ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

onal Journal of Advanced Research in Computer and Communication Engineering
Volume 2, Issue 7, July 2013

DESIGN AND SIMMULATION OF HANDWRITTEN MULTISCRIPT CHARACTER RECOGNITION

Naveed Anjum[1], Tarun Bali[2], Balwinder Raj[3]

Department of ECE, Dr. B.R Ambedkar NIT Jalandhar, Punjab India[1, 2, 3]

Abstract: The work presented in this paper focuses on recognition of isolated handwritten characters in Devanagari and Gurumukhi script. The proposed work uses four feature extraction methods like Zoning density, Projection histograms, Distance profiles and Background Directional Distribution(BDD). On the basis of these four types of features we have formed 10 feature vectors using different combinations of four basic features. This work uses two classifiers like Support Vector machines (SVM), K-Nearest Neighbor (KNN). A total of 7000 samples of characters are taken for Gurumukhi and 7200 samples for Devanagari are used and we have attain a maximum recognition accuracy of 95.79% in case of Gurumukhi recognition and 92.88% for Devanagri. In addition to it we have compared the performance of three similarity based classifiers like Euclidean distance, Manhattan distance and Normalized Histogram Intersection for Gurumukhi and Devanagari characters. Among these three Normalized Histogram Intersection gives the highest accuracy of 89.28% for Gurumukhi and Manhattan distance gave the highest recognition accuracy of 86.14% for Devanagri characters.

Keywords: Character recognition, Feature extraction, Support Vector Machine, Classification

I. INTRODUCTION

India is a multilingual country. In such country a single page of document may contain words of two or more languages. So multi-script character recognition is necessary to read these documents. Optical Character Recognition (OCR) is the process of converting the scanned images of handwritten, typewritten or printed text into machine or computer editable text. [3]. The character recognition is classified into two main categories: Online line recognition and Offline recognition .On-line character recognition deals with a data stream which comes from a transducer while the user is writing. While the Off-line character recognition is performed after the writing is finished. The offline [2] character recognition is further classified as Printed and Handwritten. Earlier OCR was widely used to recognize printed or typewritten documents. But recently, there is an increasing trend to recognize handwritten documents. The recognition of handwritten documents is more complicated in comparison to recognition of printed documents It is because handwritten documents contains unconstrained variations of written styles by different writers even different writing styles of same writer on different times and moods. The heart of the recognition process is feature extraction and the classification. The feature extraction stage extracts the information or the features from the raw data that can be used to uniquely identify characters. In feature extraction each character is represented by feature vector which becomes its identity. Selection of feature extraction is an important parameter in achieving high performance in character recognition systems. The classification is the main decision making stage of the recognition system. The input to the classification stage is the set of feature vectors that are generated in the feature extraction stage. The characters are classified using each of these feature vectors in different classifiers independently. Support Vector Machine (SVM) is a supervised machine learning technique which is a classification tool that uses machine learning theory to maximizes predictive accuracy. K-NN classifier uses the instance based learning by relating unknown pattern to the known according to some distance or some other similarity function. It classifies the object by majority vote of its neighbor**.**

Hand written OCR systems consist of five major stages as shown in figure 1:

A. Image preprocessing: The pre-processing phase normally includes many techniques applied for binarization, noise removal, skew detection, slant correction, normalization, contour making and skeletonization like processes to make character image easy to extract relevant features and efficient recognition [4].

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

onal Journal of Advanced Research in Computer and Communication Engineering
Volume 2, Issue 7, July 2013

B. Segmentation: Segmentation partitions the digital image into multiple segments. It is used to decompose an image of a sequence of characters into sub images of individual symbols by segmenting words and lines [6].

