IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 2, Ver. IV (Mar – Apr. 2015), PP 83-88
www.iosrjournals.org

Handwritten Character Recognition: A Comprehensive Review on Geometrical Analysis

Meenu Mohan [1], Jyothi R.L[2]
[1](Computer Science & Engineering, College of Engineering, Karunagappally/ Cusat, India)
[2](Computer Science & Engineering, College of Engineering, Karunagappally/ Cusat, India)

Abstract: *This paper presents a detailed review of Offline Handwritten Character Recognition. HCR is an optical character recognition, which convert the human readable character to machine readable format. In HCR, to attain 99% accuracy is very difficult. Here a detailed study on Geometrical methods of feature extraction in character recognition has been done by giving more emphasis to Zone based techniques and it has been analyzed that the efficiency of HCR depends on the selection of appropriate feature extraction methods and classifier. A comparative study in various steps in character recognition like Preprocessing, Segmentation, Feature Extraction and Classification are carried out. Various application areas of HCR like Postal address reading, mail sorting, office automation for text entry, person identification, signature verification, bank-check processing etc. are also analyzed.*

Keywords: *OCR, Preprocessing, Segmentation, Feature Extraction, Classification.*

I. Introduction

Character Recognition is an active research area in the field of image processing and pattern recognition. It is the process of converting an image representation of document into digital format. Character recognition is of 2 types: Magnetic character recognition and Optical character recognition. Optical character recognition (OCR) is the translation of scanned images of handwritten, typewritten or printed document into machine encoded form. The document image may be printed or handwritten. The printed document means that the documents are written by electronic devices, which includes all the printed materials such as book, newspaper, magazine etc. Handwritten documents are written by hand held equipments. The handwritten recognition system can be classified into online and offline hand written recognition system as shown in Fig 1.

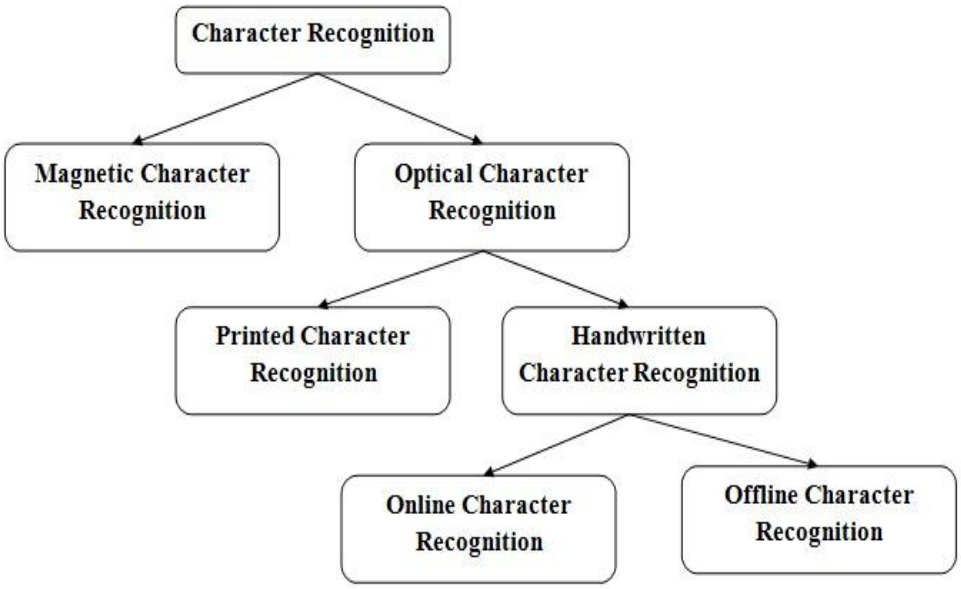

Fig 1: Classification of Character Recognition

In online handwritten character recognition (HCR), a special electronic pen samples the handwriting input where the writing is done on electronic surface. Here recognition is done in real time. Here features that are extracted depend on the dynamic information that has been used as input. In offline handwritten character the information that serves as input does not exhibit any dynamic change but the most important challenge of handwritten character recognition is the variability of writing style. Different person have their own handwriting. So the handwritten text varies from person to person. This paper discusses various methodologies

that have been analyzed based on literature study of handwritten character recognition systems. The different stages of Handwritten character recognition system are Pre-processing, Segmentation, Feature Extraction and Classification.

