International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Recognition of Roman Characters using Geometric and Regional Features

Deval Verma*, Himanshu Agarwal, A.K. Aggarwal

Abstract: ***This paper presents the Roman Characters Recognition in clean and noisy environment using geometrical and regional features. The geometrical and regional features are extracted from the standard dataset and are combined to achieve better recognition. The combination of these features is classified using neural network (NN) and random forest (RF) classifiers. In our experiments, we have achieved the recognition accuracy of 100% for some characters. However, the average recognition accuracy of 85.7% has been recorded by using NN and 88% has been recorded by using RF classifier, respectively.***

Keywords: OCR, Geometrical features, Regional features, Neural Network, Random Forest, Recognition accuracy.

I. INTRODUCTION

OCR is an electronic conversion of scanned images. It is a method of recognizing characters from the scanned image of a document by the computer. The recognition of characters is possible by extracting discriminative features from the characters [1, 2, 21]. The quality of documents does not remain same over the years. Character recognition is the most challenging task in various fields of image processing and pattern recognition problems. It has many applications in the area of document analysis recognition [25], license plate recognition [8, 10], logo and seal recognition [22] and multilingual characters recognition [17, 18].

Fig. 1. Some images of roman characters

Manuscript published on 30 September 2019
* Correspondence Author
Deval Verma*, Mathematics Department, Jaypee Institute of Information Technology, Noida, India. Email: deval09msc@gmail.com
Himanshu Agarwal, Mathematics Department, Jaypee Institute of Information Technology, Noida, India. Email: himanshu.agarwal@jiit.ac.in
A.K. Aggarwal, Mathematics Department, Jaypee Institute of Information Technology, Noida, India. Email: amrish.aggarwal@jiit.ac.in

A brief description of roman characters in English language is given in Fig 1. It contains 26 upper case and 26 lower case letters and out of them five letters in each case are vowels and rest are consonants. This paper presents a technique for recognition of offline roman characters using geometrical and regional features. Recognition of these roman characters is done using neural network and random forest classifier. Standard dataset has been taken for experimental evaluation. The accuracy of 100% for some characters and an average accuracy of 88% for all characters have been achieved with clean dataset. In noisy environment, the accuracy of 75% for some characters and overall average accuracy of 38% have been recorded.

The organization of the rest of the paper is divided into following sections. Section 2 describes the related work. Section 3 deals with a description of proposed methodology. Feature extraction techniques are discussed in section 4. Recognition methods are discussed in Section 5. Results and experimental work are discussed in Section 6 and conclusions are provided in section 7.

II. RELATED WORK

This section reviews the literature related to alphabetic character recognition in various fields. Uttal (1969) [1] used conflicted dot patterns to recognize characters by dynamic visual noise (DVN) technique. Increasing the number of dots in the form of geometric structures leads to increase in recognizability. Kahan et al. (1987) [2] presented an algorithm in which they recognized roman alphabetic characters of different font sizes. Comelli et al. (1995) [3] found some degraded pictures of vehicles, which were captured by TV camera. The captured pictures were having some imperfection like geometric distortion, presence of noise and blurring. So, it was very difficult to recognize the license plate number properly. The authors have considered template matching and cross correlation technique to overcome this problem. Heutte et al. (1998) [4] presented a combined structural and statistical features based vectors (SSFBV) for character recognition. Jain et al. (1998) [5] recorded an accuracy of 85% after using hybrid features. Perwej et al. (2011) [16] proposed a neural network based English character recognition technique and achieved an accuracy of 82.5%. Kaur et al. (2011) [17] presented a hybrid recognition technique for uppercase letters in English language. They achieved an accuracy of 91.1%. Soore et al. (2017) recorded an accuracy of 98.8% after using a geometrical shape features on English alphabets.

