Gaurav Y. Tawde, Mrs. Jayashree M. Kundargi / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 3, Issue 1, January -February 2013, pp.919-926

An Overview of Feature Extraction Techniques in OCR for Indian Scripts Focused on Offline Handwriting

[*]Gaurav Y. Tawde , [**]Mrs. Jayashree M. Kundargi
Department of Electronics and Telecommunication Engineering,
K.J. Somaiya College of Engineering, Mumbai University, India

ABSTRACT

Optical Character Recognition (OCR) is an interesting and challenging field of research in pattern recognition, artificial intelligence and machine vision and is used in many real life applications. Optical character recognition is a type of document analysis where a scanned document image that contains either machine printed or handwritten script is input to an OCR software engine, is translated into editable, machine-readable digital text format. With the spread of computers in public and private sectors and individual homes, automatic processing of tabular application forms, bank cheques, tax forms, census forms and postal mails has gained importance. Such automation needs research and development of handwritten characters/numerals recognition for different languages or scripts. The field of OCR is divided into two parts, one is recognition of machine printed characters and the second is recognition of handwritten characters. Recognizing handwritten text is an important area of research because of its various application potentials. Feature extraction is very important step in the process of OCR. This manuscript gives a review of comparative study of different feature extraction techniques used in OCR.

Keywords - **classifier, feature extraction, optical character recognition, pattern recognition, pre-processing, multilayer perceptron, multi-resolution recognition of characters segmentation,**

I. INTRODUCTION

Handwriting recognition (or HWR) is the ability of a computer to receive and interpret intelligible handwritten input from sources such as paper documents, photographs, touch-screens and other devices. Depending on the manner in which data is acquired, the domain of handwritten character recognition is divided into two types.

- On-line Handwritten Recognition
- Off-line Handwritten Recognition

On-line handwriting recognition-On-line handwriting recognition involves the automatic conversion of text as it is written on a special digitizer or PDA, where a sensor picks up the pen-tip movements as well as pen-up/pen-down switching. That kind of data is known as digital ink and can be regarded as a dynamic representation of handwriting. The obtained signal is converted into letter codes that are usable within computer and text-processing applications.

The elements of an on-line handwriting recognition interface typically include:

- A pen or stylus for the user to write with.
- A touch sensitive surface, which may be integrated with, or adjacent to, an output display.
- A software application that interprets the movements of the stylus across the writing surface, translating the resulting strokes into digital text.

Off-line handwriting recognition-The image of the written text may be sensed "off line" from a piece of paper by optical scanning (optical character recognition) or intelligent word recognition. For offline character recognition the following cases can be considered: recognition of one or affixed number of fonts (Fixed Font OCR and Multifont OCR), any printed font (Omnifont OCR), isolated hand printed characters (Handwriting OCR) and unconstrained handwriting (Script Recognition) However, handwritten character recognition is a challenging task because of variability of writing styles of different writers from different environment. The task becomes more tedious when the text document quality is poor and if the characters are written very close to each other. In addition, some of the Indian scripts have compound characters. Some characters have similar shapes that require advanced and complex techniques for recognition.

Fig.1 Sample of handwritten numerals

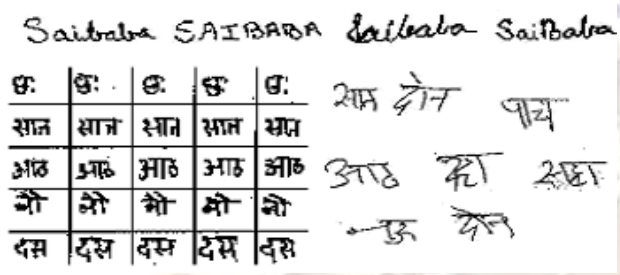

Fig.2 Sample of handwritten words in English and Devanagri script

1. Introduction

The paper is arranged to review the character recognition methodologies with respect to the stages of Character Recognition (CR) system and feature extraction methodologies. On-line and Offline character recognition techniques have different approaches, but they share many common problems and solutions. This article reviews some of the important feature extraction methods for OCR techniques. The paper is divided in four sections. Section 1 gives a brief introduction of the field of handwritten character recognition. Section 2 gives the description of the general process and operations performed by the OCR system. Section 3 gives an overview of different feature extraction techniques used for character recognition of different Indian scripts. Section 4 concludes the discussion.

