A survey on feature extraction and classification techniques for character recognition of Indian scripts

Sukhpreet Kaur[1], Simpel Rani[2]
[1]*M. Tech. Research Scholar in Computer Science & Engineering, Yadwindra College of Engineering, Talwandi Sabo, Bathinda, Punjab, India*
[1]sukhpreetkaur766@gmail.com
[2]*Associate Professor, Computer Science & Engineering, Yadwvindra College of Engineering, Talwandi Sabo, Bathinda, Punjab, India*
[2]simpel_jindal@rediffmail.com

Abstract
Much research has been done by many researchers on Optical Character Recognition system. But most of the work done is on Greek, Chinese, English and Japanese characters. There has not been adequate work on character recognition of Indian languages like Bangla, Marathi, Malayalam, Telugu, Gujarati, Kannada, Gurmukhi and Oriya. The development of handwritten character recognition (HCR) is an interesting area in pattern recognition. In HCR, the set of features are very important in selecting the appropriate feature that produces little classification error. In this paper, we have presented a survey on feature extraction and classification techniques used for character recognition of Indian scripts.

Keywords
Handwritten Character Recognition, OCR, Indian Languages, OCR Review

1. INTRODUCTION
Handwritten Character Recognition (HCR) is the capability of a computer to acquire and translate explicit handwritten input through many automated process system. HCR can be isolated into three steps namely pre-processing, feature extraction and classification (recognition). HCR is the process of changing scanned images of handwritten text into computer processing text such as ASCII code. It is generally used to improve the speed of operations, reduced errors or noise in the documents and decrease storage space needed for papers documents. It is a simple method for fast retrieval, easily searched, saved more compressed data. It is an active field of research in pattern recognition and image processing system.

In character recognition system, feature derivation is an important job. Its main task is obtaining particular information from character in order to minimize variations within class pattern. HCR is a challenging issue because there is a divergence of identical character due to the change of writing styles. The variance in writing styles make the recognition task difficult and output of the recognition of character process becomes not good. HCR has many applications in mail sorting, bank processing, document reading and postal code recognition. So, off-line handwriting recognition is a challenging research area towards exploring the newer techniques that would improve recognition accuracy.

Feature extraction stage is used to remove redundancy from data. Feature extraction methods for character recognition are based on three types of features: a) statistical features b) structural and c) transformation based features. The most statistical features have been used

for character representation are: a) zoning, where the image is divided into several zones [11-14], b) projections and c) crossings and distances.

1.1 HCR Phases

1.1.1 Image acquisition

Image acquisition is an initial phase of character recognition system. In this phase, input image scanned is scanned and is converted into electronic form in bitmap images. After acquisition, the acquired image is fed to pre-processing phase.

1.1.2 Pre-processing

Pre-processing is next phase of text recognition system. It includes noise removal, skew detection/correction and skeltonization. Pre-processing of document is required to detect and remove all unwanted bit pattern which lead to reduce the recognition accuracy. After pre-processing of text, features have been extracted using various feature extraction techniques for recognition purpose.

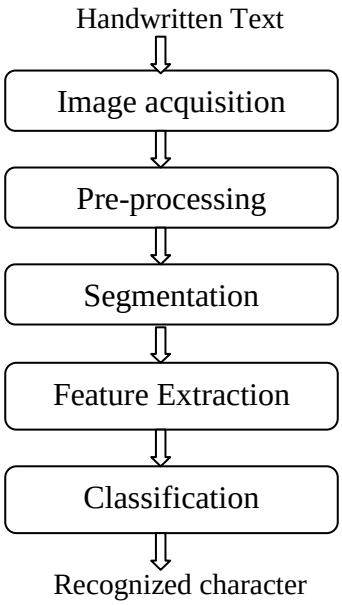

Figure 1: Handwritten Character Recognition Phases

1.1.3 Segmentation

In segmentation phase, an image of continues characters is break-down into sub-images of individual character. Segmentation is done in the following types: Line segmentation, Word segmentation and Character segmentation. Segmentation is an integral part of HCR system to find boundaries of character, word and line segmentation.

1.1.4 Feature Extraction and Classification

Feature extraction technique is the most important part of recognition system. Feature extraction phase is used to remove redundancy from data. There are various feature extraction techniques like statistical and structural features. Classification phase is a decision making part of a recognition system and features extracted in the previous phase are used to identify characters.

