Printed Text Recognition System for Multi-Script Image

Inderpreet Kaur and Kiran Jot Singh
Department of ECE, Chandigarh University, Gharuan, India
Email: {inder08preet, kiranjot.693}@gmail.com

***Abstract*—Optical Character Recognition system provides transformation of input text into editable form. Multi-script recognition systems are requisite in the countries like India where different people speak different languages in numerous states of country. In the recent time, multi-script recognition is a demanding problem and research work for expansion of optical character recognition scheme for classification of multi-scripts is needed. In this paper, a multi-script recognition system is proposed for the English, Numerals and Gurumukhi scripts. For recognition the image is processed through various stages like pre-processing, segmentation, feature extraction, and classification. After binarization of the image, it is segmented using line segmentation, word segmentation and character segmentation techniques of proposed system. Then features like number of holes, and projection histogram profiles are calculated for its classification. The system efficiency is calculated by using test images of different text sizes. Arial font is used for English script and Gurbanikalmi font is used for Gurumukhi script to train the system. Results show that proposed system provides high accuracy.**

***Index Terms*—OCR, multi-script, English, Gurumukhi, segmentation, feature extraction, recognition**

I. Introduction

Optical Character Recognition is a scheme that provides automatic conversion of handwritten and machine printed input text into editable and searchable data. Loads of presented research work of optical character recognition system clarified that this system has been advanced to versatile offline and online systems in many areas due to its interesting nature and practical importance of its applications [1], [2]. But these advances are limited to recognition of single script, for example English, Persian and Arabic etc. [3]-[5].

From literature survey, it has been observed that huge amount of research papers has been published for recognition of different scripts. Hand-printed English character recognition based on fuzzy theory [6]. Anupama *et al.* have discussed Character Segmentation for Telugu Image Document using Multiple Histogram Projections [7]. Koshti and Govilkar have purposed a Segmentation of Touching Characters in Handwritten Devanagari Script [8]. Pal *et al.* have offered recognition of English multi-oriented characters [9]. Lehal and Singh [10], [11] have presented segmentation schemes for Gurumukhi text and also developed feature extraction and classification schemes for machine recognition of Gurumukhi characters. Rajashekararadhya and Ranjan have implemented a feature extraction method based on zoning and projection distance metric for extracting the feature of Kannada numerals [12]. Shah *et al.* have presented OCR-based Chassis-Number Recognition Using Artificial Neural Networks [13]. Wong *et al.* have described a technique for recognition of Handwritten Digit Recognition using Multi-Layer Feed-forward Neural Networks with Periodic and Monotonic Activation Functions [14]. In this paper a system has been proposed to identify English, Numerals and Gurumukhi scripts. The features used in the system have been developed after a close study of characteristics of Gurumukhi and English script.

The continuous of paper is as description of some characteristics of scripts is given in Section II. Section III involves explanation of architecture of projected method and segmentation. After that description of extracted features is presented in Section IV. The process of classification of characters using extracted features is explained in Section V. The results are discussed in Section VI which show efficiency of projected method and at last conclusion of work is presented in Section VII.

II. Characteristics of Scripts

The introduction of characteristics of target scripts is essential for implementation of novel technique using OCR system. These characteristics of scripts are also helpful in the selection of features to be extracted.

A. Characteristics of English Script

English is primarily a West Germanic language. It is 3rd native language in the world and widely learned as 2nd language along with mother language. The English language consist 26 characters (both uppercase and lowercase) in which there are 5 vowels (a, e, i, o, u) and 21 consonants. This language is written from left to right. The English characters are in isolated form. The database also includes 0 to 9 numerals. The English uppercase and lowercase characters are illustrated in Fig. 1 and Fig. 2 respectively. The numeral set is shown in Fig. 3.

Manuscript received January 2, 2015; revised December 8, 2015.

doi: 10.18178/ijsps.4.5.411-416

A	B	C	D	E	F	G	H	I	J
K	L	M	N	O	P	Q	R	S	T
U	V	W	X	Y	Z				

Figure 1. The English uppercase character set

a	b	c	d	e	f	g	h	i	j
k	l	m	n	o	p	q	r	s	t
u	v	w	x	y	z				

Figure 2. The English lowercase character set

0	1	2	3	4	5
6	7	8	9		

Figure 3. The numeral set

B. Characteristics of Gurumukhi Script

Punjabi is primarily a Gurumukhi language, which was formulated by Guru Nanak Dev Ji (First Sikh Guru) during 16th century. But this was popularized by Guru Angad Dev Ji (Second Sikh Guru). Punjabi is the 14th most widely spoken language in the world.