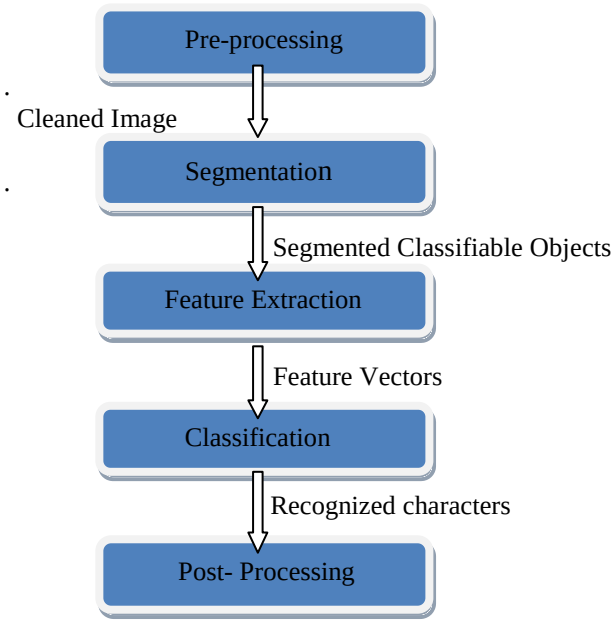

Figure 1: Flow of OCR system

C. Feature Extraction: Feature extraction is extracting information from raw data which is most relevant for classification purpose. In feature extraction stage every character is assigned a feature vector to identify it. This vector is used to distinguish the character from other characters.[7]

D. Classification: Classification is the main decision making stage of OCR system. It uses the features extracted in the previous stage to identify the characters.[3] Classifiers are first trained by a training set of pattern samples to prepare a model which is later used to recognize the test samples. The training data should consist of wide varieties of samples to recognize all possible samples during testing. Some examples of generally practiced classifiers are- Support Vector Machine (SVM), K- Nearest Neighbor (K-NN), Probabilistic Neural Network (PNN).[8]

E. Post processing :In post processing step we bind up our work to create complete machine encoded document through the process of recognition, assigning Unicode values to characters and placing them in appropriate context to make [6]characters, words, sentences, paragraphs and finally whole document

II. CHALLENGES IN OCR DESIGN

Researchers have investigated OCR for a variety of Indian scripts. The work done by many researchers towards handwritten recognition with a wide variety of techniques have been successfully applied. But still there are certain areas which are yet to be explored in order to achieve better accuracy and performance. Apart from it there are certain challenges regarding the recognition of handwritten characters. These are discussed as follows:

(a) To recognize handwritten documents, either online or offline, the character recognition is much affected by style variations of handwriting by different writers and even different styles of same writer on different times.

(b) Distortions like poorly written, degraded or overlapping characters and noise incorporated while digitization is also a major issue in character recognition that affects the recognition accuracy negatively.

(c) After careful examination of the character sets of various Indian scripts, it becomes evident that there are certain characters, having similar shape as that of other characters. Recognition of such confusing characters is a major challenge.

III. DESIGN STEPS OF IMAGE PREPROCESSING

Image Pre- processing is a process in which the scanned images of handwritten characters are first converted to binary image and then various types of techniques are applied in order to remove noise in order to make the images ready for feature extraction and classification purposes. Pre-processing involves a series of operations that are being performed on the scanned images. These are discussed as follows:

A. Binarization: The scanned images of the characters may be colour image format. So it is required to first convert it into gray level image before converting to binary image. Gray level image is one in which each pixel of the image is represented by intensity values lying between 0 and 1. But in binary images there are only two levels 0 and 1. Black pixel is represented by 0 and white with 1.

B. Noise Removing: Scanning process may introduce noise that [6] may be in the form of disconnected line segments, blurred images etc. In order to remove noise special filters are used.

C. *Skew correction*: Deviation of the baseline of the text from horizontal direction is called skew. Document [21] skew often occurs during document scanning or copying. This effect visually appears as a slope of the text lines with respect to the x-axis. Skew lines are made horizontal and making proper correction in the raw image.

D. Slant correction: The character inclination that is normally found in cursive writing is called slant. Slant correction an important step in the pre-processing stage of handwritten character recognition. To correct the slant presented first we need to estimate the slant angle, then

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

onal Journal of Advanced Research in Computer and Communication Engineering
Volume 2, Issue 7, July 2013

horizontal shear transform is applied to all the pixels of images of the character in order to shift them to the left or to the right.

