Online ISSN: 0975-4172
Print ISSN: 0975-4350

Global Journal of Computer Science and Technology

DISCOVERING THOUGHTS AND INVENTING FUTURE

9 Technology Reforming Ideas

October 2011

Pinnacles

- Disk Scheduling Algorithms
- Wireless Sensor Network
- Software Effort Estimation
- Gaussian Mixture Model

The Volume 11
Issue 18
Version 1.0

ENG

Global Journal of Computer Science and Technology

Global Journal of Computer Science and Technology

Volume 11 Issue 18 (Ver. 1.0)

Open Association of Research Society

Incorporation No.: 0423089
License No.: 42125/022010/1186
Registration No.: 430374
Import-Export Code: 1109007027
Employer Identification Number (EIN):
USA Tax ID: 98-0673427

Global Journals Inc.

(A Delaware USA Incorporation with "Good Standing"; Reg. Number: 0423089)

Sponsors: Open Association of Research Society

Open Scientific Standards

Publisher's Headquarters office

Global Journals Inc., Headquarters Corporate Office,
Cambridge Office Center, II Canal Park, Floor No. 5th, ***Cambridge (Massachusetts)***, Pin: MA 02141
United States
USA Toll Free: +001-888-839-7392
USA Toll Free Fax: +001-888-839-7392

Offset Typesetting

Open Association of Research Society, Marsh Road, Rainham, Essex, London RM13 8EU
United Kingdom.

Packaging & Continental Dispatching

Global Journals, India

Find a correspondence nodal officer near you

To find nodal officer of your country, please email us at *local@globaljournals.org*

eContacts

Press Inquiries: *press@globaljournals.org*
Investor Inquiries: *investers@globaljournals.org*
Technical Support: *technology@globaljournals.org*
Media & Releases: *media@globaljournals.org*

Pricing (Including by Air Parcel Charges):

For Authors:
22 USD (B/W) & 50 USD (Color)
Yearly Subscription (Personal & Institutional):
200 USD (B/W) & 250 USD (Color)

Editorial Board Members (HON.)

Dr. Bart Lambrecht
Director of Research in Accounting and FinanceProfessor of Finance
Lancaster University Management School
BA (Antwerp); MPhil, MA, PhD (Cambridge)

Dr. Carlos García Pont
Associate Professor of Marketing
IESE Business School, University of Navarra
Doctor of Philosophy (Management), Massachusetts Institute of Technology (MIT)
Master in Business Administration, IESE, University of Navarra
Degree in Industrial Engineering, Universitat Politècnica de Catalunya

Dr. Fotini Labropulu
Mathematics - Luther College
University of ReginaPh.D., M.Sc. in Mathematics
B.A. (Honors) in Mathematics
University of Windso

Dr. Lynn Lim
Reader in Business and Marketing
Roehampton University, London
BCom, PGDip, MBA (Distinction), PhD, FHEA

Dr. Mihaly Mezei
ASSOCIATE PROFESSOR
Department of Structural and Chemical Biology, Mount Sinai School of Medical Center
Ph.D., Etvs Lornd University
Postdoctoral Training,
New York University

Dr. Söhnke M. Bartram
Department of Accounting and FinanceLancaster University Management SchoolPh.D. (WHU Koblenz)
MBA/BBA (University of Saarbrücken)

Dr. Miguel Angel Ariño
Professor of Decision Sciences
IESE Business School
Barcelona, Spain (Universidad de Navarra)
CEIBS (China Europe International Business School).
Beijing, Shanghai and Shenzhen
Ph.D. in Mathematics
University of Barcelona
BA in Mathematics (Licenciatura)
University of Barcelona

Philip G. Moscoso
Technology and Operations Management
IESE Business School, University of Navarra
Ph.D in Industrial Engineering and Management, ETH Zurich
M.Sc. in Chemical Engineering, ETH Zurich

Dr. Sanjay Dixit, M.D.
Director, EP Laboratories, Philadelphia VA Medical Center
Cardiovascular Medicine - Cardiac Arrhythmia
Univ of Penn School of Medicine

Dr. Han-Xiang Deng
MD., Ph.D
Associate Professor and Research
Department Division of Neuromuscular Medicine
Davee Department of Neurology and Clinical NeuroscienceNorthwestern University
Feinberg School of Medicine

Dr. Pina C. Sanelli
Associate Professor of Public Health
Weill Cornell Medical College
Associate Attending Radiologist
NewYork-Presbyterian Hospital
MRI, MRA, CT, and CTA
Neuroradiology and Diagnostic Radiology
M.D., State University of New York at Buffalo,School of Medicine and Biomedical Sciences

Dr. Roberto Sanchez
Associate Professor
Department of Structural and Chemical Biology
Mount Sinai School of Medicine
Ph.D., The Rockefeller University

Dr. Wen-Yih Sun
Professor of Earth and Atmospheric SciencesPurdue University Director
National Center for Typhoon and Flooding Research, Taiwan
University Chair Professor
Department of Atmospheric Sciences, National Central University, Chung-Li, TaiwanUniversity Chair Professor
Institute of Environmental Engineering, National Chiao Tung University, Hsin-chu, Taiwan.Ph.D., MS The University of Chicago, Geophysical Sciences
BS National Taiwan University, Atmospheric Sciences
Associate Professor of Radiology

Dr. Michael R. Rudnick
M.D., FACP
Associate Professor of Medicine
Chief, Renal Electrolyte and Hypertension Division (PMC)
Penn Medicine, University of Pennsylvania
Presbyterian Medical Center, Philadelphia
Nephrology and Internal Medicine
Certified by the American Board of Internal Medicine

Dr. Bassey Benjamin Esu
B.Sc. Marketing; MBA Marketing; Ph.D Marketing
Lecturer, Department of Marketing, University of Calabar
Tourism Consultant, Cross River State Tourism Development Department
Co-ordinator , Sustainable Tourism Initiative, Calabar, Nigeria

Dr. Aziz M. Barbar, Ph.D.
IEEE Senior Member
Chairperson, Department of Computer Science
AUST - American University of Science & Technology
Alfred Naccash Avenue – Ashrafieh

Contents of the Volume

Global Journal of Computer Science and Technology
Volume 11 Issue 18 Version 1.0 October 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Sim_Dsc: Simulator for Optimizing the Performance of Disk Scheduling Algorithms

By P.K. Suri, Sumit Mittal
Kurukshetra University

Abstract - Disk scheduling involves a careful examination of pending requests to determine the most efficient way to service these requests. A disk scheduler examines the positional relationship among waiting requests, then reorders the queue so that the requests will be serviced with minimum seek. The purpose of the study is to obtain the best scheduling algorithm based on the seek time, rotation time and transfer time for moveable head disks. Keeping in view an attempt has been made to design a simulator for optimizing the performance of disk scheduling algorithms using Box-Muller transformation. The input for the simulator has been derived by using an algorithm for generating pseudo random numbers which follows box-muller transformations. Simulator takes access time which is generated using seek time, rotation time and transfer time, as the request of cylinder numbers, current position of read/write head as inputs. On the basis of these inputs, total head movement of each disk scheduling algorithm is calculated under various loads.

Keywords : *disk scheduling algorithms, seek time, rotational delay, transfer time, access time, head movement, box-muller transformation.*

GJCST-A Classification : *F.2.1,G.1.6*

Strictly as per the compliance and regulations of:

Sim_Dsc: Simulator for Optimizing the Performance of Disk Scheduling Algorithms

P.K. Suri [α], Sumit Mittal [Ω]

Abstract - Disk scheduling involves a careful examination of pending requests to determine the most efficient way to service these requests. A disk scheduler examines the positional relationship among waiting requests, then reorders the queue so that the requests will be serviced with minimum seek. The purpose of the study is to obtain the best scheduling algorithm based on the seek time, rotation time and transfer time for moveable head disks. Keeping in view an attempt has been made to design a simulator for optimizing the performance of disk scheduling algorithms using Box-Muller transformation. The input for the simulator has been derived by using an algorithm for generating pseudo random numbers which follows box-muller transformations. Simulator takes access time which is generated using seek time, rotation time and transfer time, as the request of cylinder numbers, current position of read/write head as inputs. On the basis of these inputs, total head movement of each disk scheduling algorithm is calculated under various loads.

Keywords : disk scheduling algorithms, seek time, rotational delay, transfer time, access time, head movement, box-muller transformation.

I. Introduction

Among major responsibilities of operating system disk scheduling is one of the important tasks to use disk efficiently. For meeting these objective disk drives should have fast access time and disk bandwidth. Access time is improved by scheduling the service of disk I/O in a good manner. Many processes make request for reading/writing data on disk simultaneously. As these requests sometimes makes requests faster than serviced by the disk. Therefore, a request queue has to hold disk requests. To reduce the time spent seeking records, the request queue is ordered in some manner. This process is called Disk scheduling.

A disk-scheduling algorithm decides that which request of cylinder is to be serviced when there are so many requests. Various disk-scheduling algorithms are used. However, there will be common criteria for evaluating the performance of all these algorithms that is total head movement. Each algorithm aims to minimise the total head movement. The algorithms can be evaluated by running them on a particular string of randomly generated requests and computing the access time of the moveable head disks.

Author [α] : Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, Haryana 136118, India.
Author [Ω] : M.M. Institute of Computer Technology & Business Management, M.M. University, Mullana, Ambala, Haryana, 133207, India. E-mail : sumit_amb@yahoo.com

Access Time has two major components. First one is Seek time and another on is Rotational Latency Time. The Seek Time is the time taken by read/write head to reach at a requested Cylinder/Track number and later one the time taken by the disk to rotate the desired sector under the read/write head. The disk bandwidth is defined as the total number of bytes transferred, divided by the total time between first request and completion of last transfer. Both the access time and disk bandwidth can be improved by scheduling the service of disk I/O in a good manner [7]. The time taken by a disk to move the required data under the read/write head is called rotational latency time. A disk's average rotational latency is simply half the time it takes to complete one revolution.

a) FCFS algorithm

This algorithm treats the requests of cylinders as a FIFO queue. Besides simplicity, this policy is preferred because this ensures that no request can be postponed indefinitely. This policy suffers from global zigzag effect.

b) SSTF algorithm

This algorithm selects the request, which has shortest seek from the current position of R/W head. As this policy can leads to indefinite postponement of the requests, which are not closer to R/W head. This policy gives a substantial improvement in performance, but it leads to problem of starvation.

c) SCAN algorithm

In this algorithm request is chosen for service that requires the shortest seek in preferred direction & do not change the direction until it reaches at the end of the disk. After that head moves in reverse direction and services all the requests in the opposite direction. This policy is also called as elevator algorithm.

d) C-SCAN algorithm

In C-Scan head moves only in one direction to service the requests. When head moves in reverse direction it does not service the incoming requests. When head has completed its inward sweep, it jumps to outermost cylinder without servicing the requests and then it resumes its inward sweep.

e) Look (Up/Down) algorithm

In this, head goes only as far as the final request in each direction. Then, it reverses direction immediately, without going all the way to end of disk. It

is appropriate to call the elevator algorithm as it continuous in one direction until it reaches the last request in that direction, then reverse direction.

f) C-Look algorithm

This algorithm reduces the bias against request located at the extreme ends of platters. When there is no request on a current sweep in either direction (inward or outward) the read/write head moves to the request closest to the outer/inner cylinder and again begins the next sweep.

II. Related Work

David M. Jacobson and John Wilkes [1] have discussed the disk scheduling algorithm based on rotational position in their research paper. Disk scheduling based on rotational position as well as disk arm position is shown to provide improved performance. The access time based algorithms match or outperform all the seek-time ones. The best of them is Aged Shortest Access Time First, or ASATF, which forms a continuum between FCFS and SATF. It is equal or superior to the others in both mean response time and variance over the entire useful range.

Margo Seltzer, Peter Chen and John Ousterhout [2] have jointly written a research paper "Disk Scheduling Revisited". In this paper, the invention of the movable head disk has been discussed. These techniques have been applied to systems with large memories and potentially long disk queues. Disk bandwidth utilisation can be improved by applying some traditional disk scheduling techniques, which attempt to optimise head movement and guarantee fairness in response time.

Daniel T. Joyce [3] in his article "An Investigation of Disk Scheduling Algorithms Laboratory" discussed the behaviour of disk scheduling algorithms by using a simulation program. The program is used to generate data that reflects the performance of the FCFS and SSTF algorithms under a variety of conditions. For each algorithm under each situation the program simulates how the algorithm would handle the situation and calculates the expected service time b/w requests, the expected waiting time for a request and the standard deviation of these waiting times.

Toby J. Teorey and Tad B. Pinkerton [4] has discussed five well-known scheduling policies for movable head disks. These policies are compared using the performance criteria of expected seek time and expected waiting time. The variance of waiting time is introduced as another meaningful measure of performance, showing possible discrimination against individual requests. Then the choice of a utility function to measure total performance including system oriented and individual request oriented measures is described.

Helen D. Karatza [5] has discussed scheduling in a distributed system. A simulation model is used to address performance issues associated with scheduling. Three policies which combine processor and I/O scheduling are used to schedule parallel jobs for a variety of workloads.

Hu Ming [6] has discussed disk-scheduling algorithms based on both disk arm and rotational positions. Their time-resolving powers are more precise in comparison with those for disk-scheduling algorithms based only on disk arm position. For modem disks, increase of disk rotation rate makes overhead of disk access to data transfer heavier. Therefore, it seems more important to improve both parallel processing capability of disk I/O and disk-scheduling performance at the same time.

III. Proposed Model

In this research effort, the problem under study is to optimize the performance of various disk scheduling algorithms before these are actually followed in any operating system and to design the simulator to mimic the real behaviour of the system. Because the seek distance between the position of head and position of requesting cylinder at the time of request is the basic need for evaluating the performance of the I/O system. Thus an efficient Disk Scheduling algorithm can enhance the performance of overall system whereas a poorly design scheme can degrade the performance. Thus to study the various algorithms, simulator is designed.

A simulation of any process in which there are inherently random components requires a method of generating random numbers. Thus whenever simulator is used, as a tool for research, there is need for generating random numbers that are conveniently and efficiently generated from a desired probability distribution. The present research work uses box-muller transformation for generation of cylinder numbers.

Suppose R_1 and R_2 are independent random variables that are uniformly distributed in the interval [0, 1].

$$S = (-2 \log_e (R_1))^{1/2} * \cos(2\pi R_2)$$

Here S is independent random variables with a normal distribution of standard deviation 1. In present research work, the foremost criterion for the evaluation of disk scheduling algorithms is the access time calculated by seek time, rotational delay and transfer time that are produced by each policy under same set of conditions and same workload. The workload here is the cylinder numbers whose data is to be accessed to perform I/O operation. This calculated access time is used to find out the total head movement for various disk scheduling algorithms.

$$T_A = Ts + T_R + T_T$$

Where

T_A (access time): sum of seek time, rotational latency time and transfer time.

T_S (seek time): time for the disk arm to move the heads to the cylinder containing the desired sector.

T_R (rotational delay): time waiting for the disk to rotate the desired sector to the disk head.

T_T (transfer time): the time it takes to transfer a block of bits to and from the disk.

Among these three, seek time has large significant effect on the total access time of the disk. As seek time is the time relating to cylinder number. Therefore cylinder number and number of seek movements are central point of consideration.

Simulator for Optimizing the Comparative Performance of Disk Scheduling Algorithms

N : no. of cylinders
NODE : current position of moveable read/write head
R_1/R_2 : two independent random variables in the interval [0, 1]
T_S (i) : seek time of N cylinders
T_A (i) : access time of N cylinders
T_R : rotational speed of the disk
T_T : transfer time between adjacent cylinders
RUNS : no. of times the simulation process is repeated
RAND : random number
L_TIME : latency time to move the head from one to another cylinder
CL[i] : left requests with respect to head position.
CR[j] : right requests with respect to head position.

Algorithm to compute the access time to read/write a disk

Step 1. Read no. of cylinders for different workload.
Step 2. Generate random numbers using the random number generation method in the interval of [0, 1].
Step 3. Compute the mean and standard deviation of m-pseudo random numbers.
Step 4. Apply Box-Muller transformation to calculate the value of S, using two random variates between [0, 1].
Step 5. Using the values of mean, standard deviation and S, calculate the value of x and store in an array x[i], which can use as the number of requests.
Step 6. Call modules for all seven policies named FCFS (), SSTF (), SCAN (), C-SCAN (), LOOK UP (), LOOK DOWN () and C-LOOK ().
Step 7. Compute access time based on seek time, rotational delay and transfer time produced by each policy is returned to the main module.
Step 8. Each algorithm is run for 20000 times and result of every 1000th run of each algorithm is displayed in a table.
Step 9. Stop

IV. Results

The best way to compare the result of different algorithms is to present them in form of table depicting the result in the form of rows and columns. Different test cases are simulated by varying the number of randomly generated cylinders and accordingly results are shown as in Table 1/Table 2/Table 3.

Test case 1: No. of cylinders (Low Laod) =200
Test case 2: No. of cylinders (Medium Laod) =700
Test case 3: No. of cylinders (Heavy Laod) =1200

Test Case 1: It is shown in the table 1 regarding total head movement of different disk scheduling algorithms in the case of low load on various simulation runs.

Simulation Runs	FCFS	SSTF	SCAN	C-SCAN	LOOK (UP)	LOOK (DN)	C-LOOK
1000	4065	574	289	376	187	107	194
2000	4677	459	2€84	325	199	146	229
3000	4629	1077	293	410	199	119	211
4000	3867	479	281	361	182	121	201
5000	4328	415	299	396	226	155	252
6000	4253	536	285	369	184	113	197
7000	4133	586	282	310	187	128	208
8000	4095	530	290	378	194	118	206
9000	4372	612	282	456	180	114	193
10000	4604	448	293	385	208	137	229
11000	4260	426	302	402	218	130	230
12000	4492	558	278	355	184	134	211
13000	4438	450	281	379	183	123	203
14000	3837	403	278	355	171	108	185
15000	4713	517	290	402	203	136	225
16000	4130	539	290	379	198	126	215
17000	4690	444	298	395	204	114	211
18000	4139	481	293	326	200	121	212
19000	4580	548	298	393	222	150	245
20000	4518	482	292	382	199	122	212

Table 1 : Total head movement for low load (No. of cylinders: 200)

Test Case 2: It is shown in the table 2 regarding total head movement of different disk scheduling algorithms in the case of medium load on various simulation runs.

Simulation Runs	FCFS	SSTF	SCAN	C-SCAN	LOOK (UP)	LOOK (DN)	C-LOOK
1000	15730	1057	287	1297	199	137	223
2000	15520	1069	290	1359	215	160	249
3000	14302	947	299	1415	213	129	227
4000	15615	976	285	1325	208	161	245
5000	15438	1210	292	1427	205	134	225
6000	15253	1026	296	1382	215	142	237
7000	15683	1106	294	1350	231	180	273
8000	14991	1117	297	1402	233	175	271
9000	15959	1132	304	1372	238	164	267
10000	15072	1043	289	1415	221	175	263
11000	14662	1098	293	1318	210	141	233
12000	14926	1128	300	1365	233	166	265
13000	14034	1057	288	1380	200	136	223
14000	16468	1026	297	1426	220	149	245
15000	15466	1100	289	1402	206	145	233
16000	15024	1178	290	1379	201	132	221
17000	15252	1065	284	1424	205	158	241
18000	14442	1106	286	1408	198	138	223
19000	15238	1352	291	1392	211	149	239
20000	15617	1094	289	1310	206	145	233

Table 2 : Total head movement for medium load(No. of cylinders: 700)

Test Case 3: It is shown in the table 3 regarding total head movement of different disk scheduling algorithms in the case of heavy load on various simulation runs.

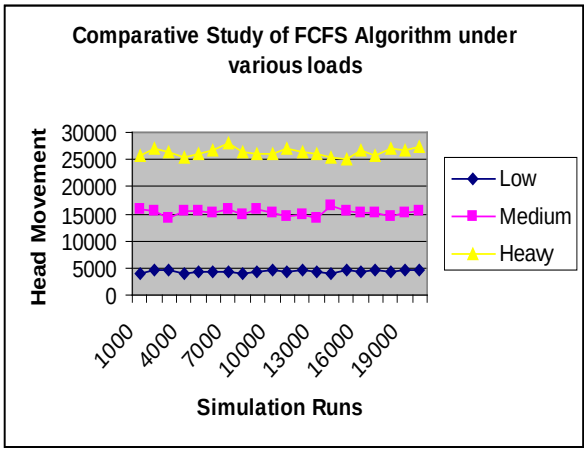
Simulation Runs	FCFS	SSTF	SCAN	C-SCAN	LOOK (UP)	LOOK (DN)	C-LOOK
1000	25629	1528	289	2397	209	151	239
2000	27118	1664	301	2382	240	177	277
3000	26256	1728	300	2356	223	146	245
4000	25234	1802	294	2326	228	174	267
5000	25969	1663	292	2415	215	154	245
6000	26546	1652	302	2340	233	160	261
7000	27861	1590	293	2502	224	169	261
8000	26404	1584	298	2448	228	162	259
9000	26019	1608	299	2397	229	161	259
10000	26055	1568	293	2415	215	151	243
11000	26978	1776	309	2345	242	157	265
12000	26299	1595	291	2300	210	147	237
13000	25891	1760	297	2417	222	153	249
14000	25360	1556	300	2397	233	166	265
15000	25035	1636	291	2396	226	179	268
16000	26601	1658	303	2318	248	187	289
17000	25792	1555	294	2368	217	152	245
18000	26916	1530	310	2420	250	170	279
19000	26671	1707	294	2382	213	144	237
20000	27463	1865	290	2392	212	154	243

Table 3 : Total head movement for heavy load (No. of cylinders: 1200)

V. Discussion and Conclusion

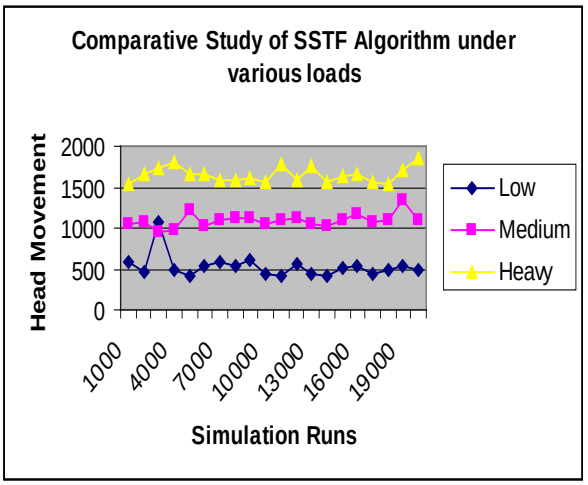
After analysing the results and findings of the simulator, it might be concluded no single policy is best in all situations. The performance do not depend upon only on the number of requests but it also depends on the position of read/write head & direction of the movement of head and it varies with the variation in number of requests even the current head position is same. It has been also observed that if there is only one outstanding request, then all the policies behave the same.

FCFS policy can be considered best for the system, which has fewer loads of Input-output requests, but in heavy load of requests, FCFS tends to saturate the device. SSTF produced least number of head movement in maximum runs as compared to FCFS. Therefore this policy is the optimal policy. But this policy can not be considered optimal as this policy has the starvation problem. LOOK has no starvation problem. But this policy has the overhead of decision variable, which is used to decide the direction (inward or outward) of read/write head. LOOK (Down) algorithm is always better than as compared to LOOK (UP) algorithm. C-Look disk scheduling algorithm performs better for those systems which puts medium and heavy load of requests on the disk. The graph 1 depicts the head movement for different number of simulation runs for FCFS algorithm under various loads.



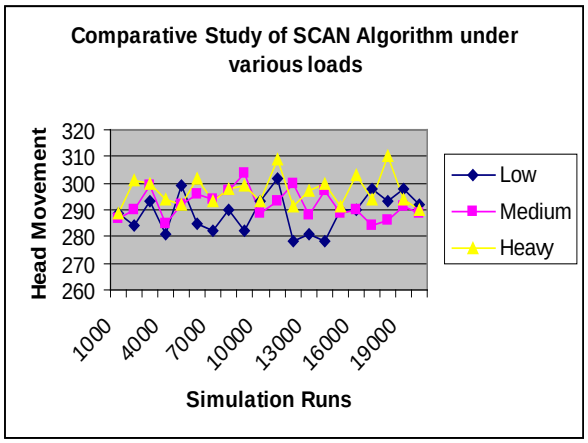
Graph No. 1

The graph 2 depicts the head movement for different number of simulation runs for SSTF algorithm under various loads.



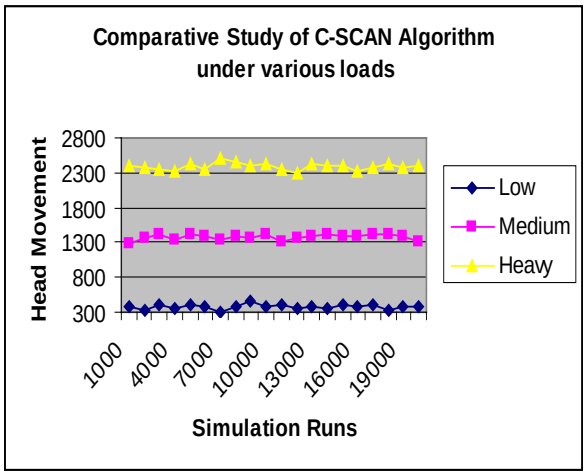
Graph No. 2

The graph 3 depicts the head movement for different number of simulation runs for SCAN algorithm under various loads.



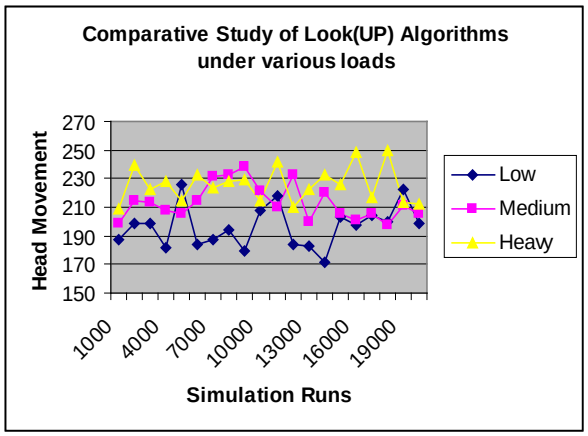
Graph No. 3

The graph 4 depicts the head movement for different number of simulation runs for C-SCAN algorithm under various loads.



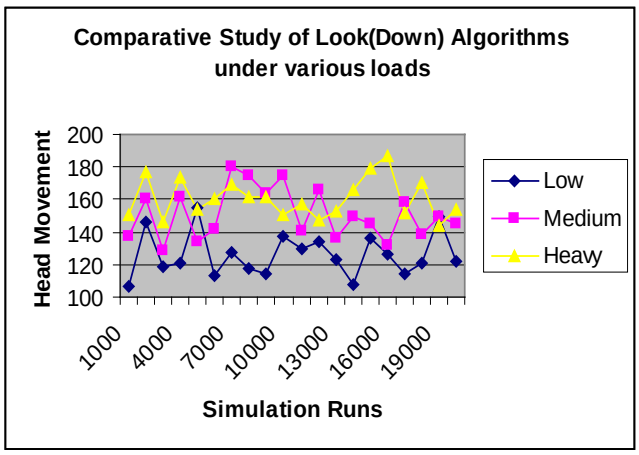
Graph No. 4

The graph 5 depicts the head movement for different number of simulation runs for Look (UP) algorithm under various loads.



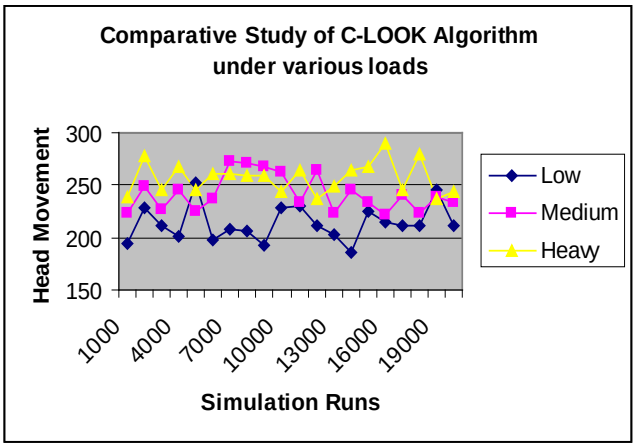
Graph No. 5

The graph 6 depicts the head movement for different number of simulation runs for Look (Down) algorithm under various loads.



Graph No. 6

The graph 7 depicts the head movement for different number of simulation runs for C-Look algorithm under various loads.



Graph No. 7

References References Referencias

1. David M. Jacobson and John Wilkes, "Disk scheduling algorithms based on rotational position" Hewlett Packard, May 1995.
2. Margo Seltzer, Peter Chen and John Ousterhout, "Disk Scheduling Revisited", Winter Usenix, Washington, January 1990.
3. Daniel T. Joyce, "An Investigation of Disk Scheduling Algorithms Laboratory", 2001.
4. Toby J. Teorey and Tad B. Pinkerton, "A comparative analysis of disk scheduling policies", March 1972, New York, NY, USA.
5. Helen D. Karatza," A Comparative Analysis of Scheduling Policies in A Distributed System Using Simulation", Thessaloniki, Greece, 2000.
6. Hu Ming, "Improved disk scheduling algorithms based on rotational position", October, 2005.
7. Silberschatz A.,P.B.Galvin et. al., "Operating System Concepts", 6th Edition, 2001.
8. Muhammad Younus Javed, Ihsan Ilah Khan, "Simulation and performance comparison of four disk scheduling algorithms", IEEE, 2000.
9. Deo. Narsingh, "System Simulation with Digital Computer", 15th edition, PHI, New Delhi, India, 2002.
10. Dietal H.M., "An Introduction to Operating Systems", Rev. 1st Edition Reading, Addition-Wesley, 1984.
11. Steven Robbins, "A Disk Head Scheduling Simulator", Norfolk, Virginia, USA, March, 2004.

Global Journal of Computer Science and Technology
Volume 11 Issue 18 Version 1.0 October 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Improved Privacy in Wireless Sensor Network Using QOS Routing Protocols

By Tenali. Nagamani, Damineni.SreeLakshmi

Dept of CSE PVPSIT Kanuru, Vijayawada

Abstract - Full network level privacy has often been categorized into four sub-categories:Identity, Route, Location and Data privacy. Achieving full network level privacy is a challenging problem due to the conditions imposed by the sensor nodes (e.g., energy, memory and computation power), sensor networks (e.g., mobility and topology) and QoS issues (e.g., packet reach-ability and timeliness). This proposed paper consists of two algorithms IRL algorithm and data privacy mechanism that addresses this problem. The proposed system provides additional trustworthiness, less computation power, less storage space and more reliability. Also, we proved that our proposed solutions provide protection against various privacy disclosure attacks, such as eavesdropping and hop-by-hop trace back attacks.

Keywords : *anonymity; eavesdropping; hop-by-hop trace back; privacy; routing; wireless sensor networks.*

GJCST-F Classification : *C.2.2*

Strictly as per the compliance and regulations of:

Improved Privacy in Wireless Sensor Network Using QOS Routing Protocols

Tenali. Nagamani [α], Damineni.SreeLakshmi [Ω]

Abstract - Full network level privacy has often been categorized into four sub-categories: *Identity*, *Route*, *Location* and *Data* privacy. Achieving full network level privacy is a challenging problem due to the conditions imposed by the sensor nodes (e.g., energy, memory and computation power), sensor networks (e.g., mobility and topology) and QoS issues (e.g., packet reach-ability and timeliness). This proposed paper consists of two algorithms IRL algorithm and data privacy mechanism that addresses this problem. The proposed system provides additional trustworthiness, less computation power, less storage space and more reliability. Also, we proved that our proposed solutions provide protection against various privacy disclosure attacks, such as eavesdropping and hop-by-hop trace back attacks.

Keywords : anonymity; eavesdropping; hop-by-hop trace back; privacy; routing; wireless sensor networks.

I. Introduction

In order to present the adversary from back-tracing, the route, location and data privacy mechanism must be enforced. With the spreading application of Wireless Sensor Networks (WSNs) in various sensitive areas such as health-care, military, habitat monitoring, etc. Network level privacy often been categorized into 4 categories:

1. Sender node identity privacy: no intermediate node can get any information about who is sending the packets except the source, its immediate neighbors and the destination.
2. Sender node location privacy: no intermediate node can have any information about the location (in terms of physical distance or number of hops) about the sender node except the source, its immediate neighbors and the destination.
3. Route privacy: no node can predict the information about the complete path (from source to destination). Also, a mobile adversary gets no clue to trace back the source node either from the contents and/or directional information of the captured packet(s).
4. Data packet privacy: no node can see the information inside in a payload of the data packet except the source and the destination.

Author [α] : M.Tech, CSE, PVPSIT, Kanuru, Vijayawada.
E-mail : tenalinagamani@gmail.com
Author [Ω] : M.Tech, Vijayawada, Asst. Professor, Dept of CSE PVPSIT, Kanuru, Vijayawada. E-mail : damineni.mtech@gmail.com

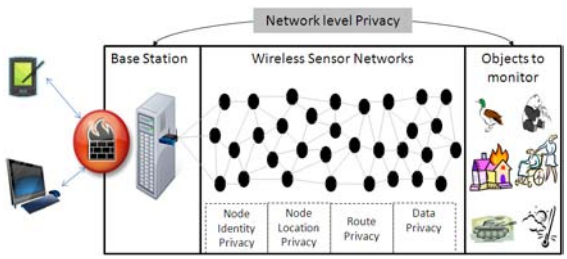
An energy-efficient privacy solution is needed to address these patterns in Wireless Sensor Network. Advanced features in cryptographic system were introduced in this paper are:

- A new Identity, Route and Location (IRL) privacy algorithm is proposed that ensures the source, identity and location. This algorithm allows the packets to destination only through trusted intermediate nodes.
- The extension of our proposed IRL algorithm is a new reliable Identity, Route and Location (r-IRL) privacy algorithm. This algorithm has the ability to forward packets from multiple secure paths to increase the packet reach-ability.
- A data privacy mechanism is used to unique in the sense that it provides secure data and packet authentication.

a) Network and Assumptions Model

A wireless sensor network (WSN) is composed of large number of small sensor nodes that are of limited resource and densely deployed in an environment. This sensor node uses IEEE 802.11 standard link layer protocol, which keeps packets in its cache until the sender receives an acknowledgment (ACK). The sender node will retransmit the packet, if the ACK does not receive within threshold.

Figure 1 : Typical WSN scenario.



II. Proposed Scheme

a) Concepts and Definitions

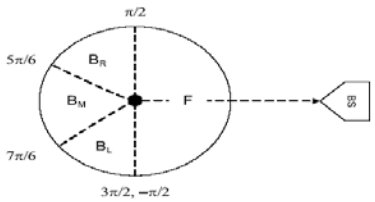
The proposed algorithms use two notions: Direction and Trust. These notions are used to provide reliable secure paths for achieving robust route privacy. Direction helps to forward packet to the destination in a timely manner and trust will help to forward the packets via reliable nodes.

Direction: The first notion used in our algorithms is that of direction. The physical location of the base station is the reference point for each sensor node. Based on this reference point, each node classifies its neighboring nodes into four categories: (1) forward neighboring nodes (F), (2) right side backward neighboring nodes (Br), (3) left side backward neighboring nodes (B_l), and (4) middle backward neighboring nodes (B_m). The objective of this categorization is to provide more path diversity as discussed in Section 4.2. A node x classifies its neighboring node y in following fashion :

$$C_{x,y} = \begin{cases} F & \frac{-\pi}{2} \leq \theta \leq \frac{\pi}{2} \\ B_r & \frac{\pi}{2} < \theta \leq \frac{5\pi}{6} \\ B_m & \frac{5\pi}{6} < \theta \leq \frac{7\pi}{6} \\ B_l & \frac{7\pi}{6} < \theta < \frac{3\pi}{2} \end{cases}$$

Where θ is the angle between the node x and its neighboring node y with respect to the line joining node x and the base station as shown in Figure 2.

