Recognition of Myanmar Handwriting Text Based on Hidden Markov Model

San San Mon and Myint Myint Sein

ssm10@wwlmail.com, mlm@wwlmail.com

University of Computer Studies (UCSM, UCSY)

Abstract

Handwriting recognition is one of the most challenging tasks and exciting areas of research in computer vision. Numerous document recognition methods have been proposed in various languages and character set such as Arabic, India, Korean, Japanese, Chinese and so on. This paper presents the recent result of the research work of Myanmar handwriting text recognition and translation. Each segmented character of handwriting text is not only recognized but also transform to Myanmar printed character and multiple languages. The technology is successfully used by business which process lots of handwritten documents, like insurance companies. The quality of recognition can be substantially increased by structuring the document.

Keyword: *Myanmar printed character, Handwriting text, Natural Handwritten Recognition, OCR*

1. Introduction

Today Natural Handwritten Recognition (NHR) has arrived that will have a major impact on reading documents. Off-line handwritten recognition involves the automatic conversion of text in an image into letter codes which are usable within computer and text processing applications. The data obtained by this form is regarded as a static representation of handwriting. Handwriting recognition is the ability of a computer to receive intelligible handwritten input. The image of written text may be sensed "off line" form a piece of paper by optical scanning (Optical character recognition). Handwritten recognition principally entails optical character recognition.

Numerous document recognitions have been proposed in various languages and character set. F.-H.Cheng et al. [1] proposed the relaxation method for recognizing the hand printed Chinese characters. Handwritten Chinese characters are recognized based on short line segments by H.-J.Lee and B.Chen [2]. F. H. Cheng et al.[3] modified the Hough transform techniques for recognition of handwritten Chinese characters. C.Y Suen et al.[4] presented the recognition of handwritten Month words on bank cheques. M.Y. Chen et al.[5] proposed the Off-line Handwritten Word Recognition using a Hidden Markov Model type Stochastic Network. N. Thein [7] proposed a recognition method for Myanmar Car license plate number. Myanmar handwritten date on bank cheque is recognized in H. H. Thaung [6]. Their approach recognize only Myanmar printed and handwritten digit characters. S. Hamar et al.[8] done the person identification from their handwritten text. Their approach can be determined the same writer or not. The difference between two texts or letters is investigated.

In our approach, we not only determine the same writer but also transform to the Myanmar printed characters. Hidden Markov Model has been developed for classification of Myanmar characters. This paper organized by 4 stages. Existing method and proposed method concerning with handwritten recognition has been introduced in Section1. In Section 2, Myanmar character pattern and general process of proposed system are discussed. Developed Hidden Markov Model is illustrated in Section 3. Experiments and results are shown in Section 4. In Section 5, discussion and future work are expressed.

2. Overview of the Proposed System Myanmar character pattern

Myanmar character pattern will be introduced in this Section. Myanmar script consists of (33) basics characters and (12) basic vowels, respectively. Basic Myanmar alphabet and vowels are shown in figure 1. Myanmar written style is left to right in nature. Most of the words are written combined with other symbols. One basic character either may be combined with one or more extended characters or may stand as a single character. Myanmar handwritten text document is illustrated in figure 2.