E. Character Normalization: Character normalization is done so that all the characters have the same size. In character normalization the size of the all handwritten characters are normalized to 32×32 matrix.

F. Thinning: Thinning is a morphological operation that is used to remove selected foreground pixels from binary images, somewhat like erosion or opening. It can be used for several applications, but is particularly useful for skeletonization

IV. PROPOSED DESIGN FLOW

The work presented in this paper focuses on recognition of isolated handwritten characters in Devanagari and Gurumukhi script. In this work the main part is the feature extraction and the classification. We have used four feature extraction techniques like Zoning density, Projection histograms, Distance profiles and Background Directional distribution(BDD). On the basis of these four types of features we have formed 10 feature vectors using different combinations of four basic features which are used for classification. We have used two classifiers like SVM and KNN for the classification

V. FEATURE EXTRACTION

Feature extraction is extracting information from raw data which is most relevant for classification purpose and that minimizes the variations within a class and maximizes the variations between classes [9]. Selection of a feature extraction method is an important factor in achieving high recognition performance in character recognition systems. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate with the least amount of confusion. Some of the feature extraction techniques used in our work are discussed below:

Zoning: In zoning, the character image is divided into several overlapping or non overlapping zones [6] From each zone features are extracted to form the feature vector.

Projection histogram: Projection histograms count the number of pixels in specified direction. We have used three types of histogram like horizontal, vertical and diagonal. [10] In this the number of foreground pixels are calculated in horizontal, vertical and diagonal direction.

Distance Profile: In distance profile the number of pixels from the bounding box of character are being calculated. we have used profiles of four sides left, right, top and bottom.

Background Directional Distribution: For these features we have considered the directional distribution of neighbouring background pixels to foreground pixels [10] We computed 8 directional distribution features. To calculate directional distribution values of background pixels for each foreground pixel, we have used the masks for each direction

VI. CLASSIFICATION

Classification stage is the main decision making stage of an OCR system and uses the features extracted in the previous stage to identify the characters. We have used two classifiers in our work. For the recognition of Gurumukhi script, SVM classifier is used and for devanagari, KNN classifier is used. In addition to them we have also used three similarity based classifiers. These are discussed below:

SVM classifier: Support Vector Machine [2] are based on statistical learning theory that uses supervised learning. In supervised learning, a machine is trained instead of programmed, to perform a given task on a number of input-output pairs[11]. Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects [12] having different class memberships. The figure 2 shows the classification of objects having class one of the two: either *triangle* or *diamond*.

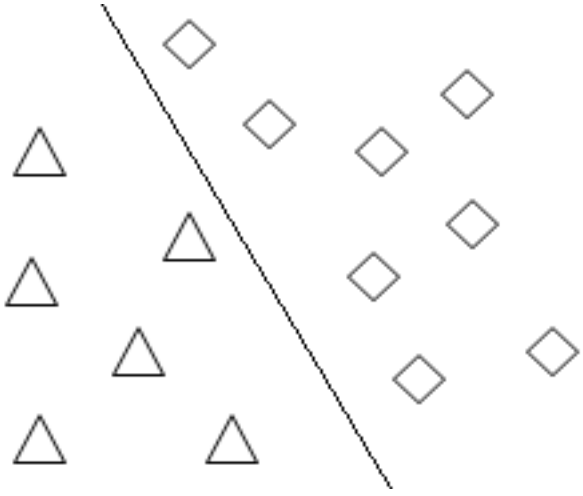

Figure 2: Linear classification of objects by SVM into two classes

The separating line defines a boundary on the right side of which all objects are diamond and to the left of which all objects are triangle. Any new object falling to the right is labeled, i.e., classified, as diamond (or classified as triangle if it falls to the left of the separating line). Classification tasks based on drawing separating lines to distinguish between objects of different class memberships are known as hyper plane classifiers. Support Vector Machines are particularly suited to handle such tasks.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

onal Journal of Advanced Research in Computer and Communication Engineering
Volume 2, Issue 7, July 2013