2. HCR WORK ON INDIAN LANGUAGES

Many researchers have proposed several techniques for handwritten as well as printed character and numerals recognition. There are 23 official languages in India namely Assamese, Bengali, Bodo, Dogri, English, Gujarati, Hindi, Nepali, Oriya, Punjabi, Sanskrit, Santhali, Sindhi, Tamil, Telugu, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Urdu and Marathi. For writing these languages, 14 scripts have been used like Assamese, Bangla, Devanagari, Oriya, Roman, Tamil, Telugu, Urdu, Gujarati, Gurmukhi, Kannada, Kashmiri and Malayalam.

2.1 Studies on Devanagari character recognition

Agnihotri et *al.* [7] have presented diagonal based features extraction technique for Devanagari script recognition. The features have been extracted from each zone by moving along their diagonals. That features of character have been changed to chromosome bit string of length has 378. For classification or recognize the characters, Genetic algorithm is used.
Aggarwal *et al.* [9] have presented isolated handwritten Devanagari character recognition using Gradient features extraction technique. They have collected 200 samples of 36 Devanagari character from 20 different writers writing 10 samples of each 36 characters. All samples of Devanagari character are normalized to 90*90 pixel size. For classification Support Vector Machine with RBF kernel is used and achieved better accuracy of 94%.

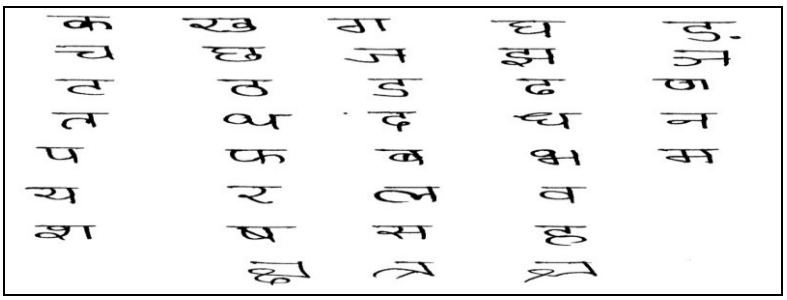

Figure 2: Handwritten Devanagari character

2.2 Studies on Malayalam character recognition

John *et al.* [1] have described Wavelet transform tool for **Malayalam** character recognition. The line segmentation is done using horizontal projection and isolated character with connected component algorithm. Haar Wavelets is used for multi-resolution feature extraction. Haar wavelet is also known as a compact orthonormal wavelet transform. SVM classifier with RBF kernel is used for classification and liner separable.

George *et al.* [4] have described the system including image acquisition, preprocessing, segmentation, feature extraction, classification & recognition stages. In first stage, original image is taken through scanner as input. In preprocessing stage noise reduction, smoothing, binarization and edge detection is done. In segmentation complete image of character is divided into sub-images of character. For feature extraction, contourlet transform method is used with ratio of grid values in horizontal and vertical direction. A feed forward back propagation neural network is used with three hidden layers for classification. The system is achieved accuracy of 97.3%.

2.3 Studies on Kannada character recognition

Niranjan *et al*. [2] have presented FLD based unconstrained Kannada character recognition. Fisher Linear Discriminate Analysis is used for feature extraction with 2DFLD and Diagonal FLD based methods. For classification different distance measures techniques have been used such as Euclidean, Squared Euclidean, Mean Square Error, Angle, Correlation co-efficient and Manhattan. The recognition of 2DFLD with angle & correlation is achieved better result than other methods. When combined that result with other modified characters an accuracy of 68% has been reported.

2.4 Studies on Telugu character recognition

Singh and Kaur [3] have proposed adaptive sampling algorithm, normalization, image binarization and image thinning (skeletonization) for preprocessing. They have represented each character as a feature vector in the feature extraction stage. The various general features have been extracted such as character width, height, closed shapes, diagonal lines, line intersections, special dots. For the classification, Back Propagation algorithm is used.

2.5 Studies on Marathi character recognition

Patil *et al.* [10] have proposed Moment invariant technique, image thinning, affine invariant moments and image in box format for features extraction. These features are independent in size; slant and orientation have been used for compare feature extraction methods. The Fuzzy Gaussian membership function is used as classifier. The mean and standard deviation is computed for each type of feature for find out maximum membership value.

2.6 Studies on Bangali numeral recognition

Rahman *et al*. [6] have presented canny method for edge detection. The method finds edges by local maxima of the gradient of binary image. The gradient is computed using the derivative of a Gaussian filter. Kirsch mask and the PCA method is used for dimension reduction. Also recognition time and training time can be reduced. Kirsch edge detector is used for detect directional feature vectors for horizontal, vertical, right-diagonal and left diagonal directions. The output of the PCA is passing to Support Vector Machine (SVM) to determine in which class the input belongs to.

