ISSN: 2233-7857

International Journal of Future Generation Communication and Networking

Volume No. 18

Issue No. 2

May - August 2025

ENRICHED PUBLICATIONS PVT. LTD

S-9, IInd FLOOR, MLU POCKET,
MANISH ABHINAV PLAZA-II, ABOVE FEDERAL BANK,
PLOT NO-5, SECTOR-5, DWARKA, NEW DELHI, INDIA-110075,
PHONE: - + (91)-(11)-47026006

ISSN: 2233-7857

International Journal of Future Generation Communication and Networking

Aims and Scope

SERSC is an international center for supporting distinguished scholars and students who are researching various areas of Science and Technology. SERSC wishes to provide good chances for academic and industry professionals to discuss recent progress in various areas of Science and Technology.

SERSC organizes many international conferences, symposia and workshops every year, and provides sponsor or technical support to researchers who wish to organize their own conferences and workshops.

SERSC also publishes high quality academic international journals in various areas of Science and Technology

ISSN: 2233-7857

International Journal of Future Generation Communication and Networking

ISSN: 2233-7857

International Journal of Future Generation Communication and Networking

(Volume No. 18, Issue No. 2, May - August 2025)

Contents

ISSN: 2233-7857

Design and Analysis of an Efficient Approach of Cluster Head Selection for Balanced Energy Consumption in Wireless Sensor Networks

Zhong Luo and Naixue Xiong
School of Computer, Hubei University of Education
Hubei Co-Innovation Center for basic education & IT service

ABSTRACT

Since wireless sensor networks(WSNs) consist of massive tiny sensors for monitoring environmental data cooperatively and the energy resource of these nodes is nonreplenishable, protocols should concentrate on communication and processing technologies with the minimum energy cost. To improve energy efficiency and maximize the network lifetime, some cluster-based routing protocols are presented. In those protocols, cluster heads selection could be regard as an important issue and it will affect the lifetime of the whole network. In this paper, we propose an efficient approach of cluster head selection for balanced energy consumption in WSNs based on sensor nodes' energy level and distance to the sink. Simulation results demonstrate our method can enlarge the life-time and balance the energy consumption well among all sensors.

Keywords: wireless sensor networks, energy consumption, cluster head selection, clustering

1. Introduction

WSNs are often composed of massive cheap and resource-limited sensor nodes for monitoring physical or environmental data cooperatively, and the energy efficiency has become one of the hottest problems in recent years. Unlike traditional wireless ad-hoc network, the sensor nodes distributed in WSNs are usually kept in a static state and equipped with limited computing capability, memory, and battery power. Since it is particularly difficult to replenish the energy of the battery, how to achieve high energy efficiency and increase the network scalability are crucial research topics [4-5].

To maximize the lifespan of the whole network and achieve high energy efficiency, sensor nodes can be organized into clusters, which consist of member nodes and the cluster-head (CH) [6]. In this hierarchical architecture, data collected by member nodes usually are forwarded by CH node to the sink or base station (BS), and the energyefficient communication protocol LEACH (Low Energy Adaptive Clustering Hierarchy). On one hand, CH is responsible for communicating with its members. On the other hand, it should receive the sensed data of other sensors in the same cluster and transmit these data to the BS [7]. Because the energy consumption of the CH is higher than of other nodes, the cluster membership and the cluster-head (CH) are changed periodically by BS so as to balance the energy consumption and the quality of the CH selection will influence the network lifetime seriously. Based on this method, many clustering protocols are presented for optimizing the selection of cluster heads and maximize data communication.

The remainder of the paper is organized as follows. In Section 2, we briefly review related work. Section 3 describes the system model. Section 4 presents the detail of the new methodology for CH selection based on sensor nodes' energy level and distance to the sink (or base station). Section 5 shows the

performance by simulations and compares it with LEACH and ECP [8]. Finally, Section 6 gives concluding remarks.

2. Literature Review

To improve the energy consumption from the arbitrary node to the sink, a wireless sensor network is organized by several clusters, which can reduce communication overhead in single-hop manner. In this way, the network's lifetime is extended and it is equivalent to increase the energy efficiency of the WSN. In [9], an optimal number of cluster-heads is proven to be existed to achieve minimum energy consumption in each round in clustering WSNs. It can't ensure that the number of cluster-heads is always equal to the optimal value owing to the random cluster-heads selection in LEACH. Meanwhile, the residual energy of candidate nodes is not taken into account and then the communication overhead in the network is not optimum [10-11]. Therefore, the nodes have more residual energy should be selected as a CH with more opportunities, which can balance the communication overhead among the sensor nodes and improve the lifetime of the network.

To improve LEACH method by changing probability, Inbo Sim et al. [12] proposed Energy Efficient Cluster Header Selection (ECS) algorithm, which take the probability function and energy parameter into consideration to choose optimal cluster heads. N.D. Tan et al. [13] proposed an LEACH-DE (LEACH- Distance Energy) routing protocol to decrease energy consumption and prolong network lifetime. During the process of Chs selection in LEACH-DE, residual energy of the nodes is classified into several levels and the geometric distance between the candidate nodes to the BS are examined for generating a key parameter. Both of those factors will influence the formation of clusters.

Minhas Akhtar [14] have extended the analytical model and present an Energy Aware Intra Cluster Routing (EAICR), which adopts multihop routing. Besides, several parameters are considered for CH selection, including number of packets sent in the network, energy consumed by the network, remaining energy level of nodes at specific time and network lifetime of the network.

O. Younis et al. [15] proposed a hybrid energy-efficient distributed clustering protocol, which attempts to deal with the CH selection problem based on a hybrid approach of the node's residual energy and the proximity distance between the node and its neighbors.

M. S. Ali et.al [16] proposed a novel scheme for cluster heads' selection, which defines current state probability and general probability of each node and employs approximation operators or priori information in each round. Sajjanhar et al. [17] proposed a Distributive Energy Efficient Adaptive Clustering (DEEAC) protocol, which focuses on the spatio-temporal variations of message transmission rates in different regions and determines the cluster head's selection according to node's hotness value and residual energy. B. Elbhiri et al. [18] proposed a Stochastic Distributed Energy-Efficient Clustering method (SDEEC) where the clusters are formulated in distributed manner and the cluster head election probability is more efficient and design a stochastic scheme detection to extend the network lifetime. Li, C. et al. [19] proposed an unequal clustering algorithm (EEUC), where the clusters far from the base station is larger In proportion than those close to the base station so as to balance the energy consumption of different cluster heads.

3. System Assumptions

We use the same "first order radio model" as presented in [20], in which the sensor nodes are equipped with transmitter and receiver antenna to obtain or amplify the transmission signals. A radio dissipates elec E_{elec} to run the transmitter or receiver circuitry,and the transmitter amplifier is E_{amp} . Thus, to transmit b bits message a distance d, the energy consumption can be calculated as:

$$E_{Tx} = \begin{cases} b(E_{elec} + \varepsilon_{fs} d^2) & d < d_0 \\ b(E_{elec} + \varepsilon_{amp} d^4) & d \geq d_0 \end{cases} \quad (1)$$

and to receive this message, the energy expends:

$$E_{Rx} = b * E_{elec} \quad (2)$$

All nodes in the network are divided into three types, including the CH, auxiliary cluster nodes and member nodes. Firstly, the CHs should be selected, and the characteristic of heterogeneous energy of each node is taken into consideration. For node i, we denote the set of neighbor nodes in range of radius R by S_i, and E_{res_i} is the residual energy. Whether the node i can be selected as CH or not depends on the ratio of its residual energy to the average energy of it neighbors. Assuming that p_i is the percentage of being selected as CH:

$$p_i = p_{opt} \times \frac{\sum_{j \in S_i} E_{res\_j} / |S_i|}{E_{res\_i}} \quad (3)$$

where p_{opt} is the optimal percentage, and it can be calculated as:

$$p_{opt} = \frac{k_{opt}}{N} \quad (4)$$

According to [21], the optimal number of CHs in the whole network is given as:

$$k_{opt} = \frac{\sqrt{N}}{\sqrt{2\pi}} \sqrt{\frac{\varepsilon_{fs}}{\varepsilon_{amp}}} \frac{M}{d_{sink}^2} \quad (5)$$

where d_{sink} is the mean value of all CHs to the sink, and $d_{sink} = 0.765 \frac{M}{2}$.

4. The Proposed Solution

4.1. Cluster Formation

Due to the impact of the nodes' energy on our system model, it is better to select the nodes with more remaining energy near the base station to reduce the energy consumption of nodes. By introducing the energy factor into the selection of the CHs, the improved threshold value can be expressed as:

$$T^*(n) = \begin{cases} \frac{p}{1 - p(r \bmod (1/p))} \times \left(\frac{E_{res\_i}}{\sum_{j \in S_i} E_{res\_j} / |S_i|} \right)^2, n \in G \\ 0, others \end{cases} \quad (6)$$

where r is the current round and G is the set of the sensors which have not been selected as CHs till the last rounds. E_{res_i} is the residual energy of node i, then $\sum_{j \in S_i} E_{res\_j} / |S_i|$ denotes the average remain energy of its neighbors.

If $E_{res\_i} > \sum_{j \in S_i} E_{res\_j} / |S_i|$, then $\frac{E_{res\_i}}{\sum_{j \in S_i} E_{res\_j} / |S_i|} > 1$ and it makes the value of $(\frac{E_{res\_i}}{\sum_{j \in S_i} E_{res\_j} / |S_i|})^2$ grows exponentially, and vice versa. Consequently, the value of $T^*(n)$ from the node which remains more energy is much larger and has the more possibility to be selected as CH.

According to the optimal probability p_{opt} , the whole network can be divided into k_{opt} equal regions. In conventional deployment stage, the CH only is responsible for aggregating the data collected from it memberships with the range of its region and then relay the aggregated solution to sink. Besides, the neighbor nodes of sink will perform direct transmission to the destination. In this paper, we assign a random initial energy level to each sensor after cluster formation. To balance the energy consumption of all nodes, the definition of energy levels is presented and used for selecting the CH-candidate nodes, which transmit packets and advertises its ID and residual energy level. A CH-candidate monitors advertisements from others and defers from acting as a CH if a higher energy level is reported by another. Finally, candidate with the highest residual energy level will become CH. Other nodes in this region will become the member of this cluster.

4.2. Selection of CH-Candidate Nodes

In the secondary stage of the election of CH-candidate nodes, the circumstances can be divided into two different situations. In the model of single hop, when the distance between the CHs and the sink meets with $d \geq d_0$, the CHs are responsible for collecting and aggregating the sensing data and the CH-candidate nodes forward this solution to the sink. However, in the multi-hop mode, while the distance accords with $d \geq \frac{1}{2} d_0$, the CH-candidate nodes will fulfill the task of collecting the data by the member nodes in same cluster, and at this point, the CH is responsible for forwarding the fusion data to the next CH near by sink.

In case of single-hop routing mode, if the distance between the CH and the sink node is greater than or equal to d_0 , then the CH-candidate nodes should be selected. To optimize the lifetime and decrease the energy consumption during the process of data relay, the selection of CH-candidate nodes are considered from three aspects, *i.e.* the residual energy E_{res_i}, the distance between the node and CH $d(i, CH)$, and the distance from node to the sink. Normally, the nodes with higher residual energy, nearer by the CH, or the distance from CH-candidate nodes less than those should be chosen. Therefore, the formula for selecting the appropriate CH-candidate nodes is shown as follows:

$$j = \arg\max_{j \in S_{CH-i}} \{ \frac{d(CH_i, sink) - d(j, sink)}{d(j, sink)} \times E_{res-j} \} \quad (7)$$

where $d_{CH_i - sink}$ denotes the distance between CH i and the sink.

4.3. Intra-cluster Multi-hop Routing Setup

In multi-hop communication, the task of CH receives the data from the various member nodes of the cluster, aggregates all the data and then sends the data to the next hop and forwards messages from neighboring cluster head. In this paper, a multi-hop communication protocol is designed to save energy. We set a value d, which is the threshold for judging the data will be transmitted to CHs directly or not. If the distance is farther than d, it will find an adjacent node as the relay one. Otherwise, it can forward the data to the destination directly. In addition, the selection of relay node depends on the distance and residual energy. According to the free space propagation channel model, the energy consumption and the energy consumption for forwarding node i to node j are defined respectively as follows.

$$\begin{cases} E(i,j) = 3lE_{elec} + l\varepsilon_{fs}(d(i,j)^2 + d(j,CH)^2) \\ E_{forward}(i,j) = d(i,j)^2 + d(j+CH)^2 \end{cases} \quad (8)$$

Since the CHs near the sink are responsible for forwarding the data both of theirs and remoter clusters in traditional multi-hop communication, it will result in the consequence that the nodes near the sink become invalid earlier than the remote and cause network segmentation. We resolve this problem by introducing the relay node to avoid the nodes near CHs depleting their energy quickly. The cost function is defined as follows.

$$cost_j = \delta \times \frac{d(i,j)^2 + d(j+CH)^2}{\max_{i,j \in S_{CH}}\{d(i,j)^2 + d(j+CH)^2\}} + (1-\delta)\frac{\max\{E_{res\_j}\} - E_{res\_j}}{\max\{E_{res\_j}\}} \quad (9)$$

Where δ is a parameter and $0 < \delta < 1$.

After each node has chosen the minimum cost node as its relay node, an intra-cluster route is constructed.

5. Simulation

In this section, we evaluate the performance of our protocol implemented with MATLAB [22-23]. For simplicity, we assume the probability of signal collision and interference in the wireless channel is ignorable. The parameters of specific experimental are shown in Table 1.

Table 1. Experimental Parameters

Parameters	Value
Simulation area(m×m)	400×400
Number of nodes	200
The station of BS	(200,200)
Initial energy	2J
ε_{fs}	10PJ/bit.m^2
E_{elec}	50nJ/bit
ε_{amp}	0.0013PJ/bit.m^4
D_0	90m
δ	0.4

In this section, we study the proposed algorithm, concentrating on the number of active nodes over time, the number of messages received by sink and the distribution of the average energy consumption of all nodes. And we define whether sensor node is considered as active or not depends its existing energy is greater than 0 and also can communicate with its adjacent nodes within given range. As a result, it can be deduced that few CHs die quickly for improper load balancing in contrast that sensor nodes may be

unable to communicate with its belonged CH even with some existing energy.

As shown in Figure 1, the number of active sensors in each round varies and our proposed method can obtain the longer lifetime than others along with the operation of the network. In LEACH, the difference of nodes' initial energy is not taken into account during the phrase of cluster head's selection, the energy consumption of all nodes can be distributed unbalancedly, which influence the overall lifetime. Hence, it can be observed from the test that our protocol improves the lifetime by 18.3% compared with LEACH and 11.2% compared with ECP. Also, since ECP adopts data fusion processing and multi hop forwarding mode, it can decrease the transmission overload from sensors to sink and more active nodes exists than LEACH.

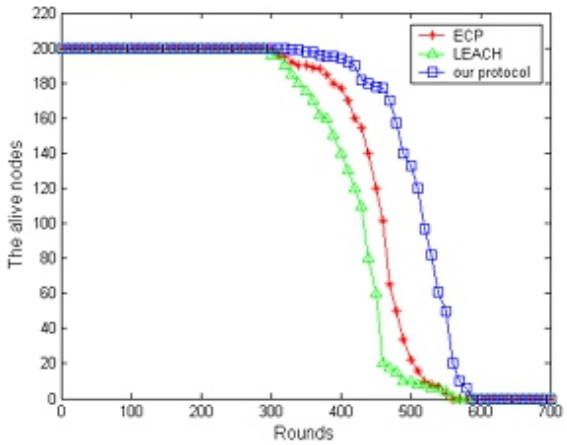


Figure 1. Number of Active Nodes Over Time

Figure 2 shows the information received by the sink and it is clear from the figures that total number of packets received by the base station in case of proposed protocol is much greater than other algorithms.

