Offline Handwritten Pre-Segmented Character Recognition of Gurmukhi Script

Munish Kumar[1], Manish K. Jindal[2], Rajendra K. Sharma[3] and Simpel R. Jindal[4]

[1] *Maharaja Ranjit Singh Punjab Technical University, Department of Computer Applications, GZS Campus College of Engineering & Technology, Bathinda, Punjab, India* munishcse@gmail.com

[2] *Panjab University Regional Centre, Department of Computer Science & Applications, Muktsar, Punjab, India* manishphd@rediffmail.com

[3] *Thapar University, Department of Computer Science & Engineering, Patiala, Punjab, India* rksharma@thapar.edu

[4] *Yadavindra College of Engineering, Talwandi Sabo, Computer Science & Engineering Section, Bathinda, Punjab, India* simpel_jindal@rediffmail.com

Abstract. In this paper, we have proposed a feature extraction technique for recognition of segmented handwritten characters of Gurmukhi script. The experiments have been performed with 7000 specimens of segmented offline handwritten Gurmukhi characters collected from 200 different writers. We have considered the set of 35 basic characters of the Gurmukhi script and have proposed the feature extraction technique based on boundary extents of the character image. PCA based feature selection technique has also been implemented in this work to reduce the dimension of data. We have used k-NN, SVM and MLP classifiers. SVM has been used with four different kernels. In this work, we have achieved maximum recognition accuracy of 93.8% for the 35-class problem when SVM with RBF kernel and 5-fold cross validation technique were employed.

Key words: feature extraction, classification, PCA, k-NN, SVM, MLP

1. Introduction

Optical Character Recognition (OCR) is the process which helps to convert the handwritten or printed text into a format that is processable by machine. Handwritten character recognition is more complicated due to the variation in styles of writing. When a text is handwritten and scanned, a large amount of noise will occur while recognizing such handwritten characters. We have divided handwritten character recognition into two categories, *viz.* online and offline. This paper focuses on offline handwriting recognition. Therefore, in this paper we have exhibited a recognition technique for offline handwritten character recognition of Gurmukhi script in light of the fact that a little work has been done on Gurmukhi script to date, to the best of our knowledge, in spite of the fact that a considerable measure of work has been done on various scripts, for example, Bangla, Devanagari and so forth.

2. Related work

Bhattacharya *et al.* [1] have presented Bangla character recognition system and they have obtained maximum recognition accuracy of 94.7%.

Bunke and Varga [2] have reviewed the state of the art in offline Roman cursive handwriting recognition. They identified the challenges in Roman cursive handwriting recognition.

Kunte and Samuel [8] have presented an efficient framework for printed Kannada text recognition. They considered invariant moments and Zernike moments as features and Neural Network (NN) as classifier. They achieved a recognition accuracy of 96.8% using 2 500 characters.

Grosicki and Abed [3] proposed a French handwriting recognition system in a competition held in ICDAR 2011. In this competition, they have presented comparisons between different classification and recognition systems for French handwriting recognition. Kacem *et al.* [4] have used structural components for recognition of Arabic names.

Kumar *et al.* [7] have presented efficient feature extraction techniques for offline handwritten Gurmukhi character recognition. They have also presented a hierarchical technique for offline handwritten Gurmukhi character recognition [5]. Using this technique, they accomplished a recognition accuracy of 91.8%. Kumar *et al.* [6] have presented a study on various transformation techniques for offline handwritten Gurmukhi character recognition.

Lorigo and Govindaraju [9] have introduced a critical review on offline Arabic handwriting recognition frameworks. They have presented various techniques employed at different stages of the offline handwritten Arabic character recognition system.

Pal *et al.* [10] dealt with recognition of offline handwritten Bangla compound characters using modified quadratic discriminant function. They have acquired 85.9% recognition precision by using five-fold cross validation technique. Pal *et al.* [11] have assimilated a comparative study of handwritten Devanagari character recognition. Sharma *et al.* [12] considered a quadratic classifier based scheme for the recognition of off-line Devanagari handwritten characters. In this system, they have used chain code directional features and achieved a recognition accuracy of 80.4%.

Tran *et al.* [13] have considered the problem of French handwriting recognition using 24 800 samples. They have worked on both online and offline handwritten character recognition. Wang *et al.* [14] have presented a technique for recognition of Roman alphabets and numeric characters. They had a recognition rate of about 86.0%.