Table- I: Work Related to Character recognition

Algorithms	Main method	Contents	Accuracy
Uttal et al. (1969) [1]	Dynamic visual noise	Capital alphabets	61.6 %
Kahan et al. (1987) [2]	Bayesian classifier, line adjacency graph based thinning	Small and capital alphabets	97%
Comelli et al. (1995) [3]	RITA software, Template matching, Cross correlation	Alphabetical capital characters	97.1%
Heutte et al. (1998) [4]	Statistical classifier, Structural and statistical feature based vector	Capital and small alphabets using different databases	90.8%, 83.4% , 52.8%
Jain et al.(1998) [5]	Mixed features	8500 English characters	85%
Iqbal et al. (2004) [11]	Distance calculated by geometrical features	4680 English characters	98.6%
Lee et al. (2004) [10]	Contour based features and area of peripheral background	1248 characters of English and numerals	95.7%
Sohn et al. (2008) [12]	Nono gram features of row and columns	4084 English characters	100%
Blumenstein et al. (2009) [9]	Hybrid features	English (CAS database, BAC database), CEDAR database characters	75.3%, 84.57%, 80%
Campos et al. (2009) [14]	English hybrid characters	English characters and Kannada characters	55.26%
Rani et al. (2011) [15]	Left diagonal line using Crossing corners	650 English characters	84.52%
Prasad.et al. (2012) [12]	Mixed features	6600 Hindi and English characters	95.41%
Soora et al. (2014) [19]	Triangle area and crossing corner and perpendicular distance features	6584 characters of English and numerals	99.031%
Roy et al. (2014) [18]	Histogram of local gradients	500 English words	80.1%
Soora et al. (2016) [20]	Angular width, crossing corners and geometrical shapes	6584 characters of English and numerals	98.81%

III. PROPOSED METHODOLOGY

This section presents the description of proposed technique that is used for roman character recognition as depicted in Fig 3. First step is the preprocessing of standard dataset. Dataset consist of total 910 uppercase roman characters of 26 classes. Second step is to extract geometrical and regional features from the binarized dataset. Third step is division of dataset feature matrix into training features submatrix and testing features submatrix. Then corresponding labels are given to training dataset to obtain trained model. After classification, it predicts the labels of corresponding training or testing dataset.

IV. EXTRACTION OF FEATURES

This section describes the method of features extraction, that are extracted from the input binarized dataset of roman characters. This work only focuses on geometrical and regional features.

A. *Geometrical Features*

Zone based features are the part of geometrical features. In this work, we extract the features from zones in two ways. Firstly, the input image is divided into 3×3 zones and secondly, the input image is divided into 1×3 and 3×1 and then extract the features from each zone.

1. Total nine element feature vectors are calculated from each part that are as follows.
- Number of horizontal lines in each zone.
- Complete length of horizontal lines in each zone.
- Number of right diagonal lines in each zone.
- Complete length of right diagonal lines in each zone.
- Number of vertical lines in each zone.
- Complete length of vertical lines in each zone.
- Number of left diagonal lines in each zone.
- Complete length of left diagonal lines in each zone.
- Number of intersecting points

2. Euler number: It is the difference between number of objects in the region and number of holes in that image.

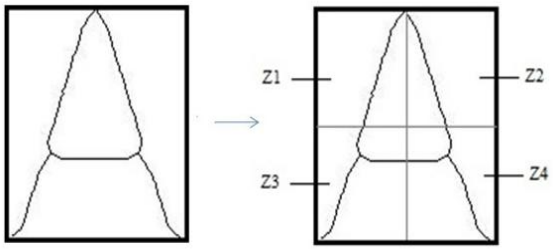

Fig 2. Zone based features

B. *Regional feature*

We have considered three regional features namely eccentricity, extent and orientation. In this study a total of 85 features ($9\times9=81$ features are calculated from 3×3 zone, 1 Euler feature and 3 regional features.) are extracted from feature1 submatrix from first part and total 58 features ($9\times3=27,\ 3\times9=27$ features are calculated from 1×3 and 3×1 zone, 1 Euler feature and 3 regional features) are extracted from feature 2 submatrix from second part. Total 143 features are calculated for each uppercase letter of the dataset.

V. CLASSIFIERS FOR CHARACTER RECOGNITION

In this work, we have used neural network (NN) [16] and random forest (RF) [8] classifiers to determine the recognition accuracy of roman characters.

A. Recognition using ANN Classifier

The fundamental unit of network is called node or neuron. ANN Classifier consists of input layer (first layer), hidden layers (second layer, third layer and so on) and output layer (last layer). In this work 45 numbers of hidden nodes have been used to determine the recognition accuracy of roman characters.

B. Recognition using Random Forest Classifier

We have tested the performance of roman characters by using another classifier i.e. RF. In this classifier, total 300 decision trees are generated. At 50 and 64 number of decision trees, we depicted the best results.