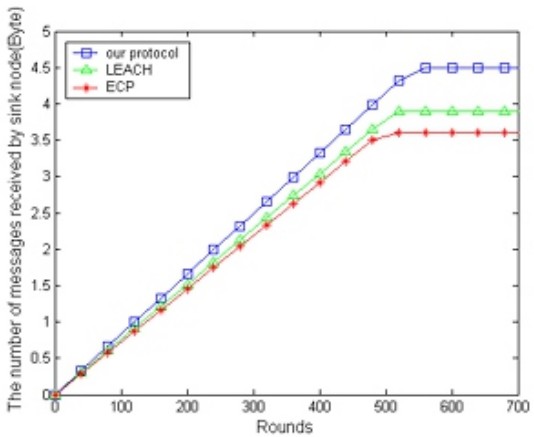


Figure 2. The Number of Messages received by Sink Over Time

The distribution of the average energy consumption of all nodes with respect to the number of rounds for each algorithm is shown in Fig. 3. The simulation results show that in our protocol the average energy consumption of all nodes varies comparative stably during the most of the rounds. The average energy consumption of other algorithms fluctuates in wide range, especially in LEACH. On the other hand, the average energy consumption of ECP can keep in the low level, which benefit from the optimization of

the selection of Chs.

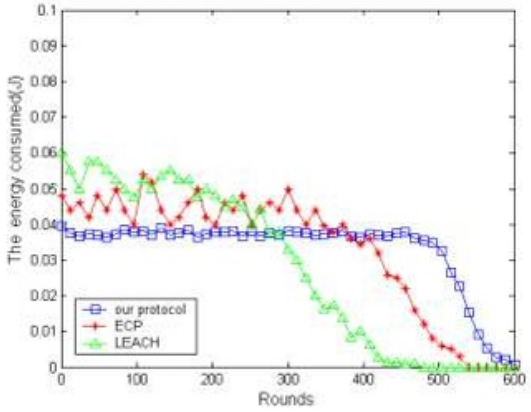


Figure 3. The Distribution of the Average Energy Consumption of all Nodes in Different Rounds

6. Conclusion

By analyzing the deficiency LEACH clustering algorithm, this paper proposes candidate cluster head selection mechanism for energy balancing. The main idea is the classification of nodes according to the residual energy for the sake of that the node with higher level are more likely to be cluster-head. In addition, for the non-cluster head nodes, they may receive the messages from multiple candidate cluster heads and choose to join the cluster according to the comprehensive evaluation of residual energy and distance. In order to make the energy consumption more balanced, the cluster-heads send the data to the sink in the manner of multi-hop transmission, and use unequal clustering mechanism to avoid the situation where the nodes near the sink deplete their energy much faster distant nodes.

Acknowledgments

This work is partially supported by the National Science Foundation of Hubei under Grant No. 2015CFB405. The authors would like to thank the anonymous reviewers for their insightful comments and constructive suggestions that have improved the paper.

References

[1] M. Presser and P. M. Barnaghi, "The SENSEI project: integrating the physical world with the digital world of the network of the future", IEEE Commun. Mag., no. 47, no. 4, (2009), pp. 1-4.

[2] V. M. Rohokale and N. R. Prasad, "Receiver sensitivity in opportunistic cooperative internet of things", Ad hoc Networks, vol. 49, no. 3, (2010), pp. 160-167.

[3] B. Shen, Z. Wang and Y. S. Hung, "Distributed consensus H-infinity filtering in sensor networks with multiple missing measurements: the finite-horizon case", Automatica, vol. 46, no. 10, (2010), pp, 1682 1688.

[4]H. S. Seo, S. J. Oh and C. W. Lee, "Evolutionary Genetic Algorithm for Efficient Clustering of Wireless Sensor Networks", the 6th IEEE Consumer Communications and Networking Conference, Las Vegas, (2009), pp. 1-5.

[5]B. P. Deosarkarl, N. S. Yadav and R. P. Yadav, "Cluster-head Selection in Clustering Algorithms for Wireless Sensor Networks: A Survey", International Conference on Computing, Communication and Networking. St. Thomas, (2008), pp. 1-8.

[6]W. B. Heinzelman, A. Chandrakasan and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks", Proceedings 33rd Hawaii Int Conf System Sciences (HICSS) , Maui, HI, (2000), pp. 20-30.

[7]J. B. Othman and B. Yahya, "Energy efficient and QoS based routing protocol for wireless sensor networks", Journal Parall. Distrib. Comput., vol. 70, no. 8, (2010), pp. 849–857.

[8] M. C. M. Thein and T. Thein, "An Energy Efficient Cluster-Head Selection for Wireless Sensor Networks", International Conference on Intelligent Systems, Modeling and Simulation, IEEE, (2009).

[9] Gupta, D. Riordan and S. Sampalli, Cluster-head Election using Fuzzy Logic for Wireless Sensor Networks, In Proceedings of IEEE Communication Networks and Services Research Conference, pp. 255-260, (2005).

[10] P. Ding, J. Holliday and A. Celik, "Distributed energy-efficient hierarchical clustering for wireless sensor networks", Distributed Computing in Sensor Systems, vol. 3560, (2005), pp. 322–339.

[11] H. Abusaimeh and S. H. Yang, "Dynamic cluster head for lifetime efficiency in WSN", International Journal of Automation and Computing, vol. 6, no. 1, (2009), pp. 48-54.

[12] I. Sim, K. Choi, K. Kwon and J. Lee, "Energy Efficient Cluster header Selection Algorithm in WSN", International Conference on Complex, Intelligent and Software Intensive Systems, (2009).

[13] N. D. Tan, L. Han, N. D. Viet and M. Jo. "An Improved LEACH Routing Protocol for EnergyEfficiency of Wireless Sensor Networks", Smart Computing Review, vol. 2, no. 5, (2012), pp. 360-369.

[14] A. A. Minhas Akhtar and S. Jabbar, "Energy Aware Intra Cluster Routing for Wireless Sensor Networks", International Journal of Hybrid Information Technology, vol. 3, no. 1, (2010), pp. 29-48.

[15] O. Younis and S. Fahmy, "HEED: A Hybrid, Energy efficient, Distributed Clustering approach for adhoc sensor networks", IEEE Transactions on mobile computing, vol. 3, no. 4, (2004), pp. 660-669.

[16] S. Ali, T. Dey and R. Biswas, "ALEACH: Advanced LEACH Routing Protocol for Wireless Microsensor Networks", ICECE 2008, (2008).

[17] U. Sajjanhar and P. Mitra, "Distributive Energy Efficient Adaptive Clustering Protocol for Wireless
Sensor Networks", Proceedings of the 2007 International Conference on Mobile Data Management, pp. 326-330, (2007).

[18] E. Brahim, S. Rachid, A. P. Zamora, D. Aboutajdine, "Stochastic Distributed Energy-Efficient Clustering (SDEEC) for heterogeneous wireless sensor Networks", ICGST-CNIR Journal, vol. 9, no. 2, (2009).

[19] C. Li, M. Ye, G. Chen and J. Wu, "An energy-efficient unequal clustering mechanism for wireless sensor networks", In Proceedings of the IEEE Intl Conf. on Mobile Adhoc and Sensor Systems Conf. (MASS), Washington, DC, USA, (2005), pp. 597–604.

[20] R. Zhu, "Intelligent Collaborative Event Query Algorithm in Wireless Sensor Networks", International Journal of Distributed Sensor Networks, vol. 2012, (2012), pp.1-11.

[21] F. Bajaber and I. Awan, "Energy efficient clustering protocol to enhance lifetime of wireless sensor network", Journal Ambient Intell Human Comput., (2010), pp. 239-248.

[22] M. Novey and T. Adalı, "Complex ICA by negentropy maximization", IEEE Trans. Neural Netw., vol. 19, no. 4, (2008), pp. 596–609.

[23] Y. M. Tashtoush and M. A. Okour, "Fuzzy Self-Clustering for Wireless Sensor networks", 2008 IEEE/IFIP International Conference on Embedded and Ubiquitous Computing, (2008), pp. 223-229.

ISSN: 2233-7857

Social Recommendation using Graph Database Neo4j : Mini Blog, Twitter Social Network Graph Case Study

Enzhi Zhang, Jinan Fiaidhi and Sabah Mohammed
Department of Computer Science, Lakehead University
*955 Oliver Rd, Thunder Bay, ON P7B 5E1, Canada

ABSTRACT

We are currently living in the age of BigData and social networking where the generated data is mainly unstructured and disorganized. This nature of such newly generated data which are growing exponentially gives importance for relationships between entities and point to the importance of using graph databases. By following the relationships between the people and properties in a meaningful manner you can determine co-occurrences, frequencies, and relevant nodes in the graph. This is the basis for many recommendation engines especially the one used for real-time recommendation engines operating on fast growing social data like Twitter. Real-time recommendation engines are key to the success of any online business. To make relevant recommendations in real time requires the ability to correlate product, customer, inventory, supplier, logistics and even social sentiment data. Moreover, a real-time recommendation engine requires the ability to instantly capture any new interests shown in the customer's' current visit – something that batch processing can't accomplish. This case study is an attempt to use a graph database Neo4j; one of the NoSQL data model, to build a mini blog prototype in order to perform efficient social recommendation. The Neo4j cypher query language is used to analyze real life social network dataset imported from twitter. The mini blog prototype has been created a web application using python flask web framework.

Keywords: Graph database, Neo4j, Python Flask, Social recommendation

1. Introduction

We believe that any developer has met a series of very complicated design problems in the use of relational database. In relational databases data is stored as rows in a table. It looks like kind of Excel spreadsheet. Each table stores a specific category of data (e.g. Persons and Friends) and has a predefined list of a properties for every element in that table (i.e. columns). With this data representation one can generate relations (e.g. making the connections between Persons and their Friends). However, adding a new column in a table is something that adds new dimension of complexity. The table is going to expand both vertically and horizontally, making it really tough to manage where the majority of these fields are going to be null for most the entries. This is an inefficient use of space. More complicated relations can be created in the relational representation using the join or self-join operations between different tables1 (see Figure 1 when applying the self-join operation on the Person table to produce a new relation of PersonFriend). In the relational database approach, friend relationship is stored using ID. In order that you look for friend relationship on the basis of this data, search all tables and then match IDs. relational database work well when the data is predictable and fits well into tables, columns, rows, and wherever queries are not very join-intensive. But there are rich, connected domains all around us including social and P2P networks, thesauri, route planning systems, recommendation systems, collaborative filtering and the World Wide Web where the relational approach isn't so well equipped at dealing with. Given such new paradigms importance, it's really worth spending some time to find better representation of

data as well as algorithms to work with them effectively. The start was the use of NOSQL wave database approaches (e.g. MongoDB, Cassandra, and Riak) which are designed to handle simple data. However, the most interesting applications deal with a complex, connected world. Such new relational complexity among data mandates the search and the development of new type of database that can significantly changes the standard direction taken by NOSQL. Graph databases, unlike their NOSQL and relational brethren, are designed for lightning-fast access to complex data found in social networks, recommendation engines and networked systems. A graph database2 is a database that uses graph structures for semantic queries with nodes, edges and properties to represent and store data. A key concept of the system is the graph (or edge or relationship), which directly relates data items in the store. The relationships allow data in the store to linked together directly, and in many cases retrieved with a single operation. Figure 2 illustrates representing the Friends relations in a graph database.

(a) The Person Relation (b) The PersonFriend Relation

Figure 1. Adding a New a New Column to a Relational Table

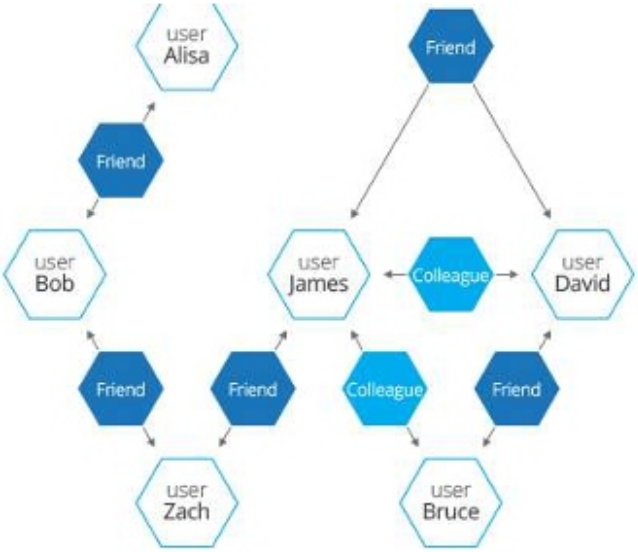


Figure 2. Representing the Friend Relations in a Graph Database

In a graph database, nodes and relationships (also called edges) are native to a graph database and this includes making sure that the graph is consistent (for example ensuring that every relationship has a start node and an end node). As a developer you can focus more on what to model and less on how to represent

it. For example, to traverse a graph is to find the nodes that are directly connected to a node in the case of directed relationships. We assume the node variable as input and then the code is as easy as follows:

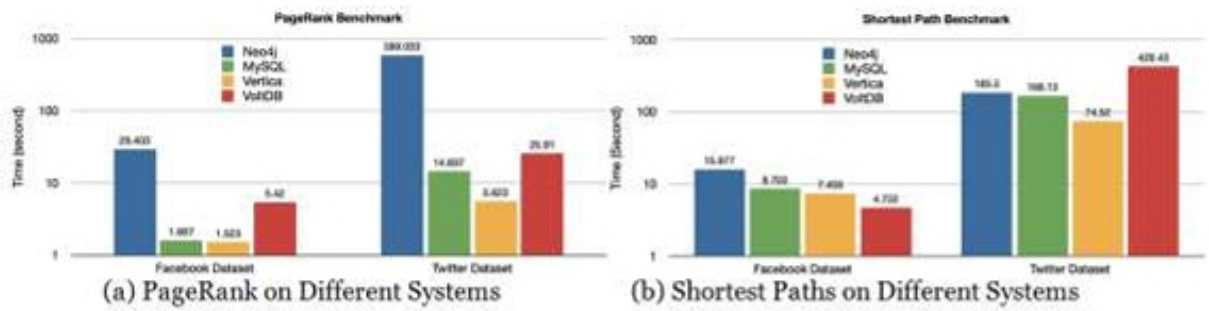
```
for ( Relationship rel : node.getRelationships( RelationshipTypes.FRIEND,
Direction.OUTGOING ) )
{
Node otherNode = rel.getEndNode();
// Here goes your code to make use of otherNode.
// You can aggregate it or whatever you want to do.
}
```

These relationships can be uni-directional or bi-directional and can even contain properties specific to that relationship (see Figure 3).

Figure 3. Graph Representation with Attributes

Comparing the performance of relational databases on graph analytics [1]. Here the leading graph database system (Neo4j 3) compared to three relational databases as described in Figure 3: a row-oriented database (MySQL4), a column-oriented database (Vertica), 5 and a main-memory database (VoltDB 6) are compared. Two queries, PageRank and Shortest Paths, on each of these systems. Considering two datasets from
the Stanford large network dataset collection:

- A Facebook dataset having 4K nodes and 88K edges, and
- A Twitter dataset having 81K nodes and 1.8Medges



(a) PageRank on Different Systems (b) Shortest Paths on Different Systems

Figure 3. Comparing the Performance of Neo4j with three Relational Databases

Based on the findings from Figure 3 we can conclude that searching social networks like Twitter will far more efficient using Neo4j graph database. Twitter in particular creates a graph of Users, Tweets, Hashtags and shared Links which makes searching for relations like a local community very difficult with any relational database. Figure 4 illustrates the various relations Twitter generates when linking users and tweets.

