RESEARCH ARTICLE OPEN ACCESS

Extract the Punjabi Word from Machine Printed Document Images

Gaurav Singla[1], Dr. Parmod Kumar[2]

[1]M.Phil in Computer Application (Research Scholar), University College of Computer Applications, Guru Kashi University, Talwandi Sabo, Punjab, India

[2]Assistant Professor, University College of Computer Applications, Guru Kashi University, Talwandi Sabo, Punjab, India

ABSTRACT

Extract the Punjabi Word from image has been a very intensive area of research during last decades due to it is wide range of solution to real world problems. A lot of work has been done in languages like Chinese, Arabic, Devnagari, Urdu and English. A neural network based Gurmukhi recognition system has been developed. Range free skew detection and correction algorithms for de-skewing Gurmukhi machine printed text skewed at any angle have been developed. If different classifiers cooperate with each other group decisions may reduce errors drastically and achieve a higher performance. The whole process consists of two stages. The first, feature extraction stage analyzes the set of isolated characters and selects a set of features that can be used to uniquely identify characters. The performance depends heavily on what features are being used. Main advantage of this system is its accuracy to extract the Punjabi word. Input to the system is the scanned images from newspaper, magazines and old books and Extract the Punjabi Word from Machine printed Document Images.

Keywords: OCR, Extraction, Punjabi Word, Machine Printed, Gurmukhi, Segmentation, Pre-Processing, Skeletonization, Noise Removal, HGCR, MATLAB

I. Introduction

During the past fifty years optical character recognition systems have come a long way from one-of-a-kind special purpose readers to the multi-purpose production and interactive on-line systems of today. This progress has lowered data capture costs and has caused development of more reliable and accurate OCR systems. Now commercial OCR systems for Latin characters are widely available on personal computers. Further systems in the market can now read a variety of writing styles (e.g. handwritten, printed Omni-font) and character sets including Chinese, Japanese, Korean, Cyrillic and Arabic. Modern OCR software is highly accurate, easy to use and affordable and for the first time OCR looks set to be adopted in all kinds of work environments on a mass scale. Research on Devanagari, Tamil and Telugu optical text recognition started around mid 70s. But the research had only theoretical importance and it did not lead to development of a practical OCR system. It was only around mid 90s that researchers started working for development of complete OCR systems for Indian scripts such as Devanagari and Bangla. The research work on OCR of Gurmukhi script is in its infancy. Lehal and Singh and Goyal et al have presented segmentation schemes for Gurmukhi text. Lehal and Singh have also developed feature extraction and classification schemes for machine recognition of Gurmukhi characters. A post processing system for an OCR of Gurmukhi script has also been presented by Lehal and Singh. Gurmukhi script is used primarily for the Punjabi language which is the world's 14th most widely spoken language. The populace speaking Punjabi is not only confined to North Indian states such as Punjab, Haryana, Delhi, Rajasthan and Jammu & Kashmir but is spread over all parts of the world. It is spoken by over 30 million people in India as well as people living in far flung countries such as UK, USA, Canada, UAE, Singapore, Kenya, Fiji and Malaysia. There is rich literature in this language in the form of scripture, books, poetry etc. Gurmukhi is the first official script adopted by Punjab state. It is also the second language in many northern states of India. It is, therefore important to develop OCR for such a rich and widely used language which may find much practical use in various areas. In this thesis we present a complete OCR system for Gurmukhi script [1].

Today many researchers have been done to recognize Punjabi characters. But the problem of interchanging data between human beings and computing machines is a challenging one. Even today many algorithms have been proposed by many researchers so-that these Gurmukhi characters can be easily recognize. But the efficiency of these algorithms is not satisfactory [2].

