Study of Different Off-line Handwritten Character Recognition Algorithms for Various Indian Scripts

Hetal R. Thaker
Atmiya Institute of Technology & science,
Kalawad Road,
Rajkot – Gujarat, INDIA

C. K. Kumbharana, PhD.
Head - Dept. of Computer Science,
Saurashtra University,
Rajkot – Gujarat, INDIA

ABSTRACT
Handwritten recognition is an area of research where many researchers have presented their work and is still an area under research to achieve higher accuracy. In past collecting, storing and transmitting information in form of handwritten script was the most convenient way and is still prevailing as a convenient medium in the era of digital technology. As technology has advanced tablet and many similar devices allows humans to input data in form of handwriting. Use of paper to write handwritten text, converting to an image using scanner, identifying handwritten characters from the image is known as off-line handwritten text recognition is a challenging area due to the fact that different people will have different style of writing and all scripts have their own character set and complexities to write text. Many researchers have presented their work and many algorithms are proposed to recognize handwritten and printed characters. One can trace extensive work for off-line handwritten recognition for English and Arabic script. This paper presents review of work to recognize off-line handwritten text for various Indian language scripts. Paper reviews methodologies with respect to the phases of character recognition.

General Terms
Pattern Recognition, Off-line Handwriting recognition

Keywords
Character Recognition, Off-line handwriting recognition, Pre-processing, segmentation, feature extraction, classification, Indian script handwriting recognition.

1. INTRODUCTION
Recognizing handwritten character written by human sometimes difficult to recognize by other human but still humans are performing far better in comparison to machines in many areas, so making an interface which is capable of recognizing characters written by human yet requires intensive research.

Handwriting recognition classified into two types as off-line and on-line handwriting recognition methods. In off-line recognition, the writing is usually captured optically by a scanner and complete writing is available as an image. But, in the on-line handwriting recognition words are generally written on a pressure sensitive surface from which real time information, such as the order of the stroke made by the writer is obtained and preserved. On-line handwriting recognition is shown as superior in comparison to off-line handwriting recognition as temporal information will be available with the on-line handwriting recognition.

Several applications including mail sorting, bank processing, document reading and postal address recognition require off-line handwriting recognition systems.[15] As a result, the off-line handwriting recognition continues to be an active area for research towards exploring the newer techniques that would improve recognition accuracy.

Many researchers have worked in the area of handwriting recognition, and numerous techniques and models have been developed to recognize handwritten text.

The study investigates that there exist five major stages in any character recognition that are as per figure. 1.

Paper reviews off-line handwritten character recognition methodologies with respect to the phases of the Character Recognition systems for research work presented by researchers in the field of handwritten character recognition for various Indian Script like Devnagari, Tamil, Telugu, Kannada, Hindi, Gurumukhi, Malyalam, Gujarati.

2. PRE-PROCESSING
The pre-processing is a series of operations performed on the scanned image taken as a input. This phase enhances the image rendering making it suitable for segmentation. Various tasks performed during this stage are : thresholding, Noise removal / reduction, binarization, edge detection, skew detection and correction, size normalization, dilation and filling.[14,15]

J. R. Prasad et.al.[4] have used median filter to remove salt and pepper noise from the scanned images stored in png file format and have applied thinning to reduce character to minimum one pixel thickness.

N.Shanthi et. al.[11] have used Otsu's global thresholding method to extract the foreground from background, hilditch algorithm is applied to do skeletonization as it is a parallel sequential algorithm.

B.V. Dhandra et. al. [13] have proposed novel approach for recognizing kannada, telugu and devnagari handwritten numerals by applying median filter to remove noise and morphological opening operations to remove scanning artifacts.

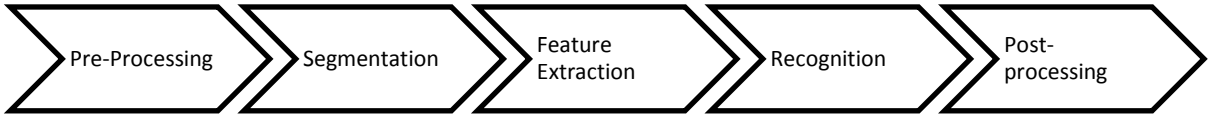

Fig 1: Character Recognition Steps

Apurva A. Desai[2] has presented his work to recognize Gujarati numerals where he has collected numerals from 300 different people of various backgrounds and different genders. On the image scanned in 300 dpi by a flatbed scanner contrast adjustment is performed using contrast limited adaptive histogram equalization algorithm, median filter is used for smoothening boundaries of image and Nearest Neighbor Interpolation algorithm is used to put all scanned digit images in uniform size. To remove skew he has rotated numerals upto 100 about center point and created five patterns rotating numeral images clock wise and anti-clock wise direction and having difference of 20 each.