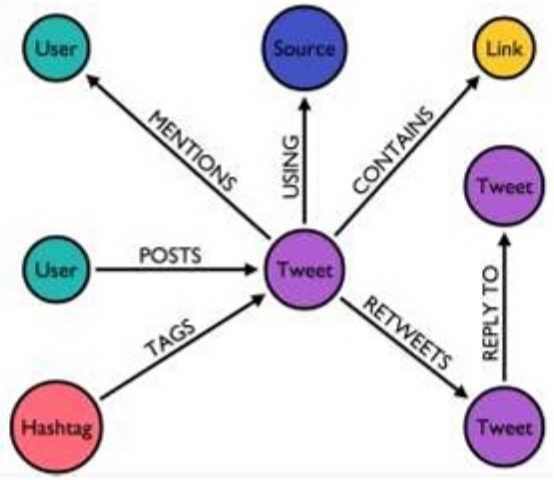


Figure 4. Representing Tweets Relations in a Graph Database

Graph databases such as Neo4j provide production grade front- or back-end social graph storage. Moreover, they offer graph analytics such as link prediction, shortest paths, clustering coefficient, and minimum spanning trees, bolstering the potential of graph tools such as NetworkX or NodeXL, machine learning frameworks such as Graphlab, and distributed processing systems such as Spark [2]. Increasingly, more important information is being shared through Twitter. New opportunities arise to use this tool to detect and extract crucial information about the scope and nature of certain emerging events. A major challenge for the extraction of such event information from Twitter is represented by the unstructured and noisy nature of tweets which requires a graph based representation that can accommodates such complexity and be practical to be used for extracting such relations [3,4].

2. Developing Mini Blog Social App

The mini blog web app is built under Python Flask web framework, Flask depends on two external libraries: the Jinja27 template engine and the Werkzeug WSGI toolkit. Jinja2 is a modern and designer-friendly templating language for Python, modelled after Django's templates. It is fast, widely used and secure with the optional sandboxed template execution environment, it has some certain features to support the project [5,6]

1.Built-in development server and debugger
2.Integrated unit testing support
3.RESTful request dispatching
4.Uses Jinja2 templating
5.Support for secure cookies (client side sessions)
6.100% WSGI 1.0 compliant
7.Unicode based

8.Extensively documented.

Beyond these features, the most important part is the connection between the front-end and the graph database, which in this project will be using Neo4j, Neo4j is a highperformance, open-source NOSQL graphical database that stores structured data into graph instead of the tables. It is sponsored by Neo Technology and implemented in Java and Scala. Because of the advantages such as embedded, high-performance and lightweight, etc., Neo4j is getting more attention by the world. The Neo4j developer manual [7] has indicate that "Neo4j implements the Property Graph Model efficiently down to the storage level. As opposed to graph processing or inmemory libraries, Neo4j provides full database characteristics including ACID transaction compliance, cluster support, and runtime failover, making it suitable to use graph data in production scenarios." All the data in Neo4j will be stored as an edge, a node, or a property. Each node and edge could have multiple properties. While processing a large number of complex, interconnected, low-structured data, the data traversal speed will not be influenced under this graph data model. Furthermore, Neo4j provides a unique SQL like Cypher query language, which has more straightforward syntax

There is a list of features of Neo4j provided by the Neo4j developer manual that make Neo4j very popular among users, developers, and DBAs:

1) Materializing of relationships at creation time, resulting in no penalties for complex runtime queries.

2) Constant time traversals for relationships in the graph both in depth and in breadth due to efficient representation of nodes and relationships.

3) All relationships in Neo4j are equally important and fast, making it possible to materialize and use new relationships later on to "shortcut" and speed up the domain data.when new needs arise.

4) Compact storage and memory caching for graphs, resulting in efficient scale-up and billions of nodes in one database on moderate hardware. Besides, Neo4j has provide powerful library Py2neo [8], which is a 3rd-party library for connecting to Neo4j from python. While the project started, the only way to communicate with a Neo4j server was over HTTP, which is impossible to dumping a lot of data to Neo4j at once. However, with the recently release of Neo4j 3, this technology now has a binary protocol that takes place over TCP, which it's now possible to dump data directly from python to the Neo4j database. Therefore, by simply import the Py2neo and set database URL to the Neo4j data browser, the mini blog will build connection n to the graph database. Same as most social network, the mini blog will support user register, login and share new post, all the HTML file will be create using Jinja2 template, the social recommendation will be demonstrating by using Neo4j CQL (Cypher query language), the Neo4j Developer Manual v3.1 is published by Neo Technology in 2016, and it provides detailed information about the CQL. Cypher is a unique query language for Neo4j which allows user for expressive and efficient querying and updating of the graph store. Cypher is a relatively simple language but very powerful. Even the most complicated database queries can be expressed easily through Cypher. This will save you from getting lost in database access and focus on your domain.

CQL has more straightforward syntax which all very complex queries in very easy manner.

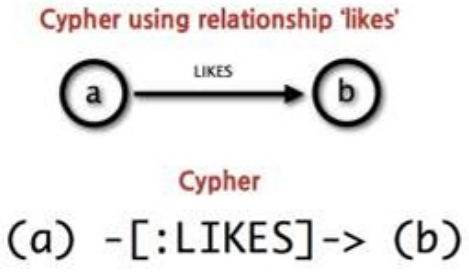


Figure 5. CQL Structure

Three recommendations will be built within the mini blog, which are the similar user by same tags, common user by the post and list most recent post by the time. And the graph data in the database will show as below:

Figure 6. Mini Blog Graph Data

3. Extending the Prototype to Twitter Recommendation

In order to store the real twitter data into the graph database Neo4j, this project will use another Python Library Requests, refer to the documentation provided by Python, Requests is the only Non-GMO HTTP library for Python, safe for human consumption [6]. Requests allows user to send organic, grass-fed HTTP/1.1 requests, without the need for manual labor. There's no need to manually add query strings to your URLs depends on the will of the developers and the objectives of the project, or to form-encode your POST data. Keep-alive and HTTP connection pooling are 100% automatic, powered by urllib3, which is embedded within Requests.

The authorization is required in order to access the Twitter API, which contains the API Key and API secret, with a Twitter app account user will be allowed to generate the access token, and with the authorization this project will be able to connect to Twitter API and import data into the graph database. This project will access to the Twitter search engine and import the keyword data into Neo4j, by construct the URL to set certain rules for the data importing, the parameter of the rules will depend on the device and the user needs. The final step is to create nodes and relationships in the graph database after the scraper got a list of dictionary back from the Twitter API, in order to generate the data graph. The process flow of the method is shown as below:

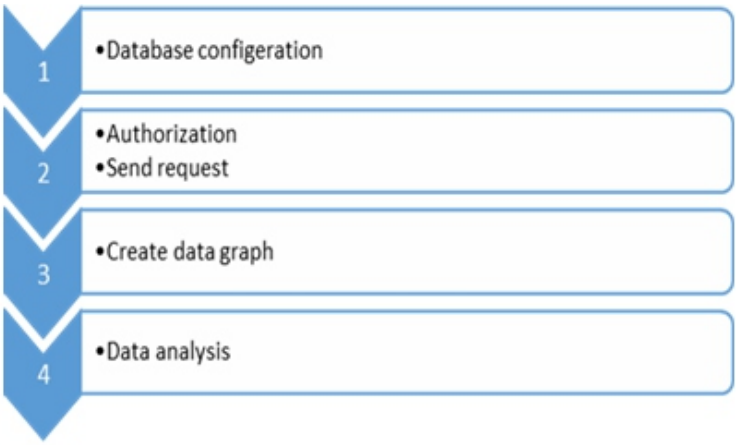


Figure 7: Process Flow of the Method

Following the authorization process, our blog allows the user to provide the arguments as a dictionary, using the params keyword argument, we could send required data in the URL's query string in order to constructing the URL, the data would be given as key/value pairs in the URL after a question mark:

```
def find_tweets(since_id):
payload["since_id"] = since_id
url = base_url +
“q={q}&count={count}&result_type={result_type}&lang={lang}&since_id={since_id}"
 .format(**payload)
r = requests.get(url, headers=headers)
tweets = r.json()["statuses"]
return tweets
```

Following that is to create graph through each of these dictionaries, by extracting the user and entities out of the dictionaries:

```
for t in tweets:
u = t['user']
e = t['entities']
```

The graph tweets will be created first, merge function will be used to find or create the node tweet, the first argument Tweet is the node label, the id is property key and the third argument is the property value. Because every tweet need a unique id, it will give an id to the tweet, merge_one function is used just in case some tweets will show more than once, it is to avoid create two same tweets. By setting the text properties to text and push it in order to create the tweet:

```
tweet = graph.merge_one("Tweet", "id", t['id'])
tweet.properties['text'] = t['text']
tweet.push()
```

Then using merge_

one function to find or create the user node, because certainly user will tweet about something multiple times, the following code will find the user with the property username and screen_name value out of the user dictionary. Then by using relationship function to create the relationship that indicate the user post the tweet:

```
user = graph.merge_one("User", "username", u['screen_name'])
graph.create_unique(Relationship(user, "POSTS", tweet))
```

Next step will go through the entities dictionary which has a key hashtag, it will give back a list of hashtag that tag the tweet, and also use the merge_one function to avoid create the hashtag nodes more than once, find or create the hashtag with hashtag as the node label, name as the property which is equal to a lower case version of the hashtag, then create the relationship for each of the hashtag that the hashtag tags the tweet:

```
for h in e.get('hashtags', []):
hashtag = graph.merge_one("Hashtag", "name", h['text'].lower())
graph.create_unique(Relationship(hashtag, "TAGS", tweet))
```

Also we will capture mentions with the same function and create relationship for the mentions that the tweet mentions other users:

```
for m in e.get('user_mentions', []):
mention = graph.merge_one("User", "username", m['screen_name'])
graph.create_unique(Relationship(tweet, "MENTIONS", mention))
```

Second last, we will get the tweet that potentially will reply to, it will refer to the tweets that reply to other tweets, to get the reply if there are any we need to find or create the tweet by the unique id then create the

relationship indicate the tweet reply to other tweet:

```
reply = t.get('in_reply_to_status_id')
if reply:
reply_tweet = graph.merge_one("Tweet", "id", reply)
graph.create_unique(Relationship(tweet, "REPLY_TO", reply_tweet))
```

Then if exist a tweet is a retweet of other tweet, we will also get the retweet return with the Twitter API as retweeted_status, then extract the id out and find or create the tweet by its id property, follow by create the relationship that the tweet retweet the other tweet if it exist:

```
r = t.get('retweeted_status', {})
r = r.get('id')
if r:
retweet = graph.merge_one("Tweet", "id", r)
graph.create_unique(Relationship(tweet, "RETWEETS", retweet))
```

Until now, the entire schema of the script has explained clearly, before start collecting data into Neo4j, we need to create some unique constrains using Py2neo create unique constrain function, the method is to make tweets unique by id, users unique by username and hashtag unique by name, it will also create an index on the node label property pair:

```
def constraint(label, property):
if property not in graph.schema.get_uniqueness_constraints(label):
graph.schema.create_uniqueness_constraint(label, property)
constraint("Tweet", "id")
constraint("User", "username")
constraint("Hashtag", "name")
```

Applications which process a timeline, wait some quantity of time, and then need to process new Tweets which have been added since the last time the timeline was processed can make one more optimization using the since_id parameter. So by steeing a since_id and update it in order to avoid finding same tweets is required, the last part is to find tweets by the keyword that specified in the command line, and send the list of dictionary to upload tweets, warn if there is no match tweets found and then sleep the Twitter API after each update:

```
since_id = -1
while True:
try:
tweets = find_tweets(sys.argv[1], since_id=since_id)
if not tweets:
print("No tweets found.")
time.sleep(60)
continue
since_id = tweets[0].get('id')
upload_tweets(tweets)
print(str(len(tweets)) + " tweets uploaded!")
time.sleep(60)
except Exception as e:
print(e)
time.sleep(60)
continue
```

The resulted data will be store as a graph into Neo4j, the final process to the data is using CQL to create recommendation Cypher. View the database via the Neo4j data browser, and use the Cypher below to view all the data:

```
MATCH (n) RETURN n
```

The data graph is shown as below:

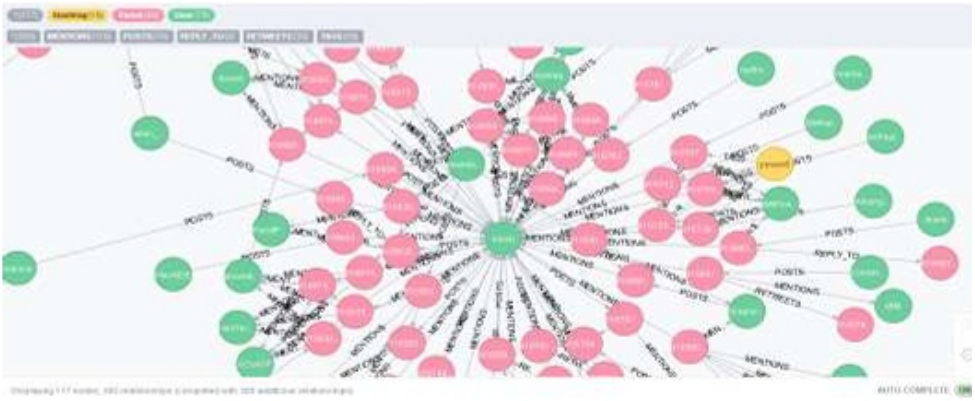


Figure 8. Twitter Graph Data

Normally, the social recommendation will base on what kind of data that user wants to get from the database, therefore the creation of the recommendation Cypher should start with asking questions, then translate the questions into Cypher, run it in the database and get the required data. For example: the question "Which has the top mentions of Neo4j?"

```
MATCH
(u:User)-[:POSTS]->(t:Tweet)-[:MENTIONS]->(m:User {username:'Neo4j'})
WHERE   u.username <> 'Neo4j'
RETURN  u.username AS username, COUNT(u.username) AS count
ORDER BY   count DESC LIMIT 10
```

username	count	u
jeanepaul	10	{username: jeanepaul}
mpyeager	3	{username: mpyeager}
importio	2	{username: importio}
sharmagourav	2	{username: sharmagourav}
FergusInLondon	1	{username: FergusInLondon}
tnarik	1	{username: tnarik}
marc_data	1	{username: marc_data}
IBMPowerSystems	1	{username: IBMPowerSystems}
jimwebber	1	{username: jimwebber}
riz_emba	1	{username: riz_emba}

Figure 9: Top Mentions of Neo4j

4. Conclusions

Due to the rapid development of the network performance, we started this case study to test the Neo4j graph database performance in both large and small range social networks. The mini blog web app indicate that the advantage of graph database is not only appears on the performance, the clear model and simple method makes the entire database extremely easy to manage. The recommendation is functional but with more entities import into the system, the power of the graph recommendation will appear better. From the Twitter data analysis, we have built social recommendations that uses the Twitter graph database we found that without joint tables the relationships and nodes traversing takes barely no time.

Acknowledgments
This case study is part from an MSc project under the supervision of Dr. J. Fiaidhi.

