

Offline Handwritten Character Recognition Techniques: A Survey

Mrs. Sapana Shailesh Dhere¹ Prof. K. K. Pandey²

^{1,2}Department of Electronics Engineering

^{1,2}PVPIT, Budhgaon, Sangali, India

Abstract — Handwriting character recognition has been one of the most interesting and challenging research areas in field of image processing and pattern recognition in the recent years. The Biggest challenge in the field of image processing is to recognize documents both in printed and handwritten format. Recognition of handwritten character is a problem since there is a variation in same character due to different types of noises or font size or font shape. An effective segmentation and Feature extraction as well as training and classification techniques are discussed in this paper. Also, highlighted on most important results reported and tried to achieve the beneficial techniques of research till date.

Keywords: Hand Printed Character Recognition, Statistical, Geometrical and Topological Features, SOM Based Classifier

I. INTRODUCTION

Today's world there is a need to convert the analog into digital. Since the introduction of digital scanners after the computer came onto the scene, there has been need to convert books or text into digital media viewable over the internet or on a computer. This is where optical character recognition is open area. It is one of the most challenging areas of pattern recognition. It improves an automation process and interface between human and machine.

Optical Character Recognition (OCR) is a field of exploration in pattern recognition, artificial intelligence and machine vision. It refers to the mechanical or electronic translation of images of handwritten, typewritten or printed text into machine-editable text. Handwritten character recognition is comparatively difficult; as different people have different handwriting shapes styles. So, handwritten OCR is still a subject of active research.

The domain of handwritten text recognition has two completely different problems of On-line and Off-line character recognition. OCR is performed in off-line recognition; it recognizes the character after the writing and printing has been completed where as in on-line recognition the computer recognizes the character as they are drawn. On-line contains more information about writing style of a person such as speed, pressure, angle of a pen, direction and pressure against the system which is called order of the stroke. Both hand printed and printed characters may be recognized, but the performance is directly dependent upon the quality of the input documents. The more constrained the input is, the better will the performance of the OCR system.

An OCR system consists of different phases as data acquisition, pre-processing, segmentation, feature extraction and classification and post processing. Fig.1 shows the major steps involved in OCR systems.

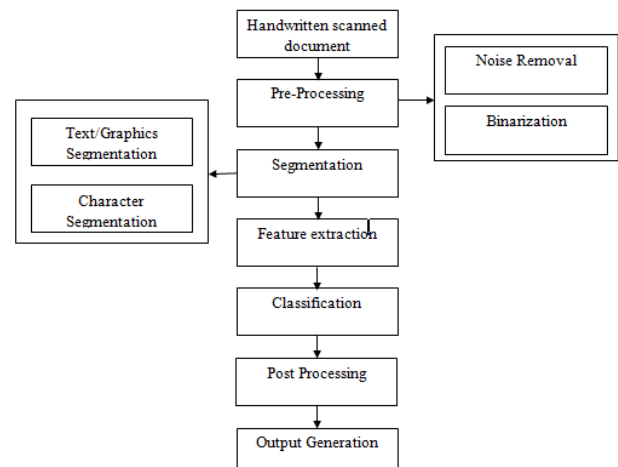


Fig. 1: Major steps involved in the process of OCR

A. Data Acquisition:

In Data acquisition, the recognition system acquires a scanned image as an input image. The image should have a specific format such as JPEG, BMT, etc. This image is acquired through a scanner, digital camera or any other suitable digital input device.

B. Pre-Processing:

The first step in character recognition process is converting the image in to Binary image. Two intensity values are available in binary image. These values are Black and White. We are use zero for Black and one for white. Thus, the color of the character is White and the background is black. The preprocessing techniques are needed on color or grey-level document images containing text and/or graphics. Images may have non-uniform background or water marks making it difficult to recognize hence; the desired result from preprocessing is a binary image containing text only.

Pre-Processing improves a document image preparing it for the feature extraction phase in the OCR system. In order to achieve high recognition rate, prior to character recognition, it is crucial to eliminate the noise and imperfections introduced in the image. Pre-processing covers all those functions of feature extraction to produce an original image.

The step involved in pre-processing are-

C. Binarization-

In binarization the gray scale image is converted into binary image with the help of thresholding. The binary image is represented as 'ones' and 'zeros.' '1' is used to represent object pixel and '0' is used to represent background.

D. Noise reduction-

The noise introduced by the optical scanning device or the writing instrument, causes disconnected line segments, bumps and gaps in lines, filled loops etc. The distortion

including local variations, rounding of corners, dilation and erosion, is also a problem.

