nor11	2	40
Abn12	17	161	Nor12	4	89
Abn13	12	145	Nor13	3	50
Abn14	38	136	Nor14	2	38
Abn15	11	107	Nor15	2	72
Abn16	11	118	Nor16	2	70
Abn17	14	94	Nor17	6	64
Abn18	13	100	Nor18	2	42
Abn19	11	124	Nor19	9	143
Abn20	11	202	Nor20	3	62
Abn21	18	176	Nor21	5	57
Abn22	20	178	Nor22	2	36
Abn23	18	99	Nor23	8	96
Abn24	14	148	Nor24	6	87
Abn25	29	168	Nor25	4	59
Abn26	16	131	Nor26	4	51
Abn27	16	139	Nor27	4	74
Abn28	10	108	Nor28	2	78
Abn29	28	252	Nor29	3	60
Abn30	24	206	Nor30	3	66
Abn31	19	131	Nor31	3	66
Abn32	18	97	Nor32	4	51
Abn33	21	148	Nor33	2	29
Abn34	27	117	Nor34	7	141
Abn35	40	153	Nor35	2	34
Abn36	20	130	Nor36	3	59
Abn37	10	158	Nor37	2	83
Abn38	10	109	Nor38	2	54
Abn39	16	159	Nor39	5	70
Abn40	13	147	Nor40	4	71
Abn41	26	151	Nor41	7	156
Abn42	11	142	Nor42	2	67
Abn43	22	168	Nor43	3	53
Abn44	17	213	Nor44	2	48
Abn45	11	151	Nor45	4	138
Abn46	15	167	Nor46	4	27
Abn47	14	117	Nor47	5	128
Abn48	33	150	Nor48	3	61
Abn49	27	138	Nor49	3	56
Abn50	31	141	Nor50	5	71

TABLE 1: Number of Regions and Zero Crossings for Normal and Abnormal Images.

Table 1 show the number of regions obtained after the segmentation by the active contours without edges and total number of zero crossings for all the edges obtained in the segmented image. The results are shown for the 50 normal and 50 abnormal test images. In the results it can be seen that the proposed segmentation for the abnormal images generate large number of regions when compared to the normal images. We can notice that few normal images have generated large number of segments and zero crossings. It is due to the presence of noise such as lumen regions and bright spots generated by the reflection of light sources.

The following Table 2 shows the classification results using the nearest neighbor and K nearest neighbor classifiers.

Image Type	NN Classifier	KNN Classifier
Abnormal Images	96%	98%
Normal Images	90%	90%

TABLE 2: Classification Results.

6. CONCLUSION

The proposed segmentation method is based on the active contours without edges using Mumford-Shah [9] technique and it does not rely on the boundaries defined by the gradients. Also it starts with one initial curve and splits itself to detect the interior curves forming number of regions depending on the smoothness of the surface. The boundary curvature depends on the roughness of the image surface. It generates less number of zero crossings for the smooth normal images where as large number of zero crossings for relatively rough texture in abnormal images. Results obtained from the NN and KNN classifier are quite encouraging. The other features of endoscopic images such as color can be used in future to improve the classification results.

7. REFERENCES

[1] http://digestive.healthcentersonline.com/digestiveimagingtest/endoscopy.cfm

[2] P. Wang, S. M. Krishnan, C. Kugean, M.P. Tjoa, *"Classifiation of Endoscopic images based on Texture and Neural Network"*, In Proceedings of the 23rd Annual EMBS International Conference, October 25-28, Intanbul, Turkey

[3] S. M. Krishnan, X. Yang, K. L. Chan, S. Kumar, P. M. Y. Goh, *"Intestinal Abnormality Detection from Endoscopic Images"*, In Proceedings of 20th Annual International Conference of IEEE EMBS 98, Hongkong, 1998

[4] P.S.Hiremath, B.V.Dhandra, Ravindra Hegadi, G.G.Rajput, *"Abnormality detection in endoscopic images using color segmentation and curvature computation"*, In Proceedings of 11th International Conference on Neural Information Processing, ICONIP-2004, ISI, Calcutta, India, LNCS, ISBN-3-540-23931-6, Springer-Verlag, 2004

[5] P. S. Hiremath, B.V. Dhandra, Iranna Humnabad, Ravindra Hegadi, G.G. Rajput, *"Detection of esophageal Cancer (Necrosis) in the Endoscopic images using color image segmentation",* In Proceedings of second National Conference on Document Analysis and Recognition (NCDAR-2003), Mandya, India, 2003

[6] B.V.Dhandra, Ravindra Hegadi, *“Classification of Abnormal Endoscopic Images using Morphological Watershed Segmentation”,* In Proceedings of International Conference on Cognition and Recognition (ICCR-2005), Mysore, India, 2005

[7] Tony F. Chan, Luminita A. Vese, *“Active Contours without Edges”*, IEEE Transactions on Image Processing, 10(2), 2001

[8] S. Osher, J. A. Sethian, *“Front propagating with curvature dependent speed: Algorithm based on Hamilton-Jacobi formulation”*, Journal of Computational Physics, 79:12-49, 1988

[9] D. Mumford, J. Shah, *“Optimal approximation by piecewise smooth functions and associated variational problems”*, Communications on Pure and Applied Mathematics, 42:577-685,1989

[10] Eric W. Weisstein, *“Jacobi Method, Technical Report”*

[11] V. Torre, T. A. Poggio, *“On Edge Detection*”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 8:147-163, 1986

[12] N. Ansari, K.W. Huang, *“Non-parametric Dominant Point Detection”*, Pattern Recognition, 24:849-862,1991

Simple Encryption/Decryption Application

Majdi Al-qdah majdi.qdah@mmu.edu.my
Faculty of Information Technology
Multimedia University
Cyberjaya, 63100, Malaysia

Lin Yi Hui
Faculty of Information Technology
Multimedia University
Cyberjaya, 63100, Malaysia

Abstract

This paper presents an Encryption/Decryption application that is able to work with any type of file; for example: image files, data files, documentation files…etc. The method of encryption is simple enough yet powerful enough to fit the needs of students and staff in a small institution. The application uses simple key generation method of random number generation and combination. The final encryption is a binary one performed through rotation of bits and XOR operation applied on each block of data in any file using a symmetric decimal key. The key generation and Encryption are all done by the system itself after clicking the encryption button with transparency to the user. The same encryption key is also used to decrypt the encrypted binary file.

Keywords: Encryption, Decryption, Key, Rotation, XOR.

1. INTRODUCTION

Encryption is the most effective way to achieve data security. The process of Encryption hides the contents of a message in a way that the original information is recovered only through a decryption process [1]. The purpose of Encryption is to prevent unauthorized parties from viewing or modifying the data [2]. Encryption occurs when the data is passed through some substitute technique, shifting technique, table references or mathematical operations. All those processes generate a different form of that data. The unencrypted data is referred to as the plaintext and the encrypted data as the ciphertext, which is representation of the original data in a difference form [2].

Key-based algorithms use an Encryption key to encrypt the message. There are two general categories for key-based Encryption: Symmetric Encryption which uses a single key to encrypt and decrypt the message and Asymmetric Encryption which uses two different keys – a public key to encrypt the message, and a private key to decrypt it. Currently, there are several types of key based Encryption algorithms such as: DES, RSA, PGP, Elliptic curve, and others but all of these algorithms depend on high mathematical manipulations [4, 5].

One simple and good way to encrypt data is through rotation of bits or sometimes called bit shifting. But, rotation of bits is more advanced than simple bit shifting. In rotation of bits operation, the bits are moved, or shifted, to the left or to the right. The different kinds of shifts typically differ in what they do with the bits [6]. Another way is to perform logical operation on the bits of the file such as XOR operation. The idea behind XOR Encryption is that it is impossible to reverse the operation without knowing the initial value of one of the two arguments [7].

2. PRELIMINARIES

There are several kinds of Encryption software in the market categorized by their functions and target groups. For example, some are single Encryption applications for files and database security; some are for messenger security or email Encryption applications that hide the actual text in the medium between the sender and the receiver [8]. One of the first types of Encryption was made by Julius Caesar. [9]. in his system, Caesar wrote B instead of A and C instead of B – so to a sentence “ABC” will be written in “BCD” [6].

dsCrypt is AES/Rijndael file Encryption software with simple, multi-file, drag-and-drop operations. It features optimal implementation, performance and safety measures. dsCrypt uses an advanced Encryption algorithm and offers unique options for enhanced security[10].

NeoCrypt is a free, open-source File Protection Utility for Windows. It helps to protect sensitive information easily by encrypting it with password or key. It yields fast, reliable and unbreakable Encryption and supports many popular Encryption algorithms. All types of files can be encrypted like Audio, Video, Documents and Executables programs [11].

Neekprotect is a software in the market right now with the ability to make Encryption on any files in the window platform, a key is set when one try to encrypt a files and the key will be used again when some one else trying to open the files been decrypted through decryption on the certain files [12].NeekProtect is a good software operated under Microsoft window because of the flexibility of this program’s advanced features integration such as double click, file icons, .npt file extension etc.

This paper reports on a similar encryption technique that uses binary rotation of bits with XOR logical operation using a custom made encryption key that operates on any type of a file.

3. METHODOLOGY

The main feature of the encryption/decryption program implementation is the generation of the encryption key. Other features are related to the design of the GUI, progress of encryption details, and user notification of the status of encryption.

3.1 Key Generation

A symmetric Encryption key is used for this application, which means the same key is shared for both Encryption and decryption. A copy of the generated key is saved in a file named .ekf during the Encryption process and the same key is used as the decryption key to retrieve the encrypted file. The technique of generating the key uses two methods: random number generation and combination. First, a long number with only digit values called A is generated. Then another long number with character values called B is generated. The size of B is twice the size of A. Then an insertion operation is performed such that each digit of A is inserted after two characters of B. The result of the insertion is called C. Then another only digit number called D is randomly generated. C is combined with D by placing alternately one character or digit from C after a character or a digit from D. The result of the combination is a relatively strong key. Then, an odd and even partitioning is performed on the key. The position of each character in the key decides it to be an even or an odd character. For example, the character at position 0 is an even one while the character at position 1 is an odd character. Similarly, position 2 is an even position while position 3 is odd one. The even part of the key is combined together and the odd part of the key is combined together. Finally, the two parts of the key are joined as an even part followed by an odd part to produce one final encryption key. Since the final key is a key that consists of all characters, another key with only ASCII values of each character is obtained. The result is a very long decimal key. Figure 1 shows the complete key generation process.

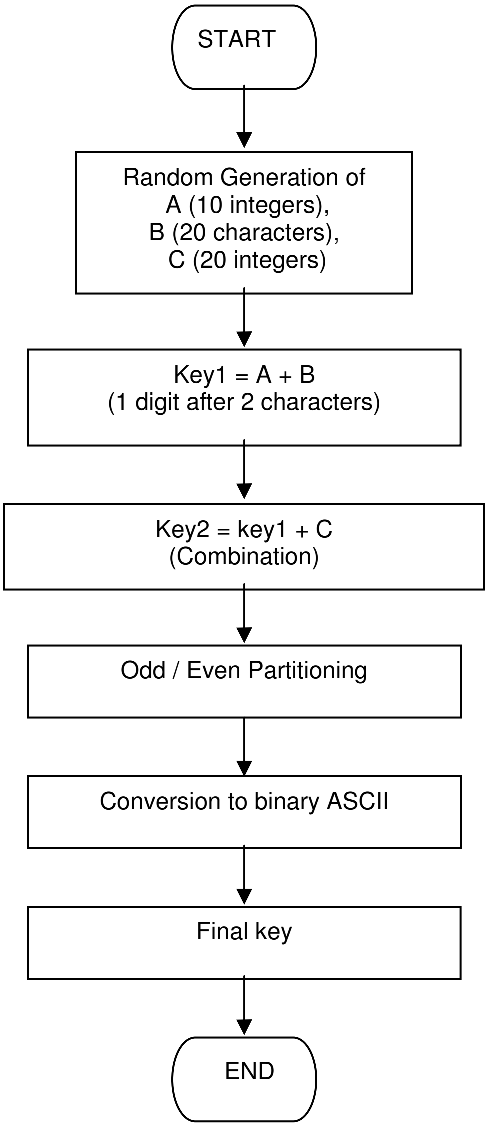


FIGURE 1: Key Generation Flow Chart.

3.2 Encryption Rules

For the Encryption method, a single digit in the decimal ASCII representation will decide which Encryption method is to be applied to the single a binary block in a file.

I. 0 in the key means a rotation of bit to the left is performed and the next integer to 0 in the ASCII code decides how many bits the block will be moved to the left.
II. 1 in the key means a rotation of bit to the right is performed and the next integer to 1 in the ASCII code decides how many bits the block will be moved to the right.
III. 2 means the block will be passed to an XOR Encryption to be performed with a binary block from the file.
IV. Else, all other numbers in the key like 3,4,5,6,7,8,9 are ignored.

The above rules are summarized in the encryption chart shown in figure 2.

3.3 File Types

There are no limitations of the type of files accepted for encryption in this application, which means any type of a file such as data files, audio files, video files or image files can be encrypted by the application. This is because all the files are encrypted at the binary level. There is also no limitation of the size of the file that can be encrypted using this application, which provides flexibility to the user. The encrypted file can only be opened and viewed after it has been decrypted to its original file using the symmetric encryption key.

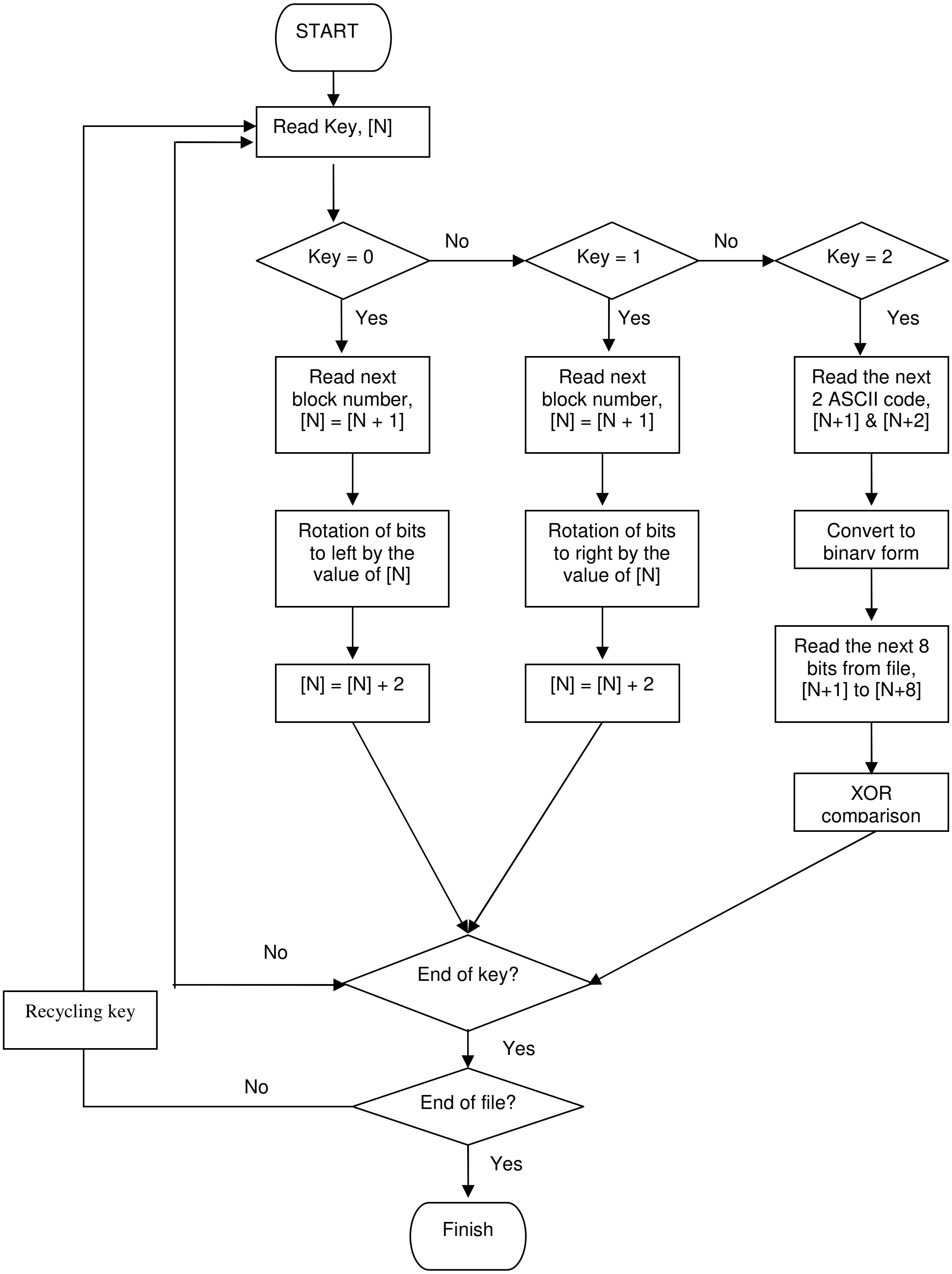
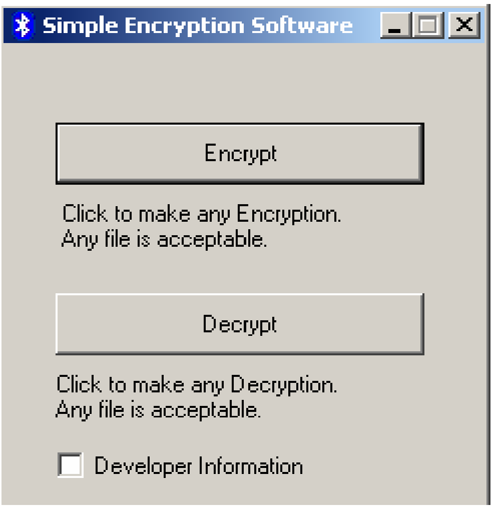

START
Read Key, [N]
Key = 0
No
Key = 1
No
Key = 2
Yes
Yes
Yes
Read next block number, [N] = [N + 1]
Read next block number, [N] = [N + 1]
Read the next 2 ASCII code, [N+1] & [N+2]
Rotation of bits to left by the value of [N]
Rotation of bits to right by the value of [N]
Convert to binary form
[N] = [N] + 2
[N] = [N] + 2
Read the next 8 bits from file, [N+1] to [N+8]
XOR comparison
No
End of key?
Recycling key
Yes
No
End of file?
Yes
Finish

FIGURE 2: Encryption Flow Chart.

4. RESULTS

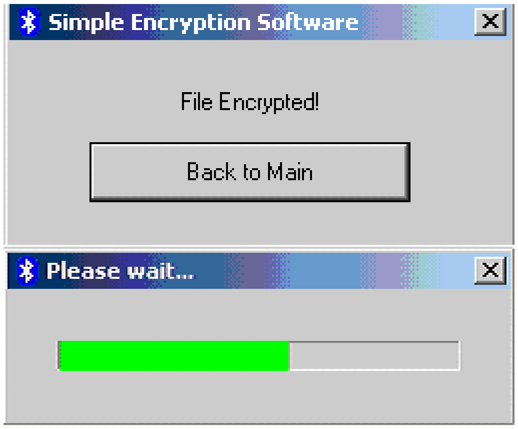
The interface of the application is simple enough to be used by any user. Figure 3 shows the interface with the encryption and a decryption buttons. The encryption is performed simply by choosing any file while decryption is executed by choosing an encrypted file with an appropriate key. Figure 4 and 5 shows a successful encryption and decryption, respectively.



Click to make any Encryption.
Any file is acceptable.

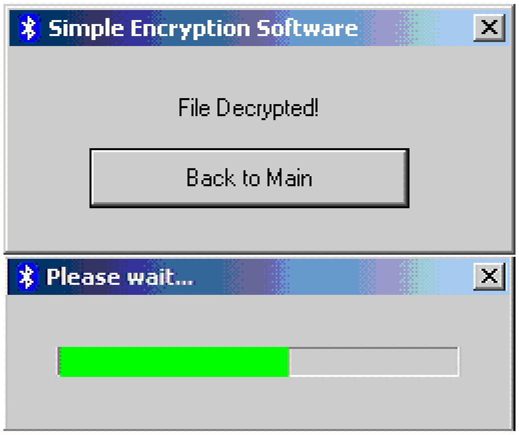
Click to make any Decryption.
Any file is acceptable.

FIGURE 3: shows the Encryption/Decryption interface.



File Encrypted!

FIGURE 4: shows a successful encryption and the encryption progress bar.



File Decrypted!

FIGURE 5: successful decryption and the decryption progress bar.

5. APPLICATION TESTING & PERFORMANCE CHECKING

Application testing is applied to the entire application with multiple application features to make sure the application can encrypt all type and all sizes of files. Successful testing means the application is user-friendly and comfortable to be used by all range of target users. For performance checking, the application was tested with different types and sizes of files and the performance of the application was rated by computing the time required for encryption of the files. Also, the reliability of the application was examined by the success rate of encryption. A successful execution means an encrypted file is not visible by others; also successful execution means a decrypted file was obtained using a key and an encrypted file. Table 1 shows the testing of different types and sizes of files. Figure 6 shows the encryption time for six different types of files that are of the same size of 5 mb. It can be seen that the encryption time is similar for all the files especially when the file size is small. The small size of files is a typical example of the use of this application as it is mainly targeted for small university campus. Most of the documents used in this environment are of text type with some figures inside of the text; therefore, the sizes of the files may not go over few mega bites.

TABLE 1: Testing of the Application with Different Types and Sized Files.

File Types	File Size(Mb)	Encryption Time (S)	Success Rate
Document	1/ 3/ 5	9/27/45	100%
Image	1/ 3/ 5	10/26/44	100%
Audio/ Video	1/ 3/ 5	18/28/45	100%
Zipped	1/ 3/ 5	10/25/44	100%
Exe.	1/ 3/ 5	12/27/48	100%

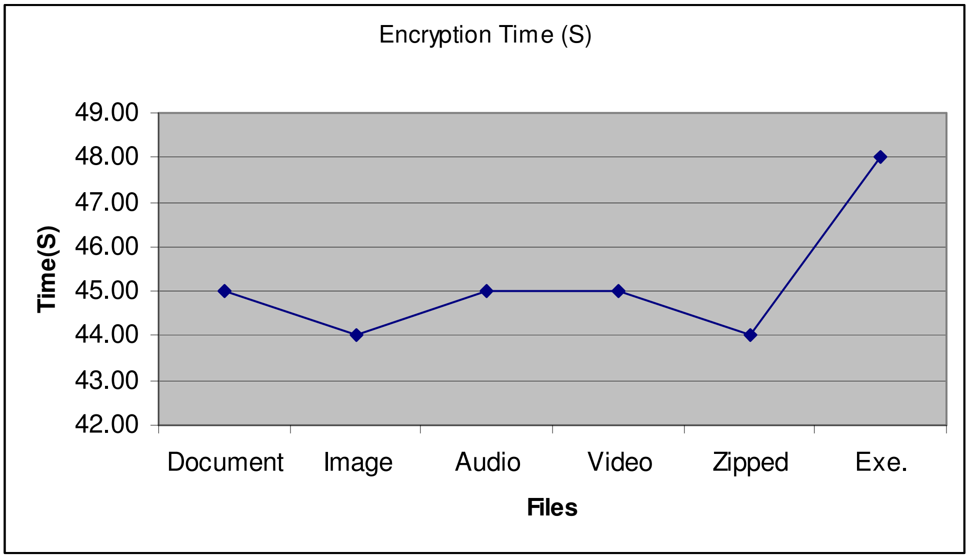


FIGURE 6: Graphical view of the encryption time for different types of files.

