ROBUST EXTRACTION AND CLASSIFICATION METHOD FOR GURMUKHI SCRIPTS

Er.Bharti[1], Mr. Manit Kapoor[2], Mr. Naveen Dhillon[3]

[1]P.G Student, [2,3] Assistant Professor, Department of ECE,
RIET, Phagwara, Punjab (India)

ABSTRACT

In this paper, a word extraction and hybrid classification scheme for recognition of Gurmukhi script is described. The extraction of Gurmukhi words from machine printed images has been a very intensive area of research during last decade due to its wide range of applications as a solution to real world problems. lot of work has been done on languages like English, Arabic, Chinese, Devnagari and Urdu. The whole process consists of two stages. The first, word extraction stage analyzes the set of isolated characters and selects a set of features that can be used to uniquely identify characters. Main advantage of this system is its accuracy to extract the Gurmukhi word. Input to the system is the scanned images from newspapers, magazines and old books and extract the Gurmukhi word from machine printed document Images. The classification process is carried out in three steps. In the first step, the characters are grouped into three sets depending on their zonal position (upper zone, middle zone and lower zone). In the second step, the characters in middle zone set are further distributed into smaller sub-sets by a binary decision tree using a set of robust and font independent features. In the third step, the nearest neighbor classifier is used and the special features distinguishing the characters in each subset are used. One significant point of this scheme, in contrast to the conventional single-stage classifiers where each character image is tested against all prototypes, is that a character image is tested against only certain subsets of classes at each stage. This enhances computational efficiency.

Keywords: Segmentation, Pre-Processing, OCR, Skeletonization, Handwritten Gurmukhi Script, Middle Zone, Upper Zone, Lower Zone.

I. INTRODUCTION

Gurmukhi script is used primarily for Punjabi language, which is the world‟s 14th most widely spoken language. The Character set of Gurmukhi script is as in Fig. 2(a), 2(b) & 2(c). Some of the properties of Gurmukhi script are: Gurmukhi script is cursive and the character set consist of 41 consonants, 9 vowels, 3 sound modifiers (semi-vowels) and 3 half characters, lie at the feet of

consonants. Most of the Gurmukhi characters have a horizontal line at the upper part. The characters of words are connected mostly by this line called head line and so there is no vertical inter-character gap in the letters of a word. There is no concept of upper or lowercase characters. A line of Gurmukhi script can be partitioned into three horizontal zones namely, upper zone, middle zone and lower zone. Consonants are generally present in the middle zone. These zones are shown in Fig.1. The upper and lower zones may contain parts of vowel modifiers and diacritical markers.

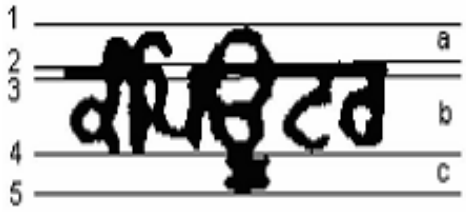

Fig. 1: a) Upper zone from line number 1 to 2, b) Middle Zone from line number 3 to 4, c) lower zone from line number 4 to 5.

In Gurmukhi Script, most of the characters, as shown in Fig. 2, contain a horizontal line at the upper of the middle zone. This line is called the headline. The characters in a word are connected through the headline along with some symbols as i, I, A etc. The headline helps in the recognition of script line positions and character segmentation. The segmentation problem for Gurmukhi script is entirely different from scripts of other common languages such as English, Chinese, and Urdu etc. In Roman script, windows enclosing each character composing a word do not share the same pixel values in horizontal direction. But in Gurmukhi script, as shown in Fig. 1, two or more characters/symbols of same word may share the same pixel values in horizontal direction. This adds to the complication of segmentation problem in Gurmukhi script. Because of these differences in the physical structure of Gurmukhi characters from those of Roman, Chinese Japanese and Arabic scripts, the existing algorithms for character segmentation of these scripts does not work efficiently for handwritten Gurmukhi script.

