ISSN 2250-3153

A Model Approach to Off-line English Character Recognition

Binod Kumar Prasad*, Goutam Sanyal**

*Department of Electronics and Communication Engineering,
Bengal College of Engineering andTechnology, Durgapur, INDIA
(binod.krpr@rediffmail.com)

**Department of Computer Science and Engineering, National Institute of Technology, Durgapur, INDIA
(nitgsanyal@gmail.com)

Abstract- Recognition rate of handwritten character is still limited due to presence of large variation of shape, scale and format in hand written characters. A sophisticated hand written character recognition system demands a better feature extraction technique that would take care of such variation of hand writing. In this paper, we propose a recognition model based on Hidden Markov Models (HMMs) supported by novel feature extraction technique. Ultimately post-processing is done to enhance the recognition rate further.
A data-base of 13000 samples is collected from 100 writers written five times for each character. 2600 samples have been used to train HMM and the rest are used to test recognition model. Using our proposed recognition system we have achieved a good average recognition rate of about 93.24 percent.

Index Terms- Hidden Markov Model, Sobel masks, gradient features, curvature features and projected histogram.

I. Introduction

The 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 application including mail sorting, bank processing, document reading and postal address recognition require offline handwriting recognition systems. Character recognition is nothing but Machine simulation of human reading [1], [2]. It is also known as Optical Character Recognition. It contributes immensely to the advancement of an automation process and can improve the interface between man and machine in numerous applications. Several research works have been focussing on new techniques and methods that would reduce the processing time while providing higher recognition accuracy. The methods of Character recognition have grown up sequentially [3], [4]. The recognition of isolated handwritten character was first investigated [5], but later whole words [6] were addressed. Most of the systems reported in literature until today consider constrained recognition problems based on vocabularies from specific domain e.g., the recognition of handwritten check amounts [7] or postal address [8]. Free handwritten recognition, without domain specific constraints and large vocabularies was addressed only recently in a few papers [9], [10]. The recognition rate of such system is still low and there is a need of improvement [11].It is now a well established fact that the direction of character strokes contain vast 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.[12]Previous researchers [13][14] demonstrated that among direction features the gradient features [15] outperform various other directional features.
That's why we have given due stress on Gradient features by finding it both globally and locally for a character image.

The tool to train the system with the obtained feature vectors is taken to be HMM because OHR systems based on HMM have been shown to outperform segmentation based approaches [16]-[19]. With the usage of HMM models for the pattern recognition or character recognition, a HMM model keeps information for a character when the model is trained properly and the trained model can be used to recognize an unknown character. The advantage with HMM based systems is that they are segmentation free that is no pre segmentation of word/line images into small units such as sub-words or characters is required [20]. However there are well-known limitations with HMM based approaches. These limitations are due to two reasons-(a) the assumptions of conditional independence of the observations given the state sequence and (b) the restriction on feature extraction imposed by frame based observations [21].

The rest of the paper has been arranged as follows-

Section 2 shows the proposed model, section 3 details out pre-processings, section 4 deals with feature extraction methods; section 5 describes the classifier whereas in section 6, post-processings are described. Section 7 is about the experiments and results. Conclusions have been drawn in section 8 and in section 9, we have discussed our future works followed by section 10 which contains a single set of collected samples.

II. PROPOSED MODEL

Features are extracted in both global and local processing.HMM model for each character has been trained by the sequence of the symbols of the features extracted from collected samples. To test a handwritten character image, we extract the similar features using 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) is compared and the highest probability concludes the highest matching of the features with the corresponding model

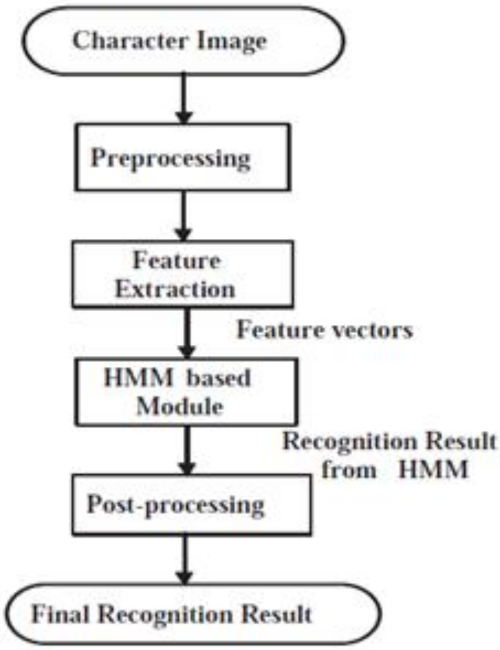


Figure 1: System Overview

III. PRE-PROCESSING

Any image processing application suffers from noise like isolated pixels. This noise gives rise to ambiguous features which results 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 flow diagram .We have used median filter for its better performance to get rid of unwanted marks or isolated pixels. Thinning is performed to get the skeleton of character image so that strokes could be conspicuous.