The Gurumukhi language includes basic 35 different characters. This language is also written from left to right. This language does not contain lowercase and uppercase characters concept. In Gurumukhi word, there is a headline at the top of characters. This headline connected the characters with each other. The characters share same pixels of headline. Thus segmentation of Gurumukhi characters is difficult as compared to English characters segmentation. The Gurumukhi character set is shown in Fig. 4.

ੳ	ਅ	ੲ	ਸ	ਹ	ਕ	ਖ	ਗ	ਘ	ਙ
ਚ	ਛ	ਜ	ਝ	ਞ	ਟ	ਠ	ਡ	ਢ	ਣ
ਤ	ਥ	ਦ	ਧ	ਨ	ਪ	ਫ	ਬ	ਭ	ਮ
ਯ	ਰ	ਲ	ਵ	ੜ					

Figure 4. The Gurumukhi character set

Each OCR system uses selected features to recognize the characters of target language. While selecting features which have to extract, researcher need to consider that what is type of script (English or Gurumukhi or else) and what is style of characters (printed or handwritten) [15]. Basically feature extraction methods are of two types structural and statistical feature extraction method. Structural feature extraction method belongs to geometry of the text such as number of endpoints, number of holes or loops, number of junctions, number of loops, concavities and convexities [11]. Statistical features extraction method belongs to topology of text such as zoning, chain code, outer profile, crossing count, histogram projection [16], [17].

Classification is a most important stage of optical character recognition system. There are various classification method has been presented by researchers such as statistical methods, structural methods, kernel methods, template matching and artificial neural network [17], [18]. In the proposed system the classifier attempts to recognize the input character on the basis of features which are extracted in the feature extraction section.

Depending upon the types of scripts and other requirements of scripts, the number of holes and histogram projection features has been implemented in this paper. The classification of the scripts is carried out on the basis of these extracted features.

III. Proposed System Architecture

A. Proposed System Architecture

The main steps followed by proposed system architecture to recognize the Punjabi as well as English script are pre-processing, segmentation, feature extraction and classification steps is shown in Fig. 5.

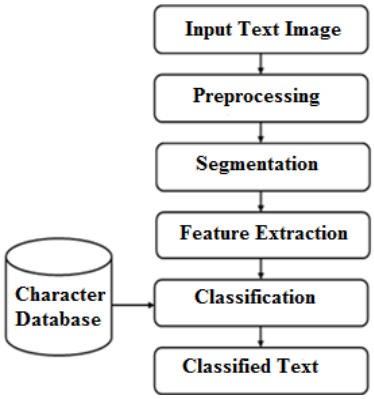

Figure 5. The proposed system architecture

The passage of input image through pre-processing steps is essential to modify raw image for further process in the forthcoming levels of system. The main objectives of pre-processing are binarization and clipping of image [19].

In pre-processing first step is to convert input RGB image into gray scale image and further into binary image. Binarization is a technique by which gray scale images are converted to binary images by selecting a global threshold value. Each image pixel of gray scale image is compared with selected threshold value. Greater pixel value is made 1 and smaller pixel value is made 0.This conversion is made for easy access of image. In binary image identification of basic units which represents text is easy as compare to gray scale and RGB image. This is also advantageous in saving memory. In the second step image is clipped to extract text from the image and deduct spare pixels around text.

B. Segmentation

The segmentation process is applied on both scripts. Segmentation is categorized in three levels as line segmentation, word segmentation and character segmentation as shown in Fig. 6. Normally in binary image value of background pixels is 1 and value of text pixels is 0. For segmentation of scripts by using proposed segmentation algorithm the input image is inverted (background pixels have 0 value and text pixels have 1 value).