6. CONSLUSION & FUTURE WORK

A new simple tool has been created, which is targeted for use inside of a small institution such as a small university for lecturers' daily use of sending exam files and sensitive material such that the material can be encrypted and the file is sent in one e-mail while the encryption key is sent in another e-mail or via any secure communication channel. The encryption application developed and described in this paper might not be comparable to well-known encryption algorithms but its simplicity and availability proves that tools can be developed that fit the needs of an institution without resorting to purchasing expensive software from the market. For future enhancement to this application public key encryption can be applied where two keys can be generated: one to encrypt a file using the public key and another private key to decrypt it. Also, other more advanced encryption operations can be included to enhance the security of the application so that it can be used to encrypt more sensitive administrative material in an institution.

7. REFERENCES

[1] Wikipedia, “*Encryption*”, http://en.wikipedia.org/wiki/Encryption, modified on 13 December 2006.

[2] Freeman J., Neely R., and Megalo L. “*Developing Secure Systems: Issues and Solutions*”. IEEE Journal of Computer and Communication, Vol. 89, PP. 36-45. 1998

[3] Agnew G. B., Mullin R. C., Onyszchuk I. M., and Vqanstone S. A. “*An Implementation for a Fast Public-Key Cryptosystems*”. Journal of Cryptology, Vol.3, No 2, PP. 63-79. 1995.

[4] Beth T. and Gollmann D. “*Algorithm Engineering for Public Key Algorithms*”. IEEE Journal on Selected Areas in Communications; Vol. 7, No 4, PP. 458-466. 1989

[5] IBM. “*The Data Encryption Standard (DES) and its strength against attacks*”. IBM Journal of Research and Development, Vol. 38, PP. 243-250. 1994

[6] Wikipedia , “*Bitwise operation*” , http://en.wikipedia.org/wiki/Bitwise_operation , last modified on10 December 2006.

[7] Andy Wilson , “*Tips and Tricks: XOR Encryption*” http://www.andyw.com/director/xor.asp , 1998.

[8] Baraka H., El-Manawy H. A., and Attiya A. “*An Integrated Model for Internet Security Using Prevention and Detection Techniques*”. IEEE Journal of Computer and Communication Vol. 99, PP. 25-33. 1998

[9] Microsoft, "*Encrypting File System for Windows 2000*", http://www.microsoft.com/windows2000/techinfo/howitworks/security/encrypt.asp, 1998.

[10] Dariusz Stanislawek , “*Free Software copyright 1997 - 2006* ” http://members.ozemail.com.au/~nulifetv/freezip/freeware

[11] NeoCrypt, “*NeoCrypt File Protection Utility*” , http://sourceforge.net/projects/neocrypt

[12] Vivek Thakur , “NeekProtect”, http://neekprotect.sourceforge.net , 2006.

[13] Artur Ekert, Carolina Moura Alves, Ajay Gopinathan, “History of Cryptography”.

[14] Cryptomathic, “E-SECURITY DICTIONARY”, http://www.cryptomathic.com/labs/techdict.html , 2003.

Morphological Reconstruction for Word Level Script Identification

B.V.Dhandra dhandra_b_v@yahoo.co.in
P.G.Department of Studies and Research in
Computer Science, Gulbarga University,
Gulbarga -585106, India

Mallikarjun Hangarge mhangarge@yahoo.co.in
P.G.Department of Studies and Research in
Computer Science, Gulbarga University,
Gulbarga -585106, India

Abstract

A line of a bilingual document page may contain text words in regional language and numerals in English. For Optical Character Recognition (OCR) of such a document page, it is necessary to identify different script forms before running an individual OCR system. In this paper, we have identified a tool of morphological opening by reconstruction of an image in different directions and regional descriptors for script identification at word level, based on the observation that every text has a distinct visual appearance. The proposed system is developed for three Indian major bilingual documents, Kannada, Telugu and Devnagari containing English numerals. The nearest neighbour and k-nearest neighbour algorithms are applied to classify new word images. The proposed algorithm is tested on 2625 words with various font styles and sizes. The results obtained are quite encouraging

Keywords: Script identification, Bilingual documents, OCR, Morphological reconstruction, regional descriptors

1. INTRODUCTION

As the world moves closer to the concept of the “paperless office,” more and more communication and storage of documents is performed digitally. Documents and files that were once stored physically on paper are now being converted into electronic form in order to facilitate quicker additions, searches, and modifications, as well as to prolong the life of such records. A great portion of business documents and communication, however, still takes place in physical form and the fax machine remains a vital tool of communication worldwide. Because of this, there is a great demand for software, which automatically extracts, analyzes, and stores information from physical documents for later retrieval. All of these tasks fall under the general heading of document analysis, which has been a fast growing area of research in recent years.

A very important area in the field of document analysis is that of optical character recognition (OCR), which is broadly defined as the process of recognizing either printed or handwritten text from document images and converting it into electronic form. To date, many algorithms have been presented in the literature to perform this task for a specific language, and these OCRs will not work for a document containing more than one language/script.

Due to the diversity of languages and scripts, English has proved to be the binding language for India. Therefore, a bilingual document page may contain text words in regional language and numerals in English. So, bilingual OCR is needed to read these documents. To make a bilingual OCR successful, it is necessary to separate portions of different script regions of the bilingual document at word level and then identify the different script forms before running an individual OCR system. Among the works reported in this direction to distinguish between various Indian languages/scripts at word level are due to [4, 9, 12 and 14]. The algorithms proposed by Dhanya et al. [4] based on Gabor filters and spatial spread features, Padma et al. [9] based on discriminating features and Pal et al. [12] based on water reservoir and conventional features have recognition rate of more than 95%. The recognition accuracy of these algorithms falls drastically when the word length is less than three characters. Hence, the algorithms are word size dependent. Peeta Basa Pati et al. [14] have proposed word level script identification for Tamil, Devnagari and Oriya scripts based on 32 features using Gabor filters. It is obvious that the time complexity of this algorithm is more as it depends on 32 features. Authors have not reported about the performance of their algorithm for various font sizes and styles. Furthermore, the algorithms discussed so far have addressed only alphabet-based script identification (i.e. English text words separation from bilingual documents), whereas numeral script identification (i.e. English numerals separation from bilingual documents) is ignored. But, the fact is that, the large number of bilingual documents contains text words in regional languages and numerals in English (printed or handwritten). For example, Newspapers, Magazines, Books, Application forms, Railway Reservation forms etc. This has motivated us to develop a method for automatic script identification by separating English numerals (printed and handwritten) from bilingual documents.

From the literature survey, it is evident that some amount of work other than word level script/language identification has been carried out. Spitz [16] proposed a method for distinguishing between Asian and European languages by examining the upward concavities of connected components. Tan et al. [10] proposed a method based on texture analysis for automatic script and language identification from document images using multiple channel (Gabor) filters and Gray level co-occurrence matrices for seven languages: Chinese, English, Greek, Koreans, Malayalam, Persian and Russian. Hochberg, et al. [5] described a method of automatic script identification from document images using cluster-based templates. Tan [18] developed rotation invariant features extraction method for automatic script identification for six languages. Wood et al. [19] described projection profile method to determine Roman, Russian, Arabic, Korean and Chinese characters. Chaudhuri et al. [1] discussed an OCR system to read two Indian languages scripts: Bangla and Devnagari (Hindi). Chaudhuri et al. [2] described a complete printed Bangla OCR. Pal et al. [11] proposed an automatic technique of separating the text lines from 12 Indian scripts. Gaurav et al. [3] proposed a method for identification of Indian languages by combining Gabor filter based techniques and direction distance histogram classifier for Hindi, English, Malayalam, Bengali, Telugu and Urdu. Basavaraj et al. [13] proposed a neural network based system for script identification of Kannada, Hindi and English. Nagabhushan et al. [15] discussed an intelligent pin code script identification methodology based on texture analysis using modified invariant moments. In this paper an attempt is made to demonstrate the potentiality of morphological reconstruction approach for script identification at word level.

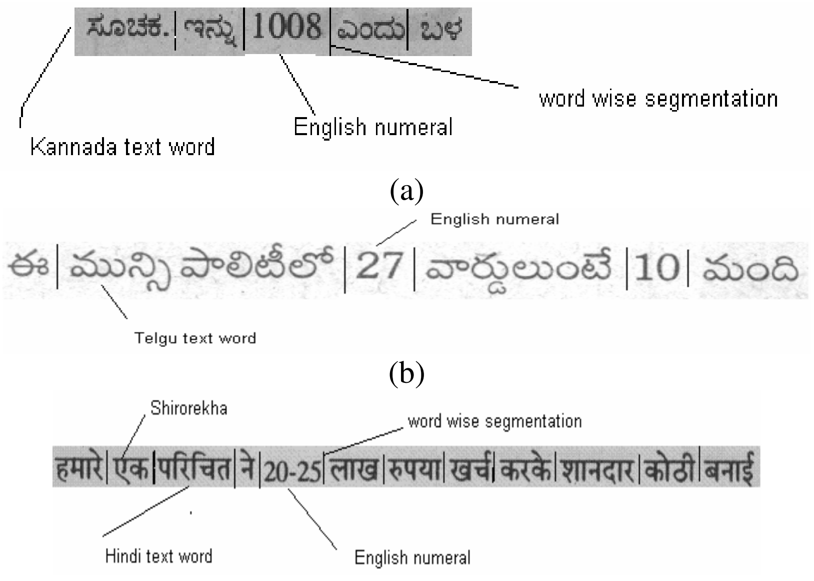
In Section 2, the brief overview of data collection, pre-processing and line and word segmentation are presented. In Section 3, the feature extraction, features computation and K nearest neighbour classifier are discussed. In Section 4, the proposed algorithm is presented. The experimental details and results obtained are presented in Section 5. Conclusion is given in Section 6.

2. DATA COLLECTION AND PREPROCESSING

In this paper three data sets are used for experimentation. The first data set is of 500 document pages of Kannada, Telugu, Devnagari are obtained from various magazines, newspapers, books and other such documents containing variable font styles and sizes. The second data set is of 175 handwritten English numerals collected from 100 writers. The third data set consists of 150 word images of varied font styles and sizes. The collected documents are scanned using HP Scanner at 300 DPI, which usually yields a low noise and good quality document image. The digitized images are in gray tone and we have used Otsu's [22] global thresholding approach to convert them into two-tone images. The two-tone images are then converted into 0-1 labels where the label 1 represents the object and 0 represents the background. The small objects like, single or double quotation marks, hyphens and periods etc. are removed using morphological opening. The next step in pre-processing is skew detection and correction. Using algorithm [7], with minor modification (i.e. vertical dilation with line structuring element of length 10 pixels) skew detection and correction has been performed before segmentation.

2.1 Line and word segmentation

To segment the document image into several text lines, we use the valleys of the horizontal projection computed by a row-wise sum of black pixels. The position between two consecutive horizontal projections where the histogram height is least denotes one boundary line. Using these boundary lines, document image is segmented into several text lines. Similarly, to segment each text line into several text words, we use the valleys of the vertical projection of each text line obtained by computing the column-wise sum of black pixels. The position between two consecutive vertical projections where the histogram height is least denotes one boundary line. Using these boundary lines, every text line is segmented into several text words. The word wise segmentation is illustrated in Fig.1. These word images are then used to compute the eight-connected components of white pixels on the image and produce the bounding box for each of the connected components.



(a)

(b)

(c)

FIGURE 1: Word-Wise Segmentation of (a) Kannada (b) Telugu and (c) Devnagari Scripts

3. FEATURE EXTRACTION

Each sample or pattern that we attempt to classify is a word. It is helpful to study the general characteristics of each of the three proposed scripts for feature extraction.

Devnagari: Most of the characters of Devnagari script have a horizontal line at the upper part. In Devnagari, this line is called sirorekha. However, we shall call them as headlines. When two or more Devnagari characters sit side by side to form a word, the sirorekha or headline will touch each other and generates a big headline [10].

Roman numerals (English numerals): The important property of the Roman numerals (English) is the existence of the vertical strokes in its characters and has less number of horizontal strokes. By the experiment, it is noticed that the vertical strokes in digits like 1, 3, 4, 6, 8, 9, and 0 are more dominant than that of horizontal strokes as compared to Devnagari script.

Kannada and Telugu Scripts: The Kannada and Telugu scripts have more horizontal, right and left diagonal strokes. Thus, the right and left diagonal strokes will also play an important rule in distinguishing Kannada, Telugu, and Devnagari script from Roman numerals. These directional stroke features are extracted from the connected components of an image or pattern using morphological opening. In the following, we describe the features and their method of computation. To extract the characters or components containing strokes in vertical, horizontal, right and left diagonal directions, we have performed the erosion operation on the input binary image with the line-structuring element. The length of the structuring element is thresholded to 70 %(experimentally fixed) of average height of all the connected components of an image. The resulting image is used for morphological opening in four directions to obtain the strokes of an input image, as illustrated in Fig.2.

Morphological Reconstruction: Reconstruction is a morphological transformation involving two images and a structuring element. One image, the marker, is the starting point for the transformation. The other mask image constraints the transformation. In this paper, a fast hybrid reconstruction algorithm [8] is used for reconstruction and erode image is used as the marker image throughout the experiment.

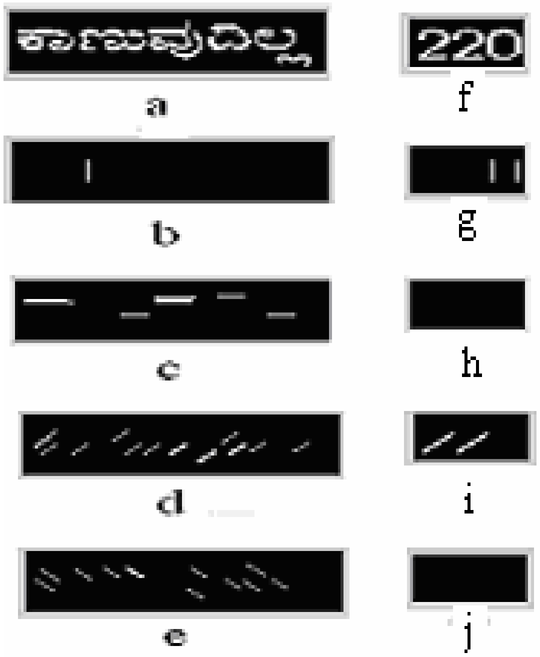


FIGURE 2: Shows the Stroke Extraction Process (a) Input Image of Kannada Script and (f) English Numeral (b), (c), (d), (e) and (g), (h), (i), (j) are Vertical, Horizontal, Right and Left Diagonal Stokes of Kannada and English Scripts Respectively

Opening by Reconstruction: It restores the shapes of the objects that remain after erosion. As an illustration the opening by reconstruction of Kannada word in vertical and horizontal directions is shown in Figure 3.

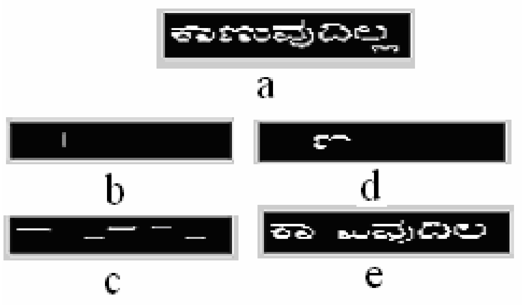


FIGURE 3: (a) Kannada Script Input Word Image, (b**)** Vertical Opening of (a), (c) Horizontal Opening of (a), (d) Vertical Reconstruction based on (b), (e) Horizontal Reconstruction based on (c).

For fill holes, we choose the marker image (erode image), f_m, to be 0 every where except on the image border, where it is set to 1-f. Here f is the image of a connected component.

$$f_m(x,y) = \begin{cases} 1 - f(x,y), & if\ (x,y) is\ on the border of\ f \\ 0, & otherwise \end{cases}$$

Then g = $[R_f^c\ (f_m)]^c$ has the effect of the filling the holes in f, where, R_f^c is the reconstructed image of f. As an example the horizontal reconstruction with fill holes of the Kannada script is shown in Figure 4.

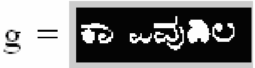


FIGURE 4: Represents Horizontal Reconstruction with Fill Holes**.**

3.1 Features Computation

1 On pixels densities (OPD) after reconstruction in vertical, horizontal, right and left diagonal directions with fill holes is given by

$$OPD(\theta) = \frac{\sum onpixels(g)}{size(g)}$$

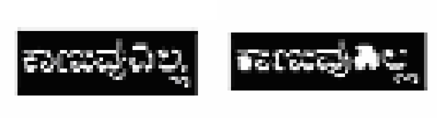
where, θ varies from 0 to 135 with the incremental bandwidth of 45 degrees. The OPD values are real numbers. Thus, we obtain a set of four features by reconstruction approach. The remaining four features considered are aspect ratio, pixel ratio, eccentricity and extent. These features computation is discussed in the following. Throughout, the discussion of section 3.1, N is referred for the number of components present in an image.

2 Aspect Ratio: - The ratio of the height to the width of a connected component of an image [6]. The average aspect ratio (AAR) is defined as

$$AAR(pattern) = \frac{1}{N} \sum_{i}^{N} \frac{height(component_i)}{width(component_i)}$$

The value of AAR is a real number. Note that the aspect ratio is very important feature for word wise script identification [21].

3 Pixels Ratio (PR): It is defined as the ratio between the on pixels of an input image after fill holes to its total number of pixels before fill holes, as illustrated in Fig. 5,. The value of the pixel ratio is a real number.



(a) (b)

FIGURE 5: Shows Hole Fill Operation (a) Input Kannada Word, (b) After Hole Fill Operation of (a)

$$PixelsRatio = \frac{\sum onpixels(b)}{Size(a)}$$

4. Eccentricity: It is defined as the length of minor axis divided by the length of the major axis of a connected component of an image [20]. This is also a real valued function.

$$Average\_eccentricity = \frac{1}{N}\sum_{i}^{N} eccentricity(Component_i)$$

5. Extent: It is a real valued function; defined as the proportion of the pixels in the bounding box that are also in the region. It can be computed as area divided by the area of the bounding box.

$$Average\_extent = \frac{1}{N}\sum_{i}^{N} extent(Component_i)$$

The sample feature vectors of Devnagari, English numerals and Kannada scripts are as under.

Devnagari = [0.3333 0.3333 0.3333 0.3333 0.3333 0.1846 0.9868 0.4872]
English = [0.1611 0 0.1384 0 0.2995 1.5515 0.7755 0.4030]
Kannada = [0.0172 0.1510 0.1682 0.1550 0.1805 0.7704 0.7184 0.4414]

3.2. K-Nearest neighbour Classifier

K-nearest neighbour is a supervised learning algorithm. It is based on minimum distance (Euclidian distance metric is used) from the query instance to the training samples to determine the k- nearest neighbours. After determining the k nearest neighbours, we take simple majority of these k-nearest neighbours to be the prediction of the query instance. The experiment is carried out by varying the number of neighbours (K= 3, 5, 7) and the performance of the algorithm is optimal when K = 3.

4. PEOPOSED ALGORITHM

The various steps involved in the proposed algorithm are as follows

1. Pre-process the input image i.e. binarisation using Otsu's method, and remove speckles using morphological opening.
2. Carry out the line wise and word wise segmentation based on horizontal and vertical projection profiles.
3. Carry out the morphological erosion and opening by reconstruction using the line structuring element in vertical, horizontal, left and right diagonal directions and perform the fill hole operation.
4. Compute the average pixel densities of the resulting images of step 3.
5. Compute the ratio of on pixels left after performing fill hole operation on input image to its size.
6. Compute the aspect ratio, eccentricity and extent of all the connected components of an input image and obtain their average values.
7. Classify the new word image based on the nearest neighbour and k-nearest neighbour classifiers.

5. RESULTS AND DISCUSSION

For experimentation, a sample image of size 256x256 pixels is selected manually from each document page and created a first data set of 2450 word images by segmentation. Out of these 2450 word images, Kannada, Devnagari are 750 each and Telugu and English numerals are 600 and 350 respectively. The second data set of 175 handwritten English numerals is used to test the potentiality of the proposed algorithm for script identification of handwritten numerals versus printed text words.

The classification accuracy achieved in identifying the scripts of first and second data set is presented in Table 2, 3, 4 and 5. Experimentally (based on principal component analysis), it is observed that, the right and left diagonal reconstruction features are not dominant and other six features are leading to retain the accuracy as reported. Although the primary aim of this paper is achieved, that is the word-wise script identification in bilingual documents; the fact is that, normally printed documents font sizes and styles are less varied. We therefore conducted a third set of experiment on 150 word images to test the sensitivity of the algorithm towards different font sizes and styles. These words are first created in different fonts using DTP packages, and then printed from a laser printer. The printed documents are scanned as mentioned earlier. On most commonly used five fonts of Kannada, Hindi and English are considered for experiment. For each font 10 word images are considered varying in font size from 10 to 36. Out of these 150 word images, Kannada, Devnagari and English numerals are 50 each. The Kannada font styles used are KN-TTKamanna, TTUma, TTNandini ,TTPadmini and TT-Pampa. The Devnagari font styles considered are DV-TTAakash, TTBhima, TTNatraj, TTRadhika, and TTsurekh. Times New Roman, Arial, Times New Roman italic, Arial Black and Bookman Old Style of English numerals are used for font and size sensitivity testing. It is noticed that, script identification accuracy achieved for third data set is consistent.

In the reported work of [4, 9, 12], it is mentioned that, the error rate is more when the word size is less than 3 characters. Our algorithm works for even single character words, but it fails when words like से ,हो, marks like “|” and broken sirorekha‘s are encountered in Devnagari. The touched, broken and bold face words of Kannada and Telugu are not recognized correctly because of loss in aspect ratio. Arial Black English numerals of size more than 16 points also misclassified. The sample test words and misclassified words are shown in Fig. 7 and 8.The proposed algorithm is implemented in MATLAB 6.1. The average time taken to recognize the script of a given word is 0.1547 seconds on a Pentium-IV with 128 MB RAM based machine running at 1.80 GHz. Since, there is no work reported for script identification of numerals at word level, to the best of our knowledge. However, the proposed method is compared with [4, 9, 12 and 14] as shown in Table 1.

6. CONCLUSION AND FUTURE WORK

In this paper, we investigated a tool of morphological opening by reconstruction of the image based on the strokes present in different directions and regional descriptors for script identification at word level. The simplicity of the algorithm is that, it works only on the basic morphological operations and shows its novelty for font and size independent script identification. The morphological reconstruction approach is efficiently used for extracting only those components or characters containing directional strokes and their densities are used for discriminating of the scripts. Furthermore, our method overcomes the word length constraint of [4, 9, 12] and works well even for single component words with minimum number of features. This work is first of its kind to the best of our knowledge. This algorithm can be generalized because of the visual appearance of every text is distinct and hence directional distribution of strokes must be distinct at an appropriate threshold. This work can also be extended to other Indian regional languages.