VI. RESULTS AND DISCUSSION

Table II shows the recognition accuracy of RF classifier for all possible ratios feature set. The peak and minimum average recognition accuracy of 88.46% and 69.46% has been recorded for the testing dataset of 90:10 and 10:90 ratio respectively using 50 number of decision trees.

In Fig 4, accuracies are computed by applying impulsive noise of different strength using RF. It shows the accuracy decreases with increasing noise level. Table III shows the recognition accuracy of NN for all possible ratios feature set. The peak and minimum average recognition accuracy of 86.57% and 45.20% has been recorded for the testing dataset of 90:10 and 10:90 ratio respectively using 45 numbers of nodes of hidden layer in neural network. In Fig 5, accuracies are computed by applying impulsive noise of different strength using NN. It shows the accuracy decreases with increasing noise level. All experiments have been performed using MATLAB version 2016.

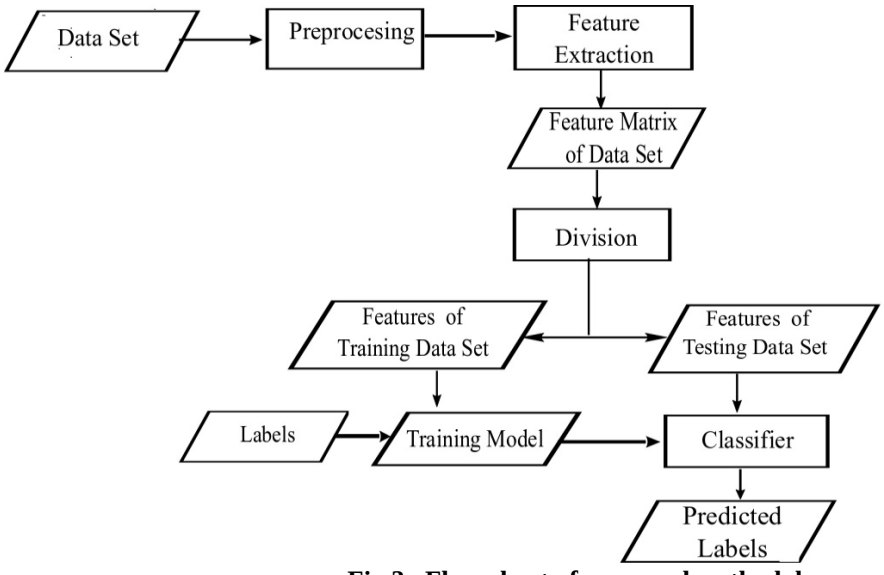

Fig 3. Flow chart of proposed methodology

Table- II: Recognition Accuracy of Characters using Random Forest (RF)

90_10	80_20	70_30	50_50	60_40	40_60	30_70	20_80	10_90
100	100	90.90909091	88.23529	100	90.47619	79.16667	66.66667	83.87097
100	75	90.90909091	82.35294	85.71429	85.71429	79.16667	77.77778	74.19355
25	62.5	72.72727273	82.35294	78.57143	76.19048	83.33333	77.77778	83.87097
50	75	81.81818182	88.23529	78.57143	66.66667	79.16667	70.37037	32.25806
100	100	90.90909091	88.23529	92.85714	90.47619	83.33333	77.77778	83.87097
100	87.5	90.90909091	94.11765	92.85714	85.71429	75	88.88889	83.87097
100	87.5	90.90909091	88.23529	100	76.19048	83.33333	81.48148	83.87097
75	87.5	90.90909091	88.23529	92.85714	76.19048	58.33333	62.96296	64.51613
50	62.5	72.72727273	82.35294	71.42857	90.47619	62.5	48.14815	25.80645
75	75	72.72727273	82.35294	78.57143	80.95238	58.33333	59.25926	64.51613
100	100	90.90909091	94.11765	92.85714	85.71429	83.33333	77.77778	67.74194
100	100	100	100	100	100	100	96.2963	96.77419
100	100	100	100	100	100	95.83333	92.59259	80.64516
100	75	81.81818182	94.11765	85.71429	90.47619	83.33333	88.88889	74.19355
75	87.5	90.90909091	70.58824	92.85714	80.95238	75	85.18519	74.19355
75	62.5	63.63636364	70.58824	71.42857	71.42857	70.83333	74.07407	64.51613
100	100	72.72727273	64.70588	71.42857	57.14286	37.5	44.44444	41.93548
100	100	100	82.35294	100	90.47619	87.5	81.48148	64.51613
100	62.5	81.81818182	82.35294	85.71429	85.71429	87.5	81.48148	61.29032
100	100	100	94.11765	100	95.2381	95.83333	70.37037	67.74194
100	62.5	72.72727273	76.47059	71.42857	80.95238	83.33333	85.18519	77.41935
100	100	81.81818182	100	100	100	95.83333	96.2963	90.32258
100	87.5	90.90909091	100	92.85714	90.47619	87.5	88.88889	90.32258
50	75	81.81818182	76.47059	85.71429	52.38095	66.66667	70.37037	64.51613
100	75	81.81818182	82.35294	85.71429	85.71429	79.16667	85.18519	74.19355
100	87.5	90.90909091	94.11765	92.85714	85.7	54.16667	37.03704	45.16129
Avg. = 87.5	**Avg. = 83.1346**	**Avg. = 85.667**	**Avg. = 86.42**	**Avg. = 88.46**	**Avg. = 83.5**	**Avg. = 77.84**	**Avg. = 75.64**	**Avg. = 69.86**

