A Survey on Handwritten Character Recognition Techniques for Various Indian Languages

Krupa Dholakia
Computer Science Department
Shri Jagdish Prasad Jhabarmal Tibrewala University, India

ABSTRACT

Handwritten character recognition is always an interesting area of pattern recognition for research in the field of image processing. Many researchers have presented their work in this area and still research is undergoing to achieve high accuracy. This paper is mainly concerned for the people who are working on the character recognition and review of work to recognize handwritten character for various Indian languages. The objective of this paper is to describe the set of preprocessing, segmentation, feature extraction and classification techniques.

Keywords

Handwritten character, Preprocessing, Segmentation, Feature extraction, Classification.

1. INTRODUCTION

India is a multi-lingual country with more than 20 languages like Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri (Meithei), Marathi, Nepali, Oriya, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu, Urdu etc[1] for which the concept of uppercase and lowercase is not present.

Handwritten Character Recognition is a process of transforming handwritten text into machine executable format. There are mainly three steps in pattern recognition: observation, pattern segmentation and pattern classification. Recognition of character has become very interesting topic in pattern recognition for the researchers during last few decades. In general, handwritten recognition is classified in to two types as on-line and off-line recognition methods [3]. Off-line handwriting recognition involves the automatic conversion of text into an image into letter codes which are usable within computer and text-processing applications. The data obtained by this form is regarded as a static representation of handwriting. But, in the on-line system, the two dimensional coordinates of successive points are represented as a function of time and the order of strokes made by the writer are also available. Offline character recognition is comparatively more challenging due to shape of characters, great variation of character symbol, different handwriting style and document quality.

Several applications including mail sorting, bank processing, document reading and postal address recognition require off-line handwriting recognition systems. As a result, the off-line handwriting recognition continues to be an active area of research towards exploring the newer techniques that would improve recognition accuracy [2].

The study defines the five major stages in any HCR which is shown in Fig.1.

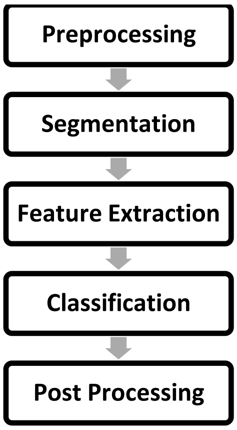

Fig.1: Stages of Character Recognition

This paper represents review of handwritten character recognition techniques with respect to the stages of character recognition systems for various Indian scripts.

2. PRE-PROCESSING

In image pre-processing we apply a series of operations like Binarization, Complement, Size normalization, Morphological Operation, Noise removal using filters, thresholding, skeletonization, thinning, cleaning techniques and filtering mechanisms on scanned image which are taken as input.It makes the input image easier to process in order to increase the overall efficiency of recognition system.

K.Singh et.al.[4] used median filtration, dilation, some morphological operations to join unconnected pixels, to remove isolated pixels, to set neighbor pixel values in majority and to remove spur pixels.

A.Desai[5] has done his work on Gujarati numerals recognition where he has collected 0-9 digits from 300 different people. In the scanned image contrast, adjustment is done by adaptive histogram equalization algorithm, smoothing of image boundaries are done using median filter and nearest neighborhood interpolation algorithm is used to put all handwritten digit in a uniform size. To deal with skew correction, digit is rotated up to 10^0 fine patterns for each digit in clock wise and anti-clock wise direction with difference of 2^0 each.

In the preprocessing phase, Md.Saidur et.al.[6]used canny method for edge detection and normalized numerals using thinning and dilation algorithm.

R.Singh and M.Kaur[7]suggested adaptive sampling algorithm, Otsu's threshold algorithm, Hilditch algorithm and its variants for normalization, image binarization and thinning of binarized image.

A.Aggarwal et.al.[9] used threshold value for converting image into binary image. Median filtering is used to remove the noise and after segmentation each character is normalized to size of 90*90.

In preprocessing stage, V.Agnihotri[12] used threshold value and sobel technique for binarization and edge detection. After binarization and edge detection, dilation on the image and filling of the holes were presented.

3. SEGMENTATION

Segmentation is process of extracting the basic constituent symbols of the script. Image is subdivided into many parts so that each part of the image is readable. To accomplish this task the image is subdivided considering three aspects, i.e. line wise segmentation, word wise segmentation and character wise segmentation.