Mainly users do Handwritten Character Recognition for interpretation of data which describes handwritten drawing. Handwritten character recognition can be differentiated into two categories i.e. Online Handwritten character recognition and Offline Handwritten character recognition. On-line handwritten character recognition deals with automatic conversion of characters which are written on a special digitizer, tablet PC or PDA where a sensor picks up the pen-tip movements as well as pen-up/pen-down switching. Off-line handwritten character recognition deals with a data set which is obtained from a scanned handwritten document. The main objective of handwritten Gurumukhi character recognition (HGCR) is to recognize the Gurumukhi characters in desirable format from image format so that they can be easily edited.

II. Character Recognition

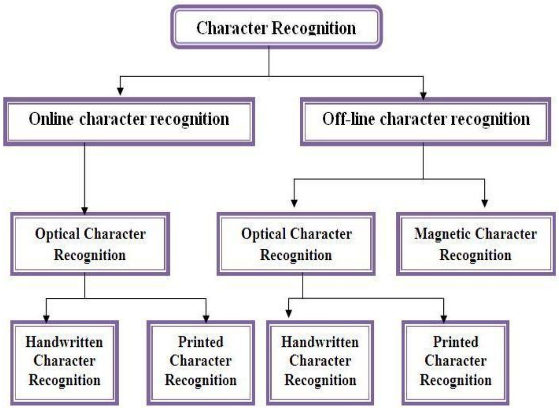

Character Recognition
Online character recognition
Off-line character recognition
Optical Character Recognition
Optical Character Recognition
Magnetic Character Recognition
Handwritten Character Recognition
Printed Character Recognition
Handwritten Character Recognition
Printed Character Recognition

Fig. 1: Classification of Character Recognition

Character recognition is a process which associates a symbolic meaning with objects (letters, Symbols and numbers) have drawn on an image, i.e. character recognition techniques associate a symbolic identity with the image of a character. Mainly, character recognition machine takes the raw data that further implements the process of preprocessing of any recognition system.

2.1 Online recognition system

The handwriting is captured and stored in digital form via different means. Usually, a special pen is used in conjunction with an electronic surface. As the pen moves across the surface, the two-dimensional coordinates of successive points are represented as a function of time and are stored in order [1].

2.2 Offline recognition system

The process of recognizing words that has been scanned from a surface (such as a sheet of paper) and is stored digitally in grey scale format. After being stored, it is conventional to perform further processing to allow superi**or recognition. The offline character recognition can** be further grouped into two types:

- Magnetic Character Recognition (MCR)
- Optical Character Recognition (OCR)

In MCR, the characters are printed with magnetic ink. The reading device can recognize the characters according to the unique magnetic field of each character. MCR is mostly used in banks for check authentication. OCR deals with the recognition of characters acquiring by optical means, typically a scanner or a camera. The characters are in the form of pixelized images, and can be either printed or handwritten, of any size, shape, or orientation. The OCR can be subdivided into handwritten character recognition and printed character recognition. Handwritten Character Recognition is more difficult to implement than printed character recognition due to diverse human handwriting styles and customs. In printed character recognition, the images to be processed are in the forms of standard fonts like Times New Roman, Arial, Courier, etc