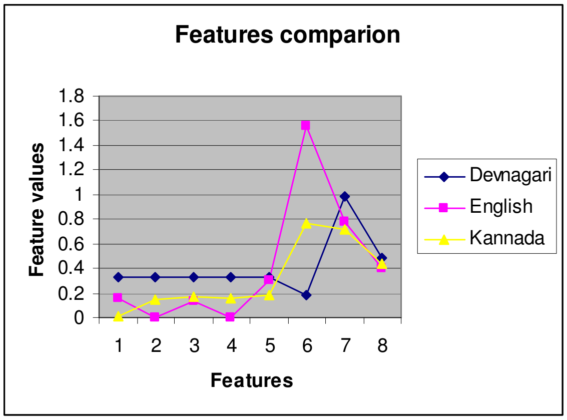


FIGURE 6: Shows Features Comparison of Three Scripts

Methods Proposed by	Scripts	Accuracy in %	Time complexity	Comments
D.Dhanya	Tamil and Roman	96.03	Not reported	Algorithm suffers ,when the word length is less than four characters
U.Pal	Devnagari, Telugu and Roman	96.72	Not reported	Algorithm suffers ,when the word length is less than three characters
M.C.Padma	Kannada, Roman Devnagari	95.66	Not reported	Algorithm suffers ,when the word length is less than three characters
Peeta Basa Pati	Devnagari, Tamil Oriya	97.33	Not reported	Not reported
Proposed Method	Telugu and Roman	97.60	0.15 seconds	Proposed algorithm works even for single character words.
	Kannada and Roman	96.68		
	Devnagari and Roman	98.58		
	Telugu, Devnagari and Roman	96.47	0.1547 seconds	
	Kannada, Devnagari and Roman	95.54		

TABLE 1: Comparative Study

Script /language	NN	KNN
Telugu Vs. English numerals	95.56% 96.38%	97.46 % 97.74%
Average	95.97%	97.6%
Kannada Vs. English numerals	97.43% 95.48%	97.43 % 95.93%
Average	96.45%	96.68%
Devnagari Vs. English numerals	98.53% 98.64%	98.53% 98.64%
Average	98.58%	98.58%

TABLE 2: Script Recognition Results of Printed Telugu, Kannada and Devnagari Text Words with English Numerals

Script	NN	KNN
Telugu	90.48%	95.24%
Devnagari	94.13%	96.44%
English	96.38%	97.74%

TABLE 3: Script Recognition Results of Telugu, Devnagari and Printed English Numerals

Script	NN	KNN
Kannada	91.53%	95.72%
Devnagari	94.13%	94.97%
English	95.53%	95.93%

TABLE 4: Script Recognition Results of Kannada, Devnagari and Printed English Numerals**.**

Script /language	KNN
Telugu Vs. English numerals	91.43% 94.68%
Average	93.05%
Kannada Vs. English numerals	93.58% 95.74%
Average	94.66%
Devnagari Vs. English numerals	98.53% 98.94%
Average	98.73%

TABLE 5: Script Recognition Results of Printed Telugu, Kannada and Devnagari Text Words with Handwritten English Numerals.

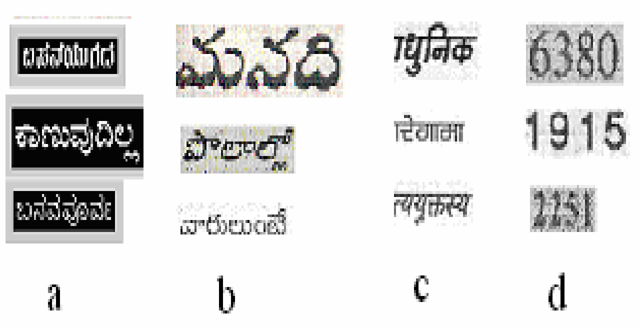


FIGURE 7: (a), (b), (c) and (d) are the Sample Test Images of Kannada, Telugu, Devnagari and English Numerals

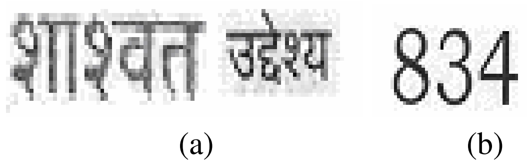


FIGURE 8: Misclassified Sample Word Images (a) Broken Sirorekha's of Devnagari Script, (b) English Numeral of Arial font of Size 24 Points

7. REFERENCES

1. B.B.Chaudhuri and U.Pal," *An OCR system to read two Indian language scripts: Bangla and Devnagari (Hindi)*", In Proceedings of 4th ICDAR, Uhn. 18-20 August, 1997

2. B.B.Chaudhuri and U.Pal. "*A complete printed Bangla OCR*", Pattern Recognition vol.31, pp 531-549, 1998

3. Santanu Chaudhury, Gaurav Harit, Shekar Madnani, R.B.Shet," *Identification of scripts of Indian languages by Combining trainable classifiers*", In Proceedings of ICVGIP 2000, Dec-20-22, Bangalore, India.

4. D Dhanya, A.G Ramakrishnan and Peeta Basa pati, "*Script identification in printed bilingual documents*," Sadhana, vol. 27, part-1, pp. 73-82, 2002

5. J. Hochberg, P. Kelly, T Thomas and L Kerns, "*Automatic script identification from document images using cluster-based templates*," IEEE Transactions Pattern Analysis and Machine Intelligence, vol.19, pp.176-181, 1997

6. Judith Hochberg, Kevin Bowers, Michael Cannon and Patrick Keely, "*Script and language identification for hand-written document images,*" IJDAR-1999, vol.2, pp. 45-52

7. B.V.Dhandra, V.S.Malemath, Mallikarjun Hangarge, Ravindra Hegadi, "*Skew detection in Binary image documents based on Image Dilation and Region labeling Approach*", In Proceedings of ICPR 2006, V. No. II-3, pp. 954-957

8. Vincent, L.," *Morphological gray scale reconstruction in image analysis: Applications and efficient algorithms*," IEEE Trans. on Image processing, vol.2, no. 2, pp. 176-201, 1993

9. M.C.Padma and P. Nagabhushan," *Identification and separation of text words of Kannada Hindi and English languages through discriminating features*", In Proceedings of NCDAR-2003, pp- 252-260. 2003

10. G.S.Peake and Tan, "*Script and language identification from document images*", In Proceedings of Eighth British Mach. Vision Conf., vol.2, pp. 230-233, Sept-1997

11. U.Pal and B.B.Chaudhuri, "*Script line separation from Indian Multi-script documents*," 5th ICDAR, pp.406-409, 1999

12. U.Pal. S.Sinha and B.B Chaudhuri, "*Word-wise Script identification from a document containing English, Devnagari and Telgu Text*," In Proceedings of NCDAR-2003, PP 213-220

13. S. Basavaraj, Patil and N.V.Subbareddy. "*Neural network based system for script identification in Indian documents*," Sadhana, vol. 27, part-1, pp. 83-97, 2002

14. Peeta Basa pati, S. Sabari Raju, Nishikanta Pati and A.G. Ramakrishnan, "*Gabor filters for document analysis in Indian Bilingual Documents*," In Proceedings of ICISIP-2004, pp. 123-126

15. P. Nagabhushan, S.A. Angadi and B.S. Anami," An Intelligent Pin code Script Identification Methodology Based on Texture Analysis using Modified Invariant Moments," In Proceedings of ICCR-2005, pp. 615-623

16. A.L.Spitz, "*Determination of the script and language content of document images*," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, pp.234-245, 1997

17. A. L. Spitz, "*Multilingual document recognition Electronic publishing, Document Manipulations*, and Typography," R. Furuta ed. Cambridge Uni. Press, pp. 193-206, 1990

18. T.N.Tan, "*Rotation invariant texture features and their use in automatic script identification*," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, pp.751-756, 1998

19. S. Wood. X. Yao. K.Krishnamurthi and L.Dang ”*Language identification from for printed text independent of segmentation*,” In Proceedings of International conference on Image Processing, pp. 428-431, 1995

20. Dengsheng Zhang, Guojun Lu, “*Review of shape representation and description techniques*,” Pattern Recognition, vol. 37, pp. 1-19, 2004

21. Annop M. Namboodri, Anil K Jain, “ *Online handwritten script identification*”, IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 26,no.1,pp. 124-130, 2004

22. N. Otsu, ” *A Threshold Selection Method from Gray-Level Histogram*” , IEEE Transaction Systems, Man, and Cybernetics, vol.9,no.1,pp.62-66,1979

A REVIEW PAPER ON AD HOC NETWORK SECURITY

Karan Singh, karancs12@yahoo.com
Computer Science and Engineering Department
Motilal National Institute of Technology, Allahabad
Allahabad, India -211004

Rama Shankar Yadav, rsy@mnnit.ac.in
Computer Science and Engineering Department
Motilal National Institute of Technology, Allahabad
Allahabad, India -211004

Ranvijay, cs0620@mnnit.ac.in
Computer Science and Engineering Department
Motilal National Institute of Technology, Allahabad
Allahabad, India -211004

Abstract

In this article we present a survey of secure ad hoc routing protocols for wireless networks. Ad hoc network is a collection of nodes that is connected through a wireless medium forming rapidly changing topologies. Attacks on ad hoc network routing protocols disrupt network performance and reliability with there solution. We briefly present the most popular protocols that follow the table-driven and the source-initiated on-demand approaches. The comparison between the proposed solutions and parameters of ad hoc network shows the performance according to secure protocols. We discuss in this paper routing protocol and challenges and also discuss authentication in ad hoc network.

KEYWORDS: Wireless Network, Ad hoc Network, Security Service, Routing Protocols, Routing Authentication, Hash function and Secure Routing Protocols.

I. INTRODUCTION

Wireless networks [34] consist of a number of nodes which communicate with each other over a wireless channel which have various types of networks: sensor network, ad hoc mobile networks, cellular networks and satellite networks. Wireless sensor networks consist of small nodes with sensing, computation and wireless communications capabilities. Many routing protocols have been specifically designed for WSNs where energy awareness is the key issue. Routing protocols in WSNs [41] differ depending on the application and network architecture. Ad-hoc networks are a new paradigm of wireless communication for mobile hosts where node mobility causes frequent changes in topology. Ad hoc networks are self-configurable and autonomous systems consisting of routers and hosts, which are able to support movablity and organize themselves arbitrarily. This means that the topology of the ad hoc network changes dynamically and unpredictably. Moreover, the ad hoc network can be either constructed or destructed quickly and autonomously without any administrative server or infrastructure. Without support from the fixed infrastructure, it is undoubtedly arduous for people to distinguish the insider and outsider of the wireless network. That is to say, it is not easy for us to tell apart the legal and the illegal participants in wireless systems. Because of the above mentioned properties, the implementation of security infrastructure has become a critical challenge when we design a wireless network system. If the nodes of ad hoc networks are mobile and with wireless communication to maintain the connectivity, it is known as mobile ad hoc network (MANET) and require an extremely flexible technology for establishing communications in situations which demand a fully decentralized network without any fixed base stations, such as battlefields, military applications, and other emergency and disaster situations.

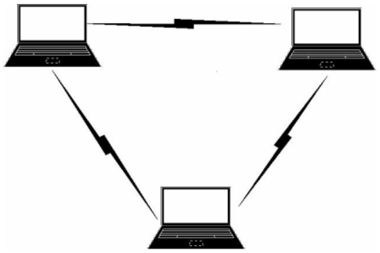


FIGURE 1: AD HOC NETWORK

Since, all nodes are mobile, the network topology of a MANET is generally dynamic and may change frequently. Thus, protocol such as 802.11 to communicate via same frequency or Bluetooth have require power consumption is directly proportional to the distance between hosts, direct *single-hop* transmissions between two hosts can require significant power, causing interference with other such transmissions [41]. To avoid this *routing problem*, two hosts can use *multi-hop* [34] transmission to communicate via other hosts in the network A router should provide the ability to rank routing information sources from most trustworthy to least trustworthy and to accept routing information about any particular destination from the most trustworthy sources first. A router should provide a mechanism to filter out obviously invalid routes. Routers must not by default redistributes routing data they do not themselves use, trust or otherwise consider valid. Routers must be at least a little paranoid about accepting routing data from anyone, and must be especially careful when they distribute routing information provided to them by another party.

Figure 1 shows three node where ad hoc network where every node is connected to wireless, and work as access point to forward and receive data. This article discuss attacks on ad hoc networks and discusses current approaches for establishing cryptographic keys in ad hoc networks. We describe the state of research in secure ad hoc routing protocols, routing challenges and its research issues.

II. ROUTING PROTOCOL AND ITS CHALLENGE IN AD HOC NETWORK

In this section we are going to discuss different approaches adopted for routing and security challenges in Ad hoc networks.

A. ROUTING PROTOCOLS

Routing in mobile ad hoc networks faces additional problems and challenges [22], [30] when compared to routing in traditional wired networks with fixed infrastructure. There are several well known protocols in the literature that have been specifically developed to cope with the limitations imposed by ad hoc networking environments. Most of the existing routing protocols follow two different design approaches to confront the inherent Characteristics of ad hoc networks: the *table-driven* and the *source-initiated on-demand* approaches.

Table-driven ad hoc routing protocols maintain at all times routing information regarding the connectivity of every node to all other nodes that participate in the network. Also known as *proactive*, [49] these protocols allow every node to have a clear and consistent view of the network topology by propagating periodic updates [27]. An alternative approach to that followed by table-driven protocols is the source-initiated on-demand routing. According to this approach, a route is created only when the source node requires a route to a specific destination. A route is acquired by the initiation of a *route discovery* function by the source node.

The data packets transmitted while a route discovery is in process are buffered and are sent when the path is established. An established route is maintained as long as it is required through a *route maintenance* procedure. Table 1 shows the various type of routing protocols according to parameter which are response time, bandwidth and energy.

Parameter	Network	Protocols	Examples
Response Time And Bandwidth	Ad hoc	Proactive protocols	Destination-sequenced Distance-Vector (DSDV)
			Optimized Link- State Routing (OLSR)
		Reactive protocols	Ad Hoc On-Demand Distance-Vector (AODV)
			Dynamic Source Routing (DSR)
			Geography-based routing
			Cluster-based (or *hierarchical*) routing
Energy	Sensor	Network structure	Flat network routing
			Hierarchical network routing
			Location based routing
		Protocol operation	Negotiation based routing
			Multi-path based routing
			Query based routing
			QoS based routing
			Coherent based routing

TABLE 1: CLASSIFICATION OF ROUTING PROTOCAL

B. SECURITY CHALLENGES IN AD HOC NETWORKS

Use of wireless links renders an Ad hoc network susceptible to link attacks ranging from passive eavesdropping to active impersonation, message replay and message distortion [9],[10],[52].Eavesdropping might give an attacker access to secret information thus violating confidentiality. Active attacks could range from deleting messages, injecting erroneous messages; impersonate a node etc thus violating availability, integrity, authentication and non-repudiation. Nodes roaming freely in a hostile environment with relatively poor physical protection have non-negligible probability of being compromised. Hence, we need to consider malicious attacks not only from outside but also from within the network from compromised nodes. Thus following are the ways by which security can be breached. [56]

- **Vulnerability of Channels:** As in any wireless network, messages can be eavesdropped and fake messages can be injected into the network without the difficulty of having physical access to network components.
- **Vulnerability of nodes:** Since the network nodes usually do not reside in physically protected places, such as locked rooms, they can more easily be captured and fall under the control of an attacker.
- **Absence of Infrastructure:** Ad hoc networks are supposed to operate independently of any fixed infrastructure. This makes the classical security solutions based on certification authorities and on-line servers inapplicable.
- **Dynamically Changing Topology**: In mobile ad hoc networks, the permanent changes of topology require sophisticated routing protocols, the security of which is an additional challenge. A particular difficulty is that incorrect routing information can be generated by compromised nodes or as a result of some topology changes and it is hard to distinguish between the two cases.

For high survivability Ad hoc networks should have a distributed architecture with no central entities, centrality increases vulnerability. Ad-hoc network is dynamic due to frequent changes in topology. Even the trust relationships among individual nodes also changes, especially when some nodes are found to be compromised. Security mechanism need to be on the dynamic and not static and should be scalable.

III. SECURITY MODEL

In this section we first discuss security goals attacks and thus secure routing protocol which are following:

A. SECURITY GOALS FOR AD HOC

- **Availability:** Ensures survivability despite Denial Of Service (DOS) attacks. On physical and media access control layer attacker can use jamming techniques to interfere with communication on physical channel. On network layer the attacker can disrupt the routing protocol. On higher layers, the attacker could bring down high level services e.g.: key management service.
- **Confidentiality:** Ensures certain information is never disclosed to unauthorized entities.
- **Integrity:** Message being transmitted is never corrupted.
- **Authentication:** Enables a node to ensure the identity of the peer node it is communicating with. Without which an attacker would impersonate a node, thus gaining unauthorized access to resource and sensitive information and interfering with operation of other nodes.
- **Non-repudiation:** Ensures that the origin of a message cannot deny having sent the message.
- **Non-impersonation**: No one else can pretend to be another authorized member to learn any useful information.
- **Attacks using fabrication:** Generation of false routing messages is termed as fabrication messages. Such attacks are difficult to detect.

B. ATTACK ON AD HOC NETWORK

There are various types of attacks on ad hoc network which are describing following:

- **Location Disclosure:** Location disclosure is an attack that targets the privacy requirements of an ad hoc network. Through the use of traffic analysis techniques [20], or with simpler probing and monitoring approaches, an attacker is able to discover the location of a node, or even the structure of the entire network.
- **Black Hole:** In a black hole attack a malicious node injects false route replies to the route requests it receives, advertising itself as having the shortest path to a destination[26]. These fake replies can be fabricated to divert network traffic through the malicious node for eavesdropping, or simply to attract all traffic to it in order to perform a denial of service attack by dropping the received packets.
- **Replay:** An attacker that performs a replay attack injects into the network routing traffic that has been captured previously. This attack usually targets the freshness of routes, but can also be used to undermine poorly designed security solutions.
- **Wormhole:** The wormhole attack is one of the most powerful presented here since it involves the cooperation between two malicious nodes that participate in the network [53]. One attacker, e.g. node A, captures routing traffic at one point of the network and tunnels them to another point in the network, to node B, for example, that shares a private communication link with A. Node B then selectively injects tunneled traffic back into the network. The connectivity of the nodes that have established routes over the wormhole link is completely under the control of the two colluding attackers. The solution to the wormhole attack is *packet leashes*.
- **Blackmail:** This attack is relevant against routing protocols that use mechanisms for the identification of malicious nodes and propagate messages that try to blacklist the offender [58]. An attacker may fabricate such reporting messages and try to isolate legitimate nodes from the network. The security property of non-repudiation can prove to be useful in such cases since it binds a node to the messages it generated.

- **Denial of Service:** Denial of service attacks aim at the complete disruption of the routing function and therefore the entire operation of the ad hoc network [15]. Specific instances of denial of service attacks include the *routing table overflow and* the *sleep deprivation torture.*. In a routing table overflow attack the malicious node floods the network with bogus route creation packets in order to consume the resources of the participating nodes and disrupt the establishment of legitimate routes. The sleep deprivation torture attack aims at the consumption of batteries of a specific node by constantly keeping it engaged in routing decisions.
- **Routing Table Poisoning:** Routing protocols maintain tables that hold information regarding routes of the network. In poisoning attacks the malicious nodes generate and send fabricated signaling traffic, or modify legitimate messages from other nodes, in order to create false entries in the tables of the participating nodes [15]. For example, an attacker can send routing updates that do not correspond to actual changes in the topology of the ad hoc network. Routing table poisoning attacks can result in the selection of non-optimal routes, the creation of routing loops, bottlenecks, and even portioning certain parts of the network.
- **Rushing Attack:** Rushing attack is that results in denial-of-service when used against *all* previous on-demand ad hoc network routing protocols [55]. For example, DSR, AODV, and secure protocols based on them, such as Ariadne, ARAN, and SAODV, are unable to discover routes longer than two hops when subject to this attack. develop *Rushing Attack Prevention (RAP)*, a generic defense against the rushing attack for on-demand protocols that can be applied to any existing on-demand routing protocol to allow that protocol to resist the rushing attack.
- **Breaking the neighbor relationship:** An intelligent filter is placed by an intruder on a communication link between two ISs(Information system) could modify or change information in the routing updates or even intercept traffic belonging to any data session.
- **Masquerading:** During the neighbor acquisition process, a outside intruder could masquerade an nonexistent or existing IS by attaching itself to communication link and illegally joining in the routing protocol do main by compromising authentication system. The threat of masquerading is almost the same as that of a compromised IS.
- **Passive Listening and traffic analysis:** The intruder could passively gather exposed routing information. Such a attack can not effect the operation of routing protocol, but it is a breach of user trust to routing the protocol. Thus, sensitive routing information should be protected. However, the confidentiality of user data is not the responsibility of routing protocol

C. ROUTING SECURITY IN AD HOC NETWORK

The contemporary routing protocols for Ad hoc networks cope well with dynamically changing topology but are not designed to accommodate defense against malicious attackers. No single standard protocols capture common security threats and provide guidelines to secure routing. Routers exchange network topology informally in order to establish routes between nodes another potential target for malicious attackers who intend to bring down the network. External attackers injecting erroneous routing info, replaying old routing info or distorting routing info in order to partition a network or overloading a network with retransmissions and inefficient routing. Internal compromised nodes - more severe detection and correction more difficult Routing info signed by each node won't work since compromised nodes can generate valid signatures using their private keys.

Detection of compromised nodes through routing information is also difficult due to dynamic topology of Ad hoc networks [22]. Routing protocols for Ad hoc networks must handle outdated routing information to accommodate dynamic changing topology. False routing information generated by compromised nodes can also be regarded as outdated routing information. As long as there are sufficient numbers of valid nodes, the routing protocol should be able to bypass the compromised nodes, this however needs the existence of multiple, possibly disjoint routes between nodes. Routing protocol should be able to make use of an alternate route if the existing one appears to have faulted

D. ROUTING AUTHENTICATION

Routing authentication is one of the important factors in ad hoc networks during route discovery because ad hoc is infrastructure less network. So it is required that a reply coming from a node against a route request must be authentic. That's why authentication protocol is required between the nodes of ad hoc network. In this section we emphasize on the ways by which these protocols can be used.

i. New key agreement scenario

Consider a group of people getting together for an Ad hoc meeting in a room and trying to establish a wireless network through their laptops. They trust one another personally; however don't have any a priori shared secret (password) to authenticate one another. They don't want anybody outside the room to get a wind of their conversation indoors. This particular scenario is vulnerable to any attacker who not only can monitor the communication but can also modify the messages and can also insert messages and make them appear to have come from somebody inside the room. This is a classic example of Ad hoc network and the simplest way to tackle this example would be through location based key agreement - to map locations to name and then use identity based mechanisms for key agreement[10][56]. e.g.: participants writing the IP addresses on a piece of paper and passing it around. Then a certificate based key agreement mechanism can be used. These public key certificates can allow participants to verify the binding between the IP address and keys of other participants.

ii. Two obvious problems

a) Difficult to determine if the certificate presented by the participant has been revoked

b) Participants may be divided into 2 or more certification hierarchies and that they don't have cross certification hierarchies.