Where, 90_10 represents the ratio of training dataset and testing dataset i.e. 90 % data is used for training and 10 % for testing. Similarly, the rest of the partition of dataset up to 10_90 is computed in Table II using RF classifier.

Table- III: Recognition Accuracy of Characters using Neural Network (NN)

90_10	80_20	70_30	60_40	50_50	40_60	30_70	20_80	10_90
100	87.5	90.90909	78.57143	80.95238	66.66667	50	51.85185	48.3871
100	87.5	90.90909	85.71429	80.95238	66.66667	50	37.03704	35.48387
25	62.5	63.63636	71.42857	71.42857	71.42857	83.33333	77.77778	61.29032
50	62.5	63.63636	71.42857	42.85714	47.61905	58.33333	40.74074	54.83871
100	100	90.90909	85.71429	95.2381	95.2381	79.16667	85.18519	48.3871
100	75	81.81818	64.28571	57.14286	80.95238	91.66667	81.48148	0
100	87.5	81.81818	71.42857	80.95238	80.95238	79.16667	55.55556	9.677419
75	87.5	90.90909	85.71429	52.38095	71.42857	50	51.85185	58.06452
50	62.5	63.63636	71.42857	47.61905	38.09524	58.33333	25.92593	25.80645
100	100	63.63636	78.57143	66.66667	76.19048	62.5	59.25926	61.29032
100	100	90.90909	78.57143	80.95238	57.14286	95.83333	81.48148	35.48387
100	100	100	92.85714	100	95.2381	100	96.2963	80.64516

100	50	63.63636	78.57143	85.71429	90.47619	79.16667	88.88889	9.677419
75	75	72.72727	64.28571	57.14286	57.14286	70.83333	70.37037	41.93548
75	62.5	72.72727	42.85714	71.42857	47.61905	66.66667	55.55556	3.225806
75	87.5	63.63636	64.28571	76.19048	90.47619	41.66667	66.66667	74.19355
100	100	90.90909	35.71429	38.09524	61.90476	37.5	33.33333	35.48387
100	75	90.90909	71.42857	71.42857	76.19048	37.5	37.03704	12.90323
100	62.5	63.63636	78.57143	85.71429	61.90476	75	55.55556	74.19355
100	87.5	81.81818	92.85714	95.2381	76.19048	83.33333	70.37037	87.09677
100	62.5	72.72727	64.28571	76.19048	61.90476	50	55.55556	54.83871
100	75	90.90909	85.71429	90.47619	85.71429	79.16667	88.88889	70.96774
75	75	81.81818	64.28571	76.19048	80.95238	62.5	66.66667	80.64516
75	75	72.72727	71.42857	52.38095	57.14286	45.83333	40.74074	41.93548
75	62.5	54.54545	50	57.14286	71.42857	62.5	55.55556	38.70968
75	75	63.63636	85.71429	80.95238	52.38095	62.5	40.74074	48.3871
Average =85.57	**Average =78.36**	**Average =77.17**	**Average =74.52**	**Average. =71.2**	**Average =69.96**	**Average. =65.86**	**Average. =60.39**	**Average =45.20**

Where, 90_10 represents the ratio of training dataset and testing dataset i.e. 90 % data is used for training and 10 % for testing. Similarly, the rest of the partition of dataset up to 10_90 is computed in Table III using NN classifier.