III. OCR

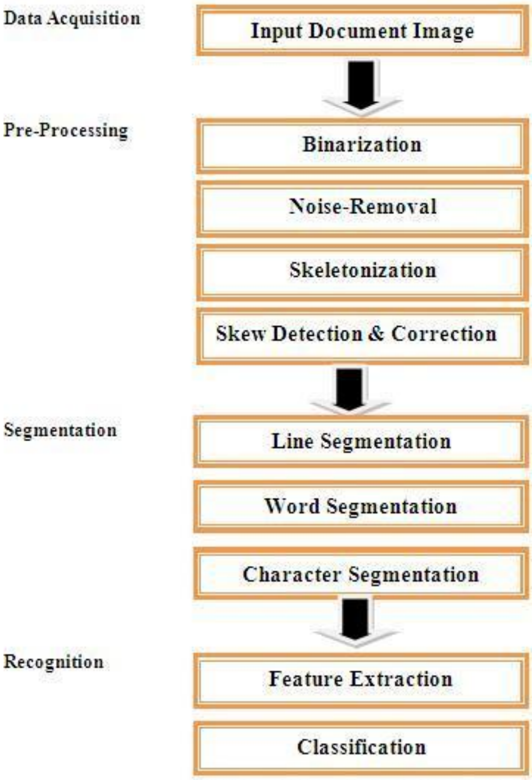


Fig. 2: Block diagram of OCR system

Optical character recognition is the prominent area of research in the world. OCR is the translation of scanned images of handwritten, typewritten or printed document into machine encoded form. This machine encoded form is editable text and compact in size. Character recognition can be applied on printed, type-written or handwritten text. Optical character recognition can be classified as offline recognition and online recognition. The offline recognition is associated with static applications in which entire document first scanned and then processed to recognize while the online recognition is associated with dynamic application as web application where we need recognized result simultaneously or within a fraction of time. Optical Character Recognition (OCR) is the process of converting scanned images of machine printed or handwritten text into a computer processable format. The practical importance of OCR applications as well as the interesting nature of the OCR problem has led to great research interest and measurable advances in this field. But these advances are limited to English, Chinese and Arabic languages [l-3] and there has been very limited reported research on OCR of the scripts of Indian languages [2]. The major Steps of OCR System is given below. The block diagram of OCR system is shown in Fig. 2.

- Data Acquisition
- Pre-Processing
- Segmentation
- Recognition

3.1 Data Acquisition

The input images are acquired from documents containing text by using scanner as an input device. Scanned images are then stored in some picture file such as BMP, JPG etc. After data acquisition, the major steps involved in OCR are Preprocessing Segmentation and Recognition.

Table 1: Comparison between online and offline handwritten characters

Sr No	**Comparisons**	**On-line characters**	**Off-line characters**
1.	Availability of no. of pen-strokes	Yes	No
2.	Raw Data Requirement	#samples/second (e.g.100)	# dots/inch(e.g. 300)
3.	Way of writing	Using digital pen on LCD surface	Paper document
4.	Recognition Rates	Higher	Lower
5.	Accuracy	Higher	Lower

In this case, no information regarding pen-strokes etc. is available. The comparison between offline and online characters is shown in table 1. Digitization produces the digital image which is fed to the pre-processing phase.

3.2 Preprocessing

3.2.1 Binarization

In the pre-processing 1st stage is to convert the input RGB image into gray scale image. Binarization is the process of converting a gray scale image (0 to 255 pixel values) into binary image (0 and 1 pixel values) by selecting a global threshold that separates the foreground from background. Each pixel is compared with the threshold and if it is greater than the threshold it is made 1 or else 0. The histogram of gray scale values of a document image typically consists of two peaks: a high peak corresponding to the white background and a smaller peak corresponding to the foreground. So the task of determining the threshold gray-scale value is one of determining as optimal value in the valley between the two peaks.

3.2.2 Noise removal

The noise introduced during scanning or due to page quality has to be cleared before further processing. It is necessary to filter this noise before process the image. The commonly used approach is to low-pass filter the image and to use it for later processing. The objective in the design of a filter to reduce noise is that it should remove as much of the noise as possible while retaining the entire signal.

3.2.3 Skeletonization

Skeletonization is also called thinning. Skeletonization refers to the process of reducing the width of a line like object from many pixels wide to just single pixel. This process can remove irregularities in letters and in turn makes the recognition algorithm simpler because they only have to operate on a character stroke which is only one pixel wide. It also reduces the memory space required for storing the information about the input characters and no doubt this process reduces the processing time too.