One obvious solution

A trusted third is party capable of locating players, however not always feasible due to non-infrastructure nature of Ad hoc networks. Physically secure channel limited to those present in the room to negotiate the session key before switching to the insecure wireless channel.

iii. Password based Authenticated Key Exchange

A fresh password is chosen and shared among those present in the room in order to capture the existing shared context. If this password is long random string, can be used to setup security association, but less user friendly. Natural language phrases are more users friendly, however vulnerable to dictionary attacks[10][16][42]. Need to derive a strong session key from a weak shared password. Desirable properties for such a protocol are following

- **Secrecy:** Only those players that know the initial shared weak secret password should learn the session key and nobody else should.
- **Perfect Forward Secrecy:** Warrants that if an attacker who succeeds in compromising one of the participants at a later time would be unable to figure out the session key resulting from previous runs of protocol.
- **Contributory Key Agreement:** If each and every player participates in the creation of the final session key, by making a contribution, then it is called contributory key agreement.
- **Tolerance to Disruption Attempts:** Not only strong attackers who can disrupt communication by jamming radio channels etc but even the weaker attackers who can insert but cannot modify or delete messages sent by players are also provided for.

iv. Password authenticated Diffie - Hellman key Exchange

- **Two Party Version:** In the elementary DH protocol, *two parties* A and B agree on a prime p and a generator g of the multiplicative group Zp^* (i.e. the set $\{1, 2 \ldots p-1\}$). A and B choose random secrets S_A and S_B such that $1 \le S_A, S_B \le p-1$
  1) A computes g^{S_A}, encrypts it with the shared secret password P and sends it to B. $A \to B: A,\ P(g^{S_A})$

2) B extracts g^{S_A} from the message computes g^{S_B} and also computes the session key $K = (g^{S_A})^{S_B}$. B then chooses a random challenge C_B and encrypts it using the key K. B encrypts S_B using P. It then sends the two quantities to A. $B \rightarrow A, P(S_B), K(C_B)$.
3) A extracts S_B from $P(S_B)$ and computes the key $K = (g^{S_A})^{S_B}$. It then extracts C_B by decrypting $K(C_B)$. A then generates challenge (random) C_A, encrypts both C_A and C_B with K and sends it to B. $A \rightarrow B, K(C_A, C_B)$.
4) This message (3) convinces B that A was able to decrypt the message in (2) correctly. B then encrypts C_A using K and sends it to A.$B \rightarrow A, K(C_A)$.

A decrypts the message to see if the plaintext is indeed C_A. This would convince A that B knew K. This would in turn convince A that B knew P.

- **Multi-party version:** There are let's just say n players M_1, M_2, …, M_n who all share a password P, each generating a random quantity S_i which is its contribution to the eventual session key K = g $^{S1S2\_\_\_Sn-1Sn}$. The protocol is divided into 3 parts. In the first part (steps 1 and 2) players M_i to M_{n-1} generate an intermediate key $PI = g^{S1S2---Sn-1}\ in\ n-1$ steps.

In the second part (steps 3 and 4) each M_i (where i = 1 to n-1) has a separate with M_n, at the end of which all the players are in a position to compute K. The third part (step 5) being the key confirmation.

1) $M_i \rightarrow M_{i+1} : PI = g^{S1S2---Si} = 1\ to\ n-2$ in sequence.
2) M_{n-1} --> ALL : PI = g $^{S1S2\_\_\_Sn-1}$, broadcast
3) $M_i \rightarrow M_n : P(C_i)\ i = 1\ to\ n-1$, in parallel, where C_i = PI $^{Si'/Si}$ and S_i' is the blinding factor that is randomly chosen by M_i.
4) $M_n \rightarrow M_i : (C_i)S_n\ i = 1\ to\ n-1$, in parallel.
5) $M_i \rightarrow ALL : K(M_i, h(M_1, M_2, \ldots\ldots\ldots M_n))$ broadcast.

Step 1 consists of (n-2) sub steps. In the first sub step player M_1 computes g^{S1} and sends it to M_2 etc. At the end of the (n-2)th sub step, M_{n-1} receives $g^{S1S2\_\_\_Sn-2}$, which it then raises by (S_{n-1}) to get the intermediate key PI = $g^{S1S2---Sn-1}$.

In step 2, M_{n-1} broadcast this PI to everyone. Now every M_i *(i = 1 to n-1)* removes its contribution i.e, S_i *(i=1 to n-1)* from the PI respectively but also inserts a randomly chosen blinding factor S_i, encrypts the whole thing with the shared password P.

In step 3, each M_i will in parallel send the encryption to M_n. M_n decrypts the received message to extract C_i. It then raises each C_i by S_n and returns the result in parallel to each M_i. At this point each player can compute the session key as follows K = g $^{S1S2\_\_\_Sn-1Sn}$. M_n raises PI by S_n : $K = (PI)^{Sn}$. Each M_i unblinds the quantity it receives from M_n and re inserts its original contribution S_i to construct the session key K = g $^{S1S2\_\_\_Sn-1Sn}$ = $(PI)^{Sn}$.

Finally, some player broadcasts a key confirmation message that allows each player to verify that at least one another player has decided on the same key K. The blinding factor S_i is needed for the following reasons.

1) Without the blinding, the quantity encrypted with P by M_{n-1} from step 3 is the same as what it receives in step 1.
2) An attacker could send g $^{S1S2\_\_\_Si}$ to M_i in step 2 instead of the broadcast message (intermediate key) PI. If M_i uses this quantity to generate its message in step 3, the resulting message is same as the message received by M_i in step 1. To thwart dictionary attacks, blinding is necessary.

This protocol does provide perfect forward secrecy. It is also quasi-resilient to disruption except when M_n is compromised.

IV. SECURE ROUTING PROTOCOLS

A. ARAN

Authenticated Routing for Ad-hoc Networks (ARAN) detects and protects against malicious actions by third parties and peers in Ad-hoc environment. ARAN introduces authentication, message integrity and non-repudiation to an Ad-hoc environment [12][30]. ARAN is composed of two distinct stages. The first stage is simple and requires little extra work from peers beyond traditional ad hoc protocols. Nodes that perform the optional second stage increase the security of their route, but incur additional cost for their ad hoc peers who may not comply (e.g., if they are low on battery resources). Brief description of stages as follows

- **Stage 1:** It contains a preliminary certification stage and a mandatory end-end authentication stage. It is a lightweight stage and does not demand too many resources.

  Preliminary Certification: ARAN requires the use of a trusted certificate server T. Before entering the Ad-hoc network, each node requests a certificate from T. For a node A the certificate contains the IP address of A, the public key of A, a timestamp t of when the certificate was created, and a time e at which the certificate expires. These variables are concatenated and signed by T. All nodes must maintain fresh certificates with the trusted server and must know T's public key.

  End-to-End Authentication: The goal of stage 1 is for the source to verify that the intended destination was reached. In this stage, the source trusts the destination to choose the return path.

  Source node: A source node A, begins route instantiation to a destination X by broadcasting to its neighbors a route discovery packet (RDP): $\rightarrow broadcast: [RDP, IP_X, Cert_A, N_A, t]K_A -$. The RDP includes a packet type identifier ("RDP"), the IP address of t the destination (IPx), A's certificate (CertA), a nonce N_A , and the current time t, all signed with A's private key. Each time A performs route discovery, monotonically increases the nonce. Nodes then store the nonce they have last seen with its timestamp.

  Intermediate node for RDP: Each node records the neighbor from which it received the message. It then forwards the message to each of I its neighbors, signing the contents of the message. This signature prevents poofing attacks that may alter the route or form loops. Let A's neighbor be B. $B \rightarrow broadcast:$ $[[RDP, IP_X, Cert_A, N_A, t]K_A -]K_B -, Cert_B$. Nodes do not forward messages for which they have already seen the (N_A ,IP_A) tuple. Upon receiving the broadcast, B's neighbor C validates the signature with the given certificate. C then rebroadcasts the RDP to its neighbors, first removing B's signature. $C \rightarrow broadcast: [[RDP, IP_X, Cert_A, N_A, t]K_A -]K_C -, Cert_C$

  Destination node: Eventually, the message is received by the destination, X, who replies to the first RDP that it receives for a source and a given nonce. There is no guarantee that the first RDP received traveled along the shortest path from the source. The destination unicasts a Reply (REP) packet back along the reverse path to the source. $X \rightarrow D: [REP, IP_A, Cert_X, N_A, t]K_X -$

  Intermediate node for REP : Nodes that receive the REP forward the packet back to the predecessor from which they received the original RDP. All REPs are signed by the sender. Let D's next hop to the source be node C. $D \rightarrow C: [[REP, IP_A, Cert_X, N_A, t]K_X -]K_D -, Cert_D$ C validates D's signature, removes the signature, and then signs the contents of the message before unicasting the RDP to B. $C \rightarrow B: [[REP, IP_A, Cert_X, N_A, t]K_X -]K_C -, Cert_C$ A node checks the signature of the previous hop as the REP is returned to the source. This avoids attacks where malicious nodes instantiate routes by impersonation and re-play of X's message.

  Source Node: When the source receives the REP, it verifies that the correct nonce was returned by the destination as well as the destination's signature. Only the destination can answer an RDP packet. Other nodes that already have paths to the destination cannot reply for the destination. While other protocols allow this networking optimization, we note that removing it also removes several possible exploits and cuts down on the reply traffic received by the source. Because only the destination can send REPs, loop freedom is guaranteed easily.

Disadvantages: ARAN requires that nodes keep one routing table entry per source-destination pair that is currently active. This is certainly more costly than per-destination entries in non-secure ad hoc routing protocols.

- **Stage 2:** It is done only after Stage 1 is over. This is because the destination certificate is required in Stage 2. This stage is primarily used for discovery of shortest path in a secure fashion. Since a path is already discovered in Stage 1, data transfer can be pipelined with Stage 2 's shortest path discovery operation.

  Source Node: The source begins by broadcasting a Shortest Path Confirmation (SPC) message to its neighbors (the same variables are used as in stage 1. $A \rightarrow broadcast: SPC, IP_X, Cert_X, [[IP_X, Cert_A, N_A, t]K_A -]K_X +$. The SPC message begins with the SPC packet identifier (“SPC"), X's IP address and certificate. The source concatenates a signed message containing the IP address of X, its certificate, a nonce and timestamp. This signed message is encrypted with X's public key so that other nodes cannot modify the contents.

  Intermediate Node: A neighbor B that receives the message rebroadcasts the message after including its own cryptographic credentials. B signs the encrypted portion of the received SPC, includes its own certificate, and re-encrypts with the public key of X. This public key can be obtained in the certificate forwarded by A. $B \rightarrow broadcast: SPC, IP_X, Cert_X, \left[[[IP_X, Cert_A, N_A, t]K_A -]K_X + \right] K_B -, Cert_B]K_X +$ Nodes that receive the SPC packet create entries in their routing table so as not to forward duplicate packets. The entry also serves to route the reply packet from the destination along the reverse path.

  Destination Node: Once the destination X receives the SPC, it checks that all the signatures are valid. X replies to the first SPC it receives and also any SPC with a shorter recorded path. X sends a Recorded Shortest Path (RSP) message to the source through its predecessor D $X \rightarrow D: [RSP, IP_A, Cert_X, N_A, route]K_X -$.The source eventually receives the packet and verifies that the nonce corresponds to the SPC is originally generated.

Advantages: The onion-like signing of messages prevents nodes in the middle from changing the path in several ways. First, to increase the path length of the SPC, malicious nodes require an additional valid certificate. Second, malicious nodes cannot decrease the recorded path length or alter it because doing so would break the integrity of the encrypted data.

- **Route Maintenance:** ARAN is an on-demand protocol. Nodes keep track of whether routes are active [58]. When no traffic has occurred on an existing route for that route's lifetime, the route is simply de-activated in the route table. Data received on an inactive route causes nodes to generate an Error (ERR) message that travels the reverse path towards the source. Nodes also use ERR messages to report links in active routes that are broken due to node movement. All ERR message must be signed. For a route between source A and destination X, a node B generates the ERR message for its neighbor C as follows: $B \rightarrow C: [ERR, IP_A, IP_X, Cert_C, N_B, t]K_B$ This message is forwarded along the path towards the source without modification. A nonce and timestamp ensures the ERR message is fresh. Because messages are signed, malicious nodes cannot generate ERR messages for other nodes. The non-repudiation provided by the signed ERR message allows a node to be verified as the source of each ERR message that it sends. A node which transmits a large number of ERR messages, whether the ERR messages are valid or fabricated, should be avoided.

B. SEAD

Our Secure Efficient Ad hoc Distance vector routing protocol (SEAD) is robust against multiple uncoordinated attackers creating incorrect routing state in any other node, in spite of active attackers or compromised nodes in the network[50]. To support use of SEAD with nodes of limited CPU processing capability and to guard against DoS attacks in which an attacker attempts to cause other nodes to consume excess network bandwidth or processing time, we use efficient one-way hash functions

- **Hash chains*:*** A one-way hash chain is built on a one-way hash function [52][58]. Like a normal hash function, a one-way hash function H maps an input of any length to a fixed-length bit string. Thus, H: $\{0,1\}^* \rightarrow \{0,1\}^p$, where p is the length in bits of the hash function's output. The function H should be simple to compute yet must be computationally infeasible in general to invert. To create a one-way hash chain, a node chooses a random $x \in \{0,1\}^p$ and computes the list of values h_0, h_1, h_2, h_3……………………..h_n, where $h_0= x$, and $h_i = H(h_i -1)$ for $0 < i \leq n$, for some n. The node at initialization generates the elements of its hash chain using this recurrence, in order of increasing subscript i; over time, it uses certain elements of the chain to secure its routing updates. In using these values, the node progresses in order of decreasing subscript i within the generated chain. Given an existing authenticated element of a one-way hash chain, we can verify elements later in the sequence of use within the chain (further on, in order of decreasing subscript). For example, given an authenticated h_i value, a node can authenticate $h_i -3$ by computing $H(H(H(h_i -3)))$ and verifying that the resulting value equals h_i. To use one-way hash chains for authentication, we assume some mechanism for a node to distribute an authentic element such as h_n from its generated hash chain.

SEAD for authenticating an entry in a routing update uses the *sequence number* in that entry to determine a contiguous group of m elements from that destination node's hash chain, one element of which must be used to authenticate that routing update. The particular element from this group of elements that must be used to authenticate the entry is determined by the *metric* value being sent in that entry. Specifically, if a node's hash chain is the sequence of values h_0, h_1, h_2, h_3……………………..h_n and n is divisible by m, then for a sequence number i in some routing update entry, let $k=n/m -i$. An element from the group of elements h_{km}, h_{km}+1,……………….h_{km+m-1} from this hash chain is used to authenticate the entry; if the metric value for this entry is j, $0 \leq j \leq m$, then the value h_{km+j} here is used to authenticate the routing update entry for that sequence number.

C. SORP

OSPF is a link state routing protocol used within one autonomous system (AS) or routing domain. It creates a global network topology in three which are following

- **Phase I**: Neighbor and Adjacency Establishment A router broadcasts periodically a Hello packet to discover its neighboring routers. After the neighboring routers establish connections, they synchronize their databases with each other through a Database Exchange Process.
- **Phase II**: Information Exchange by LSA Flooding A router assembles the link state information about its local neighborhood into a Link State Advertisement (LSA) and floods it to the whole network.
- **Phase III**: Calculate Shortest Route using Link State Database After a router collects all the link state information, it calculates a shortest path tree with itself as the root by using Dijkstra algorithm and forms a complete structure of routing in the network. OSPF divides an AS into groups of routers called *areas*.

A two level hierarchy among these areas is established, with the top level defined as the backbone area and the second level consisting of many areas attached to the backbone. Routers belonging to a single area are called *internal routers*. Routers that belong to more than one area are called Area Border Routers (ABR). All ABRs belong to the backbone and several of the routers, within an area or within the backbone, which exchange information with an external autonomous system, are known as Autonomous System Boundary Routers (ASBR). Security Strong Points of OSPF routing protocol, some inherent properties of OSPF make it very robust to failures and some attacks.

- **Flooding And Information Least Dependency:** As we mentioned above, OSPF uses flooding for the dissemination of LSAs. This makes sure that within the same *area* all the routers have the identical topological database. Even if a router goes down, other routers can still exchange their link state information provided that an alternate path exists. Furthermore the link state information propagated in the network is the raw message generated by the original router instead of the summarized information from neighbors, which is the situation for distance vector routing. This makes it easy to protect the authenticity of the information.

- **Hierarchy Routing and Information Hiding:** OSPF is a two level routing protocol which are intra-area routing and inter-area routing. ABRs connect to backbone and exchange summarized area information. Since intra-area routing depends only on information from within that area, it is not vulnerable to problems out of the area. And problems in one area will not influence the intra-area routing of other areas and inter-area routing among other areas. So hierarchy routing has security advantage.

D. SRP

Secure Routing Protocol [4][13] (Lightweight Security for DSR), which we can use with DSR to design SRP as an extension header that is attached to ROUTE REQUEST and ROUTE REPLY packets. SRP doesn't attempt to secure ROUTE ERROR packets but instead delegates the route-maintenance function to the Secure Route Maintenance portion of the Secure Message Transmission protocol. SRP uses a sequence number in the REQUEST to ensure freshness, but this sequence number can only be checked at the target. SRP requires a security association only between communicating nodes and uses this security association just to authenticate ROUTE REQUESTS and ROUTE REPLYS through the use of message authentication codes. At the target, SRP can detect modification of the ROUTE REQUEST, and at the source, SRP can detect modification of the ROUTE REPLY.

Because SRP requires a security association only between communicating nodes, it uses extremely lightweight mechanisms to prevent other attacks. For example, to limit flooding, nodes record the rate at which each neighbor forwards ROUTE REQUEST packets and gives priority to REQUEST packets sent through neighbors that less frequently forward REQUEST packets.
SRP authenticates ROUTE REPLYS from intermediate nodes using shared group keys or digital signatures. When a node with a cached route shares a group key with (or can generate a digital signature verifiable by) the initiator of the REQUEST, it can use that group key to authenticate the REPLYS. The authenticator, which is either a message authentication code, computed using the group key or a signature is called the intermediate node reply token. The signature or MAC is computed over the cache REPLY.

E. SECURE AODV

The SecAODV [54] implements two concepts secure binding between IPv6 addresses and the independent of any trusted security service, Signed evidence produced by the originator of the message and signature verification by the destination, without any form of delegation of trust. The SecAODV implementation follows Tuominen's design which uses two kernel modules ip6_queue, ip6_nf_aodv, and a user space daemon AODV. The AODV daemon then generates a 1024-bit RSA key pair. Using the public key of this pair, the securely bound global and site-local IPv6 addresses are generated.

The AODV protocol is comprised of two basic mechanisms, route discovery and maintenance of local connectivity. The SecAODV protocol adds security features to the basic AODV mechanisms, but is otherwise identical. A source node that requests communication with another member of the MANET referred to as a destination D initiates the process by constructing and broadcasting a signed route request message RREQ. The format of the SecAODV RREQ message differs from the one proposed in [18], it additionally contains the RSA public key of the source node S and is digitally signed to ensure authenticity and integrity of the message. Upon receiving a RREQ message, each node authenticates the source S, by verifying the message integrity and by verifying the signature against the provided public key. Upon successful verification, the node updates its routing table with S's address and the forwarding node's address. If the message is not addressed to it, it rebroadcasts the RREQ.

F. BISS

Building Secure Routing out of an Incomplete Set of Security Associations (BISS) [38], the sender and the receiver can establish a secure route, even if, prior to the route discovery, only the receiver has security associations established with all the nodes on the chosen route. Thus, the receiver will authenticate route nodes directly through security associations. The sender, however, will authenticate directly the nodes on the route with which it has security associations, and indirectly (by exchange of certificates) the node with which it does not have security associations. The operation of BISS ROUTE REQUEST relies on mechanisms similar to direct route authentication protocols. When an initiator sends a ROUTE REQUEST, it signs the request with its private key and includes its public key *PKI* in the request along with a certificate *cI* signed by the central authority binding its id with *PKI* .

This enables each node on the path to authenticate the initiator of the ROUTE REQUEST. The ROUTE REQUEST message contains the id of the target node. The node that receives this ROUTE REQUEST authenticates the initiator (by verifying the signature on the message), and tries to authenticate the target directly through security associations that it has. Only if a node can successfully authenticate both the initiator and the target will the node broadcast the message further. In BISS, we use similar route request data authentication mechanisms as in Ariadne.

G. SLSP

The Secure Link State Protocol (SLSP) [30] for mobile ad hoc networks is responsible for securing the discovery and distribution of link state information. The scope of SLSP may range from a secure neighborhood discovery to a network-wide secure link state protocol. SLSP nodes disseminate their link state updates and maintain topological information for the subset of network nodes within *R* hops, which is termed as their *zone* . Nevertheless, SLSP is a self-contained link state discovery protocol, even though it draws from, and naturally fits within, the concept of hybrid routing. To counter adversaries, SLSP protects link state update (*LSU*) packets from malicious alteration, as they propagate across the network.

It disallows advertisements of non-existent, fabricated links, stops nodes from masquerading their peers, strengthens the robustness of neighbor discovery, and thwarts deliberate floods of control traffic that exhausts network and node resources. To operate efficiently in the absence of a central key management, SLSP provides for each node to distribute its public key to nodes within its zone. Nodes periodically broadcast their certified key, so that the receiving nodes validate their subsequent link state updates. As the network topology changes, nodes learn the keys of nodes that move into their zone, thus keeping track of a relatively limited number of keys at every instance. SLSP defines a secure neighbor discovery that binds each node *V* to its Medium Access Control (*MAC*) address and its *IP* address, and allows all other nodes within transmission range to identify *V* unambiguously, given that they already have *EV*. Nodes advertise the state of their incident links by broadcasting periodically signed link state updates (*LSU*). SLSP restricts the propagation of the *LSU* packets within the zone of their origin node. Receiving nodes validate the updates, suppress duplicates, and relay previously unseen updates that have not already propagated *R* hops. Link state information acquired from validated *LSU* packets is accepted only if both nodes incident on each link advertise the same state of the link.

H. TIARA

Techniques for Intrusion-Resistant Ad Hoc Routing Algorithms (TIARA) mechanisms protect ad hoc networks against denial-of-service (DoS) attacks launched by malicious intruders. TIARA addresses two types of attacks on data traffic which are flow disruption and resource depletion. The innovation is following

- Routing algorithm independent approach for dealing with flow disruption and resource depletion attacks
- Fully distributed, self configuring firewall confines impact of DoS attack to immediate neighborhood of offending node
- Intrusion-resistant overlay routing reconfigures routes to circumvent malicious nodes

Wireless Router Extension implementation architecture enables TIARA survivability mechanisms to be easily incorporated within existing wireless IP routers.

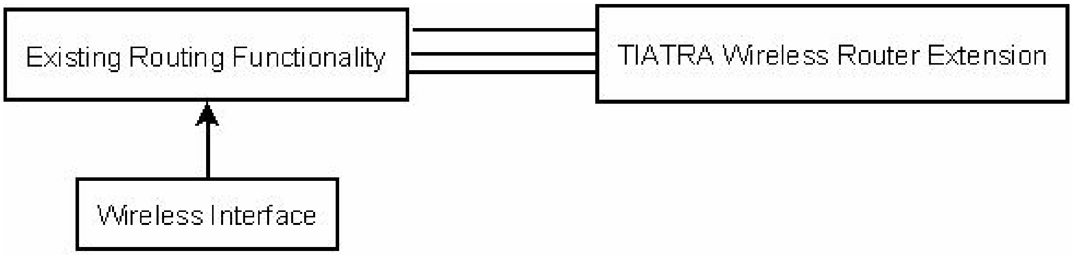

Existing Routing Functionality
TIATRA Wireless Router Extension
Wireless Interface

FIGURE 2: WIRELESS ROUTER EXTENSION IN TIARA

I. ARIADNE

A Secure On Demand Routing Protocol for Ad Hoc Networks (ARIADNE) using the TESLA [43][44] broadcast authentication protocol for authenticating routing messages, since TESLA is efficient and adds only a single message authentication code (MAC) to a message for broadcast authentication. Adding a MAC (computed with a shared key) to a message can provide secure authentication in point-to-point communication; for broadcast communication, however, multiple receivers need to know the MAC key for verification, which would also allow any receiver to forge packets and impersonate the sender. Secure broadcast authentication thus requires an asymmetric primitive, such that the sender can generate valid authentication information, but the receivers can only verify the authentication information. TESLA differs from traditional asymmetric protocols such as RSA in that TESLA achieves this asymmetry from clock synchronization and delayed key disclosure, rather than from computationally expensive one-way trapdoor functions.