Zhu *et al.* [16] have described a robust model for online handwritten Japanese text recognition. They obtained a recognition accuracy of 92.8% using 35 686 samples.

The work carried out in this paper focuses on the recognition of segmented offline handwritten character recognition for Gurmukhi script of a 35-class problem.

3. Gurmukhi script

The name Gurmukhi signifies "from the mouth of the Guru" and originates from the old Punjabi word *Guramukhi*. Gurmukhi script is utilized for composing the Punjabi dialect. A portion of the properties of the Gurmukhi script are as follows.

- In Gurmukhi script, there are 35 character constants out of which the first three are vowel bearers.
- There are six consonants which are created by placing a dot (*bindi*) at the foot (*pair*) of the consonant which is called Gurmukhi constants with subscript dots.
- Auxiliary Gurmukhi symbols denote double consonants, or conjunct adjacent constants.
- Double Consonants: These symbols are also called *adhak*.
- In Gurmukhi script, various punctuation symbols are used. These symbols signify the partition of heading from the text, or line break. Various punctuation symbols are *visarg*, *dandi*, and *dodandi*.
  - *Visarg*: it is symbolized as two circles, one circle is above the other circle, just as a colon used in English language. This symbol points towards the division of heading from the text.
  - *Dandi*: it is a single vertical line which indicates the completion of a sentence.
  - *Dodandi*: signifies two parallel lines which indicate a break in a line.

4. The proposed recognition system

Proposed recognition framework comprises of different stages: data collection, digitization, preprocessing, feature extraction, and classification. In our proposed framework, we have also used Principal Component Analysis (PCA) to reduce the length of the feature vector. The general plan of the proposed framework is depicted in Fig. 1.

4.1. Data collection

In this work, we have collected 7 000 specimens of segmented Gurmukhi characters from 200 different writers. They were requested to write each of the fundamental 35 characters of the Gurmukhi script. We have collected this type of dataset from different government offices, schools, colleges etc.

4.2. Digitization and pre-processing

Digitization is the way toward changing over the handwritten document into electronic shape. This procedure is accomplished by using HP-1400 model of scanner. Digitization stage creates the digital image which is sustained to the pre-processing stage. Pre-processing stage is the underlying phase of character recognition framework. In this,

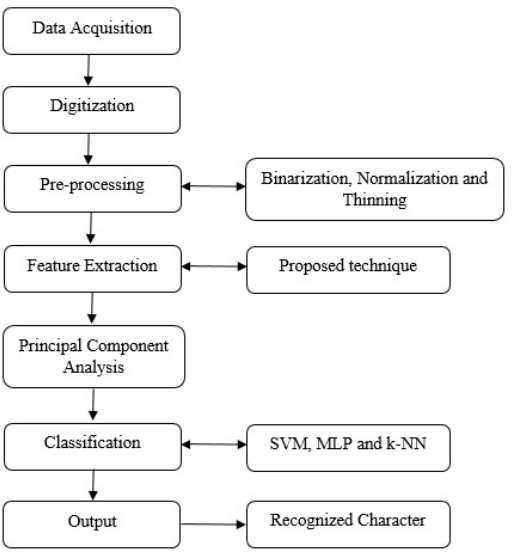

Fig. 1. Block diagram of the offline handwritten Gurmukhi character recognition system.

a **b**

Fig. 2. Image of a Gurmukhi character: (**a**) digitized; (**b**) thinned.

stage we change over the image into a gray scale format and bitmap image. From that point onward, bitmap image is changed over into the thinned image as shown in Fig. 2 by using parallel thinning algorithm proposed by Zhang and Suen [15].

4.3. Proposed feature extraction technique

In our proposed technique, a character image is horizontally divided into n horizontal regions and then the width of the character in this region, that is, the distance between the leftmost and rightmost element of the boundary of the character is taken as a feature in each region, thus forming n features. Similarly, the height of the character, that is, the vertical distance between the uppermost and lowermost element of the boundary of the character in n vertical regions is found, thus forming further n features. For the

regions that do not have a foreground pixel, the feature value is taken as zero. By using this process, $2n$ features are extracted for an image of one character. In our case it was $n = 100$ so there were $2n = 200$ features in total.

In the algorithm, the background pixel (OFF pixel) is considered as 0 and foreground pixel (ON pixel), that is the pixel belonging to the character, is considered as 1. The steps made to extract the features for horizontal regions are as follows.