3.2.4 Skew Detection & Correction

Skew Detection refers to the tilt in the bitmapped image of the scanned paper for OCR. It is usually caused if the paper is not fed straight into the scanner. A few degrees of skew (tilt) unavoidable. Skew angle is the angle that the lines of the text in the digital image make with the horizontal direction. There exist many techniques for skew estimation. One skew estimation technique is based on the projection profile of the document; another class of approach is based on nearest neighbour clustering of connected components [5] [6]. Techniques based on the Hough transform and Fourier transform are also employed for skew estimation.

3.3 Segmentation

It is an operation that seeks to decompose an image of sequence of characters into sub images of individual symbols. Character segmentation is a key requirement that determines the utility of conventional Character Recognition systems. It includes line, word and character segmentation [7].

1.3.1 Line segmentation

In a Bangla printed script the text lines are almost of same height provided that the script is written in a specific font size. If the script is composed by a type-machine surely the font size will be uniform everywhere. Between two text lines there is a narrow horizontal band with either no pixel or very few pixels. Hence applying horizontal projection profile (HPP) and detecting the valleys in it, text line bands can be retrieved [8-9].

1.3.2 Word segmentation

From the extracted text lines words get separated. Usually applying vertical projection profile (VPP) and detecting some specific threshold exceeding horizontal gaps words are separated from a text line [10]. Computer composed scripts may contain different font sizes and different styles (i.e. bold, italic etc) and adversely affect the threshold value for identifying isolated words. Hence identifying an effective threshold value is very difficult. It may change twice or more even a single text line.

1.3.3 Character segmentation

Segmentation of characters from the isolated words is the most challenging part of the script segmentation phase [9-11]. Since in computer composed scripts some characters in a container word may partially overlap with one another, it becomes very difficult to isolate those characters properly. Besides some symbols like Chandra Bindu often come between two consecutive characters in a word; then isolating those becomes a tough job. This problem can be overcome by applying contour tracing mechanism [18] or by implementing greedy search technique [9] for letter segmentation.

3.4 Recognition

3.4.1 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.

3.4.2 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.

IV. Requirement

MATLAB

V. Result & Discussion

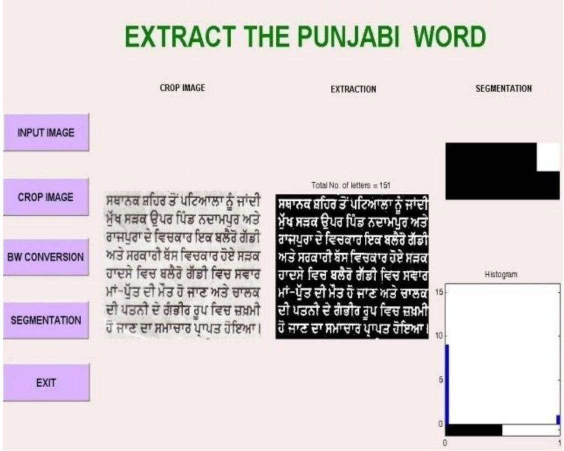
5.1 Extract the Punjabi Word

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:-

5.1.1 Input Image

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

In figure 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.



Total No. of letters = 151

ਸਥਾਨਕ ਸ਼ਹਿਰ ਤੋਂ ਪਟਿਆਲਾ ਨੂੰ ਜਾਂਦੀ
ਮੁੱਖ ਸੜਕ ਉਪਰ ਪਿੰਡ ਨਦਾਮਪੁਰ ਅਤੇ
ਰਾਜਪੁਰਾ ਦੇ ਵਿਚਕਾਰ ਇਕ ਬਲੈਰੋ ਗੱਡੀ
ਅਤੇ ਸਰਕਾਰੀ ਬੱਸ ਵਿਚਕਾਰ ਹੋਏ ਸੜਕ
ਹਾਦਸੇ ਵਿਚ ਬਲੈਰੋ ਗੱਡੀ ਵਿਚ ਸਵਾਰ
ਮਾਂ-ਪੁੱਤ ਦੀ ਮੌਤ ਹੋ ਜਾਣ ਅਤੇ ਚਾਲਕ
ਦੀ ਪਤਨੀ ਦੇ ਗੰਭੀਰ ਰੂਪ ਵਿਚ ਜ਼ਖ਼ਮੀ
ਹੋ ਜਾਣ ਦਾ ਸਮਾਚਾਰ ਪ੍ਰਾਪਤ ਹੋਇਆ।

