Automatic Extraction of Payee's Name and Legal Amount from Myanmar Cheque by Using Hidden Markov Model

Nang Aye Aye Htwe,Thi Thi Soe and Myint Myint Sein

Mandalay Technological University, University of Computer Studies, Yangon
htwe.aye@gmail.com and mlm@wwlmail.com

Abstract

This paper proposed a model to register the payee's name and legal amount on Myanmar bank cheque by using Hidden Markov Model. An innovative approach for extracting payee's name and legal amount and other documents from bank cheque images is proposed based on the integration of the crop method with the sliding window technique. The new entry documents of chequer are extracted by eliminating the background. Then individual groups of document are obtained by using region base segmentation approach. Recognition on extracted parameter writing in Myanmar character from check image demonstrates by the effectiveness of the Hidden Markov Model.

Keywords: Offline handwriting, Preprocessing, normalization, feature extraction, handwriting word recognition.

1. Introduction

Writing has been the most natural mode of collecting, storing and transmitting information through the countries, now serves not for communication among human but also serves for communication of human and machines. Machine simulation of human reading has been the subject of intensive research for the last three decades. However, the early investigations were limited by the memory and power of the computer available at that time. Handwritings can be used for forensic tasks and for person authentication. According to the way handwriting data is generated, two different approaches can be distinguished: on-line and off-line. The data are captured during the writing process by a special pen on an electronic surface in online system. In off-line recognition system, the data are acquired by a scanner after the writing process is over. Recognition of off-line character is more-complex than the on-line case due to the presence of noise in the image acquisition process and the loss of temporal information such as the writing sequence and the velocity. This information is very helpful in a recognition process.

Each day, billion of business and financial documents have to be processed by computer. The great bulk of them are still processed manually by human operators, the most common and labor-consuming operation being document amount reading and typing. A common way to automate the process is to replace the human operator with an off-line recognition system that is able to do the operator's job.

With respect to bank processing, many systems have been developed for information extraction, courtesy amount and legal amount recognition and signature verification. One of the most important tasks in automatic bank cheque processing is extraction of handwriting payee's name and legal amount from bank cheque and then feeding them to off-line handwriting character recognition. The nature of bank cheques is varied and complex and this makes the problem of automatic bank cheque processing very difficult.

Recognition on Myanmar handwritten date on bank checks using Zoning methods is presented by H.H. Taung[1]. In this work, Myanmar handwritten digit in bank check is only detected. Cheriet et al. [2] have been proposed to eliminate complex background of checks. Baselines extraction is used to localize the information of the principal items such as the amount and the payee. Djeziri et al. [3] also developed a technique to eliminate the background picture of checks. This technique can extract the line even from the low contrast images. Two fusion strategies are employed on the feature extracted using filiformity and Gabor filter techniques. In order to produce a successful cheque processing system, many sub-problems have to be solved such as background and noise removal, recognition of the immense styles of handwriting and signatures, touching and overlapping data in various field of information and errors in the recognition techniques [6]. Madasu et al. [4] implemented a system to read courtesy amount, legal amount and date fields on cheques. Oliveira el al. [5] defined verification as the postprocessing of the results produced bt recognizers. Method to recognize handwritten words are well known and widely uderd for many different languages. Hidden Markov Models (HMM) have been very successfully implemented for recognizing cursive written words. The process of recognition system involves preprocessing, segmentation, feature extraction, and classification. Several type of decision method including statistical methods, neural networks, structural matching (on trees, chains, etc) and stochastic processing (Markov chains, etc) will use

along with different types of features. The most common goal of automatic bank cheque treatment systems is the recognition of handwriting information. However, in order to do this, it is necessary to use a reliable and efficient process able to identify and to extract the information which can be submitted to further recognition phase.

In this paper, automatic data entry system for bank cheque is presented. Our experiment perform not only recognize the Myanmar handwritten characters but also transform to the printed characters. Binarization is used in order to reduce data storage and increase processing speed. It is often desirable to represent gray-scale or color image as binary image. Sliding window approach is used to avoid the segmentation. It converts the words into sequence of observation vectors and then recognizes them with Hidden Markov Models. The normalization is based on a statistical approach that assumes very general hypotheses about the words and that it is completely adaptive. The use of feature vectors is more suitable for the recognition to improve significantly the recognition rate.

Hidden Markov Models have been successfully applied in speech recognition. More recently, they have been used in handwriting word recognition. HMM word architecture is pretty well adapted to describe a word image as a sequence of observations. Some approach is based on explicit segmentation where words are split into characters or pseudo-characters to provide a grapheme sequence. In other approaches, this kind of segmentation is carried out implicitly, during the recognition phase through the HMM. In some holistic approaches, the explicit segmentation can also be used to improve the word-length estimation and to permit a better class discrimination while training the word models.