II. Phases Of Character Recognition System

A. Image Acquisition:-

The offline recognition system acquires an optically scanned image as an input image. Digitization in handwritten character recognition is the process of converting a handwritten document into a digital format. A scanner or digital camera captures an image of text and converts it to an image files format such as a bitmap, jpeg etc.

B. Preprocessing:-

Preprocessing is a series of operations that is performed on the scanned input image to improve the quality of image for effective feature extraction. Major steps under pre-processing are:

1. Noise Removal
2. Binarization
3. Morphological Operations
4. Size Normalization

Noise is introduced in an image during image acquisition. It produces a random variation of image intensity and sometimes will be visible as grains in the image. Noise removal is the process of removing or reducing the noise from the image. There exist several algorithms and filters for noise reduction and removal. The different types of noises that exist in document images are Salt and Pepper noise, Gaussian noise, Gamma noise, Uniform noise etc. Various type of filtering methods like Gaussian filtering method, Min-max filtering method etc. are applied for noise removal. Median filter is used to remove salt and pepper noise. Binarization is the process of converting colour or gray-scale image into binary image with the help of thresholding. The different methods of binarization are Global thresholding, Local thresholding, Adaptive thresholding, Otsu's method etc. Morphological operations are also used in preprocessing. Dilation and Erosion are commonly used morphological operation that increase or decrease character size of an image. Dilation is the process of adding pixels to the character boundary. In erosion, the pixels are removed from the boundary of character. Skeletonization is the process of reducing the character image to single pixel wide representation.

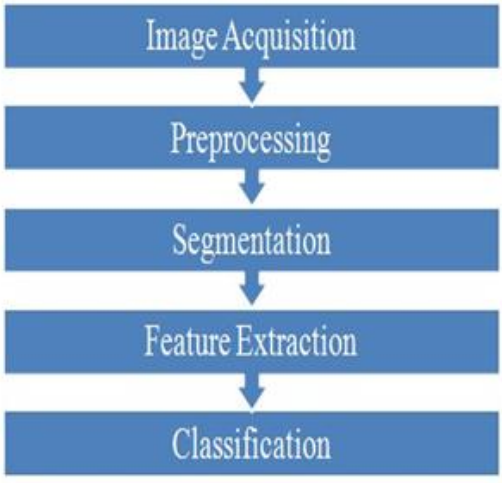

Fig 2: Offline Handwritten Character Recognition Architecture

Normalization is the process that reduces the range of pixels intensity values present in an image. Size normalization is the preprocessing step that resizes the character image into a standard size. Skew detection and correction is also a part of character image preprocessing. During document scanning, skew is introduced in the image. Skew angle is an angle that the text lines of the image make with horizontal direction. The aim of skew detection is to align an image text before processing. Commonly used skew elimination techniques are projection profile method and Hough transform method.

C. Segmentation:-

Segmentation is the process that isolates individual character from handwritten character image. Segmentation is classified into Implicit and Explicit segmentation. In implicit segmentation, the words are predicted directly without segmenting the word as individual letters but the explicit segmentation, the word is segmented into individual character. Segmentation is carried out using threshold based, edge based, region based, clustering techniques etc. The different steps in segmentation are line, word and character segmentation. In line segmentation, horizontal projection profile method is used. It separates the boundaries between lines. Word segmentation is done by applying vertical projection profile method on the separated lines. Finally, the characters are isolated from these words using connected component labeling.