To recognize Tamil handwritten characters R. Kannan et. al.[5] have applied preprocessing techniques where thresholding to extract foreground ink from background image have used, median filtering and wiener filtering to remove noise and cumulative scalar product of windows of text blocks with gabor filters at different orientation is calculated for skew angle detection and have applied it to all possible 50X50 windows and skew angle was found as median of all angles obtained.

In paper authored by R.Kannan et.al.[7] in an attempt to recognize Tamil handwritten characters have used octal graph conversion which improves the slant correction. He has explained conversion of letter into an octal graph, by representing each pixel of a given character as a node of the graph where each node has eight field, as basic form of a letter can be represented independent of the style of writing[7]. For segmentation he has proposed a strategy in which assign a number for each pixel and scanning row by row, if pixel is unvisited with neighbor pixel visited assigned a no. to the current pixel and in case of more no. of visited neighbor pixel, least no. is chosen to assign value to current pixel and for non visited neighbor pixel assigned next no. of pixel continue and assigned value to all pixels, finally using DFS or BFS assigned a same number for all the continuous patterns where each letter is assigned a unique number and hence can be separated.

R.Kannan et. Al.[7] has demonstrated conversion of input to octal graph by following steps where normalization, conversion identification of weighting factors and feature matching is used. The normalization process involves correcting slant, normalization width and vertical scaling by normalizing height of three zones to a fixed size.

M.Jangid et. al. [12] has presented his work to recognize handwritten devnagari numeral where otsu's gray thresholding is used for binarization, median filtering for removing salt and pepper noise, normalized the image into 32X32, 40X40 and 48X48.

Bikash Show et.al.[10] have smoothen image using median filter and binarized using Otsu's thresholding method.

3. SEGMENTATION

Segmentation is a process in which an image of sequence of characters is decomposed into lines, words, and even characters of a hand written document, a crucial step as it extracts the meaningful regions for analysis. In case of many Indian script characters may have modifiers called Matras. So, identifying matras is crucial step in segmentation.

R. Kannan et. al.[5] have calibrated local minima points with each component to approximate imaginary baseline for line segmentation caps between character segments and heights of character segments.

B. Show et.al. [10] have used morphology operator opening in which composition of erosion followed by dilation is used for detecting matras in sample images of handwritten devnagari script.

N.Shanthi et. al.[11] has used valleys in horizontal histogram for line segmentation, vertical histogram profile is calculated for each segmented line and inter space between histogram is used to separate characters, bilinear interpolation is used to normalize the scanned image to fixed pre-specified size.

4. FEATURE EXTRACTION

Feature extraction is the process of collecting distinguishable information of an object or a group of objects so that on the basis of this information we can classify objects with different features. I.S. Oh [16] has defined that feature extraction and selection is a process of extracting the most representative information from the raw data. For this purpose, a set of features are extracted for each class that helps to make it separate from rest of the classes.

Baheti M.J. et. al. et.al. [1] have derived feature set from affine invariant moments for each gujarati numerals.

For extracting features of Gujarati numerals Apurva A. Desai[2] has suggested technique in which he has used four different profiles horizontal, vertical, two diagonals to create a feature vector for each digit.

R. Kannan et. al.[5] have used vectoriztion process on basis of bi-dimensional image and have used oriented search process and have scanned from top to bottom and left to right and have identified starting point of first line segment, first pixel.

S. Niranjan et. Al. [6] in his attempt to recognize unconstrained Kannada handwritten characters have used Fisher linear discriminate analysis – FLD, 2DFLD and diagonal FLD methods. He has calculated within class scatter matrix, between class scatter matrix, solved eigen values and eigen vector, sorted the eigen vectors by associate eigen values from high to low and extracted feature for each sample in training set.

S. Arora et.al. [8] have proposed technique to recognize non-compound devnagari charcters in which two feature sets are created one as shadow features extracted from scaled

binarized image and second from chain code histogram feature extracted using chain coding that is the contour points of scaled character binary image.

N. Sharma et. al.[9] have described histograms of direction chain code of the contour points of the characters for extracting feature in an attempt to recognize sample handwritten devanagari characters, where 64 dimension of feature vector.

B. Show. et.al. [10] have considered two directional view based extraction of strokes that are either horizontal or vertical where 8 scalar features were extracted representing the shape, size and position of stroke with respect to pseudocharacter image.