Figure 2 : Neighbor node classification



Trust: The second notion used in our algorithms is that of trust. The definition of a trust here is based on our other paper and restated here. A node can be classified into one of the three categories: trustworthy, untrustworthy, and uncertain. A node is considered trustworthy if it interacts successfully most of the time with the other nodes. A node is considered untrustworthy if it tries to do as many unsuccessful interactions as possible with the other nodes. An untrustworthy node could be a faulty or malicious node. A node is considered uncertain if it performs both successful and unsuccessful interactions. Detailed definition if the successful and unsuccessful interactions and trust calculation methodology is available in our paper and provided here in a simplified form.

A sender will consider an interaction successful if the sender receives confirmation that the packet is successfully received by the neighbor node and forwarded towards the destination in an unaltered fashion. The first requirement of successful reception is achieved on the reception of the link layer acknowledgment (ACK). The second requirement of forwarding towards the destination is achieved with the help of enhanced passive acknowledgment (PACK) by overhearing the transmission of a next hop on the route, since they are within the radio range. If the sender node does not overhear the retransmission of the packet within a threshold time from its neighboring node or if the overheard packet is found to be illegally fabricated (by comparing the payload that is attached to the packet), then the sender node will consider that interaction as unsuccessful.

With this simple approach, several attacks can be prevented, i.e., the black hole attack is straightforwardly detected when malicious node drops the incoming packets and keeps sending self-generated packets .Similarly, sink hole attack, an advanced version of the black hole attack, is also easily detectable by looking at the passive acknowledgment. Likewise, the selective forwarding attack and gray-hole attack [27] can also be eliminated with the aid of above mentioned approach. Based on these successful and unsuccessful interactions node x can calculate the trust value of node y in following fashion:

$$T_{x,y} = \left[100\left(\frac{S_{x,y}}{S_{x,y}+U_{x,y}}\right)\left(1-\frac{1}{S_{x,y}+1}\right)\right]$$

Where [.] is the nearest integer function, $S_{x,y}$ is the total number of successful interactions of node x with y during time δt, and $U_{x,y}$ is the total number of unsuccessful interactions of node x with y during time δt. After calculating trust value, a node will quantize trust into three states as follows:

$$Mp(T_{x,y}) = \left\{ \begin{array}{cc} \text{trustworthy} & 100 - f \leq T_{x,y} \leq 100 \\ \text{uncertain} & 50 - g \leq T_{x,y} < 100 - f \\ \text{untrustworthy} & 0 \leq T_{x,y} < 50 - g \end{array} \right\}$$

Where, f represents half of the average values of all trustworthy nodes and g represents one-third of he average values of all untrustworthy nodes. Both f and g are calculated as follows:

$$f_{j+1} = \begin{cases} \left[\frac{1}{2}\left(\frac{\sum_{i \in R_x} T_{x,i}}{|R_x|}\right)\right] & 0 < |R_x| \leq n-1 \\ f_j & |R_x| = 0 \end{cases}$$

$$g_{j+1} = \begin{cases} \left[\frac{1}{3}\left(\frac{\sum_{i \in M_x} T_{x,i}}{|M_x|}\right)\right] & 0 < |M_x| \leq n-1 \\ g_j & |M_x| = 0 \end{cases}$$

The steady-state operation, these values can change with every passing unit of time which creates dynamic trust boundaries. After each passage of time, Δt, nodes will recalculate the values of f and g. This trust calculation procedure will continue in this fashion.

The time window length (Δt) could be made shorter or longer based on the network analysis scenarios. If Δt is too short, then the calculated trust value may not reflect the reliable behavior. On the other hand, if it is too long, then it will consume too much memory to store the interaction record at the sensor node. Therefore, various parameters can be used to adjust the length of Δt.

Where [.] is the nearest integer function, Rx represents the set of trustworthy nodes for node x, Mx the set of untrustworthy nodes for node x, and n is the total number of nodes that contains trustworthy, untrustworthy and uncertain nodes. The initial trust

values of all nodes are 50.The values of f and g are adaptive.

b) Identity, Route, and Location Privacy (IRL)

The proposed identity, route and location privacy scheme works in two phases. The first is neighbor node state initialization phase, and the second is routing phase.

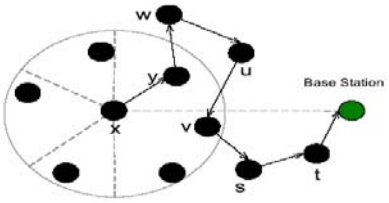
Route Privacy: In initialization phase, let the node i have m neighboring nodes in which t nodes are trusted. So, $0 \leq t \leq m$ and $M(t) = M(tF) \cup M(tBr) \cup M(tBl) \cup M(tBm)$. Here $M(tF)$, $M(tBr)$, $M(tBl)$, and $M(tBm)$ represent the set of trusted nodes that are in the forward, right backward, left backward, and middle backward directions, respectively. These neighbor sets ($M(tF)$, $M(tBr)$, $M(tBl)$,and $M(tBm)$) are initialized and updated whenever a change occur in neighborhood. For example, the entrance of a new node, change of a trust value, *etc*.

Whenever a node needs to forward a packet, the routing phase for source node and for intermediate node) of IRL algorithm is called.

Whenever a source node wants to forwards the packet, it will first check the availability of the trusted neighboring nodes in its forward direction set $M(tF)$. If trusted nodes exists then it will randomly select one node as a next hop from the set $M(tF)$ and forward the packet towards it .If there is no trusted node in its forward direction, then the source node will check the availability of a trusted node in the right ($M(tBr)$) and left ($M(tBl)$) backward sets. If the trusted nodes are available then the source node will randomly select one node as a next hop from these sets and forward the packet towards it.If the trusted node does not exist in these sets either, then the source node will randomly select one trusted node from the backward middle set ($M(tBm)$) and forward the packet towards it.If there are no trusted nodes available in all of the sets then the packet will be dropped .

When an intermediate node receives the packet (either from the source node or from another en-route node), it will first check whether the packet is new or old.If it is new, then the node will first check the availability of the trusted node from the forward direction set (*MF*) excluding the *prevhop* node if it belongs to forward set.If trusted nodes exists in the forward set then the node will randomly select any one trusted node as a next hop and forward the packet towards it. If there is no trusted node available in the forward direction, then it will check to which set the sender of the packet belongs to. For example, If the packet, forwarded by a node, belongs to the right backward set, then it will first check whether the left or middle backward sets contain any trusted nodes. If so, it will randomly select one node from those sets and forward the packet towards it. If there is no trusted node in those two sets, then the node will randomly select a trusted node from the right backward set ($M(tBr)$) excluding the one from which the node received the current packet and forward the packet towards it.Similar operations will be performed, if the packet, forwarded by a node, belongs to the left and middle backward or forward sets. An example IRL routing scenario is shown in Figure 3.

Figure 3 : Sample routing scenario of IRL scheme.

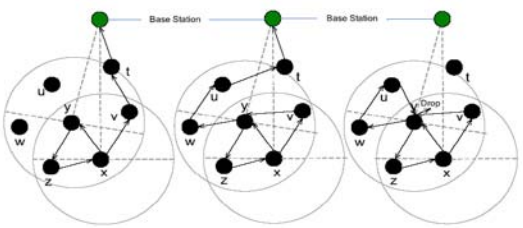


This routing strategy may result in the creation of a cycle (loop). However, due to the randomness in the selection of the next-hop and the presence of the different four direction sets, the probability of creation of any cycle is very low. Nevertheless, in order to fully avoid the occurrence of the cycles, each node (prior to forwarding of a packet) will save the signature of the packet in the buffer for the δt time,that is:

$$\delta t = 2\left(\frac{D}{d} \times p_t\right)$$

Where D is the distance between the forwarding node and the base station, d is the distance between the forwarding node and the next hop, and pt is the propagation transfer time between the forwarding node and the next hop. This signature consists of two fields: (1) sequence number of the packet, and (2) the payload. The potential of the signature to compare and identify the same packet is detailed in the later section. Corresponding to this signature, three more fields are also stored in the buffer: (1) Previous hop identity, (2) next hop identity where the packet is forwarded, and (3) Counter, that tells how many times the same packet is received by the node. This information will later be used to get rid of any cycle. The size of the buffer is mainly dependent on the network traffic conditions. However, it is expected to be low due because the sensor nodes sent data either in periodic intervals or upon the occurrence of some event.

Figure 4 : Three sample cycle detection and prevention scenarios.



If the node received the packet whose signature exists in the buffer, then including the previous hop node, two other nodes will also be excluded from the

selection of the next hop process: 1) the node from which last time the packet was received the node from which last time the packet was forwarded. If the same packet is received three times by the same node then the packet will be dropped. Three sample scenarios of the loop creation, detection and prevention are shown in Figure 4. Creation of loops and traversing of the packets in the backward direction is not a completely negative effect. Rather, it provides positive effects in terms of strengthening the route and source location privacy, because these effects will helps to increase the safety period, which is the time for an adversary to reach at the source node.

Identity Privacy: Whenever a node receives the packet *p* from the source node or en-route node then the receiving node will replace the previous hop's identity *prevhop* contained in the packet with its own.After that, the node will get the next forwarding node *nexthop* and update the header of the packet *p* = *{prevhop, nexthop, payload}*. After modification of the two header fields, the node will forward the packet.In this way, all the intermediate forwarding nodes replace the source and next hop's identity contained in the packet *p*. This process will go on until the packet reaches the base station.

Location Privacy: The neighboring nodes which are in each other's radio range can easily approximate the location of each other by measuring the received signal strength and the angle of arrival.If the adversary is within the range of the source node, then adversary can easily estimate the location of the source. Once the packet has crossed the radio range of the original source node, then becomes very difficult for an attacker to estimate the location of the node either in terms of the physical distance or in terms of the number of hops of an original source node. The main reason for this is that the path selection is random and packets are forwarded by only trusted nodes which only contain the information of the last and the next hop.

c) Reliable Identity, Route, and Location Privacy (r-IRL)

It is also possible that some applications require more reliability in terms of packet reach-ability; and the packet could be dropped due to either network congestion or malicious behavior of an en-route node. Thus, in order to achieve more reliability, the packet should be forwarded from multiple paths simultaneously, which will give trustworthiness in the sense that at least the packet should reach the base station by any one of the paths, although, this may increase some communication overhead. Our reliable IRL (r-IRL) algorithm is the extended version of our proposed IRL algorithm, in which we introduce one more parameter, reliability *r*. The source node *i* will multi-cast a packet to all *r* randomly selected neighboring trusted nodes that are in the forward direction. If there are no adequate trusted nodes present in the forward direction, then it will select the remaining trusted nodes from the backward direction. The rest of the mechanism of the r-IRL algorithm is the same as the IRL algorithm.

d) Data Privacy

The payload contains the identity of the source node (IDx) and the actual data (d). Identity is encrypted with the public key $(k+bs)$ of the base station and data is encrypted with the secret key (kx,bs) shared between the sender node and the BS. Both are appended with the payload as shown below:

$$Payload = [E(IDx, k+ bs), E(d, kx, bs)]$$

If we assume that the adversary knows the range of identities assigned to the sensor nodes, public key of the base station and information about cipher algorithm used in the network, an adversary can then successfully obtain the identity of the source by performing simple brute-force search attack by comparing the pattern of encrypted identity with a known range of identities. Therefore in order to provide protection against brute-force search attack, we append a random number (Rn) (equivalent to the size of identity) with the identity of a node and then perform encryption. Now the payload is:

$$Payload = [E(IDx||Rn, k+bs), E(d, kx, bs)]$$

Where // is the append operation. Inclusion of random number may introduce additional computational overhead. However, the amount of overhead is mainly dependent on random number generation technique. Recently, very nice random generation techniques have been specially designed for low power sensor networks, such as. These techniques could be used to generate random number for each packet. Also, overall computational overhead is dependent on the number of packets generated by the sensor nodes.

Our proposed data privacy approach provides several benefits. Firstly, data secrecy is achieved in the presence of identity anonymity. This feature is not available in earlier proposed privacy schemes. Secondly, the base station will receive both the identity of the actual source node and message authentication. If the packet has been successfully decrypted with the shared secret key, it means that packet is received from genuine sensor node.

III. Analysis and Evaluation

a) Energy Consumption Analysis

This section, shows the efficiency of our routing strategies with existing schemes. Energy is computed based on the communication overhead (including transmission and reception cost, path length) introduced by our proposed routing protocols and compared it with other existing schemes.

Table 4 : Simulation parameters

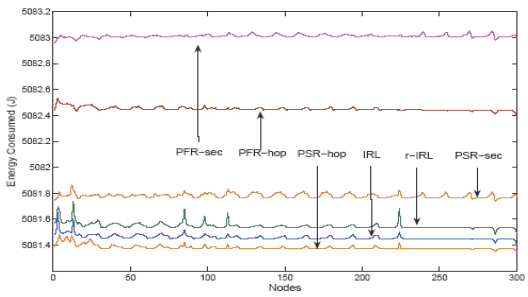
Network specific	Number of nodes	300
	Distance b/w nodes	50 units
	Mobility of nodes	zero
Node specific	Sensor node's Initial battery	1 × 106J
	Power consumption for trans.	1.6W
	Power consumption for recv.	1.2 W
	Idle power consumption	1.15W
	Carrier sense threshold	3.65e−10W
	Receive power threshold	1.55e−11W
	Frequency	9.14e8
	Trans. & Recv. antenna gain	1.0
Protocol & Application specific	Application	CBR
	Reliability param. *r* for r-IRL	3
	hwalk param. for PFR & PSR	10

The proposed paper has implemented our IRL and r-IRL routing schemes on Sensor Network Simulator and Emulator (SENSE).At the application layer we used constant bit rate component (CBR) that generate constant traffic during simulation between randomly selected source node(s) and the base station. For the simplicity, assume that both sensor nodes and the base station are static. Network consists of 300 sensor nodes that are organized into 15 by 20 grid manner.

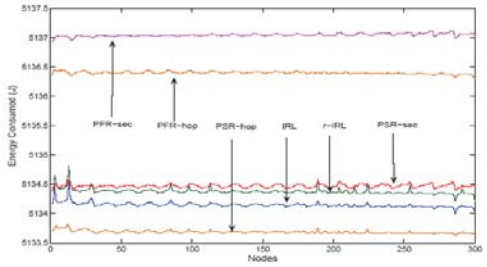
Comparison of proposed IRL and r-IRL algorithms with the four variations of phantom routing schemes that are:

1. Phantom single path routing scheme with hop-based approach (PSR-hop).
2. Phantom single path routing scheme with sector-based approach (PSR-sec).
3. Phantom flood routing scheme with hop-based approach (PFR-hop).
4. Phantom flood routing scheme with sector-based approach (PFR-sec).

Figure 6 : Energy consumption analysis: simulation time: 5,000.



(a) Source node 5

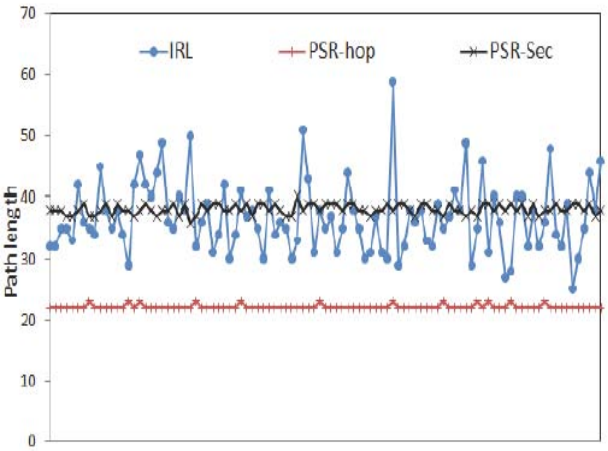


(b) Source node 10

The energy consumption analysis with different scenarios are shown in Figure 6. For the r-IRL scheme we select r = 3, which means a single packet will reach the destination via three different routes simultaneously. For phantom routing schemes, we select parameter $hwalk$=10 (as recommended. Figure 6 clearly indicates that, the IRL and r-IRL schemes consume less energy as compared to the PSR-sec, PFR-hop and PFR-sec schemes but slightly consume higher energy as compared to the PSR-hop scheme. This is due to the fact that the IRL and r-IRL algorithms provides more path diversity and packets sometimes took longer paths.

Our proposed routing strategies (IRL and r-IRL) have both features. Because of the concept of *direction* (Section 3.1), proposed schemes provide more length variation and because of the *randomness* (Section 3.2) proposed schemes provide high path variation. Incorporation of both features offer high path diversity.

Figure 7 : Path diversity of privacy schemes.



In order to analyze the path diversity behavior, assume 300 sensor nodes in a 10 by 30 grid manner. In the simulation, a single source node (ID: 224) generates 100 data packets for the base station. Figure 7 shows the path diversity (in terms of path length) of the IRL, PSR-hop and PSR-sec schemes.

The average path taken by the PSR-hop, IRL and PSR-sec is 22.12, 36.81 and 38.17, respectively. It indicates that the IRL scheme incurs more delay as compared with the PSR-hop scheme and less delay as compared with the PSR-sec scheme. This figure also indicates that the IRL scheme has more path variation as

compared with the other schemes, which creates more difficulties for the adversary to trace back the source from the captured packets.

Figure 7 also shows that some packets took longer paths in the IRL scheme as compared with others. This is due to the fact that the source or en-route node did not find any trusted node in its forward direction, so the packet is relayed back in the backward direction.

Figure 8 shows the result of 100 simulation runs in each node has equal probability to be trusted and un-trusted. It shows that, as the neighborhood size increases, the probability of the packet to move in the backward direction decreases sharply.

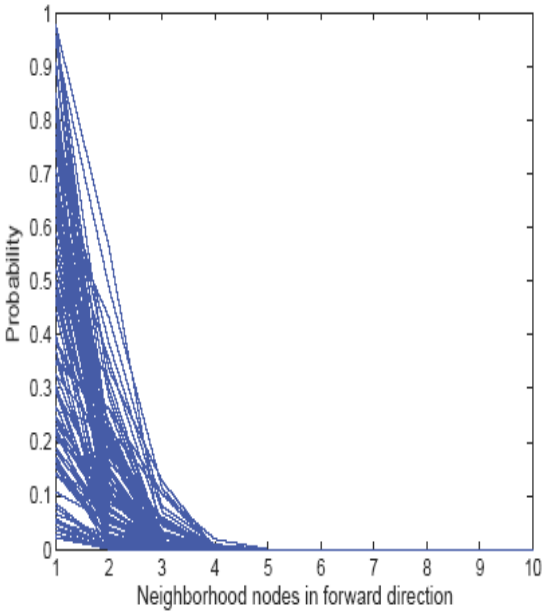


Figure 8 : Probability of a packet to move in the backward direction.

IV. Conclusions and Future Work

Existing privacy schemes of WSNs only provides partial network level privacy. Providing full network level privacy is a critical and challenging issue due to the constraints imposed by the sensor nodes (e.g.,energy, memory and computation power), sensor network (e.g., mobility and topology) and QoS issues (e.g., packet reach-ability and timeliness). Therefore, in this paper we proposed the first full network level privacy solution that is composed of two new identity, route and location privacy algorithms and data privacy mechanism. Our solutions provide additional trustworthiness and reliability at modest cost of energy and memory. Future work, will evaluate proposed schemes from the perspective of computation cost that is required to perform encryption and random number generation.

Bibliography

1. Wood, A.D.; Fang, L.; Stankovic, J.A.; He, T. SIGF: A Family of Configurable, Secure Routing Protocols for Wireless Sensor Networks. In Proceedings of the 4th ACM Workshop on Security of Ad Hoc and Sensor Networks, Alexandria, VA, USA, 2006; pp. 35–48.
2. Kamat, P.; Zhang, Y.; Trappe, W.; Ozturk, C. Enhancing Source-Location Privacy in Sensor Network Routing. In Proceedings of the 25th IEEE International conference on Distributed Computing Systems, Columbus, OH, USA, 2005; pp. 599–608.
3. Misra, S.; Xue, G. Efficient Anonymity Schemes for Clustered Wireless Sensor Networks. Int. J. Sens. Netw. 2006.
4. Habitat monitoring on Great Duck Island (Maine, USA), 2002. Available online: http://ucberkeley.
5. Ozturk, C.; Zhang, Y.; Trappe,W. Source-Location Privacy in Energy-Constrained Sensor Network Routing. In Proceedings of the 2nd ACM workshop on Security of Ad hoc and Sensor Networks, Washington, DC, WA, USA, 2004; pp. 88–93.
6. Xi, Y.; Schwiebert, L.; Shi, W.Preserving Source Location Privacy in Monitoring-Based Wireless Sensor Networks. In *Proceedings of Parallel and Distributed Processing Symposium (IPDPS2006)*, Rhodes Island, Greece.
7. Capone, A.; Pizziniaco, L.; Filippini, I.; de la Fuente, M.G. SiFT: An Efficient Method for Trajectory Based Forwarding. In Proceedings of International Symposium on Wireless Communication Systems, Siena, Italy.
8. Zorzi, M.; Rao, R.R. Geographic Random Forwarding (GeRaF) for Ad Hoc and Sensor Networks: Energy and Latency Perfor-mance. IEEE Tran. Mob. Comput. 2003.
9. Blum, B.; He, T.; Son, S.; Stankovic, J. IGF: A State-Free Robust Comm-unication Protocol for Wireless Sensor Networks; Technical Report CS-2003-11; Department of Computer Science, University of Virginia, USA, 2003.
10. Barbeau, M.; Kranakis, E.; Krizanc, D.; Morin, P. Improving Distance Based Geographic Location Techniques in Sensor Networks. In Proceedings of 3rd International Conference on Ad Hoc Networks and Wireless, Vancouver, British Columbia, 2004.
11. RYU, J.; Kim, S.G.; Choi, H.H.; An, S.S.; Ahn, S.Y.; Kim, B.J. Method and System for Locating Sensor Node in Sensor Network Using Transmit Power Control. U.S. Patent Application: 2009/0128298 A1.
12. Barbeau, M.; Kranakis, E.; Krizanc, D; Morin, P. Improving Distance Based Geographic Location Techniques in Sensor Networks. In Proceedings of 3rd International Conference on Ad Hoc Networks and Wireless, Vancouver, British Columbia, 2004.

13. Gaubatz, G.; Kaps, J.-P.; Sunar, B. Public Key Cryptography in Sensor Networks-Revisited. Lect. Note. Comput. Sci. 2006, 3313, pp. 2–18.
14. Lopez, J. Unleashing Public-Key Cryptography inWireless Sensor Networks. J. Comput. Security 2006.
15. Armenia, S.; Morabito, G.; Palazzo, S. Analysis of Location Privacy /Energy Efficiency Tradeoffs in Wireless Sensor Networks. In IFIP-Networking 2007.

This page is intentionally left blank

Global Journal of Computer Science and Technology
Volume 11 Issue 18 Version 1.0 October 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Software Effort Estimation Using Particle Swarm Optimization with Inertia Weight

By Prasad Reddy.P.V.G.D, Ch.V.M.K.Hari

Andhra University

Abstract - Software is the most expensive element of virtually all computer based systems. For complex custom systems, a large effort estimation error can make the difference between profit and loss. Cost (Effort) Overruns can be disastrous for the developer. The basic input for the effort estimation is size of project. A number of models have been proposed to construct a relation between software size and Effort; however we still have problems for effort estimation because of uncertainty existing in the input information. Accurate software effort estimation is a challenge in Industry. In this paper we are proposing three software effort estimation models by using soft computing techniques: Particle Swarm Optimization with inertia weight for tuning effort parameters. The performance of the developed models was tested by NASA software project dataset. The developed models were able to provide good estimation capabilities.

Keywords : *PM- Person Months, KDLOC-Thousands of Delivered Lines of Code, PSO - Particle Swarm Optimization, Software Cost Estimation*

GJCST-C Classification : *D.2.9*

Strictly as per the compliance and regulations of:

Software Effort Estimation Using Particle Swarm Optimization with Inertia Weight

Prasad Reddy.P.V.G.D [α], Ch.V.M.K.Hari [Ω]

Abstract - Software is the most expensive element of virtually all computer based systems. For complex custom systems, a large effort estimation error can make the difference between profit and loss. Cost (Effort) Overruns can be disastrous for the developer. The basic input for the effort estimation is size of project. A number of models have been proposed to construct a relation between software size and Effort; however we still have problems for effort estimation because of uncertainty existing in the input information. Accurate software effort estimation is a challenge in Industry. In this paper we are proposing three software effort estimation models by using soft computing techniques: Particle Swarm Optimization with inertia weight for tuning effort parameters. The performance of the developed models was tested by NASA software project dataset. The developed models were able to provide good estimation capabilities.

Index Terms : PM- Person Months, KDLOC-Thousands of Delivered Lines of Code, PSO - Particle Swarm Optimization, Software Cost Estimation.

I. Introduction

The modern day software industry is all about efficiency. With the increase in the expanse and impact of modern day software projects, the need for accurate requirement analysis early in the software development phase has become pivotal. The provident allocation of the available resources and the judicious estimation of the essentials form the basis of any planning and scheduling activity. For a given set of requirements, it is desirable to cognize the amount of time and money required to deliver the project prolifically. The chief aim of software cost estimation is to enable the client and the developer to perform a cost – benefit analysis. The software, the hardware and the human resources involved add up to the cost of a project. The cost / effort estimates are determined in terms of person-months (pm) which can be easily interchanged to actual currency cost.

The basic input parameters for software cost estimation is size, measured in KDLOC (Kilo Delivered Lines Of Code). A number of models have been evolved to establish the relation between Size and Effort [13]. The parameters of the algorithms are tuned using Genetic Algorithms [5], Fuzzy models [6], Soft-Computing Techniques [7][9][10], Computational Intelligence Techniques[8],Heuristic Algorithms, Neural Networks, Radial Basis and Regression [11][12] .

Author α : Department of Computer Science & Systems Engineering, Andhra University. E-mail : prasadreddy.vizag@gmail.com
Author Ω : Department of IT, Gitam Institute of Technology, GITAM University. E-mail : kurmahari@gmail.com

a) Basic Effort Model

A common approach to the estimation of the software effort is by expressing it is as a single variable function of the project size. The equation of effort in terms of size is considered as follows:

$$\text{Effort} = a * (\text{Size})^{b} \quad (1)$$

Where a, b are constants. The constants are usually determined by regression analysis applied to historical data.

b) Standard PSO with Inertia Weights

In order to meet the needs of modern day problems several optimization techniques have come been introduced. When the search space is too large to search exhaustively, population based searches may be a good alternative, however, population based search techniques cannot guarantee you the optimal (best) solution. We will discuss a population based search technique, Particle Swarm Optimization (PSO) with Inertia Weights [Shi and Eberhart 1998]. Particle Swarm has two primary operators: Velocity update and Position update. During each generation each particle is accelerated toward the particles previous best position and the global best position. At each iteration a new velocity value for each particle is calculated based on its current velocity, the distance from its previous best position, and the distance from the global best position. The new velocity value is then used to calculate the next position of the particle in the search space. The inertia weight is multiplied by the previous velocity in the standard velocity equation and is linearly decreased throughout the run. This process is then iterated a set number of times or until a minimum error is achieved.

The basic concept of PSO lies in accelerating each particle towards its Pbest and Gbest locations with regard to a random weighted acceleration at each time. The modifications of the particle's positions can be mathematically modeled by making use of the following equations:

$$V_i^{k+1} = w * V_i^{k} + c_1 * \text{rand}()_1 * (\text{Pbest} - S_i^{k}) + c_2 * \text{rand}()_2 * (\text{Gbest} - S_i^{k}) \quad (2)$$

$$S_i^{k+1} = S_i^{k} + V_i^{k} \quad (3)$$

Where,

S_i^k is current search point,

S_i^{k+1} is modified search point,

V_i^k is the current velocity,

V^{k+1} is the modified velocity,

V_{pbest} is the velocity based on Pbest ,

V_{gbest} = velocity based on Gbest,

w is the weighting function,

c_j is the weighting factors,

Rand() are uniformly distributed random numbers between 0 and 1.

II. The Standard PSO With Inertia Weight For Software Effort Estimation

The software effort is expressed as a function of a single variable of effort in terms of the project size as shown in equation-1. The parameters a, b are measured by using regression analysis applied to historical data. In order to tune these parameters we use the standard PSO with inertia weights. A nonzero inertia weight introduces a preference for the particle to continue moving in the same direction it was going on the previous iteration. Decreasing the inertia over time introduces a shift from the exploratory (global search) to the exploitative (local search) mode. The updating of weighting function is done with the following formula.

$$W_{new} = [(T_{mi} - T_{ci}) * (W_{iv} - W_{fv})] / T_{mi} + W_{fv} \quad (4)$$

Where

W_{new} is new weight factor,

T_{mi} is the maxium numer of iteration specified,

T_{ci} is the current iteration number,

W_{iv} is the initial value of the weight,

W_{fv} is the final value of the weight.

Empirical experiments have been performed with an inertia weight set to decrease linearly from 0.9 to 0.4 during the course of simulation. In the first experiment we keep the parameters c1 and c2 (weighting factors) fixed, while for the following experiment we change c1 and c2 (weighting factors) during subsequent iterations by employing the following equations [Rotnaweera, A. Halgamog S.K. and Watson H.C, 2004].

$C_1(t) = 2.5 - 2 * (t / max\_iter)$, which is the cognitive learning factor. (5)

$C_2(t) = 0.5 + 2* (t / max\_iter)$, which is the social coefficient. (6)

The particles are initialized with random position and velocity vectors the fitness function is evaluated and the Pbest and Gbest of all particles is found out. The particles adjust their velocity according to their Pbest and Gbest values. This process is repeated until the particles exhaust or some specified number of iterations takes place. The Gbest particle parameters at the end of the process are the resultant parameters.

III. Model Description

In this model we have considered "The standard PSO with inertia weights" with /without changing the weighting factors (c1, c2). PSO is a robust stochastic optimization technique based on the movement of swarms. This swarm behavior is used for tuning the parameters of the Cost/Effort estimation. As the PSO is a random weighted probabilistic model the previous benchmark data is required to tune the parameters, based on that data, swarms develop their intelligence and empower themselves to move towards the solution. The following is the methodology employed to tune the parameters in each proposed models following it.

a) METHODOLOGY (ALOGORITHM)

Input: Size of Software Projects, Measured Efforts, Methodology (Effort Adjustment factor-EAF).

Output: Optimized Parameters for Estimating Effort.

The following is the methodology used to tune the parameters in the proposed models for Software Effort Estimation.

Step 1: Initialize "n" particles with random positions P_i and velocity vectors V_i of tuning parameters .We also need the range of velocity between [- V_{max},V_{max}]. The Initial positions of each particle are Personally Best for each Particle.

Step 2: Initialize the weight function value w with 0.5 and weightening parameters cognitive learning factor c1, social coefficient c2 with 2.0.

Step 3: Repeat the following steps 4 to 9 until number of iterations specified by the user or Particles Exhaust.

Step 4: for i = 1,2, ……, n do // For all the Particles

For each particle position with values of tuning parameters, evaluate the fitness function. The fitness function here is Mean Absolute Relative Error (MARE). The objective in this method is to minimize the MARE by selecting appropriate values from the ranges specified in step 1.

Step 5: Here the Pbest is determined for each particle by evaluating and comparing measured effort and estimated effort values of the current and previous parameters values. If fitness (p) better than fitness (Pbest) then: Pbest = p.

Step 6: Set the best of 'Pbests' as global best – Gbest. The particle value for which the variation between the estimated and measured effort is the least is chosen as the Gbest particle.

Step 7: Update the weightening function is done by the following formula

$$W_{new} = [(T_{mi} - T_{ci}) * (W_{iv} - W_{fv})] / T_{mi} + W_{fv} \quad (7)$$

Step 8: Update the weightening factors is done with the following equations for faster convergence.

$$C_1(t) = 2.5 - 2 * (T_{ci} / T_{mi}) \quad (8)$$

$$C_2(t) = 0.5 + 2* (T_{ci} / T_{mi}), \quad (9)$$

Step 9: Update the velocity and positions of the tuning parameters with the following equations for j = 1, 2, ……m do // For number of Parameters, our case m is 2or 3 or 4

begin

$$V_{ji}^{k+1} = w * V_{ji}^{k} + c_1 * rand()_1 * (Pbest - S_{ji}^{k}) + c_2 * rand()_2 * (Gbest - S_{ji}^{k}) \quad (10)$$

$$S_{ji}^{k+1} = S_{ji}^{k} + V_{ji}^{k+1} \quad (11)$$

end;

Step 10: Give the Gbest values as the optimal solution.

Step 11: Stop

b) PROPOSED MODELS

i. *MODEL 1:*

A prefatory approach to estimating effort is to make it a function of a single variable , often this variable is project size measure in KDLOC (kilo delivered lines of code) and the equation is given as ,

$$Effort = a\ (size)^b$$

Now in our model the parameters are tuned using above PSO methodology.The Update of velocity and positions of Parameter "a" is

$$V_{ai}^{k+1} = w * V_{ai}^{k} + c_1 * rand()_1 * (Pbest - S_{ai}^{k}) + c_2 * rand()_2 * (Gbest - S_{ai}^{k}) \quad (12)$$

$$S_{ai}^{k+1} = S_{ai}^{k} + V_{ai}^{k+1}$$

The Update of velocity and positions of Parameter "b" is

$$V_{bi}^{k+1} = w * V_{bi}^{k} + c_1 * rand()_1 * (Pbest - S_{bi}^{k}) + c_2 * rand()_2 * (Gbest - S_{bi}^{k})$$

$$S_{bi}^{k+1} = S_{bi}^{k} + V_{bi}^{k+1}$$

COST FACTORS	DESCRIPTION	RATING				
		VERY LOW	LOW	NOMINAL	HIGH	VERY HIGH
	Product					
RELY	Required software reliability	0.75	0.88	1	1.15	1.4
DATA	Database size	-	0.94	1	1.08	1.16
CPLX	Product complexity	0.7	0.85	1	1.15	1.3
	Computer					
TIME	Execution time constraint	-	-	1	1.11	1.3
STOR	Main storage constraint	-	-	1	1.06	1.21
VIRT	Virtual machine volatility	-	0.87	1	1.15	1.3
TURN	Computer turnaround time	-	0.87	1	1.07	1.15
	Personnel					
ACAP	Analyst capability	1.46	1.19	1	0.86	0.71
AEXP	Application experience	1.29	1.13	1	0.91	0.82
PCAP	Programmer capability	1.42	1.17	1	0.86	0.7
VEXP	Virtual machine volatility	1.21	1.1	1	0.9	-
LEXP	Language experience	1.14	1.07	1	0.95	-
	Project					
MODP	Modern programming practice	1.24	1.1	1	0.91	0.82
TOOL	Software tools	1.24	1.1	1	0.91	0.83
SCED	Development schedule	1.23	1.08	1	1.04	1.1

Table 1 : Effort Multipliers

ii. *MODEL 2:*

Instead of having resources estimates as a function of one variable, resources estimates can depend on many different factors, giving rise to multivariable models. Such models are useful as they take into account the subtle aspects of each project such as their complexity or other such factors which usually create a non linearity. The cost factors considered are shown below. The product of all the above cost factors is the Effort Adjustment Factor (EAF).A model of this category starts with an initial estimate determined by using the strategic single variable model equations and adjusting the estimates based on other variable which is methodology.

The equation is,

$$Effort = a *(size)^b + c* (ME).$$

Where ME is the methodology used in the project. The parameters a, b, c are tuned by using above PSO methodology. The Update of velocity and positions of Parameter "a", "b" are shown in Model 1 and Parameter "c" is

$$V_{ci}^{k+1} = w * V_{ci}^{k} + c_1 * rand()_1 * (Pbest - S_{ci}^{k}) + c_2 * rand()_2 * (Gbest - S_{ci}^{k})$$

$$S_{ci}^{k+1} = S_{ci}^{k} + V_{ci}^{k+1}$$

iii. *MODEL 3*

There are a lot of factors causing uncertainty and non linearity in the input parameters. In some projects the size is low while the methodology is high and the complexity is high, for other projects size is huge but the complexity is low. As per the above two models size and effort are directly proportional. But such a condition is not always satisfied giving rise to eccentric inputs. This can be accounted for by introducing a biasing factor (d). So the effort estimation equation is:

$$Effort = a *(size)^b + c* (ME).+ d$$

a,b,c,d parameters are tuned by using above PSO methodology.

