International Journal of Students' Research in Technology & Management
eISSN: 2321-2543, Vol 10, No 3, 2022, pp 06-11

Off-line English character recognition system

Binod Kumar Prasad[1*], Om Prakash Dubey[2]
[1*]Guru Nanak Institutions Technical Campus, Hyderabad, T.S., India; [2]Veer Kunwar Singh University, Ara, Bihar, India.
Email: [1*]binod.krpr@gmail.com, [2]omprakashdubeymaths@gmail.com

Keywords
Hidden Markov Model, Sobel Masks, Gradient Features, Zonal Features, Median Filters, Preprocessing.

Article History
Received on 22nd May 2022
Accepted on 1st July 2022
Published on 25th July 2022

Cite this article
Prasad, B. K., & Dubey, O. P. (2022). Off-line English character recognition system. *International Journal of Students' Research in Technology & Management, 10*(3), 06-11. https://doi.org/10.18510/ijsrtm.2022.1032

Abstract

Purpose of the study: This paper aims to recognition of handwritten English characters in offline mode. It develops an efficient character recognition model avoiding large variations in handwriting by using better feature extraction techniques.

Methodology: The samples of characters are preprocessed by applying a sequence of operations in succession like Thickening, Thresholding, Filtering, and Thinning. Efficient features like Gradient features and Zonal features have been extracted. Gradient features are helpful to find out stroke information in the character whereas Zonal features detail out local information in a more précised way. Hidden Markov Model is the classifier.

Main Findings: Classification has been started with only a 5-state HMM model but it is observed that as the number of states of HMM model is increased, the corresponding recognition rate is also improved. Finally, with the 36 states HMM model we have got the expected result. This produces an overall average recognition rate of 92.6%. For the letters 'A' and 'W', the recognition rate is found to be very low, because of a lot of variations in writing style of these letters.

Applications of this study: HMM is a flexible tool which is capable of absorbing variations in character images. The future works will be concentrated on improvement of recognition rate of such letters by finding some demarcating features and post processing**.** The proposed method can be well used in Natural Language Processing, Signature verification, Face recognition like other Pattern Recognition applications.

Novelty/Originality of this study: Preprocessing uses Median filter which removes all stray marks in samples and hence avoids any possibility of false pixels. The combination of Gradient features and Zonal features leads to a recognition accuracy of 92.6% which may be used by researchers in any other domains for the purpose of classification. The application of HMM will motivate the readers to use it for better results of classification.

INTRODUCTION

Off-line handwriting recognition (OHR) continues to be an active area for research towards exploring the newer techniques that would improve recognition accuracy because several applications including Gmail sorting, bank processing, document reading, and postal address recognition require offline handwriting recognition systems. Character recognition is nothing but a Machine simulation of human reading. Dave et al.(2017) have presented a handwritten English character recognition system using a multilayer perceptron classifier. They have converted the original image into a grayscale image and split characters from words using segmentation. They have achieved recognition accuracy of more than 70% of handwritten English characters. Pal et al.(2010) have developed automatic English handwritten characters recognition system. They have used Fourier Descriptor to extract features and back propagation network for classification. The recognition accuracy of 94% has been claimed with 36 hidden nodes. Mathur et al. (2019) have proposed an integrated system of OCR and text-to-speech conversion to help visually impaired people to read a document. The system uses OCR and text-to-speech modules of mobile phones of the new generation. Dhande (2018) have presented a character recognition system for cursive English handwriting in a medical prescription to read the names of medicines. It uses the horizontal projection method for text-line segmentation and the vertical projection histogram method for word segmentation. Feature extraction has been done by means of a convex hull algorithm, whereas classification has been done with SVM. The system is reported to have an accuracy of 85%. Chen et al.(2016) have combined Conditional Random Fields with deep learning to recognize handwritten words. Deep features are learned and sequences are labeled in a unified framework. Deep structure is pre-trained with stacked restricted Boltzmann machines and optimization of the entire network has been done with an online learning algorithm. Celar (2015) have used structural characteristics as global features and local descriptor based on Scale–Invariant Feature Transform (SIFT) which is invariant to scaling, rotation, and even affine transformation to recognize digits in student ID. These features are organized by the Bag of Words (BoW) model. They tested the MNIST test database with MLP to achieve a correct recognition rate of 94.76%. It is now a well-established fact that the direction of character strokes contains huge important information for character recognition. If we can precisely describe that strokes in certain directions occur at certain positions in the character image, the character will be easily categorized. Many statistical features used in character recognition are designed according to this idea (Liu 2005). Previous researchers demonstrated that among direction features, the gradient features outperform various other directional features. That is why we have given due stress on Gradient features for a character image. The tool to train the system with the obtained feature vectors is taken to be HMM because systems based on HMM have been shown to outperform segmentation-based approaches (Rabiner, L. R. (1998), Mokbel, C., Abi Akl H. and Greige H. (2002), El-Hajj, R., Likforman-Sulem, L., and Mokbel, C.(2005), El Abed, H., Margner V.(2009)).