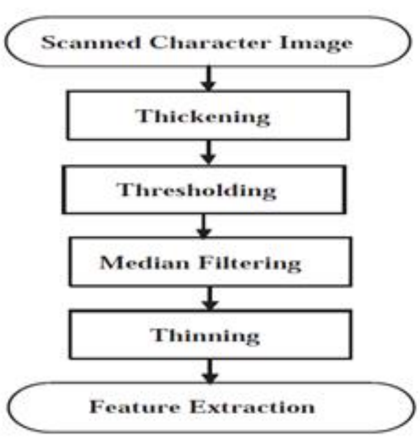


Figure 2: Block Diagram for Pre-processing

IV. FEATURE EXTRACTION METHODS

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

4.1 GLOBAL FEATURE EXTRACTION METHOD

Global feature of an image sample describe its overall structure. We extract global features as described below:

Gradient Features

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

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

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

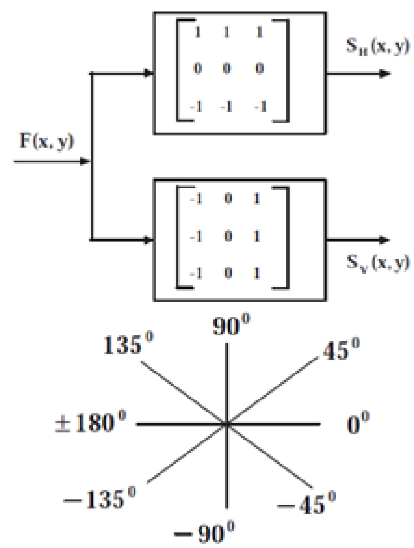


Figure 3: Gradient Feature Extraction using Sobel Masks

Phase is quantized in eight directions as shown in Fig. 6.For each quantized phase value, corresponding magnitudes are added to get the total strength in that direction. To get the feature within finite number of symbols (here eleven symbols are used), magnitudes are normalized and quantized. Finally, we consider four global gradient (G) features combining following pairs-(0°, ±180°), (45°,−135°), (90°,−90°) and (135°,−45°).

4.2 LOCAL FEATURE EXTRACTION METHODS

To extract the local information contained in the preprocessed character image, we divide the image by nine equal blocks and then, from each block four gradient features (as discussed earlier) are extracted. Thus, the local feature (L) contains 36 observations.

Therefore, our final observation sequence contains 50 observations obtained by global and local feature extraction methods, as shown below

$$O = [G(4)\ \ L(36)] \qquad (3)$$

V. 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 remain hidden within this model[16]. This HMM model can be defined as

$$\lambda = (\pi, A, B) \qquad (4)$$

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

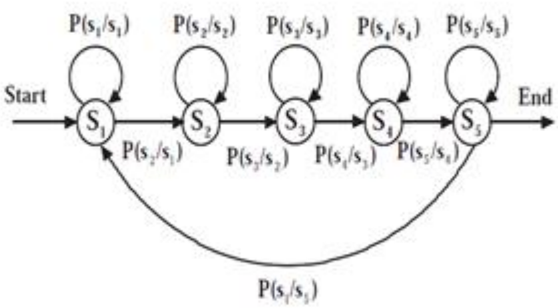


Figure 4: Left to Right Chain HMM Model with 5 States

5.1 TRAINING OF HMM MODEL

For the HMM model shown in Fig.4, π is taken as

$$\pi = [1\,0\,0\,0\,0] \qquad (5)$$

Initial value of A and B are taken as below:

$$\mathbf{A} = \begin{pmatrix} 0.8 & 0.2 & 0 & 0 & 0 \\ 0 & 0.8 & 0.2 & 0 & 0 \\ 0 & 0 & 0.8 & 0.2 & 0 \\ 0 & 0 & 0 & 0.8 & 0.2 \\ 0.2 & 0 & 0 & 0 & 0.8 \end{pmatrix}$$

and

$$\mathbf{B} = \begin{pmatrix} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{pmatrix}$$

We have used Baum-Welch algorithm to train the HMM using observation sequence obtained from the feature vectors. At the end of training process we obtain the final value of A and B which is used for recognition purpose.