I : given image $(x \times y)$
$V \leftarrow$ Allocate(n): vector of n features for horizontal regions
for $i := 1$ **to** n **step** 1 **do**
  for $j := 1$ **to** x **step** 1 **do**
    for $k := 1$ **to** y **step** 1 **do**
      if $I_{jk} = 1$ **then**
        $a := j$
        break loops for j, k
      end if
    end for
  end for
  for $j := x$ **to** 1 **step** -1 **do**
    for $k := 1$ **to** y **step** 1 **do**
      if $I_{jk} = 1$ **then**
        $b := k$
        break loops for j, k
      end if
    end for
  end for
  $V_i := b - a$
end for
Return V

4.4. Principal Component Analysis

Principal Component Analysis (PCA) is a methodology which changes various correlated features into a number of uncorrelated features called principal components. PCA has been broadly used in the field of pattern recognition. PCA is a technique which is used to reduce the dimension of data and to extract the meaningful feature subset. We covered 95% variance in this work for PCA.

4.5. Classification

Classification phase is the last phase of character recognition framework; this phase uses the feature extracted in the previous phase. The preliminary aim of the classification

Tab. 1. Recognition accuracy of the proposed feature extraction technique with different classifiers. Recognition accuracy was calculated as an average for five folds.

Feature Extraction Technique	**Classifier**	**False Positive Rate (FPR) [%]**	**False Rejection Rate (FRR) [%]**	**Recognition Accuracy [%]**
without PCA	k-NN ($k = 5$)	0.80	9.10	90.10
	MLP	1.20	11.60	87.20
	Linear SVM	1.70	8.10	90.20
	Polynomial SVM	1.60	19.60	78.80
	RBF SVM	0.60	8.80	90.60
	Sigmoid SVM	1.20	16.40	82.40
with PCA	k-NN ($k = 5$)	1.10	6.60	92.30
	MLP	0.89	9.71	89.40
	Linear SVM	0.84	7.46	91.70
	Polynomial SVM	1.40	16.20	82.40
	RBF SVM	0.60	5.60	**93.80**
	Sigmoid SVM	1.20	11.70	87.10

phase of an OCR framework is to build up the constraint for decreasing the misclassification relevant to feature extraction. In order to reduce the complexity of classifier, we have used PCA to reduce the dimension of the feature vector, as written above. In present work, we have used three different classifiers, namely, k-NN ($k = 5$), SVM and MLP. SVM classifier has also been considered with four different flavours, namely, Linear-SVM, Polynomial-SVM, RBF-SVM and Sigmoid-SVM.

5. Experimental results and discussion

In this section, we have presented the results of experiments by using proposed feature extraction technique with different classifiers. We have used 5-fold cross validation technique for obtaining the recognition accuracy. For the purpose of experiments, we have considered 7 000 samples of offline handwritten Gurmukhi characters of the 35-class problem. Various types of recognition accuracies obtained forthwith have been depicted in Tab. 1. False Positive Rate (FPR), Rejection Rate (RR) and recognition accuracy achieved with various classifiers, without PCA and with PCA, are graphically depicted in Figs. 3a, b and c, respectively.

We have accomplished the best recognition accuracy of 93.8% using the proposed feature extraction and RBF-SVM classifier. Confusion matrix for this case has been depicted in Tab. 2.