D. Feature Extraction:-

The feature extraction method is the most vital and conclusive one and therefore the features should be extracted correctly, that decides the effectiveness of the classification. Feature extraction methods are classified into three major groups:

1. Statistical features.
2. Global transformation and Series expansion
3. Structural features.

Statistical features represent the character image as statistical distribution of points. Zoning, Crossing and Distances, Projections etc. are the various methods used for statistical feature extraction. Global transformation and series expansion includes various techniques like Fourier transform, Gabor transforms, Wavelets, Moments and Karhunen-Loeve Expansion etc. Structural features are based on geometrical and topological properties of the character. Loops, curves, lines, T-point, cross, aspect ratio, strokes and their directions, inflection between two points etc. are used as structural features.

E. Classification:-

Classification is the decision making part of the any recognition system. Various approaches for classification in character recognition systems are analyzed. Most commonly seen classifiers are Artificial Neural Network, SVM, and Nearest Neighbor classifier. The classifiers compare the given vector with the stored pattern and give the best match as an output. The various pattern classification methods can be successfully applied to character recognition. The classification methods that are used in handwritten character recognition systems are categorized into statistical methods, ANN, SVM, structural methods and multiple classifier methods. In case of Statistical methods, ANN and SVM the input feature vectors should be of same dimensionality for a single recognition system. In multiple classifier methods, the classification results of multiple classifiers are combined to reorder the classes.

III. Literature Review

There are many researches that have been done in the field of image processing and pattern recognition which is related to handwritten character recognition. This section describes an extensive review for handwritten character recognition:

J.Pradeep et.al. [1] focus on recognition of English offline handwritten character using Neural Network. Noise Removal is done using median filter, Binarization is using Otsu's global technique, Detection of edges are done using Sobel filter, dilation and filling are also carried out as the part of preprocessing. Diagonal feature extraction method is used for feature extraction stage. Divide the enhanced image into 54 equal zones. So 54 features are obtained from each character. In classification, feed forward back propagation neural network is used. The diagonal features provide good recognition accuracy compared to the conventional horizontal and vertical methods of feature extraction. Using this 54 feature based system it yield a recognition efficiency of 98%.

S.V. Rajashekararadhya et. al. [2] proposed the Image centroid and Zone centroid based distance metric feature extraction system handwritten numeral recognition for four popular South Indian Scripts. The four languages are Malayalam, Tamil, Kannada and Telugu. Preprocessing stage concentrated on Noise reduction, Slant correction, Normalization and Thinning. In feature extraction stage, image centroid and zone centroid and hybrid methods are used. Feed forward back propagation neural network and nearest neighbor classifiers are used for subsequent classification and recognition purpose. The recognition accuracy obtained for Kannada and Telugu numerals is 99%. For Tamil and Malayalam numerals have obtained 96% and 95% respectively.

S.L.Mhetre et.al. [3] have proposed two different approaches for recognition of Devanagari handwritten numerals. In the first method, Grid features are used. In the second method, ICZ (Image Centroid Zone) & ZCZ (Zone Centroid Zone) features based on distance information are extracted. Here ANN and

matching score are used for classification and the accuracies obtained using two approaches are evaluated. In classification ANN provide better accuracy compared to matching score.

S. V. Rajashekararadhya et. al. [4] has described Zone based feature extraction system (ICZVDD-ICZHRD method for handwritten numeral/mixed numerals recognition of South-Indian scripts. The nearest neighbor, feed forward back propagation neural network and support vector machine classifiers are used for classification. The recognition rate obtained of 98.65 % for Kannada numerals, 96.1 % for Tamil numerals, 98.6 % for Telugu numerals and 96.5% for Malayalam numerals using the Support vector machine.