2. Major Steps in OCR

The general process pattern recognition is shown in Fig.3.The input to the system is either printed or handwritten character or numeral. The input data goes through different stages in the process.

2.1 Data Acquisition

The input to the OCR system is the scanned document image. This input image should have specific format such as .jpeg, .bmp etc. This image is acquired through a scanner, digital camera or any other suitable digital input device. After image acquisition, the image data goes through following processes.

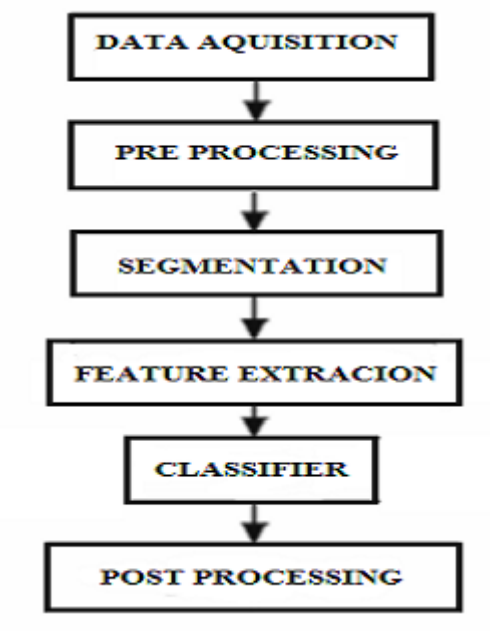

Fig.3 General Process of OCR

2.2 Pre-processing

The raw data is subjected to a number of preliminary processing steps to make it usable in the descriptive stages of character analysis. Pre-processing aims to produce data that are easy for the OCR systems to operate accurately. The main objectives of pre-processing are:

2.2.1 Binarization- Document image binarization (thresholding) refers to the conversion of a gray-scale image into a binary image.

Global binarization picks one threshold value for the entire document image which is often based on an estimation of the background level from the intensity histogram of the image.

Adaptive (local) binarazation, uses different values for each pixel according to the local area information.

Fig. 4 Example of binarization

2.2.2 Noise reduction (morphological operators)- Optical scanning devices may introduce noises, e.g. disconnected line segments, gaps in line, filled loops,

etc. Noise removal stage removes isolated specks and holes in the characters. Noise reduction improves the quality of the document. Two main approaches used are filtering and Morphological operations.

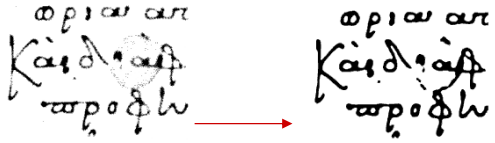

Fig.5 Example of Noise Reduction

2.2.3 Normalization-This stage removes some of the variations in the image that do not affect the identity of the input data and provides a tremendous reduction in data size. Thinning extracts the shape information of the characters.

Fig.6 Example of Normalization

2.2.4 Skew correction- Skew Correction methods are used to align the paper document with the coordinate system of the scanner. Main approaches for skew detection include correlation, projection profiles, Hough transform.

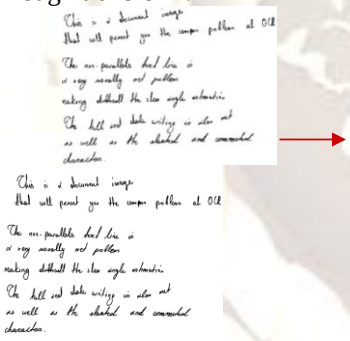

Fig.7 Example of skew correction

2.2.5 Slant removal-The slant of handwritten texts varies from user to user. One of the measurable factor of different handwriting styles is the slant angle between longest stroke in a word and a vertical direction. Slant removal methods are used to normalize the all characters to a standard form.

Fig.8 Example of slant removal

2.3 Segmentation

Segmentation is by far the most important aspect of the pre-processing stage. It allows the recognizer to extract features from each individual character. In the more complicated case of handwritten text, the segmentation problem becomes much more difficult as letters tend to be connected to each other, overlapped or distorted. Segmentation is done to break the single text line, single word and single character from the input document. For isolated characters or numerals, segmentation task is not that difficult. However, for joint and complex strings more advanced techniques required to be employed.
There are two types of segmentations:

1. External segmentation, that isolates various writing units such as paragraphs, sentences or words,
2. Internal segmentation, which is isolation of letters

Character segmentation strategies are divided into three categories:

Explicit Segmentation-This approach is used to identify the smallest possible word segments (primitive segments) that may be smaller than letters, but surely cannot be segmented further. It is done by the process of dissection.
Implicit Segmentation- This is based on recognition. It searches the image for components that match predefined classes.
Mixed Strategies- They combine explicit and implicit segmentation in a hybrid way.