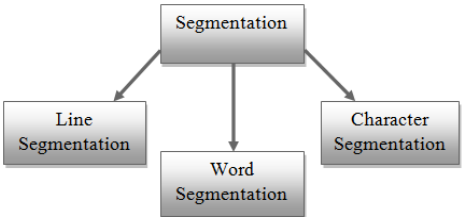

Figure 6. Segmentation technique

1) Line segmentation

In computer composed script of both languages, the height of text lines is almost same. There is also existed an empty horizontal line between two text lines with values of only background pixels. So to find the empty lines between two text lines Horizontal Projection Profile (HPP) is calculated. The image is scanned row wise starting from first row and sum of whole pixels in a row is calculated. If sum is found 0, it means all the pixels in that row are background pixels and this row is considered as empty row. Thus image from first row to this last processed row is segmented and considered as the first line of the image. The process is repeated till whole rows of the image are scanned.

2) Word segmentation

Every extracted text line from above step is processed for words separation. For words segmentation of both scripts same segmentation technique cannot be used because English characters in a word are separated by a blank space and words in each text line are also separated by a blank space but with more width. This feature of English text lines is used for its word segmentation. A threshold value is calculated by analyzing whole blank spaces between the characters as well as words of English line. Image is scanned throughout columns for calculation of Vertical Projection Profile (VPP) and detection of particular spaces which exceeded threshold value to separate the words from a text line.

In Gurumukhi script, characters in each word are connected through a connector line of same pixel value but words in each line are separated through empty columns. To find the empty columns between two words Vertical Projection Profile (VPP) is calculated. Image is scanned column wise starting from first column and sum of whole pixels in a column is calculated. If sum is found 0, it means all the pixels in that column are background pixels and this column is considered as empty space between words. Thus image from first column to this last processed column is segmented and considered as first word of processing line. The process is repeated till whole columns of line are scanned.

3) Character segmentation

Every isolated word which is segmented from above step is processed for character segmentation. For character segmentation of both scripts different techniques are followed. In English words characters are separated by empty space. Hence characters are separated by calculating vertical projection profile in the same manner as discussed above.

The Gurumukhi characters in each word are connected with a connector line. Hence the segmentation of Gurumukhi characters is a difficult task. The connector line which is used to connect the Gurumukhi characters is of same narrow width throughout the word. This feature of the connector line is used to segment the characters from Gurumukhi word.

The width of this connector line is calculated and set as threshold value. Then word image is scanned throughout columns for calculation of Vertical Projection Profile (VPP) and detecting columns having values greater than threshold. The scanning is started from first column. Those columns whose value is greater than threshold value that columns contain character pixels. The column whose value is found less than or equal to threshold value is considered as connector line. The character image from first column to this last processed column is segmented and considered as first character of processing word image.

IV. Feature Extraction

At this stage, relevant features are extracted for classification of target scripts. In this paper number of holes and histogram projection (X & Y projection) profile features are implemented. These features belong to structural and statistical approach respectively. Structural features describe geometry and statistical features describe topology of a character. The number of holes feature is illustrated in Fig. 7.

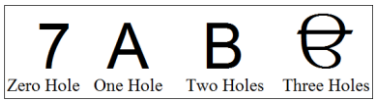

Figure 7. Number of holes

The main contribution of histogram projection features is that each letter has unique horizontal (X) and vertical (Y) projection profile which is beneficial in separating one class character from other class. The horizontal and vertical projections of different characters are shown in Fig. 8.

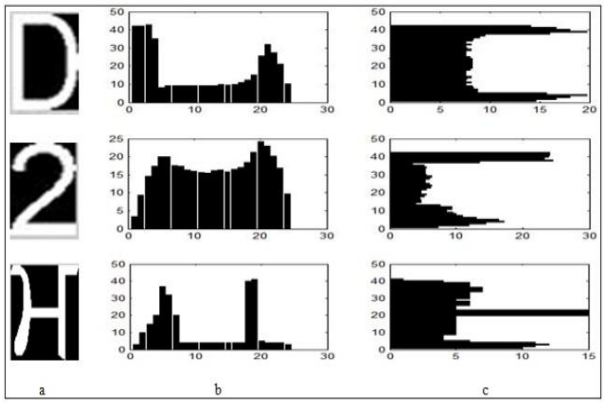

Figure 8. Histogram projection of characters (a) characters (b) X-projection (c) Y-projection

V. Classification

Classification is a main decision making stage of system and uses the features extracted in the previous stage to identify text which is segmented according to preset rules. On the basis of number of holes feature, all characters were divided into four different groups such as

characters with zero hole, characters with one hole, characters with two holes and characters with three holes. This division of characters into different groups enables system to find segmented character only inside the corresponding group. This group is in the form of a structural array which also includes following features:

- Letter value, each character has a corresponding unique value assigned to it.
- Letter horizontal projection profile
- Letter vertical projection profile

Hence incoming character is classified by comparing its features within single group of characters instead of all characters. This helps to improve the computation efficiency and increases recognition accuracy. The result is then sent to a text file.