M.Jangid et. al. [12] has demonstrated use of zoning density (ZD) where no. of foreground pixels in each zone is divided by total no. of pixels in each zone to obtain 16 zoning density feature and Background directional distribution (BDD) where 8 directional distribution feature is obtained.

B.V. Dhandra et. al.[13] has applied directional density estimation feature, features based on water reservoir principle, maximum profile distance feature and fill hole density feature structural features used for handwritten numerals.

N.Shanthi et. al.[11] has applied zoning methodology to identify features for given inputted handwritten tamil characters.

5. CLASSIFICATION & RECOGNITION

The classification stage is the decision making part of a recognition system and it uses the features extracted in the previous phase of recognizing character. [14]

Baheti M.J. et. al. et.al. [1] has compared K-Nearest Neighbor (KNN) and Principal Component Analysis (PCA) to classify Gujarati handwritten numerals in which they have observed to have higher recognition rate using KNN compare to PCA by applying both these classifier methods on data collected from 80 different people and have achieved recognition rate of 90.04% with KNN and 84.1% for PCA. In paper presented by authors[3] is an extension work where affine moment invariant moment is used as a feature extraction technique and have compared Gaussian distribution function and Support vector machine along with KNN and PCA and found that SVM have highest recognition rate of 92.28% in comparison to Gaussian distribution which is 87.2%. Authors have concluded that SVM proves to be better classifier than PCA, K-NN and Gaussian distribution function classifier for affine invariant moments as a technique for feature extraction [3].

Apurva A. Desai[2] in his work has used Feed forward back propagation neural network to classify Gujarati numerals and have demonstrated multi layer neural network with three layers (94, 50 and 10) neurons respectively and have achieved 82% of success rate.

J. R. Prasad et. al.[4] have used template matching algorithm in which various steps like template classification based on median data number and ITF level, correlation analysis, computation of cross correlation coefficient, identification of valid cross correlation coefficient and pattern matching are involved. Based on geometrical shapes of gujarati script characters six groups where formed and have achieved 71.66% of an average overall recognition rate which was examined on ten samples four good, four misaligned and two extremely disfigured samples. In proposed technique[4] of combining advantage of template matching and neural network method for achieving high recognition rate with improved speed to recognize Gujarati handwritten characters and have applied cross correlation function for pattern matching to identify similar patterns between input test image and standard trained database image.

R.Kannan et. al.[5] have used discrete Hidden Markov Model to recognize cursive handwritten tamil character recognition have created two HMM for every character for modeling horizontal information and vertical information and have reported 96.4% efficiency for 25 count of words.

S. Niranjan et. Al. [6] have used Fisher linear discriminate analysis – FLD, 2DFLD and diagonal FLD methods and have used different distance measure techniques such as correlation-coefficient, Manhattan, Mahalanobis distance between normed vectors, Mahalanobis, Euclidean, Minkowski, Modified manhattan, Modified Sq Euclidean, Mean sq error, Sq. Euclidean, Weight angle, Weight manhattan, Weight modified manhattan, Weight modified SSE, Weight SSE, Canberra, Angle etc. and have concluded that 2D-FLD with angle and correlation performs better recognition compared to other methods and distance metric for vowels and consonants for Kannada handwritten characters.

R. Kannan et. al.[7] has calculated weighing factor which includes loops, horizontal lines, vertical lines for matching input graph and character in repository. In process of comparing them result that will be displayed will be the one with highest level of confidence. Author has concluded 82% of overall success rate for ten samples of tamil character set.

S. Arora et. al.[8] have used Multi Layer Perceptron(MLP) classifier with three layers where MLP is trained using backpropagation learning algorithm with momentum and have applied to two different feature sets i.e. Chain code histogram feature set and Shadow feature set and have concluded success rate of 73.33% and 68.10% respectively. For classification of confused character set S. Arora has applied minimum edit distance on the image where corners are detected and image is divided into 25 segments where in each segment number of corners were counted. For corner detection Harris/Plessey corner detection with some modification is used. Result achieved by S. Arora et. al. [8] was 76.67% for combined MLP (top 1 choices), 93.27% combined MLP (top 5 choices) and 85% with minimum edit distance classification and overall recognition rate was found to be 90.74%.

N. Sharma et. al. [9] have presented use of modified quadratic discriminant function for quadratic classification. He has compared result with other researcher's result and concluded 98.86% accuracy for Devnagari numerals and 80.36% accuracy for Devnagari characters.

B. Show et.al. [10] have used Hidden Markov Model as classification technique at pseudocharacter level to recognize handwritten devnagari characters.