5.2. RECOGNITION BY HMMM MODEL

Consider an observation vector (O) of length L as below

$$O = [O1\quad O2\quad O3 \cdots\quad OL] \qquad (6)$$

We use Viterbi decoding algorithm 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.

VI. POST-PROCESSING

A post-processing block is included at the final stage of recognition process in order to provide special care to the highly confused group of characters due to their high structural correlation factor (similarity). Few examples of such groups are mentioned below-

(1) O and Q,
(2) M and N,
(3) V and Y,
(4) C and O,
(5) B, K,R and P etc.

For each group, one or more new features are extracted that can discriminate these characters with good accuracy. For example, O and Q can be easily differentiated using signature features, as shown in Fig.5-7.

(a) (b)

Figure 5: Samples of Character Image for Post-processing (a) Character O and (b) Character Q

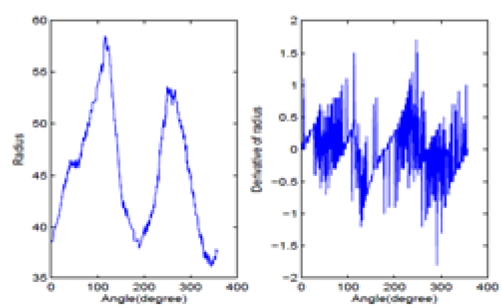


Figure 6: Signature and Derivative of Signature Plot for Character O Shown in Fig.5 (a).

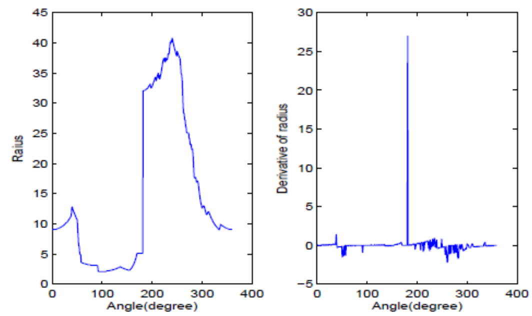


Figure 7: Signature and Derivative of Signature Plot for Character Q Shown in Fig.5 (b).

These show that the differentiation of signature plot of Q contains very large spike that can be utilized to distinguish this (Q) from the character O using a threshold criteria. Note that, this signature feature is not used to train the HMM models of all characters. The effect of this post-processing block is presented in the next section providing some experimental studies.

VII. EXPERIMENTS AND RESULTS

A total of 13000 samples are collected from 100 persons. Each writer wrote 5 set of A-Z characters. Each character image is converted to a fixed size of 150×150 pixels .We have applied our feature extraction method on these samples and then these feature values are quantized and encoded to the eleven symbols in order to create sequences of observation symbols. First 100 samples of each character are used to train the corresponding character HMM. Rest 400 samples are used to test our HMM classifier. For our experiment we have started with only 5 state model but we 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 we have got our expected result as shown in table. Table I shows the effectiveness of our proposed model.

In table 1 we have shown final recognition rate of our character recognition system using post-processing and it is compared with result obtained without post processing technique. This produces an average recognition rate of 93.24%.

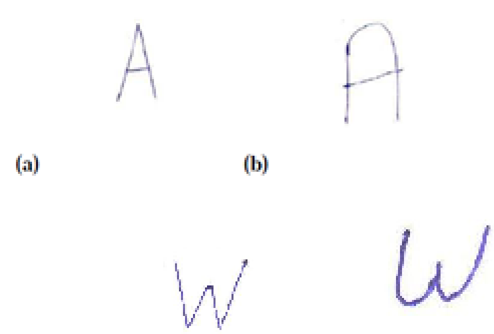
TABLE 1: Recognition Rate With or Without Post-Processing (PP) Using Proposed HMM Model