ਸਥਾਨਕ ਸ਼ਹਿਰ ਤੋਂ ਪਟਿਆਲਾ ਨੂੰ ਜਾਂਦੀ
ਮੁੱਖ ਸੜਕ ਉਪਰ ਪਿੰਡ ਨਦਾਮਪੁਰ ਅਤੇ
ਰਾਜਪੁਰਾ ਦੇ ਵਿਚਕਾਰ ਇਕ ਬਲੈਰੋ ਗੱਡੀ
ਅਤੇ ਸਰਕਾਰੀ ਬੱਸ ਵਿਚਕਾਰ ਹੋਏ ਸੜਕ
ਹਾਦਸੇ ਵਿਚ ਬਲੈਰੋ ਗੱਡੀ ਵਿਚ ਸਵਾਰ
ਮਾਂ-ਪੁੱਤ ਦੀ ਮੌਤ ਹੋ ਜਾਣ ਅਤੇ ਚਾਲਕ
ਦੀ ਪਤਨੀ ਦੇ ਗੰਭੀਰ ਰੂਪ ਵਿਚ ਜ਼ਖ਼ਮੀ
ਹੋ ਜਾਣ ਦਾ ਸਮਾਚਾਰ ਪ੍ਰਾਪਤ ਹੋਇਆ।

Fig. 3: Extract the Punjabi word from Scanned Document

5.1.2 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.

5.1.3 Black & White Conversion

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

5.1.4 Segmentation

The next left hand side window at the top corner shows the segmentation of each word. The advantage of creating bounding box is calculation of area of particular word and the bottom corner window shows the Histogram of Each segmented word of input image. In this, there is no limitation of number of characters. Any number of characters can be boxed which are present in image.

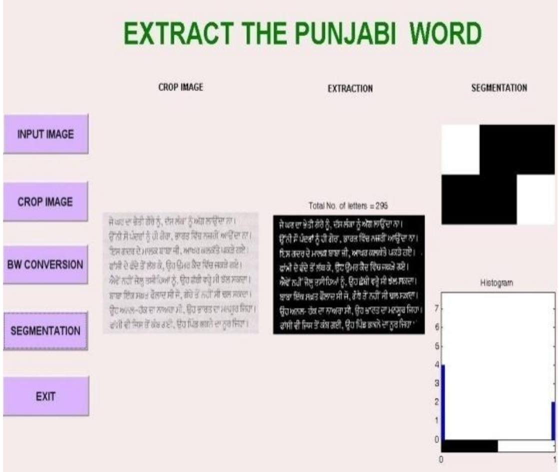


Fig.4: Extract the Punjabi Word from Scanned Newspaper

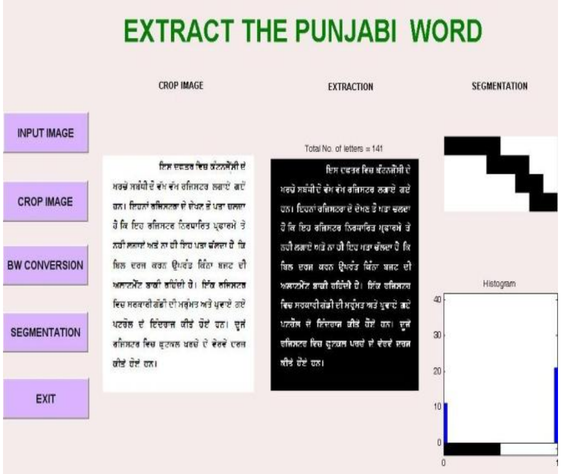


Fig. 5: Extract the Punjabi word from Machine Printed

VI. Conclusion & Future Work

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.

