International Journal of Hybrid Information Technology
Vol. 10, No.2 (2017), pp. 97-106
http://dx.doi.org/10.14257/ijhit.2017.10.2.10

Regenerate the Shredded Documents by using Memetic Algorithm

Fozia Hanif Khan[1*], Rehan Shams[2], Dur-e- shawar Agha[3] and Rabia Noor Enam[4]

[1,] Department of Mathematics,
[2]Department of Telecommunication,
[3,4]Department of Computer Science, Sir Syed University of Engineering and Technology Karachi, Pakistan
[1]ms-khans2011@hotmail.com, [2]r.shams@hotmail.com,
[3]engr.dureshwaragha@gmail.com, afaq_rabia@yahoo.com

Abstract

It seems quite interesting to reconstruct the destroyed documents. On the other hand it is possible that documents might be mistakenly destroyed by someone. Another useful application of this type is in the field of forensics and archeology for the restoring of the ancient documents. To reduce the availability of information documents are shredded. The advent of new and effective ways has developed a convenient way for the people to get rid of important information so that it could not get into the hands of others. It includes the use of shredder to render the information on the pages useless and usually eliminating of proof is been practiced. The same scenario has been dealt hereby and an efficient way of rearranging those strips, known as shreds, to recreate the original document is been designed. This paper presents an efficient way of reconstructing the documents by genetic algorithm which is an extended version of memetic algorithm by emerging the variable neighborhood search (VNS). This computer based algorithm deals with an input of multiple shreds of a single page, which are to be rearranged in order to make the text on it understandable. A few Image Processing techniques are been used to get back what was intended to be destroyed. This consumes less time as compared to manual rearranging with promising results in a form of Image. Recommendations are also indicated to improve the algorithm.

***Keywords**: Shred, Pixels, Xoring, memetic algorithm, image processing*

1. Introduction

It is very obvious now a days that many confidential documents are prepared and strolled electronically due to some legal reasons. But by the passage of time the printed form of these documents are then destroyed by using mechanical machines which makes these papers almost impossible to read. Paper shredders are easily available in stationary stores and one may use them to destroy their confidential documents, as do most organizations. In some places, shredders are used to destroy potentially criminal evidence making police and forensic services to recover lost documents using every possible way. Methodology for reconstruction of shredded documents could therefore be extremely useful to the concerned authorities, if there is a chance to retain almost all of the conversation or documentation in black and white.

In this study we are focusing on the reformation of the rectangular shaped shredded documents. As the process of information retrieval is mainly connected with the pattern recognition or image processing, therefore we are trying to make the process of reconstruction of shredded papers possible by using the combinatorial optimization and the process of pattern recognition for producing promising results.

This process, if done manually, require a lot of time and man power to achieve little result, so a motivation to automate this process is been aroused and hence been created.[1] [2] [3].

2. Virtual Requirements

Paper shredding is a common cryptographic operation, like block cipher. Most of the important documents that cannot be stored electronically but are printed on the papers are sometimes need to be shred for the sake of security. Although there are some services which claim to reconstruct your shredded documents, but all of them are very expensive and it is almost impossible to handover your secrete information to somebody else.

There are several ways by which we can destroy the confidential evidence such as burning and crushing. But destroying the secrete information either by burning procedure or by crushing, it is almost impossible to reconstruct them again. However the shredded documents are able to reconstruct even if some of the shreds are destroyed or missing.

3. Previously Developed Methodologies

This field can be classified in many sub domain, such as the restoration of hand torn paper documents [4,5] and the rejoining of cross-cut shredded (text) documents (RCCSTD) [6,7,8]. Another relevant methodology is the (computer aided) by solving jigsaw puzzles [9,10]. Currently there is a service called “Unshredder”, they claim to reassemble all your shredded documents and cite themselves as “the first commercial document reconstruction tooling the world”[11]. To subscribe these unshredder services it cost thousand off dollars per year, they use special software available with the additional costs and they will recover only a limited numbers of documents. There are several guide available [12] that shows how to reconstruct the shredded documents by hand. This is the most easily accessible procedure to assemble shredded pieces of papers. Autostitch and CleVr [13] is a software which is quite similar to the documents reconstruction. Similar to autostitch, photo stitching can combine many small pieces to make larger image [14]. Dynamic programming [24] [25] for puzzle solving has been used for matching the documents. Wolfson etal. An algorithm provided in [26] to solve large puzzles but with some constraints regarding the shape of the puzzles pieces.