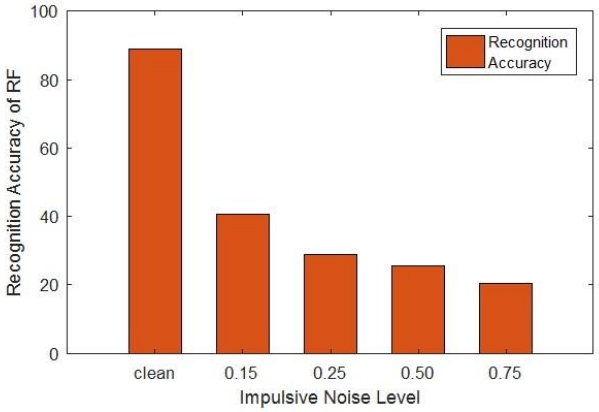

Fig:4 Recognition accuracy of RF on clean and noisy dataset at different levels

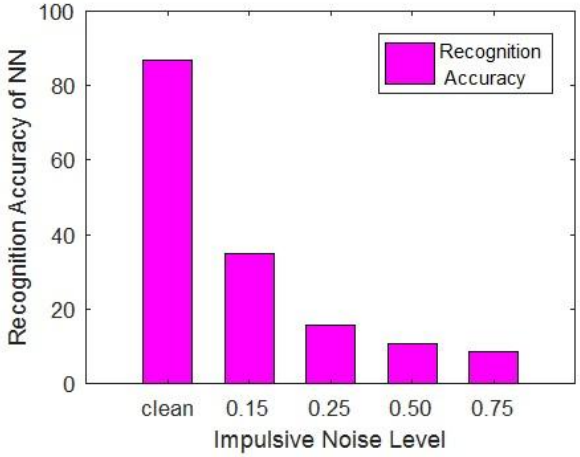

Fig:5 Recognition accuracy of NN on clean and noisy dataset at different levels

VII. CONCLUSION

In this paper, two different classifiers viz. RF and NN are used for classification of offline alphabetic uppercase roman characters with and without impulsive noise. The geometric and regional features have been extracted from the dataset taken into consideration. The experimental results show that in noiseless environment, RF and NN classifier provides highest accuracy of 88.46% and 86.5% respectively. On introducing noise of different strength, it has been observed that NN and RF do not provide significant recognition accuracy.

REFERENCES

1. Uttal, W.R.: Masking of alphabetic character recognition by dynamic visual noise (dvn). *Perception & Psychophysics*, (1969) 6(2), 121–128.
2. S. Kahan, T. Pavlidis, H.S. Baird, On the recognition of printed characters of any font and size. *IEEE Transactions on pattern analysis and machine intelligence* (1987) (2), 274–288.
3. P. Comelli, , P. Ferragina, M.N. Granieri, F. Stabile, Optical recognition of motor vehicle license plates. *IEEE transactions on Vehicular Technology*, (1995) 44(4), 790–799.
4. L. Heutte, T. Paquet, , J.V. Moreau, Y. Lecourtier, C. Olivier, A structural/statistical feature based vector for handwritten character recognition. *Pattern recognition letters*, (1998) 19(7), 629–641.
5. S. T. Jain and H. B. Dave "Handwritten text recognition using geometric features," *IETE J. Res.* . (1998). 44 (6), pp. 299–303.
6. K.K. Kim, K. Kim, J. Kim, H.J. Kim, Learning-based approach for license plate recognition. *In: Neural Networks for Signal Processing X, 2000. Proceedings of the 2000 IEEE Signal Processing Society Workshop*, vol. 2, pp. 614–623. IEEE (2000).
7. Z.C. Li, C.Y Suen, The partition–combination method for recognition of handwritten characters. *Pattern Recognition Letters*, (2000) 21(8), 701–720.
8. L. Beirman, Random forests. *Machine learning* 45(1), 5–32 (2001).
9. M. Blumenstein, B. Verma and H. Basli. "A novel feature extraction technique for the recognition of segmented handwritten characters*," in Proc. Seventh Int. Conf. Doc. Anal. Recognit., Edinburgh*, 1, 2003, pp. 137–141.
10. H. J. Lee, S. Y. Chen and S. Z. Wang. "Extraction and recognition of license plates of motorcycles and vehicles on highways," *in Proc. 17th Int. Conf. Pattern Recognit., Cambridge*, 4, 2004, pp. 356–359.
11. A. Iqbal, A. B. M. Musa, A. Tahsin, M. A. Sattar, M. M. Islam and K. Nurase. "A novel algorithm for translation, rotation and scale invariant character recognition," *in Proc. SCIS & ISIS, Nagoya*, 2008, pp. 1367–1374.
12. Y. S. Sohn and B. S. Kim, A recognition of the printed alphabet by using nonogram puzzle, *J. Korea Inst. Intell. Syst.* . (2008) 18 (4), pp. 451–455.
13. U. Bhattacharya and B. B. Chaudhuri, Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals, *IEEE Trans. Pattern Anal. Mach. Intell*. (2009) 31 (3), pp. 444–457.