E. Thinning-

Thinning is operation that is used to remove selected foreground pixels from the binary images and thin the images to single-pixel width

F. Edge Detection, Dilation, and Filling-

The boundary detection of image is done to enable easier subsequent detection of relevant features and objects of interest. After locating the edges, the image is dilated and the holes present in the image are filled.

G. Segmentation:

Segmentation is a process that determines the ingredients of an image. It is necessary to locate the regions of the document where data have been printed and distinguish them from figures and graphics.

The segmentation is divided into

- Line Segmentation – to separate the text line from image line segmentation is used.
- Text Segmentation – Word segmentation provides the space between two words.
- Character Segmentation – It provide the spacing between two characters.

Yonghong Song *et al.* [13] proposed text patch segmentation method. Text patches in document image are segmented based on an improved connected component analysis algorithm.

Shankar Mathur, *et al.* [15] used novel algorithm for segmentation. This is capable of extracting handwritten words from the input image and carries out segmentation of the selected word to generate vectors for individual character of a word.

Rakesh Kumar Mandal, N.R. Manna [16] recommended column-wise segmentation of 10x10 matrix.

H. Feature extraction:

The objective of feature extraction phase is to extract the essential and differentiable characteristics of the symbols. Feature space is much less than input image space as we extracted the features of the characters that are crucial for classifying them at recognition stage. This is an important stage as its functioning improves the recognition rate and reduces the misclassification. Different feature extraction methods are Statistical features, Geometrical features, Topological features, Directional features [1], Neighborhood foreground pixel density [2], Surf features [6] etc.

In the approach the grouping of Statistical features, Geometrical features, Topological features are implemented to uniquely categorize each character.

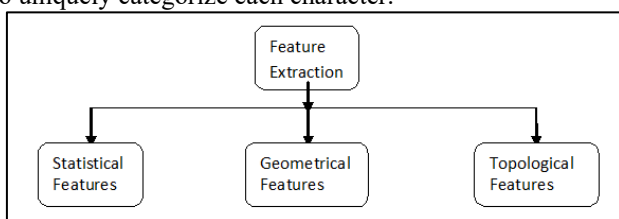


Fig. 2: Different features of image

1) Statistical Features-

The character image is divided into zones, taking data from each zone as a feature of the image. Statistical feature extraction technique is applied by calculating the percentage of black pixel in each zone. The zones are chosen to cover most of the region in an image which having information.

2) Geometrical features-

Geometrical features are obtained after performing Thinning and normalization on each character image. Center of skeleton was calculated. Then image was divided into 3x3 zones. In each zone geometrical features were calculated by using-

- Calculating average distances of each pixel present in zone from center point.
- The average of angle of each pixel present in zone from center point.

3) Topological features-

Topological features (End points and Transitions) are extracted after thinning of image and deriving end points with respect to the zone. These features have been used to eliminate confusions among various characters at the stage of post-processing.

I. Classification

The classification stage is decision making part of the OCR system. In this stage extracted features are used to recognize character. The classification of relevant features is done by using different classifier these are Artificial Neural Network, MLP NN with back propagation, SVM, SOM etc.

II. LITERATURE REVIEW

This section gives a review of offline handwritten character recognition system. A brief sketch of preprocessing, feature extraction and classification techniques for several works are summarized here:

Nisha Sharma *et al.* [1] taken care of off-line character recognition of hand printed document images containing English Characters-Uppercase and Lowercase, Numerals and Special Characters. Statistical, Geometric, Topological and Directional Feature Extraction techniques have been applied over segmented character image. Classification was done using Multilayer perception neural network (NN) with back propagation and Support vector machine (SVM) classifier. SVM was found outperform MLP back propagation neural network from the result. Thus, it can be concluded that recognition rate depend on the type of feature extraction is used. Recognition efficiency obtained for combined data set is 92.167%. The author suggested that to explore new or hybrid feature extraction methods which can be used for training the NN and SVM so that the efficiency of system can be improved.

Shilpy Bansal *et al.* [2] Proposed a feature extraction technique named as Neighborhood Foreground Pixels Density technique authors used a dimensionality reduction technique namely Principal component analysis (PCA). Principal component analysis includes a mathematical procedure which transforms a number of correlated features into a number of uncorrelated features. It means it. As there could be some insignificant feature values so to reject those eliminate insufficient features and left only sufficient features

for character recognition. so we get the maximum possible recognition rate in less time. Support Vector Machine (SVM), Naïve Bayes classifier and Multilayer Perceptron classifier three types of classifiers are used. Recognition efficiency obtained is 91.95% with SVM, 87.3% with MLP, 77.7% with Naïve Bayes.