Requirements are differentiated into ‘must’ for essential requirements, ‘should’ for desirable, but not-necessary requirements. The basic program contains all essential requirements but, further improvements may also be apply for some or all of the not-necessary requirements. The Input, Processing & Output requirements are defined as below:

4. Problem Formulation

Shredding is an n-page documents D is a process of cutting each n page $p_{0<i<n} \in D$ in to m rectangular pieces called strips by mechanical shredding. There are several types of paper shredding that comes in different shapes and sizes. Three types of shredding are the main categories which are as follows,

i. Strip-cut: The most cheapest and common shredding. In this type the paper is cut in vertical shape called strip, whose width varies according to the requirement of security level.

ii. Cross-cut: For the sake of extra security this type of shredding is used. In this type of shredding the paper is cut into both horizontal and vertical small rectangles. Again the size varies according to the requirement.

iii. Other: Several other types of industrial and special shredding which falls outsides the above mentioned category. This procedure includes the things such as grinders that have rotating blades and make very small pieces of papers. These types of shredding are very rarely used and we are not consider them right here.

4.1. Input Requirements

First of all the shredded strip of any text material that has to be reconstructed will be scanned and these scanned images are the main inputs to the system. Logically, the strips should not have to be in any particular order before they are scanned (except for the first strip), they just have to be in the left corner of the page as for perfect perpendicular image as it has to be cropped manually later on.

The strips should have the following standard:

- Scanned images must be able to accept by the program.
- Random size of scanned image should be accepted by the program (but all the strips should be same sized row into column).
- The file formats can be any image format necessary (*i.e.* jpeg, png), but it would have to be amended into program first.
- The program should accept variety of document as possible (*i.e.* printed text and written text).

4.2. Processing Requirements

These requirements deal with the variables and prerequisite demands for the processing task and also the result expectations from each of the processes involving in it. The process requirements of our project are defined as below:

4.2.1. Extraction

- Extraction process of images must be as automated as possible (*i.e.* no need to specify the location by the user of the strips on the image).
- The program must have the ability to extract all the matched relevant strips from a scanned image with the exclusion of other irrelevant un-matched pixels.
- As an initial data structure the program is able to extract strips to work from.
- All extracted strips should be identical in size and shape of the same image.

4.2.2 Fitness Calculations

- In order to calculate how it is close to the original document, the program must be able to process the combined image using feature extraction or Optical Character Recognition (OCR).
- For the fitness measure the output evaluation should be between 0 to 1 by the program.

4.2.3. Document Reconstruction

- By taking all possible combinations of strips program must use heuristic approach which is actually an attempt to improve the initial estimate, and after that it uses some sort of evaluation criteria to detect whether an improvement has been done or not.
- Double sided documents should be able to process by the program, together with the condition that user must specifies which strips matches with each other.
- Multiple documents should be able process by the program, and the result should present as separate images [4] [15][5] [16].

4.3. Output Requirement

4.3.1. Functional Requirement

The program should generate the final output to be as an image that must be readable in such case, where the original input includes text contents. The output image should be appropriate for OCR program if there is any text contents present in the original documents.

4.3.2. Non-Functional Requirements

- Reasonable amount of time must be taken by the program to run.
- For the possible removal of errors, the code must be as thoroughly checked and tested.
- There must be a well-commented, clear and modular code as per the requirement.

5. Methodology

Memetic algorithm is a method to achieve the required task, our process is been divided into four basic sections, it includes initial population by Data Extraction' from the input to make it computable then 'Scanning' of the better individual (shreds), 'Processing' of the entities in the form of crossover and finally 'Recombination' to deliver a final result in the form of mutation

Genetic algorithm is kind of evolutionary algorithm that can be adopted to achieve the optimization. In this study we are considering the memetic algorithm that can be consider as a merger of genetic algorithm (GA) and local optimum procedure. We are dealing with the local improvement with the neighborhood search criteria. Genetic algorithm has basically consists of four unique steps. First is the generation of initial population by selecting the most favorable features in the form of array of strings, after this, the newly created individuals will go through the fitness evaluation that measures which individual will be selected as a parent and will go under the procedure of crossover. The crossover procedure sometimes may provide the local optima, to achieve the global optima we perform the procedure of mutation that also defines the stopping procedure of the any genetic algorithm.

As define earlier the proposed methodology is memetic algorithm which is actually the combination of several procedures. The provided study is basically unshredding the documents by using the memetic algorithm and image processing. In the following we discuss the individual part of the algorithm in detail.