VI. Results

MATLAB (R2010a) is used to develop the system. A GUI is made for multi-script recognition system as shown in Fig. 9. Arial font is used for database of English script characters and Gurbanikalmi font is used for database of Gurumukhi script to train the system. Testing samples of various sizes were prepared to test efficiency of projected method for both the scripts. Before passing test sample user has to select type of language to be recognized and then testing sample is processed for recognition. After recognizing the sample, it is displayed in the edit text box of GUI and stored in the text file. After that for efficiency calculation purpose, user has to update the status of recognized sample as correct or wrong.

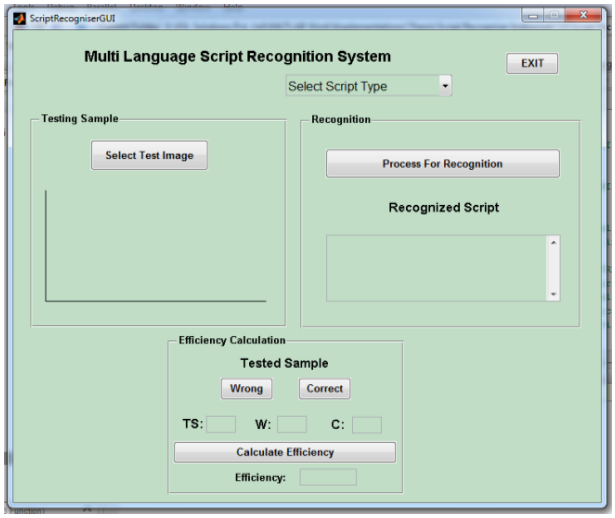

Figure 9. GUI for multi-script recognition system

The system will not take any further input sample until status is not updated. When system has updated status of recognized sample, then system is ready for further use. When user wants to calculate the efficiency of the system then "Calculate Efficiency" button can be used to calculate it.

Different testing samples of both scripts have taken for experimentation. The recognition result of two lines Gurumukhi text sample is displayed in the edit text box of GUI as illustrated in Fig. 10.

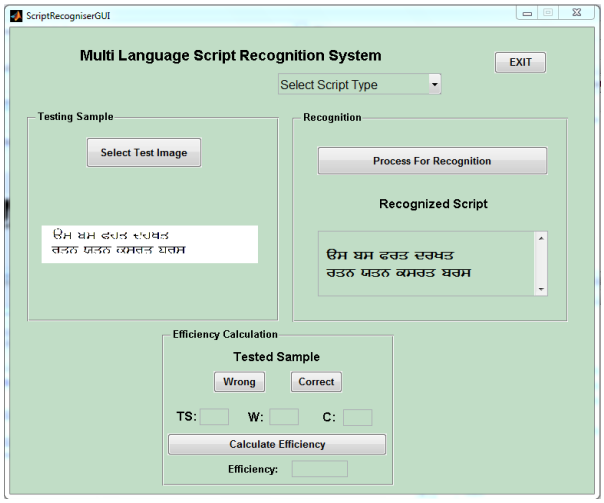

Figure 10. GUI for recognition of Gurumukhi script

Similarly correctly recognized single line English text and mathematical numeral is displayed in edit text box of GUI as shown in Fig. 11. This test sample contains English both uppercase, lowercase letters and Numerals. In the next step, effiecency of the system is calculated for different number of samples to show the overall efficiency of the system.