PROPOSED MODEL

Character images are preprocessed to filter out any noise elements present in them. Features are extracted and HMM model for each character has been trained by the sequence of the feature vectors. To test a handwritten character image, we extract the similar features using the same procedure as earlier and the corresponding sequence (observation) is processed with each HMM model P(O/λ), probability of the observation sequence (O) by the models (λ). Probabilities are compared and the highest probability concludes the highest matching of the features with the corresponding model (Figure 1).

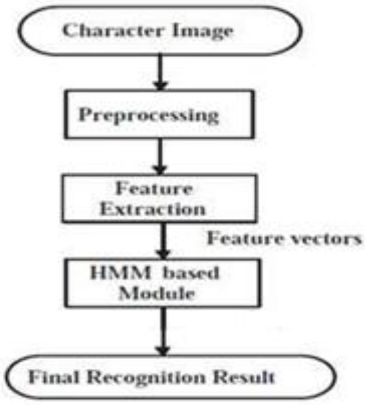

Figure 1: System Overview

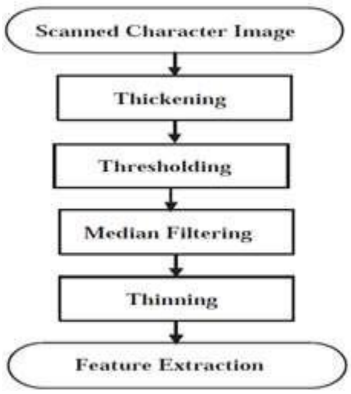

Figure 2: Block Diagram for Pre-processing

PRE-PROCESSING

Any image processing application suffers from noise like isolated pixels. This noise gives rise to ambiguous features which result in poor recognition rate or accuracy. Therefore a preprocessing mechanism has been executed before we could start with feature extraction methods. Here a sequence of operations is carried out in succession as shown in the flow diagram [Figure 2]. We have used the median filter for its better performance to get rid of unwanted marks or isolated pixels. Thinning is performed to get the skeleton of the character image so that strokes could be clear.

FEATURE EXTRACTION METHODS

Feature extraction is an important part of any type of pattern recognition. A better feature extraction method may yield a better recognition rate by a given classifier. Therefore, much attention is paid to extracting the suitable features from the preprocessed images. Our feature extraction process consists of

Gradient Features

To compute the density of line segments in the quantized direction we use only two masks - The horizontal Sobel mask and the Vertical Sobel mask. The magnitude and phase of the gradient obtained by Sobel masks are calculated as below

eISSN: 2321-2543, Vol 10, No 3, 2022, pp 06-11

$$\text{Magnitude: } M(x,y) = \sqrt{[S_H^2\ (x,y) + S_V^2\ \ (x,y)]} \qquad (1)$$

$$\text{Phase: } \quad \phi(x,y) = \tan^{-1}\frac{S_H(x,y)}{S_V\ \ (x,y)} \qquad (2)$$

The phase is quantized in eight directions as shown in Figure 3. For each quantized phase value, corresponding magnitudes are added to get the total strength in that direction. To get the feature within the finite number of symbols, magnitudes are normalized and quantized. Finally, we consider four global gradients (G) features combining the following pairs-(0°,±180°), (45°,−135°), (90°,−90°), and (135°,−45°).