Bhunia *et al.* [5] have studied of Bangla text. Zone segmentation is performed in middle zone and features like PHOG, GABOR, LGH, G-PHOG, profile feature have been extracted. This system is achieved high recognition rate 85.74% when touching of middle zone character & lower zone modifier occur at a single place.

2.7 Studies on Gujarati character and numeral recognition

Desai [11] has proposed different preprocessing methods for digits like skewness, contrast correction, resizing, thinning before the classification of digits. The author is used four different profiles i.e. vertical, horizontal and two diagonals for the feature extraction. The vector of these profiles is used for research of a digit. Through feed forward back propagation neural network (BPN) 81.66% accuracy is achieved.

2.8 Studies on Hindi character recognition

Yadav *et al.* [8] have presented preprocessing step for normalizing the input image in which noise removal, skew detection, slant identification process is performed. After that

segmentation is performed to separate the touching characters in text in three forms- line, word and character segmentation. Three types of features have been extracted- projection based histogram on mean distance, projection based histogram on pixel value and vertical zero crossing. ANN classifier is used for classification and 90% accuracy is achieved.

2.9 Studies on English alphabet and numeral recognition

Rachana *et al*. [12] have proposed zoning feature extraction technique for isolated English alphabets and numerals in different zones like upper zone, middle zone and lower zone. The required search space is reduced because character set is divided into three parts. Euler number is used with zoning method for character recognition. Accuracy for uppercase letters is 91.15%, for middle case letters is 90.57% and for lowercase letters is 90% achieved. **Gatos** *et al.* [14] have presented character and word recognition using two adaptive zoning features method. In first, features have been calculated by density of pixels. In second, features have been calculated by characteristics in each zone. For experiment, they have used CIL database with 28750 Greek characters. Euclidean distance between two feature vectors with a minimum distance is used for classification.

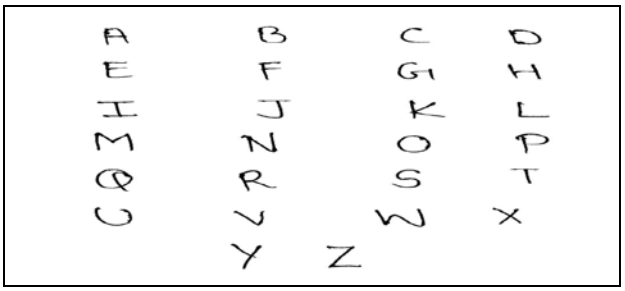

Figure 3: Handwritten English Alphabets

2.10 Studies on Gurmukhi character recognition

Aggarwal *et al.* [15] have presented Gradient Technique for feature extraction of Gurmukhi character and numerals. Gradient technique is applied by two methods on characters. First is decomposition of gradient vector into directions and second is non-decomposition of gradient vector. Though first method, they have achieved high efficiency and accuracy as compared to second method. They have using database of 7000 samples for Gurmukhi character and 2000 samples for numerals for testing methods. The result of Gurmukhi character recognition is obtained 97.38% and 99.65% for Gurmukhi numerals recognition. Singh *et al*. [13] have recognized Gurmukhi character using two Gabor filter features extraction methods namely Gabor- GABM and Gabor-GABN. SVM classifier is used for recognition of character. With the help of Gabor-GABN features extraction technique, they have achieved high recognition accuracy 94.29% with dimensionality 200 as compared to Gabor- GABM.