Char-acter	Recognition Rate (%)		Char-acter	Recognition Rate (%)	
	Without pp	With pp		Without pp	With pp
A	89.50	89.50	N	93.75	94.25
B	91.00	93.25	O	91.50	94.25
C	93.75	94.25	P	91.25	92.75
D	93.25	93.25	Q	92.25	94.00
E	93.75	93.75	R	90.75	93.25
F	94.00	94.00	S	90.50	91.50
G	92.50	92.50	T	93.25	93.25
H	93.75	93.75	U	93.50	93.50
I	94.25	94.25	V	91.75	93.75
J	93.50	93.50	W	89.25	89.25
K	92.50	93.75	X	93.25	93.25
L	93.75	93.75	Y	91.25	93.75
M	92.25	93.50	Z	94.50	94.50

VIII. CONCLUSION

In this paper, an approach has been made to increase the rate of recognition of handwritten character by finding both local and global features. In the post-processing section, a trial has been done to put a line of demarcation between similar looking characters.

All these specialties of this paper has made us obtain an average accuracy of 93.24%. 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 as shown in fig.



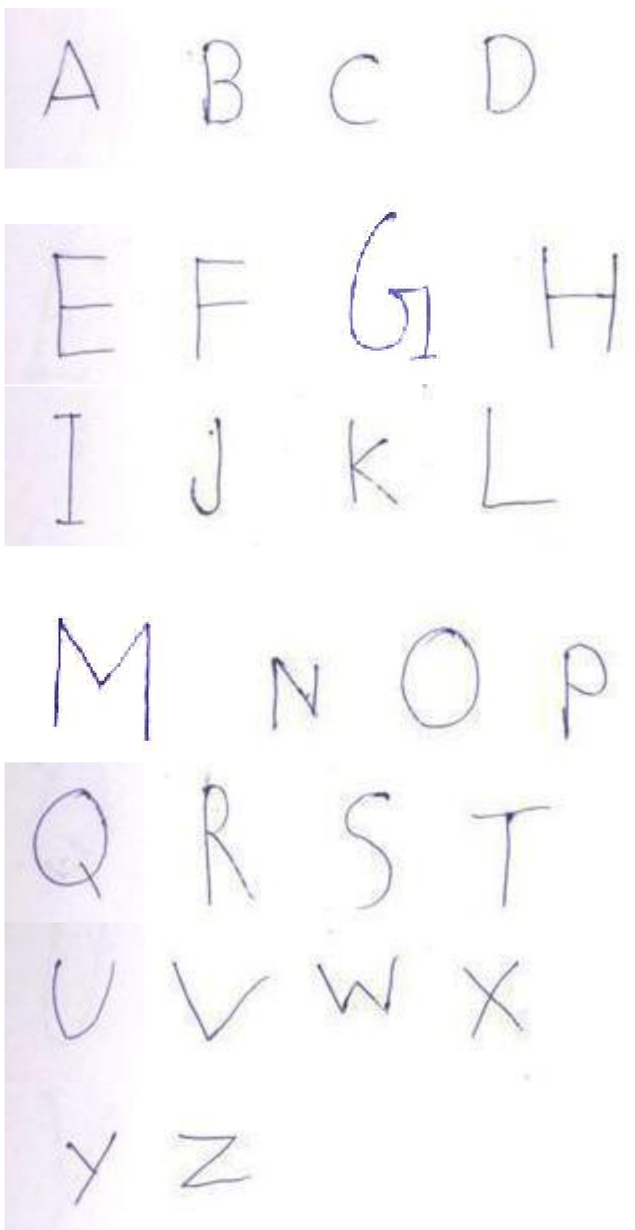
(c) (d)

Fig. 8 Two completely different formats for handwritten character A (a) A1 (b) A2 and for character W (c) W1 (d) W2

IX. CONCLUSION

HMM is a flexible tool by its inherent nature which is capable of absorbing a little bit of variations in character images. It is obvious from the samples shown above that their gradient features should differ considerably. So, our future works will be concentrated on improvement of recognition rate of such letters**.**

X. DATA-SET



ACKNOWLEDGMENT

Our thanks to the authority of Bengal College of Engineering and Technology, Durgapur (W.B.) to provide us with WI-FI connection and other facilities to carry out research works.

REFERENCES

[1] U. Bhattacharya, and B. B. Chaudhury, "Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals", IEEE Trans. Pattern analysis and machine intelligence, vol. 31, No. 3, pp. 444-457, 2009.