The Update of velocity and positions of Parameter "a", "b", "c" are shown in Model 1,2 and Parameter "d" is

$$V_{di}^{k+1} = w * V_{di}^{k} + c_1 * rand()_1 * (Pbest - S_{di}^{k}) + c_2 * rand()_2 * (Gbest - S_{di}^{k})$$

$$S_{di}^{k+1} = S_{di}^{k} + V_{di}^{k+1}$$

IV. Model Analysis

a) Implementation

We have implemented the above methodology for tuning parameters a,b,c and d in "C" language. For the parameter' a 'the velocities and positions of the

particles are updated by applying the following equations:

$$V_{ai}^{k+1} = w * V_{ai}^{k} + c_1 * rand_1 * (Pbesta - S_{ai}^{k}) + c_2 * rand_2 * (Gbest - S_{ai}^{k})$$

$$S_{ai}^{k+1} = S_{ai}^{k} + V_{ai}^{k+1}, \; w=0.5 \; , \; c1=c2=2.0.$$

And similarly for the parameters b,c and d the values are obtained for the first experiment and weight factor w changed during the iteration and C1 and C2 are constant. For the second experiment we changed the C1, C2 weighting factors by using equations 4 and 5.

b) Performance Measures

We consider three performance criterions:

1) Variance accounted – For(VAF)

$$\%VAF = \left[1 - \frac{\text{var (ME-EE)}}{\text{var (ME)}}\right] \times 100$$

2) Mean Absolute Relative Error

$$\%MARE = \text{mean}\left[\frac{\text{abs(ME-EE)}}{\text{(ME)}}\right] \times 100$$

3) Variance Absolute Relative Error (VARE)

$$\%VARE = \text{var}\left[\frac{\text{abs(ME-EE)}}{\text{(ME)}}\right] \times 100$$

Where ME represents Measured Effort, EE represents Estimated Effort.

V. Model Experimentation

EXPERIMENT – 1:

For the study of these models we have taken data of 10 NASA [13]

Project No	Size In KDLOC	Methodology (ME)	Measured Effort
13	2.1	28	5
10	3.1	26	7
11	4.2	19	9
17	12.5	27	23.9
3	46.5	19	79
4	54.5	20	90.8
6	67.5	29	98.4
15	78.6	35	98.7
1	90.2	30	115.8
18	100.8	34	138.3

Table 2 : NASA software projects data

By running the "C" implementation of the above methodology we obtain the following parameters for the proposed models.

Model 1: a=2.646251 and b=0.857612 .
The range of a is [1, 10] and b is [-5,5] .

Model 2: a=2.771722, b=0.847952 and c= -0.007171.
The range of a is [1, 10], b is [-5,5] and c is [-1,1].

Model 3: a =3.131606, b=0.820175, c=0.045208 and d= -2.020790.The ranges are a[1,10],b[-5,5], c[-1,1] and d[1,20]. respectively.

EXPERIMENT -2:

The following are the results obtained by running the above PSO algorithm implemented in "C" with changing weighting factors on each iteration.

Model 1: a=2.646251 and b=0.857612.
The range of a is [1,10] and b is[-5,5]

Model 2: a=1.982430, b=0.917533 and c= 0.056668.
The range of a, b, c is [1, 10], [-5, 5] and [-1, 1] respectively.

Model 3: a= 2.529550, b= h0.867292, c= -0.020757 and d=0.767248.
The ranges of a,b,c,d is [1,10] , [-5,5] , [-1,1] and [0,20] respectively.

VI. Results and Discussions

The following table shows estimated effort of our proposed model:

EXPERIMENT -1:

SIZE	MEASURED EFFORT	METHODOLOGY	ESTIMATED EFFORT OF OUR MODELS C1,C2 ARE CONSTANT DURING THE ITERATION (CASE-I)			ESTIMATED EFFORTOF OUR MODELS C1,C2 ARE CHANGED DURING THE ITERATION(CASE-II)		
			MODEL-I	MODEL-II	MODEL-III	MODEL-I	MODEL-II	MODEL-III
2.1	5	28	5.000002	4.998887	5.000007	5.000002	5.502722	5.000001
3.1	7	26	6.982786	7.047925	7.07543	6.982786	7.071439	6.975912
4.2	9	19	9.060186	9.222874	8.999259	9.060186	8.47359	9.154642
12.5	23.9	27	23.08629	23.40447	24.05549	23.08629	21.65101	22.82118
46.5	79	19	71.2293	71.75396	71.84614	71.2293	68.24138	71.03909
54.5	90.8	20	81.61792	82.10557	82.04368	81.61792	78.82941	81.44935
67.5	98.4	29	98.05368	98.39988	98.39998	98.05368	96.18965	97.79541
78.6	98.7	35	111.7296	111.9449	111.8526	111.7296	110.7037	111.4518
90.2	115.8	30	125.7302	125.8721	125.048	125.7302	125.0572	125.6834
100.8	138.3	34	138.3002	138.3003	137.2231	138.3002	138.523	138.2999

Table 3 : Estimated Efforts of Proposed Models

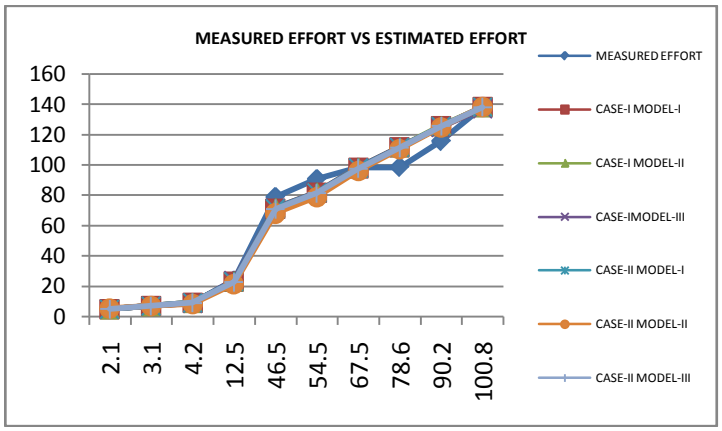


Fig 1 : Measured Effort Vs Estimated Efforts of Proposed Models

COMPARISON WITH OTHER MODELS

Refer Table 4 for the comparison with other models.

VII. Performance Analysis

Model	VAF (%)	Mean Absolute Relative Error (%)	Variance Absolute Relative Error (%)
Bailey –Basili Estimate	93.147	17.325	1.21
Alaa F. Sheta G.E. Model I Estimate	98.41	26.488	6.079
Alaa F. Sheta Model II Estimate	98.929	44.745	23.804
Harish model1	98.5	12.17	80.859
Harish model2	99.15	10.803	2.25
CASE-I MODEL -I	98.92	4.6397	0.271
CASE-I MODEL-II	98.92	4.6122	0.255
CASE-I MODEL-III	98.9	4.4373	0.282
CASE-II MODEL -I	98.92	4.6397	0.271
CASE-II MODEL-II	98.89	7.5	0.253
CASE-II MODEL-III	98.95	4.9	0.257

Table 5 : Performance Measures

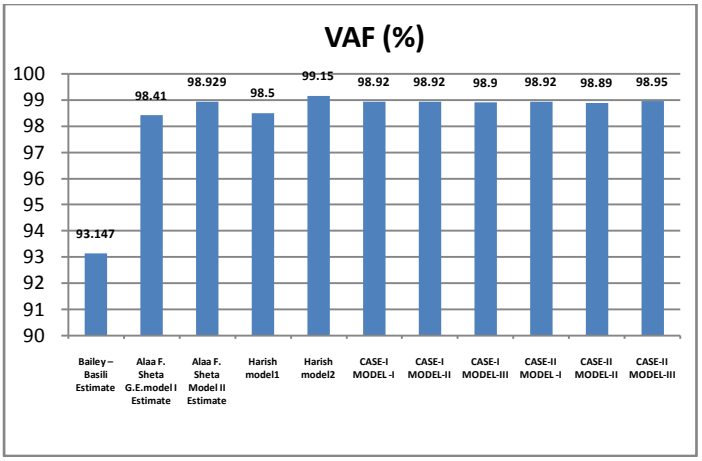


Fig 2 : Variance Accounted For %

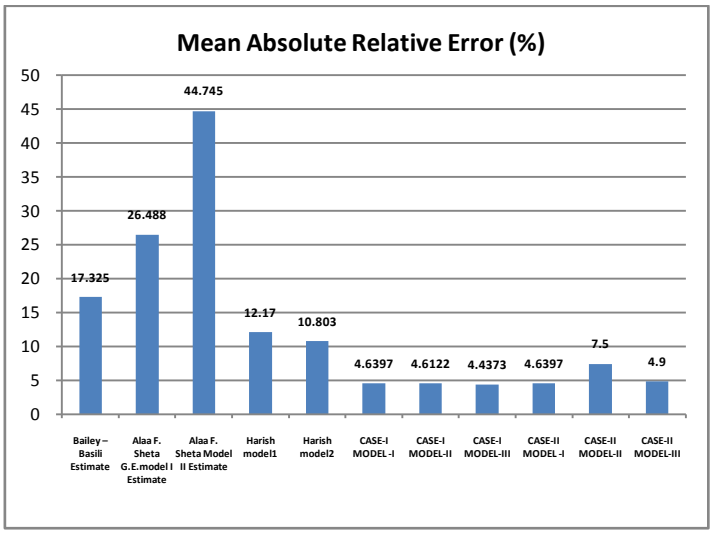


Fig 3 : Mean Absolute Relative Error (MARE)

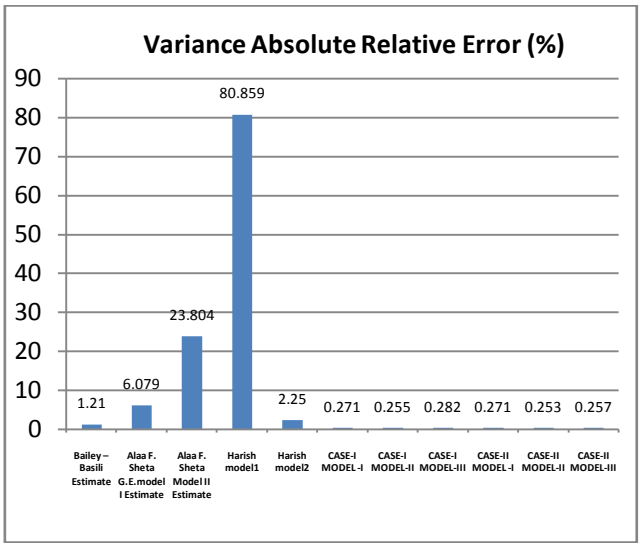


Fig 4 : Variance Absolute Relative Error %

VIII. Conclusion

Software cost estimation is based on a probabilistic model and hence it does not generate exact values. However if good historical data is provided and a systematic technique is employed we can generate better results. Accuracy of the model is measured in terms of its error rate and it is desirable to be as close to the actual values as possible. In this study we have proposed new models to estimate the software effort. In order to tune the parameters we use particle swarm optimization methodology algorithm. It is observed that PSO gives more accurate results when juxtaposed with its other counterparts. On testing the performance of the model in terms of the MARE, VARE and VAF the results were found to be futile. These techniques can be applied to other software effort models.

References References Referencias

1. D. E. Goldberg. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, 1989.
2. K. Deb. Multi-Objective Optimization Using Evolutionary Algorithms. John Wiley and Sons, 2002.
3. C.A. Coello Coello et al. Evolutionary Algorithms for Solving Multi-Objective Problems. Kluwer, 2002.
4. Robert T. F. Ah King and Harry C. S. Rughooputh, "Elitist Multi evolutionary algorithm for environmental/economic dispatch", IEEE 2003.
5. Alaa F. Sheta , "Estimation of the COCOMO Model Parameters Using Genetic Algorithms for NASA Software Projects", Journal of Computer Science 2 (2): 118-123, ISSN 1549-36362006.
6. Alaa Sheta, David Rine and Aladdin Ayesh," Development of Software Effort and Schedule Estimation Models Using Soft Computing

Techniques", 2008 IEEE Congress on Evolutionary Computation (CEC 2008), 978-1-4244-1823-7/08

7. Tad Gonsalves, Atsushi Ito, Ryo Kawabata and Kiyoshi Itoh , Swarm Intelligence in the Optimization of Software Development Project Schedule, 0730-3157/08 , 2008 IEEE.
8. J.S.Pahariya ,V. Ravi, M. Carr, Software Cost Estimation using Computational Intelligence Techniques,2009 World Congress on Nature & Biologically Inspired Computing (NaBIC 2009).
9. Parvinder S. Sandhu, Porush Bassi, and Amanpreet Singh Brar ,Software Effort Estimation Using Soft Computing Techniques, World Academy of Science, Engineering and Technology 46 2008.
10. Iman Attarzadeh and Siew Hock Ow, Soft Computing Approach for Software Cost Estimation, International Journal of Software Engineering, IJSE Vol.3 No.1 January 2010.
11. Xishi Huang, Danny Ho, Jing Ren , Luiz F. Capretz,, Improving the COCOMO model using a neuro-fuzzy approach, doi:10.1016/j.asoc.2005.06.007 ,2005 Elsevier.
12. Alaa Sheta, David Rine and Aladdin Ayesh, Development of Software Effort and Schedule Estimation Models Using Soft Computing Techniques, 978-1-4244-1823-7/08,2008 IEEE.
13. John w. Bailey and victor R.Basili,(1981) "A meta model for software development resource expenditures", Fifth International conference on software Engineering, CH-1627-9/81/0000/ 0107500. 75@ 1981 IEEE, PP 107-129,1981.

Table 4 : Measured Efforts of Various Models

Measured effort	Bailey–Basili Estimate	Alaa F. hetaG. E.Model Estimate	Alaa F. Sheta Model 2 Estimate	Harish model1	Harish model2	CASE-I MODEL-I	CASE-I MODEL-II	CASE-I MODEL-III	CASE-II MODEL-I	CASE-II MODEL-II	CASE-II MODEL-III
5	7.226	8.44	11.271	6.357	4.257	5.000002	4.998887	5.000007	5.000002	5.502722	5.000001
7	8.212	11.22	14.457	8.664	7.664	6.982786	7.047925	7.07543	6.982786	7.071439	6.975912
9	9.357	14.01	19.976	11.03	13.88	9.060186	9.222874	8.999259	9.060186	8.47359	9.154642
23.9	19.16	31.098	31.686	26.252	24.702	23.08629	23.40447	24.05549	23.08629	21.65101	22.82118
79	68.243	81.257	85.007	74.602	77.452	71.2293	71.75396	71.84614	71.2293	68.24138	71.03909
90.8	80.929	91.257	94.977	84.638	86.938	81.61792	82.10557	82.04368	81.61792	78.82941	81.44935
98.4	102.175	106.707	107.254	100.329	97.679	98.05368	98.39988	98.39998	98.05368	96.18965	97.79541
98.7	120.848	119.27	118.03	113.237	107.288	111.7296	111.9449	111.8526	111.7296	110.7037	111.4518
115.8	140.82	131.898	134.011	126.334	123.134	125.7302	125.8721	125.048	125.7302	125.0572	125.6834
138.3	159.434	143.0604	144.448	138.001	132.601	138.3002	138.3003	137.2231	138.3002	138.523	138.2999

Global Journal of Computer Science and Technology
Volume 11 Issue 18 Version 1.0 October 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Studies on Colour Image Segmentation Method Based on Finite Left Truncated Bivariate Gaussian Mixture Model with K-Means

By G.V.S. Rajkumar, K.Srinivasa Rao, P.Srinivasa Rao
GITAM University

Abstract - Colour Image segmentation is one of the prime requisites for computer vision and analysis. Much work has been reported in literature regarding colour image segmentation under HSI colour space and Gaussian mixture model (GMM). Since the Hue and Saturation values of the pixel in the image are non-negative. And may not be meso-kurtic, it is needed left truncate the Gaussian variate and is used to represent these two features of the colour image. The effect of truncation can not be ignored in developing the model based colour image segmentation. Hence in this paper a left truncated bivariate Gaussian mixture model is utilized to segment the colour image. The correlation between Hue and Saturation plays a predominant role in segmenting the colour images which is observed through experimental results. The expectation maximization (EM) algorithm is used for estimating model parameters. The number of image segments can be initialization of the model parameters are done with K-means algorithm. The performance of the proposed algorithm is studied by calculating the segmentation performance techniques like probabilistic rand index (PRI), global consistency error (GCE) and variation of information (VOI). The utility of the estimated joint probability density function of feature vector of the image is demonstrated through image retrievals. The image quality measures obtained for six images taken from Berkeley image dataset reveals that the proposed algorithm outperforms the existing algorithms in image segmentation and retrievals.

Keywords : Image Segmentation, Hue , Saturation, Finite Left Truncated Bivariate Gaussian distribution, K-means algorithm, Image Quality Metrics, EM- algorithm.

GJCST-H Classification : I.4.6

Strictly as per the compliance and regulations of:

Studies on Colour Image Segmentation Technique Based on Finite Left Truncated Bivariate Gaussian Mixture Model with K-Means

G.V.S. Rajkumar [α], K.Srinivasa Rao [Ω], P.Srinivasa Rao [β]

Abstract - Colour Image segmentation is one of the prime requisites for computer vision and analysis. Much work has been reported in literature regarding colour image segmentation under HSI colour space and Gaussian mixture model (GMM). Since the Hue and Saturation values of the pixel in the image are non-negative. And may not be meso-kurtic, it is needed left truncate the Gaussian variate and is used to represent these two features of the colour image. The effect of truncation can not be ignored in developing the model based colour image segmentation. Hence in this paper a left truncated bivariate Gaussian mixture model is utilized to segment the colour image. The correlation between Hue and Saturation plays a predominant role in segmenting the colour images which is observed through experimental results. The expectation maximization (EM) algorithm is used for estimating model parameters. The number of image segments can be initialization of the model parameters are done with K-means algorithm. The performance of the proposed algorithm is studied by calculating the segmentation performance techniques like probabilistic rand index (PRI), global consistency error (GCE) and variation of information (VOI). The utility of the estimated joint probability density function of feature vector of the image is demonstrated through image retrievals. The image quality measures obtained for six images taken from Berkeley image dataset reveals that the proposed algorithm outperforms the existing algorithms in image segmentation and retrievals.

Keywords : Image Segmentation, Hue , Saturation, Finite Left Truncated Bivariate Gaussian distribution, K-means algorithm, Image Quality Metrics, EM- algorithm.

I. Introduction

Image segmentation is a process of extracting useful information from the images through features and dividing the whole image into various homogeneous groups in which, the pixels within the group are more homogeneous and are heterogeneous between the groups. It is an important technology for image processing and understanding. The structural characteristics of objects and surfaces in an image can be determined by segmenting the image using image domain properties. One of the major advantages of image segmentation is denoising. Denoising is the process of removing unwanted noise from the image. Segmentation specifically attempts to separate structure from noise on a local scale. It is one of the most important steps in computer vision and analysis.

For the last three decades lot of work has been reported in literature regarding image segmentation methods (Lucchese L. et al (2001), Srinivas Y. and Srinivas Rao K. (2007), Majid Fakheri et al (2010), Siddhartha Bhattacharyya (2011)). The image segmentation methods can be divided into two categories depending upon the type of image. The images can be broadly categorized into two types namely, gray level images and colour images. A gray level image is usually characterized by pixel intensity (Farag A..A.. et al (2004), Seshashayee M. et al (2011), Srinivas Yerramalle et al (2010)). But in colour images the colour is a perceptual phenomenon related to human response to different wavelengths in the visible electro-magnetic spectrum. In colour images the features that represent the image pixel are highly influenced by three feature descriptions namely, intensity, colour and texture. Among these features colour is the most important one in segmenting the colour images since intensity and texture features also be embedded in colour features. (Fesharaki and Hellestrand (1992), Kato Z. et al (2006), Kang Feng et al (2009), Kaikuo Xu et al (2011)). A better colour space than the RGB space in representing the colours of human perception is the HSI space, in which the colour information is represented by Hue and Saturation values. Thus the human perception of image can be characterized through a bivariate random variable consisting of Hue and Saturation which can be measured using generic structure of a colour appearance model (Sangwine et al (1998)).

Ferri and Vidal (1992), Lee E. et al(2010), Dipti P. and Mridula J. (2011) and others have reviewed colour image segmentation techniques. Among these

Author [α] : Department of Information Technology, GITAM University, Visakhapatnam, Andhra Pradesh, INDIA,
Telephone: +91-9989888307, E-mail : gvsrajkumar@gmail.com
Author [Ω] : Department of Statistics, Andhra University, Visakhapatnam, Andhra Pradesh, INDIA, E-mail : ksraoau@yahoo.co.in
Author [β] : Department of Computer Science and Systems Engineering, Andhra University,Visakhapatnam,AndhraPradesh, INDIA,
E-mail : peri.srinivasarao@yahoo.com

model based image segmentation methods are more efficient than the edge based or threshold or region based methods (Lucchese L. et al (2001)). In model based image segmentation the whole image is divided into different image regions and each image region is characterized by a suitable probability distribution. For ascribing a probability model to the feature vector of the pixels in the image region, it is needed to study the statistical characteristics of the feature vector.

In image segmentation it is customary to consider that the whole image is characterized by a finite Gaussian mixture model. That is, the feature vector of each image region follows a Gaussian distribution (Haralick and Shapiro (1985), Shital Raut et.al (2009), Kato Z. et al (2006), Mantas Paulinas and Audrius Usinskas (2007), Rahman Farnoosh et al (2008), Sujaritha M. and Annadurai S. (2010)). The image segmentation methods based on Gaussian mixture model work well only when the feature vector of the pixels are having infinite range and the distribution of the feature vector is symmetric and meso-kurtic. But in many colour images the feature vector represented by Hue and Saturation are having finite values (say nonnegative) and may not be mesokurtic and symmetric. Hence, to have an accurate image segmentation of these sorts of colour images it is needed to develop and analyze image segmentation methods based on truncated bivariate mixture distributions.

Here, it is assumed that the feature vector in different image regions follows a left truncated bivariate Gaussian distribution and the feature vector of the whole image is characterized by a finite left truncated bivariate Gaussian mixture model. This assumption is made since the Hue and Saturation values of the pixel which represents the bivariate feature vector can take nonnegative values only. Hence, the range of the Hue and Saturation values are to be left truncated at zero. The effect of the truncated nature of Hue and Saturation cannot be ignored, since the leftover probability is significantly higher than zero in the left tail end of the distribution. This left truncated nature of the bivariate feature vector can approximate the pixels of the colour image more close to the reality.

In this method of segmentation, the number of image regions is obtained by *K*-means algorithm for which the initial value of the number of components is identified from the number of peaks in the image histogram. The model parameters are estimated by using Expectation Maximization (EM) algorithm. The EM-algorithm is one of the most preferred method of estimating the model parameters in mixture distributions (Mclanchlan G. and Krishnan T. (1997)). The EM-algorithm requires the updated equations of the model parameters which are derived for the left truncated bivariate Gaussian mixture model. The initialization of the model parameters for carrying the EM-algorithm is done through feature vector of the pixel intensities of the image regions obtained through *K*-means clustering and moment method of estimation. An image segmentation algorithm with component likelihood maximization under Bayesian frame work is also developed and analyzed.

The efficiency of the developed image segmentation algorithm is studied by conducting experimentation with six images namely, OSTRICH, POT, TOWER, BEARS, DEER and BIRD which are taken randomly from Berkeley image data set. The segmentation performance measures namely, probabilistic rand index (PRI), global consistency error (GCE) and variation of information (VOI) are computed for the six images and presented. A comparative study of these measures with those obtained from the finite Gaussian mixture model reveals that this algorithm performs better than the Gaussian mixture model with *K*-means and having clear boundaries.

Using the estimated joint probability density functions of the feature vector of pixels of each image, the images are retrieved. The efficiency of the developed algorithm in image retrieval is also studied by computing the image quality metrics like maximum distance, image fidelity, mean square error, signal to noise ratio and image quality index and the results are presented. A comparative study of these quality measures with those obtained from the Gaussian mixture model with *K*-means revealed that this algorithm performs better.

II. Finite Left Truncated Bivariate Gaussian Mixture Model

The effect of truncation in bivariate Gaussian distribution has been discussed by several researchers (Norman L.Johnson, Samuel Kotz and Balakrishnan (1994)). The probability density function of the left truncated Gaussian distribution (truncated at zero) is,

$$g(x,y;\theta)=\frac{f(x,y)}{\int_0^\infty\int_0^\infty f(x,y)\,dxdy},\ 0<x<\infty;\ 0<y<\infty \qquad (1)$$

Where, zero is the truncation point for both the Hue and saturation, $f(x,y)$ is the probability density function of the bivariate Normal distribution is

$$f(x,y)=\frac{1}{2\pi\sqrt{1-\rho^2}\sigma_1\sigma_2}\exp\left\{\frac{-1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2-2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right)+\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

$$-\infty<x<+\infty;\ -\infty<y<+\infty,$$
$$\sigma_1>0;\ \sigma_2>0;\ -1<\rho<1,$$
$$-\infty<\mu_1<+\infty;\ -\infty<\mu_2<+\infty \qquad (2)$$

The value of $\left[1-\int_0^\infty\int_0^\infty f(x,y)\,dxdy\right]$ is significant based on the values of the parameters. This distribution includes the skewed, asymmetric bivariate distributions as particular cases for limiting and specific values of the parameters. The various shapes of the frequency curves of the left truncated bivariate Gaussian distribution are shown in Figure1.

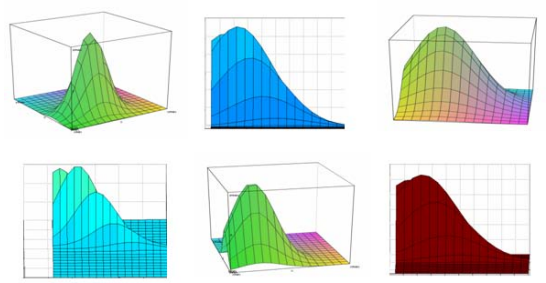


Fig1 : Shapes of left truncated bivariate Gaussian frequency surfaces

Following the heuristic arguments given by Bengt Muthen (1990), the mean value of 'X'(hue) is obtained as

$$E(X) = \mu_1 + \sigma_1 A \qquad (3)$$

Where,

$$A = \phi\left(\frac{-\mu_1}{\sigma_1}\right)\left[1-\Phi\left[\left(\left(\frac{-\mu_1}{\sigma_1}\right)-\rho\left(\frac{-\mu_1}{\sigma_1}\right)\right)c\right]\right]+\rho\phi\left(\frac{-\mu_2}{\sigma_2}\right)\left[1-\Phi\left[\left(\left(\frac{-\mu_1}{\sigma_1}\right)-\rho\left(\frac{-\mu_2}{\sigma_2}\right)\right)c\right]\right]$$

and $c = \left(1-\rho^2\right)^{-1/2}$, ϕ, Φ are the ordinate and area of standard Normal distribution. Similarly the mean value 'Y'(saturation) is

$$E(Y) = \mu_2 + \sigma_2 B \qquad (4)$$

Where,

$$B = \phi\left(\frac{-\mu_2}{\sigma_2}\right)\left[1-\Phi\left[\left(\left(\frac{-\mu_2}{\sigma_2}\right)-\rho\left(\frac{-\mu_2}{\sigma_2}\right)\right)c\right]\right]+\rho\phi\left(\frac{-\mu_1}{\sigma_1}\right)\left[1-\Phi\left[\left(\left(\frac{-\mu_2}{\sigma_2}\right)-\rho\left(\frac{-\mu_1}{\sigma_1}\right)\right)c\right]\right]$$

and c is as given in equation (3)

The Variance of X is

$$V(X) = \sigma_1^2 R - 2A\,\sigma_1\,A + A^2$$

$$= \sigma_1^2 R - A^2(2\sigma_1 - 1) \qquad (5)$$

Where,

$$R = \left[\pi+\left(\frac{-\mu_1}{\sigma_1}\right)\phi\left(\frac{-\mu_1}{\sigma_1}\right)\left[1-\Phi\left[\left(\left(\frac{-\mu_1}{\sigma_1}\right)-\rho\left(\frac{-\mu_1}{\sigma_1}\right)\right)c\right]\right]+\rho^2\left(\frac{-\mu_1}{\sigma_1}\right)\phi\left(\frac{-\mu_1}{\sigma_1}\right)\left[1-\Phi\left[\left(\left(\frac{-\mu_1}{\sigma_1}\right)-\rho\left(\frac{-\mu_1}{\sigma_1}\right)\right)c\right]\right]\right.$$
$$\left.-c^{-1}\rho\phi\left(\frac{-\mu_1}{\sigma_1}\right)\left[-\phi\left[\left(\left(\frac{-\mu_1}{\sigma_1}\right)-\rho\left(\frac{-\mu_1}{\sigma_1}\right)\right)c\right]\right]\right],$$

and c and A is given in equation (3). The Variance of Y is

$$V(Y) = \sigma_2^2 T - 2B\,\sigma_2\,B + B^2$$

$$= \sigma_2^2 T - B^2(2\sigma_2 - 1) \qquad (6)$$

Where,

$$T = \left[\pi+\left(\frac{-\mu_2}{\sigma_2}\right)\phi\left(\frac{-\mu_2}{\sigma_2}\right)\left[1-\Phi\left[\left(\left(\frac{-\mu_2}{\sigma_2}\right)-\rho\left(\frac{-\mu_2}{\sigma_2}\right)\right)c\right]\right]+\rho^2\left(\frac{-\mu_2}{\sigma_2}\right)\phi\left(\frac{-\mu_2}{\sigma_2}\right)\left[1-\Phi\left[\left(\left(\frac{-\mu_2}{\sigma_2}\right)-\rho\left(\frac{-\mu_2}{\sigma_2}\right)\right)c\right]\right]\right.$$
$$\left.-c^{-1}\rho\phi\left(\frac{-\mu_2}{\sigma_2}\right)\left[-\phi\left[\left(\left(\frac{-\mu_2}{\sigma_2}\right)-\rho\left(\frac{-\mu_2}{\sigma_2}\right)\right)c\right]\right]\right],$$

c and B are as given in equations (3) and (4) respectively. The Covariance of (X, Y) is

$$COV(X, Y) = \sigma_1\sigma_2 U - AB(\sigma_1 + \sigma_2 - 1) \qquad (7)$$

where,

$$U = \left[\rho\pi+\rho\left(\frac{-\mu_1}{\sigma_1}\right)\phi\left(\frac{-\mu_1}{\sigma_1}\right)\left[1-\Phi\left[\left(\left(\frac{-\mu_2}{\sigma_2}\right)-\rho\left(\frac{-\mu_1}{\sigma_1}\right)\right)c\right]\right]-c^{-1}\phi\left(\frac{-\mu_1}{\sigma_1}\right)\left[-\phi\left[\left(\left(\frac{-\mu_2}{\sigma_2}\right)-\rho\left(\frac{-\mu_1}{\sigma_1}\right)\right)c\right]\right]\right.$$
$$\left.+\rho\left(\frac{-\mu_2}{\sigma_2}\right)\phi\left(\frac{-\mu_2}{\sigma_2}\right)\left[1-\Phi\left[\left(\left(\frac{-\mu_1}{\sigma_1}\right)-\rho\left(\frac{-\mu_2}{\sigma_2}\right)\right)c\right]\right]\right],$$

c, A and B are as given in equations (3) and (4) respectively.

Since the entire image is a collection of regions, which are characterized by left truncated bivariate normal distribution, it can be characterized through a *K*-Component finite left truncated bivariate Gaussian distribution and its probability density function is of the form

$$h(x,y) = \sum_{i=1}^{K}\alpha_i g_i(x_i, y_i; \theta) \qquad (8)$$

Where, *K* is the number of regions, $\alpha_i > 0$ are weights such that $\sum_{i=1}^{K}\alpha_i = 1$ and $\theta = \left\{\mu_{1i}, \mu_{2i}, \sigma_{1i}^2, \sigma_{2i}^2, \rho_i\right\}$ is the set of parameters. $g_i(x_i, y_i / \theta_i)$ given in equation (1) represent the probability density function of the i^{th} image region. α_i is the probability of occurrence of the i^{th} component of the finite left truncated bivariate Gaussian mixture model (FLTBGMM) i.e., the probability that the feature belongs to the i^{th} image region.

The mean vector representing the entire image is

$$E(W^T) = \begin{bmatrix}\sum_{i=1}^{K}\alpha_i E_i(X)\\ \sum_{i=1}^{K}\alpha_i E_i(Y)\end{bmatrix} \qquad (9)$$

Where, $E(X_i)$ and $E(Y_i)$ are given in equations (3) and (4) for the i^{th} image region.

III. Estimation of the Model Parameters by EM-Algorithm

To obtain the estimation of the model parameters, we utilized the EM-algorithm by maximizing the expected likelihood function for carrying out the EM-algorithm. The likelihood function of bivariate observations $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_N, y_N)$ drawn from an image with probability density function

$$L(\theta) = \prod_{s=1}^{N} h(x_s, y_s; \theta)$$

$$= \prod_{s=1}^{N} \left(\sum_{i=1}^{K} \alpha_i g_i(x_s, y_s; \theta) \right) \tag{10}$$

$$= \prod_{s=1}^{N} \left(\sum_{i=1}^{K} \alpha_i \frac{\exp\left\{ \frac{-1}{2(1-\rho_i^2)} \left[\left(\frac{x_s - \mu_{1i}}{\sigma_{1i}} \right)^2 - 2\rho_i \left(\frac{x_s - \mu_{1i}}{\sigma_{1i}} \right) \left(\frac{y_s - \mu_{2i}}{\sigma_{2i}} \right) + \left(\frac{y_s - \mu_{2i}}{\sigma_{2i}} \right)^2 \right] \right\}}{2\pi \sqrt{1-\rho_i^2}\, \sigma_{1i} \sigma_{2i} \int_0^\infty \int_0^\infty f_i(x, y; \theta)\, dxdy} \right)$$

This implies

$$\log L(\theta) = \sum_{s=1}^{N} \log \left(\sum_{i=1}^{K} \alpha_i g_i(x_s, y_s; \theta) \right) \tag{11}$$

The updated equations of EM-algorithm for estimating the model parameters are

$$\alpha_k^{(l+1)} = \frac{1}{N} \sum_{s=1}^{N} \left[t_k(x_s, y_s; \theta^{(l)}) \right]$$

$$= \frac{1}{N} \sum_{s=1}^{N} \left(\frac{\alpha_k^{(l)} g_k(x_s, y_s; \theta^{(l)})}{\sum_{i=1}^{K} \alpha_k^{(l)} g_i(x_s, y_s; \theta^{(l)})} \right) \tag{12}$$

Where, $g_k(x_s, y_s; \theta^{(l)})$ is as given in equation (1).