2. Overview of the System

A cheque document normally comprises machine printed and written texts as well as graphic images. In layout or logical structure, each cheque is composed of the following logical objects:

- Name and address of financial institute
- Date of issue
- Name of the payee
- Amount of payment in numeric format, the courtesy amount
- Amount of payment in Myanmar character format, the legal amount
- Authorized signature
- Account number
- Logo of financial institute or company

The layout or logical structure of Myanmar bank cheque is described in the following Figure 1. The overview procedure of item extraction system is described in Figure 2.

Figure. 1 A typical Myanmar bank cheque

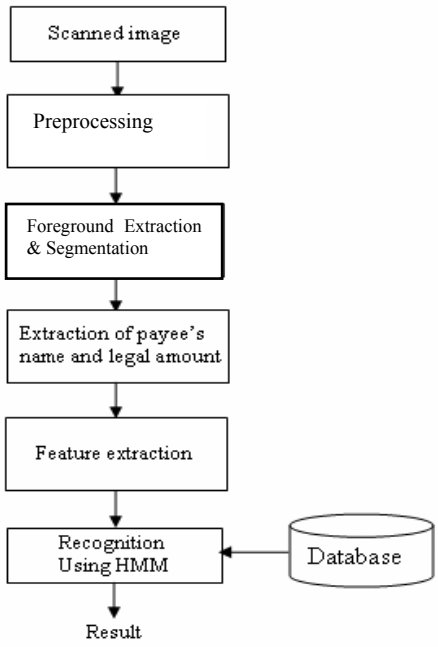


Figure. 2 The overall structure of the system

The first step is to scan the cheque or to form from which the payee's name and legal amount is to extract and apply some preprocessing steps to remove the noise. Then the foreground extraction process is performed for obtaining the new entry document in a cheque. Region base segmentation approach is applied to extract the payee's name and legal amount .The following sections describes the individual states of the system in more detail for the complete understanding of the system. Continuous Hidden Markov Modern is used for recognition.

3. Pre-processing

The goal of the preprocessing in handwriting recognition system is to reduce irrelevant information such as noise and intra-class probability. The bank

cheque is scanned by using 300 DPI (Dots per Inch) and 256 gray levels. An algorithm is developed for extracting the new entry documents of bank cheque. By removing the background pixels, the foreground image is obtained. In order to reduce data storage and increase processing speed, it is often desirable to represent gray-scale or color image as binary image by using threshold method which is based on a global thresholding technique. In such a case, one threshold value is picked for the entire image. Thresholding is used to convert a gray scale image to a binary image. All pixels with intensity less than a threshold are made black (intensity=0) and the ones above the threshold are made white (intensity=255). Threshold operates on the gray level histogram of the image. It is used to maximize the between-class variance i.e. minimize the weighted within-class).

4. Extraction of payee's name and legal amount

The payee's name and legal amount of bank cheque are written in Myanmar character. Myanmar characters are very complex. It contains in circle, curve, line and dot. So, it required the most effective method to extract the interest parameters.

Figure. 3 Extracted payee's name and legal amount on the bank Cheque

After eliminating the background from input image, the region segmentation is performed to separate the required documents which contain user name, legal amount, date and sign. Then, the segmented area for each work's group in a cheque is individually defined relative to their positions. Figure 3 shows the extracted payee's name and legal amount from the bank cheque.

5. Feature Extraction

The role of the normalization is the removal of slope and slant, two effects due to acquisition and handwriting style. The technique is based on two assumptions: The slant is the deviation of strokes from vertical, varying between words and between writers. The slope is the angle of the base line of the word if it is not written horizontally.

In sliding window based system, a fixed width window moves from left to right along the word image and stops at regular steps. At each step, a feature is extracted from the isolated frame. Once a frame is isolated, the area containing pixel is selected. The area is partitioned into 16 cells arranged in a regular 4 ×4 grid. Example of Myanmar handwritten characters are shown in Figure 4.

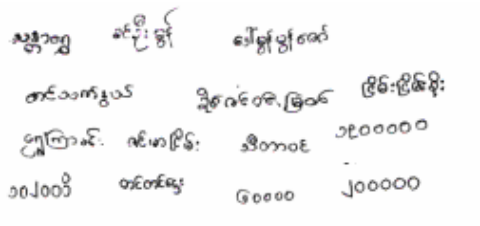


Figure. 4 Example of the Myanmar handwritten character

The number of foreground pixels is computed for every cell and the feature vector is obtained as follow:

$$F = \left(\frac{n_1}{N}, \frac{n_2}{N}, \ldots, \frac{n_{16}}{N} \right)$$

where $N = \sum n_i$

This feature extraction process has two important advantages: The first is that, being based on the local averaging rather than on a precise reconstruction of the pattern, it is robust with respect to noise. The second is that its dimension is low (this is especially important when few training samples are available).

6. Recognition

Once the feature extraction process is completed, the image is converted into a sequence of observation of vectors. This makes HMMs especially suitable for the modeling because they are probability density function over the space of the vector sequence. Since handwriting word is a sequence of symbols ordered from left to right, the most reasonable topology for handwriting recognition is called left-right topology, Figure 5.