N.Shanthi et. al.[11] has proposed a novel approach of recognizing Tamil handwritten characters using Support Vector Machine classifier and have achieved recognition

accuracy between 62.84 and 98.9% for different tamil characters.

M.Jangid et. al. [12] has used Support vector machine as classification technique to classify handwritten devanagari numerals. For 32*32 with 144 no. of features cross validation accuracy obtained is 98.94% whereas with 48*48 with 324 no. of features cross validation accuracy found was 99.08%.

B.V. Dhandra et. al.[13] have adopted Probabilistic Neural Network (PNN) classifier and average recognition rate of 99.40%, 99.60%, 98.40% was achieved for Kannada, Telugu and Devnagari numerals.

6. POST-PROCESSING

By means of post-processing one can improve handwriting recognition rate by relying on contextual post-processing or lexical post-processing, using which recognition rate can be increased by resolving ambiguities.

G S Lehal et.al.[17] have proposed shape based post processing for Gurumukhi OCR in they have partitioned Punjabi corpora based on shape and size of a word. They have designed post processor using statistical information of Punjabi language syllable combination, holistic recognition of most frequently occurring word and corpora lookup.

Research is underway for improving handwriting recognition rate using this phase for Indian script.

7. COMPARATIVE ANALYSIS OF DIFFERENT OFF- LINE HANDWRITTEN ALGORITHMS

Many researchers have proposed and presented their work in recognizing off-line handwritten text for Indian script, Table 1 represents comparative analysis of various researchers' work.

Table: 1 Comparative analysis of various researchers's off-line handwritten recognition algorithms

Researcher	Script	Sample Data	Pre-processing	Segmentation	Feature extraction	Classification & Recognition	Average recognition Rate / Efficiency
Baheti M.J. et. al. [1,3]	Isolated Gujarati Numerals	One sample of 0-9 digit from 80 persons[1] 1600 samples [3] on specials designe data sheet	Binarization, Size Normalization, Skeletonization[3]	Bounding box segmentation [3]	Affine Invariant Moments	K-Nearest Neighbor (KNN)	90.04%
						Principal Component Analysis (PCA)	84.1%
						Gaussian Distribution Function	87.2%
						Support vector machine (SVM)	92.28%
Apurva A. Desai[2]	Gujarati Numeral	0-9 from 300 people	contrast limited adaptive histogram equalization algorithm, median filter, Nearest Neighbor Interpolation algorithm, Skew-correction	---	Feature vector of four different profile – horizontal, vertical and two diagonal	Multilayer Feed forward back propagation neural network	81.66%
J. R. Prasad et. al.[4]	Gujarati	10 samples	Inversion, Gray Scaling, Median Filter, Thinning	---	Individual image pixel as feture	Template Matching & Feed forward back propogation Neural Network	71.66%
R. Kannan et. al.[5]	Tamil	---	Thresholding Median filter & Wiener filter, skew detection	Line, word and character segmentation	Measuring and approximating geometrical properties	Hidden Markov Model	96.4% for 25 count of words

S. Niranjan et. Al.[6]	Kannada	5000 characters	---	---	Fisher Linear Discriminant Analysis (FLD), 2-D FLD, Diagonal FLD	Nearest Neighbor Classifier, Distance measure techniques	For angle distance measure 68.00 for FLD, 68.00 2-D FLD, 66.00 for Dia-FLD (Vowels & Consonants)
R.Kannan et.al [7]	Tamil	---	Slant Correction	---	Topological features	Feature Matching (loop, horizontal-vertical line)	82%
S. Arora et.al.[8]	Devnagari	7154 samples	---	---	Shadow feature & chain code histogram feature	Multilayer Perceptron Classifier – backpropagation training algorithm, Minimum Edit Distance	90.74%
N. Sharma et. al. [9]	Devnagari	11270 samples	Global binarization,	Bounding Box	Histogram of chain code	Modified Quadratic Discriminant Function (MQDF) , Quadratic Classifier	Numerals - 98.86% Characters – 80.36%
Bikash Show et.al.[10]	Devnagari	22500 training 17200 testing	Median filter Otsu's thresholding	Morphology operators – erosion and dilation	Scalar feature	Hidden Markov Model at Pseudo Character level	81.63% for test set 84.31% word level accuracy
N.Shanthi et. al.[11]	Tamil	35441 characters	Otsu's thresholding Hilditch for skeletonization	Horizontal and vertical histogram	Zoning	Support Vector Machine	82.04%
M. Jangid[12]	Devnagari	22546	Otsu's thresholding	---	Zoning Density & Background Directional Distribution	Support Vector Machine	48*48 sample size with 324 feature 99.08%
B.V. Dhandra et. al. [13]	Kannada, Devnagari Telugu numeral	2500 – Kannada, 1250 Telugu, 1250 Telugu	Median filter	---	Structural feature	Probabilistic Neural Network	99.40% - Kannada, 99.60% - Telugu, 98.40%- Devnagari