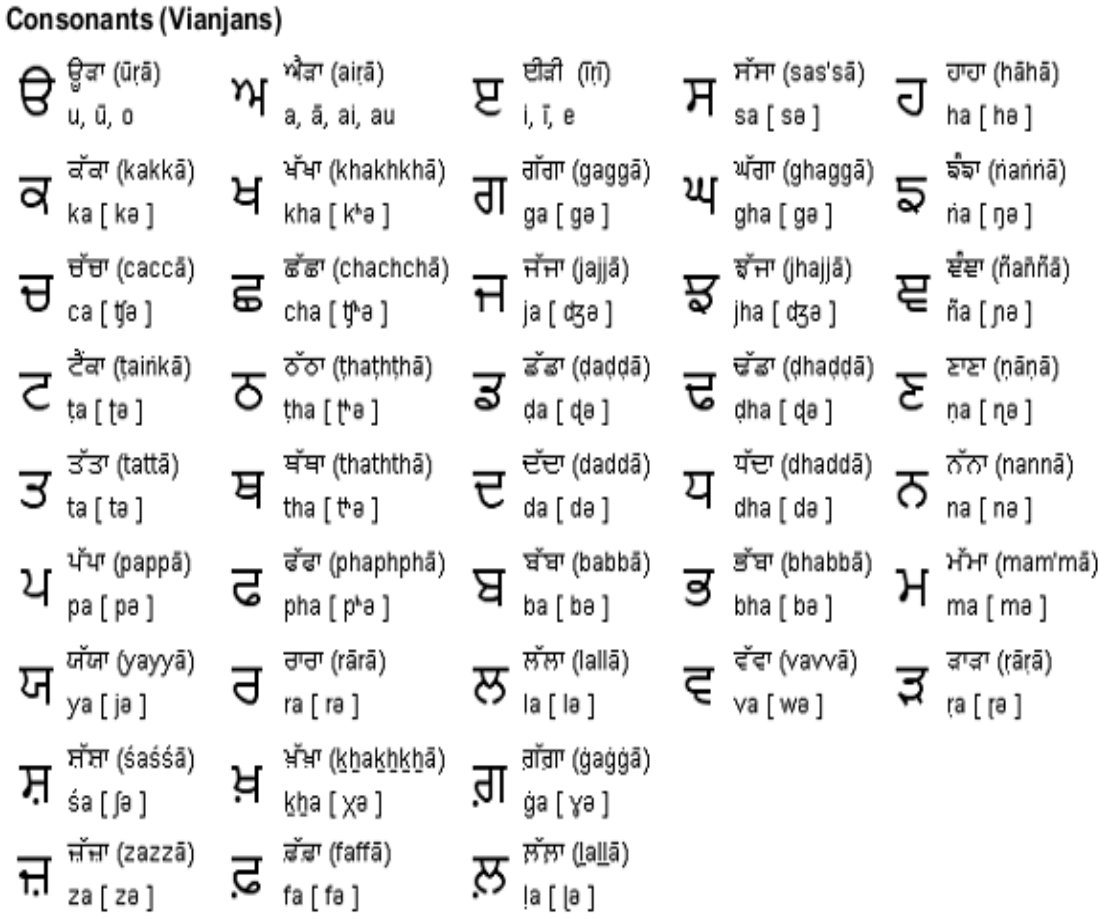

Fig. 2(a) Consonants (Vianjans) of Gurmukhi Script.

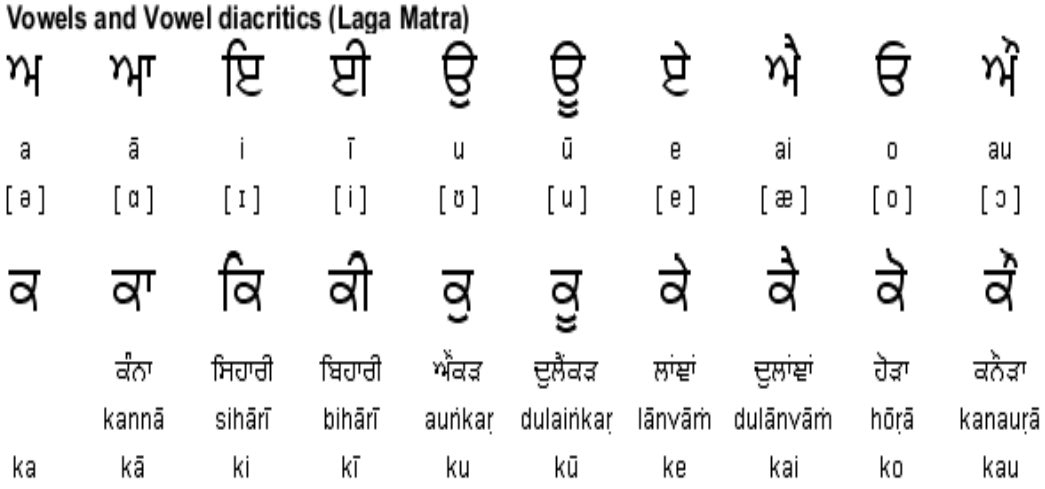

Fig. 2(b) Vowels and Vowel diacritics (Laga Matra)