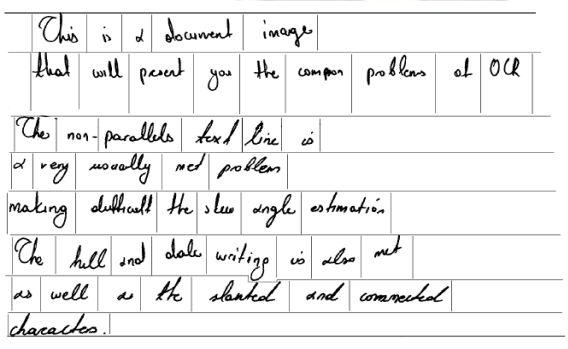

Fig.9 Example of segmentation

2.4 Feature Extraction

In feature extraction stage, each character is represented as a feature vector, which becomes its identity. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate with the least amount of elements and to generate similar feature set for variety of instances of the same symbol. Due to the nature of handwriting with its high degree of variability and imprecision obtaining these features, is a difficult task. Feature extraction methods analyze the input document image and select a set of features that uniquely identifies and

classifies the character. They are based on three types of features:

2.4.1 Statistical features- The following are the major statistical features used for character representation.

Zoning- Frame of character is divided into several overlapping and non-overlapping zones. The densities of the point or some features in different regions are analysed to form the representation. E.g. contour direction features measure the direction of the contour of the character. [9] that are generated by dividing the image into rectangular and diagonal zones and computing histograms of chain codes in these zones. Bending point features indicate high curvature points, terminal points and fork points as shown in Fig. 10

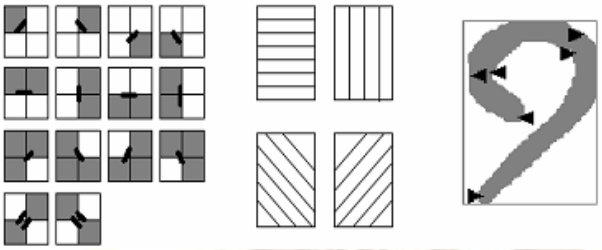

Fig. 10 Contour direction and bending features with zoning [9]

Projections and profiles- Character input data can be represented by projecting the pixel gray values onto lines in various directions giving one dimensional signal into two-dimensional image. The basic idea behind using projections is that character images, which are 2-D signals, can be represented as 1-D signal. These features, although independent to noise and deformation, depend on rotation. Projection histograms count the number of pixels in each column and row of a character image. Projection histograms can separate characters such as "m" and "n"

The profile counts the number of pixels (distance) between the bounding box of the character image and the edge of the character. The profiles describe well the external shapes of characters and allow distinguishing between a great number of letters, such as "p" and "q".

Crossings and distances-It refers to the number of crossings of a contour by a line segment in a specified direction. Distance of line segment from a given boundary can be used as one of the features. A horizontal threshold can be established above, below and through the centre of the script. The feature value is the count, the number of times the script crosses the threshold.

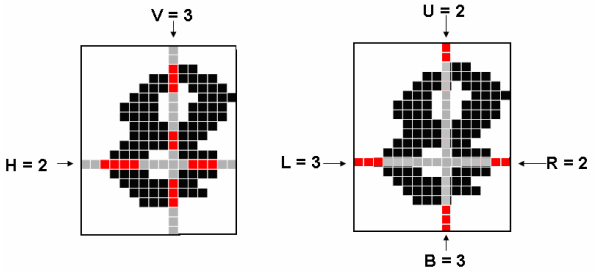

Fig. 11 Example showing Crossing and Distance

2.4.2 Structural features- These are based on topological and geometrical properties of the character, such as aspect ratio, cross points, loops, branch points, strokes and their directions, inflection between two points, horizontal curves at top or bottom, etc