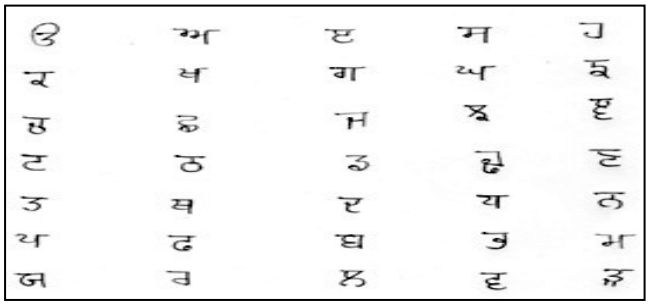

Figure 4: Handwritten Gurmukhi characters

TABLE 1: Study of various character recognition systems for Indian Languages

Author	**Feature Extraction**	**Languages**	**Classification and Accuracy**	**Advantage**	**Disadvantage**
John et al. [1]	Haar wavelet features at different Resolution scales	Handwritten Malayalam character	Support vector machine with RBF kernel and accuracy is obtained 90.25% on level 2.	In this 40-class classification problem is solved on second level decomposition.	Haar wavelet coefficients at level 3 is obtained classification only 89.64%
Niranjan et al. [2]	FLD, 2D-FLD and Diagonal FLD	Unconstrained Handwritten Kannada character	Different distance measure techniques have been using i.e. Murkowski, Angle, Manhattan, Euclidean, Mean Square Error, etc. with 68% Accuracy.	It is solving Eigen value problem and 2D-FLD with Angle & Correlation performs better recognition for consonants and vowels.	2D-FLD has not been achieving better accuracy with modifiers as compared to consonants and vowels.
Singh and Kaur [3]	It extracts the general features such as width, height, closed shapes, diagonal lines, line intersections, special dots.	Printed Telugu Script	BPN, ANN	BPN algorithm is overcome the drawback of Neural N/W algorithm, where single layer perceptron fail to solve a simple XOR problem.	It has not been working on handwritten Telugu script.

George, et al. [4]	Contourlet transform with a ratio of grid value in horizontal & vertical directions	Handwritten Malayalam character	Feed forward artificial neural network classifier is used and 97.3% accuracy is achieved	It is achieved superior recognition rate so it is suitable for handwritten name recognition & conversion of handwritten text into structural text form.	Because of the curved nature and no inherent symmetry of Malayalam characters, its feature extraction is difficult.
Bhunia et al. [5]	LGH, PHOG, GABOR, Profile feature, GPHOG	Handwritten Bangla text	SVM classifier is used and 85.74% accuracy is achieved.	This system is achieving high recognition rate when touching of middle zone characters & lower zone modifiers at a single place.	The system is fail due to unavailability of Mantra for e.g. highest peak detects at wrong place. System has not been working when touching of middle zone characters & lower zone modifiers at two different positions.
Rahman et al. [6]	PCA and four directional local feature vector for edge detection has kirsch mask	Handwritten Bengali Numerals	SVM classifier is used and recognition rate is 92.5%, Error rate is 7.5% and reliability rate is 92.5% achieved.	This approach is used in Postal System to get post code and it is extracted more features than method normalized image to SVM.	It is needed more training time for trained system. So, time performance has not been better.
Agnihotri et al. [7]	Diagonal based feature extraction	Handwritten Devnagari Script	Recognition rate is achieved 97% for 54 features by Genetic algorithm.	The precision of offline Devnagari system is 85.78% match, 13.35% is mismatch.	It has not used mutation technique in genetic algorithm for well recognize characters.
Yadav et al. [8]	Histogram of projection based on mean distance, histogram of	HandwrittenHindi text	Artificial Neural Network (ANN) classifier is used and 90% accuracy is achieved.	It is achieved high accuracy in handwritten text recognition.	This is developing an approach which has not been deal with punctuation marks and numerals.

	projection based on pixel value, and vertical zero crossing				
Aggarwal et al. [9]	Gradient Features	Handwritten Devnagari Character	SVM classifier is used and 94% accuracy is achieved	It measures direction and value of change intensity in small neighborhood of pixel.	It has not been worked on vowels and modifiers only worked on consonants
Patil et al. [10]	Moment Invariants (MIs), Affine moments Invariants (AMIs), image thinning	Handwritten Marathi Vowels	Fuzzy Gaussian membership function is used and 75% accuracy is achieved for MI, 89.09% for AMI and 52.90% for combination of MIs & AMIs.	A compound feature extraction approach based on structural analysis is achieved better performance.	They have not worked on Marathi consonants as well as has not been applied any post processing step
Desai [11]	Four different profiles horizontal, vertical and two diagonals	Gujarati Numerals	Through Feed forward back propagation neural network 81.66% of accuracy is achieved	It is achieved high success rate in recognition of zero, four and seven digits.	The success rate of this network is low because mis-identification is creating due to confusing digits
Rachana et al. [12]	Zoning	Handwritten English alphabets and numerals	Euler Number with end points is used & accuracy is achieved for uppercase alphabets 91.15%, for lowercase 90.57% and for digits 91%	It increases accuracy and speed of recognition and the search space can be reduced by dividing character set into three parts.	When different writers have been considered then different accuracy is achieved
Singh et al.[13]	Gabor-GABM and Gabor-GABN features	Handwritten Gurmukhi Character	SVM classifier with RBF kernel is used for classification and 94.29% accuracy is achieved.	Gabor features have less sensitive to noise, small range of scaling.	When kernel parameter value of SVM increases from 0.01 to 2, then accuracy decrease.
Gatos et al. [14]	Adaptive Zoning Features-pixel density and pattern	English and Greek character and word recognition	Euclidean distance is using between two feature vectors with a minimum distance classifier	It has better accuracy in adaptive zones as compared to standard zone.	It takes more processing time to recognize.