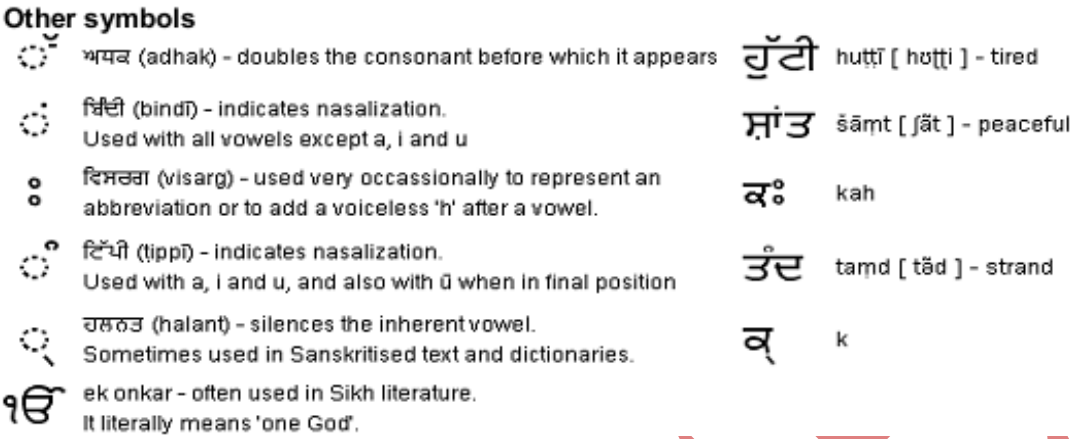

Fig. 2(c) Other symbols

II. PREPROCESSING

Preprocessing is applied on the input binary document so that the effect of spurious noise can be minimized in the subsequent processing stages. In the present study, both salt and pepper noise have been removed using standard algorithm. It is supposed that height and width of document can be known easily. The image is saved in the form of an array. For that purpose a 2-D array with number of rows equal to height of the document and number of columns equal to width of the document is created. Calculate the maximum intensity of pixels in the document using any standard function available in the tool used for the implementation, it is get RGB () method available in java. Scan every pixel of document and compare its intensity with the maximum intensity. If the intensity is equal to maximum intensity, store one in the array at that location and if it is not equal store zero in the array.

III. PROCEDURE FOR WORD EXTRACTION AND CLASSIFICATION OF GURMUKHI SCRIPT.

Following procedure is followed for extraction and classification in Gurmukhi script.

3.1 Line Detection

The following procedure is implemented to find the location of lines in the document.

i. Create an array of size equal to height of the document and with two columns.

ii. Start from the first row and count the number of 1's in that row. If it is zero, move to next row. And if it is not zero, that is the starting location of that line. Store that location in the array.
iii. Check consecutive rows until we get 0.The before we get zero is the ending location of that line. Store that value in the array.
iv. Also calculate the location of maximum intensity in each line and store it in the second column before that line. It would be used as the starting position of characters.
v. Repeat step (ii) to (iv) for the whole document.

3.2 Word Detection

The following procedure is implemented to find location of words in each line.

i. Create a 2-D array.
ii. For each line move from 0^{th} pixel up to width.
iii. Calculate the number of one's in first column from the starting location of line to the ending position of line.
iv. If numbers of 1's are not zero, that is the starting location of word. Save that location in that array. Keep on moving to the right until we get no one in any column. The column with 01's is the ending location of the word. Store that location in array too.
v. Repeat this until we reach the width.
vi. And repeat step (ii) to (v) for each line.

3.3 Character Detection

The following procedure is implemented to find the location of character in each word.

i. Create a 3-d array. Its first index will represent line number. Second index will represent word number and third index will contain the location of character. This array will be created dynamically.
ii. Repeat the step (iii) to (iv) for each line and each word detected so far.
iii. Move from starting position of the word to the ending position of the word.

iv. Start from the starting position of line and move downwards to the ending position. Count the number of one's in that column leaving the location of line with maximum intensity. If it is not zero, that is the starting position of character. Move to right until we get column with no ones. That will be the ending location of character. This process will generate the location of characters.

The above approach was put to number of documents; the image of one such scanned document is given here.