Design and evaluation of Ariadne, a new ad hoc network routing protocol that provides security against one compromised node and arbitrary active attackers, and relies only on efficient *symmetric* cryptography [49]. Ariadne operates on-demand, dynamically discovering routes between nodes only as needed; the design is based on the basic operation of the DSR protocol. Rather than generously applying cryptography to an existing protocol to achieve security, however, we carefully re-designed each protocol message and its processing. The security mechanisms we designed are highly efficient and general, so that they should be applicable to securing a wide variety of routing protocols.

This article presents the TESLA (Timed Efficient Stream Loss-tolerant Authentication) broadcast authentication protocol, an efficient protocol with low communication and computation overhead, which scales to large numbers of receivers, and tolerates packet loss. TESLA is based on loose time synchronization between the sender and the receivers. TESLA broadcast authentication protocol have the following requirements: Low computation overhead for generation and Verification of authentication information. Low communication overhead is limited buffering required for the sender and the receiver, hence timely authentication for each individual packet which are Robustness to packet loss, Scales to a large number of receivers.

J. SAR

Security-Aware ad hoc Routing (SAR) that incorporates security attributes as parameters into ad hoc route discovery. SAR enables the use of security as a negotiable metric to improve the relevance of the routes discovered by ad hoc routing protocols. We assume that the base protocol is an on demand protocol similar to AODV or DSR. In the original protocol, when a node wants to communicate with another node, it broadcasts a Route Request or RREQ packet to its neighbors.

The RREQ is propagated to neighbors of neighbors and so on, using controlled flooding. The RREQ packets set up a reverse path to the source of the RREQ on intermediate routers that forward this packet. If any intermediate node has a path already to the RREQ destination, then this intermediate node replies with a Route Reply or RREP packet, using the reverse path to the source [58]. Otherwise, if there exists a route (or connectivity) in the ad hoc network, the RREQ packet will eventually reach the intended destination. The destination node generates a RREP packet, and the reverse path is used to set up a route in the forward direction.

In SAR, we embed our security metric into the RREQ packet itself, and change the forwarding behavior of the protocol with respect to RREQs. Intermediate nodes receive an RREQ packet with a particular security metric or trust level. SAR ensures that this node can only process the packet or forward it if the node itself can provide the required security or has the required authorization or trust level. If the node cannot provide the required security, the RREQ is dropped. If an end-to-end path with the required security attributes can be found, a suitably modified RREP is sent from an intermediate node or the eventual destination. SAR can be implemented based on any on-demand ad-hoc routing protocol with suitable modification. In this paper, use AODV[40] as our platform to implement SAR.

V. COMPARISONS OF SECURE PROTOCOLS

At the last we provide the comparison of different secure routing protocols of ad hoc network using table 1 and table 2. In table 1 shows defense against different type of attack. Comparison shows which protocol is better in different type of attacks. For example replay attack cover by ARAN but it is not coverable by RAP [58].

Attack	Protocol							
	ARAN	SRP	SEAD	ARIADEAN	SAODV	SLSP	OSRP	RAP
Location Disclosure	No	No	No	No	No	No	No	No
Black- Hole	No	No	No	No	No	No	Yes	No
Replay	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Worm hole	No	No	No	No	No	No	No	No
Black mail	NA	NA	NA	NA	NA	NA	NA	NA
Denial of services	No	Yes	Yes	Yes	No	Yes	No	No
Routing table poisoning	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Rushing attacks	Yes	No	Yes	Yes	No	No	No	Yes

Table 2: DEFENSE AGAINST ATTACK

Table 3 shows the proposed solution according to the requirement as well as shows the characteristics of different routing protocols for different operation parameter. Proposed solution describe for protocols used to provide security in adhoc and routing approach used here for adhoc routing protocols used in secure routing protocol. For loop freedom use protocol a sequence no which avoid the count infinite problem. Routing algorithms have used many different metrics to determine the best route. Sophisticated routing algorithms can base route selection on multiple metrics, combining them in a single metric. All the following metrics have been used: Path length, Reliability, Delay, Bandwidth, Load and Communication cost. The shortest path problem is the problem of finding a path between two nodes such that the sum of the cost of its constituent channel is minimized.

PROPOSED SOLUTON	ROUT-ING APPROACH	LOOP FREEDOM	ROUT-ING METRIC	SHORT-EST PATH	REPLY TO ROUTE REQUESTS	REQUIREMENTS
ARAN	On-demand	Yes	None	Optional	No	Online trusted certification authority.
SAR	On-demand	Depends on the selected Security requirement.	A security requirement	No	No	Key distribution or secret sharing mechanism.
SRP	On-demand	Yes	Distance	No	Optional	Existence of a security association between each source and destination node.
SEAD	Table-driven	Yes	Distance	No	No	Clock synchronization,
ARIADNE	On-demand	Yes	Distance	No	No	TESLA keys are distributed to the participating nodes via an online key distribution center.
SAODV	On-demand	Yes	Distance	No	Optional	Online key management scheme for the acquisition and verification of public keys.
TIARA	On-demand	Depends on the basis protocol	Distance	Depends on the basis protocol	Depends on the basis protocol	Online public key infrastructure.
SLSP	Table-driven	Yes	Distance	No	No	Nodes must have their public keys certified by a TTP
BISS	On-demand	Yes	Distance	No	No	The target node of a route discovery must share secret keys with all the intermediate nodes
IPsec	NA	NA	NA	NA	NA	Prearranged common secrets between each pair of nodes, or an online trusted third party

Table 3: OPERATIONAL REQUIREMENT AND PARAMETER FOR THE PROPOSED SOLUTION

VI. CONCLUSION

We have presented an overview of the existing security scenario in the Ad-Hoc network environment. Key management, Ad-hoc routing of wireless Ad-hoc networks were discussed. Ad-hoc networking is still a raw area of research as can be seen with the problems that exist in these networks and the emerging solutions. The key management protocols are still very expensive and not fail safe. Several protocols for routing in Ad-hoc networks have been proposed. There is a need to make them more secure and robust to adapt to the demanding requirements of these networks. The flexibility, ease and speed with which these networks can be set up imply they will gain wider application. This leaves Ad-hoc networks wide open for research to meet these demanding application.

VII. REFERENCES

[1] Adrian Perrig Ran Canetti J. D. Tygar Dawn Song "*The TESLA Broadcast Authentication Protocol*", UC Berkeley and IBM Research.

[2] Ajay Mahimkar, R. K. Shyamasundar "*S-MECRA A Secure Energy-Efficient Routing Protocol for Wireless Ad Hoc Networks*" IEEE 2004

[3] Alia Fourati, Khaldoun Al Agha, Hella Kaffel Ben Ayed "Secure and Fair Auctions over Ad Hoc Networks" *Int. J. Electronic Business, 2007*

[4] Anand Patwardhan, Jim Parker, Michaela Iorga. Anupam Joshi, "*Tom Karygiannis, Secure Routing and Intrusion Detection in Ad Hoc Networks*" 3rd International Conference on Pervasive Computing and Communications (PerCom 2005), Kauai Island, Hawaii***.***

[5] Bing Wua, Jie Wua, Eduardo B. Fernandeza, Mohammad Ilyasa, Spyros Magliveras, "*Secure and efficient key management in mobile ad hoc networks*" Journal of Network and Computer Applications 30 (2007) 937–954

[6] Bissias, G.D., Liberatore, M., Jensen, D., Levine, B.N., "Privacy vulnerabilities in encrypted HTTP streams" In *Proc. Privacy Enhancing Technologies Workshop* (PET 2005).

[7] C. E. Perkins, E. M. Royer, and S. R. Das, "*Ad Hoc On-Demand Distance Vector (AODV) Routing*," IETF Mobile Ad Hoc Networks Working Group, Internet Draft, work in progress, 17 February 2003.

[8] F. Hu and N. K. Sharma, "*Security Considerations in Ad Hoc Networks*," to be appeared in Ad Hon Network, 2004.

[9] F. Anjum*,* Anup K. Ghosh*,* nada golmie*,* paul kolodzy*,* radha poovendran*,* rajeev shorey*,* d. Lee*, j-sac, "Security in Wireless Ad hoc Networks*", ieee journal on selected areas in communications, vol. 24, no. 2, February 2006.

[10] H.-A. Wen, C.-L. Lin, and T. Hwang, "*Provably Secure Authenticated Key Exchange Protocols for Low Power Computing Clients*," Computers and Security, vol. 25, pp. 106-113, 2006.

[11] Haiyun Luo, Petros Zerfos, Jiejun Kong, Songwu Lu, Lixia Zhang, "*Self-securing Ad Hoc Wireless Networks*", 7th IEEE Symp. on Comp. and Communications (ISCC), Taormina, 2002.

[12] Hongmei Deng, Wei Li, and Dharma P. Agrawal, "Routing Security in Wireless Ad Hoc Networks", IEEE Communications Magazine October 2002.

[13] Huaizhi Li Zhenliu Chen Xiangyang Qin, "*Secure Routing in Wired Networks and Wireless Ad Hoc Networks" IEEE, 2004.*

[14] Huaizhi Li, Mukesh Singha, *"Trust Management in Distributed Systems"* IEEE Computer Society February 2007.

[15] I. Aad, J.-P. Hubaux, and E-W. Knightly, *"Denial of Service Resilience in Ad Hoc Networks,"* Proc. MobiCom, 2004.

[16] J. Nam, S. Cho, S. Kim, and D. Won, *"Simple and Efficient Group Key Agreement Based on Factoring"* Proc. Int'l Conf. Computational Science and Its Applications (ICCSA '04), pp. 645-654, 2004.

[17] J. Parker, J. L. Undercoffer, J. Pinkston, and A. Joshi., "*On Intrusion Detection in Mobile Ad Hoc Networks"*. In 23rd IEEE International Performance Computing and Communications Conference Workshop on Information Assurance. IEEE, April 2004.

[18] Jeremy J. Blum, Member, IEEE, and Azim Eskandarian, Member, IEEE, "*A Reliable Link-Layer Protocol for Robust and Scalable Intervehicle Communications"* IEEE Transactions On Intelligent Transportation Systems, vol. 8, no. 1, March 2007.

[19] Jung-San Lee, Chin-Chen Chang, "*Secure communications for cluster-based ad hoc networks using node identities"* Journal of Network and Computer Applications 22 October 2006

[20] K. Balakrishnan, J. Deng, and P.K. Varshney, "*TWOACK: Preventing Selfishness in Mobile Ad Hoc Networks"* Proc. IEEE Wireless Comm. and Networking Conf. (WCNC '05), Mar. 2005.

[21] Karan Singh, Rama Shankar Yadav, Raghav Yadav, R. Shiva Kumaran, *"Adaptive Multicast Congestion Control "* HIT haldia March 2007.

[22] Kejun Liu, Jing Deng, Member, IEEE, Pramod K. Varshney, Fellow, IEEE, and Kashyap Balakrishnan, Member, IEEE, "*An Acknowledgment-Based Approach for the Detection of Routing Misbehavior in MANETs"* IEEE Transaction on Mobile Computing, VOL. 6, NO. 5, May 2007

[23] L. Buttyan and J.-P. Hubaux, "*Security and Cooperation in Wireless Networks*," http://secowinet.epfl.ch/, 2006.

[24] M. Bechler, H.-J. Hof, D. Kraft, F. Pählke, L. Wolf, "A Cluster-Based Security Architecture for Ad Hoc Networks" IEEE INFOCOM 2004.

[25] Mike Just_ Evangelos Kranakis Tao Wan, *"Resisting Malicious Packet Dropping in Wireless Ad Hoc Networks"* Internet draft: draft-ietfitrace-03.txt, January 2003.

[26] Mohammad Al-Shurman and Seong-Moo Yoo, Seungjin Park, *"Black Hole Attack in Mobile Ad Hoc Networks" ACMSE'04*, April 2-3, 2004, Huntsville, AL, USA.

[27] Muhammad Bohio, Ali Miri, E.cient, "*Identity-based security schemes for ad hoc network routing protocols*" Ad Hoc Networks 2 (2004) 309–317

[28] Nikos Komninos, Dimitris Vergados, Christos Douligeris, "*Layered security design for mobile ad hoc networks"* journal computers & security 25, 2006 , pp. 121 – 130.

[29] Nobuo Okabe, Shoichi Sakane, Kazunori Miyazawa, Ken'ichi Kamada, "*Extending a Secure Autonomous Bootstrap Mechanism to Multicast Security*" 2007 International Symposium on Applications and the Internet Workshops (SAINTW'07).

[30] P. Papadimitratos and Z.J. Haas, "*Secure Link State Routing for Mobile Ad Hoc Networks*" Proc. IEEE Workshop on Security and Assurance in Ad Hoc Networks, IEEE Press, 2003, pp. 27–31.

[31] Panagiotis Papadimitratos , Zygmunt J. Haas, "*Secure message transmission in mobile ad hoc networks, Ad Hoc Networks*" IEEE 2003, 193–209 .

[32] R. Hinden and S. Deering. RFC 3513, "*Internet Protocol Version 6 (IPv6) Addressing Architecture*" April 2003.

[33] R. Mahajan, M. Rodrig, D. Wetherall, and J. Zahorjan, "*Sustaining Cooperation in Multi-Hop Wireless Networks*," Proc. Second Symp. Networked Systems Design and Implementation, Apr. 2005.

[34] R. Shiva Kumaran, Rama Shankar Yadav, Karan Singh "*Multihop wireless LAN* " HIT haldia March 2007.

[35] *S. Holeman, G. Manimaran, J. Davis, A. Chakrabarti, Differentially secure multicasting and its implementation methods,* Computers & Security Vol 21, No 8, pp736-749, 2002.

[36] S.M. Bellovin, M. Leech, and T. Taylor. ICMP Traceback Messages. Internet draft: draft-ietfitrace 03.txt, January 2003.

[37] Seung Yi, Prasad Naldurg, Robin Kravets, "*A Security-Aware Routing Protocol for Wireless Ad Hoc Networks*" IEEE 2003.

[38] Srdjan Capkun and Jean-Pierre Hubaux, "*Building Secure Routing out of an Incomplete Set of Security Associations" WiSE'03,* September 19, 2003, San Diego, California, USA.

[39] Stallings, W., *Wireless Communications and Networks*, 2nd Ed., Prentice Hall, 2005.

[40] T. Aura. Internet Draft: Cryptographically Generated Addresses (CGA). http://www.ietf.org/proceedings/ 04mar/I-D/draftietf- send-cga-05.txt, February 2004.

[41] Thomas S. Messerges, ohnas Cukier, Tom A.M. Kevenaar, Larry Puhl, Rene truik, Ed Callaway, *"A Security Design for a General Purpose, Self-Organizing, Multihop Ad Hoc Wireless Network*" 1st ACM Workshop Security of Ad Hoc and Sensor Networks Fairfax, Virginia 2003

[42] Tzonelih Hwang, Kuo-Chang Lee, Chuan-Ming Li, "*Provably Secure Three-Party Authenticated Quantum Key Distribution Protocols*" IEEE Transactions On Dependable And Secure Computing, vol. 4, no. 1, January-March 2007.

[43] Uppsala University, The Ad hoc Protocol Evaluation (APE) test bed, release 0.3, downloaded Nov. 2005.

[44] Uppsala University, The AODV-UU implementation version 0.8.1, downloaded Nov. 2005.

[45] W. Xu, T. Wu, "*TCP Issues in Mobile Ad Hoc Networks: Challenges and Solutions*", Journal of Computer Science and Technology, 2006, 21.

[46] Weiqiang Xu and Tiejun Wu, "A Congestion Control Algorithm for Ad Hoc Networks: A Dual Decomposition Approach" 6th World Congress on Intelligent Control and Automation, June 21 - 23, 2006, Dalian, China.

[47] Wright, C.V., Monrose, F., Masson, G.M., "*HMM profiles for network traffic classification*" in *Proc.* ACM Workshop on Visualization and Data Mining for Computer Security, pp. 9–15, Oct. 2004.

[48] Wright, C.V., Monrose, F., Masson, G.M., "*Towards better protocol identification using profile HMMs"* JHU Technical Report JHU-SPAR051201, 14p., June, 2005.

[49] Y. Xue and K. Nahrstedt, "*Providing Fault-Tolerant Ad-Hoc Routing Service in Adversarial Environments,"* Wireless Personal Comm., vol. 29, nos. 3-4, pp. 367-388, 2004.

[50] Y.-C. Hu, D. B. Johnson, and A. Perrig., *"SEAD: Secure Efficient Distance Vector Routing for Mobile Wireless Ad Hoc Networks"* In Proceedings of the Fourth IEEE Workshop on Mobile Computing Systems and Applications, page 3. IEEE Computer Society, 2002.

[51] Yang Mingxi, Li Layuan, Fang Yiwei, "Securing multicast route discovery for mobile ad hoc networks" SpringerLink, February 17, 2007.

[52] Yih-chun hu, adrian perrig, "*A Survey of Secure Wireless ad hoc routing*" IEEE security & privacy May-June 2004

[53] Yih-Chun Hu, Adrian Perrig, and David B. Johnson., "*Packet Leashes A Defense against Wormhole Attacks in Wireless Ad Hoc Networks*" In Proceedings of the Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2003), April 2003. To appear.

[54] Yih-Chun Hu, Adrian Perrig, David B. Johnson Ariadne: "*A Secure On-Demand Routing Protocol for Ad Hoc Networks*" *MobiCom'02,* September 23–26, 2002, Atlanta, Georgia, USA.

[55] Yih-Chun Hu, Adrian Perrig, David B. Johnson, "Rushing Attacks and Defense in Wireless Ad Hoc Network Routing Protocols" *WiSe 2003,* September 19, 2003, San Diego, California, USA.

[56] Yuh-Ren Tsai, Shiuh-Jeng Wang, *"Routing Security and Authentication Mechanism for Mobile Ad Hoc Networks"* Chung-Shan Institute of Science and Technology, Taiwan, R.O.C., under Grant BC-93-B14P and the National Science Council, Taiwan, R.O.C., IEEE 2004.

[57] S.W. Smith, , "*A Case (Study) For Usability in Secure Email Communication*" IEEE Computer Society 2007

[58] Patroklos g. Argyroudis and donal o'mahony, "*Secure Routing for Mobile Ad hoc Networks*", IEEE Communications Surveys & Tutorials Third Quarter 2005.

A REVIEW OF STUDIES ON MACHINE LEARNING TECHNIQUES

Yogesh Singh ys66@rediffmail.com
Prof & Dean
Guru Gobind Singh IP University
Delhi, 110006, India

Pradeep Kumar Bhatia pk_bhatia2002@yahoo.com
Reader, Department of Computer Science & Engineering
Guru Jambsheshwar University of Science & Technology
Hisar, Haryana, 125001,India

Omprakash Sangwan sangwan_op@aiit.amity.edu
Head, CISCO Regional Networking Academy
Amity Institute of Information Technology
Amity University, Uttarpradesh, 201303,India

Abstract

This paper provides an extensive review of studies related to expert estimation of software development using Machine-Learning Techniques (MLT). Machine learning in this new era, is demonstrating the promise of producing consistently accurate estimates. Machine learning system effectively “learns” how to estimate from training set of completed projects. The main goal and contribution of the review is to support the research on expert estimation, i.e. to ease other researchers for relevant expert estimation studies using machine-learning techniques. This paper presents the most commonly used machine learning techniques such as neural networks, case based reasoning, classification and regression trees, rule induction, genetic algorithm & genetic programming for expert estimation in the field of software development. In each of our study we found that the results of various machine-learning techniques depends on application areas on which they are applied. Our review of study not only suggests that these techniques are competitive with traditional estimators on one data set, but also illustrate that these methods are sensitive to the data on which they are trained.

Keywords: Machine Learning Techniques (MLT), Neural Networks (NN), Case Based Reasoning (CBR), Classification and Regression Trees (CART), Rule Induction, Genetic Algorithms and Genetic Programming.

1. INTRODUCTION

The poor performance results produced by statistical estimation models have flooded the estimation area for over the last decade. Their inability to handle categorical data, cope with missing data points, spread of data points and most importantly lack of reasoning capabilities has triggered an increase in the number of studies using non-traditional methods like machine learning techniques.

Machine Learning is the study of computational methods for improving performance by mechanizing the acquisition of knowledge from experience [18]. Expert performance requires much domain specific knowledge, and knowledge engineering has produced hundreds of AI expert systems that are now used regularly in industry. Machine Learning aims to provide increasing levels of automation in the knowledge engineering process, replacing much time-consuming human activity with automatic techniques that improve accuracy or efficiency by discovering and exploiting regularities in training data. The ultimate test of machine learning is its ability to produce systems that are used regularly in industry, education, and elsewhere. Most evaluation in machine learning is experimental in nature, aimed at showing that the learning method leads to performance on a separate test set, in one or more realistic domains, that is better than performance on that test set without learning.

At a general level, there are two types of machine learning: inductive, and deductive. Deductive learning works on existing facts and knowledge and deduces new knowledge from the old. Inductive machine learning methods create computer programs by extracting rules and patterns out of massive data sets. Inductive learning takes examples and generalizes rather than starting with existing knowledge one major subclass of inductive learning is concept learning. This takes examples of a concept and tries to build a general description of the concept. Very often, the examples are described using attribute-value pairs.

Machine learning overlaps heavily with statistics. In fact, many machine-learning algorithms have been found to have direct counterparts with statistics. For example, boosting is now widely thought to be a form of stage wise regression using a specific type of loss function. Machine learning has a wide spectrum of applications including natural language processing, search engines, medical diagnosis, bioinformatics and cheminformatics, detecting credit card fraud, stock market analysis, classifying DNA sequences, speech and handwriting recognition, object recognition in computer vision, game playing and robot locomotion.

In our study we concentrate on the various paradigms, which are used in machine learning. Our review also examines the comparative study of machine learning technique with suitable application area.

This paper is organized as follows: In section 2 we discuss about the use of Neural Network in machine learning. CBR with application area is presented in section 3. CART is another efficient learning method described in section 4. Another paradigm rule induction is highlighted in section 5. In section 6 the impact of genetic algorithm and programming are discussed. Section 7 presents the discussion on various machine-learning techniques and conclusions and future direction are presented in section 8.