	characteristic in zone		& CR is improved from 85.98% to 88.35% & WR from 85.1% to 89.12% for Pixel density features and CR is improve from 66.80% to 78.76% & WR from 76.02% to 81.64% for pattern characteristics in each zone.		
Aggarwal et al. [15]	Gradient features	Gurmukhi character and numerals	SVM is used & 97.38% accuracy for Gurmukhi character & 99.65% for numerals is achieved	It efficiently recognize the character through decompose and non-decompose gradient vector.	It has not been worked on recognition of characters in word.

3. CONCLUSION AND FUTURE SCOPE

India is a nation in which various languages have been used as way of message passing between different people. In this paper, we have presented a survey on feature extraction and classification techniques for character recognition of Indian scripts. We have discussed various steps used for OCR character recognition and studies of different work is done on Indian languages. Also advantages and disadvantages of each method used has been discussed.

REFERENCES

1) John, J., Pramod, K. V. and Balakrishnan, K., 2011. Unconstrained Handwritten Malayalam Character Recognition using Wavelet Transform and Support vector Machine classifier, *International Conference on Communication Technology and System Design*, pp. 598-605.
2) Niranjan S.K., Kumar, V., Kumar G. and Aradhya V. N., 2009. FLD based Unconstrained Handwritten Kannada Character Recognition, *International Journal of Database Theory and Application*, Vol. 2(3), pp. 21-26.
3) Singh, R. and Kaur, M., 2010. OCR for Telugu Script Using Back-Propagation Based Classifier, *International Journal of Information Technology and Knowledge Management*, Vol. 2(2), pp.639-643.
4) George, A. and Gafoor, F., 2014. Contourlet Transform Based Feature Extraction for Handwritten Malayalam Character Recognition Using Neural Network, *3rd IRF International Conference Chennai*, pp. 107-110.

5) Bhunia, A. K., Das, A., Roy, P. P. and Pal, U., 2015. A Comparative Study of Features for Handwritten Bangla Text Recognition, *13th International Conference on Document Analysis and Recognition (ICDAR)*, pp.636-640.
6) Rahman, S., Atiqur Rahaman, A. K., Ahmed, A. and Salahuddin, G. M., 2008. An Approach to Recognize Handwritten Bengali Numerals for Postal Automation, *11th International Conference on Computer and Information Technology (ICCIT)*, pp.171-176.
7) Agnihotri, V., 2012. Offline Handwritten Devnagari Script Recognition, *International Journal Information Technology and Computer Science*, Vol. 8, pp.37-42.
8) Yadav, D., Sanchez-Cuadrado, S. and Morato, J., 2013. Optical Character Recognition for Hindi Language Using a Neural-network Approach, *Journal Info. Process System*, Vol. 9(1), pp.117-140.
9) Aggarwal, A., Rani, R. and Dhir, R., 2012. Handwritten Devnagari Character Recognition Using Gradient Features, *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 2(5), pp.85-90.
10) Patil, N. P., Adhiya Surendra, K. P. and Ramteke, P., 2011. A Structured Analytical Approach to Handwritten Marathi vowels Recognition, *International Journal of Computer Applications*, Vol. 31(3), pp.48-52.
11) Desai, A. A., 2010. Gujarati handwritten numeral optical character reorganization through neural network, *Pattern Recognition*, pp.2582–2589.
12) Rachana, R. H. and Dhotre, S. R., 2014. Handwritten Character Recognition Based on Zoning Using Euler Number for English Alphabets and Numerals, *IOSR Journal of Computer Engineering*, Vol.16(4), pp.75-88.
13) Singh, S., Aggarwal, A. and Dhir, R., 2012. Use of Gabor Filters for Recognition of Handwritten Gurmukhi Character, *IJARCSSE*, Vol.2 (5), pp. 234-240.
14) Gatos, B., Kesidis, A. L. and Papandreou, A., 2011. Adaptive Zoning Features for Character and Word Recognition, *International Conference on Document Analysis and Recognition*, pp.1160-1164.
15) Aggarwal, A., Singh, K. and Singh, K., 2014. Use of Gradient Technique for extracting features from Handwritten Gurmukhi Characters and Numerals, *International Conference on Information and Communication Technologies*, pp.1716-1723.