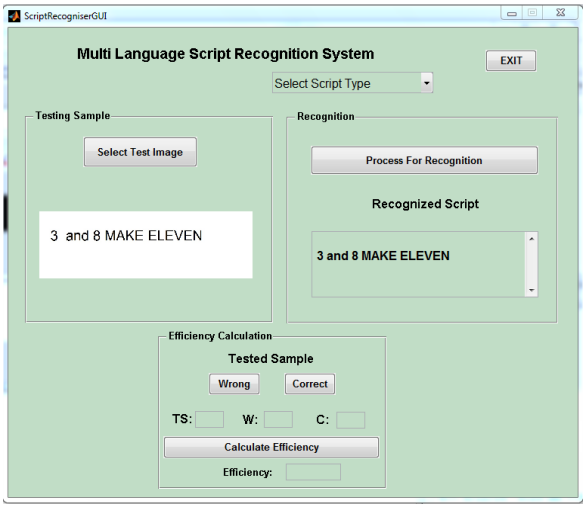

Figure 11. GUI for recognition of Numerals and English

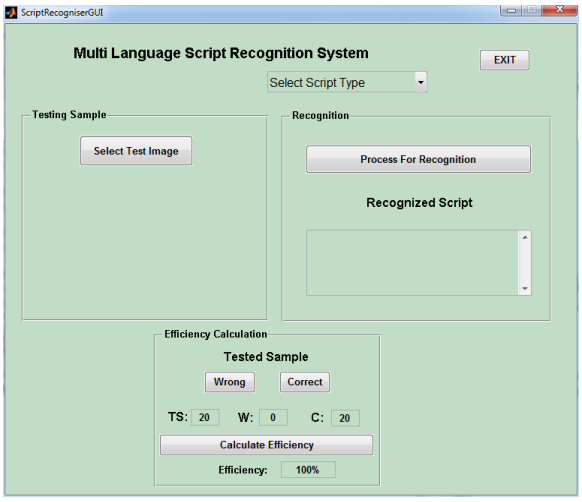

Figure 12. Recognition accuracy of 20 samples

The recognition accuracy of projected method in terms of percentage is reported for four different sets of testing samples as 20, 40, 70 and 100 of both scripts. In these different samples some samples contain single character, some samples contain single word & single line and some of them contain multi lines of different text size of both scripts. For 20 testing samples recognition rate is highest is shown in Fig. 12.

Similarly recognition rate for 40 testing samples is shown in Fig. 13, for 70 testing samples recognition rate is shown in Fig. 14 and for 100 testing samples recognition rate is shown in Fig. 15.

Total 100 samples containing single characters, single words, single lines and multi lines were made for target languages and tested using projected method. Out of these most of samples were segmented & recognized perfectly, but as number of samples increases some samples of single line a single letter is not recognized correctly as shown in Fig. 16.

All recognition results on the sets of testing samples of English, Gurumukhi and Numeral database are shown in Table I.