For the segmentation, A.George and F.Gafoor[3]used horizontal histogram profile for the line segmentation and vertical histogram profile for the word and character segmentation.

J.Johan et.al.[8] have done segmentation using projection analysis and connected component labeling. In that they have used horizontal projection profile for the line segmentation and isolated character using connected component labeling algorithm.

In segmentation, V.Agnihotri[12] segmented the preprocessed image into isolated character using labeling process. The label provides information about number of characters in image.

4. FEATURE EXTRACTION

Feature extraction is a special form of dimension reduction. This approach is useful when image sizes are large and a reduced feature representation is required to quickly complete tasks such as image matching and retrieval.

K.Singh et.al.[4] used Zoning Density(ZD) and Background Directional Distribution(BDD) features for recognition. Zoning density is computed by dividing number of foreground pixels in the zone and background directional distribution values are calculated for each foreground pixel by directional distribution of its neighboring background pixels. The value for each directional distribution is summed up for all pixels in each zone and a specific mask is used in particular direction. By combining both types of features, total 144 features are used for classification.

A.Desai[5] has suggested four different profiles, horizontal, vertical and two diagonals for the feature extraction. The vector of these four profiles is used for identification of a digit.

For the feature extraction, Md.Saidur et.al.[6] extracts four directional local feature vector by Kirsch mask and one global feature vector. Kirsch mask is used to get the edges through the horizontal, vertical, right and left diagonal.

A.George and F.Gafoor[3] used contourlet transform in addition with aspect ratio, ratios of grid value in horizontal and vertical directions. The original image is divided to a lowpass image and a bandpass image using Laplacian Pyramid (LP) decomposing and each band pass image is further decomposed by Directional filter bank (DFB).This extraction method provides high recognition accuracy and taking less time for training and classification.

R.Singh and M.Kaur[7]has represented each character as a feature vector in the feature extraction stage. The various features for the classification are the character height, character width, number of horizontal lines (long and short), number of vertical lines (long and short) and number of slop lines and special dots.

After converting the original image into gray scale and size normalization, J.John et.al.[8] used Haar Wavelet features at different resolution scales for the feature extraction.

For the feature extraction, A.Aggarwal et.al.[9] have proposed Gradient feature. This feature measures the gradient magnitude and gradient direction of greatest change in intensity in a small neighborhood of each pixel. Sobel template is used to compute the gradients.

S.Niranjan et.al.[10] used Fisher Linear Discriminate analysis (FLD), 2DFLD, and diagonal FLD based methods for feature extraction to recognize unconstrained Kannada handwritten characters. They have calculated between class scatter matrix and within class scatter matrix. They solved generalized eigenvectors and eigenvalues, sorted eigenvectors by their associate eigenvalues from high to low and from each sample of training set extracted feature.

N.Patil et.al.[11] suggested Moment Invariants (MIs), Affine moments Invariants (AMIs), image thinning, structuring the image in box format for the feature extraction. The MIs are derived by means of the theory of algebraic invariants whereas AMIs are invariants under general affine transformation.

V.Agnihotri[12] used diagonal feature extraction for extracting the features. Individual character is resized to 90*60 pixels and divided into 54 equal zones and size is 10*10 pixels. The features are extracted from each zone by moving along their diagonals. This process is repeated for zones to extraction of 54 features for each character.

5. CLASSIFICATION AND RECOGNITION

The classification stage is the decision making stage of the recognition system. The performance of a classifier depends on the quality of the features which are extracted. There are many existing techniques available for handwriting classification.

K.Singh et.al.[4] used SVM (support vector machines) classifier for the recognition. SVM classifier takes the set of input data and classifies them in one of the only two distinct classes. The effectiveness of SVM depends on kernel used and kernel parameter. SVM with RBF (Radial Basis Function) kernel, they achieved 95.04% 5-fold cross validation accuracy.

A.Desai[5] used feed forward back propagation neural network for the classification of Gujarati numerals and proposed multilayered neural network with three layers (94,50,10) neurons respectively and has achieved 81.66% of accuracy in his work.