For updating μ_{1k} we have,

$$\mu_{1k}^{(l+1)} \sum_{s=1}^{N} t_k(x_s, y_s; \theta^{(l)}) - \sum_{s=1}^{N} t_k(x_s, y_s; \theta^{(l)}) x_s + \sum_{s=1}^{N} t_k(x_s, y_s; \theta^{(l)}) \sigma_{1k}^{(l)} \left[\frac{\rho_k^{(l)} (y_s - \mu_{2k}^{(l)})}{\sigma_{2k}^{(l)}} + \left[A - \rho_k^{(l)} B \right] \right] = 0 \tag{13}$$

Similarly for updating μ_{2k}, we have ,

$$\mu_{2k}^{(l+1)} \sum_{s=1}^{N} t_k(x_s, y_s; \theta^{(l)}) - \sum_{s=1}^{N} t_k(x_s, y_s; \theta^{(l)}) y_s + \sum_{s=1}^{N} t_k(x_s, y_s; \theta^{(l)}) \sigma_{2k}^{(l)} \left[\frac{\rho_k^{(l)} (x_s - \mu_{1k}^{(l)})}{\sigma_{1k}^{(l)}} + \left[B - \rho_k^{(l)} A \right] \right] = 0 \tag{14}$$

Where,

$$t_k(x_s, y_s; \theta^{(l)}) = \frac{\alpha_k^{(l)} g_k(x_s, y_s; \theta^{(l)})}{h(x_s, y_s; \theta^{(l)})} = \frac{\alpha_k^{(l)} g_k(x_s, y_s; \theta^{(l)})}{\sum_{i=1}^{K} \alpha_i^{(l)} g_i(x_s, y_s; \theta^{(l)})} \quad ,$$

$$g_k(x_s, y_s; \theta^{(l)}) = \frac{\exp\left\{ \frac{-1}{2(1-\rho_k^2)} \left[\left(\frac{x_s - \mu_{1k}}{\sigma_{1k}} \right)^2 - 2\rho_k \left(\frac{x_s - \mu_{1k}}{\sigma_{1k}} \right) \left(\frac{y_s - \mu_{2k}}{\sigma_{2k}} \right) + \left(\frac{y_s - \mu_{2k}}{\sigma_{2k}} \right)^2 \right] \right\}}{2\pi \sigma_{1k} \sigma_{2k} \sqrt{1-\rho_k^2} \int_0^\infty \int_0^\infty f_k(x, y; \theta)\, dxdy}$$

A and B are as given in equations (3) and (4) respectively. The updated equations for σ_{1k}^2 at $(l+1)^{th}$ iteration is,

$$\sum_{s=1}^{N} t_k(x_s, y_s; \theta^{(l)}) \left[\left[\left(\frac{x_s - \mu_{1k}^{(l)}}{\sigma_{1k}^{(l+1)}} \right)^2 - \frac{\rho_k (x_s - \mu_{1k}^{(l)})(y_s - \mu_{2k}^{(l)})}{\sigma_{1k}^{(l+1)} \sigma_{2k}^{(l)}} \right] - D + \rho_k^{(l)} E \right] = 0 \quad (15)$$

Where, $t_k(x_s, y_s; \theta^{(l)})$ is given in equation (14),

$$D = \pi \sigma_{1k} \sigma_{2k} + \sigma_{1k} \sigma_{2k} c^{-1} \rho_k \left[\phi\left(\frac{-\mu_{1k}}{\sigma_{1k}}\right) \left[\phi\left[\left(\left(\frac{-\mu_{1k}}{\sigma_{1k}}\right) - \rho_k \left(\frac{-\mu_{1k}}{\sigma_{1k}}\right) \right) c \right] \right] \right]$$

$$+ \sigma_{2k} \left(\rho_k^{\,2} + 1 \right) (-\mu_{1k}) \phi\left(\frac{-\mu_{1k}}{\sigma_{1k}}\right) \left[1 - \Phi \left[\left(\left(\frac{-\mu_{1k}}{\sigma_{1k}}\right) - \rho_k \left(\frac{-\mu_{1k}}{\sigma_{1k}}\right) \right) c \right] \right],$$

And

$$E = \rho_k \pi \sigma_{1k} \sigma_{2k} + \sigma_{1k} \sigma_{2k} c^{-1} \left[-\phi\left(\frac{-\mu_{1k}}{\sigma_{1k}}\right) \left[-\phi \left[\left(\left(\frac{-\mu_{2k}}{\sigma_{2k}}\right) - \rho_k \left(\frac{-\mu_{1k}}{\sigma_{1k}}\right) \right) c \right] \right] \right]$$

$$+ \rho_k \sigma_{1k} \left[\left(\frac{-}{\mu_{2k}}\right) \phi\left(\frac{-\mu_{2k}}{\sigma_{2k}}\right) \left[1 - \Phi \left[\left(\left(\frac{-\mu_{1k}}{\sigma_{1k}}\right) - \rho_k \left(\frac{-\mu_{2k}}{\sigma_{2k}}\right) \right) c \right] \right] \right]$$

$$+ \rho_k \sigma_{2k} \left[(-\mu_{1k}) \phi\left(\frac{-\mu_{1k}}{\sigma_{1k}}\right) \left[1 - \Phi \left[\left(\left(\frac{-\mu_{2k}}{\sigma_{2k}}\right) - \rho_k \left(\frac{-\mu_{1k}}{\sigma_{1k}}\right) \right) c \right] \right] \right]$$

The updated equations for σ_{2k}^2 at $(l+1)^{\text{th}}$ iteration is

$$\sum_{s=1}^{N} t_k(x_s, y_s; \theta^{(l)}) \left[\left[\left(\frac{y_s - \mu_{1k}^{(l)}}{\sigma_{2k}^{(l+1)}} \right)^2 - \frac{\rho_k (x_s - \mu_{1k}^{(l)})(y_s - \mu_{2k}^{(l)})}{\sigma_{1k}^{(l)} \sigma_{2k}^{(l+1)}} \right] - G + \rho_k^{(l)} E \right] = 0 \quad (16)$$

where,

$$G = \pi \sigma_{1k} \sigma_{2k} + \sigma_{1k} \sigma_{2k} c^{-1} \rho_k \left[-\phi\left(\frac{-\mu_{2k}}{\sigma_{2k}}\right) \left[-\phi \left[\left(\left(\frac{-\mu_{2k}}{\sigma_{2k}}\right) - \rho_k \left(\frac{-\mu_{2k}}{\sigma_{2k}}\right) \right) c \right] \right] \right]$$

$$+ \sigma_{1k} \left(\rho_k^2 + 1 \right) (-\mu_{2k}) \phi\left(\frac{-\mu_{2k}}{\sigma_{2k}}\right) \left[1 - \Phi \left[\left(\left(\frac{-\mu_{2k}}{\sigma_{2k}}\right) - \rho_k \left(\frac{-\mu_{2k}}{\sigma_{2k}}\right) \right) c \right] \right],$$

$t_k(x_s, y_s; \theta^{(l)})$ and E are as given in equations (14) and (15) respectively and

Therefore the updated equation for estimating ρ_k is

$$\sum_{s=1}^{N} t_k(x_s, y_s; \theta^{(l)}) \left[-\frac{\rho_k}{\left(1-\rho_k^2\right)^2} \left[\left(\frac{x_s - \mu_{1k}}{\sigma_{1k}} \right)^2 + \left(\frac{y_s - \mu_{2k}}{\sigma_{2k}} \right)^2 \right] \right. \quad (17)$$

$$\left. - \frac{1+\rho_k^2}{\left(1-\rho_k^2\right)^2} \left[\left(\frac{x_s - \mu_{1k}}{\sigma_{1k}} \right) \left(\frac{y_s - \mu_{2k}}{\sigma_{2k}} \right) \right] + \frac{\rho_k (D+F)}{\left(1-\rho_k^2\right)^2} + \frac{(1+\rho_k^2) E}{\left(1-\rho_k^2\right)^2} \right] = 0$$

Where, $t_k(x_s, y_s; \theta^{(l)})$, D, E and G are as given in equations (14) ,(15) and (16) respectively and

$$F = \pi\sigma_{1k}\sigma_{2k} - \sigma_{1k}\left[(\mu_{2k})\ \phi(\frac{-\mu_{2k}}{\sigma_{2k}})\left[1-\Phi\left[\left((\frac{-\mu_{2k}}{\sigma_{2k}}) - \rho_k(\frac{-\mu_{2k}}{\sigma_{2k}})\right)c\right]\right]\right]$$

$$+\sigma_{1k}\rho_k{}^2\left[(-\mu_{2k})\phi(\frac{-\mu_{2k}}{\sigma_{2k}})\left[1-\Phi\left[\left((\frac{-\mu_{2k}}{\sigma_{2k}}) - \rho_k(\frac{-\mu_{2k}}{\sigma_{2k}})\right)c\right]\right]\right]$$

$$+\sigma_{1k}\sigma_{2k}\rho_k c^{-1}\left[-\phi(\frac{-\mu_{2k}}{\sigma_{2k}})\left[-\phi\left[\left((\frac{-\mu_{2k}}{\sigma_{2k}}) - \rho_k(\frac{-\mu_{2k}}{\sigma_{2k}})\right)c\right]\right]\right]$$

Solving equations (12), (13), (14), (15), (16) and (17) iteratively we get the revised estimates α_k μ_{1k}, μ_{2k}, σ_{1k}^2, σ_{2k}^2 and ρ_k.

IV. Initialization of Parameters by K-Means

The efficiency of the EM-algorithm in estimating the parameters is heavily dependent on the number of regions in the image. The number of mixture components taken for *K*–means algorithm is obtained, by plotting the histogram of the pixel intensities of the whole image. The mixing parameter α_k and the region parameters μ_{1k}, μ_{2k}, σ_{1k}^2, σ_{2k}^2, ρ_k are unknown as prior. A commonly used method in initializing parameters is by drawing a random sample in the entire image data (Mclanchan G. and Peel D. (2000)). This method performs well, if the sample size is large, and the computation time is heavily increased. When the sample size is small, some small regions may not be sampled. To overcome this problem, we use *K*-means algorithm to divide the whole image into various homogeneous regions. In *K*-means algorithm the centroids of the clusters are recomputed as soon as pixel joins the cluster.

The initial values of α_i can be taken as $\alpha_i = \frac{1}{K}$, where, *K* is the number of image regions obtained from the *K*-means algorithm (Rose H. Turi (2001)). *K*-means algorithm uses an iterative procedure that minimizes the sum of distances from each object to its cluster centroid, over all clusters. This procedure consists of the following steps.

1) Randomly choose *K* data points from the whole dataset as initial clusters. These data points represent initial cluster centroids.
2) Calculate Euclidean distance of each data point from each cluster centre and assign the data points to its nearest cluster centre.
3) Calculate new cluster centre so that squared error distance of each cluster should be minimum.
4) Repeat step 2 and 3 until clustering centers do not change.
5) Stop the process.

In the above algorithm, the cluster centers are only updated once all points have been allocated to their closed cluster centre. The advantage of *K* -means are that it is a very simple method, and it is based on intuition about the nature of a cluster, which is that the within cluster error should be as small as possible. The disadvantage of this method is that the number of clusters must be supplied as a parameter, leading to the user having to decide what the best number of clusters for the image is (Rose H.Turi, (2001)) .Success of *K*-means algorithm depends on the parameter *K*, number of clusters in image. After determining the final values of *K* (number of regions), we obtain the initial estimates of the parameters μ_{1k}, μ_{2k}, σ_{1k}^2, σ_{2k}^2, ρ_k and α_k for each image region and with the method of moments given by Bengt Muthen (1990) for Truncated Bivariate Normal Distribution with initial parameters as $\alpha_i = 1/K$ for i=1,2,…,*K*

μ_{1k} = $\bar{x}_{1k}$ is the k^{th} region sample mean of the Hue angle.

μ_{2k} = $\bar{y}_{1k}$ is the k^{th} region sample mean of the Saturation.

σ_{1k} = s_{1k} (Sample Standard Deviation of the k^{th} segment of Hue - angle)

σ_{2k} = s_{2k} (Sample Standard Deviation of the k^{th} segment of - Saturation)

ρ_k is the correlation coefficient between Hue and Saturation of the k^{th} image region.

Substituting these values as the initial estimates, we obtain the refined estimates of the parameters by using the EM-algorithm.

V. Segmentation Algorithm

After refining the parameters the prime step is image segmentation, by allocating the pixels to the segments. This operation is performed by segmentation algorithm. The image segmentation algorithm consists of four steps

Step 1) Plot the histogram of the whole image.

Step2) Obtain the initial estimates of the model parameters using *K*-Means algorithm and moment estimators as discussed in section IV.

Step3) Obtain the refined estimates of the model parameters μ_{1k}, μ_{2k}, σ_{1k}^2, σ_{2k}^2, ρ_k and α_k for i= 1,2,…,*K* using the EM-algorithm with the updated equations given in section III.

Step4) Assign each pixel into the corresponding j[th] region (segment) according to the maximum likelihood of the j[th] component L_j.

That is

$$L_j = \max_{j \in k} \left\{ \frac{\exp\left\{ \frac{-1}{2(1-\rho_k^2)} \left[\left(\frac{x_s - \mu_{1k}}{\sigma_{1k}} \right)^2 - 2\rho_k \left(\frac{x_s - \mu_{1k}}{\sigma_{1k}} \right) \left(\frac{y_s - \mu_{2k}}{\sigma_{2k}} \right) + \left(\frac{y_s - \mu_{2k}}{\sigma_{2k}} \right)^2 \right] \right\}}{2\pi\sigma_{1k}\sigma_{2k}\sqrt{1-\rho_k^2} \int_0^\infty \int_0^\infty f_k(x,y,\theta)\,dxdy} \right\}$$

VI. Experimental Results

To demonstrate the utility of the image segmentation algorithm developed in this paper, an experiment is conducted with six images taken from Berkeley images dataset (http://www.eecs.berkeley.edu /Research/Projects/CS/vision/bsds/BSDS300/html/dataset/images.html.). The images namely, OSTRICH, POT, TOWER, BEARS, DEER and BIRD are considered for image segmentation. The feature vectors of the whole image is taken as input for image segmentation. The feature vector of the image are assumed to follow a mixture of left truncated bivariate Gaussian distribution. That is, the image contains K regions and the feature vector of the each image region follow a left truncated bivariate Gaussian distribution with different parameters. The number of segments in each of the six images considered for experimentation is determined by the histogram of pixel intensities. The histograms of the six images are shown in Figure 2.

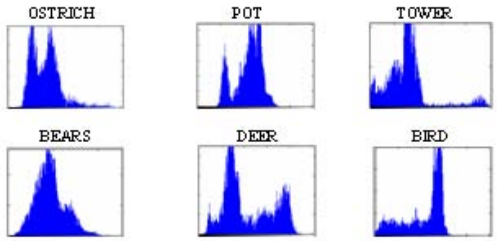


Figure 2 : Histograms Of The Images

The initial estimates of the number of regions K in each image are obtained and given in Table 1.

Table1 : Initial Estimates of K

IMAGE	OSTRICH	POT	TOWER	BEARS	DEER	BIRD
Estimate of K	2	3	4	2	4	3

After assigning these initial values of K to each image data set, the K-means algorithm is performed. The initial values of the model parameters $\mu_{1i}, \mu_{2i}, \sigma_{1i}^2, \sigma_{2i}^2, \rho_i$ and α_i for i=1,2,…,K for each image region of the images are computed by using the method given in section IV. Using these initial estimates, the refined estimates of the model parameters for each image region are obtained by using EM-algorithm given in section III. The computed values of the initial estimates and the final estimates of the model parameters K, $\mu_{1i}, \mu_{2i}, \sigma_{1i}^2, \sigma_{2i}^2, \rho_i$ and α_i for i=1,2…,K for each image are shown in Tables -**2a, 2b, 2c, 2d, 2e and 2f** for different images.

Table-2a
Estimated Values of The Parameters For OSTRICH Image
Number of image regions (K=2)

Parameters	Estimation of Initial Parameters by K-means		Estimation of Final Parameters by EM-Algorithm	
	Regions(i)		Regions(i)	
	1	2	1	2
α_i	1/2	1/2	0.2627	0.7373
μ_{1i}	0.1781	0.1940	0.2054	0.2798
μ_{2i}	0.3321	0.7613	0.2505	0.7775
σ_{1i}^2	0.0016	0.0004	0.0287	0.0772
σ_{2i}^2	0.0126	0.0207	0.0747	0.0768
ρ_i	-0.4310	0.6996	-0.6163	0.2840

Table-2b
Estimated Values of The Parameters For POT Image
Number of image regions (K=3)

Parameters	Estimation of Initial Parameters by K-means			Estimation of Final Parameters by EM-Algorithm		
	Regions(i)			Regions(i)		
	1	2	3	1	2	3
α_i	1/3	1/3	1/3	0.4888	0.2019	0.3093
μ_{1i}	0.5532	0.4946	0.1517	0.5505	0.5223	0.3089
μ_{2i}	0.2168	0.1125	0.1219	0.1958	0.0810	0.1106
σ_{1i}^2	0.0004	0.0027	0.0029	0.0248	0.0269	0.6202
σ_{2i}^2	0.0008	0.0018	0.0035	0.0358	0.0328	0.0469
ρ_i	0.1666	0.3570	-0.7230	-0.4604	0.9867	0.1373

Table-2c
Estimated Values of The Parameters For TOWER Image
Number of Image Regions (K=4)

Parameters	Estimation of Initial Parameters by K-means				Estimation of Final Parameters by EM-Algorithm			
	Regions(i)				Regions(i)			
	1	2	3	4	1	2	3	4
α_i	1/4	1/4	1/4	1/4	0.1999	0.1523	0.1872	0.4606
μ_{1i}	0.1519	0.5699	0.1505	0.5738	0.4011	0.6276	0.1616	0.5743
μ_{2i}	0.1937	0.2789	0.6176	0.7724	0.2408	0.3519	0.5047	0.7721
σ_{1i}^2	0.0033	0.0073	0.0011	0.0006	0.1268	0.1045	0.0043	0.0213
σ_{2i}^2	0.0148	0.0148	0.0291	0.0059	0.1055	0.1696	0.6936	0.0539
ρ_i	0.1561	-0.0259	0.0386	-0.1086	0.9061	0.7744	-0.1902	0.1944

Table-2d
Estimated Values of The Parameters For BEARS Image
Number of Image Regions (K=2)

Parameters	Estimation of Initial Parameters by K-means		Estimation of Final Parameters by EM-Algorithm	
	Regions(i)		Regions(i)	
	1	2	1	2
α_i	1/2	1/2	0.4531	0.5469
μ_{1i}	0.4787	0.2364	0.4867	0.4067
μ_{2i}	0.4532	0.2600	0.4171	0.3375
σ_{1i}^2	0.0027	0.0154	0.0560	0.0667
σ_{2i}^2	0.0129	0.0170	0.0786	0.1439
ρ_i	0.2044	-0.6378	0.1263	0.6274

Table-2e
Estimated Values of The Parameters For DEER Image
Number of Image Regions (K=4)

Parameters	Estimation of Initial Parameters by K-means				Estimation of Final Parameters by EM-Algorithm			
	Regions(i)				Regions(i)			
	1	2	3	4	1	2	3	4
α_i	1/4	1/4	1/4	1/4	0.0703	0.4769	0.2775	0.1753
μ_{1i}	0.1299	0.1144	0.2324	0.3016	0.1388	0.1969	0.2349	0.3185
μ_{2i}	0.6989	0.4560	0.2354	0.1262	0.6847	0.4538	0.2045	0.1039
σ_{1i}^2	0.0004	0.0001	0.0015	0.0026	0.1153	0.1372	0.0366	0.0650
σ_{2i}^2	0.0104	0.0018	0.0019	0.0011	0.1435	0.0125	0.0525	0.1987
ρ_i	-0.1355	-0.0338	-0.0833	-0.0591	-0.0868	-0.0264	-0.2712	-0.2248

Table-2f
Estimated Values of The Parameters For BIRD Image
Number of Image Regions (K=3)

Parameters	Estimation of Initial Parameters by K-means			Estimation of Final Parameters by EM-Algorithm		
	Regions(i)			Regions(i)		
	1	2	3	1	2	3
α_i	1/3	1/3	1/3	0.1029	0.6941	0.2031
μ_{1i}	0.1290	0.5948	0.1425	0.1677	0.6833	0.1946
μ_{2i}	0.6899	0.1143	0.2136	0.6031	0.0965	0.0722
σ^2_{1i}	0.0047	0.0029	0.0048	0.0242	0.2669	0.0237
σ^2_{2i}	0.0330	0.0015	0.0150	0.1369	0.0135	0.5790
ρ_i	0.0834	-0.0504	-0.1409	-0.4398	0.1672	-0.0101

Substituting the final estimates of the model parameters, the probability density function of the feature vector of each image are estimated. Using the estimated probability density functions and the image segmentation algorithm given in section V, the image segmentation is done for each of the six images under consideration. The original and segmented images are shown in Figure 3.

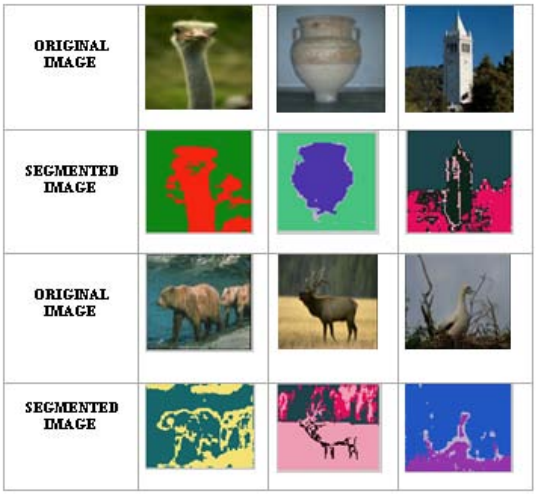


Figure 3 : Original and Segmented Images

VII. Performance Evaluation

After conducting the experiment with the image segmentation algorithm developed in this paper, its performance is studied. The comparison is based on three performance measures namely, Probabilistic Rand Index (PRI) given by Unnikrishnan R. and et.al (2007), the Variation of Information (VOI) given by Meila M. (2005), and Global Consistency error (GCE) given by Martin D. and et al (2001). The objective of the segmentation methods are based on regional similarity measures in relations to their local neighborhood.

The performance of developed algorithm using finite left truncated bivariate Gaussian mixture model (FLTBGMM) is studied by computing the segmentation performance measures namely, PRI, GCE and VOI for the six images under study. The computed values of the performance measures for the developed algorithm and the earlier existing finite Gaussian mixture model (GMM) with K-means algorithm are presented in Table 3 for a comparative study.

Table 3 : Segmentation performance measures

IMAGE	METHOD	PERFORMANCE MEASURES		
		PRI	GCE	VOI
OSTRICH	GMM	0.9234	0.4317	2.2761
	FLTBGMM-K	0.9782	0.4037	1.7611
POT	GMM	0.9456	0.4281	2.5973
	FLTBGMM-K	0.9796	0.4131	1.9263
TOWER	GMM	0.9615	0.4469	3.7121
	FLTBGMM-K	0.9816	0.4302	2.8194
BEARS	GMM	0.9121	0.4418	3.2693
	FLTBGMM-K	0.9831	0.4337	2.6421
DEER	GMM	0.9774	0.4829	2.2863
	FLTBGMM-K	0.9847	0.4030	1.3947
BIRD	GMM	0.9673	0.4671	2.7197
	FLTBGMM-K	0.9705	0.4226	2.3244

From the above Table 3, It is observed that the PRI values of the proposed algorithm for the six images considered for experimentation are more than that of the values from the segmentation algorithm based on finite Gaussian mixture model with K-means. Similarly GCE and VOI values of the proposed algorithm are less than that of finite Gaussian mixture model. This reveals that the proposed algorithm outperforms the existing algorithm based on the finite Gaussian mixture model.

After developing the image segmentation method it is needed to verify the utility of segmentation in model building of the image for image retrieval. Using the estimated probability density function of the images under consideration the retrieved images are obtained and are shown in Figure 4.

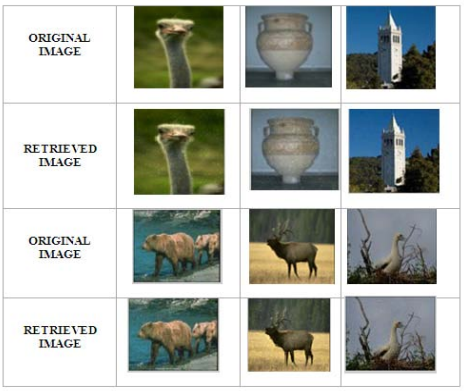


Figure 4 : Original and Retrieved Images

The Performance Evaluation of the retrieved image is done by Subjective Image Quality testing or by Objective Image Quality testing. The Objective Image Quality testing methods are often used since the numerical results of an objective measure are readily computed and allow a consistence comparison of different algorithms. There are several Image Quality

measures available for Performance Evaluation of the Image Segmentation method. An extensive survey of Quality Measures is given by Eskicioglu A.M. and Fisher P.S. (1995). For the Performance Evaluation of the developed Segmentation algorithm, we consider the Image Quality Measures namely (a) Maximum Distance, (b) Image Fidelity, (c) Mean Square Error, (d) Signal to Noise Ratio and (e) Image Quality Index are computed for all the Six images with respect to the developed method and earlier methods and presented in Table- 4.

Table 4 : Comparative study of Image Quality Metrics

IMAGE	Quality Metrics	GMM	FLTBGMM-*K*	Optimal Criteria
OSTRICH	Maximum Distance	0.5013	0.5067	Close to 1
	Image Fidelity	0.7910	0.8076	Close to 1
	Mean Square Error	0.0932	0.0793	Close to 0
	Signal to Noise Ratio	13.3781	13.9959	As big as possible
	Image Quality Index	0.8102	0.8492	Close to 1
POT	Maximum Distance	0.3290	0.3957	Close to 1
	Image Fidelity	0.6729	0.6786	Close to 1
	Mean Square Error	0.0738	0.0467	Close to 0
	Signal to Noise Ratio	11.7401	13.0240	As big as possible
	Image Quality Index	0.6075	0.6174	Close to 1
TOWER	Maximum Distance	0.8481	0.8757	Close to 1
	Image Fidelity	0.5217	0.3884	Close to 1
	Mean Square Error	0.2101	0.1792	Close to 0
	Signal to Noise Ratio	8.8724	8.8488	As big as possible
	Image Quality Index	0.6271	0.5173	Close to 1
BEARS	Maximum Distance	0.5387	0.8765	Close to 1
	Image Fidelity	0.4277	0.6586	Close to 1
	Mean Square Error	0.0872	0.0484	Close to 0
	Signal to Noise Ratio	9.1217	10.7550	As big as possible
	Image Quality Index	0.5951	0.5906	Close to 1
DEER	Maximum Distance	0.6217	0.6474	Close to 1
	Image Fidelity	0.3982	0.4470	Close to 1
	Mean Square Error	0.0828	0.0547	Close to 0
	Signal to Noise Ratio	10.0629	11.8918	As big as possible
	Image Quality Index	0.3763	0.3840	Close to 1
BIRD	Maximum Distance	0.8429	0.9129	Close to 1
	Image Fidelity	0.1920	0.2349	Close to 1
	Mean Square Error	0.2013	0.0900	Close to 0
	Signal to Noise Ratio	8.9231	9.3864	As big as possible
	Image Quality Index	0.3481	0.4160	Close to 1

From the Table 4, it is observed that all the image quality metrics for the six images are meeting the standard criteria. This implies that using the proposed algorithm the images are retrieved accurately. A comparative study of proposed algorithm with that of algorithm based on finite Gaussian mixture model (GMM) and Finite left truncated bivariate Gaussian mixture model with *K*-means reveals that the mean square error of the proposed model is less than that of the finite GMM and FLTBGMM. Based on all other quality metrics also it is observed that the performance of the proposed model in retrieving the images is better than the finite Gaussian mixture model.

VIII. Conclusion

In this paper we introduce a novel and new colour image segmentation method based on left truncated bivariate Gaussian mixture model. Here it is assumed that the colour image is characterized by HSI colour space, in which the Hue and Saturation values are non negative. they are characterized by left truncated Bivariate Gaussian mixture model. The left truncated bivariate Gaussian distribution includes the Bivariate Gaussian distribution is a limiting case when the truncation points tends to infinite. It also includes several platy, meso, lefty and skewed distributions as particular cases for different values of the parameters. The model parameters are estimated by using EM-algorithm. The initialization and the number of image segments are determined through K-means algorithm and moment method of estimation. The segmentation algorithm is developed with component maximum likelihood. The experimentation with six colour images reveals that this algorithm outperforms the existing algorithms in both image segmentation and image retrievals. The image quality metrics also supported the utility of the proposed algorithm. It is possible to develop image segmentation algorithm with finite mixture of doubly truncated multivariate Gaussian distribution with more image features which require further investigations.

References

1. Bengt Muthen (1990) “Moments of the censored and truncated bivariatenormal distribution”, British Journal of Mathematical and Statistical psychology, No.43, pp.131-143.
2. Dipti Patra, Mridula J. and Kumar K. (2011), "Combining GLCM Features and Markov Random Field Model for Colour Textured Image Segmentation", Int. Conf. on Devices and Communications (ICDeCom),pp.1-5.
3. Eskicioglu A.M. and Fisher P.S. (1995) “Image Quality Measures and their Performance”, IEEE Transactions On comm.., Vol.43, No.12, pp.2959-2965.
4. Farag A.A., El-Baz A. and Gimelfarb G. (2004),"Precise Image Segmentation by Iterative EM-Based Approximation of Empirical Grey Level Distributions with Linear Combinations of Gaussians", Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW’04).
5. Ferri F. and Vidal E. (1992), "Colour image segmentation and labelling through multiedit-condensing", pattern Recognition Letters, Vol.13, No.8,pp.561-568.
6. Fesharaki M.N. and Hellestrand G.R. (1992), "Real-time color image segmentation", Technical Report SCS&E 9316,Univ. of New South Wales,Australia.
7. Haralick and Shapiro (1985), “Survey: Image segmentation Techniques”, CVGIP, Vol.29, pp. 100-132.
8. Kaikuo Xu,HongWei Zhang,Tianyun Yan,Wei Wei,Shaomin and Wen Qiang (2011), "An MDL Approach to Color Image Segmentation", International Conference on Multimedia and Signal Processing (CMSP), Vol.2,pp.341-345.

9. Kang Feng, Wang Yaming and Zhao Yun (2009),"Flame Color Image Segmentation Based on Neural Network", International Forum on Computer Science-Technology and Applications, pp.404-407.
10. Kato Z., Pong Ting-Chuen (2006), "A Markov random field image segmentation model for color textured images", Image and Computing Vision, 24(10), pp 1103-1114.
11. Lee E., Kang W., Kim S. and Paik J. (2010),"Color shift model-based image enhancement for digital multifocusing based on a multiple color-filter aperture camera ", IEEE Trans. On Consumer Electronics,Issue-2,pp.317-323.
12. Lucchese L. and Mitra S. K.(2001) "Color image segmentation: A state-of art survey," in Proc. Indian National Science Academy (INSA-A), vol. 67-A, pp.207-221.
13. Majid Fakheri., Sedghi T. and Amirani M.C. (2010), "EM segmentation algorithm for colour image retrieval", 6th Iranian Conference on Machine Vision and Image Processing,pp.1-3.
14. Mantas Paulinas and Andrius Usinskas (2007), "A survey of genentic algorithms applications for image enhancement and segmentation", Information Technology and control, Vol.36, No.3, pp. 278-284.
15. Martin D., Fowlkes C., Tal D., and Malik J., (2001) " A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in proc. 8th Int. Conference Computer vision, vol.2, pp.416- 423.
16. Mclanchlan G. and Krishnan T. (1997), "The EM Algorithm and Extensions", John Wiley and Sons, New York -1997.
17. Mclanchlan G. and Peel D.(2000), " The EM Algorithm For Parameter Estimations", John Wileyand Sons, New York -2000.
18. Meila M. (2005) "Comparing Clustering – An axiomatic view," in proc.22nd Int. Conf. Machine Learning, pp. 577-584.
19. Norman L. Johnson, Samuel Kortz and Balakrishnan (1994), "Continuous Univariate Distributions" Volume-I, John Wiley and Sons Publications, New York.
20. Rahman Farnoosh, Gholamhossein Yari, Behnam Zarpak (2008), " Image Segmentation using Gaussian Mixture Models", IUST International Journal of Engineering Science, Vol. 19, No.1-2, 2008, pp.29-32.
21. Rose H.Turi (2001), "Cluster Based Image Segmentation", phd Thesis, Monash University, Australia.
22. Sangwine S.J. and Horne R.E.N. (1998), "The Colour Image Processing Hand Book," Chapmann and Hall, UK.
23. Seshashayee M., Srinivasa Rao K., Satyanarayana Ch., and Srinivasa Rao P. (2011), "Image segmentation based on a finite generalized new symmetric mixture model with K– means", Int. J. Computer Science Issues, Volume.8, Issue 3, No.2, pp.324-331.
24. Shital Raut Raghuvanshi M., Dharaskar R.; Raut A. (2009), "Image Segmentation- A state-of-Art Survey for Prediction", Advanced Computer control, ICACC'09. International Conference, pp.420-424.
25. Siddhartha Bhattacharyya (2011), "A Brief Survey of Color Image Preprocessing and Segmentation Techniques", Journal of Pattern Recognition Research,pp. 120-129.
26. Srinivas. Y and Srinivas. K (2007), "Unsupervised image segmentation using finite doubly truncated gaussian mixture model and Hierarchical clustering", Journal of Current Science,Vol.93,No.4, pp.507-514.
27. Srinivas Y., Srinivasa Rao K. and Prasad Reddy P.V.G.D. (2010), "Unsupervised Image Segmentation Based on Finite Generalized Gaussian Mixture Model With Hierarchical Clustering, International journal for Computational vision and Biomechanics, Vol.3, No.1, pp.73-80.
28. Sujaritha M. and Annadurai S. (2010), Color image segmentation using Adaptive Spatial Gaussian Mixture Model", International journal of signal processing 6:1, pp. 28-32.
29. Unnikrishnan R., Pantofaru C., and Hernbert M. (2007), "Toward objective evaluation of image segmentation algorithms, " IEEE Trans. Pattern Annl.Mach.Intell, Vol.29,No.6, pp. 929-944.

Global Journal of Computer Science and Technology
Volume 11 Issue 18 Version 1.0 October 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Assessing the Quality of a Software Service at the Time of Project Development by Identifying its Reputation

By Panchamukesh Ch, Venkateswarlu B

Avanthi Institute of Engineering & Technology, Visakhapatnam, India

Abstract - At the time of integration of the software while developing a project the reputation and the quality of execution is tough to identify and which is very risky. As the software industry is introduced with a new type of service delivery model known as SaaS(Software as a service),the problem has increased a lot . Existing system be inclined to rely on rating from customer to experiences of past service which may create major issues in terms of subjectivity and rating unfairness. Few previous works have been considered quality and reputation for selection of services bur none have done service rating process through automation. We proposed an automated quality and reputation framework for rating and selecting a service. In this paper the management of risk has been formulated in context of development of the project using third party software service components and credibility is calculated by a measured reputation system.

Keywords : *Reputation, Service Vendor, Automation, SaaS, Service rating.*

GJCST-D Classification : *H.4.1*

Strictly as per the compliance and regulations of:

Assessing the Quality of a Software Service at the Time of Project Development by Identifying its Reputation

Panchamukesh Ch [α], Venkateswarlu B [Ω]

Abstract - At the time of integration of the software while developing a project the reputation and the quality of execution is tough to identify and which is very risky. As the software industry is introduced with a new type of service delivery model known as SaaS(Software as a service),the problem has increased a lot . Existing system be inclined to rely on rating from customer to experiences of past service which may create major issues in terms of subjectivity and rating unfairness. Few previous works have been considered quality and reputation for selection of services bur none have done service rating process through automation. We proposed an automated quality and reputation framework for rating and selecting a service. In this paper the management of risk has been formulated in context of development of the project using third party software service components and credibility is calculated by a measured reputation system.

Keywords : Reputation, Service Vendor, Automation, SaaS, Service rating.

I. Introduction

The application of a quantifiable, systematic, disciplined approach to the development, process, and maintenance of software can be stated as software engineering. As software industry has huge competition it has shaped a strong motivation for developing solutions to support more responsive and more competitive businesses. Even with long-standing success of COTS (commercial off-the-shelf) software as a time-effective alternative to custom "in-house" developed solutions is still being compromised by the implicated cost of ownership, installation and maintenance time, and effort.