2. NEURAL NETWORKS

Neural networks have been established to be an effective tool for pattern classification and clustering [8, 15]. There are broadly two paradigms of neural learning algorithms namely supervised and unsupervised. Unsupervised neural algorithms are best suited for clustering patterns on the basis of their inherent characteristics [8, 14]. There are three major approaches for unsupervised learning: -

(a) Competitive Learning
(b) Self Organizing feature Maps
(c) ART Networks

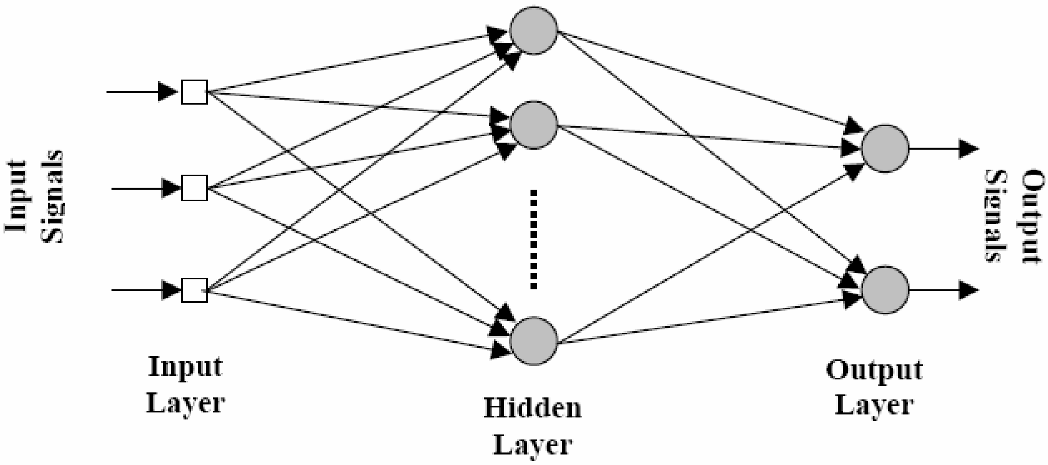


Figure 1: The architecture of the neural network

The other paradigm of neural learning is the so-called supervised learning paradigm. These networks have been established to be universal approximators of continuous/discontinuous functions and therefore they are suitable for usage where we have some information about the input-output map to be approximated. A set of data (Input-Output information) is used for training the network. Once the network has been trained it can be given any input (from the input space of the map to be approximated) and it will produce an output, which would correspond to the expected output from the approximated mapping. The quality of this output has been established to correspond arbitrarily close to the actual output desired owing to the generalization capabilities of these networks.
The activation function used is the log-sigmoid function as given in [9] can be expressed as: -

$$\Phi(a) = \frac{1}{1+e^{-a}} \qquad (1)$$

Where

$$a = \sum_{i=1}^{N} \mathbf{w}\,\mathbf{x} \qquad (2)$$

w's are the synaptic coefficients and x's are the outputs of the previous layer. For the hidden layer x's correspond to the input of the network while for the output layer x's correspond to the output of the hidden layer. The network is trained using the error back propagation algorithm [9] .The weight update rule as given in [9] can be expressed as: -

$$\Delta w_{ji}(n) = \alpha\, \Delta\, w_{ji}(n\text{-}1) + \eta\, \delta_j(n)\, y_i(n) \qquad (3)$$

where α is usually a positive number called the momentum constant , η is the learning rate, ∆wji (n) is the correction applied to the synaptic weight connecting the output of neuron i to the input of neuron j at iteration n, δj (n) is the local gradient at nth iteration, yi (n) is the function signal appearing at the output of neuron i at iteration n.

From experimental results we conclude that neural network can be used as test oracle, effort estimation, cost estimation, size estimation & other application areas of software engineering [1,7, 12, 13]. However the percentage error that can be tolerated will depend on the specific application for which test case is being design. The architecture and training algorithm will depend upon the space spanned by the test case parameters. There are some other systems like complex simulation in mechanical design, weather and economic forecasting and geological exploration that are built to solve unsolved problems using neural network for which there is no analytical solution.

The primary advantage of using neural network approach is that they are adaptable and nonparametric; predictive models can be tailored to the data at a particular site.

3. CASE BASED REASONING (CBR)

Case Based Reasoning is a technique by which we solve new problems by adapting the solutions from similarly solved problems. We take the instances of solutions from problems that have happened in the past and try to solve new problems by using these cases. Each such solution available to us can be termed as a case [11].

3.1 CBR Process

A general CBR process includes the following four processes.

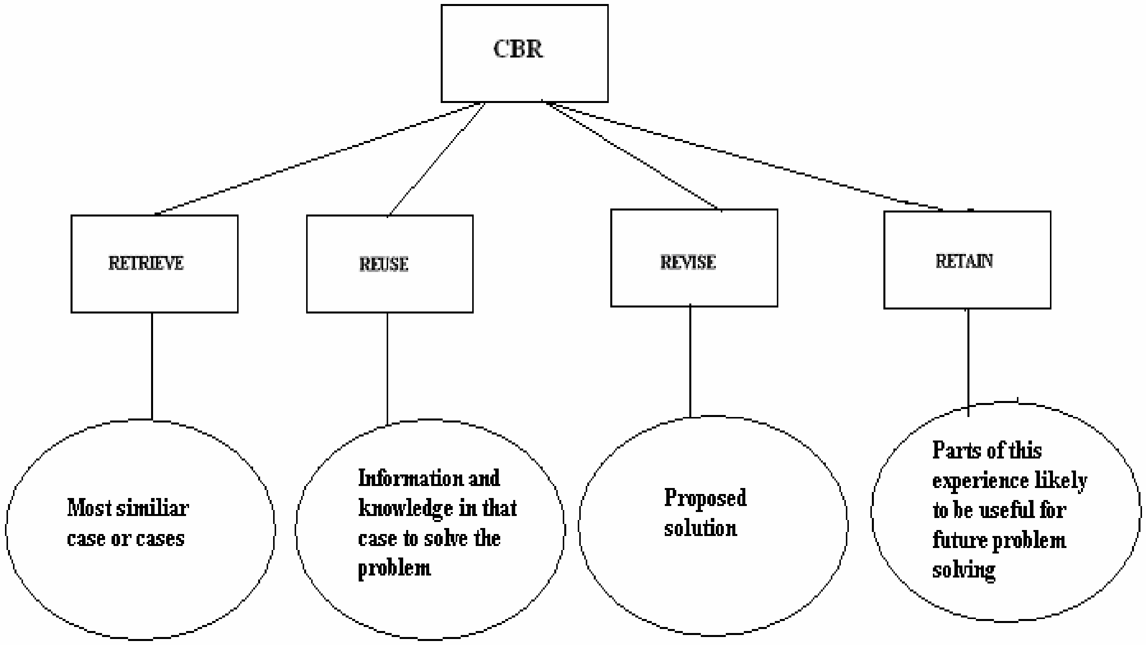

CBR
RETRIEVE
REUSE
REVISE
RETAIN
Most similiar case or cases
Information and knowledge in that case to solve the problem
Proposed solution
Parts of this experience likely to be useful for future problem solving

FIgure 3 A General CBR Process

A new case is defined by the initial description of any problem. This new case is *retrieved* from a collection of previous cases and this retrieved case is then combined with the new case through reuse into a *solved case*. This solved case is nothing but a proposed solution to the defined problem. Once this solution is identified, applying it practically to the real world tests it. This process of testing is termed as *revision* of the problem. Then comes the process of *retain* where useful experience is retained for future reuse and the case base is updated by a new learned case or by modification of some existing cases.

Thus we can say that CBR is a four-step process:
1. RETRIEVE
2. REUSE

3. REVISE
4. RETAIN

The figure: 4 give a brief illustration of the CBR Cycle**:**

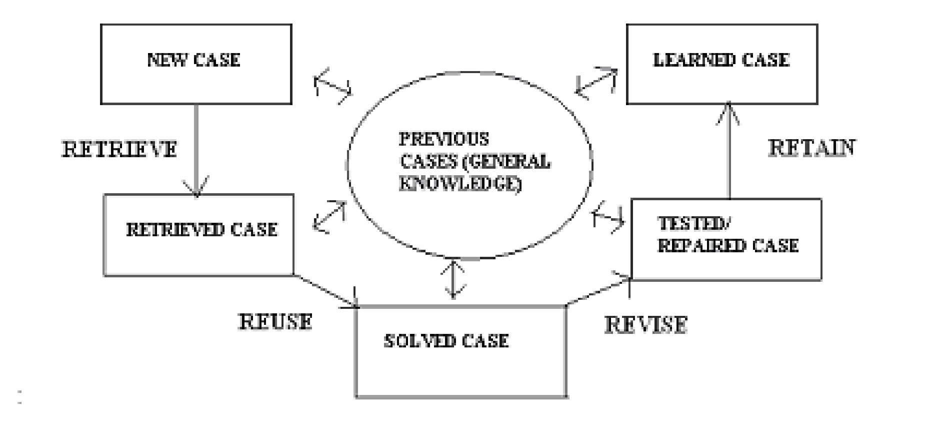


Figure 4 The CBR Cycle

It is clear from the figure that general knowledge plays a crucial in CBR. It supports all the CBR processes. General knowledge here implies domain dependent knowledge as opposed to specific knowledge embodied by cases. For instance in diagnosing a patient by retrieving and reusing the case of a previous patient, a model of anatomy together with casual relationships between pathological states may constitute the general knowledge used by a CBR system.

3.2 Fundamentals of Case Based Reasoning

3.2.1 Case Retrieval

The process of retrieval in CBR cycle begins with the problem description and ends when the best possible case from the set of previous cases has been obtained. The subtasks involved in this particular step include identifying features, matching, searching and selecting the appropriate ones executed in that order. The identification task finds a set of relevant problem descriptors, then the matching task returns those cases that are similar to the new case and finally the selection task chooses the best possible match. Among well-known methods for case retrieval are: nearest neighbor, induction, knowledge guided induction and template retrieval. These methods can be used alone or combined into hybrid retrieval strategies**.**

1) Nearest Neighbour (NN)

NN approach involves the assessment of similarity between stored cases and the new input case, based on matching a weighted sum of features.

2) Induction

This involves generating a decision tree structure to organize the cases in memory by determining which features do the best job in discriminating cases.

3) Knowledge guided induction

By applying knowledge to the induction process by manually identifying case features that are known or thought to affect the primary case feature we perform case retrieval. This approach is frequently used in conjunction with other techniques, because the explanatory knowledge is not always readily available for large case bases.

4) Template retrieval

Template retrieval returns all cases that fit within certain criteria often used before other techniques, such as nearest neighbour, to limit the search space to a relevant section of the case-base.

3.2.2 Case Reuse

This involves obtaining the solved case from a retrieved case. It analyses the differences between the new case and the past cases and then determines what part of the retrieved case can be transferred to the new case. CBR is essentially based on the concept of analogy wherein by analyzing the previous cases we formulate a solution for the new cases [5].

3.2.3 Copy

In the trivial cases of reuse we generally copy the solution of the previous cases and make it the solution for the new cases. But many systems take into consideration the differences between the two cases and use the adaptation process to formulate a new solution based on these differences.

3.2.4 Adaptation

The adaptation process is of two kinds:
Structural adaptation- Adaptation rules are applied directly to the solution stored in cases i.e. reuse past case solution.
Derivational adaptation- Reuse the method that constructed the solution to a past problem.
In structural adaptation we do not use the past solution directly but apply some transformation parameters to construct the solution for the new case. Thus this kind of adaptation is also referred to as transformational adaptation. In derivational adaptation we use the method or algorithm applied previously to solve the new problem [17].

3.2.5 Case Revision

After reusing the past cases to obtain a solution for the new case we need to test that solution. We must check or test to see if the solution is correct. If the testing is successful then we retain the solution, otherwise we must revise the case solution using domain specific knowledge.

3.2.6 Case Retainment- Learning (CRL)

The solution of the new problem after being tested and repaired may be retained into the existing domain specific knowledge. This process is called Case Retainment Learning or CRL. Retaining information involves selecting what information to retain, in what form to retain it, how to index the case for later retrieval from similar problems, and how to integrate the new case in the memory structure.

3.2.7 Case Based Learning

An important feature of CBR is its coupling to learning [2]. Case-based reasoning is also regarded a sub-field of machine learning. Thus, the notion of case-based reasoning does not only denote a particular reasoning method, irrespective of how the cases are acquired, it also denotes a machine learning paradigm that enables sustained learning by updating the case base after a problem has been solved. Learning in CBR occurs as a natural by-product of problem solving. When a problem is successfully solved, the experience is retained in order to solve similar problems in the future. When an attempt to solve a problem fails, the reason for the failure is identified and remembered in order to avoid the same mistake in the future. CBR can be applied to solve real world problems for instance handling of multiple disorders [16] or for engineering sales support [23].

4. CLASSIFICATION AND REGRESSION TREES (CART)

CART is a very efficient machine learning technique. The difference between this technique and other machine learning technique is that CART requires very little input from the analyst. This is in contrast to other technique where extensive input from the analyst, the analysis of interim results and modification of method used is needed. Before going into the details of CART we identify the three classes and two kinds of variables, which are important while defining classification and regression problems.

There are three main classes of variables:

1) Target variable -- The "target variable" is the variable whose values are to be modeled and predicted by other variables. It is analogous to the dependent variable in linear regression. There must be one and only one target variable in a decision tree analysis.

2) Predictor variable -- A "predictor variable" is a variable whose values will be used to predict the value of the target variable. It is analogous to the independent in linear regression. There must be at least one predictor variable specified for decision tree analysis; there may be many predictor variables.

3) Weight variable -- You can specify a "weight variable". If a weight variable is specified, it must a numeric (continuous) variable whose values are greater than or equal to 0 (zero). The value of the weight variable specifies the weight given to a row in the dataset.

There are 2 main kinds of variables:

1) Continuous variables -- A continuous variable has numeric values such as 1, 2, 3.14, -5, etc. The relative magnitude of the values is significant (e.g., a value of 2 indicates twice the magnitude of 1). Examples of continuous variables are blood pressure, height, weight, income, age, and probability of illness. Some programs call continuous variables "ordered" or "monotonic" variables.

2) Categorical variables -- A categorical variable has values that function as labels rather than as numbers. Some programs call categorical variables "nominal" variables. For example, a categorical variable for gender might use the value 1 for male and 2 for female. The actual magnitude of the value is not significant; coding male as 7 and female as 3 would work just as well

CART builds classification and regression trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification).

Regression-type problems: These are generally those where one attempts to predict the values of a continuous variable from one or more continuous and/or categorical predictor variables.

Classification-type problems: These are generally those where one attempts to predict values of a categorical dependent variable from one or more continuous and/or categorical predictor variables.

CART is a non-parametric statistical methodology developed for analyzing classification issues either from categorical or continuous dependent variables [24, 25]. If the dependent variable is categorical, CART produces a classification tree. When the dependent variable is continuous, it produces a regression tree.

4.2 Binary Recursive Partitioning

Consider the problem of selecting the best size and type of laryngoscope blade for pediatric patients undergoing intubations [20]. The outcome variable, the best blade for each patient (as determined by a consulting pediatric airway specialist), has three possible values: Miller 0, Wis-Hipple 1.5, and Mac 2. The two-predictor variables are measurements of neck length and or pharyngeal height. The smallest patients are best incubated with the Miller 0, medium sized patients with the Wis-Hipple 1.5, and the largest patients with the Mac 2.

CART is basically used to avoid the disadvantage of the regression techniques. CART analysis is a form of binary recursive partitioning [20]. The term "binary" implies that each node in a decision tree can only be split into two groups. Thus, each node can be split into two child nodes, in which case the original node is called a parent node. The term "recursive" refers to the fact that the binary partitioning process can be applied over and over again. Thus, each parent node can give rise to two child nodes and, in turn, each of these child nodes may themselves be split, forming additional children. The term "partitioning" refers to the fact that the dataset is split into sections or partitioned.

The figure:5 illustrates this kind of a partitioning. This tree consists of a root node (Node 1), containing all patients. This node is split based on the value of the neck length variable. If the neck length is < 2.45 centimeters, then those patients are put in the first terminal node, denoted Node -1, and the best blade is predicted to be a Miller 0. All other patients are placed in Node 2. The group of patients in Node 2 is initially assigned a Wis-Hipple 1.5 blade but they are also split based on there or pharyngeal height. Those patients with an or pharyngeal height less than 1.75 are placed in terminal Node -2, and assigned a Wis-Hipple 1.5 blade, while those with an or pharyngeal heighṭ 1.75 are placed in terminal Node –3 and assigned a Mac 2 blade.

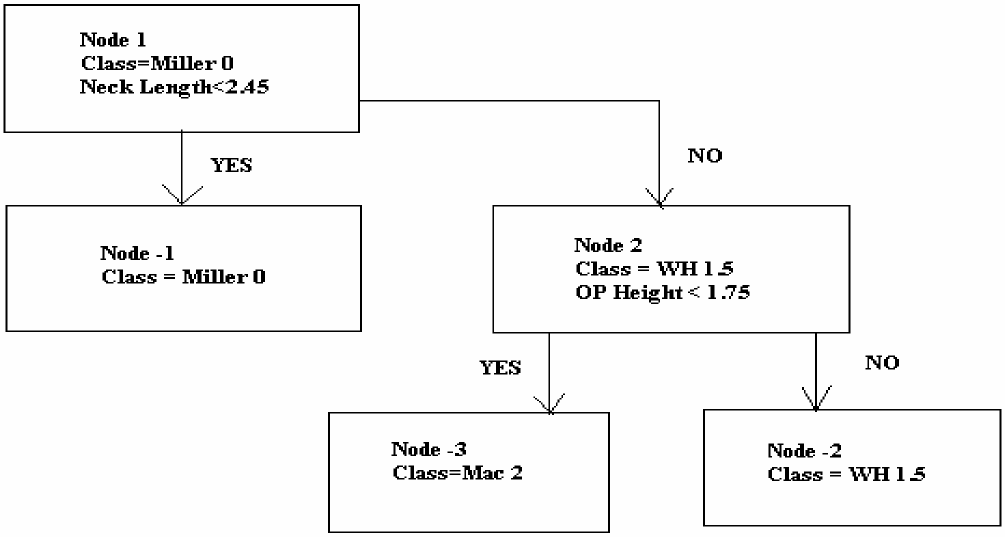


Figure 5 A CART Analysis Tree

4.3 CART Analysis

CART analysis is a tree-building technique, which is unlike traditional data analysis methods. It is ideally suited to the generation of clinical decision rules.

CART Analysis consists of four basic steps: -

1. It consists of tree building, during which a tree is built using recursive splitting of nodes. Each resulting node is assigned a predicted class, based on the distribution of classes in the learning dataset, which would occur in that node and the decision cost matrix. The assignment of a predicted class to each node occurs whether or not that node is remove space subsequently split into child nodes.

2. CART Analysis consists of stopping the tree building process. At this point a “maximal” tree has been produced, which probably greatly over fits the information contained within the learning dataset.

3. It consists of tree “pruning,” which results in the creation of a sequence of simpler and simpler trees, through the cutting off increasingly important nodes.

4. This step consists of optimal tree selection, during which the tree that fits the information in the learning dataset, but does not over fit the information, is selected from among the sequence of pruned trees.

5. RULE INDUCTION

Rule Induction is another very important machine learning method and it is easier because the rules in rule induction are transparent and easy to interpret than a regression model or a trained neural network. This paradigm employs condition-action rules, decision trees, or similar knowledge structures. Here the performance element sorts instances down the branches of the decision tree or finds the first rule whose conditions match the instance, typically using an all-or-none match process [19]. Information about classes or predictions is stored in the action sides of the rules or the leaves of the tree. Learning algorithms in the rule induction framework usually carry out a greedy search through the space of decision trees or rule sets, typically using a statistical evaluation function to select attributes for incorporation into the knowledge structure. Most methods partition the training data recursively into disjoint sets, attempting to summarize each set as a conjunction of logical conditions.

Rule Learning process

If we are given a set of training examples i.e. instances for which classification is known we find a set of classification rules which are used to predict new cases that haven't been presented to the learner before. While deriving these instances the bias imposed by languages must be taken into account such as restrictions imposed while describing data and we must also consider the language used to describe the induced set of rules.
Consider a binary classification problem of classifying instances into classes positive and negative. We are given a data description language, which impose a bias on the data, training examples, a hypothesis language imposing a bias on the induction rules and a coverage function defining when an instance is covered by a rule. Given the above data we need to find a hypothesis defined by a set of rules in a language, which is consistent that it does not cover any negative examples and is complete that it covers all positive examples. Thus in this manner, given the required data and the problem we can determine a set of rules, which classify the instances in that problem. This forms the basis of rule induction.

There are two main approaches to rule induction namely propositional learning and relational rule learning.

Propositional Rule Learning

Propositional rule learning systems are suited for problems in which no substantial relationship between the values of the different attributes needs to be represented. A set of instances with known classifications where each instance is described by values of a fixed collection of attributes is given. The attributes can have either a fixed set of values or take real numbers as values. Given these instances we then construct a set of IF-THEN rules. The output of learning is a hypothesis represented by a set of rules. After the rules have been defined determining the accuracy of such rules and then applying these rules to practical problems analyze their quality. In propositional learning the available data has a standard form with rows being individual records or training examples and columns being properties or attributes to describe the data.

Relational Rule Learning/ Inductive logic Programming (ILP)

When data is stored in several tables then it has a relational database form. In such cases the data has to be transformed into a single table in order to use standard data mining techniques. The most common data transformation approach is to select one table as the main table to be used for learning, and try to incorporate the contents of other tables by summarizing the information contained in the table into some summary attributes, added to the main table. The problem with such single-table transformations is that some information may be lost while the summarization may also introduce artifacts, possibly leading to inappropriate data mining results. What one would like to do is to leave data conceptually unchanged and rather use data mining tools that can deal with multi-relational data. ILP is intended at solving multi-relational data mining tasks.

Thus ILP is to be used for data mining in multi-relational data mining tasks with data stored in relational databases and tasks with abundant expert knowledge of a relational nature. Another important concept within the realm of relational rule learning is that of boosting. Boosting is a particularly robust and powerful technique to enhance the prediction accuracy of systems that learn from examples [22]. Thus boosting helps to improve the overall efficiency of the results obtained.

An example to illustrate Rule Induction

Case Study (Making Credit Decisions)

Loan companies regularly use questionnaires to collect information about people applying for credit, which they then use in deciding whether to make loans. This process has long been partially automated. For example, American Express UK used a statistical decision process based on discriminated analysis to reject applicants falling below a certain threshold and to accept those exceeding another. The remaining 10 to 15 percent of the applicants fell into a borderline region and were referred to higher authorities giving loan for a decision. However, records showed that these authorities were no more than 50% accurate in predicting whether these borderline applicants would default on their loans. These observations motivated American Express UK to try methods from machine learning to improve the decision process. Starting with 1014 training cases and 18 descriptive attributes (such as age and years with an employer), Michie and his colleagues used an induction method to produce a decision tree, containing around 20 nodes and ten of the original features, that made correct predictions on 70% of the borderline applicants. In addition to achieving improved accuracy, the company found the rules attractive because they could be used to explain the reasons for decisions to applicants. American Express UK was so impressed that they put the resulting knowledge base into use without further development.

6. GENETIC ALGORITHMS AND GENETIC PROGRAMMING

The genetic approach to machine learning is a relatively new concept. Both genetic algorithms and Genetic Programming (GP) are a form of evolutionary computing which is a collective name for problem solving techniques based on the principles of biological evolution like natural selection. Genetic algorithms use a vocabulary borrowed from natural genetics in that they talk about *genes* (or bits), chromosomes (individuals or bit strings), and population (of individuals) [10]. Genetic algorithm approach is centered around three main processes- crossovers, mutation and selection of individuals. Initially many individual solutions are gathered together to make a randomly generated population. Genetic algorithms are based upon the Darwin theory of " The survival of the Fittest" depending upon the fitness function the best possible solutions are selected from the pool of individuals. The fitter individuals have greater chances of its selection and higher the probability that its genetic information will be passed over to future generations. Once selection is over new individuals have to be formed. These new individuals are formed either through crossover or mutation. In the process of crossover, combining the genetic make up of two solution candidates (producing a child out of two parents) creates new individuals. Whereas in mutation, we alter some individuals, which means that some randomly chosen parts of genetic information is changed to obtain a new individual. The process of generation doesn't stop until one of the conditions like minimum criteria is met or the desired fitness level is attained or a specified number of generations are reached or any combination of the above [21].