References

1.S.Patil, G. Vaswani and A. Bhatia, "Graph Databases-An Overview", International Journal of Computer Science Technologies, vol. 5, no. 1, (2014), 657-660, http://www.ijcsit.com/docs/Volume%205/vol5issue01/ijcsit20140501141.pdf
2.Robinson, Ian, J. Webber and E. Eifrem, "Graph databases: new opportunities for connected data", O'Reilly Media, Inc., (2015).
3.Drakopoulos, Georgios, A. Kanavos and A. Tsakalidis, "Evaluating twitter influence ranking with system theory", In Proceedings of the 12th International Conference on Web Information Systems and Technologies, WEBIST, (2016).
4.Klein, Bernhard, X. Laiseca, D. C. Mansilla, D. López-de-Ipiña, and A. P. Nespral, "Detection and extracting of emergency knowledge from twitter streams", In International Conference on Ubiquitous Computing and Ambient Intelligence, (2012), pp. 462-469.
5.J. Webber, "A programmatic introduction to neo4j", In Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity, (2012), pp. 217-218.
6.A. Ronacher, "Welcome to Flask", (2015), https://media.readthedocs.org/pdf/flask-russian docs/0.9/flask-russian-docs.pdf
7.Neo Technology, The Neo4j Developer Manual v3.1, (2017), https://neo4j.com/docs/developer manual/current/
8.Neo Technology, The Py2neo 2.0 Handbook, (2017), http://py2neo.org/2.0/
9.K. Reitz, "Requests: Http for humans", Online: http://docs. pythonrequests. org/. (24 December, 2012.) (2014).

ISSN: 2233-7857

4G Wireless Networks Architecture an Overview and Security Issues On 4G

Javed Ahmad Shaheen
Virtual University of Pakistan, Lahore

ABSTRACT

In this paper comprehensive study on current proceeds in wireless network security problems or issues for 4G are being presented. Mobility is the most invigorating character, and it has a vast impact as how communication is embryonic into the future 4G long term Evolution networks and the mobility in 4G networks demands new mobility support. LTE is designed with strong cryptographic techniques, mutual authentication between LTE network elements with security mechanisms built into its architecture. The paper has a lot of contribution in different field of wireless network as in current era wireless network technologies have expanded rapidly and the emergence of novel applications such as tablet, Smartphone, mobile TV, web 2.0 and streaming contents such as videos, games and more, led to the demand of a faster network of next generation technology, the fourth Generation (4G) the new technology or generation of mobile communication standards in telecommunication emerging from future wireless networks which works under LTE (Long Term Evolution) and WiMAX (Worldwide Interoperability for Microwave Access). 4G provides five times faster mobile broadband internet access of discussed devices and Security is considered as the most important aspect in 4G LTE and WiMAX technologies, in these two standards a significant amount of attention has been given for the security design architecture. First the paper, presents the 4G architecture and its technical description and 4G wireless security with network security of upcoming generation. Second it discussed some security issues and possible threats on 4G. Third the paper proposed security model of 4 layers which ensures secure transmission of packets by adopting several compulsory security measures.

Keywords: 4G description, Security for Network, Architecture of Security, 4G Architecture and 3GPP, Wi-Max, Wireless security

1. Introduction

Mobility is the most stimulating nature, and it has a vast impact as how communication is embryonic into the future 4G networks and the mobility in 4G networks demands new mobility support. If we compared it to usual mobility especially in wireless network, it has severe and vigorous security risks. In recent there is many advancements in wireless network. The wireless network technologies have included many rising applications like mobile television, web2-0 and streaming contents which is standardization feature of 3GPP (3rd generation partnership project). The 4G is a completely new fully IPbased incorporated system of systems and also network of networks accomplished following junction of wired and wireless networks. It includes computers, electronics and communication technology for their consumer and many other convergences which give it capacity to provide 100 Mbps and 1 Gbps respectively in external and internal surroundings with uninterrupted Quality of Service (QoS) and high security. It is also offering many services of different kind at any moment of any time according to requirements of user and also anywhere and seamless interoperability always on at affordable cost.

As the 4G is the next generation wireless communication systems of worldwide and it also has to become standardized as it has increased. security and too much trustworthy communication. The architectural design of 4G is very interoperability across the HetNet environments and it is also operate on the TCP/IP architecture procedure [3,].As the 3rd generation communication is going to shift to the.4th generation communication (4G) and many organizations like IEEE802.16m, International Telecommunication union (ITU), Vodafone, China mobile communications as well a lot of next generation mobile network vendors like Motorola and Samsung.[2,] are all repairing for their 4G technology. Now many definitions of 4G developed. which provide bandwidth of 1000Mbps in mobile devices and in normal 1Gbps. It is adjoining with heterogeneous networks which number of Radio Access Technology and Radio Access network (RAN) [1,]. enabling technologies interconnected for 4G are OFDM (orthogonal, frequency division multiplexing), vertical handover protocols, and in advance multiple input & multiple output and cognitive radio network is also added in 4G technology.
Different security issues problems and challenges in 4G technologies described in 4th section of this paper are being studied in this paper. The rest of, this paper is structured as, the section 2 consist of 4G technical overview and in Section 3 I, discuss, 4G network technology architecture [7,]. In Section 5 I have to studied proposed, four-layer security model which supervises and ensure more secure packet transmission by taking all necessary security procedures like taking the, form of interference, detection systems, Firewalls, and IPSec, and operating network resources in an intellectual way by using refined and authentication protocols.

2. Technical Description

The standard for 4G constituted of set of requirements has been specified by ITUR (International Telecomm Union Radio communication sector in March 2008 which are further named as ITM-A (International mobile Telecommunications advanced specification for $g standards. It has also specified peak speed for 4G Service 1Gbps to 100Mbps. [1]
As the first version of mobile Wi-Max and LTE is having very low bit rate rather than 1Gbps peak rate and not fully compatible with IMT-Advanced but many services provider branded it. Later on, in December 2010, ITU-R accepted, the said two technologies could still be considered as 4G, provided they represent ancestor to IMT-Advanced compliant versions. Later on WiMAX for has released its version for mobile called mobile Wi-Max released 2 also known as Wireless MAN Advanced or IEEE 802.16m and LTE has also released LTE Advanced called LTE-A. These are the diffident compatible versions for IMTAdvanced acquiescent for these two systems, and they have been promising speeds of 1 Gbit ps. It should also in our mind that 4G system does not support circuit switch telephony service. It provides only Internet Protocol (IP) based communication e.g. IP telephony. In 3G the previous technology used spread spectrum radio technology is dumped in all 4G systems and it is replaced by OFDMA which is multicarrier transmission and also replaced with frequency domain equalization (FDE) schemes, which consequently transfer very high bit rates regardless of extensive multi path radio propagation. So the peak bit rate is more enhanced through smart antenna arrays for multiple input multiple output (MIMO) communication.

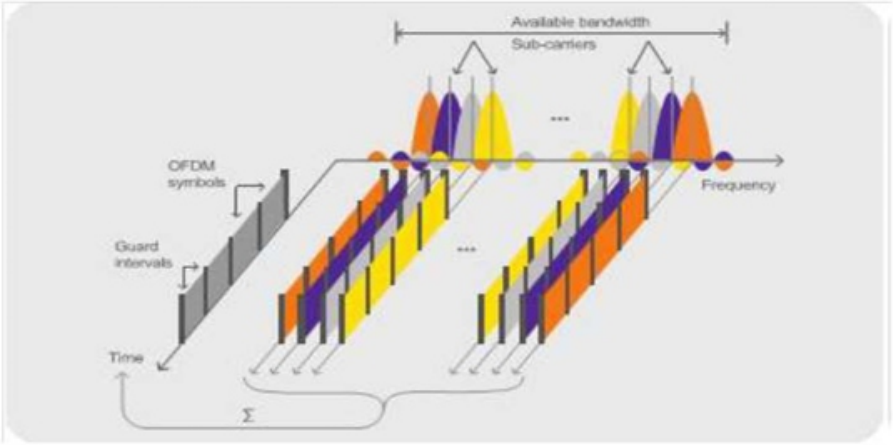


Figure 2.1.

As we have discussed 4G is based on packet switching only rather than circuit switching and it will require low latency for data transmission. When the 4G was deployed, the Ipv4 address overtiredness was probable in its final stages. Therefore, for 4G IPv6 is necessary for supporting a huge number of wireless permitted devices. IPv6 has the capability to remove the need for network address translation (NAT) if there is availability for increasing the number of IP addresses available, IPv6 is a method of sharing a few number of addresses between bigger groups of devices, although NAT will communicating with devices which are using existing IPv4 networks.

3. 4G, Network Architecture

The 4G network architecture constitute of multiple varied networks, such as Wimax and 3G [, 8]. In between the multiple access networks, anyone can use through the service subscriber and it also supports services from the same service unit like IP Multimedia subsystems. The Wimax architecture has an access serving network which has responsibility to provide access to the service stations or mobile stations which has a connection to network service provider. But the 3GPP LTE architecture has two core networks that is GPRS and EPC network. The GPRS core network provide network connections for existing RANs and Evolved packet core network (EPC) provide network connections to evolved RAN and 3 GPP, IP Access.

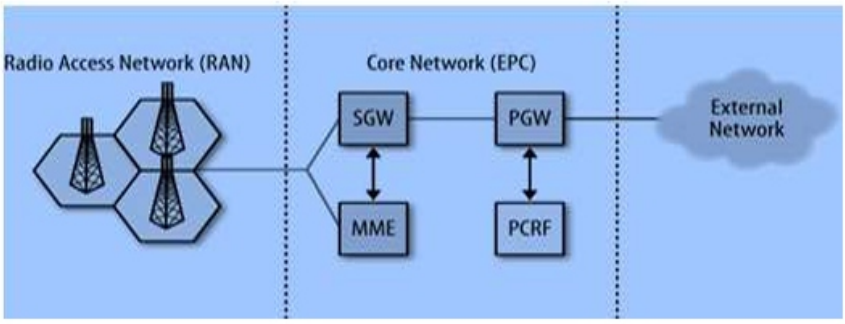


Figure 3.1. It is Showing Communication /Access of RAN with EPC

The basic network architecture for 4G is clearly elaborated as well explained in this figure 3.2

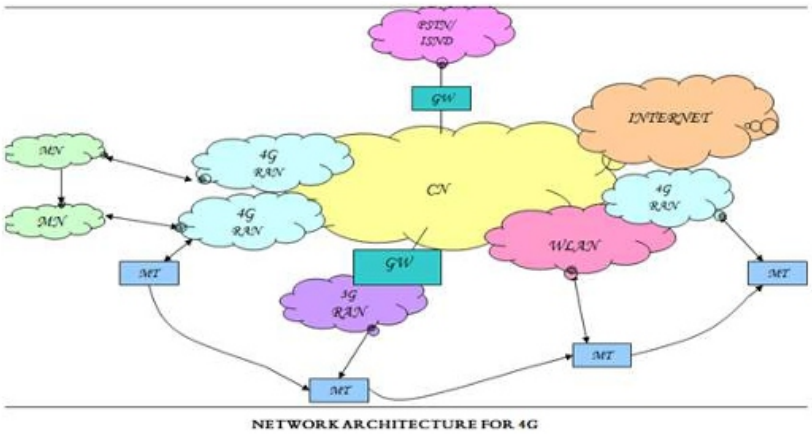


Figure 3.2.

3.1. IP Multimedia Subsystem Security Architecture

The IMS (IP Multimedia, subsystems) is an overlie on top of the network infrastructure like 3 GPP overlay on IMP. It wants to defend the all IMS sessions in between the endusers and IMS servers. It is also offering its authenticated and authorized mechanisms. There are two parts of IMS security and are described below.

3.1.1. First-hop Security: First Hop security secures the first hop from the end user to the proxy call session control function.

3.1.2. Network Domain Security: Network domain security protects the rest of hops between call session control functions inside IME core.

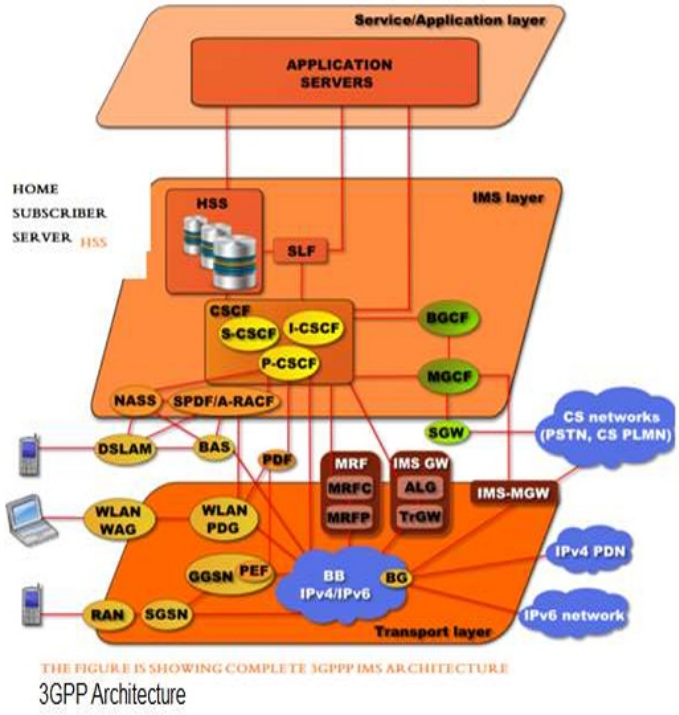


Figure 3.3.

3.2. Next Generation Network Security Architecture

The network security for next generation mostly secures the IMS security and it has two security domains.

Access view security: this view secures the first hop for end user device to access the network.

• The Core view security: the core view covers security within intra operated domain.

3.3. The 8 Security Dimensions for 4G

The dimensions of 8-security are used in 4G to take care and to measures applied counter threats and attacks some of them are describe as under

I. Access controls: it controls and measures protection level against unauthorized use of network resources.

ii. Authentication: this is used to measure confirmation level for uniqueness of each entity using the network.

iii. Nonrepudiation: it is used to demonstrate the data from its origin or recognizes the grounds of an event or action.

iv. Data confidentiality: it is major security dimension which is used for ensuring that data is not exposed to prohibited users.

v. Communication security: Communication in security is used to allow information flow only between permitted end points.

vi. Data integrity: it is used to ensure the correctness of data in such as, data can be tailored, erased, created or simulated without authorization and it also provides a intimation if unauthorized effort has put forward for the data change.

vii. Availability: it is also used to ensure that there is no denial of authorized access to elements of network, information that is stored, Information flows, services and applications suitable for network impacting events.

viii. Privacy: in the last the privacy is the main security dimension for providing the protection of derived information from the examination of network actions.

4. Wireless/4G Security Issues

4.1. Physical Layer Issue for 4G

Both WIMAX and LTE have focused on two key inclinations at the physical layer. By intentionally inserting manmade interference on a medium, the functioning of communication system can stop because of high signal to noise ratio.

There are two types of interference that can be carried out: Noise and multicarrier.

4.1.1. Noise Interference: It is performed using white Gaussian noise (WGN).

4.1.2. Multicarrier Interference: In it the attacker identifies carriers which are used by the system and then injects very narrowband signal on these carriers [4]. As the Interference attacks can easily be carried out as if the piece of equipment as well the knowledge to carry out such attacks is available without difficulty and widely. In this review on security, it indicates that interference is easy to detect using radio spectrum monitoring equipments. By using radio direction finding apparatus, the interfering source can be traced [1]. In addition, the power of the source signal has to increase and if spreading techniques are used, it may result to increase its resilience in opposition to interference. While the possibility of interference is significant, therefore it is easily being detected and addressed and we believe in its impact on the WIMAX/LTE network and users will also be limited [13].

4.2. Wimax MAC Security Issues

In WiMAX-MAC layer firstly it has to establish initial access with base station then to IEEE802.16 [11]. The Radio interface standard of WiMax describes several steps, if we say about a mobile station it includes seven steps. These are initial ranging and time synchronization, Upper level parameter acquisition, basic capabiliiities negotiation, scanning and synchronization, mobile station authorization and key exchange and registration with the serving base station is the last step by which connection established. Among these all steps five steps involved non secure traffic and other two steps involved secure traffic exchange based on the device authentication standards of WiMax [6].