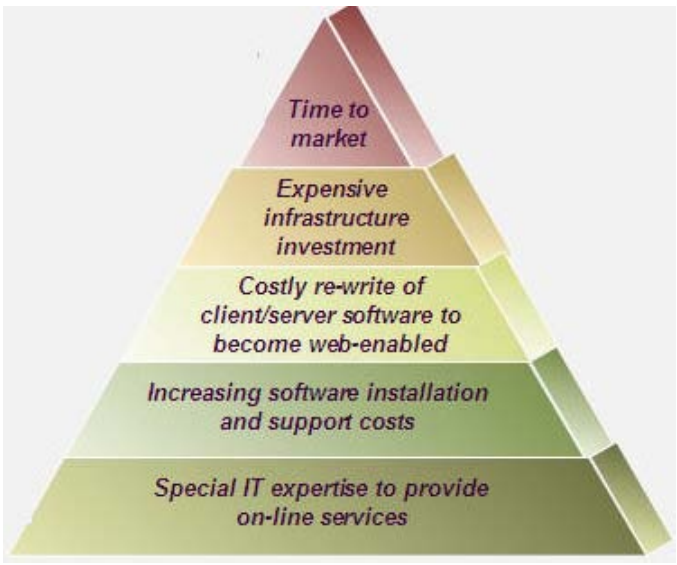


Fig 1 : Challenges for software vendors

That's the reason why software industry has started moving toward a new kind of software delivery model called SaaS(Software as a Service) and which made the things easy to install, maintenance-free, and money-spinning. In Software as a Service (SaaS) software delivery model the software is delivered on-demand and priced on-use, which made it to be widespread implementation of fast Internet access, combined with the widespread acceptance of SOA based solutions. SaaS has gained popularity by reducing the cost of tenure and alleviating the burden of software installation and maintenance. SaaS contributions has expanded dramatically as some of the enterprises have started to outsource their software infrastructure and development projects to SaaS vendors, and the competition has been increased even among vendors of traditional on premises software as in fig 1.

In the world of Software development using service delivery by SaaS model the quality of the software and software provider's credibility is tough and risky. So, the integration of external software in project development is challenging.In this paper risk management has been addressed in context of project development using external software service components. Reputation must be computed on the

Author [α Ω] : M.Tech, Department of Information Technology. Associate. Professor, Avanthi Institute of Engineering & Technology, Visakhapatnam, India .
E-mails : mukesh_1229@yahoo.com, iambondu@gmail.com

basis of fair and objective feedbacks. Most of the works that addressed until now are on evaluating the fairness of existing Feedbacks. Work in this paper focuses instead on the process of generating objective and fair feedbacks. Feedback can be individual since it is based on consumers' "personal" expectations and opinions. Consumers may have an obstructed view of a service reputation systems are prone to attacks by malicious consumers who may give false ratings and subvert service reputation. Consumers may have little incentive to leave a feedback. In this perspective a framework an automated quality and Reputation based framework for service rating and selection has been proposed.

The main objectives of this paper are:

a) In order for a reputation mechanism to be fair and objective, it is essential to compute reputation on the basis of fair and objective feedbacks.
b) The simulation results have demonstrated that the devised system has successfully met our primary objectives and can be an important component in a risk management strategy for software development with SaaS.
c) A computational model is provided to objectively evaluate the delivered service based on the actual measurement of the conformance of the execution quality to the contracted SLA. A novel algorithm is also devised to automate the rating process based on the expectancy-disconfirmation theory from market science.

II. Related Work

What is the main correlation stuck between "reputation" and "trust"? The major difference between reputation and trust can be illustrated by the following statements: (a) "Because of your good reputation I trust you" (b) "I trust you despite your bad reputation." Here the reputation is a collective measure of trustworthiness and is measured based on the referrals or ratings from other members in a community. According to A.josang and R.ismail, reputation is believed about a person's or thing's character or standing. Hence, trust for an individual is measured from the personal reputation and combination of received referrals, as in the Fig 2.

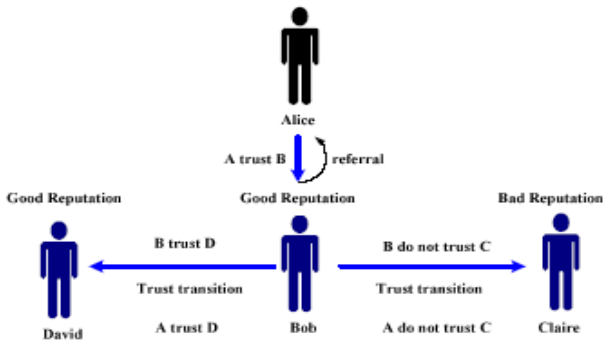


Fig 2 : A Transitive model for consumer reputation

In a centralized reputation management system, the synthetic rating of QoS of web services is aggregated by each rating in the community. To avoid the inapt evaluation by dishonest consumers, it need identify the reputable and disreputable members with their historical comments. Our idea is that consumer reputation is decided by the historical quality of comment, that is, more positive comments gain higher reputation, versa. In other words, lower reputations will worse his/her performance rating on QoS evaluation of web services. When consumers jointing the voting activity can raise their reputation by positive comments and avoid the negative comments. In this work, we proposed a centralized reputation measure for quantifying consumer reputation to properly select the service alternatives, as illustrated in Fig 3.

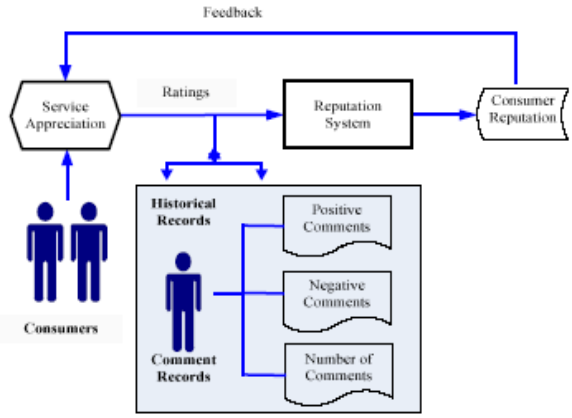


Fig 3 : Consumer Reputation Measure

III. System Analysis & Descripton

For selection of the service many previous works have measured the reputation and quality of the software, but the measurement has been done using some manual tools but none have considered the service rating process in the form of automation. WE introduce a framework for selecting and rating software to provide software service. The important point of the framework which is proposed is to automate both the rating and selection software services which is potentially increasing the objectivity of the service quality reports and concentrating on time-consumption and which finally reduces the risk associated utilization of external software services in development projects.

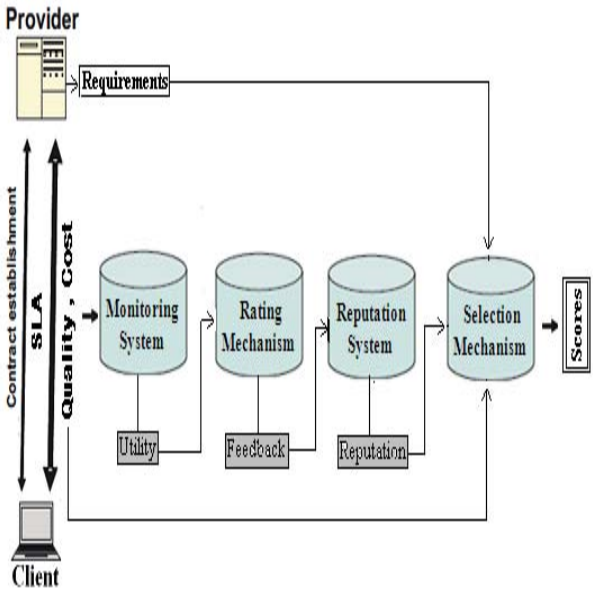


Fig 4 : System Architecture of the framework

While determining a service's suitability to a particular user's preferences in terms of quality and cost the service selection algorithm acts as a user-centric and reputation-aware service recommender. In order for a reputation mechanism to be fair and objective, it is essential to compute reputation on the basis of fair and objective feedbacks. Our work focuses instead on the process of generating objective and fair feedbacks, while most of the works that addressed this latter issue are on evaluating the fairness of existing feedbacks. Here concentrated the calculation of the reputation on works in the area of Service Level Agreement (SLA) monitoring where a computational model is provided to neutrally assess the delivered service based on the actual measurement of the execution quality to the contracted SLA.

In this paper we proposed a framework which has four major modules like Consumer, SLA (Service Level Agreement), Service Providers and Reputational System. Consumer can start the selection based on the trustworthiness features. Consumer selection information will be stored inside database like reputation table. SLA maintains some of the requirements about that particular service. These requirements can be co-inside with SLA requirements and for those services only the service certificate will be approved and that Certificate can be used as Trustworthiness certificate. The services which are provided by SLA can also be present in the service providers itself. User can be satisfied with certified services or trustworthy services. All the user behaviors features can be located inside the trustworthy services. To start the selection at the consumer side we should place the all the features inside that particular service. Reputation can be defined based on the frequent item selection procedure to define the utility measurement identification. Based on utility measure the feedback about that particular service will be defined. The proposed reputational framework is as shown in fig 4.

And the functional requirements of the proposed frame work will be as Enter Consumer Details, Update Consumer Required Services, and Enter Service Provider Details, Service updated to SLA, Retrieve Services, Select Service, Utility Measure of Service, Rating Function, Retrieve Feedback, Consumer Preference Updated, Select service and Calculate Score. An empirical study of the risk factors related to the development using external software (COTS-like) components along with associated risk reduction activities has been reported in. It showed that risk reduction at software selection time is negatively correlated with occurrences of most project development-related risks. In fact, selection must be driven by quality constraints, with selection time evaluation of component quality and choice of appropriate service providers all essential to successful integration. However, in practice, the evaluation of service quality cannot be performed until the service is acquired. Consequently, quality evaluation is typically limited to the evaluation of quality offers by comparing the quality level that providers promise to the quality requirements. Compliance cannot be guaranteed at selection time, so it is essential to choose a provider that is trusted to respect its commitments.

IV. System Design & Implementation

a) Designing of the Framework

In the system design of a system, a number of classes are identified and grouped together in a class diagram which helps to determine the static relations between those objects. With detailed modeling, the classes of the conceptual design are often split in a number of subclasses.

In order to further describe the behavior of systems, these class diagrams can be complemented by state diagram or UML state machine. Where in our framework we have four classes Service Provider, SLA, Consumer and reputational System as in Fig 5.Here Service provider will check for the service name, cost of service, utility of service and value of time. In SLA class it will monitor the service and measures the utility and produces the rating function and identifies the feedback. In Consumer class consumer will select the category, finds utility, cost and selects the service. In Reputation System it identifies the user preferences then select the service and maintains the time.

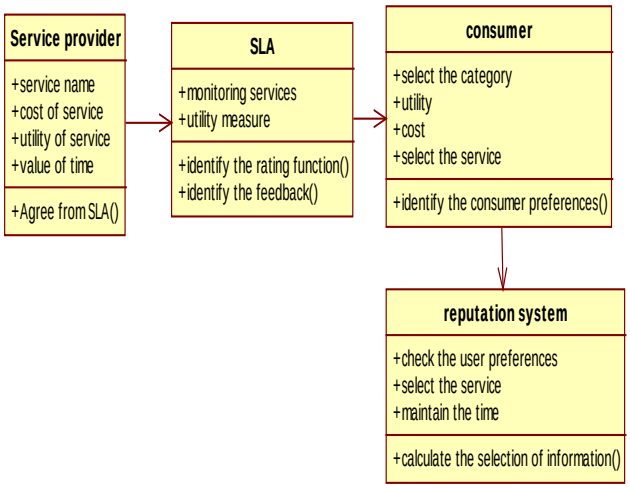


Fig 5 : Inter-operational Class diagram for framework

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted. Use Case diagrams are formally included in two modeling languages defined by the OMG: the Unified Modeling Language (UML) and the Systems Modeling Language (SysML).Major two components for a use case diagram are as follows:

- **Use cases**

A use case describes a sequence of actions that provide something of measurable value to an actor and is drawn as a horizontal ellipse. Where in our Framework we have set of use-cases like Service Name, Category, Cost of Service, Utility, Value of Time, Monitoring the services, Service Rating, User Preferences, Reputation, Feedback as in Fig 6.

- **Actors**

An actor is a person, organization, or external system that plays a role in one or more interactions with the system. Where in our framework we have Service Provider and Consumer as actors.

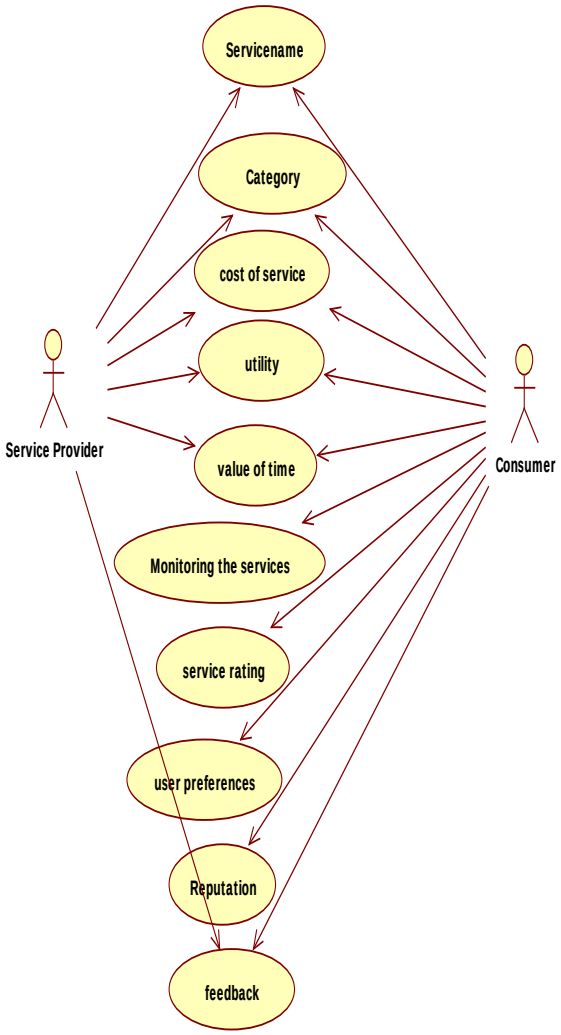


Fig 6 : Inter-operation Use Case diagram for the framework.

b) Implementation of the Framework

The Algorithm representation of the framework is detailed below:

Step 1: Start

Step 2: Enter the Customer Details.

Step 3: Update the Consumer Required Services.

Step 4: Service will be updated into SLA.

Step 5: Consumer will retrieve the services.

Step 6: Consumer choose to select the service.

Step 7: Measurement of Utility for the service.

Step 8: Calculating the Rating Function.

Step9: Retrieving the Feedback for the Service selected.

Step 10: Consumer Preference will be updated.

Step 11: Calculate the Score depending the selected service.

Step 12: Select the service depending on the score achieved.

Step 13: Stop.

V. Results

The following are the screen shots of the system.

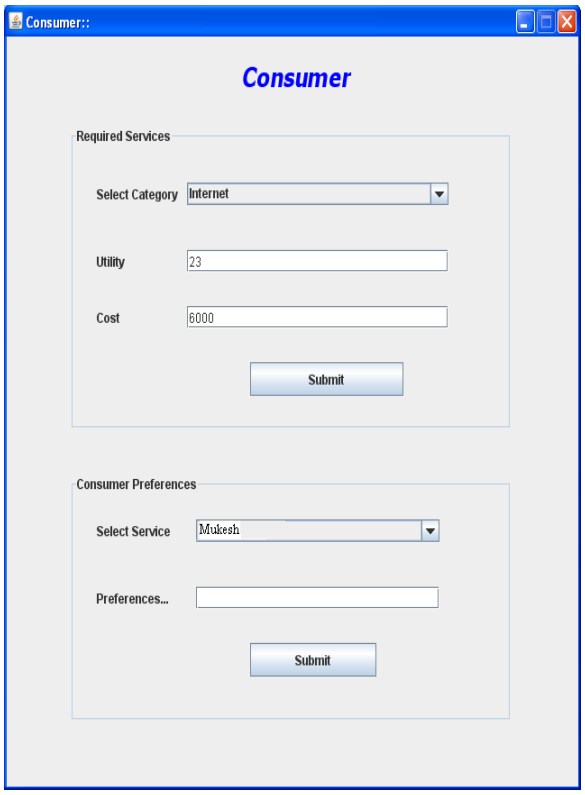


Fig 7 : Selection of required service by consumer

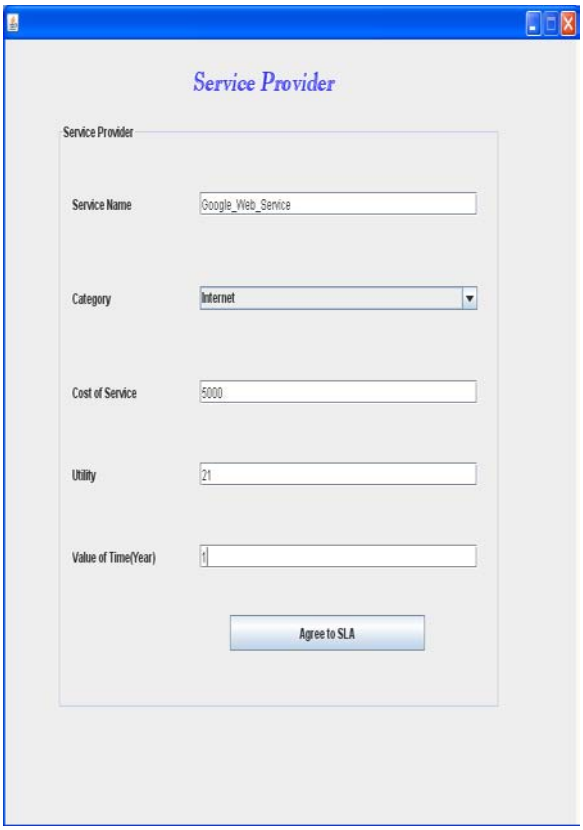


Fig 8 : Updating the type of service provided by service provider

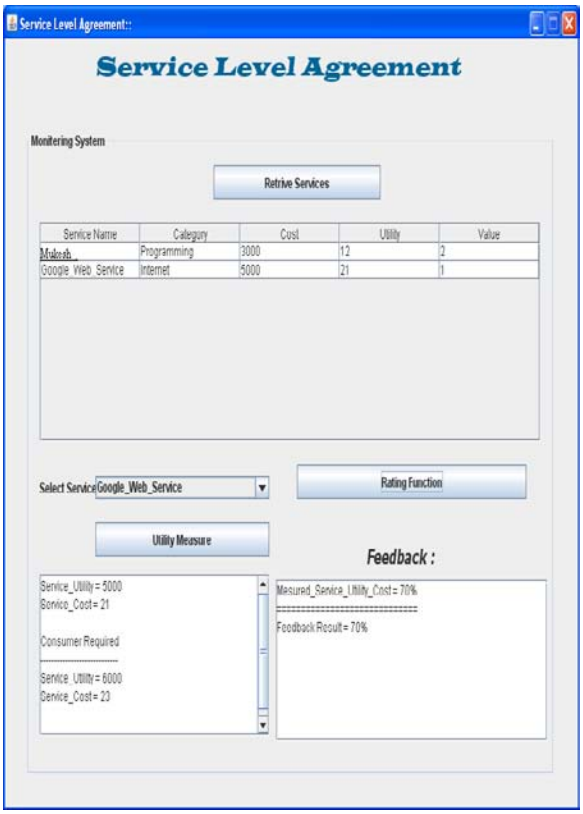


Fig 9 : The SLA between Consumer and Service Provider

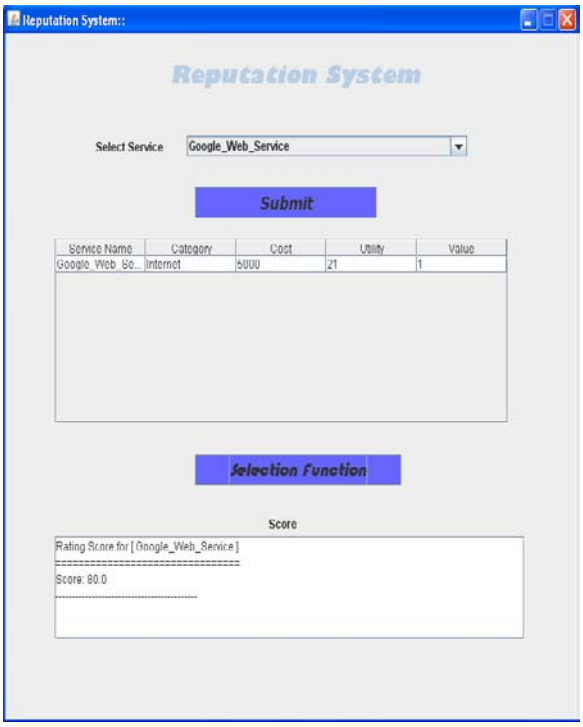


Fig 10 : Calculating the Score of the service using Reputation System

VI. Conclusion

In this paper we addressed the risk of incorporating a third party software for development of a project. To overcome the risk factor, proposed an outstanding framework Identifying the Reputation and Assessing the Quality of a Software Service at the Time of Project Development. We highlighted the framework by adding enhanced features like consumer, SLA, Service Provider and Reputation System which made as added additional advantage in rating and selecting the software to be used for integration. The proposed framework have accomplished in confining the service behaviors and translating them into probable customers choice.

References References Referencias

1. A. Jøsang, R. Ismail, C. Boyd, "A survey of trust and reputation systems for online service provision", Decision Support Systems, Vol.43, No.2, 2007, pp.618-644.
2. J. Li, R. Conradi, O.P. Slyngstad, M. Torchiano, M. Morisio, and C.Bunse, "A State-of-the Practice Survey of Risk Management in Development with Off-the-Shelf Software Components," IEEE Trans.
3. T. SaaS, "Trust Saas: Putting the Trust in Software as a Service (SaaS)," http://trustsaas.com/, 2008.
4. J. Skene, F. Raimondi, and W. Emmerich, "Service-Level Agreements for Electronic Services," IEEE Trans. Software Eng., vol. 36,no. 2, pp. 288-304, Mar./Apr. 2010, http://doi.ieeecomputersociety.org/10.1109/TSE.2009.55.
5. A.Keller and H. Ludwig, "The WSLA Framework: Specifying and Monitoring Service Level Agreements for Web Services,"J. Network and System Management, vol. 11, no. 1, pp. 57-81, 2003.
6. Software Eng., vol. 34, no. 2, pp. 271-286, Mar./Apr. 2008.
7. J. Anselmi, D. Ardagna, and P. Cremonesi, "A QoS-Based Selection Approach of Autonomic Grid Services," Proc. Workshop Service-Oriented Computing Performance: Aspects, Issues, and Approaches, 2007.
8. L. Zeng, B. Benatallah, A.H. Ngu, M. Dumas, J. Kalagnanam, and H. Chang, "QoS-Aware Middleware for Web Services Composition," IEEE Trans. Software Eng., vol. 30, no. 5, pp. 311-327, May 2004.
9. L.-H. Vu, M. Hauswirth, and K. Aberer, "QoS-Based Service Selection and Ranking with Trust and Reputation Management," Proc. 13th Conf. Cooperative Information Systems, 2005.
10. J. Skene, F. Raimondi, and W. Emmerich, "Service-Level Agreementsfor Electronic Services," IEEE Trans. Software Eng., vol. 36,no. 2, pp. 288-304, Mar./Apr. 2010, http://doi.ieeecomputersociety .org/10.1109/TSE.2009.55.
11. J.R. Douceur, "The Sybil Attack," Proc. First Int'l Workshop Peer-to-Peer Systems, 2002.
12. J. Skene, A. Skene, J. Crampton, and W. Emmerich, "The Monitorability of Service-Level Agreements for Application-Service Provision," Proc. Sixth Int'l Workshop Software and Performanc, pp. 3-14, 2007.
13. BelGOnet,"Service Level Agreement,"http: //www .belgonet.com/website/UK/Service_Level_Agreeme nt_UK. Pdf, 2008.[9] EZSM, "EZSM Service Level Agreement," http:// www.easyserver manage ment.com/sla .php, 2008.

Global Journal of Computer Science and Technology
Volume 11 Issue 18 Version 1.0 October 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

A Robust Approach to Find the Control Points for Wide Variety of 3rd Order Bézier Curves

By Alok Kumar Chowdhury, Prithwi Raj Chakraborty, Md. Ibrahim Khan, Sujan Chowdhury

Premier University, Chittagong, Bangladesh

Abstract - This paper represents a new approach that can recover the control points for wide variety of 3rd order Bézier curves. In this regards, the two stage approximation learning algorithm is adopted with some modifications. At 1st stage our key feature is segmentation of the curve which can determine intermediate points of the wide variety of curves. In this respect, an efficient recursive algorithm is used to find out the height of the curve (h) with less iteration. The proposed approach introduced horizontal segmentation rather than vertical segmentation. Different height (H), where the 2nd and 3rd control point are assumed, and also the step-size (∂), at which the control points are moved toward the actual direction, are used to find out the exact location of the control points. Experimental results demonstrate that our proposing method can recover control points for wide variety of curves with minimum error level and less iteration. Wide variety of curve shapes are used to test the proposing approach and results are presented to prove its effectiveness.

Keywords : *Bézier curve, curve fitting, segmentation of curve, learning algorithms.*

GJCST-B Classification : *I.3.5*

Strictly as per the compliance and regulations of:

A Robust Approach to Find the Control Points for Wide Variety of 3rd Order Bézier Curves

Alok Kumar Chowdhury[α], Prithwi Raj Chakraborty[Ω], Md. Ibrahim Khan[β], Sujan Chowdhury[ψ]

Abstract - This paper represents a new approach that can recover the control points for wide variety of 3rd order Bézier curves. In this regards, the two stage approximation learning algorithm is adopted with some modifications. At 1st stage our key feature is segmentation of the curve which can determine intermediate points of the wide variety of curves. In this respect, an efficient recursive algorithm is used to find out the height of the curve (h) with less iteration. The proposed approach introduced horizontal segmentation rather than vertical segmentation. Different height (H), where the 2nd and 3rd control point are assumed, and also the step-size (∂), at which the control points are moved toward the actual direction, are used to find out the exact location of the control points. Experimental results demonstrate that our proposing method can recover control points for wide variety of curves with minimum error level and less iteration. Wide variety of curve shapes are used to test the proposing approach and results are presented to prove its effectiveness.

Keywords : Bézier curve, curve fitting, segmentation of curve, learning algorithms.

I. Introduction

The development of Computer Graphics has made computers easier to interact with, and better for understanding and interpreting many types of data which has put a profound impact on many types of media and have revolutionized animation, movies and the video game industry. Generally, Computer Graphics are the representation and manipulation of image data by a computer with help from specialized software and hardware [1].

In Computer Graphics and related fields, a frequently used parametric curve is Bézier curve. Bézier curves are used to model smooth curves that can be scaled indefinitely. As the curve is completely contained in the convex hull of its control points, the points can be graphically displayed and used to manipulate the curve intuitively. When more complex shapes are needed, Bézier curves are patched together. In Animation applications, such as Adobe Flash and Synfig, Bézier curves are used to outline, for example, movement. Besides,True-type Fonts use Bézier curves. Bézier curve is also a very powerful tool for shape approximation and shape comparison. In identifying of actors drawn in ukiyoe pictures, Bézier curves are widely used [2]. They can be used in face recognition and facial emotion detection [3].

A Bézier curve is defined by its order (Linear, Quadratic, Cubic, etc.) and a set of control points. An nth order Bézier curve has n+1 control points (P_1 to P_{n+1}). The first and last control points are always the end points of the curve. These two end points are can be called initial and terminating point of the curve respectively. The intermediate control points generally do not lie on the curve; they define the shape and direction of the curve. A 1st order (Linear) Bézier curve is simply a straight line between those two given points P_1 and P_2. A 2nd order (Quadratic) Bézier curve is the path traced by the function B(t), given points P_1, P_2, and P_3. Four points P_1, P_2, P_3 and P_4 in the plane or in three-dimensional space define a 3rd order (Cubic) Bézier curve. The 3rd order curve starts at P_1 going toward P_2 and arrives at P_4 coming from the direction of P_3. P_1 and P_4 are end points [4-5]. Here P_1 and P_4 are end points as well as P_2 and P_3 are intermediate control points. A 3rd order Bézier curve given by the following equation is shown in fig. 1.

$$B(t) = (1-t)^3 P_1 + 3t(1-t)^2 P_2 + 3t^2 (1-t) P_3 + t^3 P_4 \qquad (1)$$

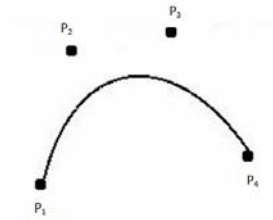


Fig.1 : A sample Bézier Curve

Our discussions will be limited to 3rd order Bézier curve. Though there are several algorithms and approaches to find out control point of 3rd order Bézier curve, these approaches have limitations like incorrect result, requirement of compatible boundary for control points etc. In our previous work [6], we tried to overcome these limitations and it finds control points of 3rd order Bézier curve in an efficient way. But it is still unable to find out control point of 3rd order Bézier curves of some certain shapes. In this paper we are going to propose a modified approach of our previous work to erase the existing limitations. Our new approach is capable of finding out the control points of a large variety of 3rd order Bézier curve shapes efficiently and accurately with minimum error and less iterations.

Author α : Department of Computer Science & Engg. Premier University, Chittagong, Bangladesh. Email : alokchy04@yahoo.com
Author Ω : Department of Computer Science & Engg. Premier University, Chittagong, Bangladesh. Email : stranger_prithwi@yahoo.com
Author β : Department of Computer Science & Engg. Chittagong University of Engg. & Tech. (CUET), Chittagong, Bangladesh. Email : muhammad_ikhancuet@yahoo.com
Author ψ : Department of Computer Science & Engg. Chittagong University of Engg. & Tech. (CUET), Chittagong, Bangladesh. Email : sujan_cse_04@yahoo.com

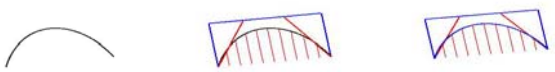
II. Background

There has been several works in the field of recovering control points of Bézier curves. All the familiar algorithms and characteristics are associated with Bézier curves are generation [7], approximation [8], interpolation [9], subdivision [10-12], degree elevation [13-16], blossoming [17], implication [18] etc. X. Ye represented an approach for directly generating Bézier points of curves and surfaces explicitly from the given compatible arbitrary order boundary information of Hermite curves, Coons-Hermite Cartesian sum patches and Coons-Boolean sum patches [19]. But for Hermite curves, this approach requires the end positions, tangents and higher order derivatives at the end points. Moreover it requires the corner points and the compatible arbitrary order partial derivatives at these points for Coons–Hermite Cartesian sum patches.

In determining inner control points of 3rd order rational Bézier curve, the interpolation method developed by J. Chou and L. A. Piegl works good for special type of 3rd order Bézier curves, not all kind of Bézier curves [20]. Besides, this approach relies on the convex hull and on the variation diminishing properties of Bézier curves. And the most important drawback is that if the curve segment is highly curved in one region and relatively flat in another, the approximation of this method is not good.

The previous approach, proposed by us, contains none of these problems. It can successfully find out control points of 3rd order Bézier curve with minimum iteration and error. Fig. 2(b) and 2(c) are showing the way how the previous approach successfully finds out the control points for the curve of a particular shape indicated in fig. 2(a). The details are described at the previous work [6].

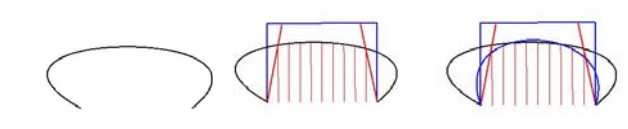
However, for some curves of shapes like fig. 3(a) and 4(a), the previous approach is unable to find out the accurate control points. It is because the segmentation process of the previous approach, which uses vertical segmentation, is different from the current one (described at Methodology section).



(a) (b) (c)

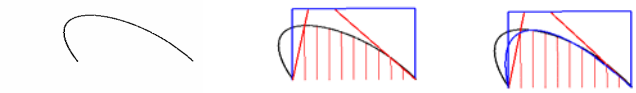
Fig.2 : (a) A typical Bézier curve, (b) Segmentation, (c) Successful-recovery of the curve (previous approach)

Fig. 3 and 4 are showing limitation of the previous approach in finding out the control points for some special shaped curves. The previous approach cannot rescue the whole area for these special shaped curves, indicated in fig. 3(c) and 4(c). Thus the control points found are also not correct for these curves. The curves drawn in blue color are the recovered curves.



(a) (b) (c)

Fig.3 : (a) A typical Bézier curve, (b) Segmentation, (c) Failed-recovery of the curve (previous approach)



(a) (b) (c)

Fig.4 : (a) A typical Bézier curve, (b) Segmentation, (c) Failed-recovery of the curve (previous approach)

But the new approach, with modified segmentation method, is able to find out the actual control points very successfully with minimum iteration and error for all the 3rd order Bézier curves of shapes indicated in fig. 2(a), 3(a) and 4(a). It can rescue the whole area of special shaped curves which is described later in Simulation Results section.

III. Methodology

As it is mentioned before, a 3rd order Bézier curve contains four control points. The objective is to find out them. In order to do this, the new proposing approach is divided into two stages: modified first stage and second stage.

a) Modified first stage

The new proposing approach is actually different from our previous one because of this stage. This stage is described as following:

1) Finding out of the end points: The end points will be lying on the Bézier curve. End points must be those two which have minimum y values. Now it is to be decided which one of these end points is initial point and which one is terminating point. Among these two end points the point which is nearest from Y axis (i.e. x value is minimum) is our initial point, $P_1(x_1, y_1)$. So, the remaining one is for sure the terminating point, $P_4(x_4, y_4)$. These two end points are shown in fig. 5.

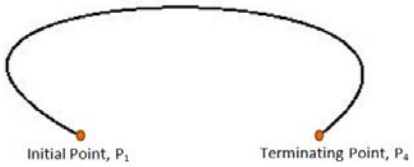


Fig.5 : End points of a Bézier curve

2) Finding out of the height: After finding out the initial and terminating point, the height of the Bézier curve, h is to be found by the following steps:

- A straight line between initial point $P_1(x_1, y_1)$ and terminating point $P_4(x_4, y_4)$ is drawn using the following equation:

$$(y - y_1)(x_4 - x_1) = (x - x_1)(y_4 - y_1) \qquad (2)$$

This line is called base line, $P_1 P_4$ shown in fig. 6(a).

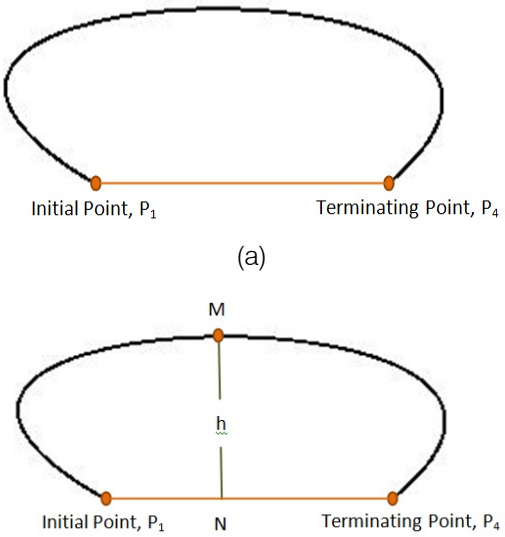


Fig.6 : (a) Base line, (b) Height and peak intersection point of a Bézier curve

- Now the height, h and the peak intersection point, M of the Bézier curve and the parallel line drawn at distance h from the base line are found by a recursive algorithm which makes the new approach faster than the previous one. This algorithm draws some parallel lines at some assumed distances, d from the base line. If a parallel line intersects the curve at two points, this line is stored and the next parallel line is drawn above the currently stored parallel line at a distance d. If the drawn parallel line does not intersect the curve then that line is not stored and the next parallel line is drawn above at a distance d / 2 from the previously stored parallel line. This algorithm contains three methods – draw_parallel(_line, _d), intersections(_curve, _line), and distance(_line1, _line2). The draw_parallel(_line, _d) method takes a line equation and a distance as inputs and returns the parallel line drawn at a distance _d from the line _line. If the equation ax + by + k = 0 be the value of input variable _line and d is the value of input variable _d, then draw_parallel(_line, _d) returns a parallel line using the following equation:

$$ax + by + [d(\sqrt{a^2 + b^2}) + k] = 0 \qquad (3)$$

The next method intersections(_curve, _line) takes a curve equation and a line equation and returns the number that represents at how many points the line intersects the curve. The peak most segment of the Bézier curve may be appeared as a tiny line rather than a curve. Therefore, the parallel line may intersect the curve at several consecutive intersection points. In that case, the middle one from the consecutive points is taken as the desired single intersection point and the number of intersections is considered as one. The last method distance(_line1, _line2) returns the distance between two lines. Now we should look at the recursive algorithm.

a. Set temp_base_line = base_line.
b. Repeat step c to g.
c. Set parallel_line = draw_parallel(temp_base_line, d).
d. Set no_of_intersections = intersections (curve, parallel_line).
e. If no_of_intersections = 1, then: Set M = intersection point of curve and parallel_line. Set h = distance(base_line, parallel_line) and Return h and M.
f. Else If no_of_intersections = 0, then: Set d = d / 2.
g. Else If no_of_intersections = 2, then: Set temp_base_line = parallel_line.
h. Return h and M

Thus fig. 6(b) shows the height, h and the peak intersection point, M found by the above algorithm.