S. V. Rajashekararadhya et. al. [5] have described Image Centroid Zone (ICZ) based Angle feature extraction for handwritten numeral recognition of Kannada script. The numerals image centroid is computed and the image is further divided into n equal zones. Average angle from the character centroid to the pixels present in the zone is computed. This procedure is repeated sequentially for all zones present in the numeral image. Finally n such features are extracted. For classification purpose, nearest neighbor classifier and support vector machines are used. The recognition accuracy achieved 96.05% for Kannada numerals using Support vector machines.

Gita Sinha et. al. [6] had taken care of Arabic numeral recognition. In preprocessing stage, binarization, dilation, erosion, noise removal and normalization are used. Three features extraction techniques that has been used are Image Centroid Zone (ICZ), Zone Centroid Zone (ZCZ) and Hybrid feature extraction techniques. Hybrid feature extraction techniques are combination of ICZ+ZCZ. SVM classifier is used for classification. The recognition rate is 97.21% on handwritten Arabic numeral.

Seema A. Dongare et.al. [7] have proposed Devanagari character recognition works in stages as document preprocessing, segmentation using line segmentation, word and character segmentation, feature extraction using zone based approach followed by recognition using feed forward neural network. Recognition of handwritten Devanagari character is quite difficult due to presence of shirorekha, conjunct characters and similarity in shapes for multiple characters. Here an attempt is carried out to increase the accuracy and performance.

Gita Sinha et.al. [8] presented Gurumukhi handwritten character recognition. Preprocessing stage includes steps like Gray scale conversion, Binarization using Otsu's method, filtering and morphological operation, Noise removal, Skeletonization, Skew detection. Zone-based feature extraction technique is used for extracting the feature and SVM classifier is used for Gurumukhi handwritten character recognition. The recognition accuracy obtained is 95.11%.

Sandeep Saha et.al.[9] proposed 40-point feature extraction for English handwritten character recognition using multilayer feed forward neural network. The whole image is divided into 16 zones and then computed the average intensities of each zones. Then the entire image is divided diagonally from left top to bottom, right top to bottom, left bottom to top and right bottom to top and innermost cell features are extracted. Finally features vectors consisting of 40 features are tested using the artificial neural network and has a better recognition efficiency is reported.

Sangeetha Sasidharan et. al. [10] describes that segmentation of offline Malayalam handwritten character recognition. Preprocessing stage includes noise removal and binarization. In segmentation stage, line segmentation using horizontal projection profile method is used. Character segmentation focus on the segmentation of untouched characters, segmentation of consonants touching to Valli (special Malayalam character) and segmentation of consonants touching to Chandrakala(special Malayalam character). The efficiency obtained in this work is 94.08%.

Anita Pal et. al. 11] have proposed boundary tracing along with Fourier Descriptor for handwritten English character recognition. In preprocessing stage skeletonization and normalization is performed. In feature extraction stage, boundary detection is done using 8-neighbor adjacent method. Neural Network is used for classification.

Reetika Verma et. al. [12] describes the surf feature extraction and neural network .This paper demonstrated capability for solving complex problems of character recognition. Preprocessing stage includes noise removal and image enhancement. Surf feature technique and neural network is used for feature extraction and classification. This technique is fast, low cost and more accurate result can be obtained.

In Abdul Rahiman M et. al.[13] proposed a handwritten character recognition system based on vertical and horizontal line positional analyzer algorithm. In preprocessing, median filter is used for noise removal. In segmentation stage, line and character separation are used, which gives isolated character. The features are extracted based on horizontal and vertical line count and position. Decision tree classifier is used for classification. Recognition accuracy obtained 91%.

In Pranchi Mukherji et. al. [14] proposed a Shape feature extraction techniques for handwritten character recognition. Preprocessing includes noise removal using Gaussian filter, binarization using Ostu's method, skeletoniztion. Average Compressed Direction Coding Algorithm for stroke is used for feature extraction method. In classification, Decision Tree classifier is used. 86.4% is the overall recognition accuracy.