[2] U. Pal, T. Wakabayashi and F. Kimura, "Handwritten numeral recognition of six popular scripts", Ninth International Conference on Document Analysis and Recognition, ICDAR07, Vol.2, pp.749-753, 2007.

[3] V. K. Govindan and A. P. Shivaprasad,"Character Recognition A review", Pattern recognition, vol.23, no.7, pp.671-683, 1990

[4] J. Pradeep, E. Srinivasan and S. Himavathi, "Diagonal Based Feature Extraction for Handwritten Alphabets Recognition System Using Neural Network", International Journal of Computer Science and Information Technology (IJCSIT),, vol. 3, no. 1, pp. 27-38, Feb 2011.

[5] C. Suen, C. Nadal, R. Legault, T. Mai, and L. Lam, "Computer recognition of unconstrained handwritten numerals", *Proc. IEEE* , 80(7):1162-80.

[6] J. C. Simon , "Off-line cursive word recognition", *Proc. IEEE* , 80(7):1150-61.

[7] S. Impedovo, P. Wang and H. Bunke,editors, *Automatic bank check processing*, Singapore: World scientific; 1997.

[8] S. Srihari,"Handwritten address interpretation: a task of many pattern recognition problems", International Journal of Pattern Recognition and Artificial Intelligence, 2000; 14:663-74

[9] G. Kim, V. Govindaraju, and S. Srihari, "Architecture for handwritten text recognition systems",International Journal on document Analysis and Recognition (IJDAR), vol. 2, pp. 37-44, 1999.

[10] U. V. Murti, and H. Bunke, " Using a statistical language model to improve the performance of an HMM-basis cursive handwriting recognition system", International Journal of Pattern Recognition and Artificial Intelligence, 2001;15:65-90.

[11] S. Gunter, and H. Bunke, "Off-line cursive handwriting recognition using multiple classifier systemson the influence of vocabulary, ensemble, and training set size", Optics and Lasers in Engineering, vol. 43, pp. 437-454,2005.

[12] Hailong Liu and Xiaoqing Ding "Handwritten Character recognition Using Gradient feature and Quadratic Classifier with Multiple Discrimination Schemes,Proceedings of the 2005Eight International Conference on Document Analysis and Recognition(ICDAR'2005)

[13] H.Fujisawa,C.L.Liu ,"Directional Pattern Matching for Character Recognition",Proc.7th ICDAR,Edinburgh,Scotland,2003,pp794-798

[14] K.Ding,Z.B.Liu,L.W.Jin ,X.H.Zhu,"A comparative study of Gabor Feature And Gradient Feature For Handwritten Chinese Character Recognition",ICWAPR,Beijing ,China,2007,pp.1182-1186

[15] C.L.Liu ,K.Nakashima,H.Sako,H.Fujisawa,"Handwritten Digit Recognition :Investigation of Normalisation and Feature Extraction Techniques",Pattern Recognition ,2004,37(2),PP.265-279

[16] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition", Proceedings of The IEEE, vol. 77, no. 2, pp. 257-286, Feb. 1998.

[17] C. Mokbel, H. Abi Akl, and H. Greige, "Automatic speech recognition of Arabic digits over Telephone network", Proc.Research Trends in Science and Technology, 2002.

[18] R. El-Hajj, L. Likforman-Sulem, and C. Mokbel, "Arabic Hand- writing Recognition Using Baseline Dependent Features and Hidden Markov Modeling, Proc. Eighth Intl Conf. Document Analysis and Recognition, pp. 893-897, 2005.

[19] H. El Abed and V. Margner, "ICDAR 2009 - Arabic handwriting recognition competition", Inter. Journal on Document Analysis and Recognition, vol. 1433-2833, 2010.

[20] P. Natarajan, S. Saleem, R. Prasad, E. MacRostie, and K. Subramanian,"Multilingual Off-line Handwriting Recognition Using Hidden Markov Models: A script independent Approach", Springer Book

Chapter on Arabic and Chinese Handwriting Recognition, ISSN:0302-9743, VOL. 4768, pp. 235-250, March 2008.

[21] Zhang Hong lin, "Visiual C++Digital image pattern recognition technology and engineering practice,"Beijing: Posts & Telecom Press, 2008,pp. 52-58.

[22] R. C. Gonzalez and P. Wintz, "DIGITAL IMAGE PROCESSING," 2nd Edition, ADDISON WESLEY, Reading Mass, 198.