3) Segmentation of the Bézier curve: The segmentation process of the new proposing approach follows horizontal approach rather than vertical approach followed by the previous approach. In segmentation, the following steps are followed:

- A normal is drawn on the base line from the point M. This normal obviously intersects the base line or extended base line and the intersection point between normal and the base line can be easily found. Let this point is N. Now the normal (MN) is divided into n equidistant points, (x_1'', y_1''), (x_2'', y_2'')... (x_n'', y_n''). The normal and these n equidistant points are shown by fig. 7(a). If $M(x_1''', y_1''')$ and $N(x_2''', y_2''')$ are the two terminal points of the normal then the normal (MN) is divided using the following equation:

$$x_i'' = (m_1 x_1''' + m_2 x_2''')/(m_1 + m_2) \qquad (4)$$

$$y_i'' = (m_1 y_1''' + m_2 y_2''')/(m_1 + m_2) \qquad (5)$$

The greater value of n is taken, the more accurate segmentation and thus more accurate result is found but in that case computational cost gets higher.

- Now n parallel lines PL_1, PL_2… PL_n of the base line are drawn which are going through n equidistant points, (x_1'', y_1''), (x_2'', y_2''), ….., (x_n'', y_n'') on the normal (MN). Each of these n parallel lines of the base line intersects the Bézier curve into two points on both sides of the normal (MN). So, there will be 2n intersection points (except initial point, terminating point and M) lying on the Bézier curve for n parallel lines. Fig. 7(b) shows these 2n points, which are (x_1', y_1'), (x_2', y_2')…(x_{2n}', y_{2n}') These 2n points and M are the desired segmentation points.

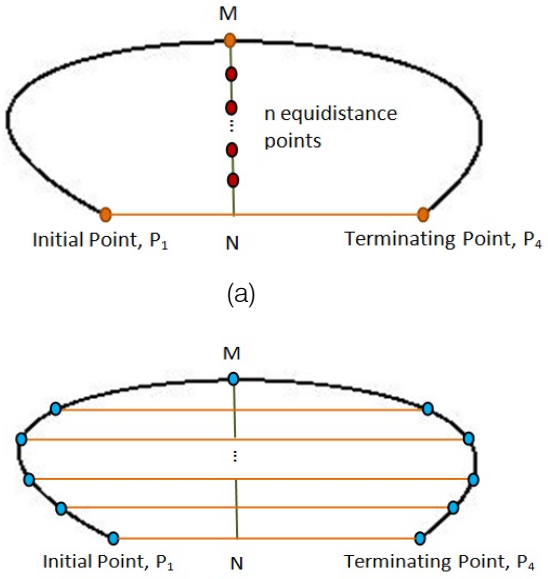


Fig.7 : (a) Normal, equidistant points, (b) Segmentation points of a Bézier curve

4) Assumption of the intermediate control points: This portion of this proposal is same as our previous work. A 3rd order Bézier curve contains two intermediate control points *(P_2 and P_3)*. In this portion these two intermediate control points are assumed. The steps are:

- From the base line, several parallel lines at distance H are drawn. H indicates assumed height with different values - *(3 / 4) h, h, (4 / 3)h* and *1.5h.* Here h is the height of the Bézier curve. Fig. 8 is showing such a parallel line drawn at H = 1.5h distance.
- After each parallel line is drawn, two tangents from initial and terminating points are drawn shown in fig. 8. The tangents must intersect the parallel line at two points. These two points are the assumed intermediate control points. Let one point is P_2', the assumed 2nd control point and other one is P_3', the assumed 3rd control point.
- So, for four values of H (four different distances) - *(3 / 4)h, h, (4 / 3)h* and *1.5h,* four pair of assumed intermediate control points *(P_2', P_3')* are found. Such a pair of points found at distance 1.5h is shown by fig. 8.

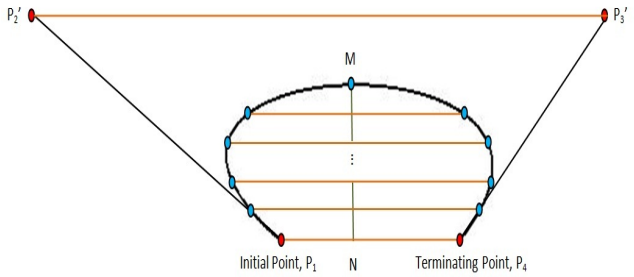


Fig.8 : Parallel line, tangent and assumed intermediate control points of a Bézier curve

From observation it has been found that the original intermediate control points (P_2, P_3) generally lie somewhere around the distances - (3 / 4)h, h, (4 / 3)h and 1.5h. The accurate assumption reduces the number of iterations [6]. It is the reason for drawing parallel lines from the base line at these different values of H. Each distance give us a pair of assumed intermediate control points *(P_2', P_3')*.

b) Second Stage

This stage is same as our previous work [6]. As proposing method focus on minimization of error, so the 2nd and 3rd control point must be in the desired location. In order to find the exact location the approach always calculate the error between the two curves iteratively. If the error becomes minimum, according to algorithm it finds the exact location of 2nd and 3rd control points. The value of the error is the summation of the difference between the points obtained from given curve *Q(t)* and their corresponding points generating from newly found control points in the approximated Bézier curve *Q'(t)* which is shown in fig. 9. Since Bézier curves can be obtained using a parameter t, where *0<=t<=1*, the error between two curves *Q(t)* and *Q'(t)* can be computed by considering the curves in parametric expressions and finding the corresponding points. For *N+1* corresponding points, we are getting locations of corresponding points *Q(t)=(Qx(t),Qy(t))* and *Q'(t)= (Q'x(t),Q'y(t)), (t=0,1/N,2/N,…,1).* The equation that is used for calculating error as follows:

$$Error, E_1 = \sum_{k=0}^{N} | Q(t = k/N) - Q'(t = k/N) |$$

$$= \sum_{k=0}^{N} \sqrt{\{Q_x(k/N) - Q'_x(k/N)\}^2 + \{Q_y(k/N) - Q'_y(k/N)\}^2}$$

The reason why summation is used instead of integration is for computational simplicity. It is applicable when the original Bézier curve is known.

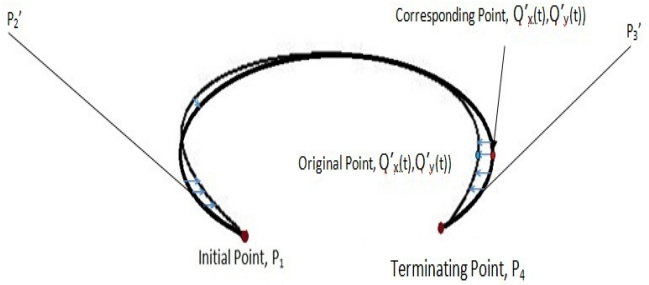


Fig.9 : Computation of the sum of the differences between the corresponding points.

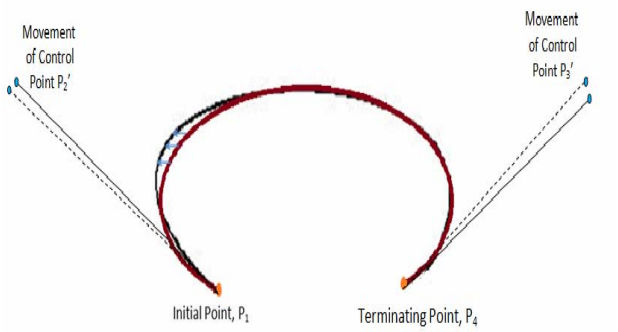


Fig.10 : Movement of the control point to recover error

In order to recover the exact control points the proposing approach try to move those points at a step ∂ towards both X and Y coordinates which shown in fig. 10. The value of the control point will be updated if the calculated error is minimized, otherwise it will not update. If the step fails to minimize error according to algorithm it reduces the step by half and recursively performing the same operation as stated above.

Increasing the efficiency of the overall program was the main contribution as well as overcoming the limitation of the step-size. Although the approach able to get the accurate result for some Bézier curve by considering step size ∂x=∂y=1 [21], but it takes more iterations compared to our propose step size, which reduce the efficiency of the overall performance. Here proposing modification works well for most of the Bézier curve and get the error level less than 0.000001 which increase the dynamicity of overall program because it need less iteration. With variable step ∂x=∂y=5, 50, 75 and 100 in our simulation result we get near to 100% accuracy. The algorithm followed in the second stage is given below.

[Step 1]
Initialization: allsteps [] = {1, 5, 50, 75,100}
Count=0(Learning time)
MAX=100 (Maximum learning time)
Stepdeterminer = 0 (Variable to determine step)
∂x=∂y=allsteps [stepdeterminer] (Variable displacement for the movement of the control points)
x=assumed_x; y=assumed_y (Initial assumption of control points at 1st stage)
e=0.000001(Minimum permissible error)

[Step 2]
Count++; if (Count>MAX), then goto step 3
if E (x+∂x, y) or E (x-∂x,y) is minimum then goto step 3
if E (x,y+∂y) or E (x,y-∂y) is minimum then goto step 4
if E (x,y) is minimum then $\partial x = \partial x/2$, $\partial y = \partial y/2$, goto step 2

[Step 3]
Searching minimum error for x-direction
While (E (x+∂x,y) < E (x ,y)) { Count++, x=x+∂x }
While (E (x-∂x,y) < E (x ,y)) { Count++, x=x-∂x }
if (Count>MAX||E(x,y)<e),then goto Step 5 else goto Step 2

[Step 4]
Searching minimum error for y-direction
While (E (x,y+∂y) < E (x ,y)) { Count++, y=y+∂y }
While (E (x,y - ∂y) < E (x ,y)) { Count++, y=y-∂y }
if (Count>MAX||E(x,y)<e),then goto Step 5 else goto Step 2

[Step 5]
If (E(x,y) > e AND stepdeterminer < allsteps.Length) { Count = 0; stepdeterminer ++; ∂x = ∂y = allsteps [stepdeterminer];
x=assumed_x; y=assumed_y; goto Step 2}
Else { Exit }

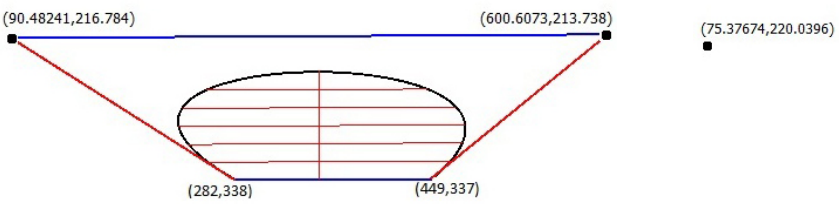
IV. Simulation Results

The main contribution of the proposed scheme is to find control points of large number of Bézier curve shapes while keeping the efficiency and accurateness of our existing method. We checked several Bézier curves of different shapes with our computer implemented simulation and got desired results as we expected.

Consider the simulation result shown in fig. 11 and fig. 12. Figures of simulation results are picked directly from our computer implemented simulation. At first stage we segment given curve to know its internal points and assume some control points at some probable height using our new proposed schema. Then using the algorithm of 2nd stage we successfully find the control points. Previous Method [6] was unable to extract control points of those shaped curves because of its limited segmentation technique.

Those shaped curves, whose control points can be found by existing method, can also be found by the newly proposed technique while the accuracy and efficiency of both methods remain same. Example of fig. 13 shows it clearly.

a) Input Image

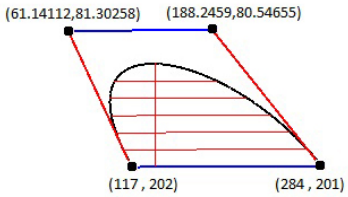


b) Assumption of control points at 1st stage

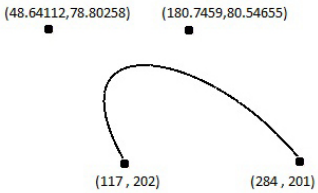
c) Recovered control points

Fig.11 : Simulation Results

a) Input Image



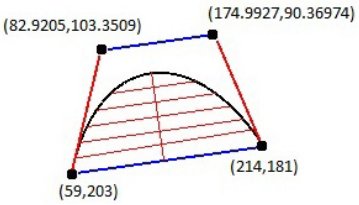
b) Assumption of control points at 1st stage



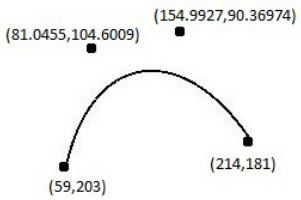
c) Recovered control points

Fig.12 : Simulation Results

a) Input Image



b) Assumption of control points at 1st stage



c) Recovered control points

Fig.13 : Simulation Results

To obtain our desired minimum error we checked our input image of Bézier curve with different combination of H (height) and ∂ (step size). Simulation automatically checks Bézier curve in different combination of H and ∂ until the error < 0.001. Recovered control points for a combination of H and ∂ is chosen as best result if both error and number of iteration are minimum. Table 1 shows recovered control points on each combination with their corresponding error for the example shown in fig. 11.

Table.1 : Calculated error in different combination of height and step-size

Height	Step Size	1st control point(P_1), 2nd control point(P_2), 3rd control point(P_3), 4th control point(P_4)	Iteration	Error
0.75h	5	(282,338),(79,239.6442), (579.5,237.4855), (449,337)	41	0.134
0.75h	50	(282,338),(77.125,216.5192), (606.375,211.2355), (449,337)	26	4E-3
0.75h	75	(282,338),(75.5625,216.9099), (609.5,213.5793), (449,337)	23	4E-3
0.75h	100	(282,338),(77.125,216.5192), (606.375,211.2355), (449,337)	22	4E-3
h	5	(282,338),(78.00368,219.3613), (602.9881,209.3237), (449,337)	40	4E-3
h	50	(282,338),(75.50368,221.8613), (600.4881,206.8237), (449,337)	10	6E-3
h	75	(282,338),(74.72243,225.1816), (598.1443,202.1362), (449,337)	23	14E-3
h	100	(282,338),(75.50368,221.8613), (600.4881,206.8237), (449,337)	11	6E-3
1.33h	5	(282,338),(76.73241,216.784), (609.3573,213.738), (449,337)	10	4E-3
1.33h	50	(282,338), (74.85741,216.784), (606.8573,213.738), (449,337)	8	8E-3
1.33h	75	(282,338),(76.41991,216.784), (607.6385,212.5662), (449,337)	13	4E-3
1.33h	100	(282,338),(74.85741,216.784), (606.8573,213.738), (449,337)	9	8E-3
1.5h	5	(282,338),(73.50174,221.2896), (604.9869,209.2318), (449,337)	16	2E-3
1.5h	50	(282,338),(78.50174,226.2896), (601.2369,204.2318), (449,337)	11	16E-3
1.5h	75	(282,338), (75.37674,220.0396), (603.5806,208.5286), (449,337)	15	2E-3
1.5h	100	(282,338),(78.50174,226.2896), (601.2369,204.2318), (449,337)	12	16E-3

Data from Table 1 shows that minimum error is found at different combination of step-size and height. In this case when height=*1.5h* and step-size= *5*, or height *=1.5h* and step-size=*75* the error is minimum that is *2E-3*. But when height=*1.5h* and step-size=*75*, our simulation takes less iteration. That's why we consider *(282, 338), (75.37674, 220.0396), (603.5806, 208.5286), (449, 337)* as 1st, 2nd, 3rd and 4th control points respectively.

V. Conclusion

The new modified approach can recover the control points of 3rd order Bézier curves more accurately and efficiently than those methods which was proposed earlier. In addition, the emphasis of this paper is on the implementation of a new method that can recover the more different shaped Bézier curve which is not possible according to earlier methods. While conducting the experiment different shapes of 3rd order Bézier curves are taken for simulation and method is proved by the simulation result.

References Références Referencias

1. Computer graphics. [Online]. Available at: http://en.wikipedia.org/wiki/Computer_graphics/
2. MD. Al- Amin Bhuiyan and Hiromitsu Hama, 'Identification of Actors drawn in Ukiyoe picture'. Pattern Recognition, Vol 35, Issue 1, pp. 93-102, 2002.
3. Mohammad Ibrahim Khan and Md. Al-Amin Bhuiyan, 'Facial Expression Recognition for Human-Robot Interface'. IJCSNS International Journal of Computer Science and Network Security, Vol 9, No 4, pp. 300-306, April 2009.
4. G. Farin, 'Curves and surfaces for computer aided geometric design: a practical guide'. Academic Press, 1997.
5. F. Yamaguchi, 'Curves and Surfaces in Computer Aided Geometric Design'. Springer Verlag, 1988.
6. Khan, M.I. Chowdhury, S. Chowdhury, A.K. Deb, K., 'An Efficient Algorithm for Finding the Control Point of 3rd Order Bézier Curve'. International Forum on Strategic Technology, Ulsan. pp. 441-445, 2010.
7. H. Hama and T. Okamoto, 'Fast generation method of Be´zier curves and anti-aliasing'. J. TV Engineers Japan 47(12), 1629–1636, 1993.
8. T. W. Sederberg and R. T. Farouki, 'Approximation by interval Be´zier curves', IEEE Comput. Graphics Appl. 12(5), 87–95, 1992.
9. J. Barry and R. N. Goldman, 'Interpolation and approximation of curves and surfaces using po´lya polynomials'. Graph. Models Image Process. 53(2), 137–148, 1991.
10. E. Catmull, 'A subdivision algorithm for computer display of curved surfaces'. Report UTEC-CSc-74-133, University of Utah, December, 1974.
11. S. Hu, G. Wang, and T. Jin, 'Generalized subdivision of Be´zier surfaces'. Graph. Models Image Process. 58(3), 218–222, 1996.
12. J. Gravesen, 'Adaptive subdivision and the length and energy of Be´zier curves.' Comput. Geometry 8(1), 12–31, 1997.
13. Bogacki, S. E. Weinstein, and X. Ye, 'Degree reduction of Be´zier curves by uniform approximation with endpoint interpolation'. Comput. Aided Design 27(9), 651–661, 1995.
14. S. Lodha and J. Warren, 'Degree reduction of Be´zier simplexes'. Comput. Aided Design 26(10), 735–745, 1994.
15. M. Eck, 'Least squares degree reduction of Be´zier curves'. Comput. Aided Design 27(11), 845–851, 1995.
16. L. Piegl and W. Tiller, 'Algorithm for degree reduction of B-spline curves'. Comput. Aided Design 27(2), 101–110, 1995.

17. L. Ramshaw, 'Blossoming: A connect-the-dots approach to splines'. Digital Systems Research Center Report No. 19, CA, 1987.
18. T. W. Sederberg, J. Zheng, K. Klimaszewski, and T. Dokken, 'Approximate implicitization using monoid curves and surfaces'. Graph. Models Image Process. 61(4), 177–198, 1999.
19. X. Ye, 'Generating Be´zier points for curves and surfaces from boundary information'. Comput. Aided Design 27(12), 875–885, 1995.
20. J. J. Chou and L. A. Piegl, 'Data reduction using cubic rational B-splines'. IEEE Computer Graphics Applications, 12(3), 60–68, 1992.
21. MD. Al- Amin Bhuiyan and Hiromitsu Hama, 'On recovering the control points of Bézier curves for line image indexing'. J. Electron Imaging, Vol 11, Issue 2, pp. 177, 2002.

Global Journal of Computer Science and Technology
Volume 11 Issue 18 Version 1.0 October 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

A Comprehensive Analysis of Congestion Control Using Random Early Discard (RED) Queue

By Md. Abdullah Al Mamun, Momotaz Begum, Mridul Kanti Das, Md. Rubel

Dhaka University of Engineering & Technology (DUET), Gazipur, Bangladesh

Abstract - Normally all the congestion control method discard the received packet when the queue is full but it is a great problem for speed of data transfer at present. There are many ways to solve this problem. Random Early Detection (RED) algorithm is one of the most famous and powerful method to improve the performance for TCP Connection. In terms of queue management RED drops packet in considered router buffer to adjust the network traffic behavior according to the queue size. We want to investigate how high priority user datagram protocol (UDP) traffic affects the performance of lower priority Transmission Control Protocol (TCP) and proof that RED is the better for controlling the Traffic when they share the same bottleneck link with one or two classes of service.

General Terms : *Network Congestion Control, TCP Variants, Network Parameters, Queuing, Drop Tail and User Data Gram Protocol.*

Keywords : *IETF, RED, AQM, BW, TCP Variants, NS-2, TCL and OTCL.*

GJCST-F Classification : *C.2.2*

Strictly as per the compliance and regulations of:

A Comprehensive Analysis of Congestion Control Using Random Early Discard (RED) Queue

Md. Abdullah Al Mamun[α], Momotaz Begum[Ω], Mridul Kanti Das[β] , Md. Rubel[ψ]

Abstract - Normally all the congestion control method discard the received packet when the queue is full but it is a great problem for speed of data transfer at present. There are many ways to solve this problem. Random Early Detection (RED) algorithm is one of the most famous and powerful method to improve the performance for TCP Connection. In terms of queue management RED drops packet in considered router buffer to adjust the network traffic behavior according to the queue size. We want to investigate how high priority user datagram protocol (UDP) traffic affects the performance of lower priority Transmission Control Protocol (TCP) and proof that RED is the better for controlling the Traffic when they share the same bottleneck link with one or two classes of service.

General Terms : Network Congestion Control, TCP Variants, Network Parameters, Queuing, Drop Tail and User Data Gram Protocol.

Keywords : IETF, RED, AQM, BW, TCP Variants, NS-2, TCL and OTCL.

I. Introduction

A Random Early Detection (RED) is the first active queue management algorithm proposed for deployment in TCP/IP networks. The basic idea behind an active queue management algorithm is to convey congestion notification early to the TCP end points so that they can reduce their transmission rates before queue overflow and sustained packet loss occur. "It is now widely accepted that the RED controlled queue performs better than a drop-tail queue. It is an active queue management algorithm" [1]. "The tail drop algorithm, a router buffer as many packets as it can, and drops the packet when it cannot buffer. If buffers are constantly full, the network is congested" [2]. RED addresses these issues. It monitor the average queue size and drops packets based on statistical probabilities. If the buffer is almost empty, all incoming packets reaccepted. As the queue grows, the probabilities for dropping incoming packet are dropped too. RED is more fair than trail drop in the sense of it does not possess a bias against burst traffic that use only a small portion of the bandwidth. The more the more a host transmits, likely it is that packets are dropped. The most common technique of queue management is a trail drop. In this method packets are accepted as long as there is space in the buffer when it becomes full, incoming packets are dropped. This approach results in dropping large number of packets in the time congestion. This can result in lower throughput and TCP synchronization [3]. However TCP includes eleven variants (Tahoe, FullTcp, TCP/Asym, Reno, Reno/Asym, Newreno/Asym, Sack1, DelAck and Sack1/DelAck) as source and five (TCPSink, TCPSink/Asym, Sack1, DelAck and Sack1/DelAck) as destination, implementation in NS-2 [4, 5]. The base TCP has become known as TCP Tahoe. TCP Reno attaches one novel mechanism called Fast Recovery to TCP Tahoe [4]. In addition, TCP Newreno employs the most recent retransmission mechanism of TCP Reno. [6]. The use of Sacks allows the receiver to stipulate several additional data packets that have been received out-of-order within one dupack, instead of only the last in order packet received [5]. TCP Vegas offers its own distinctive retransmission and congestion control strategies. TCP Fack is Reno TCP with forward acknowledgment [7]. Transmission Control Protocol (TCP) Variants Reno, NewReno, Vegas, Fack and Sack1 are implemented in NS-2. RED supervises the average queue size and drops packets based on statistical likelihoods [3].

Author[α] : Bachelor of Science in Engineering degree from Department of Computer Science and Engineering (CSE), Dhaka University of Engineering & Technology (DUET), Gazipur, Bangladesh.
PH-8801717871852. E-mail : aamcse@gmail.com
Author[Ω] : Lecturer, Department of CSE, DUET and she achieved her B. Sc. in Engineering degree from Department of Computer Science and Engineering (CSE),DUET, Bangladesh.
PH-8801913184657. E-mail : momotaz03-duet@yahoo.com
Author[β] : Bachelor of Science in Engineering degree from Department of Computer Science and Engineering (CSE), Dhaka University of Engineering & Technology (DUET), Gazipur, Bangladesh.
PH-+8801715543754. E-mail : mridul.duet@yahoo.com
Author[ψ] : Bachelor of Science in Engineering degree from Department of Computer Science and Engineering (CSE), Dhaka University of Engineering & Technology (DUET), Gazipur, Bangladesh.
PH-+8801719972044. E-mail : rubel_cse_064056@yahoo.com

II. Random Early Detection

a) RED Parameter Setting

Average queue size *avg* is formulated [1] as:

$$avg \leftarrow (1 - wq) \times avg + w_q \times q \qquad (I)$$

Where, *wq* is the queue weight, *q* is current queue size. *wq* should have lower value for bustier traffic; more weight is given in this case for the historic

size of the queue. As *avg* varies from *minth* to *maxth*, the packet-marking probability *pb* varies linearly from *0* to *maxp*..

$$p_b \leftarrow \frac{max_p \times (avg - min_{th})}{max_{th} - min_{th}} \qquad \text{(II)}$$

The final packet-marking probability p_a increases slowly as the count increases since the last marked packet [1]:

$$p_a \leftarrow \frac{p_b}{1 - count \times p_b} \qquad \text{(III)}$$

III. Performance Analysis of Red Model

a) *Variation in Threshold Value*

Table 1 : Number received packet for various TCP variants with respect to threshold for simulation time 70s

TCP variants	15	20	25	30	35
Reno	854	1185	845	711	733
Newreno	721	763	752	774	741
Vegas	821	777	685	686	625
Fack	800	721	713	644	761
Sack1	864	870	749	813	786

Table 2 : Number received packet for various TCP variant with respect to threshold for simulation time 140s

TCP variants	15	20	25	30	35
Reno	1452	1532	1333	1778	1398
Newreno	1458	1465	1501	1631	1538
Vegas	1345	1578	1350	1498	1538
Fack	1412	1754	1252	2379	1422
Sack1	1501	1339	1595	1358	1179

Table 3 : Number received packet for various TCP variants with respect to threshold for simulation time 210s

TCP variants	15	20	25	30	35
Reno	2659	2635	2376	1946	2300
Newreno	2701	2546	2032	2169	2303
Vegas	2254	2255	2301	2432	2178
Fack	2802	2462	2897	2131	2376
Sack1	2269	2416	2201	2554	2082

Table 4 : Number received packet for various TCP variant with respect to threshold for simulation time 270s

TCP variants	15	20	25	30	35
Reno	3142	3403	3312	3323	2902
Newreno	3383	3220	3204	3265	2928
Vegas	2624	2749	2778	2538	2799
Fack	3545	3088	2856	2681	4298
Sack1	3888	3216	3051	3232	3409

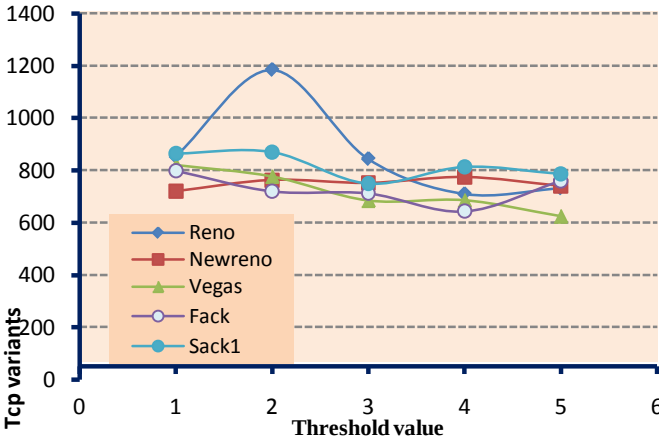


Figure1 : Graph of received packet for various TCP variants with respect to threshold for simulation time 70s

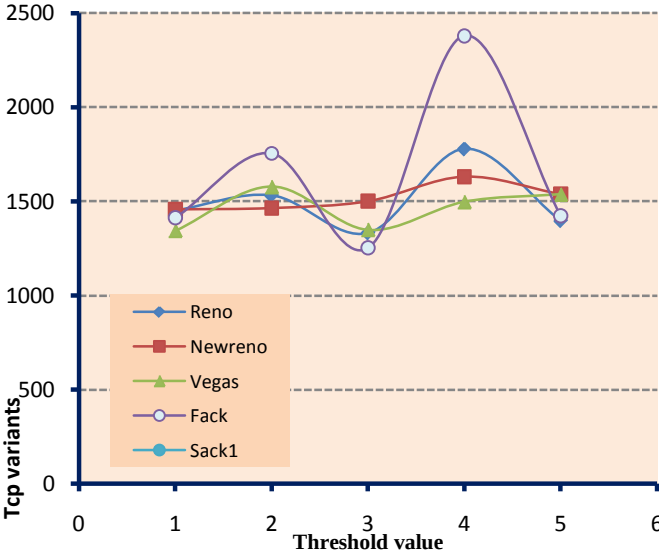


Figure2 : Graph of received packet for various TCP variants with respect to threshold for simulation time 140s

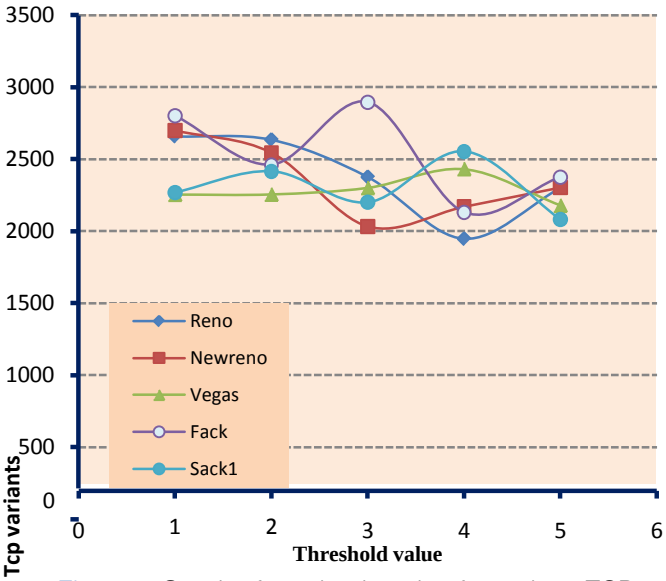


Figure3 : Graph of received packet for various TCP variants with respect to threshold for simulation time 210s

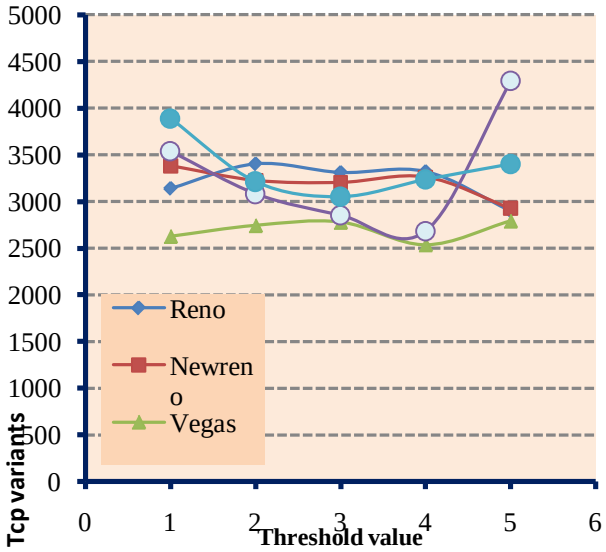


Figure4 : Graph of received packet for various TCP variants with respect to threshold for simulation time 280s

b) Performance Comparison

We that when threshold increase then variation course in received among various TCP variants and all arriving packets are received when average queue size exceeds max threshold or less than minimum threshold then packets are dropped which is shown in above all tables and corresponding figure. We found that Newreno TCP variants is the best because mean number of received packet is high mean number of dropped packet is low.

c) Comparison of TCP and UDP

i. *Comparison of Received Packet*

Table 5 : Comparison of received packet between UDP and TCP

Times		70s	140s	210s	280s
U	15	675	1294	1996	2586
	20	797	1222	1803	2694
D	25	758	1187	2127	2633
	30	795	1484	2085	2794
P	35	749	1336	1963	2783
T	15	566	1352	2725	2457
	20	665	1606	2374	3284
C	25	637	1438	2425	3694
	30	548	1656	2247	2832
P	35	834	1614	2413	3438

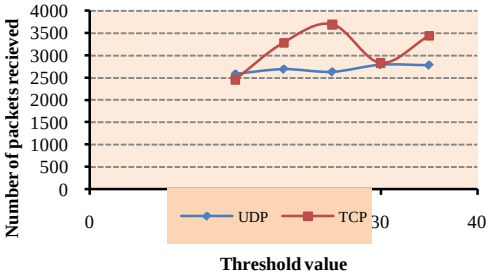


Figure5 : comparison graph of received packet between UDP and TCP for simulation time 280s

ii. *Comparison of Dropped Packet*

Table 6 : Comparison of received packet between UDP and TCP

Times		70s	140s	210s	280s
U	15	25	126	246	374
	20	67	104	113	354
D	25	24	53	426	696
	30	135	113	162	344
P	35	36	34	433	357
T	15	0	26	37	73
	20	0	14	33	46
C	25	0	5	36	43
	30	0	8	23	17
P	35	0	4	14	12

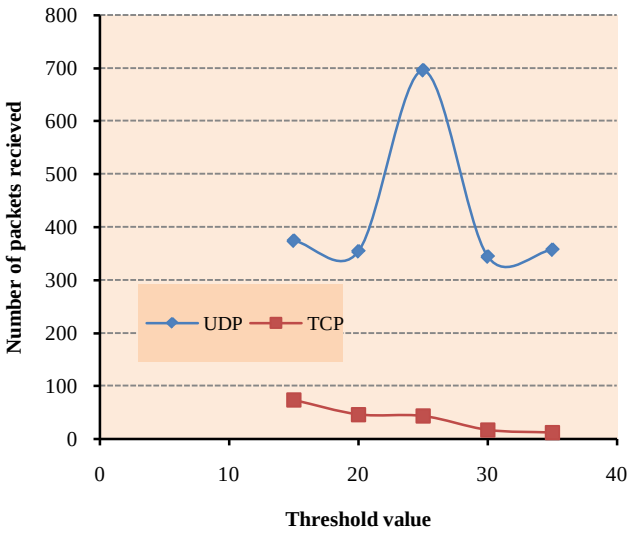


Figure6 : comparison graph of dropped packet between UDP and TCP for simulation time 280s

IV. Conclusion

From the aforementioned comparison of the performance it is found that TCP is better than UDP because packet received is higher in it with respect to UDP. That is why packet loss is lower in TCP. In case of packet drop, it is clear those packet drop is higher in UDP than TCP and also occur more congestion in it. It is possible to control congestion in TCP using RED model.