K- Nearest Neighbor Classifier: K-NN classifier uses the instance based learning by relating unknown pattern to the known according to some distance or some other similarity function.[1] The distance function used to find the nearest neighbor is Euclidian distance. The Euclidean distance between an input feature vector *X* and a library feature vector *C* is given by

$$D=\sqrt{\sum_{i=1}^{N}(Ci - Xi)(Ci - Xi)} \quad (1)$$

where *Ci* is the *i*th library feature and *Xi* is the *i*th input feature and *N* is the number of features used for classification. K specifies the number of nearest neighbors to be considered and the class of majority of these neighbors is determined as the class of unknown pattern

Similarity measure based classifiers: The distance metric can be termed as similarity measure, which is the key component in content-based image retrieval. Some of such classifiers are discussed below:

Euclidean or L2 metric: It calculates the best which the best distance metric for content based image retrieval. If **x** and y are two d-dimensional feature vectors of database image and query image respectively, then [22] Euclidean or **L2** metric are defined **as:**

$$d\,E(x,y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2} \quad (2)$$

Manhattan or L1 metric: If **x** and y[20] are two d-dimensional feature vectors of database image and query image respectively then Manhattan or L1 metric are defined **as:**

$$d\,M(x,y) = \sum_{i=1}^{d}|x_i - y_i| \quad (3)$$

Normalized Histogram intersection: The histogram intersection was proposed by Swain and Ballard. The objective was to find known objects [21] within images using color histograms. It is able to handle partial matches when the object (with feature Q) size is less than the image (with feature T) size. The histogram distance is defined as:

$$d\,HI(Q,T) = \frac{\sum_{i=1}^{n}\min[Q[i],T[i]]}{\sum_{i=1}^{n}T[i]} \quad (4)$$

VII. RESULTS AND DISCUSSION

An annotated sample image database of isolated handwritten characters in Gurumukhi script and Devanagri script has been prepared.In our work a total of 7000 samples of characters are taken for Gurumukhi and 7200 samples for Devanagari. One-fifth of the samples are used for training purposes and four-fifth are used for testing purpose The recognition accuracy obtained by using different combinations of feature extraction methods and classifiers for Gurumukhi characters are given in the Table below

Table 1: Comparison of Gurumukhi character recognition with Reported data

Proposed by	Features used	Classifier	Accuracy
Naveen Garg et.al[12]	Structural Features	Neural Network	83.32%
Anuj Sharma et.al [13]	Strokes recognition and matching	Elastic matching	90.08%
Puneet Jhaji et.al[14]	Zoning Density	SVM	73.83%
Ubeeka Jain et.al[15]	Profiles width, height, aspect ratio	Neocognitro, Neural Network	92.78%
Our Proposed approach	ZD, profile ,histogram, BDD	SVM,KNN, similarity based	95.79%

The recognition accuracy for devanagri characters obtained by using four features extraction techniques (ZD, profile, histogram, BDD) and KNN Classifier are compared with four earlier papers as shown below

Table 2: Comparison of Devanagri character recognition with earlier approaches

S.No	Proposed by	Accuracy (%)
1	N.Sharma et al.[16]	80.36
2	P.S.Deshpande et al [17]	82
3	S.Arora et al[18]	89.12
4	M.Hanmandlu et al.[19]	90.65
5	Proposed Approach	92.885

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

onal Journal of Advanced Research in Computer and Communication Engineering
Volume 2, Issue 7, July 2013