Ankit Sharma *et.al.*[3] used feed forward back propagation Neural Network to classify the characters. The ANN is trained using back propagation algorithm. NN followed by the back propagation algorithm which support in training NN. 85% accuracy is achieved by using the Neural network classifier.

Ashok Kumare *et.al.*[4] Described neural network technique to recognize the offline handwritten characters. Fourier descriptor is used for feature extraction. Characters are recognized through Feed Forward Multi-Layer Perceptron Network (MLPN) with one hidden layer. Back-propagation algorithm has been used to train the neural network. The experimental result shows that the improved BP can speed up convergence of training process. The author suggest that the system become adaptive if it learns the different pattern of same character under the same label.

Prasad. P. Chaudhari *et.al.* [5] Recommended Grid approach for recognition of offline handwritten character. In this technique first extract feature from pattern and then extracted features are used to train neural network. The proposed work is based on the pattern matching. A multilayer feed forward network is used for classification. The average success rate of algorithm in recognition is 96.9%.

Reetika Verma *et.al.*[6] Mentioned surf feature extraction, Surf feature matching and Neural network techniques for character. SURF adopts nearest neighbor. The purposed of feature point matching is to find up the feature point from the same location in two images and matches a couple of feature points. In this paper it presents a fast-matching method. It detected the feature points to match the selected part. The better accuracy is obtained by combining NN and SURF.

Theingi Htike *et. Al* [7] used Competitive Neural Trees (CNeT). Shape features descriptors are extracted from processed image which are used in CNeT. The trained dataset are 660 and testing datasets are 330. Recognition accuracy rate is 97% With testing dataset 330.

Sandeep Saha *et.al* [8] presented 40-point feature extraction and Artificial Neural Network. Training set consist 780 images and 260 images testing set. Classification is done using a Neural Network (MLP: multi-layer perceptron). MLP consists of 40-elements feature vector for each character and 26 outputs for alphabets. The overall accuracy obtained is 83.84%

Recognition of English handwritten characters without feature extraction using multilayer feed forward Neural Network is described by J. Pradeep *et.al.* [9]. In the proposed system each character is resized into 30x20 pixels, which is used to train the neural network. Each resized character has 600 pixels and each pixel value are used as a feature for training the neural network. Recognition accuracy of this technique is 90.19%. This method is less complex compared to the feature extraction method.

Amit Goyal *et.al.* [10] Proposed learning of alphabets using Kohonen's self-organizing Neural Network.

Uses of self-organizing map for feature extraction of alphabets, thus the relative feature of them are automatically defined based on arrangement of the competing units. Proposed data structure and algorithm reduced the SOM computational complexity by several orders of magnitude. Kohonen maps or SOMs are one of the most popular learning strategies among the several Artificial Neural Network algorithms. The result of this system shows better recognition performance in terms of accuracy and speed as compared to NN. A general trend of increase in performance with increase in samples is observed. The author suggests to explore the necessity to create optimization algorithm to increase the performance of the system.

SupriyaDeshmukh *et.al.*[11] suggested two feature extraction methods based on directional features. This method is used for English alphabets and also for Marathi vowels. The first method uses stroke distribution of a characters and the second method uses counter extraction. The two directional features are compared with two different correlation techniques separately. First correlation technique calculates the dissimilarity between reference pattern and test pattern and the other calculates the similarity between reference pattern and test pattern. Direct template matching technique is used as a classifier. The author concluded that the stroke length method gives good performance for character that having straight lines and counter method performs good for the character that having curves.

Rachana R. Herekar *et.al.* [12] Described Feature extraction using zoning method together with the concept of euler number. Zoning is based on two types of topologies. First is Static Topology and second is Adaptive topology. Zoning based methods are able to find the local characteristics instead of global characteristics. Concept of euler number is helps to classify the characters. Euler number is the difference between number of characters in the image and the number of holes present in the characters. Dataset consist of total 1550 number of samples of Uppercase alphabets, lowercase alphabets and numbers. Three sets of each Uppercase, lowercase and numbers are tested for recognition. 91% accuracy is achieved by using this method.

Yonghong Song *et.al* [13] defined a novel method for extraction handwriting characters from Multilanguage document images. Firstly, text patches in document image are segmented based on connected component analysis algorithm. Generic algorithm is applied for feature fusion and patch type classification. Finally, the Markov Random Field model is utilized as a post processing step to further correct the misclassification of text patch type. 100 images of multiple languages are used with 25 images of each languages Chinese, English, Japanese and mixed languages. Accuracy of the system for Chinese is 98.09%, for English 99.25%, for Japanese 97.58% and for Multiple language its 97.10%.