References References Referencias

1. S. FLOYD and V. JACOBSON: „Random Early Detection gateways for congestion avoidance ". *IEEE/ACM Transactions on Networking*, 1993, 1 (4), pp. 397-413.
2. MD. SHOHIDUL ISLAM, MD. NIAZ MORSHED, MD. SHARIFUL ISLAM AND MEJBAHUL AZAM: "An experimental analysis of random early discard (RED) queue for congestion control", 2011.
3. S. Floyd. RED: Discussions of setting parameters. *http://www.aciri.org/floyd/REDparameters.txt*, 1997.
4. NS, The network simulator-ns-2.27 URL:http://www.isi.edu/nsnam/ns
5. "The ns Manual", The VINT Project, A Collaboration between researchers at UC Berkeley, LBL, USC/ISI, and Xerox PARC.
6. Tanja Lang, (2002), "Evaluation of different TCP versions non-wireline environments", *The University of South Australia, Institute for Telecommunications Research*.
7. Merida, Venezuela and ESSI, Sophia-Antipolis (2003), "NS Simulator for beginners", Lecture notes, 2003-2004, Univ. de Los Andes, France.
8. C.V. HOLLOT, VISHAL MISRA, DON TOWSLEY AND WEI-BO GONG: A Control Theoretic Analysis of RED. *IEEE INFOCOM 2001.*

Global Journal of Computer Science and Technology
Volume 11 Issue 18 Version 1.0 October 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Clustering Method for categorical and Numeric Data sets

By Simmi Bagga, Dr. G.N. Singh

Sant Hira Dass Kanya Maha Vidyalaya, Kala Sanghian, Distt Kpt. Punjab, India

Abstract - Many issues concerned with clustering process are due to large datasets involves. In clustering computation become expensive when there are large data sets involved and work efficiently when there is limited number of cluster with relatively small data set. This paper will present a new technique for clustering for large datasets. That will work efficiently equally with large data set as well as with small data sets. The main idea behind this method is to divide the whole process in two steps. The first step uses a cheap approximate distance measure that divide the data into overlapped subsets we call it stubs. Then in second step clustering is performed for measuring exact distances only between points that occur in common stubs. The stub based clustering approach reduces computation time over a traditional clustering and also increases its efficiency.

Keywords : *Clustering,stubs, cetroid, K-Means Clustering.*

GJCST-C Classification : *I.5.3*

Strictly as per the compliance and regulations of:

Clustering Method for categorical and Numeric Data sets

Simmi Bagga[α], Dr. G.N. Singh[Ω]

Abstract - Many issues concerned with clustering process are due to large datasets involves. In clustering computation become expensive when there are large data sets involved and work efficiently when there is limited number of cluster with relatively small data set. This paper will present a new technique for clustering for large datasets. That will work efficiently equally with large data set as well as with small data sets. The main idea behind this method is to divide the whole process in two steps. The first step uses a cheap approximate distance measure that divide the data into overlapped subsets we call it stubs. Then in second step clustering is performed for measuring exact distances only between points that occur in common stubs. The stub based clustering approach reduces computation time over a traditional clustering and also increases its efficiency.

Keywords: *Clustering, Stubs, Cetroid K-Means Clustering.*

I. Introduction

From the last few years, the development of Information technology become very fast. So there has been widespread change in the adoption and utilization of new technologies in every field. That generated huge amount of data in the various fields. So there has been widespread change in the adoption and utilization of new technologies in every field that generated huge amount of data in the various fields.

To handle such large amount of data having complex structure effectively and efficiently for decision making, we use the concept of Data Mining. Data Mining is the process of identifying valid, useful and understandable patterns in data. There are various techniques have been developed and used in Data Mining including association, classification, clustering, prediction and sequential patterns etc. Clustering techniques are used for combining group that are similar to each other. Each cluster should be different from other clusters.

Clustering is one of the oldest and effective techniques of Data Mining. Basically clustering algorithms are divided in to two types these are Partitional and Hierarchical.

Author α : Assistant Professor, Sant Hira Dass Kanya Maha Vidyalaya, Kala Sanghian, Distt Kpt.(Punjab), India.
Email : simmibagga12@gmail.com
Author Ω : Department of Physics and Computer Science, Sudarshan Degree College, Lalgaon, Rewa(M.P), India.

a) Partitional clustering algorithms

In Partitional clustering algorithm firstly they compute partitions of the data based on similarity of data, and then they chose one that optimizes the criteria. These kinds of algorithms are highly complexity.

b) Hierarchical Clustering Algorithms

Hierarchical clustering algorithm creates hierarchical decomposition of the objects set. They are either agglomerative or divisive:

(a) Agglomerative algorithm starts with object as a separate cluster and then merges group according to a distance measure. These algorithm stops when all object converts in to single object or at the point where user wants to stop. These algorithms follow bottom-up merging.

(b) Divisive algorithms works opposite to agglomerative strategy. They start with one group of all objects and then split groups into smaller ones. This process repeats until each object falls in one cluster, or according to the user desire. These types of algorithm follows top down merging.

Further there are various categories of clustering algorithm. These categories are mainly focused on specific kind of data set or with some specific problems.

i. *Density-Based Clustering:*

Density Based clustering algorithms group objects according to the functions that deal with density objectives. Density is defined as number of objects resides near the data objects. In this approach cluster grows longer as the density increases i.e. number of object in neighborhood increases. These are mainly hierarchical in nature.

ii. *Grid-Based Clustering:*

Grid based Clustering algorithms deals with the spatial data i.e. about the objects related to space. Spatial data includes structure of objects in space, its relationships and properties. In these types of algorithm, we quantize data in to cells. Then we with only with those objects that's belongs to cell.

iii. *Model-Based Clustering:*

These algorithm deals with the approximations of model i.e. deals with the various parameters of the model that best fit the data. These types of methods can either be partitional or hierarchical depends on the

model. They are also very closer to density-based algorithms.

iv. *Categorical Data Clustering:*

These algorithms are mainly developed for data where numerical-oriented, distance measures cannot be applied. These are very close to both partitional and hierarchical methods.

Clustering algorithms become computationally expensive when the data set is to be large. There are mainly three reasons in which the data set can be large:

- When large number of elements in the data set involved
- When data element can have many features.
- When many clusters to be discover from the data set.

Recent cluster algorithms have focused on the efficiency issues. K-means clustering algorithm is very efficient when we start by finding good initial points, but is not efficient when the size of cluster become large. There is no such algorithm works efficiently when any of above three kinds of data is there in data set because in that case we have millions of data element with many features and cluster.

In this paper we introduce a new clustering technique that work well even in the case of large data set and large clusters. The main idea of this new technique is to perform clustering in two steps. In first step, divide the data into overlapping subsets called stubs, then in final step in which expensive distance measured only among points that occur in a common stub. First step can be performed very quickly and roughly where as second step is little rigorous. During the first step we mainly built stub by using approximate distance measure, the second step can be performed by any standard algorithm of clustering. The computation time and complexity is saved by the approximate distance measure used to create stubs. This computation is saved by eliminating all of the distance comparisons among the data set. We have found small accuracy increases due to the usage of two different distance measures. Clustering based on stubs applied to many different standard clustering algorithms, like K-means, Greedy Agglomerative Clustering and Expectation-Maximization.

II Stub Based Clustering

The basic idea behind the stub based algorithm is that one can reduce the number of distance computations for clustering. The first step cheaply partitioning the data set into overlapping subsets called stubs and then in second step the distance is measuring only among the data points that belong to a common subset. The stub based technique uses two different sources to cluster items: a cheap and approximate similarity measure. For example, in first step we just measure approximate distances among the data set like proportional comparison between data set and calculate the similarities between the data set that will reduce the size of the data set up to some extent. This step can be performed cheaply and in little time. In the final step, the more accurate similarity measures are performed that is more expensive in nature. In this step detailed distance measurement is performed.

We divide this clustering process into two steps. In the first step we use the short distance measure to create some number of overlapping subsets, called stubs. These are calculated by the proportionate measure. Stub is just a subset of the data elements find by approximate measure of its similarities. These distances are the distance performed from a central point. A data element can appear in more than one subset or stub but every data element must belong to at least one stub. Stubs are created with the intention that data element appearing in common stub may be far different that they could not possibly be in the same cluster. The method used for calculating the distance to create stubs is approximate. There exist many overlapping stubs in the data set, because we choose a large enough distance to ensure each and every data element should belong to any stub. These stubs are just made by measuring approximate distance and it's a very cheap method of calculating and it reduces the size of data set.

In the second step, we perform traditional clustering algorithm on that filtered stubs, such as K-means, Greedy Agglomerative Clustering or by using any accurate distance measure algorithm. The main restriction impose in this method is that we do not calculate the distance between two points that do not belong to same stub. This restriction is imposed because we assume the distance between the two different stubs to be infinite. The expensive distance measurements will only be made between the same stubs. This is will overall reduce the number of calculation.

If the first step is not properly performed that is if stubs are not properly made then it degrade the performance of the second step also. So stubs should be created carefully. If stubs are not too large and do not overlap much, then we cannot avoid expensive computation for clustering. The constraints imposed on the clustering imposed by the stubs may not lose accuracy but will increase computation efficiently. If distance to a cluster is measured to the centroid of the cluster, then clustering accuracy will be preserved exactly.

For every cluster, there exists a set of stubs. Expensive distance measurements will only made between pairs of data points in the same stubs.

a) Creating Stubs

In the case of stub based clustering, user will be able to focus domain-specific features in order to design a short distance measures. User efficiently creates stubs using these measures. For example, if we have large data of patients of number of hospitals that contain information of diagnosis, treatments and payment histories. A cheap measure calculates the similarity between diagnoses of the patients. Result might be 1, if they have the similarities and 0 if they do not have any similarity. In this case stub creation is small and the common diagnosis results fall in the same stub. If the same patient falls in multiple diagnoses then he will fall into multiple stubs and also some stubs will overlap. The small number features are sufficient to built stub, even if the data item may have thousands of features.

b) Cheap Distance Measurement

There are various methods to calculate the cheap distance measure. One of the methods for distance measure for text is based on the inverted index. An inverted index is in the form of sparse matrix in which, each word can directly access the list of documents containing that word. When we want to find all documents according to some, we need not to measure the distance to all documents, but only have to examine the list of documents associated with each word in the query. The documents which have no words in common with the query will never be considered. The inverted index can efficiently calculate a distance metric.

Using the above distance measure stub can creates as follows. Start with a list of the data items, and with two distance thresholds, lets say T1 and T2, where T1 > T2. Pick a data point from the list and approximately measure distance to all points. Put all similar data points to the distance threshold T1 into a stub. Remove from the list all points that are within distance threshold T2. Repeat until the list is empty. The inverted index can be applied to real-valued data item.

c) K-Means Stubs

One can also use the stubs idea to speed up prototype based clustering methods like K-means. K-Mean is well known partitioning based method of clustering. It is simple and iterative method works around one artificial point which represent the average location of the cluster is called Centroid.

This algorithm takes an input a number of clusters that is the k from. Means is an average location of all the members of a cluster. In this algorithm we have to partition n object set to k clusters. Cluster is measured on the basis of its mean or average location. The basic idea behind this algorithm is to first randomly select k of the object that is centroid of cluster. Using this k of clusters, we optimize intra cluster similarity and inter-cluster dissimilarity. Each remaining object, the most similar object is assigned to cluster based the distance between object and centroid or cluster mean. Then we compute new mean. This process repeats until the whole function overages. This is an iterative method in which we always redefine the center point or centroid until cluster detection is finished.

This approach is basically based on prototypes that are associated with the stubs that contain them. The prototypes are only influenced by data inside the associated stubs. After creating the stubs, we decide how many prototypes will be created for each stub. Then we place prototypes into each stub. For each prototype, we find the stubs that contain in it (computed by using cheap distance measures) and then calculate distances from that prototype to points within those stubs.

K-means algorithm not gives just similar results for stub. In K-means each data point is assigned to a single prototype. As long as the cheap and expensive distance measures are sufficiently similar that the nearest prototype is within the boundaries of the stubs that contain that data point, then the same prototype will win.

III. Computational Complexity

We can simply say that computational time is saving using stub based method. This technique has two step and mainly done two types of comparisons. One is relative fast step where stubs are created. The other one is a slow clustering process in which we apply K-Mean clustering algorithm. If we create stubs by using inverted index method then there is no need to perform pair wise distance comparison.

In the case of K-means or Expectation-Maximization, clustering without stubs requires O (nk) distance comparisons per iteration of clustering. Consider the K-mean method with stubs where each cluster belongs to one or more stubs. Assume that clusters have the same overlap factor f as data points do. Then, each cluster needs to compare itself to the fn/c points in f different stubs.

IV. Conclusion

Clustering large data sets having large cluster is a very tedious task. It is very expensive and inefficient to deal with large data set having large number of clusters using traditional clustering methods. The goal of this paper is to describe a new stub based method for clustering that takes relatively less computation time and performs result more effectively in the case of large data set. In this paper we described how can we create stubs and how can we apply traditional cluster methods like K-mean clustering method to perform effective clustering.

References Références Referencias

1. M. R. Anderberg: Cluster Analysis for Application. Academic Press, 1973.
2. Hartigan, John A, "Clustering Algorithms". 1975. John Wiley. New York.
3. Han Jiawei and Kamber, Micheline. "Data Mining: Concepts and Techniques". 2001. Morgon Kaufmann. Sanfransico, CA.
4. Grossman, Robert. L., Hornick, Mark. F. and Meyer, Gregor. "Data Mining Standards Initiatives." (Communications of the ACM, Vol. 45, No. 8, 2002,59-61).
5. Z. He, X Xu, S Deng: An Efficient Algorithm for Clustering Categorical Data, Journal of Computer Science and Technology(2002), vol 17.
6. Tajunisha, Saravanan: An efficient method to improve the clustering performance for high dimensional data by Principal Component Analysis and modified K-means, International Journal of Database Management Systems (IJDMS), Vol.3, 2011.
7. Singh G.N, Bagga Simmi: Three Phase Iterative Model of KDD, International Journal of Information Technology and Knowledge Management(2011), Volume 4, No. 2, pp. 695-697.

Global Journal of Computer Science and Technology
Volume 11 Issue 18 Version 1.0 April 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Comparison of Time Taken and Compression Efficiency for Different Sizes of Databases

By Dr. Mrs Pushpa Suri, Mrs. Meenakshi Sharma
Kurukshetra University

Abstracts - Data compress for object oriented data warehousing. A data warehouse is an essential component to the decision support system. The traditional data warehouse provides only numeric and character data analysis. But as information technologies progress, complex data such as semi-structured and unstructured data become vastly used. Data Compression is of interest in business data warehousing, both because of the cost saving it offers and because of the large volume of data manipulated in many business application.[3],[5]. The entropy is used in many areas such as image processing, document images. But in our research we used the entropy in object oriented data warehousing. Creation of different sizes of databases in oracle. Employment of object oriented programming for compression using Datawarehousing.

Keywords : *Data warehousing, Data compression, Object oriented, Entropy.*

GJCST-F Classification : *C.2.4, E.4*

Strictly as per the compliance and regulations of:

Comparison of Time Taken and Compression Efficiency for Different Sizes of Databases

Dr. Mrs Pushpa Suri[α], Mrs. Meenakshi Sharma[Ω]

Abstract - Data compress for object oriented data warehousing. A data warehouse is an essential component to the decision support system. The traditional data warehouse provides only numeric and character data analysis. But as information technologies progress, complex data such as semi-structured and unstructured data become vastly used. Data Compression is of interest in business data warehousing, both because of the cost saving it offers and because of the large volume of data manipulated in many business application.[3],[5]. The entropy is used in many areas such as image processing, document images. But in our research we used the entropy in object oriented data warehousing. Creation of different sizes of databases in oracle. Employment of object oriented programming for compression using Datawarehousing.

Keywords : Data warehousing, Data compression, Object oriented, Entropy.

I. Introduction

One of the hottest topics in the industry today is data warehousing and on-line analytical processing (OLAP). Although, data warehousing has been around in some form or another since the inception of data storage, people were never able to exploit the information that was wastefully sitting on a tape somewhere in a back room. Today, however, technology has advanced to a point to make access to this information an interactive reality. Organizations across the country and around the world are seeking expertise in this exploding field of data organization and manipulation. It is not a surpise, really, that business users want to get a better look at their data. Today, business opportunities measure in days, instead of months or years, and the more information empowering an entrepreneur or other business person, the better the chances of beating a competitor to the punch with a new product or service. The task of transitioning from a procedural mindset to an object-oriented paradigm can seem overwhelming; however, the transition does not require developers to step into another dimension or go to Mars in order to grasp a new way of doing things. In many ways, the object-oriented approach to development more closely mirrors the world we've been living in all along: We each know quite a bit about objects already. It is that knowledge we must discover and leverage in transitioning to object-oriented tools and methodologies.

A data warehouse is a mechanism for data storage and data retrieval. Data can be stored and retrieved with a multidimensional structure--hypercube or relational, a star schema structure or several other data storage techniques.

II. Data Compression

Data compression is of interest in business data warehousing, both because of the cost savings it offers and because of the large volume of data manipulated in many business applications. The types of local redundancy present in business data files include runs of zeros in numeric fields, sequences of blanks in alphanumeric fields, and fields which are present in some records and null in others. Run length encoding can be used to compress sequences of zeros or blanks. Null suppression may be accomplished through the use of presence bits. Another class of methods exploits cases in which only a limited set of attribute values exist. Dictionary substitution entails replacing alphanumeric representations of information such as bank account type, insurance policy type, sex, month, etc. by the few bits necessary to represent the limited number of possible attribute values.

The problem of compressing digital data can be decoupled into two subproblems: modeling and entropy coding. Whatever the given data may represent in the real world, in digital form it exists as a sequence of symbols, such as bits. The modeling problem is to choose a suitable symbolic representation for the data and to predict for each symbol of the representation the probability that it takes each of the allowable values for that symbol. The entropy-coding problem is to code each symbol as compactly as possible, given this knowledge of probabilities. (In the realm of lossy compression, there is a third subproblem: evaluating the relative importance of various kinds of errors.)

For example, suppose if it is required to transmit messages composed of the four letters a, b, c, and d. A straightforward scheme for coding these messages in bits would be to represent a by \00", b by \01", c by \10" and d by \11". However, suppose if it is known that for any letter of the message (independent of all other letters), a occurs with probability .5, b occurs with probability .25, and c or d occur with probability

Author α : Associate Professor, Deptt. Of Computer Science. & Applications, Kurukshetra University, Kurukshetra, Haryana, India
Author Ω : Assistant Professor Deptt of Computer Science & Engineering in H.C.T.M, Kaithal, Haryana, India.
E-mail : minny_kaushik@yahoo.com

.125 each. Then a shorter representation might be chosen for a, at the necessary cost of accepting longer representations for the other letters. a could be represented by \0", b by \10", c by \110", and d by \111". This representation is more compact on average than the first one; indeed, it is the most compact representation possible (though not uniquely so). In this simple example, the modeling part of the problem is determining the probabilities for each symbol; the entropy-coding part of the problem is determining the representations in bits from those probabilities; the probabilities associated with the symbols play a fundamental role in entropy coding.

One well-known method of entropy coding is Huffman coding, which yields an optimal coding provided all symbol probabilities are integer powers of .5. Another method, yielding optimal compression performance for any set of probabilities, is arithmetic coding. In spite of the superior compression given by arithmetic coding, so far it has not been a dominant presence in real data-compression applications. This is most likely due to concerns over speed and complexity, as well as patent issues; a rapid, simple algorithm for arithmetic coding is therefore potentially very useful.

An algorithm which allows rapid encoding and decoding in a fashion akin to arithmetic coding is known as the Q-coder. The QM-coder is a subsequent variant. However, these algorithms being protected by patents, new algorithms with competitive performance continue to be of interest. The ELS algorithm is one such algorithm.

The ELS-coder works only with an alphabet of two symbols (0 and 1). One can certainly encode symbols from larger alphabets; but they must be converted to a two-symbol format first. The necessity for this conversion is a disadvantage, but the restriction to a two-symbol alphabet facilitates rapid coding and rapid probability estimation.

The ELS-coder decoding algorithm has already been described. The encoder must use its knowledge of the decoder's inner workings to create a data stream which will manipulate the decoder into producing the desired sequence of decoded symbols.

As a practical matter, the encoder need not actually consider the entire coded data stream at one time. One can partition the coded data stream at any time into three portions; from end to beginning of the data stream they are: preactive bytes, which as yet exert no in seuence over the current state of the decoder; active bytes, which affect the current state of the decoder and have more than one consistent value; and postactive bytes, which affect the current state of the decoder and have converged to a single consistent value. Each byte of the coded data stream goes from preactive to active to postactive; the earlier a byte's position in the stream, the earlier these transitions occur. A byte is not actually moved to the external _le until it becomes postactive. Only the active portion of the data stream need be considered at any time. Since the internal buffer of the decoder contains two bytes, there are always at least two active bytes. The variable backlog counts the number of active bytes in excess of two. In theory backlog can take arbitrarily high values, but higher values become exponentially less likely.

III. Related Work

2-D Compression of ECG Signals Using ROI Mask and Conditional Entropy Coding," have given a novel 2-D compression scheme which employs 1-D discrete wavelet transform, the region of interest mask, and the conditional entropy coding based on context models. Experimental results on records selected from the Massachusetts Institute of Technology-Beth Israel Hospital arrhythmia database show that the proposed method outperforms some existing compression schemes [5]. Lossless Compression Using Conditional Entropy-Constrained Subband Quantization," have proposed Lossless Compression Using Conditional Entropy-Constrained Subband Quantization [13]. Sang et al in their paper "A novel approach to scene change detection using a cross entropy ," have shown that in huge video databases, an effective video indexing method is required. While manual indexing is the most effective approach to this goal, it is slow and expensive. Thus automatic indexing is desirable, and previously various indexing tools for video databases have been developed. For efficient video indexing and retrieval, the similarity measure is an important factor. This paper presents new similarity measures between frames and proposes a new algorithm to detect scene changes using a cross entropy defined between two histograms. Experimental results show that the proposed algorithm is fast and effective compared with several conventional algorithms to detect abrupt scene changes and gradual transitions including fade in/out and flash light scenes [12].

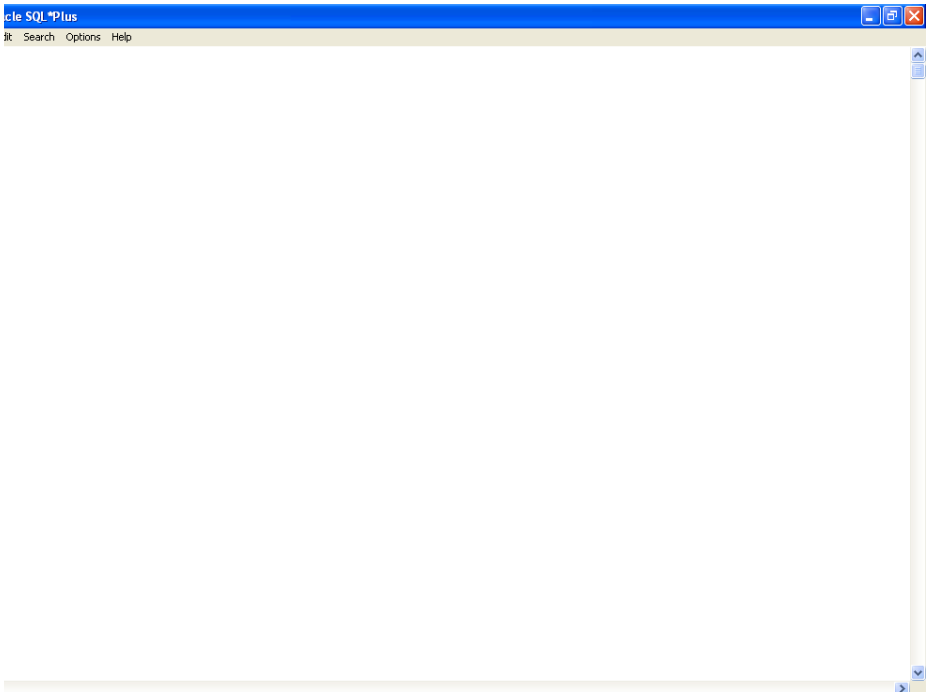
IV. Objective

The objective of the present study is to

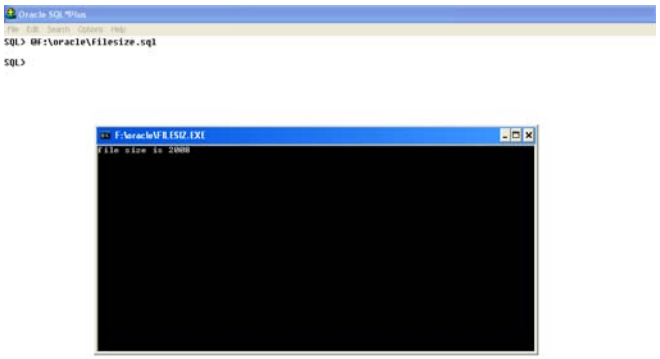
1. Develop data of compression for object oriented data warehousing.
2. Devise efficient compression algorithms in data warehousing to enhance the efficiency of the data warehousing packages so that less CPU time and less Memory is consumed.

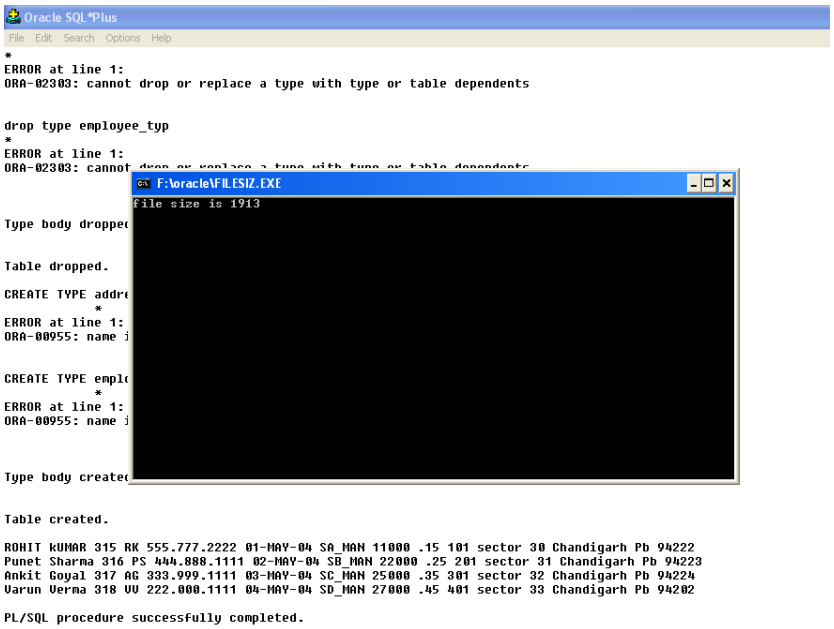
Implement compressor and expander using entropy algorithm and test its effectiveness on different sized databases

V. Results



Main File size





VI. Conclusion

In this paper we have discuss the data compression and how the data is compresses in oracle 10g using object oriented language. Data Compression is of interest in business data warehousing, both because of the cost saving it offers and because of the large volume of data manipulated in many business application. The entropy is used in many areas such as image processing, document images. But in our research we used the entropy in object oriented data warehousing. Creation of different sizes of databases in oracle. Employment of object oriented programming for compression using data warehousing. Further compression of database .csv files using C++. Comparison of time taken and compression efficiency for different sizes of databases

References Références Referencias

1. Wei-Chou Chen; Tzung-Pei Hong; Wen-Yang Lin; , "Using the compressed data model in object-oriented data warehousing ," Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference on , vol.5, no., pp.768-772 vol.5, 1999.
2. Wei-Chou Chen; Tzung-Pei Hong; Wei-Yang Lin; , "A composite data model in object-oriented data warehousing," Technology of Object-Oriented Languages and Systems, 1999. TOOLS 31. Proceedings , vol., no., pp.400-405, 1999.
3. Shieh, J. C.; Lin, H. W.,"The Novel Model of Object-Oriented Data Warehouses", Workshop on Databases and Software Engineering, 2006.
4. Chen Wei-Chou, Hong Tzung-Pei (2), Lin Wen-Yang "Three maintenance algorithms for compressed object-oriented data warehousing."
5. Boqiang Huang; Yuanyuan Wang; Jianhua Chen; , "2-D Compression of ECG Signals Using ROI Mask and Conditional Entropy Coding," Biomedical Engineering, IEEE Transactions on , vol.56, no.4, pp.1261-1263, April 2009.
6. Gong, Y.; Fan, M.K.H.; Huang, C.-M.; , "On entropy-constrained residual vector quantization design," Data Compression Conference, 1999. Proceedings. DCC '99 , vol., no., pp.526, 29-31 Mar 1999.
7. De, I.; Sil, J.; , "Wavelet entropy based no-reference quality prediction of distorted/decompressed images," Computer Engineering and Technology (ICCET), 2010 2nd International Conference on , vol.3, no., pp.V3-245-V3-250, 16-18 April 2010.
8. De, I.; Sil, J.; , "ANFIS tuned no-reference quality prediction of distorted/decompressed images featuring wavelet entropy," Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference on , vol., no., pp.126-131, 8-10 Oct. 2010.
9. Liu, L.; Dong, Y.; Song, X.; Fan, G.; , "An entropy based segmentation algorithm for computer-generated document images," Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on , vol.1, no., pp. I- 541-4 vol.1, 14-17 Sept. 2003.
10. Tu, C.; Tran, T.D.; , "Context-based entropy coding of block transform coefficients for image compression," Image Processing, IEEE Transactions on , vol.11, no.11, pp. 1271- 1283, Nov 2002.
11. Chen, S.; Reif, J.H.; , "Using difficulty of prediction to decrease computation: fast sort, priority queue and convex hull on entropy bounded inputs," Foundations of Computer Science, 1993. Proceedings., 34th Annual Symposium on , vol., no.,

pp.104-112, 3-5 Nov 1993.
12. Sang Hyun Kim; Rae-Hong Park; , "A novel approach to scene change detection using a cross entropy ," Image Processing, 2000. Proceedings. 2000 International Conference on , vol.3, no., pp.937-940 vol.3, 2000.
13. Scales, A.; Roark, W.; Kossentini, F.; Smith, M.J.T.; , "Lossless Compression Using Conditional Entropy-Constrained Subband Quantization," Data Compression Conference, 1995. DCC '95. Proceedings , vol., no., pp.498, 28-30 Mar 1995.
14. Jegou, H.; Guillemot, C.; , "Entropy coding with variable length re-writing systems," Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on , vol., no., pp.1529-1533, 4-9 Sept. 2005.
15. Hua Xie; Ortega, A.; , "Entropy- and complexity-constrained classified quantizer design for distributed image classification," Multimedia Signal Processing, 2002 IEEE Workshop on , vol., no., pp. 77- 80, 9-11 Dec. 2002.
16. Ambhore, P.B. (August 20-22, 2007), A Implementation of Object Oriented Database Security, in 5th ACIS International Conference on Software Engineering Research (2007), Institute of Electrical and Electronic Engineers, IEEE Computer Society.
17. Park M (1996), Data Warehouse Designing on Relational Database Systems, Informix Co., Stanford.
18. Molina G H (1995), Maintenance in Data Warehousing Environment, San Jose Co., California.
19. Roussopoulos N (1997), Data Warehouses and Materialized Views, Leander Press, Greece.
20. Swift R (1996), Building Advanced Data Warehouse, NCR Corporation, California.
21. Marlin E S (1992), ODBMS vs. Relational Object-Oriented Programming, SAGE, London.
22. Jae Jin Koh,(3-6 October, 2007), Relational database schema integration by overlay and redundancy elimination methods, in International Forum on Strategic Technology(2007), Institute of Electrical and Electronic Engineers, IEEE Computer Society.
23. Michael S (1991), The next generation DBMS, Pearson Education, New York.
24. Bertino E.(1991): Method precomputation in object–oriented databases. Proceedings of ACM– SIGOIS and IEEE–TC–OA International Conference on Organizational Computing Systems.
25. Eder J., Frank H., Liebhart W. (1994): Optimization of Object–Oriented Queries by Inverse Methods. Proceedings of East/West Database Workshop, Austria.

Global Journals Inc. (US) Guidelines Handbook 2011

www.GlobalJournals.org

FELLOW OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (FARSC)

- 'FARSC' title will be awarded to the person after approval of Editor-in-Chief and Editorial Board. The title 'FARSC" can be added to name in the following manner. eg. **Dr. John E. Hall, Ph.D., FARSC or William Walldroff Ph. D., M.S., FARSC**
- Being FARSC is a respectful honor. It authenticates your research activities. After becoming FARSC, you can use 'FARSC' title as you use your degree in suffix of your name. This will definitely will enhance and add up your name. You can use it on your Career Counseling Materials/CV/Resume/Visiting Card/Name Plate etc.
- 60% Discount will be provided to FARSC members for publishing research papers in Global Journals Inc., if our Editorial Board and Peer Reviewers accept the paper. For the life time, if you are author/co-author of any paper bill sent to you will automatically be discounted one by 60%
- FARSC will be given a renowned, secure, free professional email address with 100 GB of space eg.johnhall@globaljournals.org. You will be facilitated with Webmail, Spam Assassin, Email Forwarders, Auto-Responders, Email Delivery Route tracing, etc.
- FARSC member is eligible to become paid peer reviewer at Global Journals Inc. to earn up to 15% of realized author charges taken from author of respective paper. After reviewing 5 or more papers you can request to transfer the amount to your bank account or to your PayPal account.
- Eg. If we had taken 420 USD from author, we can send 63 USD to your account.
- FARSC member can apply for free approval, grading and certification of some of their Educational and Institutional Degrees from Global Journals Inc. (US) and Open Association of Research, Society U.S.A.
- After you are FARSC. You can send us scanned copy of all of your documents. We will verify, grade and certify them within a month. It will be based on your academic records, quality of research papers published by you, and 50 more criteria. This is beneficial for your job interviews as recruiting organization need not just rely on you for authenticity and your unknown qualities, you would have authentic ranks of all of your documents. Our scale is unique worldwide.
- FARSC member can proceed to get benefits of free research podcasting in Global Research Radio with their research documents, slides and online movies.
- After your publication anywhere in the world, you can upload you research paper with your recorded voice or you can use our professional RJs to record your paper their voice. We can also stream your conference videos and display your slides online.
- FARSC will be eligible for free application of Standardization of their Researches by Open Scientific Standards. Standardization is next step and level after publishing in a journal. A team of research and professional will work with you to take your research to its next level, which is worldwide open standardization.

- FARSC is eligible to earn from their researches: While publishing his paper with Global Journals Inc. (US), FARSC can decide whether he/she would like to publish his/her research in closed manner. When readers will buy that individual research paper for reading, 80% of its earning by Global Journals Inc. (US) will be transferred to FARSC member's bank account after certain threshold balance. There is no time limit for collection. FARSC member can decide its price and we can help in decision.

MEMBER OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (MARSC)

- 'MARSC' title will be awarded to the person after approval of Editor-in-Chief and Editorial Board. The title 'MARSC" can be added to name in the following manner. eg. Dr. John E. Hall, Ph.D., MARSC or William Walldroff Ph. D., M.S., MARSC
- Being MARSC is a respectful honor. It authenticates your research activities. After becoming MARSC, you can use 'MARSC' title as you use your degree in suffix of your name. This will definitely will enhance and add up your name. You can use it on your Career Counseling Materials/CV/Resume/Visiting Card/Name Plate etc.
- 40% Discount will be provided to MARSC members for publishing research papers in Global Journals Inc., if our Editorial Board and Peer Reviewers accept the paper. For the life time, if you are author/co-author of any paper bill sent to you will automatically be discounted one by 60%
- MARSC will be given a renowned, secure, free professional email address with 30 GB of space eg.johnhall@globaljournals.org. You will be facilitated with Webmail, Spam Assassin, Email Forwarders, Auto-Responders, Email Delivery Route tracing, etc.
- MARSC member is eligible to become paid peer reviewer at Global Journals Inc. to earn up to 10% of realized author charges taken from author of respective paper. After reviewing 5 or more papers you can request to transfer the amount to your bank account or to your PayPal account.
- MARSC member can apply for free approval, grading and certification of some of their Educational and Institutional Degrees from Global Journals Inc. (US) and Open Association of Research, Society U.S.A.
- MARSC is eligible to earn from their researches: While publishing his paper with Global Journals Inc. (US), MARSC can decide whether he/she would like to publish his/her research in closed manner. When readers will buy that individual research paper for reading, 40% of its earning by Global Journals Inc. (US) will be transferred to MARSC member's bank account after certain threshold balance. There is no time limit for collection. MARSC member can decide its price and we can help in decision.