5.1. Initial Population

Generate the initial population as an array of bits in which each strip can be taken as long vertical strip. The size and the width of each strip can be vary. In this proposed algorithm we generate the initial population with emergence of different techniques. It is really very important that the initial population provides better result in the process of genetic algorithm because the better individual will further consider for the process of crossover. The procedure starts with the Row building Heuristic (RBH) technique [17], according to this technique we start with the blank side of the page, since every page must have the blank side either on the left or right. Therefore we start our searching procedure with the left or right side of the page.

Consider a shredded document Ð = {S1, S2,. . ., Sn} baised of ***n*** small fragments. The algorithm will try to match the fragment Ғ1 with all the other fragments for the best matching. The aim is to match all four sides of the strip, *i.e.*, form the left, right and making the upside down. In this way the two strip with the most feature is supposed to be the adjacent strip. The fragments Si,p and $S_{i,q}$ with the maximum matching features will become $S_{p.q..}$

5.1.1. Matching of Two Strip $S_{i,p}$ and $S_{i,q}$:

i. Find match between the right edge of $S_{i,p}$ to the left edge of $S_{i,q.}$
ii. Match between the right edge of $S_{i,p}$ and the inverted right edge of $S_{i,q.}$
iii. Match between the inverted left edge of $S_{i,p}$ and the left edge of $S_{i,q}$
iv. Match between the inverted left edge of $S_{i,p}$ and the inverted right edge of $S_{i,q.}$

5.2. Fitness Evaluation

For the procedure of fitness evaluation we use the process of scanning or image processing in which the strips are placed on the scanner and see the result on the screen. The process of scanning, as we all know, is to convert a hard copy of a document into a digital form for further processing. Use of a scanner is necessary for the purpose as this device allows the conversion of real world documentation, images or any type of written information into a digital format. We collect all the paper strips and laid them on the scanner by making sure that all the strips do not touch or overlap each other.

Scanning of the strips at this stage is to be a preliminary part of our algorithm, meaning that pre-scanned images are to be inputted to the algorithm to be processed further.

5.3. Data Extraction

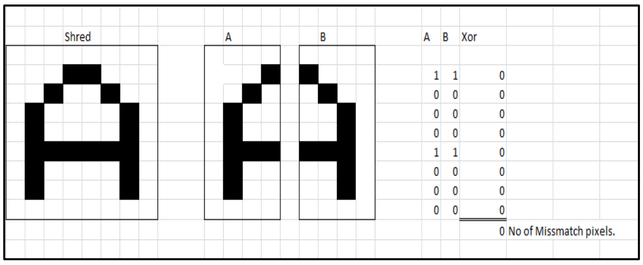
Here we introduce the idea of the similarity ***Sim*** (S_{ip}, S_{iq}). After the procedure of fitness evaluation or scanning the strip all adjacent strips are suppose to be the single strip. This will reduce the number of parent population for the further procedure of crossover.

This step is one of the most important amongst all the steps of our algorithm, which is to extract the data from each shred at the prescribed location so that the algorithm is capable to compute that data and come to a result. The correctness of data extraction is very important because the flaws in the input will definitely create problems in the output and so the desirable result could not be achieved.

To get the data from our input, the algorithm first need to convert the input into a form that allows compatibility to the processing, so for such reason, the shred we input in the form of image is to be converted into gray scale and then to a binary mapping of 0s and 1s for simple comparison to result in either ‘Match’ or ‘Mismatch’ (True or False condition).

For the algorithm, we need to extract data from two places on our shred, these two places enables the matching of the strips with each other; the right most column of the strips, which are previously converted into the binary code, is been stored in a variable and same is been done with the left most column of those strips. This extraction of the data is saved into 2 distinct variables, which will be dealt with in the next stage.

Colors comparison will be done by using the basic method of comparison the difference in color values of each pixel, those with White would gain a low value, and Black color a high one, giving a range of 0 and 1 [18] [19].



A	B	Xor
1	1	0
0	0	0
0	0	0
0	0	0
1	1	0
0	0	0
0	0	0
0	0	0
		0 No of Missmatch pixels.

Figure 1. The Pixel Position and Their Respective Values

5.4. Crossover as Processing

At this stage we can further improve our matching procedure and which is called the process of crossover in the genetic algorithm. For the procedure of crossover we take two parent chromosomes which are obtained from the previous stage. These parent chromosomes generate more fitted child chromosomes. Here we use the two point crossover that randomly selected any two individual from the previous step and exchange them according to the predefined criteria. This is the main and basic part of the whole algorithm, where result oriented decision are been taken. Our shreds are been patterned so as to get the best and final decision making. Lastly we converted the whole of a shred into a binary format, which makes it easier for comparing the values on the edges to come to a decision by using the two point crossover as shown in Figure 3.