4.3. DoS Security

The Denial-of-Service (DoS) is also a main security issue; it is attack which is a concern with WiMax network. These attacks are being initiated through simple flooding, attacking on authenticated management frames [2].

4.4. Security Issues for Wi-Fi, under 4G

Wireless LANs based, on WI-FI technology is available from a decade. However, the WIFi technology is, mostly used in homes and public places such as, airports, hostels, and, shopping malls and now in universities where security is seeming less critical, while the cost benefits of Wi-Fi could be attractive to enterprise environments thanks to improved mobility, costs for lower operational, and flexibility. Consequently, security researchers have focused on security threats and their solutions in Wi-Fi networks to make it applicable to the enterprise environments. The security mechanism for Wi-Fi so called wired equivalent privacy (WEP), had a number of security flaws arising from the misapplication of cryptography, e.g. the utilization of RC4 stream cipher and the authentication of CRC32 [3]. For this a comprehensive security assessment based on the ITU-T X.805 standard has been performed [9]. To remedies the security flaws of WIFi, there are many solutions have been proposed. The Robust Security Network (RSN) is used for the IEEE 802.Ix standard which is port based network access control that is a layer 2 authentication mechanism and it also specifies how EAP can be encapsulated in the Ethernet frames. RSA Laboratory and Cisco have developed TKIP to diminish the RC4 weakness through repeated replenishment of encryption key [5].

4.5. Potential Threats in 4G

The 4G technology may have to face many possible potential security Risks and threats. The various heterogeneous technologies have capability to access the infrastructure, so it must have needed potential security to secure technologies. 4G technology may also collapse to the entire network infrastructure when multiple service providers share the core network infrastructure. In 4G wireless technologies, end user equipments (either MS or any Wi-Fi routers) can also become a source of malicious attacks, malware, worms, viruses, calls and spam mails and so on. The spam over internet and the new spam for VoIP results a serious problem like the today's E-mail spam [2]. As like the above VoIP threats on the technology, 3 more VoIP Threats are:

(1) Spoofing that misdirects communications, modifies data, or transfers cash from a stolen credit card number,

(2) Standard input point, registration hijacking that, substitutes the IP address of packet, header with attacker's IP,

(3) Dropping of private, conversation that intercepts, and CRYPT arises, on IP packets,

5. Four Layer Security Model for 4G Networks

The 4 layer purposed security model integrated into two frameworks, peripheral and core which allow exploring new security concepts. In this model there are three separate security layers i.e. networking architecture security, network transport security & service and application security. This model is designed to create available coherent heterogeneous communication on a global scale and it is also responsible to supply continuous connectivity through the seamless operation of multiple mobile networks which are accessible by mobile nodes, providing features like cognitive radio and vertical handover. Peripheral network and Core network proposed four layer, security models which, are integrated in to the two frameworks. This proposed, model has two frame works, which are used as:
1.Peripheral frame: it can work on the mobile node and interacts with wireless access networks.
2.Core framework: it can work in a, distributed fashion in the core infrastructure, By this double, organized, frame work, a multi layer security, system arises that interacts, with these two, frameworks to provide, a secure environment.

6. Conclusion

For best perceptive, security of 4G networks, paper has represented different security aspects of 4G networks. This paper has also given some depiction on technical issues for 4G wireless network deployment and discussed physical layer issues, WiMAX_MAC layer, issues, DoS, issues and 4G WiFi security issue. As requirements apparently designates that there is a need for an integrated security model to protect data across different networks as well as the need to target the security models to protect different entities such as network infrastructures, servers and users. So far majority of research works have focused on studies or preliminary simulations. And they suggest that there is a vital need to expand researches with emulation and test-bed related studies to reveal further issues and challenges that needs to be addressed. Apparently there is still work to be completed; and researchers believe that there is a strong need for continuing studies on the fourth generation (4G) security threats and development of appropriate counter measures as in this paper 8security dimension for 4G networks are discussed and also discussed some possible potential threats, on 4G wireless networks but still there are new one exists. In the end paper has also proposed four layer 4G security model in which it they have to try to avoid unsecureness of 4G/ wireless and communication.

References

[1] N. Borisov, I. Goldberg and D. Wagner, "Intercepting Mobile Communications: The Insecurity of 802.11", Proceedings of ACM MobiCom'2001, Rome, Italy, July 2001.
[2] T. Park, H. Wang, M. Cho and K. G. Shin, "Enhanced Wired Equivalent, Privacy for IEEE 802.11. Wireless LANs", CSE-TR-469-02, University of Michigan, (2002).
[3] Bell Labs, "The Bell Labs Security, Framework: Making the Case for, End-to-End, Wi-Fi Security", (2006).
[4] IEEE Draft 802.1x./D1., "Port Based Network Access Control", available from http://www.ieee802.org/1/mirror/8021/docs99/PortNACIEEE.pdf",.
[5] Z. Shietal, "Layered Security Approach in LTE and Sim-ulation", Proceedings 3rd Int'l Conference Anticounterfeiting, Security and Identification in Communication, (ASID 09), IEEE, (2009), doi:10.1109/ICASID.2009.5276930.
[6] C. Vintila, V. Patriciu and I. Bica, "Security Analysis of ,LTE Access Network" Proceedings 10th Int'l Conf,. Networks (ICN 11), Intl Academy, Research, and ,Industry Assoc., (2011), pp. 29-34.

[7] Network Architecture, tech,. specification 3GPP, TS 23.002, V9.1.0, ,3GPP, (2009).
[8] L. Huang, "Performance of Authentication Protocols, ,in LTE_ Environments", Proceedings ,Int'l Conf,. Computational Intelligence and Security (CIS 09,), IEEE, 2009. doi:10.1109/CIS.2009.50
[9] L. Hui and B. Shuo, "Research and Implementation of LTE NAS Security", Proceedings Int'l . doi:10.1109/ICEIT.2010.5607551
[10] 4G Wireless Systems in Virtex-II by James A. Watson -- Manager, Applications Engineering, Xilinx, Inc. (7/1/01 -- Issue 40) jim.watson@xilinx.com
[11] Y. Raivio, "4G - Hype or Reality", IEE 3G Mobile Communication Technologies, Conference Publication, no. 477, (2001), pp. 346-350.
[12] J. Hu and W. W. Lu, "Open Wireless Architecture - The Core to 4G Mobile Communications", In Proceedings of ICCT, (2003).
[13] J. Pesola and S. Pönkänen, "Location-aided Handover in Heterogeneous Wireless Networks", In Wireless Personal Communications, vol. 30, iss. 2-4, (2004).
[14] Technical Specification Group Services and System Aspects, IP Multimedia Subsystem (IMS), Stage 2, TS 23.228, 3rd Generation Partnership Project, (2006).

ISSN: 2233-7857

Enhancement of Degraded Manuscript Images using Adaptive Gaussian Thresholding

Nikunj Khetan*, Lakshya Kejriwal and Dr. S. Indu
Delhi Technological University

ABSTRACT

Manuscripts in physical form are easily damaged over time resulting in loss of important information. Hence, there is a need to preserve the knowledge these manuscripts hold by enhancing the readability of the damaged manuscript by applying various image analysis techniques and storing them in digital form to prevent further deterioration of manuscript information. There are multiple well-known methods for document enhancement but, they are not suitable for use in enhancing damaged manuscript images. We propose a novel method to address the problem of enhancement and binarization of degraded manuscript images that applies a dual filtering technique for noise removal, Gaussian based adaptive thresholding technique and post processing using morphological operations that enhances readability of manuscript images. Our method showed good performance on qualitative as well as quantitative evaluation performed on 27 digital manuscript images with uniform character formation and an overall pseudo F-measure of 60.12%. Furthermore, our method is also compared with other well-known document enhancing techniques to establish the better applicability of our technique to preservation of manuscript information.

Keywords: Manuscript, Gaussian, Thresholding, Binarization, Enhancement

1. Introduction

Ancient manuscripts contain information pertaining to fields like art, science, astronomy, astrology, architecture, law, music and medicine. They are an important part of our cultural heritage. Parchments and palm leaf have been generally used as materials for manuscript writings but the ravages of time have affected legibility of characters in ancient manuscripts in the form of ink bleeding, smears, blotches, etc. There is a lot of ongoing effort to preserve manuscripts in physical form to prevent further deterioration but this has resulted in restricted access to the public. Hence, there is a need to store manuscripts in an enhanced digital form. Enhancement of manuscript images is an important digital heritage application and has many researchers working on methods to enhance and binarize historical documents. A lot of work has been done to enhance digital images of historic documents and binarize them so that OCR readability is feasible. The main objective is to separate foreground (text layer) from degraded(noisy) background layer and binarize the document so that readability is significantly enhanced.

Most commonly used method is the thresholding technique that separates these layers. Non-thresholding techniques [1][2][3] have also been used but, have been seen to show less promising results as compared to thresholding techniques [4]. Otsu [5], Kapur [6] and Kittler [7] use global thresholding technique in which a single threshold value is used to segment foreground and background.

Bernsen's method [8] computes an adaptive threshold by using the maximum and minimum gray-level value of the local block. The local contrast is given by:

$$C(i,j) = I_{max}(i,j) - I_{min}(i,j) \quad (1)$$

Where $C(i,j)$ denotes the contrast of an image pixel (i,j), $I_{max}(i,j)$ and $I_{min}(i,j)$ denote the maximum and minimum intensities within a local neighborhood windows of (i,j), respectively. If the local contrast $C(i,j)$ is smaller than a threshold, the pixel is set as background. Otherwise it is classified as text or background after comparing with the mean of $I_{max}(i,j)$ and $I_{min}(i,j)$. This method fails to work properly on degraded documents with complex background.

Local thresholding techniques used by Niblack [9] and Sauvola [10] calculate threshold value for a local block. Niblack's [9] method uses the mean and standard deviation to calculate the threshold. The threshold T of any pixel is given by

$$T(i,j) = m(i,j) + k.s(i,j) \quad (2)$$

where $m(i,j)$ and $s(i,j)$ refer to the mean and the variance of the gray values in the neighborhood, respectively. Sauvola's method [10] used a dynamic method to calculate the threshold by using the grayscale average and the standard deviation of the current pixel in the neighborhood. Here T is given by

$$T(i,j) = m(i,j).\left[1 + k.\left(\frac{s(i,j)}{R} - 1\right)\right] \quad (3)$$

where $m(i,j)$ and $s(i,j)$ are the same as those in Niblack's method. R is the dynamic range of standard deviation and k is the correction factor from 0 to 1. Sauvola's method does not perform well in very light or dark background. The block size is set according to the character property in the image.

Huang [11] employed a local thresholding method in which Otsu's method is applied on non-overlapping blocks. Gato [12] introduced an adaptive thresholding method in which Niblack's method is utilized to estimate foreground and then sequentially the background region is estimated. B. Gangamma [13] demonstrated bilateral filter and mathematical morphology to restore degraded historical image. Quraishi [14] employed a two-way approach using Particle Swarm Optimization (PSO) and bilateral filter.

B.M. Singh [15] proposed an adaptive binarization method for enhancement of severely degraded document images. The proposed method could deal with degradations that included low contrast, variable background, blur *etc.* but, the method produced noise where there was a large gap between text lines, high contrast above the text or in the entire image and also in areas where amount of text was comparatively less.

Kale [16] employed a hybrid approach in which first a global thresholding approach is applied (using Iterative Global Thresholding (IGT)) and then IGT is re-applied to areas with detected background noise. Tomar [17] illustrated a technique to extract text from historic images by considering the problem as a blind source separation. It aims to calculate the independent components from a linear mixture of source signals by maximizing a contrast function based on higher order cumulants.

Lu Di [18] proposed a method for binarization of degraded images to tackle degradations like bleeding-through, uneven illumination and contrast variation. The approach concentrated on the difference in image grayscale contrast in various areas.

In the proposed technique, the threshold value to segment foreground from background is calculated adaptively using Gaussian function. Two filtering methods are used in the pre-processing phase and morphological operations are used in the post-processing phase to further enhance the manuscript text readability. The section 2 in the paper explains the proposed method, its further compared with other methods in section 3.

2. Methodology

2.1. Proposed Method

The main goal is to separate the foreground (text) layer from the degraded background(noisy) layer but, manuscripts suffer from a range of damages with some parts of the same manuscript heavily affected compared to other parts. Hence, the use of a global thresholding method is generally unsuitable and therefore we employed the adaptive Gaussian Thresholding technique to separate the text layer from the noisy layer. We also used a novel dual-filtering method to eliminate noise and make further thresholding more effective. Our methodology consists of three steps- Pre-processing, Adaptive Thresholding and Post-processing. The various outputs after applying every step on manuscript image (shown in Figure 1a) is illustrated (Figure 1).

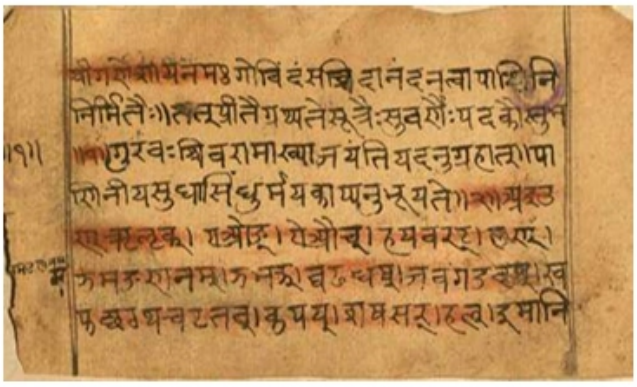


Figure 1a. Original Manuscript

2.1.1. Pre-processing: In order to remove noise and make further processing more efficient, Gaussian filter is used. Gaussian blurring technique convolves each point in an input array with a Gaussian kernel and thereafter sums them all to give the output array. See Figure 1b. A 2D Gaussian can be represented as:

$$G(x,y) = Ae^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}-\frac{(y-\mu_y)^2}{2\sigma_y^2}} \quad (4)$$

where μ is the mean (the peak) and σ represents the variance. The kernel size used was 9 × 9 and σ was calculated from kernel size using formula below:

$$\sigma = 0.3 * \big((kernel\ size - 1) * 0.5 - 1\big) + 0.8 \quad (5)$$

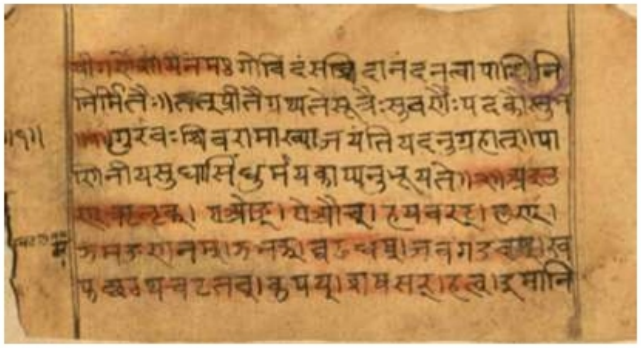


Figure 1b. After Applying Gaussian Blur

For further smoothing of background surfaces, we employed the mean-shift filter which is the initial step of the meanshift segmentation of an image. D. Comaniciu et. al. [19] expanded the vector (r,g,b) that represents each pixel as a 5-dimensional vector (x,y,r,g,b) which includes location information (x,y). The mean shift analysis is applied to all points in the image from which we obtain a set of convergence points. After obtaining set of convergence points, we replace the color component of the original point by the color component of the convergence point to generate a new point. The image constructed from the set of new points is the final result of the mean shift analysis. The output is the filtered posterized image with color gradients and fine-grain texture flattened as can be seen in Figure 1c. Mathematically, the meanshift filter works as follows. At every pixel (X,Y) of the input image the function executes meanshift iterations, i.e. , the pixel (X,Y) neighborhood in the joint space-color hyperspace is considered:

$$\begin{aligned}(x,y): X - sp \le x \le X + sp,\\ Y - sp \le y \le Y + sp,\\ ||(R,G,B) - (r,g,b)|| \le sr\end{aligned} \qquad (6)$$

Where sp is the spatial window radius and sr is the color window radius, (R,G,B) and (r,g,b) are the vectors of color components at (X,Y) and (x,y), respectively (though, the algorithm does not depend on the color space used, so any 3-component color space can be used instead). Over the neighborhood the average spatial value (X',Y') and average color vector (R',G',B') are found and they act as the neighborhood center on the next iteration. After the iterations are over, the color components of the initial pixel (that is, the pixel from where the iterations started) are set to the final value (average color at the last iteration). Also, Gaussian pyramid of multiple levels can be built, and the above procedure is then run on the smallest layer first. After that, the results are propagated to the larger layer and the iterations are run again only on those pixels where the layer colors differ by more than sr from the lower-resolution layer of the pyramid. That makes boundaries of color regions sharper[22]. The value of the parameters we used for the meanshift filter was fairly constant over all images. Both sp and sr were set to 8. Also, Gaussian pyramid of 3 levels was built. The values determined give optimal results for majority of images in our dataset but would need tuning for application to different types of manuscript images.