IV. RECOGNITION

The recognition process consists of two main steps, which are explained as follows:

4.1 Feature Extraction

The segmented Punjabi characters are converted into a real valued vector called feature form of 0's and 1's that characterizes the essential information content of the pattern. Each character has some features which play an important role in pattern recognition. Handwritten Punjabi characters have many particular features. Feature extraction describes the relevant shape information contained in a pattern so that the task of classifying the pattern is made easy by a formal procedure. Feature extraction stage in OCR system analysis these character segment and selects a set of features that can be used to uniquely identify that character segment. Mainly this stage is heart of OCR system because output depends on these features.

4.2 Classification

Classification stage is the main decision making stage of the system and uses the features extracted in the previous stage to identify the text segment according to preset rules. Classification is concerned with making decisions concerning the class membership of a pattern in question. The task in any given situation is to design a decision rule that is easy to compute and will minimize the probability of misclassification relative to the power of feature extraction scheme employed. Patterns are thus transformed by feature extraction process into points in dimensional feature space. A pattern class can then be represented by a region or sub-space of the feature space.

Classification then becomes a problem of determining the region of feature space in which an unknown pattern falls.

V. RESULT & DISCUSSION

5.1 Extract the Punjabi Word

As defined in previous chapter, the problem of recognition of characters can be solved using neural networks. A scheme is proposed to extract the Punjabi word from image. Using neural network, extract the Punjabi word is done in following steps:-

5.2 Input Image

Firstly, input digitized image. Further, this image is used to extract Punjabi word. Fig. 4 represents the step of loading of image.

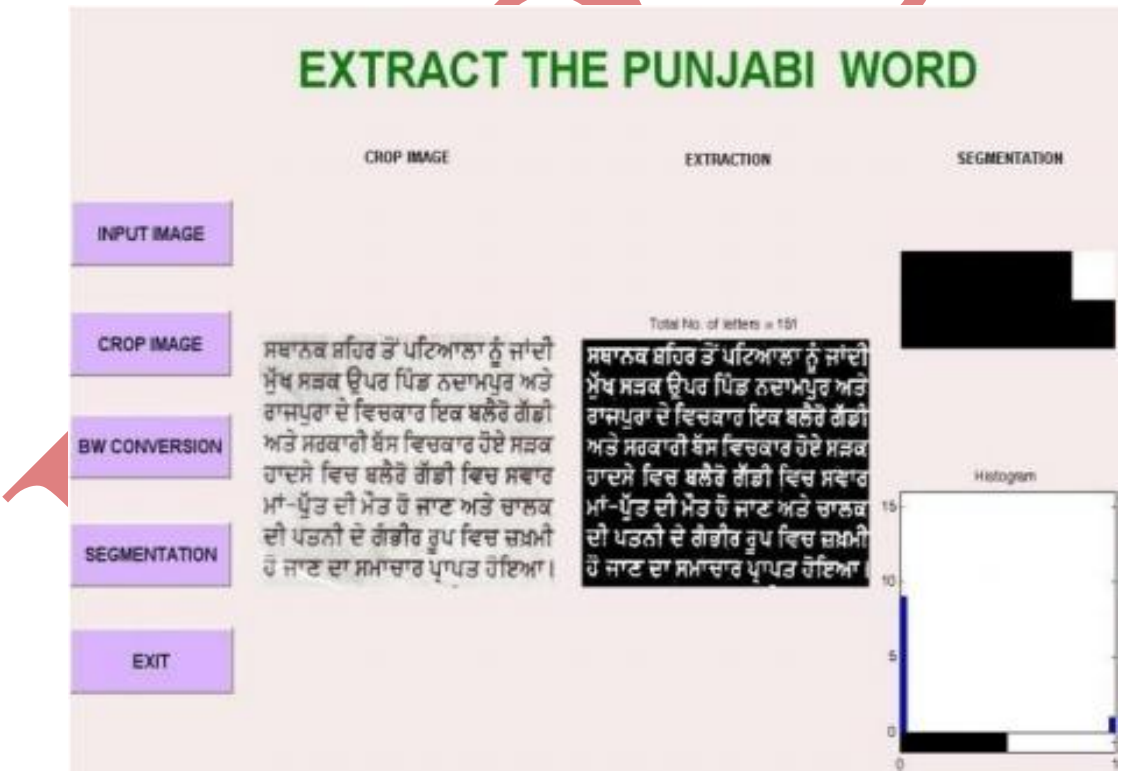

Fig. 3 Punjabi word extraction from the Gurmukhi script.