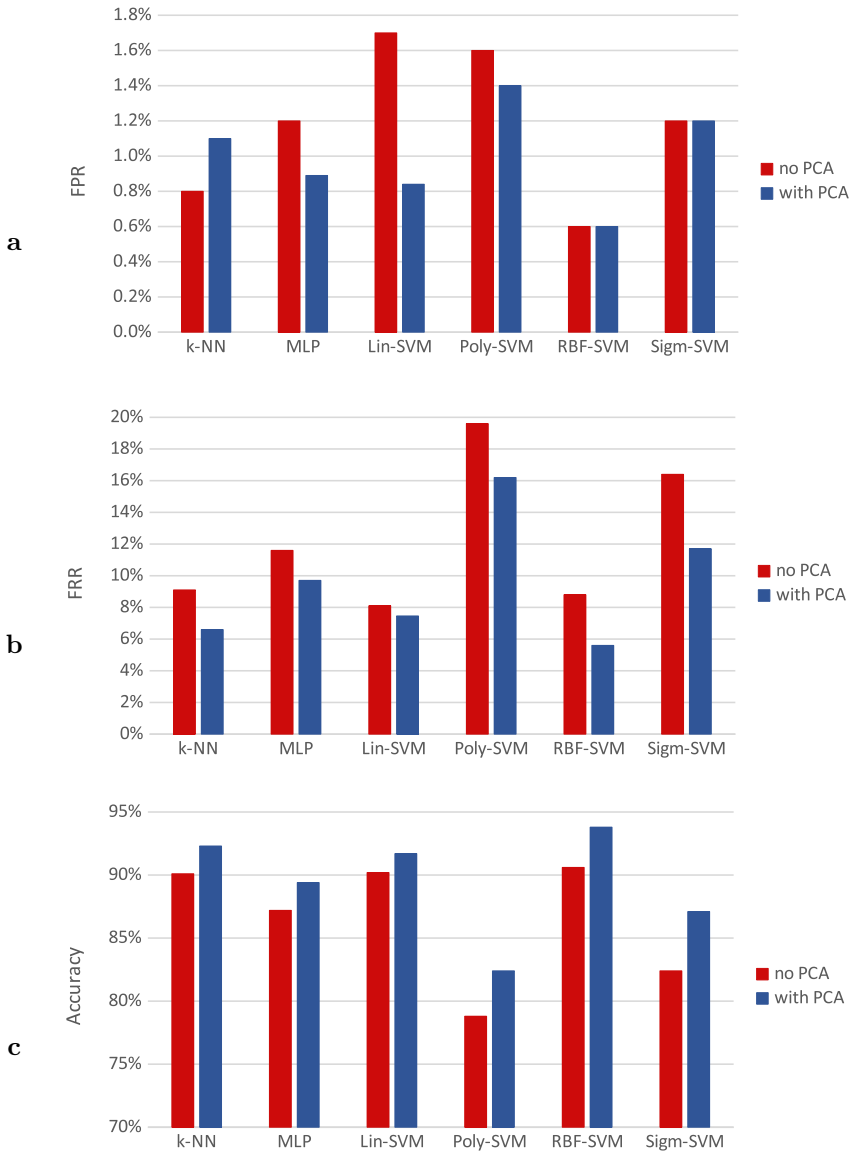

Fig. 3. Quality measures of classification with classifiers as listed in Tab. 1 (here, names abbreviated w.r.t. table for clarity). (**a**) False Positive Rate (FPR); (**b**) False Rejection Rate (FRR); (**c**) Recognition Accuracy (average of five folds).

Tab. 2. Confusion matrix for proposed feature extraction with PCA and RBF-SVM classifier.