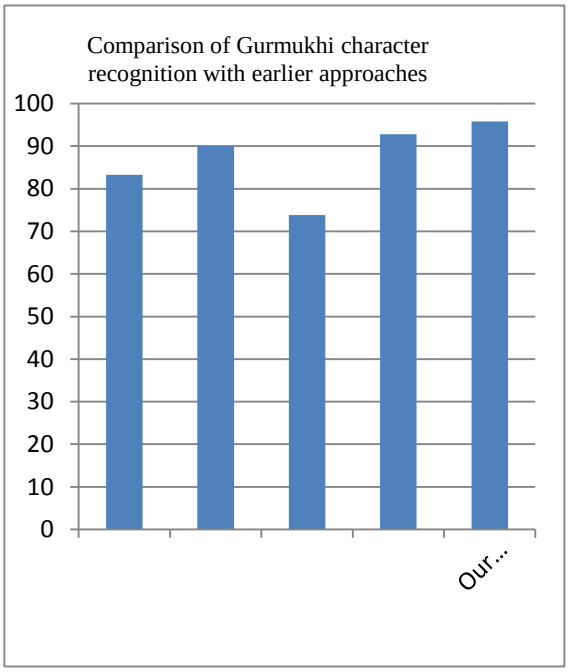

Figure 3: Comparison of Gurmukhi character recognition with earlier approaches

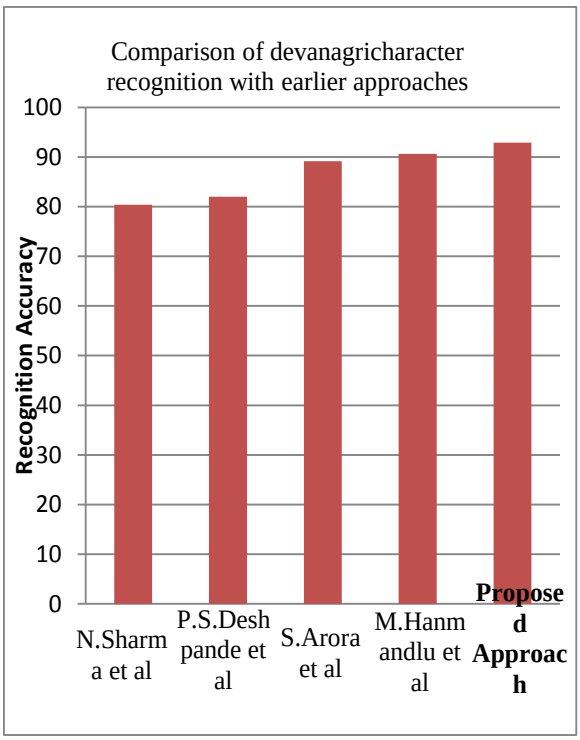

Figure 4: Comparison of Devanagri character recognition with earlier app

We have also compared the performance of three similarity based classifiers for Gurumukhi and Devagari characters the result of which are shown in the table below:

Classifiers	Gurumukhi Characters	Devanagri Characters
Euclidean distance metric	82.71	85.92
Manhattan Distance Metric	88.85	86.14
Normalized Histogram Intersection	89.28	85.28