Dileep Kumar Patel *et. al.* [14] Presented DWT (Discrete wavelet transform) for features extraction. DWT used with appropriate level of multi-resolution technique and then each pattern class is characterized by a mean vector. Distance from input pattern vector to all the mean vectors are computed by EDM (Euclidean distance metric). 3900 characters of samples are collected; 2600 characters are used to train the proposed HCR system and 1300 characters are used for training purpose. When average recognition

accuracy is optimal for particular level of multi resolution and appropriate resolution of character image then accuracy is 90%. Then, any further increment in level of multi resolution result in the decrease of the average accuracy. The resolution of input character image and the level of multiresolution are dependent upon each other so the challenge is to find out the exact relation between them.

ShashankMathuret.al.[15] Mentioned combination of artificial neural network and genetic algorithm. In this method the segmented character is converted into column vectors of 625 values that are used to fed into advance neural network. The network has been designed with 625 input and 26 output neurons corresponding to each character A-Z. This output for all four networks is fed into the genetic algorithm. The genetic algorithm which is applied is accepts the output from the four artificial neural networks. Optimized recognized output is 71%.

Rakesh Kumar Mandalet.al. [16] Proposed the compressed column wise segmentation of Image matrix (CSIM) using Neural network. The input Image matrix is compressed into a lower dimension matrix in order to reduce non-significant elements of the image matrix. The compressed matrix is segmented column wise. Each column of particular image matrix is mapped to identical patterns for recognizing a particular character. The result shows CMIS training allows very fast convergence. The number of epochs required for the training is very less. The compression of the input matrix helps to improve the performance of CMIS.CMIS is better as compared to other segmentation methods. Column wise segmentation is better than row wise segmentation.

SubhashPanwaret.al [17] used bottom-up grouping approach for segmentation. Also used a novel connectivity strength parameter with depth search approach for extraction of connected components of the same line. The cursive stroke sequencing technique is also used. The proposed technique is tested on the IAM database. The accuracy of proposed method is 98%.

Isha Vatset.al. [18] proposed classical technique Template matching. This technique is also called as correlation template matching. In this work filtration is performed using Weiner filter. Correlation matrix is used for feature extraction. Extracted features are used for number recognition. Efficiency rate of proposed work is 97%.

Unconstrained handwritten word recognition using a combination of Neural networks is proposed by Rodolfo Luna-Perez [19]. In this paper author proposed a novel method for classification based on three components: a self-organizing map (SOM) for non-supervised classification, a simple recurrent network (SRN) for temporal classification and a multi-layer perceptron (MLP). The system was tested using a lexical with 10 words taken from database IAM. Training set contained 20 images of each word and testing set consists 5 images of each word. The overall result of proposed work is 78.2%.

III. COMPARATIVE ANALYSIS

In this section, we analyze the result from different techniques. We studied the different methods of segmentation, feature extraction and classification. The

below table shows the performance of previous work related to handwritten character recognition.

Sr.No	Feature extraction Technique	Classification Method	Accuracy
1	Statistical, Geometric, Topological and Directional Feature Extraction techniques	SVM	92.8%
		Neural Network	89%
2	Neighborhood Foreground Pixels Density technique	SVM	91.95%
		MLP	87.3%
		Naïve Bayes classifier	77.7%
3	surf feature point matching	Neural Network	98.77
4	surf feature extraction	Competitive Neural Tree	97%
5	40-point feature extraction	Artificial Neural Network	83.84%
6	without feature extraction	multilayer feed forward Neural Network	90.19%
7	Two feature extraction method: strock distribution, counter extraction	Direct templet matching	83%
8	Discrete wavelet transforms	Euclidean distance matric	90%

Feature extraction plays very important role in character recognition. The feature space is much less than the input image space as we extract only essential properties for higher recognition rate. The combinations of statistical, geometrical and topological features are used to identify each character uniquely and to avoid the confusions in post processing.

Classification is the process of identifying each character and assigning to the correct character class. The efficiency of system is depending on the classification technique. As we studied the verity of classification techniques, the SOM is better than all other classification technique. SOM learn on their own through unsupervised competitive learning. It is modifying the algorithm itself hence it shows better performance in terms of accuracy and speed.

IV. CONCLUSION

This Paper represents detailed survey of different feature extraction and classification algorithms used for character recognition. From this survey it can be observed that the accuracy of recognition system depends on methods used for feature extraction and classification. This compressive discussion will provide insight into the concepts involved in this area. Further work can explore the necessity to integrate into an optimization algorithm to further enhance the performance of the system.