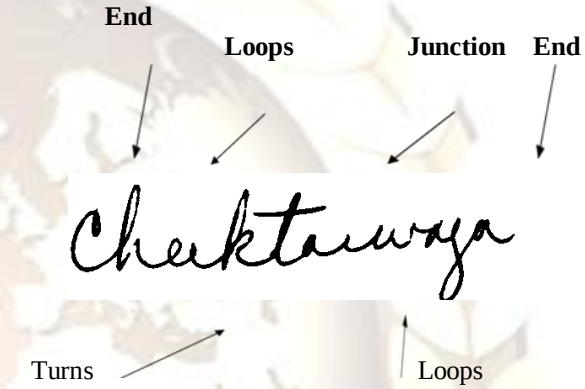

Fig. 12 Example of Structural Features

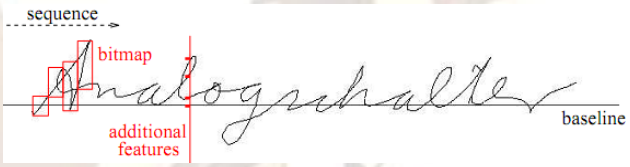

Fig. 13 Example of feature extraction in offline recognition

2.4.3 Global transformations and moments- A continuous signal contain more information that can be represented for the purpose of classification. The Fourier Transform (FT) of the contour of the image is calculated. Since the first *n* coefficients of the FT can be used in order to reconstruct the contour, then these *n* coefficients are considered to be a *n*-dimensional feature vector that represents the character.

Gabor transform, which is the variation of windowed Fourier transform, is another approach.

Wavelet transform is a tool that cuts up data, functions or operator into different frequency component, and studies each component with resolution matched to its scale. The segments of document image can be represented as wavelet coefficients with multiple resolutions

2.5 Classification

Feature extraction stage gives us the feature vector that is used for classification. Classification is the decision making step in the OCR system that makes use of the features extracted from the previous stage in the process. To do the classification we must have a data bank to compare with many feature vectors. A classifier is needed to compare the feature vector of input and the feature vector of data bank. The selection of classifier depends upon training set and number of free parameters. There are many existing classical and soft computing techniques for handwritten recognition.

2.6 Post-processing

The purpose of this step is the incorporation of context and shape information in all the stages of OCR systems is necessary for meaningful improvements in recognition rates. A dictionary can be used to correct minor errors.

3. Approaches to Feature Extraction Methods

Feature Extraction and Classification techniques are very important steps in character recognition process to achieve high recognition performance [1]. A number of feature extraction techniques are available in the literature of pattern recognition. Out of these, which feature extraction method will best suit for the application, is the key issue. Feature extraction is defined as the problem of "extracting from the raw data the information which is most relevant for classification purposes, in the sense of minimizing the within-class pattern variability while enhancing the between-class pattern variability".[2] Different feature extraction methods can fulfil this requirement based on specific recognition problem and the input data for recognition. Also different feature extraction techniques are well suited for different domain. A detail survey of various feature extraction techniques is given in [1] [3]

A novel approach to feature extraction based on fractal theory is presented. [4] A new feature that can be applied to extract the feature of two-dimensional objects. It is constructed by a hybrid feature extraction combining wavelet analysis, central projection transformation and fractal theory. A multiresolution family of the wavelet is used to compute information conserving micro-features. A central projection method is used to reduce the dimensionality of the original input pattern. A wavelet transformation technique is employed to transform the derived pattern into a set of sub-patterns. Its fractal dimension can be computed and used as feature vector. The important characteristic of fractals, the fractal dimension, contains information about their geometrical structure. This concept is applied to feature extraction technique.

Gabor filter possess optimal localization properties in both spatial and frequency domain. Gabor filter gives a chance for multi-resolution analysis by giving coefficient matrices [5]. In this approach, a 2D Gabor filter gives a extracted features. A Gabor is Gaussian modulated sinusoid in the spatial domain and a shifted as a shifted Gaussian in frequency domain. It can be represented by:

$$g_{\gamma,\eta,\varphi,\lambda} = exp\left(\frac{x'^2+\gamma 2y'^2}{2\sigma^2}\right).\cos\left(\frac{2\pi x'}{\lambda}+\varphi\right) \quad (1)$$

$$x' = x\cos\theta - y\sin\theta$$

$$y' = x\sin\theta - y\cos\theta \quad (2)$$

$$\iint I(\varepsilon,\eta)g(x-\varepsilon,y-\eta)d\varepsilon d\eta \quad (3)$$

The Gabor filter can be better used by varying the parameters like λ, γ, φ and θ. In equation (2), x and y are image coordinates. λ is the wavelength of cosine equation, γ characterizes the shape of Gaussian,
γ= 1 for circular shape, γ<1 for elliptical shape.
θ represents the channel orientation and takes values in the
interval (0,360). The response of Gabor filter is convolution given by equation (3).