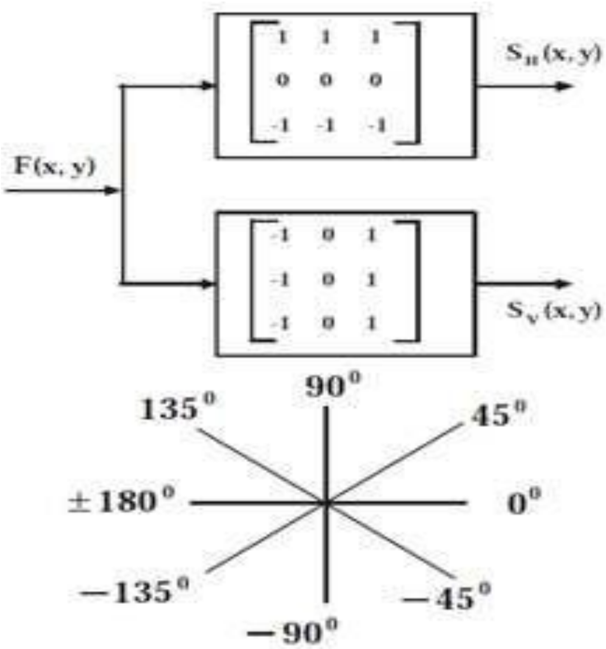

Figure 3: Gradient Feature Extraction using Sobel Masks

Zonal features

The whole image is divided into 25 zones by using the non-overlapping window of size 30×30. Features are extracted in two ways- First, the feature is extracted from each zone by finding the total number of white pixels of the zone and then dividing this number by the total number of pixels of the zone.

$$F_n = \frac{1}{N}\sum_{i=0,j=0}^{10,10} I_{ij} \ \ [\text{for } n^{th} \text{ zone}] \qquad (3)$$

The second set of features is extracted as vector distance. The vector distance of each white pixel of a zone is calculated from the element at the bottom left corner of the same zone. All such distances are summed up and finally, the normalized vector distance is calculated by dividing the sum by the sum of all such distances for all pixels in the zone from the same reference.

$$D_{i,j} = \sqrt{i^2 + j^2} \qquad (4)$$

$$F_n = \frac{\sum_{I_{i,j}=1} D_{i,j}}{\sum_{I_{i,j}=0,1} D_{i,j}} \ \ [I_{i,j} = \text{intensity at pixel position } (i\ ,\ j)] \ (5)$$

To find the global feature, the same process is followed by taking the whole image as a single zone. Therefore, our final observation sequence contains 52 observations obtained by global and local Zonal feature extraction methods:

$F(ZF) = [Z_G\ (2)\ Z_L\ (50)]$

RECOGNITION BY HIDDEN MARKOV MODEL

Hidden Markov Model (HMM) is a finite state machine in which a sequence of observations (O) is produced by this model but the corresponding sequence of states remains hidden within this model. This HMM model can be defined as

$$\lambda = (\pi, A, B) \qquad (6)$$

where π is the initial state probability vector, A is the final state transition probability matrix and B is the final observation probability matrix. The HMM model was initially used for speech recognition purposes, but later it has been proved that the HMM model can be efficiently utilized for other recognition processes like character recognition, pattern recognition, etc. In this paper, I have used a closed left-to-right chain HMM model for handwritten English character recognition. A sketch of 5 states HMM model is shown in Figure 4.

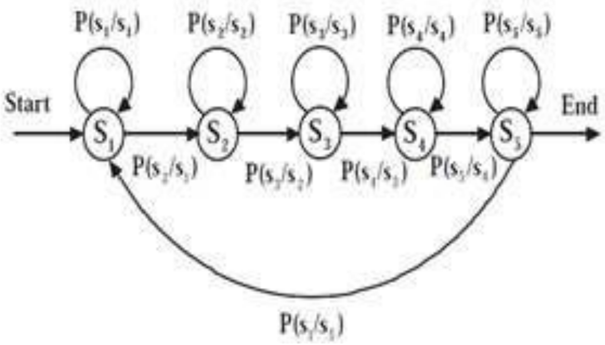

Figure 4: Left to Right Chain HMM Model with 5 States

I have used the Baum-Welch algorithm to train the HMM using the observation sequence obtained from the feature vectors. At the end of the training process, we obtain the final value of A and B which is used for recognition purposes. The Viterbi decoding algorithm has been applied to decode the sequence of states of the HMM model λ for the sequence of observation O and it returns P (O/λ), the probability of generating the given sequence O by the HMM model λ.