In Parikh Nirav Tushar et. al.[15] describes that a Chain Code based handwritten character recognition. In preprocessing stage, binarization and slant correction are used. In feature extraction stage Chain Codes are constructed which form the features of character. ANN is used for classification. Recognition accuracy is 80% is reported. Amritha Sampath et.al [16] proposed the same method of feature extraction.
Comparison between the various literatures that is mentioned in the section is summarized in the following table1.

Table 1: Comparison of Various Geometrical Techniques in HCR

Author	Preprocessing	Segmentation	Feature Extraction	Classification	Recognition Accuracy
J.Pradeep[1]	Noise Removal, Binarization, Edge detection, Dilation and filling.	-	Diagonal feature	Feed Forward Back propagation Neural Network.	98%
S.V.Rajashekararadhya [2]	Noise Removal, Slant correction, Normalization, Thinning.	-	Image centroid zone Zone centroid zone Hybrid centroid zone	Feed Forward Back propagation Neural Network and Nearest Neighbor	99% for Kannada 96% for Tamil 95% for Malayalam
S.L.Mhetre [3]	Colour to gray conversion, Noise Removal, Thresholding, Thinning, Size Normalization.	-	Grid based method; ICZ &ZCZ method	ANN & Matching Score	-
S.V.Rajashekararadhya [4]	Noise Removal, Slant correction, Normalization, Thinning	-	Image centroid zone Zone centroid zone Hybrid centroid zone	SVM	98.65% for Kannada 98.6% for Telugu 96.1% for Tamil 96.5% for Malayalam.
S.V.Rajashekararadhya [5]	Noise Removal, Slant correction, Normalization, Thinning.	-	Zone based angle	SVM	96.05%
Gita Sinha[6]	Binarization, Dilation, Erosion, Noise removal, Normalization.	-	Image centroid zone Zone centroid zone Hybrid centroid zone.	SVM	97.21%
Seema A Dongare[7]	Colour to gray conversion, Noise removal, Binarization.	Line Word Character	Image centroid zone Zone centroid zone Hybrid centroid zone	ANN	-
Gita Sinha [8]	Colour to gray conversion, Noise Removal, Binarization, Filtering operation, Contour smoothing, Skew detection, Skeletonization.	Line Word Character	Image centroid zone Zone centroid zone Hybrid centroid zone	SVM	95.11%
Sandeep Saha[9]	Colour to gray conversion, Binarization.	Image cropping	40-point feature	ANN	-
Sangeetha Sasidharan[10]	Noise Removal, Binarization	Line segmentation using projection profile; Character segmentation of untouched characters, consonants touching to Valli and Chandrakala.	-	-	94.08%
Anita Pal[11]	Skeletonization, Normalization.	-	Fourier descriptors (8-Neighbour Adjacent Method)	Multilayer Perceptron Network	94%
Reetika Verma [12]	Noise Removal	-	Surf feature extraction	Back Propagation Neural Network	-

Abdul Rahiman M[13]	Noise Removal	Line & Character separation	Horizontal and Vertical Line count and position	Decision Tree	91%
Pranchi Mukherji [14]	Noise Removal, Binarization, Skeletonization.	-	Average Compressed Direction Coding Algorithm	Decision Tree	86.4%
Parikh Nirav Tushar [15]	Binarization Slant correction	-	Chain code	ANN	80%

IV. Conclusion

This paper presented a detailed study of offline handwritten character recognition systems developed in different languages. From the literature review it has been analyzed that the recognition accuracy mainly depends on proper selection of feature extraction methods. This work mainly concentrated on geometrical based character analysis methods. The recognition efficiency of Feature vectors are can be improved by selection of appropriate preprocessing methods. It has been also analyzed that in case of character recognition neural network and SVM provide better classification efficiency compared to other classification methods.

References

[1]. J.Pradeep, E.Srinivasan and S.Himavathi, Diagonal Feature Extraction Based Handwritten Character System Using Neural Network, International Journal of Computer Applications (0975-8887), vol.8, no.9, pp.17-22, October 2010.