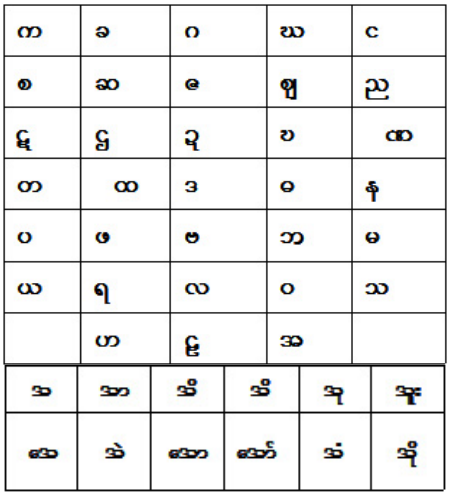

Figure 1 Basic Myanmar character and Basic Myanmar vowel

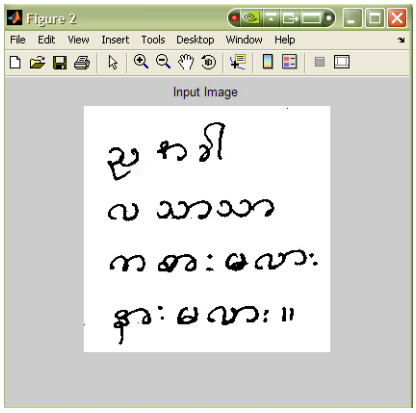

Figure 2 Myanmar handwritten text document

2.1 Preprocessing and Segmentation

It is necessary to perform several document analysis operations prior to recognizing text in scanned documents. Some operations, the task of converting a gray scale, binarilization, noise removing and line segmentation, are performed in preprocessing step. Here, lines of text might undulate up and down and ascenders and descenders frequently intersect characters of neighboring lines. Line separation is usually followed by a procedure that separates the text line into words. For word segmentation issues, we assume that the gaps between words are larger than the gaps between the characters. The author's writing style, in terms of spacing, is captured by characterizing the variation of spacing between adjacent characters as a function of the corresponding characters themselves. The notion of expecting grater space between characters with leading and trailing ligatures is enclosed into the segmentation scheme. Segmentation points are determined using features like ligatures and concavities. Gaps between character segments (a character segment can be a character or a part of character or a part of character) and heights of character segments are used in the algorithm.

The block diagram of the system is shown in figure 1.

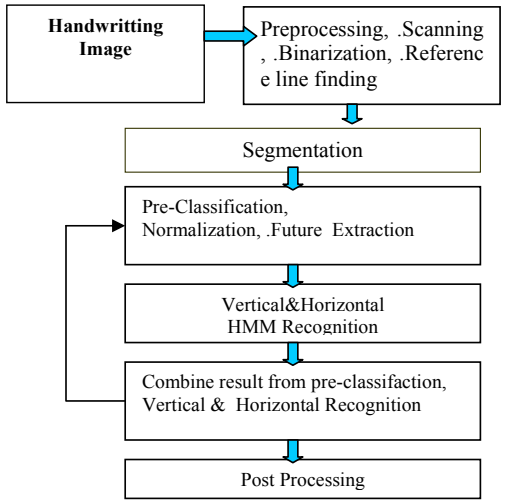

Figure 3. Block diagram of the system

3. Hidden Markov Model (HMM)

Hidden Markov Models (HMMs) have been very successful in the field of Optical Character Recognition (OCR) due to their ability to model the passage of time. This Section shall present an introduction to the theories and algorithms behind Hidden Markov Models, as well as the issues involved in their use in practical applications. Markov recognition techniques to off-line handwriting and have shown promising results in the recognition of cursive handwriting.

The Hidden Markov Models (HMMs) is a doubly stochastic variant of Markov Model, with an underlying stochastic process that is not observable (hidden), but can not be observed through another set of stochastic process that produce the sequence of observed symbols. The underlying stochastic process of the HMM is a finite automaton. It contains state, state transitions, and transition probabilities. Initial state probabilities exist for each state which defines the chances of the model being found in the particular state at the beginning of an observation sequence. Observation densities, which indicate the probability of a observation being produced by the model, are defined for each state, and these make up the observable stochastic process. In a given state, the observation probability density function for that state defines the expectancy of observations in that state; while the sequence of states traveled define a movement from one expected observations to another. It is unknown to an outside observer what the source or model representing the source is in at any given time. All that can be observed is the sequence of symbols produced that make up the signal. Therefore, this type of model is referred to as Hidden. A sample of the left-right Hidden Markov model is shown in figure 4.