AUXILIARY MEMBERSHIPS

ANNUAL MEMBER

- Annual Member will be authorized to receive e-Journal GJMBR for one year (subscription for one year).
- The member will be allotted free 1 GB Web-space along with subDomain to contribute and participate in our activities.
- A professional email address will be allotted free 500 MB email space.

PAPER PUBLICATION

- The members can publish paper once. The paper will be sent to two-peer reviewer. The paper will be published after the acceptance of peer reviewers and Editorial Board.

Process of submission of Research Paper

The Area or field of specialization may or may not be of any category as mentioned in 'Scope of Journal' menu of the GlobalJournals.org website. There are 37 Research Journal categorized with Six parental Journals GJCST, GJMR, GJRE, GJMBR, GJSFR, GJHSS. For Authors should prefer the mentioned categories. There are three widely used systems UDC, DDC and LCC. The details are available as 'Knowledge Abstract' at Home page. The major advantage of this coding is that, the research work will be exposed to and shared with all over the world as we are being abstracted and indexed worldwide.

The paper should be in proper format. The format can be downloaded from first page of 'Author Guideline' Menu. The Author is expected to follow the general rules as mentioned in this menu. The paper should be written in MS-Word Format (*.DOC,*.DOCX).

The Author can submit the paper either online or offline. The authors should prefer online submission.Online Submission: There are three ways to submit your paper:

(A) (I) First, register yourself using top right corner of Home page then Login. If you are already registered, then login using your username and password.

(II) Choose corresponding Journal.

(III) Click 'Submit Manuscript'. Fill required information and Upload the paper.

(B) If you are using Internet Explorer, then Direct Submission through Homepage is also available.

(C) If these two are not convenient, and then email the paper directly to dean@globaljournals.org.

Offline Submission: Author can send the typed form of paper by Post. However, online submission should be preferred.

Preferred Author Guidelines

MANUSCRIPT STYLE INSTRUCTION (Must be strictly followed)

Page Size: 8.27" X 11'"

- Left Margin: 0.65
- Right Margin: 0.65
- Top Margin: 0.75
- Bottom Margin: 0.75
- Font type of all text should be Swis 721 Lt BT.
- Paper Title should be of Font Size 24 with one Column section.
- Author Name in Font Size of 11 with one column as of Title.
- Abstract Font size of 9 Bold, "Abstract" word in Italic Bold.
- Main Text: Font size 10 with justified two columns section
- Two Column with Equal Column with of 3.38 and Gaping of .2
- First Character must be three lines Drop capped.
- Paragraph before Spacing of 1 pt and After of 0 pt.
- Line Spacing of 1 pt
- Large Images must be in One Column
- Numbering of First Main Headings (Heading 1) must be in Roman Letters, Capital Letter, and Font Size of 10.
- Numbering of Second Main Headings (Heading 2) must be in Alphabets, Italic, and Font Size of 10.

You can use your own standard format also.
Author Guidelines:

1. General,

2. Ethical Guidelines,

3. Submission of Manuscripts,

4. Manuscript's Category,

5. Structure and Format of Manuscript,

6. After Acceptance.

1. GENERAL

Before submitting your research paper, one is advised to go through the details as mentioned in following heads. It will be beneficial, while peer reviewer justify your paper for publication.

Scope

The Global Journals Inc. (US) welcome the submission of original paper, review paper, survey article relevant to the all the streams of Philosophy and knowledge. The Global Journals Inc. (US) is parental platform for Global Journal of Computer Science and Technology, Researches in Engineering, Medical Research, Science Frontier Research, Human Social Science, Management, and Business organization. The choice of specific field can be done otherwise as following in Abstracting and Indexing Page on this Website. As the all Global

Journals Inc. (US) are being abstracted and indexed (in process) by most of the reputed organizations. Topics of only narrow interest will not be accepted unless they have wider potential or consequences.

2. ETHICAL GUIDELINES

Authors should follow the ethical guidelines as mentioned below for publication of research paper and research activities.

Papers are accepted on strict understanding that the material in whole or in part has not been, nor is being, considered for publication elsewhere. If the paper once accepted by Global Journals Inc. (US) and Editorial Board, will become the copyright of the Global Journals Inc. (US).

Authorship: The authors and coauthors should have active contribution to conception design, analysis and interpretation of findings. They should critically review the contents and drafting of the paper. All should approve the final version of the paper before submission

The Global Journals Inc. (US) follows the definition of authorship set up by the Global Academy of Research and Development. According to the Global Academy of R&D authorship, criteria must be based on:

1) Substantial contributions to conception and acquisition of data, analysis and interpretation of the findings.

2) Drafting the paper and revising it critically regarding important academic content.

3) Final approval of the version of the paper to be published.

All authors should have been credited according to their appropriate contribution in research activity and preparing paper. Contributors who do not match the criteria as authors may be mentioned under Acknowledgement.

Acknowledgements: Contributors to the research other than authors credited should be mentioned under acknowledgement. The specifications of the source of funding for the research if appropriate can be included. Suppliers of resources may be mentioned along with address.

Appeal of Decision: The Editorial Board's decision on publication of the paper is final and cannot be appealed elsewhere.

Permissions: It is the author's responsibility to have prior permission if all or parts of earlier published illustrations are used in this paper.

Please mention proper reference and appropriate acknowledgements wherever expected.

If all or parts of previously published illustrations are used, permission must be taken from the copyright holder concerned. It is the author's responsibility to take these in writing.

Approval for reproduction/modification of any information (including figures and tables) published elsewhere must be obtained by the authors/copyright holders before submission of the manuscript. Contributors (Authors) are responsible for any copyright fee involved.

3. SUBMISSION OF MANUSCRIPTS

Manuscripts should be uploaded via this online submission page. The online submission is most efficient method for submission of papers, as it enables rapid distribution of manuscripts and consequently speeds up the review procedure. It also enables authors to know the status of their own manuscripts by emailing us. Complete instructions for submitting a paper is available below.

Manuscript submission is a systematic procedure and little preparation is required beyond having all parts of your manuscript in a given format and a computer with an Internet connection and a Web browser. Full help and instructions are provided on-screen. As an author, you will be prompted for login and manuscript details as Field of Paper and then to upload your manuscript file(s) according to the instructions.

To avoid postal delays, all transaction is preferred by e-mail. A finished manuscript submission is confirmed by e-mail immediately and your paper enters the editorial process with no postal delays. When a conclusion is made about the publication of your paper by our Editorial Board, revisions can be submitted online with the same procedure, with an occasion to view and respond to all comments.

Complete support for both authors and co-author is provided.

4. MANUSCRIPT'S CATEGORY

Based on potential and nature, the manuscript can be categorized under the following heads:

Original research paper: Such papers are reports of high-level significant original research work.

Review papers: These are concise, significant but helpful and decisive topics for young researchers.

Research articles: These are handled with small investigation and applications

Research letters: The letters are small and concise comments on previously published matters.

5.STRUCTURE AND FORMAT OF MANUSCRIPT

The recommended size of original research paper is less than seven thousand words, review papers fewer than seven thousands words also.Preparation of research paper or how to write research paper, are major hurdle, while writing manuscript. The research articles and research letters should be fewer than three thousand words, the structure original research paper; sometime review paper should be as follows:

Papers: These are reports of significant research (typically less than 7000 words equivalent, including tables, figures, references), and comprise:

(a)Title should be relevant and commensurate with the theme of the paper.

(b) A brief Summary, "Abstract" (less than 150 words) containing the major results and conclusions.

(c) Up to ten keywords, that precisely identifies the paper's subject, purpose, and focus.

(d) An Introduction, giving necessary background excluding subheadings; objectives must be clearly declared.

(e) Resources and techniques with sufficient complete experimental details (wherever possible by reference) to permit repetition; sources of information must be given and numerical methods must be specified by reference, unless non-standard.

(f) Results should be presented concisely, by well-designed tables and/or figures; the same data may not be used in both; suitable statistical data should be given. All data must be obtained with attention to numerical detail in the planning stage. As reproduced design has been recognized to be important to experiments for a considerable time, the Editor has decided that any paper that appears not to have adequate numerical treatments of the data will be returned un-refereed;

(g) Discussion should cover the implications and consequences, not just recapitulating the results; conclusions should be summarizing.

(h) Brief Acknowledgements.

(i) References in the proper form.

Authors should very cautiously consider the preparation of papers to ensure that they communicate efficiently. Papers are much more likely to be accepted, if they are cautiously designed and laid out, contain few or no errors, are summarizing, and be conventional to the approach and instructions. They will in addition, be published with much less delays than those that require much technical and editorial correction.

The Editorial Board reserves the right to make literary corrections and to make suggestions to improve briefness.

It is vital, that authors take care in submitting a manuscript that is written in simple language and adheres to published guidelines.

Format

Language: The language of publication is UK English. Authors, for whom English is a second language, must have their manuscript efficiently edited by an English-speaking person before submission to make sure that, the English is of high excellence. It is preferable, that manuscripts should be professionally edited.

Standard Usage, Abbreviations, and Units: Spelling and hyphenation should be conventional to The Concise Oxford English Dictionary. Statistics and measurements should at all times be given in figures, e.g. 16 min, except for when the number begins a sentence. When the number does not refer to a unit of measurement it should be spelt in full unless, it is 160 or greater.

Abbreviations supposed to be used carefully. The abbreviated name or expression is supposed to be cited in full at first usage, followed by the conventional abbreviation in parentheses.

Metric SI units are supposed to generally be used excluding where they conflict with current practice or are confusing. For illustration, 1.4 l rather than 1.4 × 10-3 m3, or 4 mm somewhat than 4 × 10-3 m. Chemical formula and solutions must identify the form used, e.g. anhydrous or hydrated, and the concentration must be in clearly defined units. Common species names should be followed by underlines at the first mention. For following use the generic name should be constricted to a single letter, if it is clear.

Structure

All manuscripts submitted to Global Journals Inc. (US), ought to include:

Title: The title page must carry an instructive title that reflects the content, a running title (less than 45 characters together with spaces), names of the authors and co-authors, and the place(s) wherever the work was carried out. The full postal address in addition with the e-mail address of related author must be given. Up to eleven keywords or very brief phrases have to be given to help data retrieval, mining and indexing.

Abstract, used in Original Papers and Reviews:

Optimizing Abstract for Search Engines

Many researchers searching for information online will use search engines such as Google, Yahoo or similar. By optimizing your paper for search engines, you will amplify the chance of someone finding it. This in turn will make it more likely to be viewed and/or cited in a further work. Global Journals Inc. (US) have compiled these guidelines to facilitate you to maximize the web-friendliness of the most public part of your paper.

Key Words

A major linchpin in research work for the writing research paper is the keyword search, which one will employ to find both library and Internet resources.

One must be persistent and creative in using keywords. An effective keyword search requires a strategy and planning a list of possible keywords and phrases to try.

Search engines for most searches, use Boolean searching, which is somewhat different from Internet searches. The Boolean search uses "operators," words (and, or, not, and near) that enable you to expand or narrow your affords. Tips for research paper while preparing research paper are very helpful guideline of research paper.

Choice of key words is first tool of tips to write research paper. Research paper writing is an art.A few tips for deciding as strategically as possible about keyword search:

- One should start brainstorming lists of possible keywords before even begin searching. Think about the most important concepts related to research work. Ask, "What words would a source have to include to be truly valuable in research paper?" Then consider synonyms for the important words.
- It may take the discovery of only one relevant paper to let steer in the right keyword direction because in most databases, the keywords under which a research paper is abstracted are listed with the paper.
- One should avoid outdated words.

Keywords are the key that opens a door to research work sources. Keyword searching is an art in which researcher's skills are bound to improve with experience and time.

Numerical Methods: Numerical methods used should be clear and, where appropriate, supported by references.

Acknowledgements: Please make these as concise as possible.

References

References follow the Harvard scheme of referencing. References in the text should cite the authors' names followed by the time of their publication, unless there are three or more authors when simply the first author's name is quoted followed by et al. unpublished work has to only be cited where necessary, and only in the text. Copies of references in press in other journals have to be supplied with submitted typescripts. It is necessary that all citations and references be carefully checked before submission, as mistakes or omissions will cause delays.

References to information on the World Wide Web can be given, but only if the information is available without charge to readers on an official site. Wikipedia and Similar websites are not allowed where anyone can change the information. Authors will be asked to make available electronic copies of the cited information for inclusion on the Global Journals Inc. (US) homepage at the judgment of the Editorial Board.

The Editorial Board and Global Journals Inc. (US) recommend that, citation of online-published papers and other material should be done via a DOI (digital object identifier). If an author cites anything, which does not have a DOI, they run the risk of the cited material not being noticeable.

The Editorial Board and Global Journals Inc. (US) recommend the use of a tool such as Reference Manager for reference management and formatting.

Tables, Figures and Figure Legends

Tables: Tables should be few in number, cautiously designed, uncrowned, and include only essential data. Each must have an Arabic number, e.g. Table 4, a self-explanatory caption and be on a separate sheet. Vertical lines should not be used.

Figures: Figures are supposed to be submitted as separate files. Always take in a citation in the text for each figure using Arabic numbers, e.g. Fig. 4. Artwork must be submitted online in electronic form by e-mailing them.

Preparation of Electronic Figures for Publication

Even though low quality images are sufficient for review purposes, print publication requires high quality images to prevent the final product being blurred or fuzzy. Submit (or e-mail) EPS (line art) or TIFF (halftone/photographs) files only. MS PowerPoint and Word Graphics are unsuitable for printed pictures. Do not use pixel-oriented software. Scans (TIFF only) should have a resolution of at least 350 dpi (halftone) or 700 to 1100 dpi (line drawings) in relation to the imitation size. Please give the data for figures in black and white or submit a Color Work Agreement Form. EPS files must be saved with fonts embedded (and with a TIFF preview, if possible).

For scanned images, the scanning resolution (at final image size) ought to be as follows to ensure good reproduction: line art: >650 dpi; halftones (including gel photographs) : >350 dpi; figures containing both halftone and line images: >650 dpi.

Color Charges: It is the rule of the Global Journals Inc. (US) for authors to pay the full cost for the reproduction of their color artwork. Hence, please note that, if there is color artwork in your manuscript when it is accepted for publication, we would require you to complete and return a color work agreement form before your paper can be published.

Figure Legends: Self-explanatory legends of all figures should be incorporated separately under the heading 'Legends to Figures'. In the full-text online edition of the journal, figure legends may possibly be truncated in abbreviated links to the full screen version. Therefore, the first 100 characters of any legend should notify the reader, about the key aspects of the figure.

6. AFTER ACCEPTANCE

Upon approval of a paper for publication, the manuscript will be forwarded to the dean, who is responsible for the publication of the Global Journals Inc. (US).

6.1 Proof Corrections

The corresponding author will receive an e-mail alert containing a link to a website or will be attached. A working e-mail address must therefore be provided for the related author.

Acrobat Reader will be required in order to read this file. This software can be downloaded

(Free of charge) from the following website:

www.adobe.com/products/acrobat/readstep2.html. This will facilitate the file to be opened, read on screen, and printed out in order for any corrections to be added. Further instructions will be sent with the proof.

Proofs must be returned to the dean at dean@globaljournals.org within three days of receipt.

As changes to proofs are costly, we inquire that you only correct typesetting errors. All illustrations are retained by the publisher. Please note that the authors are responsible for all statements made in their work, including changes made by the copy editor.

6.2 Early View of Global Journals Inc. (US) (Publication Prior to Print)

The Global Journals Inc. (US) are enclosed by our publishing's Early View service. Early View articles are complete full-text articles sent in advance of their publication. Early View articles are absolute and final. They have been completely reviewed, revised and edited for publication, and the authors' final corrections have been incorporated. Because they are in final form, no changes can be made after sending them. The nature of Early View articles means that they do not yet have volume, issue or page numbers, so Early View articles cannot be cited in the conventional way.

6.3 Author Services

Online production tracking is available for your article through Author Services. Author Services enables authors to track their article - once it has been accepted - through the production process to publication online and in print. Authors can check the status of their articles online and choose to receive automated e-mails at key stages of production. The authors will receive an e-mail with a unique link that enables them to register and have their article automatically added to the system. Please ensure that a complete e-mail address is provided when submitting the manuscript.

6.4 Author Material Archive Policy

Please note that if not specifically requested, publisher will dispose off hardcopy & electronic information submitted, after the two months of publication. If you require the return of any information submitted, please inform the Editorial Board or dean as soon as possible.

6.5 Offprint and Extra Copies

A PDF offprint of the online-published article will be provided free of charge to the related author, and may be distributed according to the Publisher's terms and conditions. Additional paper offprint may be ordered by emailing us at: editor@globaljournals.org .

the search? Will I be able to find all information in this field area? If the answer of these types of questions will be "Yes" then you can choose that topic. In most of the cases, you may have to conduct the surveys and have to visit several places because this field is related to Computer Science and Information Technology. Also, you may have to do a lot of work to find all rise and falls regarding the various data of that subject. Sometimes, detailed information plays a vital role, instead of short information.

2. Evaluators are human: First thing to remember that evaluators are also human being. They are not only meant for rejecting a paper. They are here to evaluate your paper. So, present your Best.

3. Think Like Evaluators: If you are in a confusion or getting demotivated that your paper will be accepted by evaluators or not, then think and try to evaluate your paper like an Evaluator. Try to understand that what an evaluator wants in your research paper and automatically you will have your answer.

4. Make blueprints of paper: The outline is the plan or framework that will help you to arrange your thoughts. It will make your paper logical. But remember that all points of your outline must be related to the topic you have chosen.

5. Ask your Guides: If you are having any difficulty in your research, then do not hesitate to share your difficulty to your guide (if you have any). They will surely help you out and resolve your doubts. If you can't clarify what exactly you require for your work then ask the supervisor to help you with the alternative. He might also provide you the list of essential readings.

6. Use of computer is recommended: As you are doing research in the field of Computer Science, then this point is quite obvious.

7. Use right software: Always use good quality software packages. If you are not capable to judge good software then you can lose quality of your paper unknowingly. There are various software programs available to help you, which you can get through Internet.

8. Use the Internet for help: An excellent start for your paper can be by using the Google. It is an excellent search engine, where you can have your doubts resolved. You may also read some answers for the frequent question how to write my research paper or find model research paper. From the internet library you can download books. If you have all required books make important reading selecting and analyzing the specified information. Then put together research paper sketch out.

9. Use and get big pictures: Always use encyclopedias, Wikipedia to get pictures so that you can go into the depth.

10. Bookmarks are useful: When you read any book or magazine, you generally use bookmarks, right! It is a good habit, which helps to not to lose your continuity. You should always use bookmarks while searching on Internet also, which will make your search easier.

11. Revise what you wrote: When you write anything, always read it, summarize it and then finalize it.

12. Make all efforts: Make all efforts to mention what you are going to write in your paper. That means always have a good start. Try to mention everything in introduction, that what is the need of a particular research paper. Polish your work by good skill of writing and always give an evaluator, what he wants.

13. Have backups: When you are going to do any important thing like making research paper, you should always have backup copies of it either in your computer or in paper. This will help you to not to lose any of your important.

14. Produce good diagrams of your own: Always try to include good charts or diagrams in your paper to improve quality. Using several and unnecessary diagrams will degrade the quality of your paper by creating "hotchpotch." So always, try to make and include those diagrams, which are made by your own to improve readability and understandability of your paper.

15. Use of direct quotes: When you do research relevant to literature, history or current affairs then use of quotes become essential but if study is relevant to science then use of quotes is not preferable.

16. Use proper verb tense: Use proper verb tenses in your paper. Use past tense, to present those events that happened. Use present tense to indicate events that are going on. Use future tense to indicate future happening events. Use of improper and wrong tenses will confuse the evaluator. Avoid the sentences that are incomplete.

17. Never use online paper: If you are getting any paper on Internet, then never use it as your research paper because it might be possible that evaluator has already seen it or maybe it is outdated version.

18. **Pick a good study spot:** To do your research studies always try to pick a spot, which is quiet. Every spot is not for studies. Spot that suits you choose it and proceed further.

19. Know what you know: Always try to know, what you know by making objectives. Else, you will be confused and cannot achieve your target.

20. Use good quality grammar: Always use a good quality grammar and use words that will throw positive impact on evaluator. Use of good quality grammar does not mean to use tough words, that for each word the evaluator has to go through dictionary. Do not start sentence with a conjunction. Do not fragment sentences. Eliminate one-word sentences. Ignore passive voice. Do not ever use a big word when a diminutive one would suffice. Verbs have to be in agreement with their subjects. Prepositions are not expressions to finish sentences with. It is incorrect to ever divide an infinitive. Avoid clichés like the disease. Also, always shun irritating alliteration. Use language that is simple and straight forward. put together a neat summary.

21. Arrangement of information: Each section of the main body should start with an opening sentence and there should be a changeover at the end of the section. Give only valid and powerful arguments to your topic. You may also maintain your arguments with records.

22. Never start in last minute: Always start at right time and give enough time to research work. Leaving everything to the last minute will degrade your paper and spoil your work.

23. Multitasking in research is not good: Doing several things at the same time proves bad habit in case of research activity. Research is an area, where everything has a particular time slot. Divide your research work in parts and do particular part in particular time slot.

24. Never copy others' work: Never copy others' work and give it your name because if evaluator has seen it anywhere you will be in trouble.

25. Take proper rest and food: No matter how many hours you spend for your research activity, if you are not taking care of your health then all your efforts will be in vain. For a quality research, study is must, and this can be done by taking proper rest and food.

26. Go for seminars: Attend seminars if the topic is relevant to your research area. Utilize all your resources.

27. Refresh your mind after intervals: Try to give rest to your mind by listening to soft music or by sleeping in intervals. This will also improve your memory.

28. Make colleagues: Always try to make colleagues. No matter how sharper or intelligent you are, if you make colleagues you can have several ideas, which will be helpful for your research.

29. **Think technically:** Always think technically. If anything happens, then search its reasons, its benefits, and demerits.

30. Think and then print: When you will go to print your paper, notice that tables are not be split, headings are not detached from their descriptions, and page sequence is maintained.

31. Adding unnecessary information: Do not add unnecessary information, like, I have used MS Excel to draw graph. Do not add irrelevant and inappropriate material. These all will create superfluous. Foreign terminology and phrases are not apropos. One should NEVER take a broad view. Analogy in script is like feathers on a snake. Not at all use a large word when a very small one would be

sufficient. Use words properly, regardless of how others use them. Remove quotations. Puns are for kids, not grunt readers. Amplification is a billion times of inferior quality than sarcasm.

32. Never oversimplify everything: To add material in your research paper, never go for oversimplification. This will definitely irritate the evaluator. Be more or less specific. Also too, by no means, ever use rhythmic redundancies. Contractions aren't essential and shouldn't be there used. Comparisons are as terrible as clichés. Give up ampersands and abbreviations, and so on. Remove commas, that are, not necessary. Parenthetical words however should be together with this in commas. Understatement is all the time the complete best way to put onward earth-shaking thoughts. Give a detailed literary review.

33. Report concluded results: Use concluded results. From raw data, filter the results and then conclude your studies based on measurements and observations taken. Significant figures and appropriate number of decimal places should be used. Parenthetical remarks are prohibitive. Proofread carefully at final stage. In the end give outline to your arguments. Spot out perspectives of further study of this subject. Justify your conclusion by at the bottom of them with sufficient justifications and examples.

34. After conclusion: Once you have concluded your research, the next most important step is to present your findings. Presentation is extremely important as it is the definite medium though which your research is going to be in print to the rest of the crowd. Care should be taken to categorize your thoughts well and present them in a logical and neat manner. A good quality research paper format is essential because it serves to highlight your research paper and bring to light all necessary aspects in your research.

Informal Guidelines of Research Paper Writing

Key points to remember:

- Submit all work in its final form.
- Write your paper in the form, which is presented in the guidelines using the template.
- Please note the criterion for grading the final paper by peer-reviewers.

Final Points:

A purpose of organizing a research paper is to let people to interpret your effort selectively. The journal requires the following sections, submitted in the order listed, each section to start on a new page.

The introduction will be compiled from reference matter and will reflect the design processes or outline of basis that direct you to make study. As you will carry out the process of study, the method and process section will be constructed as like that. The result segment will show related statistics in nearly sequential order and will direct the reviewers next to the similar intellectual paths throughout the data that you took to carry out your study. The discussion section will provide understanding of the data and projections as to the implication of the results. The use of good quality references all through the paper will give the effort trustworthiness by representing an alertness of prior workings.

Writing a research paper is not an easy job no matter how trouble-free the actual research or concept. Practice, excellent preparation, and controlled record keeping are the only means to make straightforward the progression.

General style:

Specific editorial column necessities for compliance of a manuscript will always take over from directions in these general guidelines.

To make a paper clear

· Adhere to recommended page limits

Mistakes to evade

- Insertion a title at the foot of a page with the subsequent text on the next page

- Separating a table/chart or figure - impound each figure/table to a single page
- Submitting a manuscript with pages out of sequence

In every sections of your document

· Use standard writing style including articles ("a", "the," etc.)

· Keep on paying attention on the research topic of the paper

· Use paragraphs to split each significant point (excluding for the abstract)

· Align the primary line of each section

· Present your points in sound order

· Use present tense to report well accepted

· Use past tense to describe specific results

· Shun familiar wording, don't address the reviewer directly, and don't use slang, slang language, or superlatives

· Shun use of extra pictures - include only those figures essential to presenting results

Title Page:

Choose a revealing title. It should be short. It should not have non-standard acronyms or abbreviations. It should not exceed two printed lines. It should include the name(s) and address (es) of all authors.

Abstract:

The summary should be two hundred words or less. It should briefly and clearly explain the key findings reported in the manuscript--must have precise statistics. It should not have abnormal acronyms or abbreviations. It should be logical in itself. Shun citing references at this point.

An abstract is a brief distinct paragraph summary of finished work or work in development. In a minute or less a reviewer can be taught the foundation behind the study, common approach to the problem, relevant results, and significant conclusions or new questions.

Write your summary when your paper is completed because how can you write the summary of anything which is not yet written? Wealth of terminology is very essential in abstract. Yet, use comprehensive sentences and do not let go readability for briefness. You can maintain it succinct by phrasing sentences so that they provide more than lone rationale. The author can at this moment go straight to

shortening the outcome. Sum up the study, with the subsequent elements in any summary. Try to maintain the initial two items to no more than one ruling each.

- Reason of the study - theory, overall issue, purpose
- Fundamental goal
- To the point depiction of the research
- Consequences, including definite statistics - if the consequences are quantitative in nature, account quantitative data; results of any numerical analysis should be reported
- Significant conclusions or questions that track from the research(es)

Approach:

- Single section, and succinct
- As a outline of job done, it is always written in past tense
- A conceptual should situate on its own, and not submit to any other part of the paper such as a form or table
- Center on shortening results - bound background information to a verdict or two, if completely necessary
- What you account in an conceptual must be regular with what you reported in the manuscript
- Exact spelling, clearness of sentences and phrases, and appropriate reporting of quantities (proper units, important statistics) are just as significant in an abstract as they are anywhere else

Introduction:

The **Introduction** should "introduce" the manuscript. The reviewer should be presented with sufficient background information to be capable to comprehend and calculate the purpose of your study without having to submit to other works. The basis for the study should be offered. Give most important references but shun difficult to make a comprehensive appraisal of the topic. In the introduction, describe the problem visibly. If the problem is not acknowledged in a logical, reasonable way, the reviewer will have no attention in your result. Speak in common terms about techniques used to explain the problem, if needed, but do not present any particulars about the protocols here. Following approach can create a valuable beginning:

- Explain the value (significance) of the study
- Shield the model - why did you employ this particular system or method? What is its compensation? You strength remark on its appropriateness from a abstract point of vision as well as point out sensible reasons for using it.
- Present a justification. Status your particular theory (es) or aim(s), and describe the logic that led you to choose them.
- Very for a short time explain the tentative propose and how it skilled the declared objectives.

Approach:

- Use past tense except for when referring to recognized facts. After all, the manuscript will be submitted after the entire job is done.
- Sort out your thoughts; manufacture one key point with every section. If you make the four points listed above, you will need a least of four paragraphs.
- Present surroundings information only as desirable in order hold up a situation. The reviewer does not desire to read the whole thing you know about a topic.
- Shape the theory/purpose specifically - do not take a broad view.
- As always, give awareness to spelling, simplicity and correctness of sentences and phrases.

Procedures (Methods and Materials):

This part is supposed to be the easiest to carve if you have good skills. A sound written Procedures segment allows a capable scientist to replacement your results. Present precise information about your supplies. The suppliers and clarity of reagents can be helpful bits of information. Present methods in sequential order but linked methodologies can be grouped as a segment. Be concise when relating the protocols. Attempt for the least amount of information that would permit another capable scientist to spare your outcome but be cautious that vital information is integrated. The use of subheadings is suggested and ought to be synchronized with the results section. When a technique is used that has been well described in another object, mention the specific item describing a way but draw the basic

principle while stating the situation. The purpose is to text all particular resources and broad procedures, so that another person may use some or all of the methods in one more study or referee the scientific value of your work. It is not to be a step by step report of the whole thing you did, nor is a methods section a set of orders.

Materials:

- Explain materials individually only if the study is so complex that it saves liberty this way.
- Embrace particular materials, and any tools or provisions that are not frequently found in laboratories.
- Do not take in frequently found.
- If use of a definite type of tools.
- Materials may be reported in a part section or else they may be recognized along with your measures.

Methods:

- Report the method (not particulars of each process that engaged the same methodology)
- Describe the method entirely
- To be succinct, present methods under headings dedicated to specific dealings or groups of measures
- Simplify - details how procedures were completed not how they were exclusively performed on a particular day.
- If well known procedures were used, account the procedure by name, possibly with reference, and that's all.

Approach:

- It is embarrassed or not possible to use vigorous voice when documenting methods with no using first person, which would focus the reviewer's interest on the researcher rather than the job. As a result when script up the methods most authors use third person passive voice.
- Use standard style in this and in every other part of the paper - avoid familiar lists, and use full sentences.

What to keep away from

- Resources and methods are not a set of information.
- Skip all descriptive information and surroundings - save it for the argument.
- Leave out information that is immaterial to a third party.

Results:

The principle of a results segment is to present and demonstrate your conclusion. Create this part a entirely objective details of the outcome, and save all understanding for the discussion.

The page length of this segment is set by the sum and types of data to be reported. Carry on to be to the point, by means of statistics and tables, if suitable, to present consequences most efficiently.You must obviously differentiate material that would usually be incorporated in a study editorial from any unprocessed data or additional appendix matter that would not be available. In fact, such matter should not be submitted at all except requested by the instructor.

Content

- Sum up your conclusion in text and demonstrate them, if suitable, with figures and tables.
- In manuscript, explain each of your consequences, point the reader to remarks that are most appropriate.
- Present a background, such as by describing the question that was addressed by creation an exacting study.
- Explain results of control experiments and comprise remarks that are not accessible in a prescribed figure or table, if appropriate.
- Examine your data, then prepare the analyzed (transformed) data in the form of a figure (graph), table, or in manuscript form.

What to stay away from

- Do not discuss or infer your outcome, report surroundings information, or try to explain anything.
- Not at all, take in raw data or intermediate calculations in a research manuscript.

- Do not present the similar data more than once.
- Manuscript should complement any figures or tables, not duplicate the identical information.
- Never confuse figures with tables - there is a difference.

Approach

- As forever, use past tense when you submit to your results, and put the whole thing in a reasonable order.
- Put figures and tables, appropriately numbered, in order at the end of the report
- If you desire, you may place your figures and tables properly within the text of your results part.

Figures and tables

- If you put figures and tables at the end of the details, make certain that they are visibly distinguished from any attach appendix materials, such as raw facts
- Despite of position, each figure must be numbered one after the other and complete with subtitle
- In spite of position, each table must be titled, numbered one after the other and complete with heading
- All figure and table must be adequately complete that it could situate on its own, divide from text

Discussion:

The Discussion is expected the trickiest segment to write and describe. A lot of papers submitted for journal are discarded based on problems with the Discussion. There is no head of state for how long a argument should be. Position your understanding of the outcome visibly to lead the reviewer through your conclusions, and then finish the paper with a summing up of the implication of the study. The purpose here is to offer an understanding of your results and hold up for all of your conclusions, using facts from your research and generally accepted information, if suitable. The implication of result should be visibly described. Infer your data in the conversation in suitable depth. This means that when you clarify an observable fact you must explain mechanisms that may account for the observation. If your results vary from your prospect, make clear why that may have happened. If your results agree, then explain the theory that the proof supported. It is never suitable to just state that the data approved with prospect, and let it drop at that.

- Make a decision if each premise is supported, discarded, or if you cannot make a conclusion with assurance. Do not just dismiss a study or part of a study as "uncertain."
- Research papers are not acknowledged if the work is imperfect. Draw what conclusions you can based upon the results that you have, and take care of the study as a finished work
- You may propose future guidelines, such as how the experiment might be personalized to accomplish a new idea.
- Give details all of your remarks as much as possible, focus on mechanisms.
- Make a decision if the tentative design sufficiently addressed the theory, and whether or not it was correctly restricted.
- Try to present substitute explanations if sensible alternatives be present.
- One research will not counter an overall question, so maintain the large picture in mind, where do you go next? The best studies unlock new avenues of study. What questions remain?
- Recommendations for detailed papers will offer supplementary suggestions.

Approach:

- When you refer to information, differentiate data generated by your own studies from available information
- Submit to work done by specific persons (including you) in past tense.
- Submit to generally acknowledged facts and main beliefs in present tense.

Administration Rules Listed Before Submitting Your Research Paper to Global Journals Inc. (US)

Please carefully note down following rules and regulation before submitting your Research Paper to Global Journals Inc. (US):

Segment Draft and Final Research Paper: You have to strictly follow the template of research paper. If it is not done your paper may get rejected.

- The **major constraint** is that you must independently make all content, tables, graphs, and facts that are offered in the paper. You must write each part of the paper wholly on your own. The Peer-reviewers need to identify your own perceptive of the concepts in your own terms. NEVER extract straight from any foundation, and never rephrase someone else's analysis.

- Do not give permission to anyone else to "PROOFREAD" your manuscript.

- Methods to avoid Plagiarism is applied by us on every paper, if found guilty, you will be blacklisted by all of our collaborated research groups, your institution will be informed for this and strict legal actions will be taken immediately.)
- To guard yourself and others from possible illegal use please do not permit anyone right to use to your paper and files.

Criterion for Grading a Research Paper (Compilation)
by Global Journals Inc. (US)

Please note that following table is only a Grading of "Paper Compilation" and not on "Performed/Stated Research" whose grading solely depends on Individual Assigned Peer Reviewer and Editorial Board Member. These can be available only on request and after decision of Paper. This report will be the property of Global Journals Inc. (US).

Topics	Grades		
	A-B	**C-D**	**E-F**
Abstract	Clear and concise with appropriate content, Correct format. 200 words or below	Unclear summary and no specific data, Incorrect form Above 200 words	No specific data with ambiguous information Above 250 words
Introduction	Containing all background details with clear goal and appropriate details, flow specification, no grammar and spelling mistake, well organized sentence and paragraph, reference cited	Unclear and confusing data, appropriate format, grammar and spelling errors with unorganized matter	Out of place depth and content, hazy format
Methods and Procedures	Clear and to the point with well arranged paragraph, precision and accuracy of facts and figures, well organized subheads	Difficult to comprehend with embarrassed text, too much explanation but completed	Incorrect and unorganized structure with hazy meaning
Result	Well organized, Clear and specific, Correct units with precision, correct data, well structuring of paragraph, no grammar and spelling mistake	Complete and embarrassed text, difficult to comprehend	Irregular format with wrong facts and figures
Discussion	Well organized, meaningful specification, sound conclusion, logical and concise explanation, highly structured paragraph reference cited	Wordy, unclear conclusion, spurious	Conclusion is not cited, unorganized, difficult to comprehend
References	Complete and correct format, well organized	Beside the point, Incomplete	Wrong format and structuring

INDEX

Global Journal of Computer Science and Technology

Visit us on the Web at www.GlobalJournals.org | www.ComputerResearch.org
or email us at helpdesk@globaljournals.org

ISSN 9754350