Optical Character Recognition system for English and Tamil
script is presented. [6] Gabor filter is used for feature extraction and SVM for classification. After segmentation, the image is passed through a bank of 24 Gabor filters and for each image 50 features are obtained. The average accuracy of recognition for English is 97% and for Tamil it is 84% is achieved.
In [7] Mahesh Jangid and Kartar singh Siddhart have proposed statistical features like zonal density, projection histogram, distance profiles, Background Directional Distribution(BDD) for handwritten Gurumukhi isolated character recognition. With SVM classifier they achieved 95.04% accuracy.
Munish Kumar presents Offline handwritten Gurumukhi character recognition based on k-NN classifier.[8] used diagonal features and transitional features with k-nn classifier for the same script. Diagonal features are used to give high recognition accuracy and reduction in misclassification. Feature extraction is done from the pixels of each zone by moving along its diagonals as shown in Fig. 14.

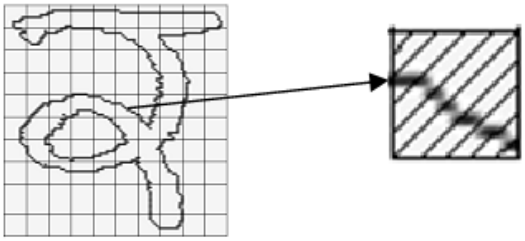

Fig. 14 Diagonal Feature Extraction [8]

Another feature extraction approach is based on calculation and location of transition features from background to foreground pixels in the vertical and horizontal directions is also discussed. The proposed system achieves accuracy of 94.12%which can be further improved by adding more features.

An overview of Feature Extraction techniques for off-line recognition of isolated Gurumukhi characters recognition is presented in [13]. In this paper Zone based approach is discussed. The centroid of image (numeral/character) is calculated. This image is further divided into 100×100 equal zones where size of each zone is (10×10). The average distance from image centroid to each pixel present in the zones/block is found. They got 100 feature vector of each image. For Zone Centroid Zone (ZCZ) approach, they divided image into n equal zones and calculated centroid of each zones. Then the average distance from the zone centroid to each pixel present in zones is computed. For empty zone the value of that particular zone is assumed to be zero. The procedure is repeated for all zones present in image(numeral/character) With SVM and K-NN classifiers respectively, the accuracy of 95.11% and 90.64%. is obtained

A recognition based OCR system is proposed in [9]. After acquisition, preprocessing, segmentation, structural features are extracted. Hybrid edge detector is used to extract contour of a character. Character fragment is traced from top right hand black pixel through its whole contour based on 2x2 window. By this feature extraction process a particular sequence of freeman code is produced. This code chain is concentrated by dividing the run-length of a code with a threshold. Experimental results on printed text yield 90% of accuracy.
Freeman Chain Code is a means to represent lines or boundaries of shapes by a connected sequence of straight line. This chain code is generated by using changing direction of connected pixels contained in a boundary. The representation based on 4-connectivity or 8-connectivity of the segments. The character boundary of the connected components is established to calculate chain code. Freeman chain code is based on observation that each pixel has four or eight neighbor pixels.

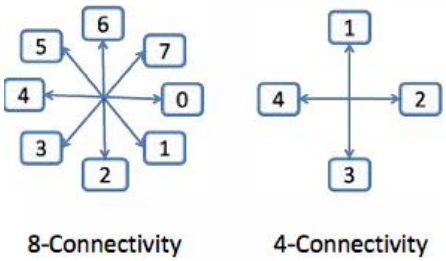

Fig. 15 Slope convention of Freeman Chain Code

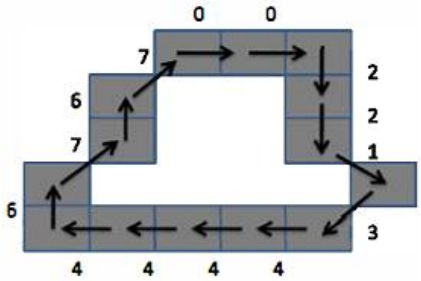

Fig. 16 Example of 8-directional Chain code

Optical Bangla character recognition system using Freeman Chain Code and feed forward back propagation neural network is presented. [10] The recognition rate of about 99.953% with 0.0073 training error is obtained.