John Koza popularized GP, an offset of Genetic Algorithm in 1992. It aims at optimizing computer programs rather than function parameters.

GP is a supervised machine learning technique where algorithms are modeled after natural selection. These algorithms are represented as function trees where these trees are intended to perform a given task [6]. In GP the fitter individuals are retained and allowed to develop whereas others are discarded [4].

GP works in a manner similar to genetic algorithm. It also follows the principles of natural evolution to generate a solution that maximizes (or minimizes) some fitness function [3]. GP differs from GA in the sense that GP tends to find the solution of a given problem by representing it as a array of integers while the goal of a GP process is to produce a computer program to solve the optimization problem at hand. GP cycle works as any evolutionary process. New individuals are created; tested and fitter ones succeed in creating their own children. The unfit individuals are removed from the population. The figure:6 illustrates how GP cycle works.

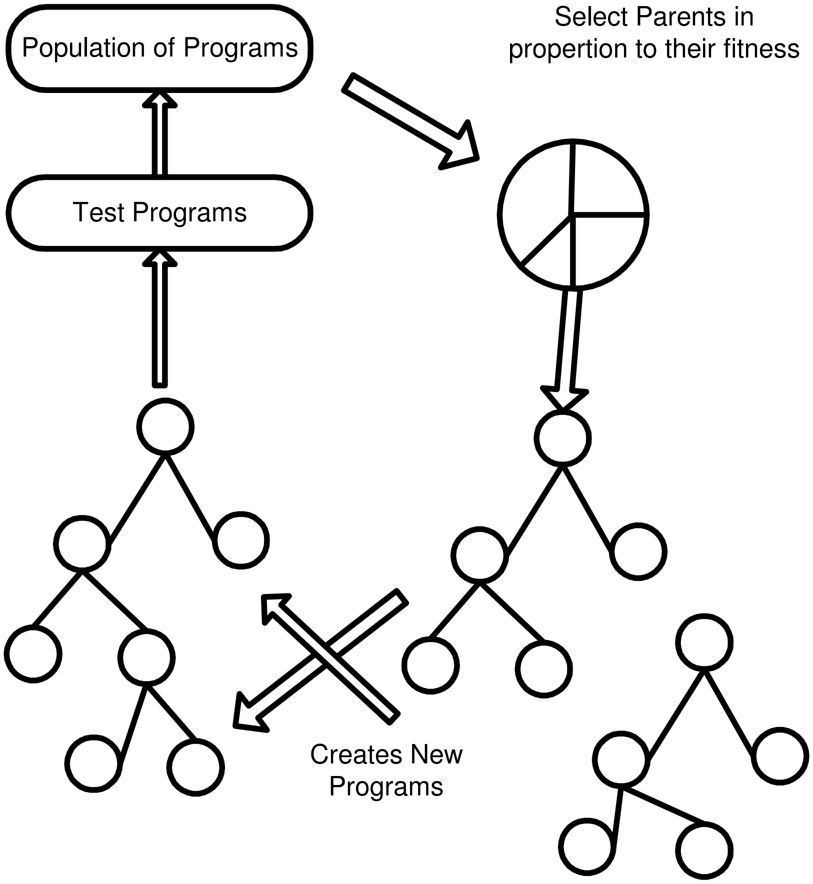


Figure 6: Genetic Programming Cycle

7. Discussion on Various Machine Learning Techniques

Technique	Application Areas	Potential Benefits	Limitations
Neural Networks (NN)	Testing Effort Estimation Function Point Analysis Risk Management Reliability Metrics Sales Forecasting	Adaptive learning: An ability to learn how to do tasks based on the data given for training or initial experience. Self-Organization: An ANN can create its own organization or representation of the information it receives during learning time. Real Time Operation: ANN computations may be carried out in parallel, and special hardware devices are being designed and manufactured which take advantage of this capability. Fault Tolerance via Redundant Information Coding: Partial destruction of a network leads to the corresponding degradation of performance. However, some network capabilities may be retained even with major network damage.	Minimizing over fitting requires a great deal of computational effort. The individual relations between the input variables and the output variables are not developed by engineering judgment so that the model tends to be a black box or input/output table without analytical basis. The sample size has to be large.
Case Based Reasoning (CBR)	Help-Desk Systems Software Effort Estimation Classification and Prediction Knowledge Based Decision systems.	No Expert is Required The CBR Process is more akin to human thinking. CBR can handle failed cases (i.e. those cases for which accurate prediction cannot be made) No extensive maintenance is required.	Case data can be hard to gather. Predictions are limited to the cases that have been observed.
Classification and Regression Trees (CART)	Financial applications like Customer Relationship Management (CRM)	It is inherently non-parametric in other words no assumptions are made regarding the underlying distribution of values of the	Relatively new and somewhat unknown. Since CART is a new technique it is

	Effort Prediction (used in models like COCOMO)	predictor variables. CART identifies splitting variables based on an exhaustive search of all possibilities. It has methods for dealing with missing variables. It is a relatively automatic machine learning technique. CART trees are easy to interpret even for non-statisticians.	difficult to find statisticians with significant expertise in this technique. CART may have unstable decision trees. CART splits only by one variable.
Rule Induction	Making Credit Decisions (in various loan companies) Diagnosis of Mechanical Devices Classification of Celestial Objects Preventing breakdowns in transformers	Simplicity of input variables. The representation in rule-based technique is easier to depict and understand.	No sufficient background knowledge is available. It is deduced from examples. Hard to maintain a complex rule-base.
Genetic Algorithms (GA) and Genetic Programming (GP)	Optimization Simulation of economic processes Scientific research purposes (Biological Evolution) Computer Games	GA and GP techniques can be applied to a variety of problems. GP is based on the 'Survival of the Fittest Scheme' allowing fitter individuals to develop and discarding unfit ones. GA is easy to grasp and can be easily applied without much difficulty	Resource requirements are large. It can be a time consuming process. GA practitioners often run many copies of the same code with the same inputs to get statistically reliable results.

8. Conclusions and Future Directions

The main contribution of this review is to discuss the various Machine-Learning Techniques employed in effort estimation, cost estimation, size estimation and other field of Software Engineering. The paper also gives a relative comparison of all the techniques based on their applications, advantages and limitations. After analysis of all the techniques, we cannot state as any one technique being the best. Each technique has different application areas and is useful in different domains based on its advantages. Thus, keeping in mind the limitations of each of the techniques and also the prime focus being the improvement in performance and efficiency we should use that technique, which best suits a particular application. For instance GA and GP prove to be useful in the area of scientific research involving biological evolution whereas rule based techniques and CART analysis may be useful in many financial applications. Similarly CBR is being developed for use in Help- Desk Systems, a relatively new application and NN may be employed for Risk Management or Sales Forecasting.

Our study also encourages that no one technique can be classified as being the perfect machine learning technique. For this reason there is a strong need for better insight into the validity and generality of many of the discussed techniques. In particular we plan to continue with research on: -
When to use machine-learning techniques and estimation models.
How to select and combine a set of test cases for effective estimation technique & to get better results?

9. REFERENCES:

[1] Aggarwal K.K., Yogesh Singh, A.Kaur, O.P.Sangwan "A Neural Net Based Approach to Test Oracle" ACM SIGSOFT Vol. 29 No. 4, May 2004.
[2]. Agnar Aamodt, Enric Plaza. "Foundational Issues, Methodological Variations, System approaches." AICom -Artificial Intelligence Communications, IOS Press Vol. 7: 1, pp. 39-59.
[3] Al Globus. "Towards 100,000 CPU Cycle-Scavenging by Genetic Algorithms." CSC at NASA Ames Research Center, September 2001.
[4] Chris Bozzuto. "Machine Learning: Genetic Programming." February 2002.
[5] Dr. Bonnie Morris, West Virginia University "Case Based Reasoning" AI/ES Update vol. 5 no. 1 Fall 1995.
[6] Eleazar Eskin and Eric Siegel. "Genetic Programming Applied to Othello: Introducing Students to Machine Learning Research" available at http://www.cs.columbia.edu/~evs/papers/sigcse-paper.ps.
[7] Gavin R. Finnie and Gerhard E. Wittig, “AI Tools for Software Development Effort Estimation”, IEEE Transaction on Software Engineering, 1996.
[8] Haykin S., “Neural Networks, A Comprehensive Foundation,” Prentice Hall India, 2003.
[9]. Howden William E. and Eichhorst Peter. Proving properties of programs from program traces. In Tutorial: Software Testing and Validation Techniques: E Miller and W.E.howden(eds.0. new York:IEEE Computer Society Press, 1978.
[10] Hsinchun Chen. "Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms" available at http://ai.bpa.arizona.edu/papers/mlir93/mlir93.html#318.
[11] Ian Watson & Farhi Marir. "Case-Based Reasoning: A Review " available at http://www.ai-cbr.org/classroom/cbr-review.html.
[12] Juha Hakkaarainen, Petteri Laamanen, and Raimo Rask, “ Neural Network in Specification Level Software Size Estimation”, IEEE Transaction on Software Engineering, 1993.
[13] Krishnamoorthy Srinivasan and Douglas Fisher, “Machine Learning Approaches to Estimating Software Development Effort”, IEEE Transaction on Software Engineering, 1995.
[14] Kohonen T., “Self Organizing Maps”, 2nd Edition, Berlin: Springer- Verlag, 1997.
[15]. Mayrhauser A. von, Anderson C. and Mraz R., “Using A Neural Network to Predict Test Case Effectiveness”' – Procs IEEE Aerospace Applications Conference, Snowmass, CO, Feb.1995.

[16] Martin Atzmueller, Joachim Baumeister, Frank Puppe, Wenqi Shi, and John A. Barnden " Case Based Approaches for handling multiple disorders" Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society, 2004.
[17] Nahid Amani, Mahmood Fathi and Mahdi Rehghan. "A Case-Based Reasoning Method for Alarm Filtering and Correlation in Telecommunication Networks" available at http://ieeexplore.ieee.org/iel5/10384/33117/01557421.pdf?arnumber=1557421.
[18] Pat Langley, Stanford and Herbert A. Simon, Pittsburgh. "Application of Machine Learning and Rule Induction." available at http://cll.stanford.edu/~langley/papers/app.cacm.ps.
[19]Peter Flach and Nada Lavrac. "Rule Induction" available at www.cs.bris.ac.uk/Teaching/Resources/COMSM0301/materials/RuleInductionSection.pdf.
[20] Roger J. Lewis. "An Introduction to Classification and Regression Tree (CART) Analysis" Presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California.
[21] Stephen M Winkler, Michael Aenzeller and Stefan Wagner. "Advances in Applying Genetic Programming to Machine Learning, Focusing on Classification Problems" available at http://www.heuristiclab.com/publications/papers/winkler06c.ps.
[22] Susanne Hoche. "Active Relational Rule Learning in a Constrained Confidence-Rated Boosting Framework" PhD Thesis, Rheinische Friedrich-Wilhelms-Universitaet Bonn, Germany, December 2004.
[23] Watson, I. & Gardingen, D. " A Distributed Case-Based Reasoning Application for Engineering Sales Support". In, Proc. 16th Int. Joint Conf. on Artificial Intelligence (IJCAI-99), Vol. 1: pp. 600-605, 1999.
[24] Yisehac Yohannes, John Hoddinott " Classification and Regression Trees- An Introduction" International Food Policy Research Institute, 1999.
[25] Yisehac Yohannes, Patrick Webb " Classification and Regression Trees" International Food Policy Research Institute, 1999.

Texting satisfaction: does age and gender make a difference?

Vimala Balakrishnan vimala.balakrishnan@mmu.edu.my
Faculty of Information, Science and Technology
Multimedia University
Melaka, 75450, Malaysia.

Paul H. P. Yeow hpyeow@mmu.edu.my
Faculty of Business and Law
Multimedia University
Melaka, 75450, Malaysia.

Abstract

This study investigated the effect of age and gender on mobile phone users' texting satisfaction, focusing on text entry factors. Structured questionnaire interviews were used to interview 18 subjects of both genders, aged between 17–37 years. Analysis of variance was computed. Gender effect was found on the speed of text entry method ($p = 0.027$), with the females being more satisfied than males. Age has a significant effect on navigation ($p = 0.026$) and learnability ($p < 0.001$), with the younger groups being more satisfied than the older groups. A significant interaction was found between age and gender for learnability ($p = 0.039$), with a clear difference between genders in the twenties and thirties. Results suggest that age and gender affect users' texting satisfaction, with varying text entry factors. Results obtained can be used to improve text entry methods by (i) catering to local dialects (ii) reducing key overloading or introducing a better method for text entry. It is concluded that texting satisfaction awaits improvements to the text entry methods.

Keywords: Age; Gender; Texting satisfaction; Text entry factors; Structured questionnaire interview

1 INTRODUCTION

Texting on mobile phones refers to the activity of composing short character based messages and exchanging it between mobile phone subscribers. Text messages or popularly known as SMS (Short Message Service) is an offshoot of the mobile phone which has evolved to service a number of unanticipated different uses (Lewis 2005). Teenagers originally started the textual use of the mobile as a form of cheap and accessible social communication. Today texting is a world wide phenomenon that has now well and truly spread beyond teenagers. Mobile Data Association (MDA) reported that over 3 billion text messages were sent in the UK during January 2006. This represents a 25% increase over the same period in 2005, where over 100 million SMS are now being sent per day (Upside Wireless Text Messaging 2006).

One of the major benefits of texting on a mobile phone is that it is cheaper than a mobile call, which appeals to people with limited incomes and to those who do not have access to the Internet for electronic communication. Text messages are being sent for various reasons all over the world. Some interesting ones are: a Malaysian man who divorced his wife by text (Kent 2003), a

Singaporean man being fired by text (Soh 2001) and an African shepherd alerting another shepherd regarding other high quality grazing areas (Gwin 2005).
Due to its overwhelming popularity, SMS and mobile phones have always been a subject of study among many researchers. One area that has been focus of many studies is the text entry method. Mobile phones were initially designed for the sole purpose of making and receiving calls. Today, it is overloaded with many utilities such as, games, browsing and chatting. Text input and the mobile phones are not a very good pairing. The standard keypad on mobile phones consists between 12–15 keys that can be used for text entry (capitalization, symbols, numbers and punctuations) with two to four keys for navigation purpose (arrow keys). This design obviously creates problem when a text needs to be entered in English, a language that involves 26 letters, punctuations and an intricate grammatical structure.

Typically text is entered on a mobile phone in one of two ways: *multitap* or predictive text entry. As the name implies, in a *multitap* system the user needs to make multiple key presses to make a letter selection. For example, the key '2' is loaded with the letters 'A', 'B' and 'C', thus if a user wants to enter a 'C', then he or she has to press the key three times (2–2–2) as 'C' is the third letter placed on the key. Things become more complicated when the intended letters are placed on the same key. For example, to text '*cab'* the key presses will be 222–2–22. Segmentation takes place to determine the correct letter. Most of the mobile phones employ a time-out process, in which the user is required to wait for a specified time (typically one – two seconds) before attempting to enter the next letter; hence *multitap* is often criticized for being slow.

On the other hand, the predictive text entry method uses linguistic knowledge to predict the intended words of the user. Most mobile phones have licensed the T9 input method which uses a dictionary as the basis for disambiguation. Each key is pressed only once. For example, the user enters 8–4–3–0 to key in '*the*' whereby the 0-key delimits words and terminates disambiguation of the preceding keys (Silfverberg *et al.* 2000). However, this process of disambiguation must sometimes automatically choose between more than one words produced by the same set of key presses, or present the user with a series of word choices. For example, the combination of 3–6–4 ('def'–'mno'–'ghi') might mean dog or fog. If the algorithm suggests a wrong word, the user has to manually cycle through the possible options by pressing a *next* key. Predictive text entry was found to expedite messaging, only when one really knows how to use it. An experiment using a mobile phone (James and Reischel 2001) found that experts and novices reached about 8 words per minute (wpm) with *multitap*. In comparison, predictive text entry was used by novices at 9.1 wpm and experts at 20.4 wpm. However, all these results were based on English-based text only.

User input is a crucial issue concerning mobile devices, thus a lot of studies have been conducted on the efficiency of text entry methods and ways to improve them. Some researchers have done comparison studies based on the text entry methods (Silfverberg *et al*. 2000, Buchanan *et al*. 2001, James and Reischel 2001, Cockburn and Siresena 2003, Wigdor and Balakrishnan 2004) whereas others have tried to introduce new techniques to enter text via mobile phone's limited interface (Mackenzie 2002, Wigdor and Balakrishnan 2003, Gong and Tarasewich 2005). Mackenzie *et al*. (1999) explored the text entry rates for several variations of soft keyboards. Studies have also been conducted to enhance the existing predictive text entry methods (Maragoudakis *et al.* 2002). Though numerous studies have been conducted related to text entry methods, but to the best of our knowledge, none focused on users' texting satisfaction based on age and gender.

People of different age communicate differently as they grew up with different genres of technologies. A number of studies of technology usage and age have found that usage diminishes with age, and sometimes this is linked with attitudes towards specific technologies like the Internet (Pew-Internet 2001, US Department of Commerce 2002). The Department of Commerce study found that computer and Internet use is highest among children and teenagers, while people over 50 years old are less likely to use computers. The e-Living project (Ling *et al.* 2002) investigated technology uptake across age and found that Internet usage, mobile ownership and household Internet connections rapidly decreased for users over 50 years old, whilst being high for younger adults. A study examining the text entry on handheld computers by older workers found that younger people were faster but less accurate than older people at using a touch-screen keyboard (Wright *et al.* 2000). Age factor was studied in many other studies as well, e.g. Bunce and Sisa (2002) investigated age differences in the perceived workload associated with the performance of a demanding, high event rate, vigilance task. Lightner (2003) studied the impact of age, income and education in e-commerce designs and found that sensory impact of e-commerce sites became less important as respondents increase in age.

Gender has also been used as an important moderating variable in many studies. These include the influence of gender on grip strength and endurance (Nicolay and Walker 2005), the effects of chronic low back pain, age and gender on vertical spinal creep (Kanlayanaphotporn *et al.* 2003), psychology of SMS messaging between male and female users (Reid and Reid 2004), maturation and gender identity of the adoption of mobile phones by teenagers in Norway (Ling 2001) and the moderating effect of gender in explaining the intention to use mobile chat services (Nysveen *et al.* 2005). From a marketing perspective, findings suggest that females and males process advertisements differently and respond differently to marketing communication efforts (Wolin 2003). Gender differences were also noted in studies of technology use. For example, Gefen and Straub (1997) reveal that women and men differ in their perception of e-mail, while Venkatesh and Morris (2000) find gender differences in the motives for using a new software system at a workplace.

Identifying age and gender effect is important to improve technology product designs, so as to increase users' subjective satisfaction. Thus, the aim of this study is to investigate if gender and age influence mobile phone users' texting satisfaction with respect to the text entry factors.

2 RESEARCH FRAMEWORK

In this study, texting satisfaction was identified as the dependent variable and text entry factors such as speed, learnability, simplicity and navigation are the independent variables. All these factors were identified from literature reviews (Silfverberg *et al.* 2000, Buchanan *et al.* 2001, James and Reischel 2001, Wigdor and Balakrishnan, 2004, Soriano *et al.* 2005). Age and gender are the moderating variables. Table 1 shows the description for each of the text entry factors and Figure 1 shows the research framework used in this study.

Table 1- Text entry factors

Text entry factors	Explanation
Speed/Efficiency	The speed in which a text can be keyed in either by using *multitap* or predictive text entry system
Learnability	The ease in which users can learn the text entry mechanism
Simplicity	The simplicity of using the text entry mechanism
Navigation	The ease in which key selections can be made while texting (capitalization, punctuation, blank space etc.)

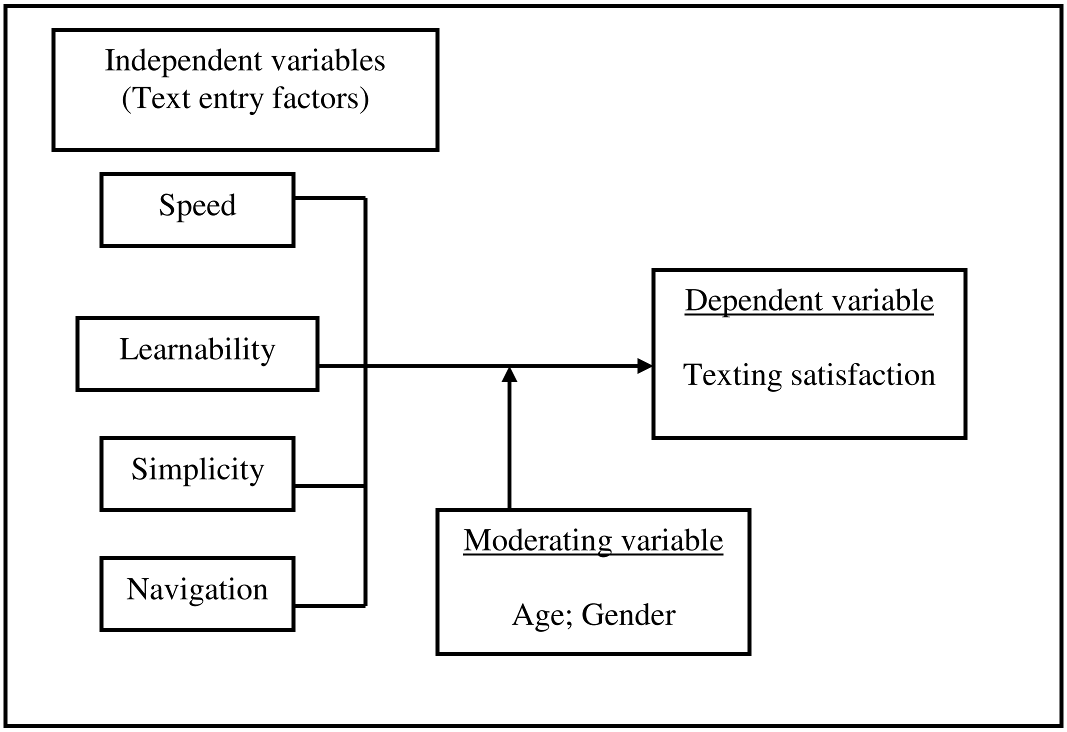


Figure 1-Research framework.

3 METHODS

3.1 Subjects

A total of 18 subjects were interviewed, consisting of nine males and nine females, aged between 17 – 37 years old (mean = 24.3 years, SD = 5.8). All subjects were recruited from Multimedia University (Malaysia), comprising of students and staff. Only a small number of interviews were conducted to obtain in-depth qualitative data. By way of comparison, in a study of teens' text messaging behaviour by Grinter and Eldridge (2001), five males and five females were interviewed. Similarly, in another study of people's monitor usage, only 18 participants were interviewed (Grudin 2001). The subjects were then grouped into three age categories, i.e. teens, twenties and thirties. This resulted in six subjects in each group with three males and three females. All the subjects have used SMS before, with an average of 3.7 years of experience and SD = 1.18. Almost 61.1% (11) of the subjects use *multitap* technique for text entry, 22.2% (4) use predictive text entry and only three teens use predictive text entry and *multitap* technique interchangeably. All 18 subjects use Nokia mobile phones of different models but with a similar keypad layout (Nokia 8250, 3120 and 6610), except for Nokia Communicator 9210™ that was used by one male subject. This particular model displays a smaller version of the QWERTY style keypads that allow for text entry with techniques similar to typing on a regular keyboard.