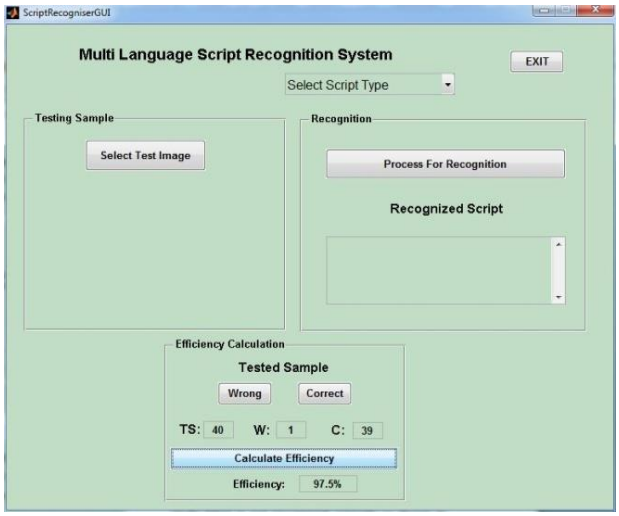

Figure 13. Recognition accuracy of 40 samples

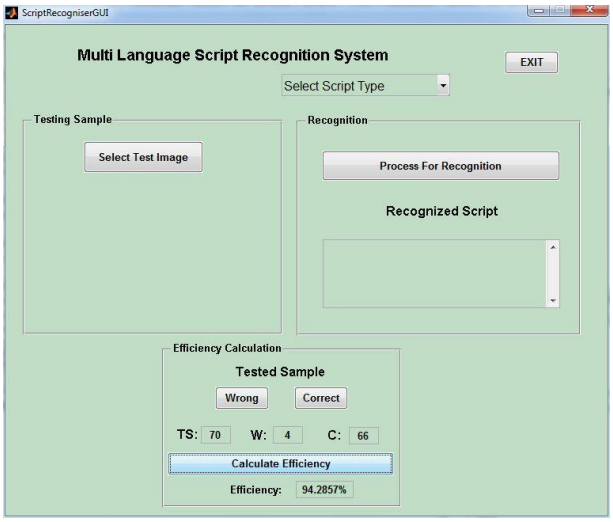

Figure 14. Recognition accuracy of 70 samples

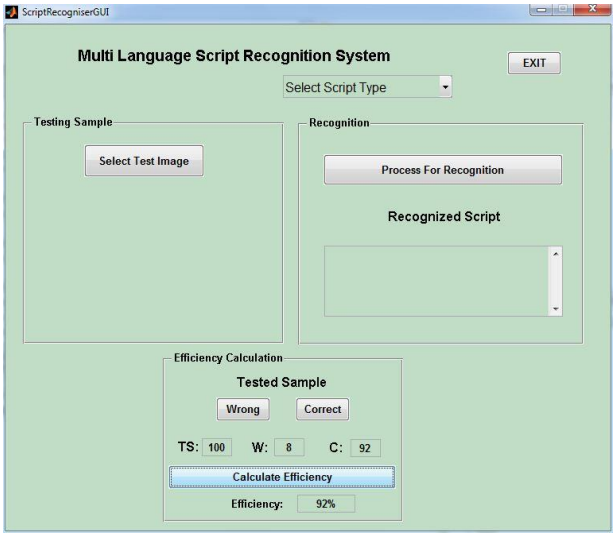

Figure 15. Recognition accuracy of 100 samples

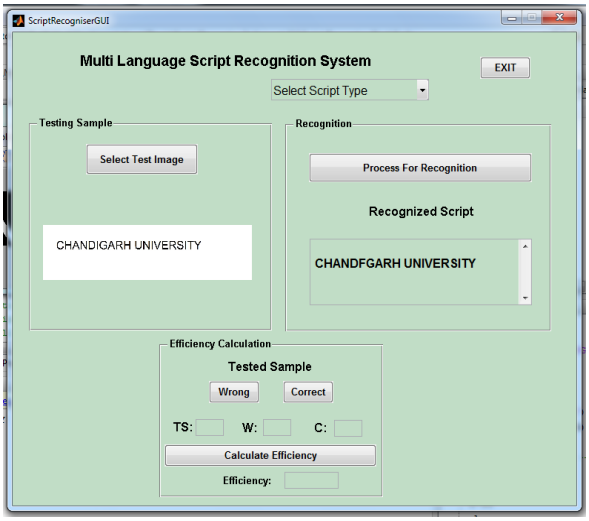

Figure 16. Incorrect recognition

TABLE I. RECOGNITION RESULTS FOR ENGLISH, GURUMUKHI AND NUMERAL

Sr. No.	Number of Samples	Recognition Rate (%)
1	20	100
2	40	97.50
3	70	94.29
4	100	92

VII. CONCLUSION

Our anticipated work has explained a fundamental and capable method for recognizing the multi-scripts. The foremost purpose of this projected method is to provide the multi-script recognizer which is capable to recognize more than one script as English, Gurumukhi text and Numerals. The features such as number of holes and projection histogram are extorted from the segmented characters of any above cited scripts which were segmented using projected segmentation algorithm.

In proposed system, segmentation of English word is difficult as compare to Gurumukhi word segmentation due to puzzlement of space between the characters and

words. The above complication is resolve by calculating the maximum space length between characters. If the space length is less than 25 then it is consider as line has one word. The extorted features are then used for classification purpose. The experimental results show that the proposed method is efficient to recognize English, Numerals and Gurumukhi text. The designed system worked on specific type of font. In future, the system can be trained for more fonts and more features can be added to system for rotational and irregular text lines to achieve high accuracy. Further system can be advanced for handwritten documents.

REFERENCES

[1] N. Arica and F. Y. Vural, "An overview of character recognition focused on off-line handwriting," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 31 pp. 216-233, 2001.

[2] R. Plamondon and S. N. Srihari, "On-Line and off-line handwritten character recognition: A comprehensive survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63-84, 2000.

[3] H. Soltanzadeh and M. Rahmati, "Recognition of Persian handwritten digits using image profiles of multiple orientations," *Pattern Recognition Letters*, pp. 1569-1576, 2004.