The values in those two variables are here been checked upon and compared, this comparison provides the basis of whether the strips are in a series or not. It is more prominent and clearly been defined in the Figure 2, where we can easily sort out the serial of the shreds without much intervention. Where it makes prominent for the algorithm to seek the areas for an accurate match. [20] [21].

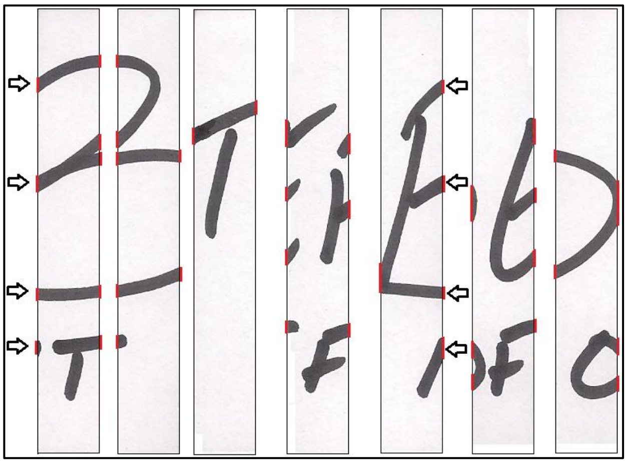
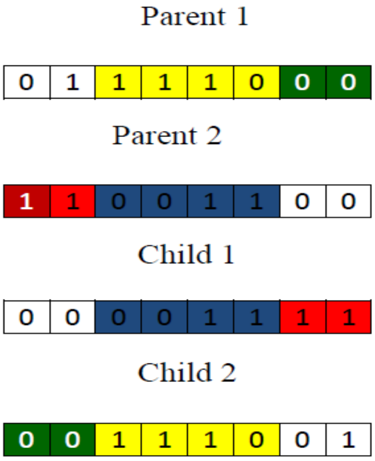


Figure 2. A Sample of Shreds from a Document Scanned and Marked For the Areas to Be Checked



Parent 1

0 1 1 1 1 0 0 0

Parent 2

1 1 0 0 1 1 0 0

Child 1

0 0 0 0 1 1 1 1

Child 2

0 0 1 1 1 0 0 1

Figure 3. Showing the Process of Crossover in Processing Step

From the mathematical prospective, we are performing the Xoring function with the values we get in our variables. This allows the same value to produce a zero whereas different values give a one (mark). By this it can easily be checked out that which of the strips are the most probable match.

In some other cases, we have coincidently same features in more than 1 strip so for that purpose, the result might be wrong as the program is just design to pick up the First strip with matching features. This is why this Algorithm is termed as Heuristic (Experimental algorithm with evident possibility of errors).

When this matching is done, the strip selected is then put to test, and its ending edge features are taken into the variable and so the comparison continues until all the strips are been matched to form the whole document in its correct combination.

5.5. Recombination by Mutation

This is the final stage where these strips are rearranged if required and then show on the screen as a picture for the user to view it, save it or either way. This is called the processes of mutation in GA where we obtained the global optima instead of local optima. In the process of mutation there is random selection of un match strip and make it swap with any other strip. The basic idea is to randomly swap two shred which is called (S2M) [17]. The process may repeat 100 times according to the predefine randomly chosen value. This is in accordance with the output requirement because we require an image form of output, so that the recombination is been seen and deductions made.

6. Tools Used

Due to Image processing efficiency, we have used MatLab® environment for compiling the whole code. This allowed us to use predefined functions and dedicated image processing commands that are unavailable or need to be generated in other coding languages. Despite the predefined commands, we have ensured the code to simplicity as far as possible to make it easy and understandable for the user. "When working with images in MatLab ®, there are many things to keep in mind such as loading an image, using the right format, saving the data as different data types, how to display an image, conversion between different image formats, and so on. [9]". Also Six standard common categories of strip or cross-cut shredders we have used here are based on different security level as shown in Table 1.