Md.Saidur et.al.[6] used PCA and SVM to enhance the accuracy. PCA decrease the dimension and extract more significant feature. The output of PCA is then passed to a SVM to determine appropriate class. They achieved 92.5% of accuracy.

Feed forward back propagation neural network algorithm used by A.George and F.Gafoor[3] as a classifier. The three hidden layer is used to perform the classification in this feed forward back propagation neural network. It gives 97.3% of recognition accuracy with total 32 features.

For the classification of the Telugu characters, R.Singh and M.Kaur[7]used Back Propagation algorithm. It is based on supervised learning. It consists of three layers: input, hidden and output. There are two phases in that: forward phase and backward phase.

J.John et.al.[8] used SVM classifier with RBF (Radial Basis Function) kernel for the classification of Malayalam characters. The feature space is linearly inseparable so using RBF kernel it is mapped into a high dimensional space and becomes linearly separable.

A.Aggarwal et.al.[9] used SVM with RBF kernel as a classifier. Basically SVM is two classes classifier. Margin width between the classes is the optimization criterion that is the empty area around the decision boundary defined by the distance to the nearest training pattern. This pattern called support vector which define classification function. They achieved 94% of recognition accuracy.

For the classification purpose, S.Niranjan et.al.[10] used different distance measure techniques such as, Minkowski, Manhattan, Euclidean, Squared Euclidean, Mean Square Error, Angle, Correlation co-efficient, Mahalonobis between normed vector, Weighted Manhattan, Weighted SSE, Weighted angle, Canberra, Modified Manhattan, Modified SSE, Weighted Modified SSE and Weighted Modified Manhattan are used which defines that combination of 2D-FLD with Angle and Correlation performs better recognition of vowels and consonants for Kannada handwritten characters compared to other methods and distance metric.

N.Patil et.al.[11] used Fuzzy Gaussian Membership function for the classification of Marathi handwritten characters. The template is formed and it consists of mean and standard deviation for each feature.

V.Agnihotri[12] used Feed Forward Back Propagation neural network for the classification. The neural network consists of 54/69 inputs layers, two hidden layer with 100 neurons and output layer with 44 neurons. He achieved 97% of recognition accuracy for 54 features and 98% for 69 features.

6. POST-PROCESSING

The last stage of the character recognition system is Post-processing. It prints the corresponding recognized characters in the structured text form. The accuracy of character recognition stages can be improved if the semantic information is available up to great extent.

7. STUDY OF VARIOUS CHARACTER RECOGNITION SYSTEMS FOR INDIAN SCRIPTS

Table 1. Study of various character recognition systems for Indian scripts

Researcher	Language	Dataset	Pre-Processing	Segmentation	Feature Extraction	Classification & Recognition	Accuracy
A.George et.al.[3]	Malayalam	-	-	Horizontal histogram profile, vertical histogram profile	Contourlet transform	Feed forward back propagation neural network algorithm	97.3%
K.Singh et.al.[4]	Gurumukhi	7000 samples	Median filtration, dilation, some morphological operations	-	Zoning Density (ZD) and Back-ground Directional Distribution (BDD) features	SVM (support vector machines) classifier	95.04%
A. Desai[5]	Gujarati Numerals	samples from 300 people	Adaptive histogram equalization algorithm, median filter and nearest neighborhood interpolation algorithm, skew correction		Feature vector of four different profiles- horizontal, vertical and two diagonals	Feed forward back propagation neural network	81.66%
Md.Saidur et.al.[6]	Bengali Numerals	1600 numerals	Canny method, using thinning and dilation algorithm		Four directional local feature vector by kirsch mask and one global feature vector	PCA and SVM	92.5%