REFERENCES

- [1] Nisha Sharma, Bhupendra Kumar, Vandita Singh, "Recognition of Off-line Hand printed English Characters, Numerals and Special Symbols" 978-1-4799-4236-7/14/\$31.00_c 2014 IEEE, pp. 640-645
- [2] Shilpy Bansal, Mamta Garg, Munish Kumar, "A Technique for Offline Handwritten Character Recognition" IJCAT International Journal of Computing and Technology, Volume 1, Issue 2, March 2014 ISSN: 2348 – 6090, pp. 210-215
- [3] Ankit Sharma, Dipti R Chaudhary, "Character Recognition Using Neural Network" International Journal of Engineering Trends and Technology (IJETT) - Volume 4 Issue 4- April 2013, pp 662-667
- [4] Ashok Kumar, Pradeep Kumar Bhatia, "Offline Handwritten Character Recognition Using Improved Back-Propagation Algorithm" International Journal of Advances in Engineering Sciences Vol.3 (3), July, 2013 e-ISSN: 2231-0347 Print-ISSN: 2231-2013
- [5] Prasad. P. Chaudhari, K.R. Sarode "Offline Handwritten Character Recognition by using Grid approach" International Journal of Application or Innovation in Engineering & Management (IJAIEEM), ISSN 2319 – 4847, Volume 3, Issue 4, April 2014
- [6] Reetika Verma, Mrs. Rupinder Kaur, "Enhanced Character Recognition Using Surf Feature and Neural Network Technique" International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 5 (4), 2014, 5565-5570, ISSN:0975-9646
- [7] Theingi Htike and Yadana Thein, "Handwritten Character Recognition Using Competitive Neural Trees" IACSIT International Journal of Engineering and Technology, Vol. 5, No. 3, June 2013, pp.352-356
- [8] Sandeep Saha, Nabarag Paul, Sayam Kumar Das, 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 (IJARCSSE) Volume 3, Issue 4, April 2013 ISSN: 2277 128X
- [9] J. Pradeep, E. Srinivasan, S. Himavath, "Neural Network based Handwritten Character Recognition system without feature extraction" International Conference on Computer, Communication and Electrical Technology – ICCET 2011, 18th & 19th March, 2011, 978-1-4244-9394-4/11/\$26.00 ©2011 IEEE
- [10] Amit Goyal, Ankita Lakhanpal and Shaveta Goyal, "Learning of alphabets using Kohonen's Self organized featured map" International Journal of Application or Innovation in Engineering & Management (IJAIEEM) Volume 2, Issue 12, December 2013 ISSN 2319 – 4847
- [11] Supriya Deshmukh, Leena Ragha, "Analysis of Directional Features - Stroke and Contour for Handwritten Character Recognition" 978-1T-4244-1888-6/08/f\$25.00 Q 2008 IEEE. pp.114-118
- [12] Rachana R. Herekar, Prof. S. R. Dhotre, "Handwritten Character Recognition Based on Zoning Using Euler Number for English Alphabets and Numerals" e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 16, Issue 4, Ver. III (Jul - Aug. 2014), PP 75-88
- [13] Yonghong Song, Guilin Xiao, Yuanlin Zhang, Lei Yang, Liulu Zhao, "A Handwritten Character Extraction Algorithm for Multi-language Document Image", 1520-5363/11 \$26.00 © 2011 IEEE DOI 10.1109/ICDAR.2011.28
- [14] Dileep Kumar Patel, Tanmoy Som, Sushil Kumar Yadav, Manoj Kumar Singh, "Handwritten Character Recognition Using Multiresolution Technique and Euclidean Distance Metric" Journal of Signal and Information Processing, 2012, 3, 208-214
- [15] Shashank Mathur, Vaibhav Aggarwal, Himanshu Joshi, Anil Ahlawat, "OFFLINE HANDWRITING RECOGNITION USING GENETIC ALGORITHM" International Book Series "Information Science and Computing", pp.21-27
- [16] Rakesh Kumar Mandal, N R Manna, "Hand Written English Character Recognition using Column-wise Segmentation of Image Matrix (CSIM)", E-ISSN: 2224-2872, Issue 5, Volume 11, May 2012
- [17] Subhash Panwar, Neeta Nain, "Cursive Stroke Sequencing for Handwritten Text Documents Recognition"
- [18] Isha Vats, Shamandeep Singh, "Offline Handwritten English Numerals Recognition using Correlation Method" International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 3 Issue 6, June – 2014.
- [19] Rodolfo Luna-Perez, Polar Gomez-Gil, "Unconstrained Handwritten Word Recognition Using a Combination of Neural Network" ISBN:978-988-17012-0-6, ISSN: 2078-0966, WCECS2010, San Francisco, USA