Table 1. Showing the Different Level of Shredders

	Strip cut		Cross cut	
Security level	**Shred size**	**No. Shred**	**Shred size**	**No. shreds**
Level 1	**10 mm**	**20**	**11×40mm**	**150**
Level 2	**8 mm**	**30**	**8×40 mm**	**216**
Level 3	**4 mm**	**100**	**4×30 mm**	**530**
Level 4	**N.A**	**N.A**	**2×15mm**	**1600**
Level 5	**N.A**	**N.A**	**0.6×10mm**	**6575**
Level 6	**N.A**	**N.A**	**0.6× 6mm**	**19725**

5. Conclusion

The proposed study is an attempt or reconstructs the destroyed documents by the experiments it can be concluded that the provided scheme has given some good results on several images, therefore we may conclude that system does not produce 100% correct results. Most likely, this might be due to introduction of noise while scanning the image, cropping of the desirable area, the extraction process can also is poor at times.

Despite this, most of the aspects of the system work well, which concludes that, when two color tones is used as input (Black & White), a system will produce good results.

Due to the manual input procedure, the system can be very slow to run in the extraction stages, particularly while dealing with the image of A4-size. This is reasonable though, as A4 images contains lrge pixels, and usually contain long strips consisting of many pixels to be considered. When comparing the performance of the system to reconstruction by hand however, it is not possible. “This is a problem, particularly when one of the aims of the project was to try and find a system comparable to reconstruction by hand”. [22] [23]. The proposed memetic methodology which is the combination of different genetic properties provides better results in most of the case.

6. Limitations

This algorithm has certain limitations, which can be thought as the predefined requirements that should be met for the algorithm to work in the intended way, most of them are defined:

- The input method for the algorithm is still manual, meaning that individual shreds are to be scanned and inputted for the algorithm to start its process.
- One of an important consideration that has to be kept in mind is that the first shred is to be recognized and distinguished in all the shreds. This is because the comparison has to start from a shred that needs to be told by the user prehand. If this shred is not defined then the first shred input is considered as the initial, so it is necessary to save the first shred in first position.
- This algorithm only deals with text in black and white, so it limits the use of this algorithm to two toned image files; including typed text, hand written text, maps or drawings only.
- Another restriction is that only the shreds of a single document can be reconstructed, that is if the shreds of other page are included in the database, then it produces many out of order arrangements.
- Color images or pictures even of shades of gray would produce a wrong combination.

References

[1] Jigsaw puzzle solver using shape and color, by Min Gyo Chung,MargaretM.Fleck, David A. Forsyth. Proceedings of igsp'1998: http://130.203.133.150/viewdoc/summary?doi=10.1.1.127.9475.

[2] What is Unshredding? ShredExFlorida's Premier-Paper Shredding-Company: http://www.shredexonline.com/unshredding.php.

[3] J.C. Russ, “The image processing handbook”, London: CRC Press, ISBN 0-8493-2516-1, **(1995)**.

[4] P. De Smet, “Reconstruction of ripped-up documents using fragment stack analysis procedures”, Forensic science international, vol. 176, no. 2, **(2008)**, pp. 124-136.

[5] E. Justino, L.S. Oliveira and C. Freitas, “Reconstructing shredded documents through feature matching”, Forensic Science International**,** vol. 160, no. 2-3, **(2006)**, pp. 140-147.

[6] M. Prandtstetter and G.R. Raidl, “Meta-heuristics for reconstructing cross cut shredded text documents”, In: Raidl, G.R., *et al.* (eds.) GECCO '09: Proceedings of the 11th annual conference on Genetic and evolutionary computation, M Press, **(2009)**, pp. 349-356.

[7] M. Prandtstetter, “Hybrid Optimization Methods for Warehouse Logistics and the Reconstruction of Destroyed Paper Documents”, Ph.D. thesis, Vienna University of Technology, **(2009)**.

[8] C. Schauer, “Reconstructing Cross-Cut Shredded Documents by means of Evolutionary Algorithms”, Master's thesis, Vienna University of Technology, Institute of Computer Graphics and Algorithms, **(2010)**.

[9] D. Goldberg, C. Malon and M. Bern, “A global approach to automatic solution of jigsaw puzzles”, Computational Geometry, vol. 8, no. 2-3, **(2004)**, pp. 165-174.

[10] M.G. Chung, M.M. Fleck and D.A. Forsyth, “Jigsaw puzzle solver using shape and color”, In: Fourth International Conference on Signal Processing Proceedings 1998, Signal Processing Proceedings, vol. 2, **(1998)**, pp. 877-880.

[11] “Unshredder”, Safe Guard Ltd. Available at: http://www.unshredder.com/, **(2010)**.