Free style handwriting recognition is difficult task because of large variations in writing styles and shapes and due to overlapping and interconnection of neighboring characters.[11] From the investigation of psychology of reading, the words are identified directly from their global shapes. The shapes of letter are used to reduce the complex nature of contour. Polygon box fitting concept is used according to which five variety of boxes has been used that is closed box, lower open, left open, down open and a line characters with no boxes. This novel box based method showed better results compared to statictical based approaches.

S. V. Rajashekararadhya et.al.[12] have used the zone based feature extraction method on handwritten numeral/mixed numerals recognition of south-indian scripts. Feed forward back propagation neural network, and Support Vector Machine for classifier. They have obtained overall 98.9% of accuracy.

In paper [14] Rajbala Tokas, Aruna Bhadu presented different feature extraction methods to classify the 26 handwritten capital alphabets written by 25 different writers with their advantage& disadvantage &comparison to each other. They came out with result shown in Table 1

In 2007 M. Hanmandlu et.al [15] using Input Fuzzy Modeling for the Recognition of Handwritten Hindi Numerals. This paper presents the recognition of Handwritten Hindi Numerals based on the modified exponential membership function fitted to the fuzzy sets derived from normalized distance features obtained using the Box approach method . They have obtained 95% recognition accuracy.

In 2012 Shailendra Kumar et al[16] used moment invariant and Affine moment invariant techniques for feature extraction for handwritten Devanagari numerals. Using SVM classifier with the extracted features, 99.48% overall recognition rate is obtained.

A novel CR technique based on the fuzzy descriptive feature using cross correlation for character classification is proposed. [17] A maximized fuzzy descriptive feature is for each pixel is obtained by following function:

$$S_{ij=}\ \max_{x=1} N1\ \left(\max_{y=1} N2\,(w[i-x, j-y] f_{xy})\right) \quad (4)$$

For *i= 1 to N1*, and
j= 1 to N2

S_{ij} gives maximum fuzzy membership pixel value using the fuzzy function *w[m,n]*, equation (4) where f_{xy} is the (x,y) binary pixel value of an input pattern ($0 \leq f_{xy} \leq 1$).

N1 and N2 are the height and width of the character window.

$$w[m,n] = \exp\left(-\beta^2(m^2+n^2)\right) \quad (5)$$
$$for \quad m — (N_1-1) to (N_1-1)$$
$$n — (N_2-1) to (N_2-1)$$

The recognition rate obtained using limited testing set is 100% for skewing less than 1degree during scanning.

In [18], Curvelet Transform and Character Geometry methods are used for feature extraction. Performance with SVM, RBF and *k*-NN classifier is studied resulting accuracy with Curvelet features with *k*-NN classifier is 93.8%.

Table 1 Comparison of different feature extraction methods[14]

S.no	Method	Merits	Demerits
1	Binary	Easy to implement and good for low resolution images	Training is slow for large size images, Redundant values in feature space
2	13-region	Feature space is Small	Feature space is Small
3	15-region	Information loss in Feature Space	Attributes of different classes have little Difference
4	16-region	Good accuracy	Zoning need Complete Understanding of character
5	25-region	less information Lost	Redundancy in feature space
6	Diagonal	This gives more exact information	This is more specific to characters build by strainght lines
7	Vertical	More simple than Diagonal	Range of values is Large
8	Centroid	It can be merged with others to give good accuracy	It work good for only characters having maximum Curves
9	Direction	Zoning does not Required	Feature values are Redundant
10	Topology	More accurate Features	Large feature Space
11	Cross-Corner	Substantial increase in Accuracy	Feature space is Large
12	Distance & Crossings	Accuracy is Increased	Features are not related to each other
13	Hybrid	Testing is confirmed by Boosting	Selection of features is critical

4. CONCLUSION

Feature extraction is the most crucial & important part of handwritten character recognition. This paper reviews some of the important feature extraction techniques employed for different Indian handwritten scripts. Various feature extraction techniques based on zone, centroid, moment invarients, shadows, chain code, fuzzy descriptive

features for different Indian scripts like Devanagari, Hindi, Bangla, Tamil, Gurumukhi, English are reviewed. Feature extraction method that is best suited for one particular recognition application may not give optimum performance for the other application. Therefore, selection of proper feature extraction technique and classifier or multiple classifiers is the key issue in OCR system.