Reference

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

[2] V. K. Govindan and A. P. Shivaprasad, Character recognition – A survey ,*Pattern Recognition, Vol. 23, pp 671-683 (1990).*

[3] B. Al-Badr and S.A. Mahmoud, Survey and bibliography of Arabic optical text Recognition, *Signal Processing, Vol. 41, pp. 49-77(1995).*

[4] 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. 147-152, (1999)*

[5] 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)*

[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, 1995, pp. 90-93.*

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

[8] J. Zhou and T. Pavlidis, Discrimination of characters by a multi-stage recognition process, *Pattern Recognition, Vol. 27, 1994, pp 1539-1549.*

[9] H. S. Baird, Feature identification for hybrid structural/statistical pattern classification, *Computer Vision, Graphics, and Image Processing, 42, 1988, pp. 318-333.*

[10] B. B. Chaudhuri and U. Pal, A complete printed Bangla OCR system, *Pattern Recognition, Vol. 31, pp 531-549 (1998).*

[11] L. Heutte, T. Paquet, J.V. Moreau, Y. Lecourtier and C. Olivier, A structural/statistical feature based vector for handwritten character recognition, *Pattern Recognition Letters, Vol. 19, 1998, pp. 629-641.*

[12] H. Freeman, Boundary encoding and processing, *Picture Processing and Psycholopictorics, B. S. Lipkin and A.*

Rosenfeld, Eds. New York Academic Press. 1970, pp. 210-221.

[13] K. Sethi and B. Chatterjee, Machine recognition of constrained hand printed Devanagari, *Pattern Recognition, Vol. 9, pp. 69-75(1977).*

[14] R. Chandrasekaran, M. Chandrasekaran and G. Siromony, Recognition of Tamil,Malayalam and Devanagari characters, *J. Inst. Electron. Telecom. Engg. (India), Vol. 30, pp. 150-154 (1984).*

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

[16] G. L. Cash and M. Hatamian, Optical Character Recognition by the method of moments, *Computer Vision, Graphics, and Image Processing, 39, 1987, pp. 291-310.*

[17] F. Kimura and M. Shridhar, Handwritten numerical recognition based on multiple algorithms, *Pattern Recognition, Vol. 24, 1991, pp. 969-983.*

[18] S. Shlien, Nonparametric classification using matched binary decision trees, *Pattern Recognition Letters, Vol. 13, 1992, pp. 83-87.*

[19] S. Kumar, A technique for recognition of printed text in Gurmukhi script, *M.Tech. thesis, Punjabi University, (1997).*

[20] 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, 215-224 (1998).*

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

[22] K. Kaur, An approach towards the recognition of machine printed Gurmukhi script, *M.Tech. thesis, Punjabi University, (1999).*

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

[24] A.F.R. Rahman and R. Rahman, Recognition of handwritten Bengali characters: a novel multistage approach, *In Proc. of 9th Biennial Conference of International Graphonomics Society, Singapore, 1999, pp. 299-304.*

[25] Kartar Singh Siddharth , Mahesh Jangid, Renu Dhir, Rajneesh Rani, Handwritten Gurmukhi Character Recognition Using Statistical and Background Directional Distribution Features, *International Journal on Computer Science and Engineering, Jalandhar , Vol. 3 No. 6, 2011.*

[26] 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.*

[27] Dr. Shivaprakash Koliwad, Jayanth.J, A Comparative Study of Segmentation in Mixed-Mode Images, *International journal of Computer Application, Karnataka, Volume 31– No.3, October 2011.*

[28] Rehna, V. J, R. Neha, Sampada. H. K, Character extraction and recognition from document images using segmentation and feature extraction (2012*), IRNet Transactions on Electrical and Electronics Engineering, Bangalore, Volume-1, Issue-2, 2012.*

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

[30] 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.*