[12] "How to Reconstruct Shredded Documents", eHow Journal Available at: http://www.ehow.com/how_4768399_reconstruct-shredded-documents.htm1, **(2009)**.

[13] "Clevr Photo Stitching", Available at http://www.clevr.com/, **(2010)**.

[14] M. Brown and D. Lowe, "Automatic Panoramic Image Stitching using Invariant Features", University of British Columbia, Avaliable at http://cvlab.epfl.ch/~brown/papers/ijcv2007.pdf, (**2007**).

[15] M. Seul, L. O'Gorman, M.J. Sammon, "Practical algorithms for image analysis: description, examples, and code", Cambridge: Cambridge University Press. ISBN 0-5216-6065-3, **(2000)**.

[16] E. Justino, L.S. Oliveira and C. Frei, "Reconstructing shredded documents through feature matching", Forensic Science International, In Press Available from: http://www.sciencedirect.com/science/article/B6T6W-4HDG9H9-1/2/70d5747569ae5f9bd1840ba4f301c0d9, **(2005)**.

[17] P. Matthias and R. Günther, "A Memetic Algorithm for Reconstructing Cross- Cut Shredded Text Documents", DOI: 10.1007/978-3-642-16054-7_8 · Source: DBLP, https://www.researchgate.net/publication/22141117, **(2010)**.

[18] M. Nixon and A. Aguado, "Feature Extraction and Image Processing", Oxford: Newnes. ISBN 0-7506-5078-8, **(2002)**.

[19] J. Canny, "A Computational Approach to Edge Detection", IEEE Transactions on Pattern Analysis and Machine Intelligence,**(1986)**.

[20] "An approach to a pictorial representation of shreds while and after reconstruction", by Roel:http://roel.reijerse.net/unshredder/.

[21] "MatLab product and programming language", The MathWorks, Inc: http://www.mathworks.com/matlab.

[22] Digital Image Processing Using MATLAB: 2nd Ed. by Rafael C. Gonzalez, Richard E. Woods, andEddins.

[23] "An Investigation into Automated Shredded Document Reconstruction using Heuristic Search Algorithms Anna Skeoch", BSc (Hons) Mathematics and Computing: **(2006),** http://www.cs.bath.ac.uk/~mdv/courses/CM30082/projects.bho/2005-6/skeoch-al-dissertation-2005-6.pdf.

[24] H. Bunke and G. Kaufmann, "Jigsaw puzzle solving using approximate string matching and best-first search", in: Proceedings of the Fifth International Conference on Computer Analysis of Images and Patterns, **(1993)**, pp. 299–308.

[25] H. Bunke and U. Buehler, "Applications of approximate string matching to 2-D shape recognition", Pattern Recognit, vol. 26 **(1993)**, pp. 1797–1824.

[26] H. Wolfson, E. Schonberg, A. Kalvin and Y. Lamdan, "Solving jigsaw puzzles by computer vision", Ann. Operat. Res., vol.12, **(1988)**, pp. 51–64.

Authors

Fozia Hanif khan, she is working as an Associate Professor in the Department of Mathematics of Sir Syed University of Engineering and technology, Karachi, Pakistan. She has done her Ph. D from University of Karachi University in Operations Research. Her fields of interest are cryptography, graph theory, Optimization Network Security and Wireless sensors Networks.

Rehan Shams, he is a Ph. D scholar working as an Assistant Professor in the Department of Telecommunication Engineering of Sir Syed University of Engineering and technology, Karachi, Pakistan, he had done his MS from the University of Plymouth UK. His fields of interest are cryptography, Network Security and Wireless sensors Networks.

Dur-e-Shawar Agha, she is working as a Junior Lecturer in the Department of Computer Engineering of Sir Syed University of Engineering and Technology, Karachi, Pakistan. She did her BS in Computer Engineering and MS in Computer Engineering (Specialization in Computer Networks) from Sir Syed University of Engineering and Technology, Karachi, Pakistan. Her research interests are Cryptography, Artificial Intelligence and Wireless Sensor Network.

Rabia N. Enam, she received her PhD and Masters in Computer Engineering from Sir Syed University of Engineering and Technology (SSUET) Pakistan. She did Bachelors in Computer Engineering from N.E.D. University, Pakistan. Rabia also did Bachelors and Masters in Applied Mathematics from Karachi University. She is an Associate Professor at the Department of Computer Engineering at SSUET. Her research interests include the conceptual frameworks and algorithms used in communication