14. T. E. de Campos, B. R. Babu and M. Varma. Character recognition in natural images, *in Proc. Int. Conf. Comput. Vis. Theory Appl., Lisbon*, 2009, pp. 273–280.
15. M. Rani and Y. K. Meena, An efficient feature extraction method for handwritten character recognition, *in Swarm Evolutionary Memetic Comput. Conf., Visakhapatnam*, (2011)7077, pp. 302–309.
16. Y. Perwej and A. Chaturvedi, Neural networks for handwritten English alphabet recognition, *International Journal of Computer Applications*, April 2011 (0975 – 8887) Volume 20– No.7, pp. 1-5.
17. R. Kaur and B. Singh, A hybrid neural approach for character recognition system, *International Journal of Computer Science and Information Technologies*, 2011, Vol. 2 (2) , pp. 721-726.
18. P. P. Roy, S. Roy and U. Pal. "Multi-oriented text recognition in graphical documents using HMM, *in 11th IAPR Int. Workshop Doc. Anal. Syst., Tours*, (2014) pp. 136–140.
19. N. R. Soora and P. S. Deshpande, Robust feature extraction technique for license plate characters recognition, *IETE J. Res*. (2014) 61 (1), pp. 72–79.
20. N. R. Soora and P. S. Deshpande. Novel geometrical shape feature extraction techniques for multilingual character recognition, *IETE Tech. Rev.*, (2016). pp. 1–10. doi:10.1080/02564602.2016.1229583.
21. N. R. Soora and P. S. Deshpande Review of Feature Extraction Techniques for Character Recognition, *IETE Journal of Research*, (2017) DOI: 10.1080/03772063.2017.1351323
22. A. Alaei, P. P. Roy, & U. Pal, Logo and seal based administrative document image retrieval: a survey. *Computer Science Review*, 2016, 22, 47-63.
23. R. Sarkhel , N. Das , A. Das , M. Kundu , M. Nasipuri , A Multi-scale Deep Quad Tree Based Feature Extraction Method for the Recognition of Isolated Handwritten Characters of popular Indic Scripts, *Pattern Recognition*(2017), doi:10.1016/j.patcog.2017.05.022.
24. C. Arora, N. Arora, H .Goyal and V. Gaurav , *Sixth sense dictionary. Proceedings of the IEEE International Conference on Computing for Sustainable Global Development"*, (2017) 5860–5864.
25. A. K Bhunia, G.Kumar, P. P. Roy, R. Balasubramanian, & U. Pal, Text recognition in scene image and video frame using Color Channel selection. *Multimedia Tools and Applications*, 2018, 77(7), 8551-8578.

AUTHORS PROFILE

Deval Verma is pursuing PHD in Department of Mathematics at Jaypee Institute of Information Technology, Noida sector 62, India. Her area of research includes Digital Watermarking, Palmprint Matching and Optical Character Recognition (OCR). She has published several research papers in conferences and journals of international repute.

Dr. Himanshu Agarwal is an Assistant Professor in the Department of Mathematics, Jaypee Institute of Information Technology, Noida, India. He has received Ph.D. degree from the Department of Mathematics, Indian Institute of Technology Roorkee in 2015. He has published several research papers in reputed international journals and conferences. His area of research includes image processing, computer vision, data visualization, information security and statistical learning.

Dr. Amrish Kumar Aggarwal is working as Professor in the Department of Mathematics, Jaypee Institute of Information Technology, Noida since 2007. Prof. Aggarwal has 27 years of teaching and research experience. He has published 43 papers in international journals and conference proceedings of high repute. His research interests are image processing, optimization, and continuum mechanics.