Character		**Confused with characters**				
ੳ	ੳ 95%	ੜ 2%	੩ 1%	ਰ 2%		
ਅ	ਅ 98%	ਘ 2%				
ੲ	ੲ 91%	ੲ 3%	ਟ 3%	ਰ 2%	ਵ 1%	
ਸ	ਸ 89%	ਭ 1%	ਗ 6%	ਣ 1%	ਮ 3%	
ਹ	ਹ 94%	ਕ 1%	ਭ 2%	੩ 1%	ਰ 2%	
ਕ	ਕ 93%	ਧ 1%	ਰ 4%	ੜ2%		
ਖ	ਖ 96%	ਘ 1%	ਥ 2%	ਪ 1%		
ਗ	ਗ 91%	ਸ 7%	ਧ 1%	ਰ1%		
ਘ	ਘ 95%	ਭ 2%	ਅ 3%			
ਙ	ਙ 91%	ਥ 4%	ਭ 2%	ਪ 2%	ੜ 1%	
ਚ	ਚ 92%	ਜ 2%	ਧ 5%	ਬ %		
ਛ	ਛ 91%	ੳ 2%	ਭ 2%	ਢ 3%	ਯ 2%	
ਜ	ਜ 93%	ਝ 2%	ਢ 4%	ਨ 1%		
ਝ	ਝ 91%	ਕ 4%	ਙ 1%	ਭ 3%	ਢ 1%	
ੲ	ੲ 91%	ਚ 2%	ਧ 2%	ਵ 5%		
ਟ	ਟ 92%	ਫ 6%	ਵ 2%			
ਠ	ਠ 92%	ਚ 2%	ਛ 1%	ਟ 2%	੩ 3%	
ਡ	ਡ 92%	੩ 3%	ਰ 3%	ੜ 2%		
ਢ	ਢ 91%	ਸ 3%	ਚ 2%	ਛ 4%		
ਣ	ਣ 93%	ੲ 4%	ਛ 3%			
੩	੩ 92%	ਭ 2%	ਬ 3%	ੜ %		
ਥ	ਥ 87%	ਧ 8%	ਖ 4%	ਪ 1%		
ਦ	ਦ 92%	ਚ 4%	ਢ 2%	ਵ 1%	ੜ 1%	
ਧ	ਧ 83%	ਬ 5%	ਥ 4%	ਪ 5%	ਵ 2%	ੜ 1%
ਨ	ਨ 92%	ਗ 1%	ਠ 2%	ਫ 2%	ਲ 2%	ਵ 1%
ਪ	ਪ 92%	ਖ 4%	ਥ 3%	ਵ 1%		
ਫ	ਫ 95%	ਗ 1%	ਛ 3%	ਟ 1%		
ਬ	ਬ 91%	ਸ 1%	ਧ 2%	ਨ 4%	ਭ 2%	
ਭ	ਭ 93%	ਖ 2%	ਙ 3%	ਝ1%	ਨ 1%	
ਮ	ਮ 95%	ਸ 1%	ਖ 2%	ਙ 1%	ਝ 1%	
ਯ	ਯ 93%	ਠ 2%	ਫ 3%	ਢ2%		
ਰ	ਰ 89%	ਠ 5%	ੲ 2%	ਭ 2%	ਨ 1%	ੜ 1%
ਲ	ਲ 92%	ੲ 2%	ਨ 4%	ਠ 2%		
ਵ	ਵ 94%	ੲ 1%	ਛ 1%	ੲ 3%	ਣ 1%	
ੜ	ੜ 92%	ੲ 4%	ਜ 1%	ਭ 3%		

6. Conclusion and scope of future research

A feature extraction technique has been presented for offline handwritten Gurmukhi character recognition of 35-class problem. For classification, we have used k-NN, SVM and MLP classifiers. The best recognition accuracy of 93.8% has been accomplished using 5-fold cross validation technique with the proposed feature extraction technique and RBF-SVM classifier. We have used 7 000 specimens of isolated offline handwritten Gurmukhi characters in experimentation work. This precision can likely be increased by considering a larger data set while training the classifier. This work can likewise be extended to other Indian scripts: Bengali, Devanagari, Tamil, and so forth.

References

[1] U. Bhattacharya, M. Shridhar, and S. K. Parui. On recognition of handwritten bangla characters. In P. K. Kalra and S. Peleg, editors, *Proc. 5th Indian Conf. Computer Vision, Graphics and Image Processing ICVGIP 2006*, pages 817–828. Springer Berlin Heidelberg, Madurai, India, December 13-16, 2006. doi:10.1007/11949619_73.

[2] H. Bunke and T. Varga. Off-line Roman cursive handwriting recognition. In B. B. Chaudhuri, editor, *Digital Document Processing: Major Directions and Recent Advances*, pages 165–183. Springer, London, 2007. doi:10.1007/978-1-84628-726-8_8.

[3] E. Grosicki and H. E. Abed. ICDAR 2009 Handwriting Recognition Competition. In *Proc. 10th Int. Conf. Document Analysis and Recognition ICDAR 2009*, pages 1398–1402, July 2009. doi:10.1109/ICDAR.2009.184.

[4] A. Kacem, N. Aouïti, and A. Belaïd. Structural features extraction for handwritten arabic personal names recognition. In *Proc. Int. Conf. Frontiers in Handwriting Recognition ICFHR 2012*, pages 268–273, September 2012. doi:10.1109/ICFHR.2012.276.

[5] M. Kumar, M. K. Jindal, and R. K. Sharma. A novel hierarchical technique for offline handwritten Gurmukhi character recognition. *National Academy Science Letters*, 37(6):567–572, December 2014. doi:10.1007/s40009-014-0280-1.

[6] M. Kumar, M. K. Jindal, and R. K. Sharma. Offline handwritten Gurmukhi character recognition: Analytical study of different transformations. *Proceedings of the National Academy of Sciences, India Section A: Physical Sciences*, 87(1):137–143, March 2017. doi:10.1007/s40010-016-0284-y.

