International Journal of Engineering Research and Development
***e**-ISSN: 2278-067X, **p**-ISSN: 2278-800X, www.ijerd.com*
Volume 3, Issue 7 (September 2012), PP. 43-49

A Simplistic Way of Feature Extraction Directed towards a Better Recognition Accuracy

Binod Kumar Prasad
Department of Electronics and Communication Engineering, Bengal College of Engineering and Technology, Durgapur, INDIA

Abstract**—*Feature extraction is an important component of a Recognition system. A better combination of features regarding an object gives rise to a system with promising result. Several feature extraction methods have been reported till date. In this paper, projection features along with curvature features have been utilised in unison to yield an acceptable overall recognition rate. Altogether only ten features are referred here. The aim is to reduce the algorithm complexity, so that the processing time could be minimised and the recognition system could be a real time system. Ultimately, HMM has been used as recognition tool to get a recognition accuracy of 95.21%.***

Keywords**— *Projection features, curvature features, Handwritten character recognition, Viterbi algorithm, Baum-Welch algorithm.***

I. INTRODUCTION

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 been found to grow 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].

Projection features consist of mean, variance and entropy of the projected histograms on both X- and Y-axes. Curvature features quantifies the trend of bending as seen towards the character image from Top, Bottom, Right and Left.

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 [12],[13],[14],[15]. With the usage of HMM models for the pattern recognition or character recognition, an 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 [16]. On the other hand, HMM based approaches have been found to possess some limitations also. 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 [17]. However, the rest of the paper has been arranged as follows-

Section II shows the proposed model, section III details out pre-processing, section IV deals with feature extraction methods; section V describes the classifier whereas in section VI, post-processings are described. Section VII is about the experiments and results. Conclusions have been drawn in section VIII and finally, in section IX, a single set of collected data has been shown.

II. PROPOSED MODEL

Features have been extracted globally. Generally for each character, a single HMM model is considered and trained by feature vectors. But it has been observed that some handwritten characters (e.g. A, W) show two completely different formats as shown in Fig.2.Multiple HMM models have been selected for these characters whereas for other characters, a single HMM model is adopted as shown in fig.3. Models are trained by the sequence of the symbols of the features extracted from some of the samples. To test a handwritten character image, we extract the similar features using same procedure as earlier and the corresponding sequence (observation) is compared 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.

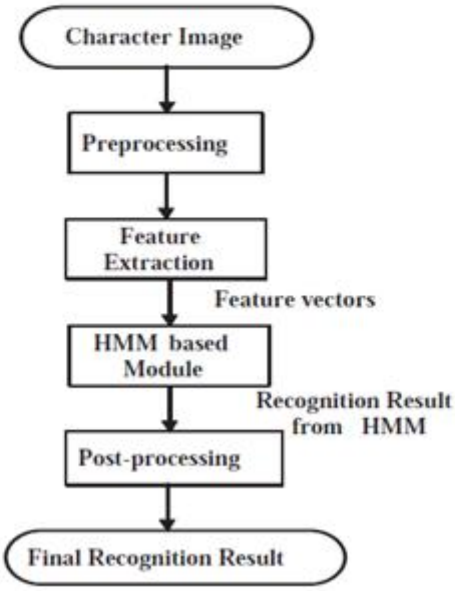
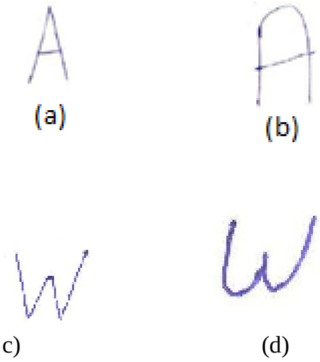


Fig. 1 **System Overview**



c) (d)

Fig.2 **Two completely different formats for handwritten character A (a) A1 (b) A2 and for character W (c) W1 (d) W2**

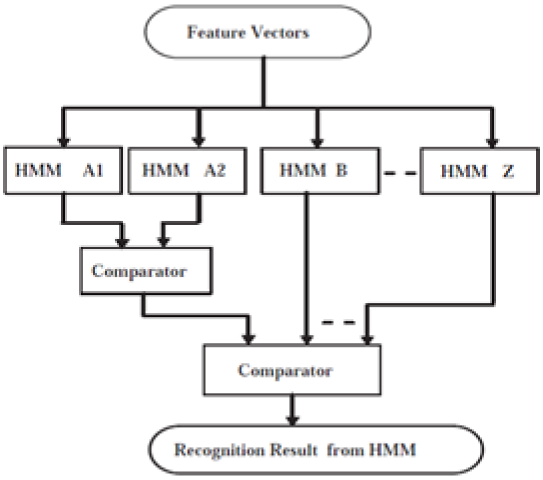


Fig.3 **Proposed HMM Module for Character Recognition**

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 pre-processing mechanism has been executed before features could be extracted. 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.