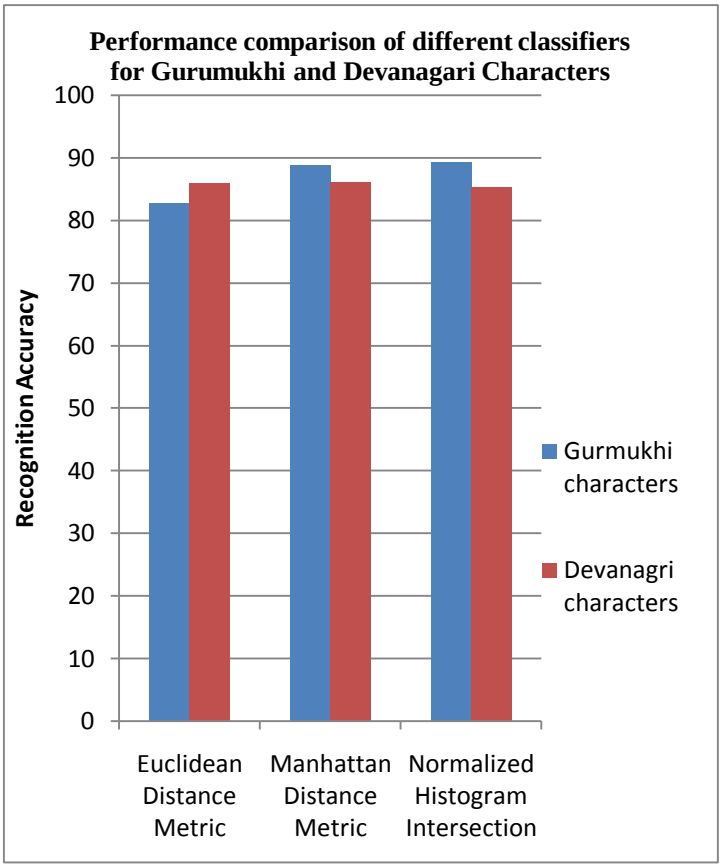

Figure 5: Performance comparison of different classifiers for Gurumukhi and Devanagari Characters

Table 4: Performance comparison of Recognition Accuracy different for different feature vectors Gurumukhi and Devanagari Characters

Feature Vector	Gurumukhi characters	Devanagri characters
FV1	78.71	74.60
FV2	84.35	66.58
FV3	90.83	82.51
FV4	93.21	78.80

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

onal Journal of Advanced Research in Computer and Communication Engineering
Volume 2, Issue 7, July 2013

FV5	90.17	71.91
FV6	95.79	92.88
FV7	89.47	75.87
FV8	93.16	84.80
FV9	90.74	76.26
FV10	94.27	84.47

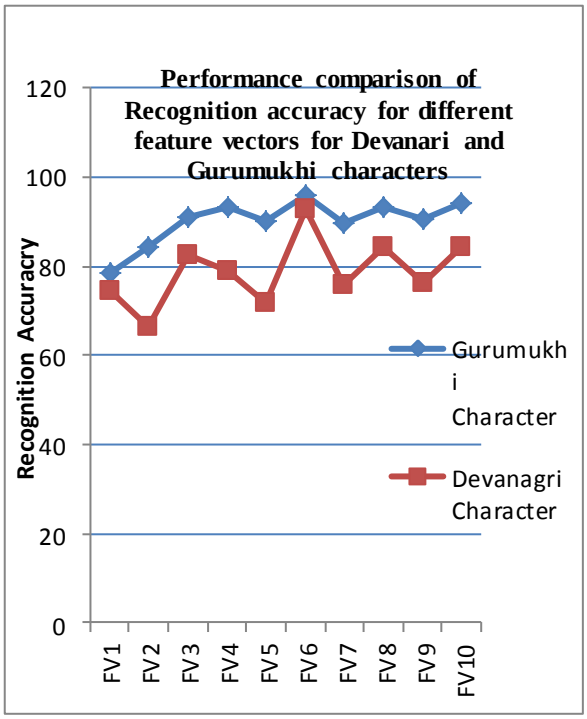

Figure 6: Performance comparison of Recognition Accuracy different for different feature vectors Gurumukhi and Devanagari Characters

CONCLUSION

In this paper, we have presented feature extraction and classification schemes for optical character recognition of Gurumukhi script and devanagari script. For the validation of our result, we compared our result with reported data and significant improvements are observed. We have attain a maximum recognition accuracy of 95.79% in case of Gurumukhi recognition and 92.88% for devanagri for feature vector FV6 in both of them as given in the table 4. The work can be extended to increase the recognition accuracy by adding some more relevant features

REFERANCES

[1] Li Lei , Zhang Li-liang, Su Jing-fei,“ Handwritten character recognition via direction string and nearest neighbor matching” The Journal of China Universities of Posts and Telecommunications, pp 160–165 October 2012

[2]Jomy John, Pramod K. V., Kannan Balakrishnan, “ Unconstrained Handwritten Malayalam Character Recognition using Wavelet Transform and Support vector Machine Classifier” International Conference on Communication Technology and System Design pp 598-605 ,2011

[3] Dharamveer Sharma and Puneet Jhajj “Recognition of Isolated Handwritten Characters in Gurumukhi Script” International Journal of Computer Applications Volume 4– No.8, pp 09-17, August 2010

[4] Muhammad Imran Razzak S. A. Hussain Muhammad Sher “ Numeral Recognition for Urdu Script in Unconstrained environment” International Conference on Emerging Technologies pp 44-47, 2009