[2]. S.V.Rajashekararadhya and P.Vanaja Ranjan, Handwritten Numeral/Mixed Numerals Recognition of South-Indian Scripts: The Zone Based Feature Extraction Method, Journal of Theoretical and Applied Information Technology, vol.7, no.1, pp. 063-079, 2005 - 2009.

[3]. S.L.Mhetre and M.M.Patil, A Comparative Study of Two Methods for Handwitten Devanagari Numeral Recognition, IOSR Journal of Computer Engineering, vol.15, pp. 49-53, Nov-Dec.2013.

[4]. S.V. Rajashekararadhya and P. Vanaja Ranjan, Efficient Zone Based Feature Extraction Algorithm for Handwritten Numeral Recognition of Four Popular South Indian Scripts, Journal of Theoretical and Applied Information Technology, pp. 1171-1181, 2005 - 2008.

[5]. S.V. Rajashekararadhya and P. Vanaja Ranjan, Handwritten numeral recognition of Kannada script, Proceedings of the International Workshop on Machine Intelligence Research, pp. 80-86, 2009.

[6]. Gita Sinha and Jitendra kumar, Arabic Numeral Recognition Using SVM Classifier, International Journal of Emerging Research in Management \&Technology, vol.2, pp.62-67, May 2013.

[7]. Seema A.Dongare, Dhananjay B.Kshirsagar and Snehal V. Waghchaure, Handwritten Devanagari Character Recognition using Neural Network, IOSR Journal of Computer Engineering (IOSR-JCE), vol.16, pp. 74-79, Mar-Apr. 2014.

[8]. Gita Sinha, Anita Rani, Renu Dhir and Rajneesh Rani, Zone-Based Feature Extraction Techniques and SVM for Handwritten Gurmukhi Character Recognition, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6, pp. 106-111, June 2012.

[9]. Sandeep Saha, Nabarag Paul, Sayam Kumar Das and Sandip Kundu, Optical Character Recognition using 40-point Feature Extraction and Artificial Neural Network, International Journal of Advanced Research in Computer Science and Software Engineering, vol.3, pp. 495-502, Apr. 2013.

[10]. Sangeetha Sasidharan, Anjitha Mary Paul, Segmentation of Offline Malayalam Handwritten Character Recognition, International Jouurnal of Advanced Research in Computer Science and Software Engineering, vol.3, pp. 761-766, Nov. 2013.

[11]. Anita Pal & Dayashankar Singh, Handwritten English Character Recognition Using Neural Network, International Journal of Computer Science & Communication, vol.1, pp. 141-144, Jul.-Dec. 2010.

[12]. Reetika Verma, Rupinder Kaur, An Efficient Technique for Character Recognition using Neural Network & Surf Feature Extraction, International Journal of Computer Science & Information Technologies, vol.5, pp. 1995-1997, 2014.

[13]. Abdul Rahiman M and M.S. Rajasree, Recognition of Handwritten Malayalam Character using Vertical & Horizontal Line Positional Analyzer Algorithm, in Proc. ICMLC, 2013.

[14]. Prachi Mukherji and Priti.P.Rege, Shape Feature and Fuzzy Logic Based Offline Devanagari Handwritten Optical Character Recognition, Proc. Journal of Pattern Recogntiion Research 4, Jun.2009.

[15]. Parikh Nirav Tushar and Saurabh Upadhyay, Chain Code Based Handwritten Cursive Character Recogntiion System with Better Segmentation Using Neural Network , in Proc. International Journal of Computational Engineering Research, vol.3, May.2013.

[16]. Amritha Sampath, Tripti.C and Govindaru.V, Freeman Code Based Online Handwritten Techiniques for Handwritten Character Recognition for Malayalam using Backpropagation Neural Networks, ACIJ, vol.3, no.4, Jul.2012.