RESULTS/FINDINGS

A total of 13000 samples are collected from 100 persons. Each writer wrote 5 sets of A-Z characters. Each character image is converted to a fixed size of 150×150 pixels. We have applied our feature extraction method to these samples and then these feature values are quantized and encoded to the eleven symbols in order to create sequences of observation symbols. For our experiment, I have started with only 5 state model but I observed that as the no. of states of HMM model is increased, the corresponding recognition rate is also improved. Finally, with 36 states HMM model I have got the expected result as shown in table. In Table 1, I have shown final recognition rate of character recognition system for each of 26 letters of the English alphabet. This produces an overall average recognition rate of 92.6%. It shows the effectiveness of the proposed model.

Table 1: Recognition Rate Using Proposed HMM Model

Character	Recognition Rate (%)	Character	Recognition Rate (%)
A	89.8	N	93.5
B	91.2	O	92.5
C	93.7	P	91.1
D	93.4	Q	92.4
E	93.0	R	91.7
F	94.2	S	91.0
G	92.3	T	93.2
H	92.8	U	93.7
I	95.0	V	92.0
J	93.6	W	90.01
K	93.0	X	93.2
L	93.9	Y	91.5
M	92.4	Z	94.4

The results produced by this system have been compared with those in Pal, A and Singh, D.(2010). Here, authors used Multilayer Perceptron with number of hidden neurons 12, 24 and 36 for recognition of handwritten English characters and applied on their self-collected data to achieve average recognition accuracy of 89%, 94% and 94% respectively. The bar diagram in Figure 5 shows the pictorial comparison of results obtained in these papers.

Table 2: Comparison of results

Paper	Classifier Description	% Accuracy
Pal et al Ref. [2]	MLP with 12 hidden neuron	89
	MLP with 24 hidden neuron	94
	MLP with 36 hidden neuron	94
Proposed	Hidden Markov Model	92.6

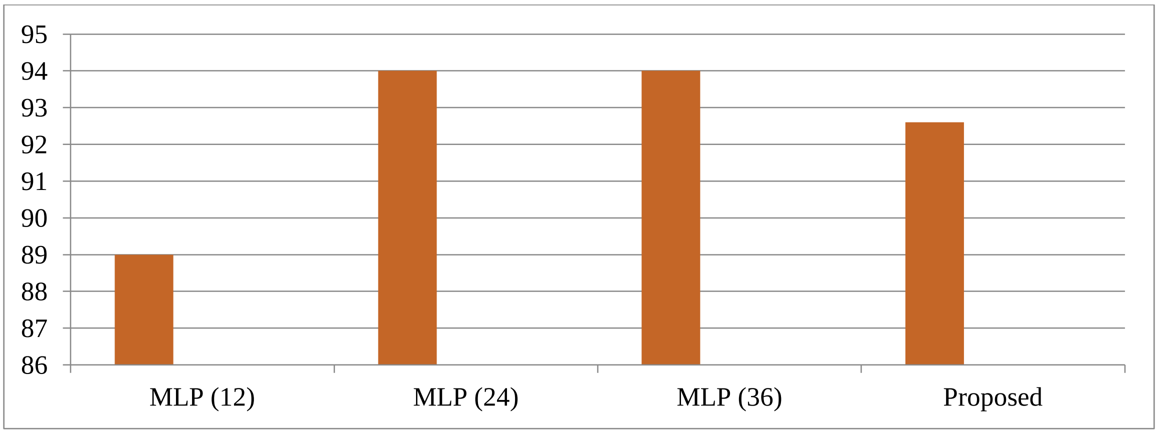

Figure 5: Comparison of recognition accuracy

CONCLUSION

In this paper, an approach has been made to increase the rate of recognition of handwritten characters by finding Gradient and Zonal features along with HMM as a classifier to obtain anaverage accuracy of 92.6%.