In fig. 3, there are 5 buttons for processing whole the document. First button is known as input image. When input image button is pressed, a window opens. This window is used to specify the path where the character image is located. After this process, the image is shown

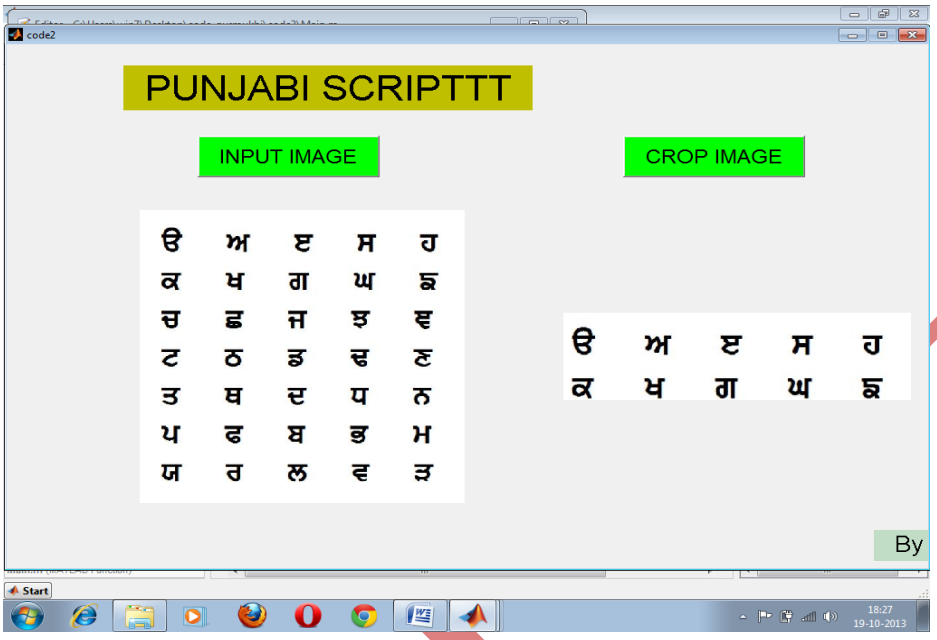

Fig.4 Segmentation of Gurmukhi script.

5.3 Crop Image

After loading of image, selection of characters is performed. After loading image, a particular area of characters is to be cropped for recognition. When a particular area of characters from image is selected, a window represents those characters.

5.4 Black & White Conversion

A separate window is also shown in which bounding box of Image is converted into Black & White using thresholding unit.

VI. CONCLUSION & FUTURE SCOPE

A small set of all characters using back propagation neural network is trained then testing was performed on other character set. The accuracy of network was very low. Then some other character images in the old character set are added and trained the network using new sets. Then

again testing was performed on some new image sets written by different fonts and it was found that accuracy of the network increases slightly in some cases. Again some new character images into old character set are added (on which network was trained) and trained the network using this new set. The network is presented new character images and it has been seen that recognition increases, although at a slow rate. The result of the last training by 25 character set and testing with the 6 character set are presented. It can be concluded that as the network is trained with more number of sets, the accuracy of extraction of Punjabi word will increase definitely. In future work, this can be implemented for recognition & extraction of complete Gurmukhi words including lower & upper Zone Characters.

REFERENCES

[1] R. M. K. Sinha, "Rule based contextual post processing for Devnagari text recognition", Pattern Recognition, Vol. 20, pp. 475-485 (1985).

[2] J. Mantas, "An overview of character recognition methodologies", Pattern Recognition, Vol. 19, pp. 425-430 (1986).

[3] H. Almuallim and S. Yamagochi, "A method of recognition of Arabic cursive handwriting", Pattern Recognition, Vol. 9, pp. 715-722 (1987).

[4] V. K. Govindan and A. P. Shivaprasad, "Character recognition – A survey", Pattern Recognition, Vol. 23, pp. 71-683 (1990).

[5] S. Shlien, "Nonparametric classification using matched binary decision trees", Pattern Recognition Letters, Vol. 13, pp. 83-87 (1992).

[6] Y. S. Huang and C. Y. Suen, "A method of combining multiple experts for the recognition of unconstrained handwritten numerals", IEEE Trans. Pattern Analysis Mach. Intelligence, Vol. 17, No.1, pp. 90-93 (1995).