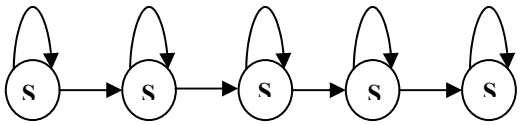

Figure 4. **A Sample Left-Right Hidden Markov Models**

Each state in the model has a number of parameters associated with it which describe the probability of making a transition from that state to another state in the model, the probability of particular observation being produced while in that state, as well as the probability that a particular state was the start state for the sequence under observation. A notation is defined as follow:

$O = \{O_1, O_2, ..., O_T\}$ as the observation sequence

T = the length of the observation sequence

$Q = \{q_1, q_2, ..., q_N\}$ as the set of states in the model

N = the number of states in the model.

In addition, for discrete HMMs:

$V = \{v_1, v_2, ..., v_M\}$ is the set of possible observations

M = the number of observation symbols.

For continuous HMMs, the number of different possible observation symbols is infinite.

In the discrete case, a Hidden Markov Model can be described by a set of parameters $\lambda = (\mathbf{A}, \mathbf{B}, \mathbf{\Pi})$, where **A** and **B** are matrices, and **Π** is a vector.

Matrix **A** is defined as the set of all state transition probability distributions for the model:

$$a_{ij} = P\{q_j \text{ at } t+1 / q_i \text{ at } t\},$$

Where q_i is the state the model is in at some time t, q_j is a possible next state, and a_{ij} is the probability that the model will make the transition to this next state at the next time increment.

Matrix **B** is defined as the set of all observation probability distribution functions for the model:

$$b_j(k) = P\{v_k \text{ at } t / q_j \text{ at } t\},$$

where q_j is the state the model is in at some time t, v_k is an observation that may be seen by the model, and $b_j(k)$ is the probability of this observation occurring in the observation sequence while in state q_j. Vector **Π** is defined as the initial state distribution for the model: z $\pi_i = P\{q_i$ at $t = 1\}$. Where $\boldsymbol{q_i}$ is a state in the model and $\boldsymbol{\pi_i}$ is the probability that the model will begin in state q_i at the beginning of an observation sequence.

Observations are feature vectors extracted from horizontal and vertical strips. Observation feature in character '**U**' illustrates in figure 5. Observation is the number of pixels in the strips. And the structure of the hidden states is shown in figure 6.

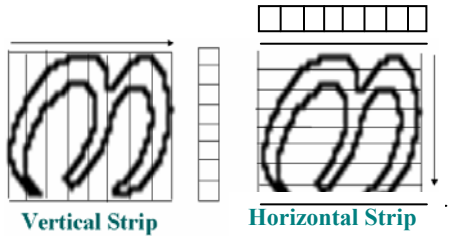

Figure.5:**Observation Feature in Character "ဆ"**

The structure of hidden states is chosen for characters 'Kagyi '. Where $S = \{1,2, \ldots, N\}$, the state at time *t*.

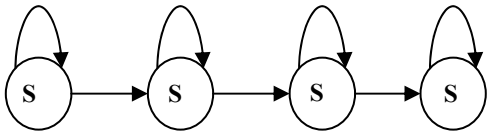

Figure 6 The structure of Hidden States

Maximum likelihood parameter estimation for HMM is obtained by the iterative procedure, with multiple observation sequences. Then, for a given observation sequence $O = \{O_1, O_2, \ldots, O_T\}$, the HMM is used to compute $P[O / \lambda]$, where *T* is the length of the sequence.

Two HMMs are created for every character, one for modeling the horizontal information and the other for modeling the vertical information. The discrete Hidden Markov character Models are trained using standard procedures. The numbers of states for all the character HMMs is fixed and no skip states are allowed. Only the pre-classified candidate characters are passed on for HMM recognition.

Two log probabilities for each candidate character are calculated using the horizontal direction HMMs and vertical direction HMMs. Then, the log probabilities are added together to obtain a final 3-best list for recognizing the 300 handwriting Myanmar characters. The results of the HMMs recognition are shown table 1.

Table 1. Results of the HMMs recognition

Top 1	Top 2	Top 3
66.67%	84.12%	92.1%

5. Experiment and Results

An experiment has been done to confirm the effectiveness of proposed method. After preprocessing step include noise filtering, skewing and resizing, foreground handwritten characters or text are extracted. For character segmentation, line segmentation is firstly performed. And then column wise segmentation has been done to obtain the segmented characters. Line segmentation of the scanned image (figure 2) is shown in figure 7. Figure 8 described the matching result for transforming the handwritten character to printed character. Final result of the interpretation of handwritten text by printed characters is illustrated in figure9.