REFERENCES

[1] Ivind Due Trier, Anil K. Jain and Torfinn Ta, *Feature Extraction Methods For Character Recognition- A Survey, Pattern Recognition, Vol.39, No.4* pp. 641- 662, 1996

[2] P.A. Devijver and J. Kittler, *Pattern Recognition: a Statistical Approach* (London: Prentice-Hall, 1982)

[3] Govindan, V.K. Shivaprasad, A.P. Character Recognition-A Review, *Pattern Recognition. Vol. 23 N0. 7*, pp 671-683, 1990.

[4] Yuan Y. Tang, Yu Tao, Ernest C. M. Lam, New Method for Feature Extraction Based on Fractal Behavior, *Pattern Recognition 35* (2002) 1071-1081

[5] R. Ramnathan, L. Thaneswaran, V. Vinkesh, T. Arunumar, P. Yuvaraj, K.P. Soman, A Novel Technique for English Font Recognition Using Support Vector Machines, *International Conference on Advances in Recent Technologies in Communication and Computing ARTCom 2009*, India, 27-28 Oct 2009

[6] R. Ramnathan, S. Ponmathavan, L. Thaneswaran, N. Valliappan, Arun S. Nair.Dr. K.P. Soman ,*Optical Character Recognition for English and Tamil Using Support Vector Machines, International Conference on Advances in Computing, Control and Telecommunication Technologies*, 2009

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

[8] Munish Kumar, M. k. Jindal, R.K. Sharma, k-nearest Neighbor based Offline Handwritten Gurumukhi Character Recognition, *International Conference on Complex Image Information processing* 2011(ICIIP)

[9] K. M. Mohiuddin, J. Mao, A Comparative Study of Different Classifers For Handprint Character Recognition, *Pattern Recognition in Practice IV*, pp. 437-448.1994

[10] Nawrin Binte Nawab, M. N. Hassan, Optical Bangla Character Recognition using Chain Code, IEEE/OSA/IAPR *International Conference on Informatics, Electronics & Vision*, 2012

[11] C. Namrta Mahender, K.V. Kale, Structured Based Feature Extraction of Handwritten Marathi Word, *International Journal of Compuetr Applications(0975- 8887), Volume 16-No. 6*, February 2011

[12] V. Rajashekararadhya, P. Vanaja ranjan, *Handwritten Numeral/Mixed Numeral Recognition of South Indian: Zone Based Feature Extraction Method*. 2005-2009 JATIT

[13] Gita Sinha, Mrs. Rajneesh Rani, Prof. Renu Dhir, Offline Handwritten Gurumukhi Character Recognition using k- Nearest Neighbor and SVM Classifier, *International Journal of Advanced Research in Computer Science and Software Engineering,Vol.2, Issue 6*, June 2012

[14] Rajbala Tokas, Aruna Bhadu, A Comparative Analysis Of Feature Extraction Techniques for Handwritten Character Recognition, *International Journal of Advanced Technology & Engineering Research (IJATER), Volume 2, Issue 4*, July 2012

[15] M. Hanmandlu, J. Grover, V. K. Madasu, S. Vasikarla, Input Fuzzy Modelling for the Recognition of Handwritten Hindi Numerals, *International Conference on Information Technology (ITNG,07) 0-7695-27760/07, 2007 IEEE*

[16] Shailendra Kumar Sharivastava, Sanjay S. Gharde, Suuport Vector Machine for Handwritten Devanagari Numeral Recognition, *International Journal of Computer Application (0975-8887), vol. 7, Issue11*, October 2010

[17] Y. Alginahi, I. El-Fei, M. Ahmadi and M. A. Sid-Ahmed, Optical Character Recognition System Based On A Novel Fuzzy Descriptive Features, *ICSP'04 Proceedings, 0-7803-8406-7/04,IEEE*, 2004

[18] Brijmohan Singh, Ankush Mittal, Debashush Ghosh, An Evaluation of Feature Extractors and Classifiers for Offline Handwritten Dvanagari Character Recognition, *Journal of Pattern Reconition Research* 2(2011) 269-277