[5] S. Chanda and U. Pal “English, Devnagari and Urdu Text Identification” Proceedings of the International Conference on Cognition and Recognition”

[6] Vikas.j. Dongre and Vijay.H.Mankar “ A review of research on devanagari character recognition” International journal of computer applicationsvolume 12 November 2010

[7] Sandhya Arora, Debotosh Bhattacharjee, Mita Nasipuri, “Combining Multiple Feature Extraction Techniques for Handwritten Devnagari Character Recognition” IEEE Region 10 Colloquium and the Third ICIIS, Kharagpur, INDIA December 8-10, 2008

[8] Malik Waqas Sagheer, Chun Lei He, Nicola Nobile, Ching Y. Suen “Holistic Urdu Handwritten Word Recognition Using Support Vector Machine” 2010 International Conference on Pattern Recognition.

[9] Anil k. .jain and Torfinn Taxt “ Feature Extraction method for character recognition-A Survey ”Elsvier Science Pattern Recognition vol 29 no. 4 pp 641- 662 1996

[10] Anita Rani ,Rajneesh Rani ,Renu Dhir “Combination of Different Feature Sets and SVM Classifier for Handwritten Gurumukhi Numeral Recognition” International Journal of Computer Applications Volume 47– No.18, June 2012

[11] C. J. C. Burges, “A tutorial on support vector machines for pattern recognition”,Knowledge Discovery and Data Mining, Vol. 2(2), pp. 121-167, 1998

[12]Naveen Garg, Karun Verma, “Handwritten Gurmukhi Charcter Recognition Using Neural Network”, M.Tech. Theis, Thapar University, 2009

[13] Anuj Sharma, Rajesh Kumar, R. K. Sharma, "Online Handwritten Gurmukhi Character Recognition Using Elastic Matching",Conference onImage and Signal Processing (CISP), Vol.2, pp.391-396, May 2008

[14] Puneet Jhajj, D. Sharma, “Recognition of Isolated Handwritten Characters in Gurmukhi Script”, International Journal of Computer Applications, Vol. 4, No. 8, pp. 9-17, August 2010

[15] Ubeeka Jain, D. Sharma, “Recognition of Isolated Handwritten Characters of Gurumukhi Script using Neocognitron”, International Journal of Computer Applications, Vol. 4, No. 8, pp. 10-16, November 2010

[16] N. Sharma, U. Pal, F. Kimura, and S. Pal, “Recognition of offline handwritten Devnagari characters using quadratic classifier,” in Proc. Indian Conference Computer Vision Graph. Image Process, pp. 805–816, 2006.

[17] P. S. Deshpande, Latesh Malik, Sandhya Arora, “Recognition of Hand Written Devnagari Characters with Percentage Component Regular Expression Matching and Classification Tree”,IEEE, 2007.

[18] Sandhya Arora, D. Bhattacharjee, M. Nasipuri, D.K. Basu, M. Kundu, “Combining Multiple Feature Extraction Techniques for Handwritten Devanagari Character Recognition‖,Industrial and Information Systems, IEEE Region 10 Colloquium and the Third ICIIS, pp. 1-6, December, 2008.

[19] M. Hanmandlu, O.V. Ramana Murthy, Vamsi Krishna Madasu, ―Fuzzy Model based recognition of handwritten Hindi characters‖, Digital Image Computing Techniques and Applications, pp. 7695-3067-IEEE. Feb-2007.

[20] Mohamed Cheriet, Nawwaf Kharma, Cheng-Lin Liu, Ching Y. Suen, “Character Recognition Systems: A Guide for Students and Practioners”, Wiley Inter-Science, 2007

[21] Manimala Singlia and K.Hemacllandran “Performance analysis of Color Spaces III Image Retrieval” Assam University Journal of Science & Technology: Physical Sciences and Technology Vol. 7 Number II pp-94-104 ,2011

[22] Manesh Kok'are, B.N. Chatterji and **P.K.** Biswas “Comparison of Similarity Metrics for Texture Image Retrieval” IEEE pp 571-575, 2003