Figure 7. **Line Segments**

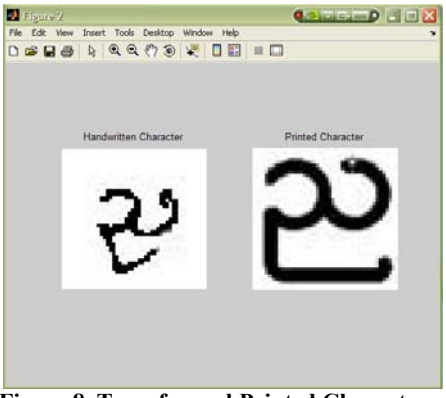

Figure 8. **Transformed Printed Characters**

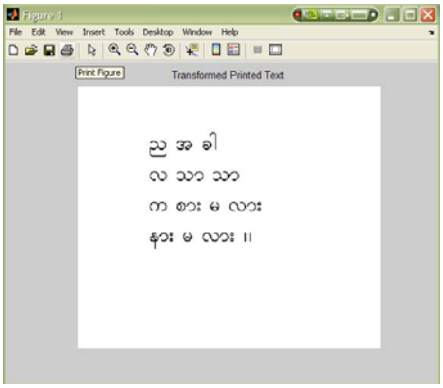

Figure 9. **Transformed Printed text**

6. Conclusion

Handwritten recognition and printed character translation techniques are presented in this paper. The recognition process is based on the Hidden Markov Model. The output of the system is machine printed character in Myanmar language. This approach is can be extended to other languages. In next steps, we aim to recognize the Myanmar handwritten characters and transform to the printed characters in multiple languages and interpret by voice. Through the experimental results, we see the proposed system woks adequately.

References

[1] F.-H.Cheng, W.-H.Hsu, and M.-Y.Chen,, "Recognition of Handprinted Chinese Characters via Stroke and Relaxation", Pattern Recognition, Vol.26, No.4, pp.579-593, 1993.

[2] H.-J.Lee and B.Chen, "Recognition of Handwritten Chinese Characters via Short Line Segments", Pattern Recognition, Vol.25, No.5, pp.543-552, 1992.

[3] F. H. Cheng, W.-H.Hsu and M.-Y.Chen, "Recognition of Handwritten Chinese Characters by Modified Hough Transform Techniques", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.11,pp .429-439, April, 1989.

[4] C.Y Suen, L. Lam, D. Guillevic, NW Strathy, M. Cheriet, " Recognition of handwritten Month Words on Bank Cheques ", September 2001. Systems, 3(1), pp. 83-95.

[5] M.Y. Chen, A. Kundu, and J. Zhou, Off-line Handwritten Word Recognition using a Hidden Markov Model Type Stochastic Network, IEEE Transaction on Pattern Analysis and Machine Intelligence,16 (May 1994),pp. 481-496.

[6] H. H. Thaung, "Recognition on Handwritten Date on Bank cheque using Zoning Methods", Ph.D Dissertation, University of Computer Studies, Yangon, 2004.

[7] N Thein, Car License Plate Recognition using Neural Network, Ph.D Dissertation, University of Computer Studies, Yangon, 2004.

[8] S. Hamar , " Recognition of Myanmar Handwritten Text Characters", Ph.D Dissertation, University of Computer Studies, Yangon, 2005.

[9] R. M. Bozinovic and S. N. Srihari, “Off-line cursive script word recognition,” IEEE trans. Pattern Anal. Mach. Intell., vol. 11, pp. 68-83, Jan. 1989.

[10] V. Bouletreau, N. Vincent, R. Sabourin, and H. Emptoz, “Handwriting and Signature: One or Two Personality Identifiers?” Proc. 14th Int'l Conf. Pattern Recognition, pp. ,1758-1,760, Brisbane, Australia, Aug. 1998.