R.Singh&M. Kaur[7]	Telugu	-	Adaptive sampling algorithm, Otsu's threshold algorithm and hilditch algorithm		Character height and width, the number of horizontal, vertical lines and slop lines, special dots	Back Propagation algorithm	
J.John et.al.[8]	Malayalam	10,000 samples	-	Projection analysis, connected component labeling	Haar Wavelet features	SVM classifier with RBF (Radial Basis Function) kernel	90.25%
A.Aggarwal et.al.[9]	Devanagari	7200 samples	Threshold value, Median filtering		Gradient feature	SVM with RBF kernel as a classifier	94%
S.Niranjan et.al.[10]	Kannada	5000 samples	-	-	Fisher Linear Discriminate analysis (FLD), 2DFLD, and diagonal FLD based methods	Different distance measure techniques	For angle distance measure 68% for FLD, 68% for 2D-FLD, 66% for Dia-FLD
N.Patil et.al.[11]	Marathi	1100 samples			Moment Invariants (MIs), Affine moments Invariants (AMIs), image thinning, structuring the image in box format	Fuzzy Gaussian membership function	MIs gives 75 %,AMI gives 89.09% and combination approach of MIs & AMIs gives 52.90 %
V.Agnihotri[12]	Devanagari	570 samples for testing	Threshold value, sobel technique, dilation, filling holes	Labeling process	Diagonal feature extraction	Feed Forward Back Propagation neural network	97% for 54 features and 98% for 69 features

8. CONCLUSION

India is a country in which different types of languages are used as medium of communication among various masses. All individual languages contain unique set of characters. Majorly this study deals with various methodologies of different phases of character recognition. Many researchers have proposed their work in this area and achieved good accuracy rate. Very few researchers have explored their research pointing out complexities involved in Indian script such as characters with modifiers, disconnected characters and conjugant characters. Still there is huge scope of research into field of character recognition for a researcher.

9. REFERENCES

[1] Ratnashil N Khobragade, Dr. Nitin A. Koli and Mahendra S Makesar, A Survey on Recognition of Devnagari Script, International Journal of Computer Applications & Information Technology, Vol. 2, Issue 1, 2013, pp. 22-16.

[2] Vijay Laxmi Sahu and Babita Kubde, Offline Handwritten Character Recognition Techniques using Neural Network: A Review, International Journal of Science and Research (IJSR), Vol. 2 Issue 1, 2013, pp.87-94.

[3] Aji George and Faibin Gafoor, Contourlet Transform based Feature Extraction for Handwritten Malayalam Character Recognition using Neural Network, International Journal of Industrial Electronics and Electrical Engineering, Vol. 2, Issue-4, 2014, pp.19-22.

[4] Kartar Singh Siddharth, Renu Dhir and Rajneesh Rani, Handwritten Gurumukhi Character Recognition using Zoning Density and Background Directional Distribution Features, International Journal of Computer Science and Information Technologies, Vol.2(3), 2011, pp.1036-1041.

[5] Apurva A. Desai, Gujarati handwritten numeral optical character reorganization through neural network, Pattern Recognition, Vol. 43, 2010, pp. 2582–2589

[6] Md.Saidur Rahman, G. M. Atiqur Rahaman, Asif Ahmed and G.M. Salahuddin, An Approach to Recognize Handwritten Bengali Numerals for Postal Automation,

International Conference on Computer and Information Technology, 2008, pp. 171-176.

[7] Rinki Singh and Mandeep Kaur, OCR for Telugu Script Using Back-Propagation Based Classifier, International Journal of Information Technology and Knowledge Management, Vol. 2, No. 2,2010,pp. 639-643

[8] Jomy John, Pramod K. V. and Kannan Balakrishnan, Unconstrained Handwritten Malayalam Character Recognition using Wavelet Transform and Support vector Machine Classifier, International Conference on Communication Technology and System Design 2011, Vol. 30, 2012,pp. 598-605

[9] Ashutosh Aggarwal, Rajneesh Rani and Renu Dhir, Handwritten Devanagari Character Recognition Using Gradient Features, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 5, 2012, pp. 85-90.

[10] Niranjan S.K, Vijaya Kumar, Hemantha Kumar G and Manjunath Aradhya V N, FLD based Unconstrained Handwritten Kannada Character Recognition, International Journal of Database Theory and Application, Vol. 2, No. 3, 2009, pp. 21-26.

[11] Nilima P. Patil K. P. Adhiya and Surendra P. Ramteke, A Structured Analytical Approach to Handwritten Marathi vowels Recognition, International Journal of Computer Applications, Vol. 31– No.3, 2011, pp.48-52.

[12] Ved Prakash Agnihotri, Off-Line Handwritten Devanagari Script Recognition using Diagonal Feature Extraction Method, International Journal of Research in Science And Technology, Vol.1, Issue No. 5, 2012.