[7] S. Kumar, "A technique for recognition of printed text in Gurmukhi script", M.Tech Thesis, Punjabi University, (1997).

[8] G. S. Lehal and S. Madan, A New Approach to Skew detection and Correction of Machine Printed Gurmukhi Script,Proceedings 2nd International Conference on Knowledge Based Computer Systems, Mumbai, India, pp. 215-224 (1998).

[9] G. S. Lehal and P. Singh, "A Technique for Segmentation of Machine Printed Gurmukhi Script", Proceedings 4th International Conference on Cognitive Systems, Delhi, India, pp. 283-287 (1998).

[10] G. S. Lehal and R. Dhir, "A Range Free Skew Detection Technique for Digitized Gurmukhi Script Documents", In Proceedings 5th International Conference of Document Analysis and Recognition, IEEE Computer Society Press, California, pp. 147-52 (1999).

[11] A. K. Goyal, G. S. Lehal and S. S. Deol, "Segmentation of Machine Printed Gurmukhi Script", Proceedings 9th International Graphonomics Society Conference, Singapore, pp. 293-297 (1999).

[12] A. K. Goyal, G. S. Lehal and J. Behal, "Machine Printed Gurmukhi Script Character Recognition Using Neural Networks", Proceedings 5th International Conference on Cognitive Systems, Delhi, India (1999).

[13] G. S. Lehal, C. Singh, "A Gurmukhi Script Recognition System", Proceedings 15th ICPR, vol. 2, pp. 557-560, Barcelona, Spain (2000).

[14] G. S. Lehal and Chandan Singh, "Text segmentation of machine printed Gurmukhi script", Document Recognition and Retrieval VIII, Proceedings SPIE, USA, vol. 4307, pp. 223-231 (2001).

[15] Veena Bansal and R. M. K. Sinha, "Segmentation of touching and Fused Devanagari characters", Pattern recognition, vol. 35, pp. 875-893 (2002).

[16] Devasar, N. M, Madan, "A Hybrid Approach to Character Segmentation of Gurmukhi Script Characters", Proceedings of the 32nd Applied Imagery Pattern Recognition Workshop (AIPR'2003).

[17] M. K. Jindal, G. S. Lehal, "Segmentation Problems and Solutions in Printed Degraded Gurmukhi Script", IJSP, Vol. 2(4): ISSN 1304-4494 (2005).

[18] Rajiv K. Sharma & Dr. Amardeep Singh, "Segmentation of Handwritten Text in Gurmukhi Script", International Journal of Computer Science and Security, volume 2, issue 3 (2006).

[19] Anuj Sharma, Rajesh Kumar, R. K. Sharma, "Online Handwritten Gurmukhi Character Recognition Using Elastic Matching", Image and Signal Processing, CISP '08, vol. 2, pp. 391-396, 27-30 May (2008).

[20] D. Sharma, G. S. Lehal, Preety Kathuria, "Digit Extraction and Recognition from Machine Printed Gurmukhi Documents", MORC, Spain (2009).

[21] J. Tripathy, "Reconstruction of Oriya alphabets using Zernike Moments", IJCA, Vol. 8 (8), pp. 26-32 (2010).

[22] Kartar Singh Siddharth et al, "Handwritten Gurmukhi Character Recognition Using Statistical and Background Directional Distribution Features", International Journal on Computer Science and Engineering, Jalandhar, Vol. 3, No. 6 (2011).

[23] Rajiv Kumar and Amardeep Singh, "Character Segmentation in Gurumukhi Handwritten Text using Hybrid Approach", International Journal of Computer Theory and Engineering, Vol. 3, No. 4, August (2011).

[24] Mandeep Kaur, Sanjeev Kumar, "A Recognition System for Handwritten Gurmukhi Characters", International Journal of Engineering Research & Technology, Amritsar, Vol. 1, Issue 6, August (2012).

[25] Usha Rani, Er. Balwinder Singh, Er. Ravinder Singh, "Machine Printed Punjabi Character Recognition Using Morphological Operators on Binary Images", International Journal of Engineering Research & Technology, Patiala, Vol. 1, Issue 3, May (2012).

[26] Gurpreet Singh, et al, "Feature Extraction of Gurmukhi Script and Numerals: A Review of Offline Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, pp. 257-263, January (2013).

[27] Gaurav Singla et al, "Extract the Punjabi Word from Machine Printed Document Images", International Journal of Engineering Research and Application, Vol. 3, Issue 5, pp. 343-348, Sep-Oct (2013).