LIMITATIONS AND STUDY FORWARD

For the letters 'A' and 'W', the recognition rate is found to be very low, because of a lot of variations in the writing style of these letters. HMM is a flexible tool that is capable of absorbing variations in character images. So, my future works will be concentrated on the improvement of the recognition rate of such letters by finding some demarcating features and post-processing.

ACKNOWLEDGEMENT

I thank the authority of Guru Nanak Institutions Technical Campus, Hyderabad to provide me with all the research facilities to prepare this research paper.

AUTHORS CONTRIBUTION

The execution of experiments and paper writing work have been done by the sole author.

REFERENCES

1. Celar, S., Stojkic, Z., Seremet, Z.(2015). Classification of Test Documents Based on Handwritten Student ID's Characteristics. *25th DAAAM International Symposium on Intelligent Manufacturing and Automation, DAAAM, Procedia Engineering (Elsevier)*, 782 – 790. https://doi.org/10.1016/j.proeng.2015.01.432
2. Chen, G., Li, Y., Srihari, S.(2016). Word recognition with deep conditional random fields. International *Conference on Image Processing (ICIP), IEEE*, 1928-1932. https://doi.org/10.1109/ICIP.2016.7532694
3. Dave, H., Patel, M.(2017). Handwritten English Character Recognition using Multilayer Perceptron Neural Network. *IJARIIE*, *3*(3), 713-719.
4. Dhande P.S., Kharat, R.(2018). Character recognition for cursive English handwriting to recognize medicine me from doctor's prescription, *International conference on computing, communication, control and automation (ICCUBEA)*, Pune, India, IEEE. https://doi.org/10.1109/ICCUBEA.2017.8463842
5. Ding, K., Liu, Z.B., Jin, L.W., Zhu, X.H. (2007). A comparative study of Gabor Feature And Gradient Feature For Handwritten Chinese Character Recognition, *ICWAPR*, Beijing, China, 1182-1186.
6. El Abed, H., Margner V.(2009). Arabic handwriting recognition competition. *Inter. Journal on Document Analysis andRecognition*, 9(2). https://doi.org/10.1109/ICDAR.2009.284
7. El-Hajj, R., Likforman-Sulem, L., and Mokbel, C.(2005). Arabic Hand- writing Recognition Using Baseline Dependent Features and Hidden Markov Modeling, Proc. *Eighth Intl Conf. Document Analysis and Recognition*, 893-897. https://doi.org/10.1109/ICDAR.2005.53
8. Fujisawa, H., Liu, C.L. (2003). Directional Pattern Matching for Character Recognition, Proc. *7th ICDAR*, Edinburgh, Scotland, 794-798.
9. Liu, C.L., Nakashima, K., Sako., Fujisawa, H.(2004). Handwritten Digit Recognition: Investigation of Normalisation and Feature Extraction Techniques, *Pattern Recognition*, *37*(2), 265-279. https://doi.org/10.1016/S0031-3203(03)00224-3
10. Liu, H., Ding, X.(2005). Handwritten Character recognition Using Gradient feature and Quadratic Classifier with Multiple Discrimination Schemes, Proceedings of the *Eighth International Conference on Document Analysis and Recognition(ICDAR*).
11. Mathur, A., Pathare, A., Sharma, P., Oak, S. (2019). AI based reading system for blind using OCR, *International conference on electronics, communication and aerospace technology (ICECA)*, Coimbatore, India, IEEE. https://doi.org/10.1109/ICECA.2019.8822226

12. Mokbel, C., Abi Akl H. and Greige H. (2002). Automatic speech recognition of Arabic digits over Telephone network, Proc.*Research Trends in Science and Technology*.
13. Pal, A, Singh D.(2010). Handwritten English Character Recognition Using Neural Network, *International Journal of Computer Science & Communication*, *1*(2), 141-144.
14. Rabiner, L. R. (1998). A tutorial on Hidden Markov Models and selected applications in speech recognition, *Proceedings of The IEEE*, *77*(2), 257-286. https://doi.org/10.1109/5.18626