[7] M. Kumar, R. K. Sharma, and M. K. Jindal. Efficient feature extraction techniques for offline handwritten Gurmukhi character recognition. *National Academy Science Letters*, 37(4):381–391, August 2014. doi:10.1007/s40009-014-0253-4.

[8] R. S. Kunte and R. D. S. Samuel. A simple and efficient optical character recognition system for basic symbols in printed Kannada text. *Sadhana*, 32(5):521–533, October 2007. http://www.ias.ac.in/article/fulltext/sadh/032/05/0521-0533.

[9] L. M. Lorigo and V. Govindaraju. Offline Arabic handwriting recognition: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5):712–724, May 2006. doi:10.1109/TPAMI.2006.102.

[10] U. Pal, T. Wakabayashi, and F. Kimura. Handwritten Bangla compound character recognition using gradient feature. In *Proc. 10th Int. Conf. Information Technology ICIT 2007*, pages 208–213, December 2007. doi:10.1109/ICIT.2007.62.

[11] U. Pal, T. Wakabayashi, and F. Kimura. Comparative study of Devnagari handwritten character recognition using different feature and classifiers. In *Proc. 10th Int. Conf. Document Analysis and Recognition ICDAR 2009*, pages 1111–1115, July 2009. doi:10.1109/ICDAR.2009.244.

[12] N. Sharma, U. Pal, F. Kimura, and S. Pal. Recognition of off-line handwritten Devnagari characters using quadratic classifier. In P. K. Kalra and S. Peleg, editors, *Proc. 5th Indian Conf. Computer Vision, Graphics and Image Processing ICVGIP 2006*, pages 805–816. Springer Berlin Heidelberg, Madurai, India, December 13-16, 2006. doi:10.1007/11949619_72.

[13] D. C. Tran, P. Franco, and J. M. Ogier. Accented handwritten character recognition using SVM – application to French. In *Proc. 12th Int. Conf. Frontiers in Handwriting Recognition ICFHR 2010*, pages 65–71, November 2010. doi:10.1109/ICFHR.2010.16.

[14] X. Wang, V. Govindaraju, and S. Srihari. Holistic recognition of handwritten character pairs. *Pattern Recognition*, 33(12):1967–1973, 2000. doi:10.1016/S0031-3203(99)00204-6.

[15] T. Y. Zhang and C. Y. Suen. A fast parallel algorithm for thinning digital patterns. *Commun. ACM*, 27(3):236–239, March 1984. doi:10.1145/357994.358023.

[16] B. Zhu, X.-D. Zhou, C.-L. Liu, and M. Nakagawa. A robust model for on-line handwritten Japanese text recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, 13(2):121–131, June 2010. doi:10.1007/s10032-009-0111-y.

Munish Kumar received his Master's degree in Computer Science & Engineering from Thapar University, Patiala, India, in 2008. He received his Ph.D. degree in Computer Applications from Thapar University, Patiala, India in 2015. He started his career as an Assistant Professor in computer application at Jaito centre of Punjabi University, Patiala. Presently, he is working as Assistant Professor in Department of Computer Applications, GZS Campus College of Engineering & Technology, Bathinda, Punjab, India. His research interests include Character Recognition and Pattern Recognition.

Prof. Manish Kumar Jindal received his Bachelor's degree in science in 1996 and Post Graduate degree in Computer Applications from Punjabi University, Patiala, India, in 1999. He holds a Gold Medal in his Post graduation. He received his Ph.D. degree in Computer Science & Engineering from Thapar University, Patiala, India in 2008. He is working as Professor in Panjab University Regional Centre, Muktsar, Punjab, India. His research interests include Character Recognition.

Prof. Rajendra Kumar Sharma received his Ph.D. degree in Mathematics from the University of Roorkee (Now, IIT Roorkee), India, in 1993. He is currently working as Professor at Thapar University, Patiala, India, where he teaches, among other things, statistical models and their usage in computer science. He has been involved in the organization of a number of conferences and other courses at Thapar University, Patiala. His main research interests are statistical models in computer science, Neural Networks, and Pattern Recognition.

Simpel Rani Jindal received her Ph.D. degree in Computer Science from Punjabi University, Patiala, India, in 2016. She is working as Associate Professor in Yadavindra College of Engineering, Talwandi Sabo, Bathinda, Punjab, India. Her research interests include Character Recognition.