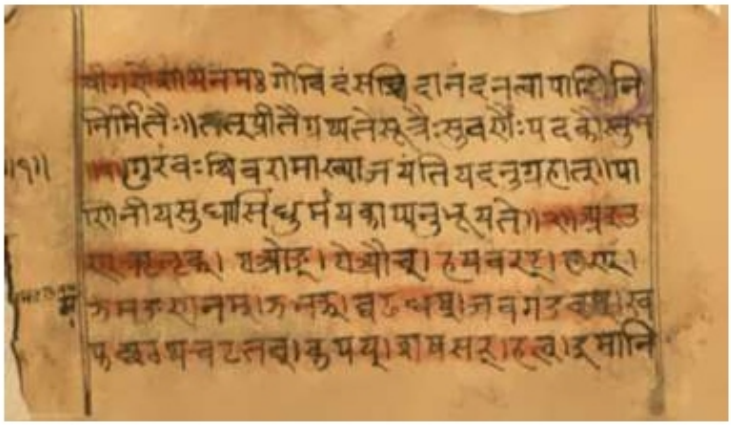


Figure 1c. After Applying Mean Shift Filter

2.1.2. Adaptive thresholding: In this stage, we apply thresholding to segment foreground from background and binarize the image. Image pixels belonging to the foreground areas are detected if the intensity value exceeds a threshold. In our case, we use adaptive thresholding to calculate the threshold value such that the value changes in accordance with the neighborhood of the pixel being examined. The

threshold value T(x,y) is the weighted sum of the blockSize × blockSize neighborhood of (x,y) where weights are calculated using a Gaussian window function w[n], (x,y) is the pixel being considered, blockSize is the size of neighborhood of pixel and C is a constant which is subtracted from the weighted mean calculated. Since the Gaussian function extends to infinity, it must either be truncated at the ends of the window, or itself windowed with another zero-ended window to get the Gaussian window. Since the log of a Gaussian produces a parabola, this can be used for nearly exact quadratic interpolation in frequency estimation [20].

$$w[n] = e^{-\frac{1}{2}\left(\frac{n-\frac{N-1}{2}}{\frac{\sigma(N-1)}{2}}\right)^2} \quad (7)$$

where $\sigma \leq 0.5$

N represents the width, in samples, of a discrete-time, symmetrical window function w[n], $0 \leq n \leq N-1$. The Gaussian window w[n] is used to calculate the weights. The blockSize used was 15 and constant C was set to 5 although some amount of variation is required to give optimal results for a particular manuscript image due to variability in the type of manuscript degradation. See Figure 1d.

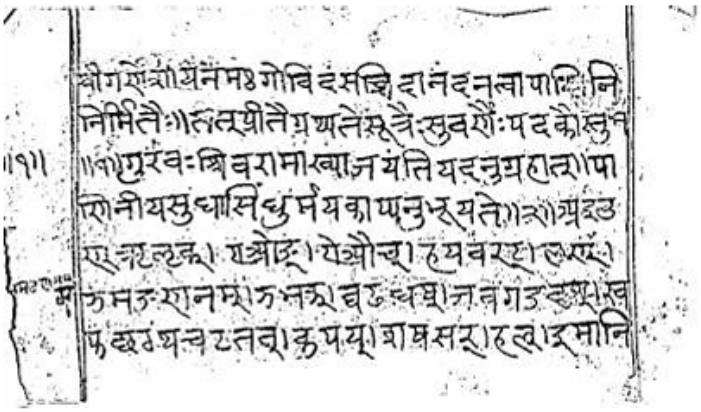


Figure 1d. After Applying Gaussian Thresholding

2.1.3. Post-processing: In this phase morphological operations- dilation and erosion are used, with a 3 × 3 rectangular structuring element, to reduce object noise from the binary image and further enhance the readability of the image and give the final result which can be seen in Figure 1e.

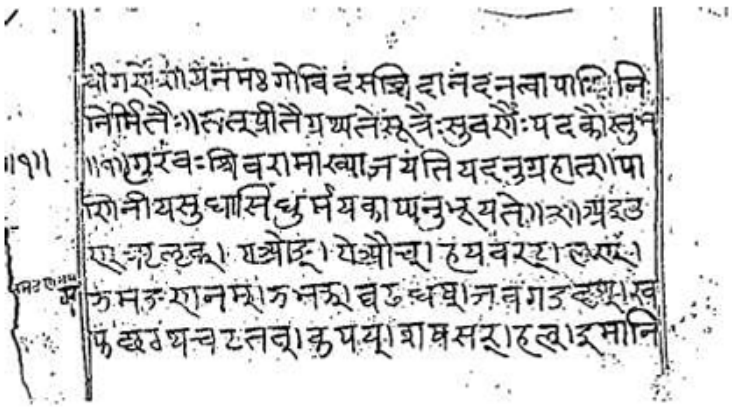


Figure 1e. Final Result After Pre-procesing

3. Performance Analysis with other Methods

We have compared the results of our method with that of Bernsen's[8], Niblack's[9], Sauvola's[10] and Gato's[12] methods. The results of the proposed methods were better qualitatively as well as quantitatively. Bernsen's method computes an adaptive threshold by using the maximum and minimum gray-level value of the local block which faces discrete threshold value problems. Niblack's method calculates a local threshold by using the mean and standard deviation value of gray-level image in a local block. In uneven illumination images, Niblack's method usually generates poor quality result. Sauvola's method is essentially an improved Niblack's method such that a variable r is added to Niblack's formula to change behavior from static to dynamic range standard deviation giving better results. Gato's method applies adaptive thresholding to deal with uneven illumination and degraded document images but, does not give satisfactory results when applied to manuscript images. Even though it eliminates noisy background effectively it suffers from non-uniform character formation which makes text legibility an issue when this method is applied, as can be seen in Figure 2d. Our proposed method gains its advantage over other methods from the use of adaptive Gaussian thresholding along with the use of a dual filtering technique which enhances legibility and effective binarization of text from background layer. See Figure 1e. The qualitative results that can be seen in Figure 2 correspond to various methods applied on manuscript image shown in Figure For quantitative estimation we used the well-known measures pseudo F-measure [21], PSNR[23], NRM[23-24], DRD[26], MPM[25] and geometric accuracy[27].

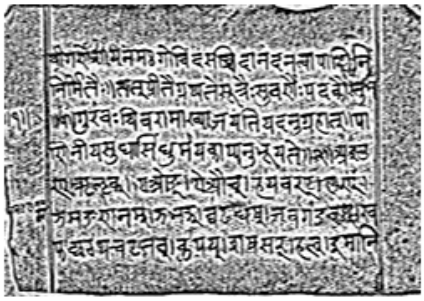


Figure 2a. Result from Bernsen's Method

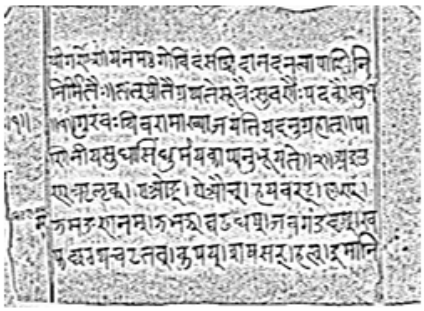


Figure 2b. Result from Niblack's Method

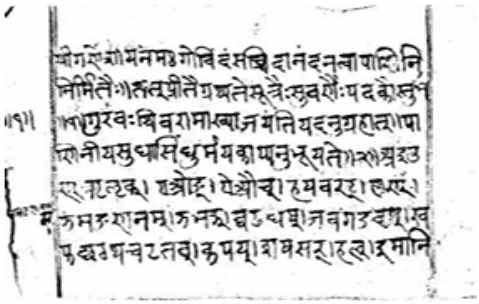


Figure 2c. Result from Sauvola's Method

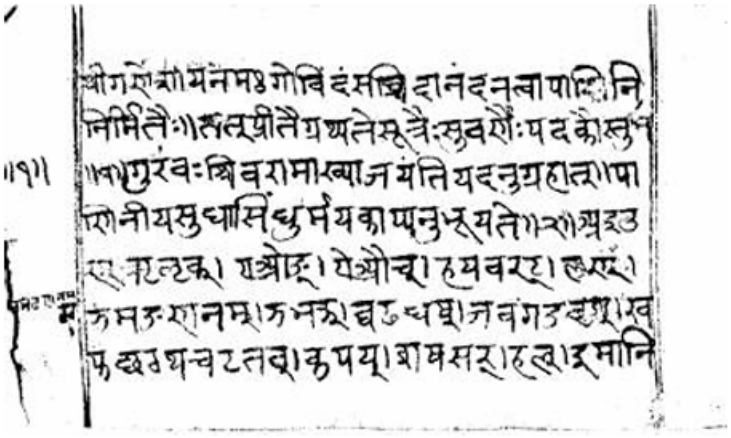


Figure 2d. Result from Gato's Method

Pseudo F-Measure

$$Pseudo\ F-measure = \frac{2 \times \text{p} - \text{Recall} \times \text{p} - \text{Precision}}{\text{p} - \text{Recall} + \text{p} - \text{Precision}} \quad (8)$$

where p-Recall is the pseudo-Recall and p-Precision is the pseudo-Precision

PSNR (Peak signal-to-noise ration)

$$PSNR = 10log\frac{C^2}{MSE} \quad (9)$$

$$\text{where } MSE = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N}\left(I(x,y)-I'(x,y)\right)^2}{MN}$$

C is the difference between the foreground and background. The higher the value of PSNR, the higher the similarity of the two MxN images is.

DRD (Distance reciprocal distortion)

The distortion of a processed image g(x,y) compared with the original image f(x,y) is measured by using a weighted matrix with each of its weights determined by the reciprocal of a distance measured from the center pixel.

$$DRD = \frac{\sum_{k=1}^{S}\text{DRD}_k}{NUBN} \quad (10)$$

where NUBN is to estimate the valid (non-empty) area in the image and it is defined as the number of non-uniform 8x8 blocks, DRDk is defined as:

$$DRD_k = \sum_{i,j}\left(D_k(i,j) * W_{N_m}(i,j)\right) \quad (11)$$

where WNm is the normalized weight matrix and Dk is given by

$$D_k(i,j) = |B_k(i,j) - g(x,y)_k| \quad (12)$$

Thus, DRDk equals to the weighted sum of the pixels of the block Bk of the original image that differ from the flipped pixel g(x,y)k in the processed image.

NRM (Negative rate metric)

The negative rate metric NRM is based on the pixel-wise mismatches between the ground truth image and processed image. It combines the false negative rate NRFN and the false positive rate NRFP. It is denoted as follows:

$$NRM = \frac{1}{2}(NR_{FN} + NR_{FP}) \quad (13)$$

where $NR_{FN} = N_{FN}/(N_{FN}+N_{TP})$, $NR_{FP} = N_{FP}/(N_{FP}+N_{TN})$

NTP denotes the number of true positives, NFP denotes the number of false positives, NTN denotes the number of true negatives, NFN denotes the number of false negatives. Binarization quality is better for lower NRM.

MPM (Misclassification Penalty Metric)

The Misclassification penalty metric MPM evaluates the prediction against the ground truth on an object-by-object basis. Misclassification pixels are penalized by their distance from the ground truth object's border.

$$MPM = \frac{1}{2}(MP_{FN} + MP_{FP}) \quad (14)$$

where $MP_{FN} = \frac{1}{D}(\sum_{i=1}^{N_{FN}} d_{FN}^{i})$, $MP_{FP} = \frac{1}{D}(\sum_{j=1}^{N_{FP}} d_{FP}^{j})$, d_{FN}^{i} and d_{FP}^{j} denote the

distance of the ith false negative and the jth false positive pixel from the contour of the text in the ground truth image. The normalization factor D is the sum over all the pixel-to-contour distances of the ground truth object. A low MPM score denotes that the algorithm is good at identifying an object's boundary. The evaluation was performed using the above described evaluation measures. The parameters for the other methods were tuned to give best possible result for a particular image. Table 1 shows the evaluation results for each method applied on the manuscript image shown in Figure 1a based on ground truth created semi-automatically.

Table 1. Evaluation Results

Method	Pseudo F-measure (%)	PSNR	DRD	NRM	MPM (x1000)	Geometric Accuracy
Bernsen	16.26	8.39	20.26	0.41	64.89	0.49
Niblack	20.06	10.09	11.65	0.41	33.17	0.47
Sauvola	23.27	11.23	7.33	0.40	2.33	0.46
Gato	32.22	9.94	12.23	0.30	4.53	0.65
Proposed Method	47.87	9.70	14.51	0.07	5.53	0.92

Table 1 shows the qualitative comparison of different methods. Our proposed method was applied on 27 digital manuscript images, the qualitative results of two can be seen in Figure 3 and Figure 4. The result after applying Gato's method is also shown to illustrate the better suitability of our method compared to

Gato's method applied on manuscript image. The quantitative comparison with other methods applied on manuscript image shown in Fig 3a and 4a can be seen in Table 2 and 3, respectively.

Figure 3a. Original Manuscript Image

Figure 3b. Result from Proposed Method

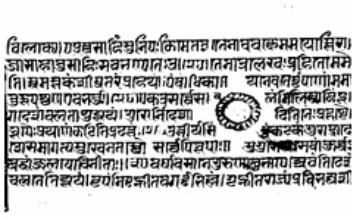


Figure 3c. Result from Gato's Method

Table 2. Evaluation results

Method	Pseudo F-measure (%)	PSNR	DRD	NRM	MPM (x1000)	Geometric Accuracy
Bernsen	38.29	7.29	13.26	0.36	63.02	0.57
Niblack	41.03	8.04	9.87	0.35	3.21	0.56
Sauvola	47.05	8.18	9.42	0.32	2.44	0.62
Gato	61.62	9.05	7.37	0.22	2.45	0.75
Proposed Method	62.17	9.31	6.58	0.23	2.60	0.74

Figure 4a. Original Manuscript Image

Figure 4b. Result from Proposed Method

Figure 4c. Result from Gato's Method

Table 3. Evaluation Results

Method	Pseudo F-measure (%)	PSNR	DRD	NRM	MPM (x1000)	Geometric Accuracy
Bernsen	28.91	6.48	7.90	0.43	18.35	0.44
Niblack	31.85	6.82	7.00	0.41	16.41	0.46
Sauvola	28.29	6.68	7.22	0.42	11.45	0.43
Gato	53.25	6.62	8.70	0.28	30.47	0.70
Proposed Method	65.02	8.11	5.89	0.21	23.17	0.78

The quantitative results for all methods applied on our dataset of 27 images can be seen in Table 4. Our method has improved uniform character formation compared to other methods although, in some manuscript images it suffers from inability to eliminate some background noise which manifests itself in the form of small black dots which can be seen in Figures 1e and 4b. This inability to eliminate all noisy pixels is a minor drawback in our method but, the improved character formation is a major advantage and could lead to better OCR readability compared to other methods which can be tested as future work.