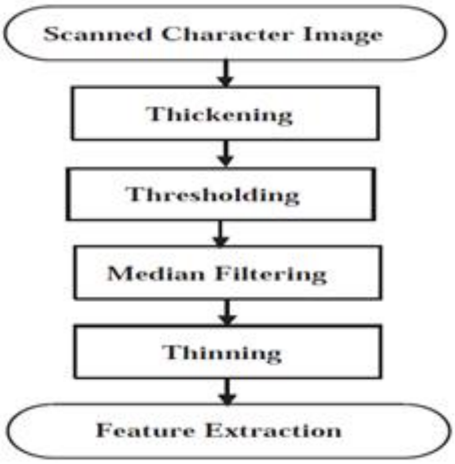


Fig. 4 Block Diagram for Pre-processing

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 pre-processed images. Our feature extraction process consists of –

1. Projection features 2.Curvature features

A. PROJECTION FEATURES

After pre-processing the two dimensional binary handwritten character image f(x, y) is projected to X and Y axis respectively and we get following two histograms.

Projection on X-axis:

$$U(i) = \sum_{k=1}^{N} f(i,k), \qquad 1 \le i \le N \qquad (1)$$

Projection on Y-axis:

$$V(i) = \sum_{k=1}^{N} f(k,j), \qquad 1 \le j \le N \qquad (2)$$

From each histogram , three features namely mean, variance and entropy are calculated as given below

$$\mu_X = \frac{1}{N}\sum_{i=1}^{N} X(i) \qquad (3)$$

$$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(X(i) - \mu_X)^2 \qquad (4)$$

and

$$E_X = -\sum_{i=1}^{N} e_{xi} \log(e_{xi}) \qquad (5)$$

Where

$$e_{xi} = \frac{X(i)}{\sum_{i=1}^{N}(X(i))} \;; \quad X = U \text{ and } V \qquad (6)$$

Therefore, from projection method we find total six features as below

$$P = [\mu_U \ \sigma_U^2 \ \ E_U \ \mu_V \ \sigma_V^2 \ \ E_V] \qquad (7)$$

B. CURVATURE FEATURES

Four curvatures measured from top (C_T), bottom (C_B), right side (C_R) and left side (C_L) have been taken. So,the curvature feature (C) consists of four observations as below

$$C = [C_T \;\; C_B \;\; C_R \;\; C_L] \qquad (8)$$

Therefore, our final observation sequence contains 10 observations as shown below

$$O = [P\ (6)\ C\ (4)] \qquad (9)$$

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 remains hidden within this model [12]. This HMM model can be defined as

$$\lambda = (\pi, A, B) \qquad (10)$$

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, a closed left to right chain HMM model has been used for handwritten English characters recognition.

Baum-Welch algorithm is used to train the HMM using observation sequence obtained from the feature vectors. At the end of training process, the final values of A and B are obtained which are used for recognition purpose.

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 almost 100 percent accuracy. For example, O and Q can be easily differentiated using signature features [14] , as shown in Fig.8-10.

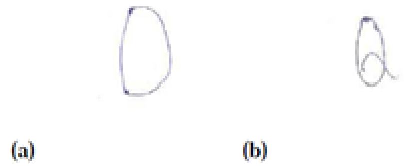

(a) (b)

Fig. 8 **Samples of Character Image for Post-processing (a) Character O and (b) Character Q**

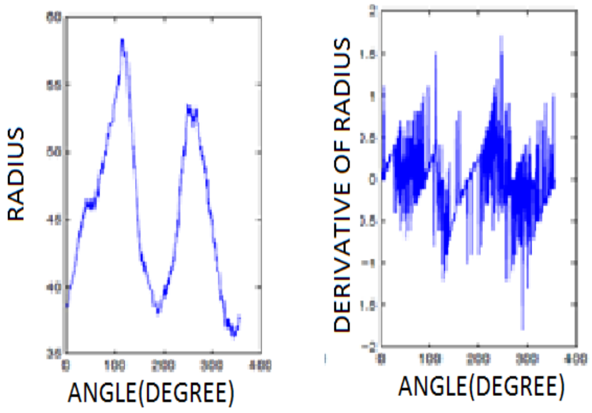


Fig. 9 **Signature and Derivative of Signature Plot for Character O Shown in Fig.8 (a).**

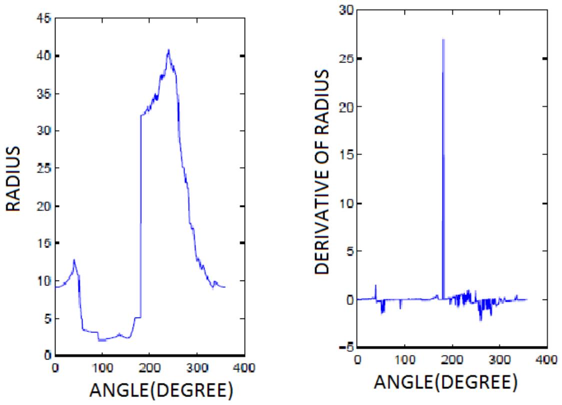


Fig. 10. **Signature and Derivative of Signature Plot for Character Q Shown in Fig.8(b).**

These show that the differentiation of signature plot of Q contains very large spike that can be utilised to distinguish this (Q) from the character O using a threshold criteria. It should be noted 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 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 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 the experiment, the training is started with only 5 states model but we observed that as the no. of states of HMM model is increased, the corresponding recognition rate is also improved. Finally, the expected result has been obtained with 24 states HMM model as shown in tables. Table 1, shows the effectiveness of our proposed model of taking multiple HMM for a single character A.