[4] J. Pradeep, E. Srinivasan, and S. Himavathi, "Performance analysis of hybrid feature extraction technique for recognizing English handwritten characters," in *Proc. IEEE World Congress on Information and Communication Technologies*, 2012, pp. 4673-4805.

[5] S. Taha, Y. Babiker, and M. Abbas, "Optical character recognition of Arabic printed text," in *Proc. IEEE Conference on Research and Development*, Dec. 2012, pp. 235-240.

[6] P. Mahasukhon and H. Mousavinezhad, "Hand-Printed English character recognition based on fuzzy theory," in *Proc. IEEE International Conference on Electro/Information Technology*, May 2012, pp. 1-4.

[7] N. Anupama, C. Rupa, and E. S. Reddy, "Character segmentation for Telugu image document using multiple histogram projections," *Global Journal of Computer Science and Technology Graphics & Vision*, vol. 13, 2013.

[8] D. V. Koshti and S. Govilkar, "Segmentation of touching characters in handwritten Devanagari script," *International Journal of Computer Science and Its Applications*, vol. 2, 2012.

[9] U. Pal, F. Kimura, K. Roy, and T. Pal, "Recognition of English multi-oriented characters," in *Proc. IEEE International Conference on Pattern Recognition*, 2006, pp. 873-876.

[10] G. S. Lehal and C. Singh, *A Technique for Segmentation of Gurumukhi Text*, Springer Berlin Heidelberg, Sep. 2001, pp. 191-200.

[11] G. S. Lehal and C. Singh, "Feature extraction and classification for OCR of Gurmukhi script," *VIVEK-BOMBAY*, vol. 12, pp. 2-12, 1999.

[12] S. V. Rajashekararadhya and P. V. Ranjan, "Zone based feature extraction algorithm for handwritten numeral recognition of Kannada script," in *Proc. IEEE International Advance Computing Conference*, Mar. 2009, pp. 6-7.

[13] P. Shah, S. Karamchandani, T. Nadkar, N. Gulechha, K. Koli, and K. Lad, "OCR-Based chassis-number recognition using artificial neural networks," in *Proc. IEEE International Conference on Vehicular Electronics and Safety*, Nov. 2009, pp. 31-34.

[14] K. Wong, C. Leung, and S. Chang, "Handwritten digit recognition using multi-layer feed forward neural networks with periodic and monotonic activation functions," in *Proc. 16th International Conference on Pattern Recognition*, 2002, pp. 106-109.

[15] A. K. Jain and T. Taxt "Feature extraction methods for character recognition - A survey," *Proceedings of the Pattern Recognition*, vol. 29, no. 4, pp. 641-662, July 1995.

[16] J. Sadri, Y. Akbari, M. J. Jalili, A. Farahi, and M. Habibi, "A new system for recognition of handwritten Persian bank checks," in *Proc. IEEE International Conference on Document Analysis and Recognition*, Sept. 2012, pp. 925-930.

[17] R. Verma and J. Ali, "A-Survey of feature extraction and classification techniques in OCR systems," *International Journal of Computer Applications & Information Technology*, vol. 1, Nov. 2012.

[18] S. G. Dedgaonkar, A. A. Chandavale, and A. M. Sapkal, "Survey of methods for character recognition," *International Journal of Engineering and Innovative Technology*, vol. 1, May 2012.

[19] R. Mithe, S. Indalkar, and N. Divekar, "Optical character recognition," *International Journal of Recent Technology and Engineering*, vol. 2, Mar. 2013.

Inderpreet Kaur was born in Chander Bhan, Punjab, India in 1989. She received the B.Tech degree in Electronics and Communication Engineering from Yadavindra College of Engineering, Talwandi Sabo in 2012 and M.Tech degree in same stream from Chandigarh University, Gharuan in 2014. Her main research interests are in image processing and optical character recognition.

Kiran Jot Singh received B.Tech and M.Tech degree in Electronics and Communication Engineering from Punjab Technical University in the year 2010 and 2012 respectively. He is currently working as Assistant Professor in Chandigarh University, Gharuan (Punjab). He has keen interest in image processing, embedded systems and analog electronics. He has guided various small as well as large scale projects. He has also filed patents in different fields of interest.