Table 4. Evaluation Results

Method	Pseudo F-measure(%)
Bernsen	27.54
Niblack	30.21
Sauvola	32.78
Gato	50.05
Proposed Method	60.12

4. Conclusion

A novel technique for enhancing and binarizing digital manuscript images has been proposed which is shown to have much improved readability of characters. Our method is much more suitable for enhancing manuscripts compared to other document enhancing methods due to their unsuitability for use in damaged manuscript images. It also performs much better compared to other well-known methods qualitatively as well as quantitatively. Our work can be further expanded for enhancing other types of documents with different types of degradation for better OCR readability. However, our method suffers the drawback of not being able to eliminate noise completely and hence, as future work further modifications to the proposed method could be tested.

References

[1] J. M. C. Sousa, J. M. Gil, C. S. Ribeiro and J. R. C. Pintom, "Old Document Recognition Using Fuzzy Methods", International Journal of Intelligent Systems Technologies and Applications, vol. 1, no. ¾, (2006), pp. 263-279.

[2] Y. Leydier, F. Lebourgeois and H. Emptoz, "Text search for medieval manuscript images", Pattern

Recognition, vol. 40, no. 12, (2007), pp. 3552-3567.
[3] L. L. Sulem and M. Sigelle, "Recognition of degraded characters using dynamic Bayesian networks", Pattern Recognition, vol. 41, no. 10, (2008), pp. 3092-3103.
[4] S. R. Yahya, S. N. H. S. Abdullah, K. Omar, M. S. Zakaria and C. Y. Liong, "Review on Image Enhancement methods of Old Manuscripts with Damaged background", International Journal on Electrical Engineering and Informatics, vol. 2, no. 1, (2010). [5] N. Otsu, "A threshold selection method from gray-level histogram," IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, (1979), pp. 62-66.
[6] B. J. N. Kapur, P. K. Sahoo and A. K. C. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram", Computer Vision, Graphics, and Image Processing, vol. 29, (1985), pp. 273-285.
[7] J. Kittler and J. Illingworth, "Minimum error thresholding", Pattern Recognition, vol. 19, (1986), pp. 4147.
[8] J. Bernsen, "Dynamic thresholding of gray-level images", in Proceeding of the 8th International Conference on Pattern Recognition, (1986), pp. 1251-1255.
[9] W. Niblack, "An Introduction to Digital Image Processing", Strandberg, Denmark, (1986).
[10] J. Sauvola and M. Pietikainen, "Adaptive document image binarization", Pattern Recognition, vol. 33, (2000), pp. 225-236
[11] Q. Huang, W. Gao and W. Cai, "Thresholding technique with adaptive window selection for uneven lighting image", Pattern Recognition Letter, vol. 26, (2005), pp. 801-808.
[12] B. Gatos, I. Pratikakis and S. Perantonis, "Adaptive degraded document image binarization", Pattern Recognition, vol. 39, (2006), pp. 317-327.
[13] B. Gangamma, K. S. Murthy and A. Vikas Singh, "Restoration of Degraded Historical Document Image", Journal of Emerging Trends in Computing and Information Sciences, vol. 3, no. 5, (2012).
[14] I. Quraishi, M. De, K. G. Dhal, S. Mondal and G. Das "A Novel Hybrid Approach To Restore Historical Degraded Documents", ISSP, vol. 1, (2013), pp. 185-189.
[15] B. M. Singh, R. Sharma, D. Ghosh and A. Mittal, "Adaptive binarization of severely degraded and non uniformly illuminated documents", International Journal on Document Analysis and Recognition (IJDAR), vol. 17, no. 4, (2014), pp. 393-412
[16] Kale, Gandhe, Phade and Dhulekar, "Enhancement of Old Images and documents by Digital Image Processing techniques", ICCICT, (2015).
[17] A. Tomar, A. Raj, S. Chaudhury and S. Indu, "Enhancement and Retrieval of Historic Inscription Images", In ComputerVision-ACCV 2014 Workshops, Springer International Publishing, (2014).
[18] D. Lu, X. Huang, C. Liu, X. Lin, H. Zhang and J. Yan, "Binarization of degraded document image based on contrast enhancement", Proceedings of the 35th Chinese Control Conference, Chengdu, China, (2016).
[19] D. Comaniciu and P. Meer, "Mean Shift: A robust approach toward feature space analysis", IEEE Transactions on Pattern Analysis and Machine Intelligence, (2002).
[20] Gaussian Window and Transform, ccrma.stanford.edu, (2016).
[21] K. Ntirogiannis, B. Gatos and I. Pratikakis, "Performance Evaluation Methodology for Historical Document Image Binarization", IEEE Transactions On Image Processing, vol. 22, no. 2, (2013).
[22] http://docs.opencv.org/3.0beta/modules/imgproc/doc/filtering.html#pyrmeanshiftfiltering
[23] I. Pratikakis, B. Gatos and K. Ntirogiannis, "H-DIBCO 2010 - handwritten document image binarization competition", in Proceedings Int. Conf. Frontiers Handwrit. Recognit., Kolkata, India, (2010), pp. 727732.
[24] B. Gatos, K. Ntirogiannis and I. Pratikakis, "DIBCO 2009: Document image binarization

contest", International Journal Document Anal. Recognit., vol. 14, no. 1, (2011), pp. 35–44.
[25] D. P. Young and J. M. Ferryman, "PETS Metrics: On-Line Performance Evaluation Service", ICCCN '05 Proceedings of the 14th International Conference on Computer Communications and Networks, (2005), pp. 317-324.
[26] H. Lu, A. C. Kot and Y. Q. Shi, "Distance-reciprocal distortion measure for binary document images", IEEE Signal Process. Lett., vol. 11, no. 2, (2004), pp. 228–231.
[27] R. Paredes, E. Kavallieratou and R. D. Lins, "ICFHR 2010 contest: Quantitative evaluation of binarization algorithms", in Proceedings Int. Conf. Frontiers Handwrit. Recognit., Kolkata, India, (2010), pp. 733–736.

ISSN: 2233-7857

Detection and Classification of Leaf Disease using Machine Learning Approach

Arpana Mahajan, Sheshang Degadwala, Anuragsinh Gohil, Rohan Brahmbhatt
Department of Computer Engineering
Sigma Instiute of Engineering

ABSTRACT

We propose and tentatively assess a product answer for programmed recognition and characterization of plant leaf diseases. Moreover, even though that person has always been bad in biology but would like to know more about that plant. It simply means that he/she is interested to explore his/her knowledge in this particular area. He might be interested to know its name or about its specific features. Sometimes, he/she might be interested to search a plant if it is rare or on the verge of the existence. Indeed, even today, distinguishing proof and arrangement of obscure plant species are performed physically by master individual who are not many in numbers. Here, we are introducing another acknowledgment approach dependent on Leaf Features Fusion and Random forest (RF) Classification calculations for characterizing the various kinds of plants. The proposed methodology comprises of four stages that are preprocessing, division, highlight extraction and grouping stages. Since most sort of plants have one of a kind leaves. Leaves are not exactly equivalent to each other by attributes, for example, the shape, shading, surface and vein. There are numerous highlights of leaf, for example, Color highlights, Vein highlights, GLCM highlights, Shape highlights and Gabor highlights. These all highlights are melded by idea linking of two vectors. Along these lines, order approach displayed in this exploration relies upon plant leaves. Test results show precision and different parameters estimated in this methodology with combination of every one of these highlights or their various mix. This is a clever framework which can recognize tree species from photos of their leaves and it gives precise outcomes in less time. Here we are developing Asp.net and C# base Web application that support small farmers in analyzing the damage to their plants quickly and easily.

Keywords— Histogram Equalization, Gaussian Filter, Haralick, Random forest.

I. INTRODUCTION

Reliable, exact appraisals from claiming ailment force level are incredulous for A large number investigate ranges for plant pathology, including assessments about social What's more crop misfortune models, dispersal gradients, malady forecasting, understanding associations between manifestations and the environment, Also assessing germplasm for imperviousness with pathogens. In spite of the vitality from claiming guaranteeing that phytopathometric appraisals are both exact What's more precise, these appraisals need aid commonly inferred starting with visual appraisals during Different levels from claiming determination (plots, plants, or tissues). A regular approach to quantifying malady force level may be with gauge its severity, characterized here Likewise those. Amount (e. G., area) from claiming ailing plant tissue relative of the aggregate add up of powerless tissue accessible. In addition, seriousness might a chance to be communicated as those amount about lesions for every testing unit for rusts and other maladies. Sickness incidence, characterized concerning illustration the ailing extent of a population's inspecting units (plant, organs) might additionally demonstrate illness force. Frequency Furthermore different check information need aid naturally lesquerella liable to inclination and errors of

observation toward raters, likewise compared for visual estimates about infection seriousness. This undertaking serves farmers and the other clients who takes consideration of the plants Anyway some way or another there are large portions instances Previously, which plants endure a considerable measure from claiming sickness like virus, microscopic organisms Other than progressions, etc all through along these lines, watching and stock course of action of all instrumentation might be enha. And a number different should beat this are malady our one task aides the farmers and the other clients on assistance plants On recouping its wellbeing.

II. RELATED WORKS

Explaining the research model, theory, the technique of collecting the data, a technique of analyzing the S. Kaur, S. Pandey, and S. Goel proposed for self-loader leaf malady location and arrangement framework for soybean group. From the investigation, grayscale pictures are anything but difficult to process and execute for different applications since they have better clearness and appropriate for examination than RGB pictures. Histogram night out is used to redesign the separation of the photos and gives a clear picture to human eyes. Histogram leveling is used to achieve a better quality picture in grayscale which is used in the various helpful applications, normal application, for instance, electronic X-bars, plant leaves sickness, and so forth. In this way, this kind of pictures will be utilized to investigation and determination the plant leaves maladies and decides the sicknesses dimension of the plant leaves [1].

J. Garcia and A. Barbedo proposed technique for Digital picture preparing procedures for recognizing, measuring and ordering plant ailments. This paper endeavored to exhibit an extensive overview of the issue. Because of the enormous number of references, the depictions are short, giving a smart graph of the thoughts fundamental age of the arrangements. Highlight that the work regarding the matter isn't obliged to what was appeared. Numerous papers regarding the matter couldn't be incorporated into a request to monitor the paper length - the papers were chosen as to think about the biggest number of various issues as could reasonably be expected. In this way, if the peruse wishes to achieve a progressively complete comprehension on a given application or issue, he/she can allude to the lists of sources of the separate articles [2].

R. Anand, S. Veni, and J. Aravinth proposed methodology for usage of picture getting ready framework for recognizable proof on Brinjal leaves using K-infers Clustering Method. A system for area and portrayal of leaf affliction is realized. The division of the incapacitated is figured it out. The division of the lamentable part is finished utilizing the K-Means division. By at that point, GLCM surface parts are extraction and a depiction is finished utilizing SVM. The structure has looked for after a zone of sullying in citrus leaves. Future work is to be produced for offers of planning in various plant species and to improve the outline precision [3].

R. Meena Prakash, G. P. Saraswathy, G. Ramalakshmi, K. H. Mangaleswari, and T. Kaviya proposed theory for Detection of Leaf Disease and Classification using Digital Image managing. This paper, for the most part, thinks about to the changed region of tomato aggravations and illness subject to the leaf surface. The affirmation models are set up to depict the tomato ailment and pasts in terms of professional career learning improvement, which accomplishes a standard game-plan accuracy of 89%. By and by, the general world-class depends upon relative astounding test pictures, future research will concentrate on the tangle figurings to see tomato bugs and sickness subject to relative low-quality leaf pictures [4].

S. Goel, S. Pandey, and S. Kaur proposed a technique for Plants Disease Identification and Classification through Leaf Images: A Survey. In this paper, disease conspicuous confirmation is somewhat less troublesome than its authentic plan. Once in a while it winds up difficult for a master to arrange a particular infection with 100% conviction. Improvement of systems that can arrange

distinctive infectious, viral and bacterial diseases precisely may in like manner be locked in. Composing considers minerals or enhancements needs as another sort of plant sickness. The improvement of systems that can suitably isolate among defilement and an absence of thought as an incredibly inconvenient focus in light of the way that from the ace perspective detaching a polluted leaf from a lacking leaf is a capricious undertaking[5].

Jailani, M. T. Nooritawati, and U. Mara proposed methodology for Orchid Leaf Disease Detection using Border Segmentation Techniques. An image setting up the to count to find the disturbance assertion and seeing evidence is proposed. The pepper plant leaves are gone in as the heading of leaves in seeing leaf contamination. The count passes on better results and solid and terrible plants can be separated with the assistance of this calculation. With this picture examination system, unprecedented sound pepper plants can be expelled out from a making ranch which expands the capability the closeness of disease by survey the visual side effects seen on the leaves of the plants [6].

III. PROPOSED APPROACH

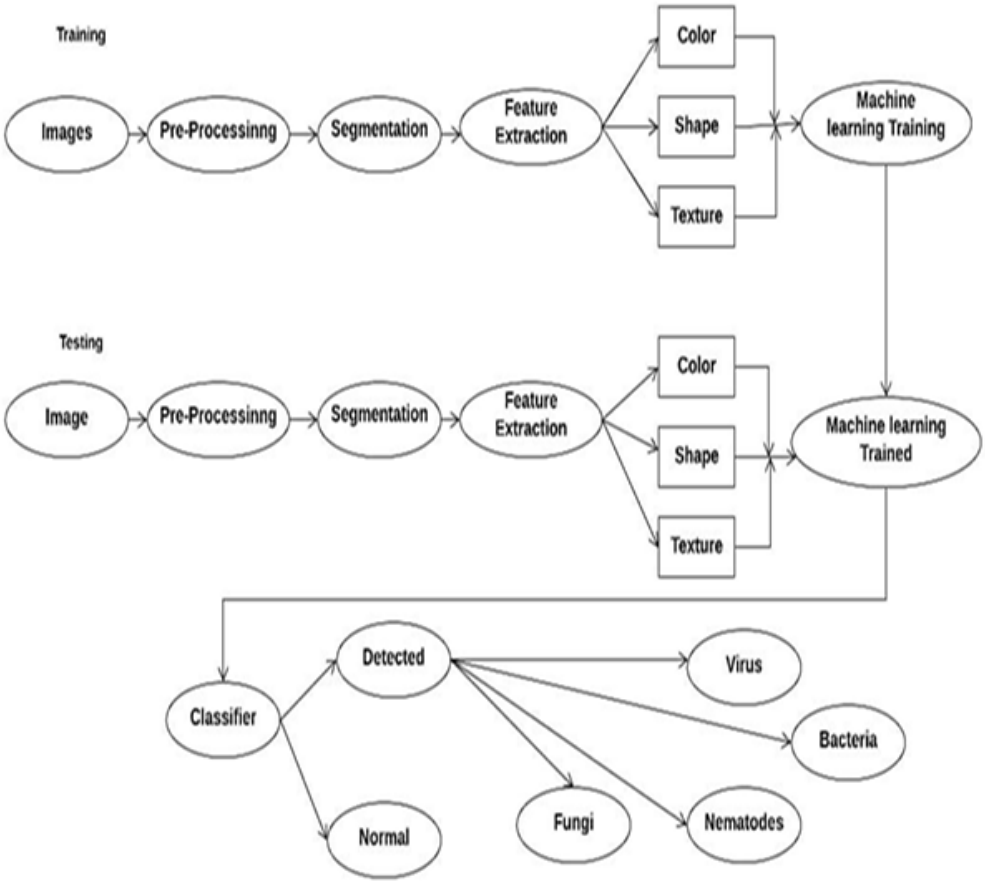


Fig. 1. Proposed Block Diagram

Training

Step 1: Select or upload images and its Label.