TABLE 1 : **Improvement of Recognition Rate for Character A Using Proposed Model**

Character	Single HMM Model Recognition Rate (%)	Multiple HMM Model Recognition Rate (%)
A	81.25	91.5
W	82.5	91.25

In Table 2, 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 95.21%.

TABLE 2. **Recognition Rate With or Without Post-Processing (PP) Using Proposed Multiple HMM Model**

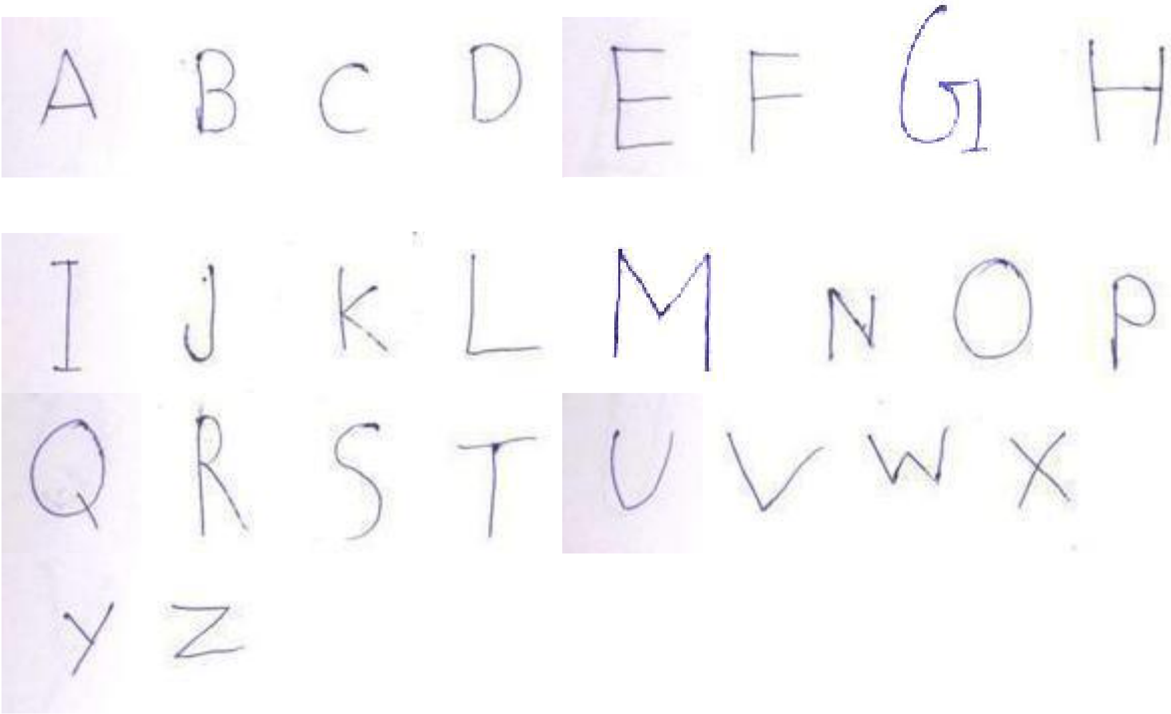
Character	Recognition Rate(%)		Character	Recognition Rate(%)	
	Without pp	With pp		Without pp	With pp
A	91.5	91.5	N	97.75	98.25
B	88.25	90.12	O	95.50	97.55
C	97.75	98.25	P	96.25	98.25
D	82.08	82.08	Q	96.25	99.00
E	89.5	89.5	R	90.75	91.25
F	89.34	89.34	S	91.50	91.50
G	98	98	T	98.85	98.85
H	90.75	90.75	U	91.50	91.50
I	98.25	98.25	V	96.55	98.75
J	99.50	99.50	W	91.25	91.25
K	95.50	97.75	X	98.35	98.35
L	97.9	97.9	Y	98.25	99.75
M	97.25	99.73	Z	98.50	98.50

VIII. CONCLUSION

In this paper, an approach has been made to increase the rate of recognition of handwritten character by good combination of projection and curvature features only. Multiple level HMM model is designed for some specific letters having wide range of variations from writer to writer. In the last section, a trial has been made to put a line of demarcation between similar looking characters. All these specialties of this paper have made us obtain an average accuracy of 95.21%. For some letters, the accuracy rate is even close to 100%.

IX. DATA-SET

One set of collected data has been shown for reference



ACKNOWLEDGMENTS

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 like scanning etc. to carry out our 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]. 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.

[13]. 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.

[14]. 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.

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

[16]. 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.

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

[18]. R. C. Gonzalez and P. Wintz,“Digital Image Processing,”2nd Edition, Addison Wesley,Reading Mass,1987