8. CONCLUSION

India is a Muti-lingual and Multi-script country [15] where every Indian script have their own character set this paper studies various methodologies proposed by researchers to recognize off-line handwritten characters for various Indian script. Many researchers have proposed their work in this area and have achieved better accuracy rate. Very few researchers have presented their work which points out complexities involved in many Indian scripts like recognizing conjunct characters (half consonant with other), character with modifiers (called matras). So, Off-line handwriting character recognition is an open area where still there is a scope of a research is there for proposing methodologies by identifying complexities of Indian script.

9. REFERENCES

[1] M. J. Baheti, K. V. Kale, and M. E. Jadhav, 2011 COMPARISON OF CLASSIFIERS FOR GUJARATI NUMERAL RECOGNITION, International Journal of Machine Intelligence, vol. 3.

[2] Apurva A. Desai, 2010 Gujarati handwritten numeral optical character reorganization through neural network, Pattern Recognition, vol. 43, pp. 2582-2589.

[3] M. J. Baheti, Kale K.V.,2012 Gujarati Numeral Recognition : Affine Invariant Moments Approach, *Soft Computing*, pp. 140-146.

[4] J. R. Prasad, U.V. Kulkarni, 2003. Offline Handwritten Character Recognition of Gujrati Script using Pattern Matching, *Computer Engineering*.

[5] R. J. Kannan, R. Prabhakar, 2008. Off-Line Cursive Handwritten Tamil Character Recognition, *Signal Processing*, vol. 4, no. 6, pp. 351-360.

[6] S. K. Niranjan, V. Kumar, H. K. G, and M. A. V. N, 2009 FLD based Unconstrained Handwritten Kannada Character Recognition, *International Journal*, vol. 2, no. 3, pp. 21-26.

[7] R. Jagdeesh Kannan, R. Prabhakar, 2008 An Improved Handwritten Tamil Character Recognition System using Octal Graph, Department of Computer Science and Engineering , Department of Computer Science and Engineering , Coimbatore Institute of Technology , Co, Journal of Computer Science, vol. 4, no. 7, pp. 509-516.

[8] S. Arora, D. Bhattacharjee, M. Nasipuri, D.K.Basu, M.Kundu Recognition of Non-Compound Handwritten Devnagari Characters using a Combination of MLP and Minimum Edit Distance, *Journal of Computer Science*, no. 4, pp. 107-120.

[9] N. Sharma, U. Pal, F. Kimura, and S. Pal, 2006. Recognition of Off-Line Handwritten Devnagari Characters Using Quadratic Classifier, *Language*, pp. 805-816.

[10] B. Shaw, Swapan Kr. Parui, Malayappan Shridhar, 2008 "Offline Handwritten Devanagari Word Recognition : A Segmentation Based Approach," Stroke.

[11] N. Shanthi, K. Duraiswamy, 2010, A novel SVM-based handwritten Tamil character recognition system, *New York*, pp. 173-180.

[12] M. Jangid, R. Dhir, R. Rani, and K. Singh, 2011 SVM Classifier for Recognition of Handwritten Devanagari Numeral, *Processing*, no. Iciip.

[13] B. V. Dhandra, R.G.Benne, M. Hangarge Kannada , Telugu and Devanagari Handwritten Numeral Recognition with Probabilistic Neural Network : A Novel Approach, *Architecture*, pp. 83-88, 2010.

[14] J.Pradeep, E.Srinivasan, and S.Himavathi, 2011 DIAGONAL BASED FEATURE EXTRACTION FOR HANDWRITTEN ALPHABETS RECOGNITION, *International Journal of Computer Science & Information Technology*, vol. 3, no. 1.

[15] V. J. Dongre, V. H. Mankar, and G. Suganya, 2010 A Review of Research on Devnagari Character Recognition, *International Journal of Computer Applications*, vol. 12, no. 2, pp. 8-15.

[16] I. S. Oh, J. S. Lee, C. Y. Suen, 1999 Analysis of class separation and Combination of Class-Dependent Features for Handwriting Recognition, IEEE Trans. Pattern Analysis and Machine Intelligence, vol.21, no.10, pp.1089-1094.