Table 2-Summary statistics for the majority number of users for each gender/age categories

Gender/Age	Time (%: N)	Sent (%: N)	Frequency of Abbreviations (%: N)	Frequency of Slang (%e: N)
Male/teens	1–3 (100:3)	3–5 (66.7:2)	Always (100:3)	Always (66.7:2)
Male/20s	3–5 (66.7:2)	3–5 (66.7:2)	Sometimes (66.7:2)	Sometimes (66.7:2)
Male/30s	<1 (100:3)	1–3 (66.7:2)	Sometimes (66.7:2)	Never (66.7:2)
Female/teens	5–7 (66.7:2)	3–5 (66.7:2)	Always (66.7:2)	Always (66.7:2)
Female/20s	3–5 (100:3)	3–5 (100:3)	Always (66.7:2)	Sometimes (66.7:2)
Female/30s	1–3 (66.7:2)	1–3 (66.7:2)	Sometimes (66.7:2)	Never (66.7:2)

Time: time spent to SMS in a day (minutes); Sent: number of SMS sent in a day; N: number of subjects; Total number of subjects for each category is 3

Table 2 shows some of the summary statistics for the average time spent in messaging in a day, average number of SMS sent in a day, frequency of using abbreviations and slang (local dialect such as *'eh'* and *'lah'*) based on gender and age.

Females spend more time texting in a day than males, hence sending more messages in a day. Moreover, texting activities also decline as the participants' age increases. Females use more abbreviations than males. It can also be noted that the younger the subject, the higher the frequency of using abbreviations and slang in messages.

3.2 Materials

An interview questionnaire was designed based on Sinclair's (1995) guidelines. The questionnaire was developed in English and had two major sections: Section A to obtain the demographic profile of the subjects (gender, age and experience in using SMS) whereas Section B is for the subjects to rate their satisfaction/dissatisfaction levels to statements using Likert's five-point scale, whereby 1 means 'Strongly dissatisfied', 2 means 'Dissatisfied', 3 means 'Neutral', 4 means 'Satisfied' and 5 means 'Strongly Satisfied'.

3.3 Interviews

Face-to-face interviews were conducted with each of the subject. Each interview session lasted for about 15–20 minutes. Subjects were encouraged to discuss problems, voice out opinions, suggestions and recommendations. All verbal comments were recorded. All interviews took place in Multimedia University involving undergraduate students and working professionals. Only one interviewer was involved in this exercise and all 18 interviews were completed within two days.

4 RESULTS AND DISCUSSION

The data collected were analysed using Statistical Package for the Social Sciences (SPSS) software. Analysis of variance (ANOVA) and Tukey Post-Hoc analysis were used to analyse the significant differences (if any) between gender and age groups, with respect to text entry factors effect on texting satisfaction. All results are considered significant at $p < 0.05$ level.

Table 3-ANOVA test for text entry factors satisfaction, based on gender

Text entry factors	F	*p*
Speed or Efficiency	5.95	0.027*
Learnability	0.05	0.821
Simplicity	0.05	0.819
Navigation	0.06	0.810

F: F statistic; *p*: *p*-value; *: significant at $p < 0.05$

Table 3 shows that there is a significant effect of gender with respect to users' satisfaction towards speed or efficiency of text entry mechanism. The females were found to be more satisfied (mean = 4.2) than males (mean = 3.3). About 44.4% (4/9) of the males reported that texting using the current text entry methods can be tedious and time consuming, especially when they are on the move or when they couldn't pay full attention to the screen and keypads while texting (crowded place, looking elsewhere etc.). Both the text entry mechanisms require a significant amount of visual searching to find a needed letter or word. This results in them not to adopt using SMS at times as making a call would be much faster and less cumbersome. Slow text entry mechanism which includes the multiple key presses involved in accessing characters has also been cited as one of the usability issues of mobile phones by Axup *et al.* (2005). Moreover, two males who use *multitap* and predictive text entry system interchangeably agreed that texting using the latter system is faster. However, having to cycle through to select the correct word can be frustrating, especially when texting needs to be done in a hurry. When this situation arises, *multitap* would be the preferred method. Interestingly, females being more satisfied than males could also be contributed to the fact that females generally have smaller fingers, thus they are able to make multiple key presses on the keypads with lesser error and faster. Another possible reason could be that females spend more time in texting and send more messages in a day than males (see table 2). This statistic is also consistent with some other studies (Ling 2003, Reid and Reid 2003, Faulkner and Culwin 2005). Heavy texting among the females might have made them an expert in keying in the text.

Table 4-ANOVA test for text entry factors satisfaction, based on age

Text entry factors	F	*p*
Speed or Efficiency	2.50	0.116
Learnability	28.81	0.000*
Simplicity	2.67	0.102
Navigation	4.67	0.026*

F: F statistic; *p*: *p*-value; *: significant at $p < 0.05$

In table 4, age was found to have significant effect with users' satisfaction towards learnability and navigation. Tukey Post-Hoc analysis revealed that teenagers are more satisfied with learnability ($p < 0.001$) and navigation ($p = 0.003$) compared to users in their thirties. Moreover, users in their twenties are also more satisfied with the navigation than those in their thirties ($p = 0.023$). Younger users have the capability and the interest in learning new techniques fast compared to their older counterparts. Moreover, they are keener in adapting to new environment or changes as well. This could be the reason for their satisfaction towards learnability compared to the older users. One participant in his late thirties commented that "*I remember that I started* *using SMS only seven months after purchasing my mobile phone, I found it to be complicated and cumbersome…and my SMS coach was my 12-year old nephew!!*". This comment hints at a fact that it is the younger generation that has taken the lead in SMS messaging. This notion is also supported by findings from other studies (Ling 2003, Reid and Reid 2003, Faulkner and Culwin 2005). The male participant who uses Nokia Communicator™ commented "*I love messaging using my Communicator as it is the fastest way to send romantic text to my girl-friend*". However, upon further discussion, he also mentioned that he only uses the QWERTY keyboard instead of the standard 12-key keypad to text as it is much faster and easier. Learnability is not an issue here but unfortunately the majority of mobile phone users carry standard 12-key mobiles.

Apart from facing the difficulty in learning text entry methods, mobile phone users need to know the language and syntax of SMS communication. Using text relies on the creative use of language in order to maximize the 160 characters that are allowed per message. Once the maximum characters are reached, then the sender is charged for a second message. This is the main reason why text message users truncate and alter conventionally written language. For example, it is very common to type '*hru?'* instead of '*How are you?*' or '*c u l8r'* instead of '*See you later*'. One must be able to decipher the meaning of all the abbreviations used to understand the content of a message. Younger users who text frequently feels very comfortable in using these SMS language, however, this might not be the case for the older users. One male in his early twenties responded that he once felt very embarrassed with himself as he couldn't understand what '*tc*' means; only to find out from a friend later that it simply means '*take care*'. Thus, having to master the technique of text entry and also being able to communicate successfully using SMS language affects learnability of mobile phone users, especially among the older ones.

The younger users were also found to be more satisfied with the navigation while texting than users in their thirties. Navigation problem occurs due to the nature of the keypads that are overloaded to accommodate all the available letters, numbers and punctuations. It is obvious to know that the letters 'A', 'B' and 'C' are placed on the 2-key, however, symbols and punctuations are not presented clearly. Users have to 'look' for these characters by making multiple key presses and not knowing which key to press complicates the procedure. Knowing which key to press might not be a problem among the younger users due to their active messaging, but it will slow down texting process among the older mobile phone users. Moreover, the pattern of messages written further complicates navigation problems among the older generations. Older subjects tend to pay more attention to capitalization, spacing and punctuations than teenagers. This results in them having to access the key that is mapped to these characters frequently. Mapping of the appropriate navigational keys to the desired object was also found to be cumbersome by a study conducted among middle-aged users (Soriano *et al.* 2005), e.g. locating the key to access the 'ABC' (non-predictive) menu that allows a user to change from different character input types, i.e. from alphabetical to numerical is not a straight-forward and clear process. However, this finding is solely based on Samsung T400 model.

Table 5-ANOVA test for text entry factors satisfaction, based on *age x gender*

Text entry factors	F	*p*
Speed or Efficiency	0.08	0.921
Learnability	4.30	0.039*
Simplicity	0.41	0.671
Navigation	1.59	0.11

F: F statistic; *p*: *p*-value; *: significant at $p < 0.05$

Table 5 shows that the interaction effect of *age x gender* was found to be significant for learnability.

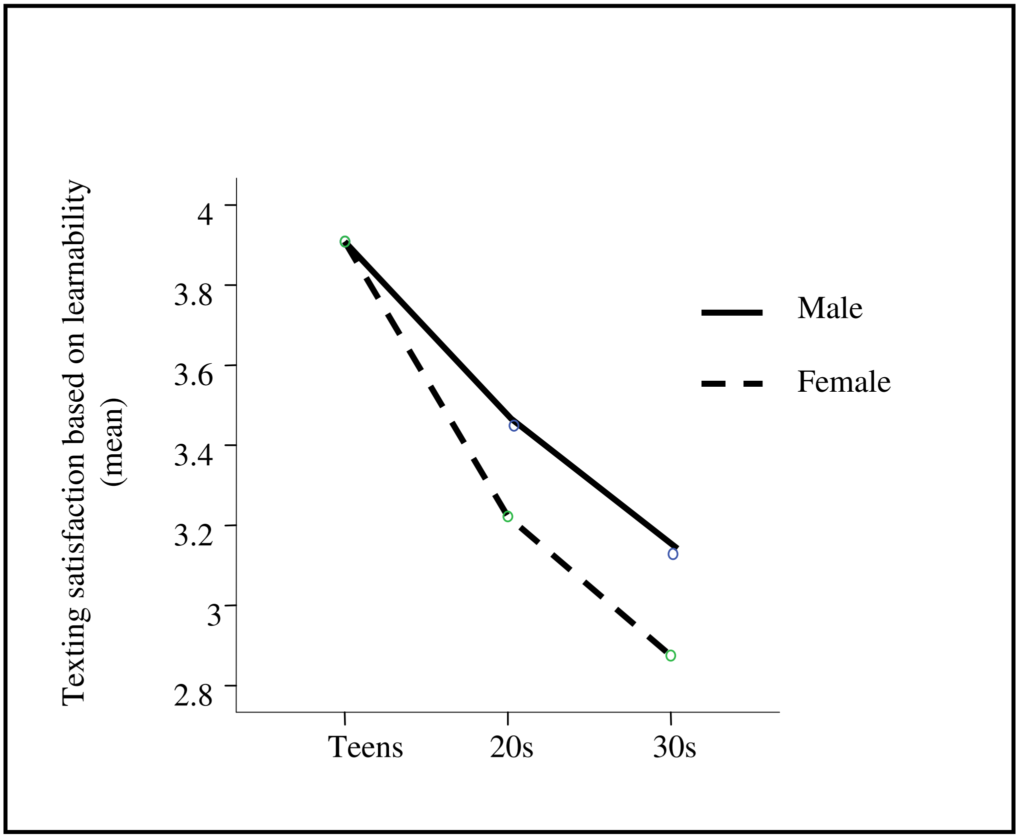


Figure 2- Interaction effect of *age x gender* on learnability

Figure 2 shows a clear difference between the genders in the twenties and thirties categories, with the males being more satisfied than females with respect to learnability. The males have the capability of learning and understanding technical things faster than the females, thus making them to be more satisfied than females. All six females in their twenties and thirties agreed that learning the art of texting was difficult as it seems to be a complex feature. Learning the method to make key presses and knowing the mapping of other characters to the keys were cited as the major obstacles. Females also have the tendency to include more literary flourishes like proper salutations, capitalization and punctuations in their messages. They are also more likely to write longer and complex messages that include emotional elements in their communications (emoticons to reflect happiness and sadness). In addition, their messages tend to retain more of the traditional conventions associated with other written forms than men (Sattle 1975, Rosenthal 1985). This requires them to make more key presses, and most importantly, they also agree that they tend to forget the correct key to press especially when they are stressed. Men on the contrary writes short text that is simple and straight to the point, often only using a single sentence or a single word (Ling 2003).

It can also be noted that the users' satisfaction declines as the age increases. The older a person gets, the more difficult it is to learn and adapt to new technological products. A similar finding was reported by Ling (2003) and Faulkner and Culwin (2005). The majority of the subjects in their thirties (4/6) stated that they rather make phone calls instead of texting as texting can be tedious and time consuming at times. Two males in this category responded that they only text when it is really necessary or unavoidable (when in meetings or conference, cinema or when pre-paid credit is low), otherwise they prefer phone calls. Moreover, subjects in their twenties and thirties are working professionals who have a steady income, thus they can afford to make phone calls unlike the teens. Subjects in twenties and thirties also rarely or never use predictive text entry as they feel that *multitap* technique is straight-forward, thus faster. Subjects also commented that no unnecessary interruptions take place while messaging using *multitap*, unlike having to press the *next* key to make word selections in predictive text entry. Learning to use predictive text entry was found to be more difficult than *multitap* system. Moreover, one has to practice using it to understand the mechanism behind it. The older users neither have the time nor the interest to learn this mechanism. The younger generations, teens in this case, have the capability of learning new things faster compared to the older generations, regardless of their gender. This explains as to why the difference is not pronounced between the males and females in their teens. Though they are more satisfied than subjects in their twenties and thirties, they also feel that predictive text entry software should be further enhanced to cater to their needs, such as, to support more SMS languages or interestingly, to support local dialects (Chinese, Malay or Tamil) and even slang frequently used among the teens like 'kewl' (cool), 'omigod' (Oh my God), '*eh*', '*lah*' (local slang) and many more. Data in this study (see table 2) also indicate that teens use abbreviations and slang more frequently compared to users in their twenties and thirties. Data obtained by Ling (2003) for Norwegian mobile users also indicate that teens are more inclined to use dialects than older users.

5 CONCLUSION

Structured questionnaire interviews' results involving 18 subjects comprising of mobile phone users from three age categories (teens, twenties and thirties) were presented. All subjects use Nokia mobile phones, thus the text entry methods and keypad layouts are similar (except for one male participant who uses a Nokia Communicator™). Focus of this study was mainly on text entry factors that could affect mobile phone users' satisfaction in texting, seen from the perspectives of gender and age variations. Females were found to be more satisfied with the speed or efficiency of text entry mechanism (*multitap* or predictive text entry) than males. Older subjects (twenties and thirties) are less satisfied with the navigation and learnability of the text entry method than teenagers. Moreover, the interaction between *age x gender* was found to be significant for learnability as well, with the differences between genders being clear for those who are in their twenties and thirties. Males were found to be more satisfied with learnability than females in the same categories. This study revealed that gender and age do influence texting satisfaction among mobile phone users. Two factors that were found to significantly affect their satisfaction are learnability and navigation. These factors were prominent among older users and they took a longer period to learn the text entry mechanism as it was found to be complicated, hence resulting in them using SMS on a rare basis only. Subjects who use predictive text entry system indicated that support for abbreviations, slang, emoticons and dialects would increase their texting satisfaction. Mobile phone designers should look into incorporating other possibilities of text entry mechanism (Mackenzie 2002, Wigdor and Balakrishnan 2004, Gong *et al.* 2005). Perhaps customized mobile phones can be designed to cater for specific needs for different targets.

6 REFERENCES

Axup, J., Viller, S. and Bidwell, N., 2005, Usability of a Mobile, Group Communication Prototype While Rendezvousing. In *CTS'05 International Symposium on Collaborative Technologies and Systems-Special Session on Mobile Collaborative Work* (St. Louis).

Buchanan, G., Jones, M., Thimbleby, H., Farrant, S. and Pazzani, M., 2001, Improving Mobile Internet Usability. In *Web 2001 Conference* (Hong Kong, ACM Press).

Bunce, D., and Sisa, L, 2002, Age differences in perceived workload across a short vigil. *Ergonomics*, **45**, pp. 949–960.

Cockburn, A. and Siresena, A., 2003, Evaluating Mobile Text Entry with the Fastap Keypad. In *British Computer Society Conference on Human Computer Interaction*, England.

Faulkner, X. and Culwin, F., 2005, When Fingers Do The Talking: A Study of Text Messaging. *Interacting with Computers*, **17**, pp. 167–185.

Gefen, D. and Straub, D.W., 1997, Gender differences in the perception of e-mail: An extension to the technology acceptance model. MIS Quarterly, December 1997, pp.389–400.

Gong, J. and Tarasewich, P., 2005, Alphabetically Constrained Keypad Designs for Text Entry on Mobile Devices. In *CHI 2005, PAPERS: Small Devices 1*, pp. 211–220.

Gong, J., Haggerty, B. and Tarasewich, P., 2005, An Enhanced Multitap with Predictive Next-Letter Highlighting. In *CHI 2005,* Oregon, USA.

Grinter, R. and Eldridge, M., 2001, y do tngrs luv 2 txt msg?. In *Seventh European Conference on Computer Supported Cooperative Work, ECSCW,* Bonn, Germany.

Grudin, J., 2001, Partitioning Digital World: focal and peripheral awareness in multiple monitor use. In *Human Factors and Computing Systems, CHI'01,* Seattle, USA.

Gwin, P., 2005, Making the Connection: The Jump From Wired to Wireless is Changing Africa, National Geographic (September 2005).

James, C.L. and Reischel, K.M., 2001, Text Input for Mobile Devices: Comparing Model Prediction to Actual Performance. *CHI 2001*, **3**, pp. 365–371.

Kanlayanaphotporn, K., Trott, P., Williams, M. and Fulton, I., 2003, Effects of chronic low back pain, age and gender ion vertical spinal creep. *Ergonomics*, **46**, pp. 561–573.

Kent,J., 2003, Malaysia Reviews Texting Divorce. Available online at:www.news.bbc.co.uk/2/hi/asia-pacific/3112151.htm (accessed August 2006).

Lewis, A., 2005, Hve I told U l8tly that I luv u?: Mobile phones: The New Domesticated Artefact. *Australian Counseling Association Journal "Counseling Australia"***, 5.**

Lightner, N.J., 2003, What users want in e-commerce design: effects of age, education and income. *Ergonomics*, **46**, pp. 153 – 168.

Ling, R., 2001, “We Release Them Little by Little”: Maturation and Gender Identity as Seen in the Use of Mobile Telephony. *Personal and Ubiquitous Computing*, **5**, pp. 123–136.

Ling, R., 2003, The Socio-linguistic of SMS: An Analysis of SMS Use by a Random Sample of Norwegians. In *Mobile Communications: Renegotiation of the Social Sphere* , Ling, R. and Pedersen, P. (Eds.), pp. 335 –349 (London: Springer, 2003).

Ling, R., Yttri, B., Anderson, B. and Diduca, D., 2002, Age, gender and social capital- a cross sectional analysis. Available at: www.eurescom.de/e-living/ (accessed August 2006).

Mackenzie, S., Zhang, S.X. and Soukoreff, R.W., 1999, Text entry using soft keyboards. *Behaviour & Information Technology*, **18,** pp. 235–244.

Mackenzie, S.I., 2002, Mobile Text Entry Using Three Keys. In *Proceedings of the Second Nordic Conference on Human-Computer Interaction- NordiCHI 2002*, pp. 27–34.

Maragoudakis, M., Tselios, N.K., Fakotakis, N. and Avouris, N.M., 2002, Improving SMS Usability Using Bayesian Networks. In Methods and Applications of Artificial Intelligence, Vlahavas, I.P. and Spyropoulos, C.D. (Eds), pp. 179–190 (Berlin: Springer-Verlag, 2002).

Nicolay, C.W. and Walker, A,L., 2005, Grip Strength and Endurance: Influence of Anthropometric Variation, Hand Dominance and Gender. *International Journal of Industrial Ergonomics*, **35**, pp. 605–618.

Nysveen, H., Pedersen, P.E. and Thorbjornsen, H., 2005, Explaining Intention to use Mobile Chat Services: Moderating Effects of Gender. *Journal of Consumer Marketing*, **22**, pp. 247–256.

Pew-Internet, 2001, Wired Seniors: A Fervent Few, Inspired by Family Ties. Available at: www.perinternet.org/reports/pdfs/PIP_Wired_Seniors_Report.pdf (accessed July 2006)

Reid, D.J. and Reid F.J.M., 2003, Text mates and text circles: insights into the social ecology of SMS text messaging. In *The Mobile Revolution: A Retrospective-Lesson on Social Shaping,* Lasen, A. (Ed.)*,* Proceedings of the 4th Wireless World Conference.

Reid, F.J.M. and Reid, D.J., 2004, Text Appeal: The Psychology of SMS Texting and Its Implications for the Design of Mobile Phone Interfaces. *Campus Wide Information Systems*, **21**, pp. 196–200.

Rosenthal, C., 1985, Kinkeeping in the Familial Division of Labor. *Journal of Marriage and the Family,* **47,** pp. 965–974.

Sattle, J.W., 1985, The Inexpressive Male: Tragedy or Sexual Politics. *Social Problems,* **23**, pp. 469–477.

Silfverberg, M., Mackenzie, S.I. and Korhonen, P., 2000, Predicting Text Entry Speed on Mobile Phones. *CHI 2000*, **2**, pp. 9–16.

Sinclair, A.M., 1995, Subjective Assessment. In *Evaluation of Human Work-A Practical Ergonomics Methodology,* Wilson, J.R. and Corlett, E.N. (Eds.), pp. 69–100 (London:Taylor & Francis, 1995).

Soh, N., 2001, First Case of Sacking Via SMS. Available at: www.it.asia1.com.sg/newsarchive/07/news004_20010724.html (accessed August 2006).

Soriano, C., Raikundalia, G.K. and Szajman, J., 2005, A Usability Study of Short Message Service on Middle-Aged Users. In *Proceedings of OZCHI 2005*, Canberra, Australia.

Upside Wireless Text Messaging, 2006, 2006: Year of the Text Message?. Available at: www.upsidewireless.com/blog/2006/2006-year-of-the-text-message/ (accessed December 2006).

US Department of Commerce, 2002, A Nation Online: How Americans Are Expanding Their Use of the Internet. National Telecommunications and Information Administration. Availbale at: www.ntia.doc.gov/ntiahome/dn/html/Chapter2.htm (accessed June 2006).

Venkatesh, V. and Morris, M.G., 2000, “Why don’t men ever stop to ask for directions: gender, social influence, and their role in technology acceptance and usage behavior”. MIS Quarterly, **24**, pp. 115–39.

Wigdor, D. and Balakrishnan, R., 2003, TiltText: Using Tilt for Text Input to Mobile Phones. *CHI Letters,* **5**, pp. 81–89.

Wigdor, D. and Balakrishnan, R., 2004, A Comparison of Consecutive and Concurrent Input Text Entry Techniques for Mobile Phones. *CHI 2004*, **6,** pp. 81–88.

Wolin, L.D., 2003, Gender Issues in Advertising – An Oversight Synthesis of Research: 1970 – 2002. *Journal of Advertising Research*, pp. 111–129.

Wright, P., Bartram, C., Rogers, N., Emslie, H., Evans, J., Wilson, B. and Belt, S., 2000, Text entry on handheld computers by older users, *Ergonomics*, **43,** pp. 702–716.