Step 2: Apply Pre-Processing using Histogram Equalization and Noise Removing Filter on whole image datasets.

Step 3: Apply Color and Cluster Based Combine Segmentation Approach.

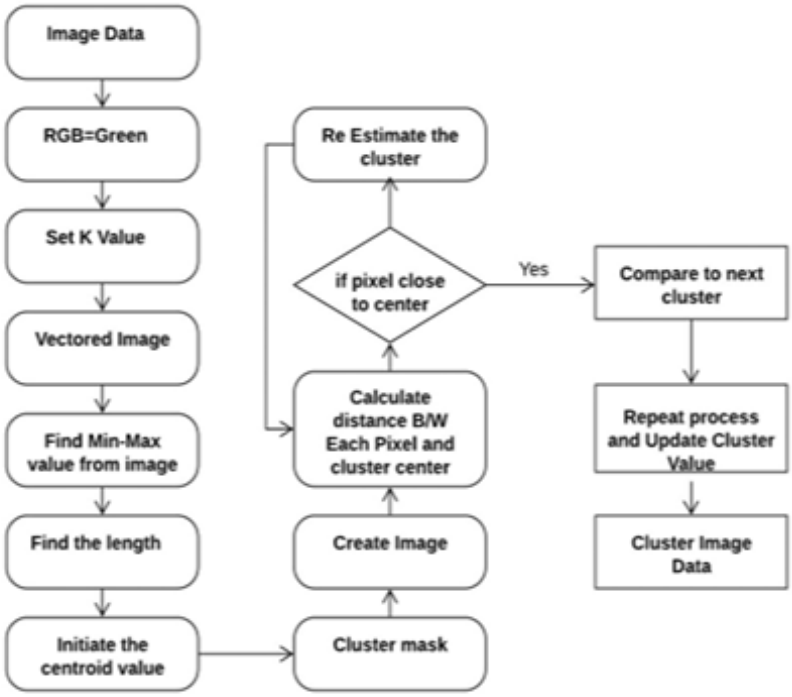


Fig. 2. Segmentation Process Diagram

Step 4: Extract Shape, Color, Vein and Texture Features for all images.

Table I: Comparison of different features

Feature	Type	Description
Shape	Area, Perimeter, Major and Minor Axis	Works with Binary Image only
Color	Color Moment	Consider Mean, STD, Skewness, Kurtosis
Texture	GLCM	Contrast, Correlation, Entropy and Variance
Vein	Vein Length	Area
Invariant	Zernike Moment	Angle and Value

Step 5: Apply Machine Learning Approach RF and make database.

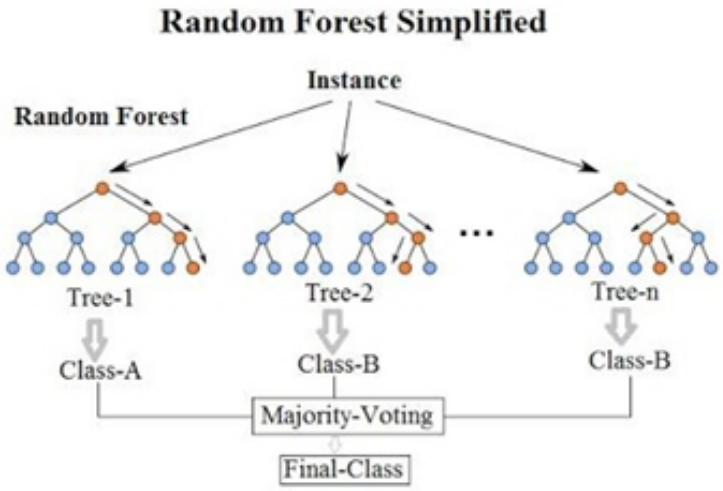


Fig. 3. Classification Random Forest

RF is an accumulation of tree-organized classifiers where each tree relies upon the estimations of arbitrary vector examined freely and the conveyance of all trees in the woodland.

Testing

Step 1: Select or upload image.
Step 2: Apply Pre-Processing using Histogram Equalization and Noise Removing Filter.
Step 3: Apply Color and Cluster Based Combine Segmentation approach.
Step 4: Extract Shape, Color, Vein and Texture Features.
Step 5: Apply Machine Learning classification Approach RF using database.
Step 6: Classify Disease type.

IV. RESULTS

Table II: Pre-Processing Results

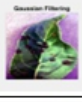
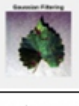
Type	Original	Filtering	Equailization
Normal			
Virus			
Fungi			
Bacteria			
Nematodes			

Table III: Segmentation and Feature Extraction

Type	Green Segmentation	Segmentation	Features
Normal			
Virus			
Fungi			
Bacteria			
Nematodes			

Table IV: Result Analysis

Algorithm	Accuracy	Precision	Recall
RF	95.83%	96%	96%
SVM	79.17%	84.50%	79%

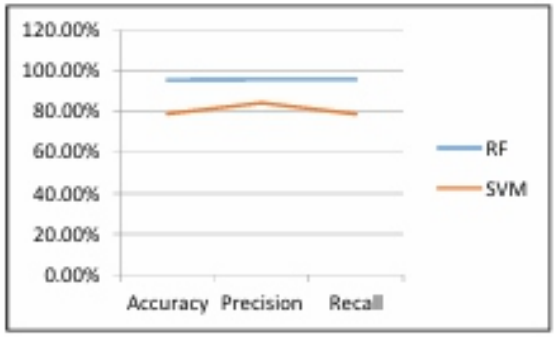


Fig. 4. Graph of Accuracy, Precision and Recall

V. CONCLUSION

An acknowledged activity of Computer Aided Plant deformity recognizable proof frameworks contains four stages, to be specific, engineering the leaf database, blessed messenger improvement, investigation (extraction of leaf), highlight extraction and characterization. The capital aim of this analysis assignment is to propose techniques that enhance anniversary operation of bulb defect identification through leaf features. A technique for leaf grouping has been created. The strategy consolidates shape and vein, shading, and surface highlights and uses RF as a classifier. We infer that joining vein for

highlight descriptors is a plausible option for grouping basically complex pictures. GLCM offer excellent invariance highlights and uncover upgraded execution than other minute based arrangements. Arbitrary Forest Classifier gives preferred exactness over some other classifier. We have use highlights combination to perceive plant deformity with precision over 90%. Regardless of the way that show of the system is satisfactory, we believe that the introduction still can be improved. Consequently, different highlights will be inquired about later on.

REFERENCES

[1] H. Q. Cap, K. Suwa, E. Fujita, S. Kagiwada, H. Uga, and H. Iyatomi, "A deep learning approach for on-site plant leaf detection," Proc. - 2018 IEEE 14th Int. Colloq. Signal Process. It's Appl. CSPA 2018, no. March, pp. 118–122, 2018.

[2] R. Anand, S. Veni, and J. Aravinth, "An application of image processing techniques for detection of diseases on brinjal leaves using k-means clustering method," 2016 Int. Conf. Recent Trends Inf. Technol. ICRTIT 2016, 2016.

[3] R. Meena Prakash, G. P. Saraswathy, G. Ramalakshmi, K. H. Mangaleswari, and T. Kaviya, "Detection of leaf diseases and classification using digital image processing," Proc. 2017 Int. Conf. Innov. Information, Embed. Commun. Syst. ICIIECS 2017, vol. 2018-Janua, pp. 1–4, 2018.

[4] J. Shijie, J. Peiyi, H. Siping, and Sl. Haibo, "Automatic detection of tomato diseases and pests based on leaf images," Proc. - 2017 Chinese Autom. Congr. CAC 2017, vol. 2017-January, pp. 3507–3510, 2017.

[5] R. Jailani, M. T. Nooritawati, and U. Mara, "Orchid Leaf Disease Detection using Border Segmentation Techniques Faculty of Electrical Engineering," no. December, pp. 12–14, 2014.

[6] J. Francis, AntoSahayaDhas D, and Anoop B K, "Identification of leaf diseases in pepper plants using soft computing techniques," pp. 168–173, 2016.

[7] N. Krithika and A. Grace Selvarani, "An individual grape leaf disease identification using leaf skeletons and KNN classification," Proc. 2017 Int. Conf. Innov. Information, Embed. Commun. Syst. ICIIECS 2017, vol. 2018-January, pp. 1–5, 2018.

[8] P. B. Padol and A. A. Yadav, "SVM classifier based grape leaf disease detection," Conf. Adv. Signal Process. CASP 2016, pp. 175–179, 2016.

[9] P. B. Padol and S. D. Sawant, "Fusion classification technique used to detect downy and Powdery Mildew grape leaf diseases," Proc. - Int. Conf. Glob. Trends Signal Process. Inf. Comput. Commun. ICGTSPICC 2016, pp. 298–301, 2017.

[10] F. T. Pinki, N. Khatun, and S. M. M. Islam, "Content based paddy leaf disease recognition and remedy prediction using support vector machine," 20th Int. Conf. Comput. Inf. Technol. ICCIT 2017, vol. 2018January, pp. 1–5, 2018.

[11] P. R. Rothe and R. V. Kshirsagar, "Automated extraction of digital images features of three kinds of cotton leaf diseases," 2014 Int. Conf. Electron. Commun. Comput. Eng. ICECCE 2014, pp. 67–71, 2014.

[12] H. Anand Kulkarni and R. K. Ashwin Patil , "Applying Image Processing Technique to Detect Plant Diseases," International Journal of Modern Engineering Research, vol. 2, no. 5, pp. 3661-3664, September October 2012.

[13] M. Suresha, K. N. Shreekanth, and B. V. Thirumalesh, "Recognition of diseases in paddy leaves using knn classifier," 2017 2nd Int. Conf. Converg. Technol. I2CT 2017, vol. 2017-January, pp. 663–666, 2017.

[14] K. Thangadurai and K. Padmavathi, "Computer visionimage enhancement for plant leaves

disease detection," Proc. - 2014 World Congr. Comput. Commun. Technol. WCCCT 2014, pp. 173–175, 2014.

[15] S. Kaur, S. Pandey, and S. Goel, "Semi-automatic leaf disease detection and classification system for soybean culture," IET Image Process., vol. 12, no. 6, pp. 1038–1048, 2018.

Instructions for Authors

Essentials for Publishing in this Journal

1 Submitted articles should not have been previously published or be currently under consideration for publication elsewhere.

2 Conference papers may only be submitted if the paper has been completely re-written (taken to mean more than 50%) and the author has cleared any necessary permission with the copyright owner if it has been previously copyrighted.

3 All our articles are refereed through a double-blind process.

4 All authors must declare they have read and agreed to the content of the submitted article and must sign a declaration correspond to the originality of the article.

Submission Process

All articles for this journal must be submitted using our online submissions system. http://enrichedpub.com/ . Please use the Submit Your Article link in the Author Service area.

Manuscript Guidelines

The instructions to authors about the article preparation for publication in the Manuscripts are submitted online, through the e-Ur (Electronic editing) system, developed by **Enriched Publications Pvt. Ltd**. The article should contain the abstract with keywords, introduction, body, conclusion, references and the summary in English language (without heading and subheading enumeration). The article length should not exceed 16 pages of A4 paper format.

Title

The title should be informative. It is in both Journal's and author's best interest to use terms suitable. For indexing and word search. If there are no such terms in the title, the author is strongly advised to add a subtitle. The title should be given in English as well. The titles precede the abstract and the summary in an appropriate language.

Letterhead Title

The letterhead title is given at a top of each page for easier identification of article copies in an Electronic form in particular. It contains the author's surname and first name initial .article title, journal title and collation (year, volume, and issue, first and last page). The journal and article titles can be given in a shortened form.

Author's Name

Full name(s) of author(s) should be used. It is advisable to give the middle initial. Names are given in their original form.

Contact Details

The postal address or the e-mail address of the author (usually of the first one if there are more Authors) is given in the footnote at the bottom of the first page.

Type of Articles

Classification of articles is a duty of the editorial staff and is of special importance. Referees and the members of the editorial staff, or section editors, can propose a category, but the editor-in-chief has the sole responsibility for their classification. Journal articles are classified as follows:

Scientific articles:

1. Original scientific paper (giving the previously unpublished results of the author's own research based on management methods).

2. Survey paper (giving an original, detailed and critical view of a research problem or an area to which the author has made a contribution visible through his self-citation);

3. Short or preliminary communication (original management paper of full format but of a smaller extent or of a preliminary character);

4. Scientific critique or forum (discussion on a particular scientific topic, based exclusively on management argumentation) and commentaries. Exceptionally, in particular areas, a scientific paper in the Journal can be in a form of a monograph or a critical edition of scientific data (historical, archival, lexicographic, bibliographic, data survey, etc.) which were unknown or hardly accessible for scientific research.

Professional articles:

1. Professional paper (contribution offering experience useful for improvement of professional practice but not necessarily based on scientific methods);
2. Informative contribution (editorial, commentary, etc.);
3. Review (of a book, software, case study, scientific event, etc.)

Language

The article should be in English. The grammar and style of the article should be of good quality. The systematized text should be without abbreviations (except standard ones). All measurements must be in SI units. The sequence of formulae is denoted in Arabic numerals in parentheses on the right-hand side.

Abstract and Summary

An abstract is a concise informative presentation of the article content for fast and accurate Evaluation of its relevance. It is both in the Editorial Office's and the author's best interest for an abstract to contain terms often used for indexing and article search. The abstract describes the purpose of the study and the methods, outlines the findings and state the conclusions. A 100- to 250-Word abstract should be placed between the title and the keywords with the body text to follow. Besides an abstract are advised to have a summary in English, at the end of the article, after the Reference list. The summary should be structured and long up to 1/10 of the article length (it is more extensive than the abstract).

Keywords

Keywords are terms or phrases showing adequately the article content for indexing and search purposes. They should be allocated heaving in mind widely accepted international sources (index, dictionary or thesaurus), such as the Web of Science keyword list for science in general. The higher their usage frequency is the better. Up to 10 keywords immediately follow the abstract and the summary, in respective languages.

Acknowledgements

The name and the number of the project or programmed within which the article was realized is given in a separate note at the bottom of the first page together with the name of the institution which financially supported the project or programmed.

Tables and Illustrations

All the captions should be in the original language as well as in English, together with the texts in illustrations if possible. Tables are typed in the same style as the text and are denoted by numerals at the top. Photographs and drawings, placed appropriately in the text, should be clear, precise and suitable for reproduction. Drawings should be created in Word or Corel.

Citation in the Text

Citation in the text must be uniform. When citing references in the text, use the reference number set in square brackets from the Reference list at the end of the article.

Footnotes

Footnotes are given at the bottom of the page with the text they refer to. They can contain less relevant details, additional explanations or used sources (e.g. scientific material, manuals). They cannot replace the cited literature.
The article should be accompanied with a cover letter with the information about the author(s): surname, middle initial, first name, and citizen personal number, rank, title, e-mail address, and affiliation address, home address including municipality, phone number in the office and at home (or a mobile phone number). The cover letter should state the type of the article and tell which illustrations are original and which are not.

Address of the Editorial Office:

Enriched Publications Pvt. Ltd.
S-9,IInd FLOOR, MLU POCKET,
MANISH ABHINAV PLAZA-II, ABOVE FEDERAL BANK,
PLOT NO-5, SECTOR -5, DWARKA, NEW DELHI, INDIA-110075,
PHONE: - + (91)-(11)-45525005