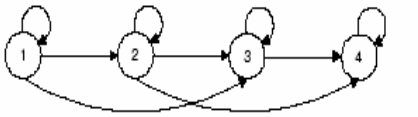


Figure. 5 Left- right HMM

In HMM, the outputs are generated by states and the alternative definition in which outputs are generated by transitions into states. Transition matrix has the following property:

$$\alpha_{ij} \geq 0 \quad \text{for } j = i,\ j = i+1$$
$$\alpha_{ij} = 0 \quad \text{otherwise}$$

The following parameters are used to define in Hidden Markov Model:

- T: Length of the observation sequence $O = (O_1, O_2, \ldots, O_T)$
- N: Number of states in the model;

- *M*: Number of possible observation symbols;
- *S* : Set of possible states of the model;
- q t : State of the process at time t
- $V = \{v_1, v_2, ..., v_m\}$: Discrete set of possible observation symbols;

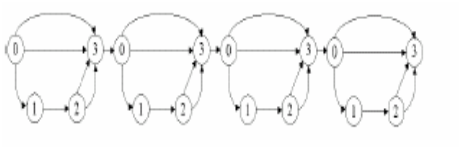


Figure. 6 Training model

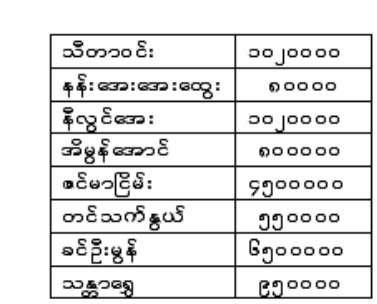
- $A = \{a_{ij}\}$: State transition probability distribution, in which aij denotes the Probability of going from state si at time t to state sj at time t + 1;
- $B = \{b_j(k)\}$: Observation symbol probability distribution, in which bj(k) denotes the output symbol probability in state sj of producing a real observation
- Symbol ot = vk;
- $\prod = \{\pi_i\}$: The initial state distribution, in which π_i denotes the probability of being in state i at time t = 1.

Given a model, to be represented by the compact notation $\lambda = \{A, B, \Pi\}$ three basic problems of interest must be solved for the model. A continuous density HMMs is used because of the emission probability distribution is continuous and estimates the probability of an observation vector being emitted when being in a certain state. The training model and recognized result are shown in figure 6 and figure 7, respectively.

Figure. 7 Recognized the handwriting characters as machine printed characters

7. Conclusion

In this paper, an algorithm is developed for automatic extraction of payee's name and legal amount from Myanmar bank cheque. A continuous Hidden Markov Model is used for recognition of payee's name and legal amount written in Myanmar language. Region base segmentation method and the effective sliding window method to improve the recognition rate. The output of the system is machine printed character. Figure 8 described the output result of the experiment. In next steps, we aim to recognize the Myanmar handwritten characters and transform to the printed characters in multiple languages.



သီတာဝင်း	၁၀၂၀၀၀၀
နန်းအေးအေးထွေး	၈၀၀၀၀
နီလွင်အေး	၁၀၂၀၀၀၀
အိမွန်အောင်	၈၀၀၀၀၀
ဇင်မာငြိမ်း	၄၅၀၀၀၀၀
တင်သက်နွယ်	၅၅၀၀၀၀
ခင်ဦးမွန်	၆၅၀၀၀၀၀
သန္တာရွှေ	၉၅၀၀၀၀

Figure. 8 The final output of the system

References

[1] H. H. Thaung, "Recognition on Handwritten Date on Bank checks using Zoning Methods", PhD Dissertation, University of Computer Studies, Yangon, 2004.

[2] M. Cheriet, J. N. Said and C. Y. Suen, "*A formal model for document processing of business forms*", Proceedings of the Third International Conference on Document Analysis and Recognition, Vol.1, pp. 210-213, August 14-16, 1995, Montreal, Canada.

[3] S. Djeziri, F. Nouboud, and R. Plamondon, "*Extraction of signatures from check background based on a filiformity criterion*," IEEE Trans. Image Processing, vol. 7, pp. 1425–1438, Oct. 1998.

[4] V.K. Madasu, M.H.M. Yusof, M. Hanmandlu and K. Kubik "*Automatic extraction of signatures from bank cheques and other documents*," Proceedings of the VIIth Biennial Australian Pattern Recognition Conference – DICTA 2003, 10-12 December 2003, Sydney, Australia, vol. 2, pp. 591-600.

[5] L.S. Oliveira, R.Sabourin, F.Bortolozzi, and C.Y.Suen, "*A modular system to recognize numerical amounts on brazilian bank cheques*," Proc 6th International Conference on Document Analysis and Recognition, Seattle, WA, 2001; 389–394.

[6] C.Y. Suen, Q. Xu and L. Lam, "*Automatic recognition of handwritten data on cheques – Fact or Fiction*," Pattern Recognition Letters, Vol.20, pp. 1287-1295, 1999.

[7] M. Pechwitz, V. Maergner, "*HMM Based Approach for Handwritten Arabic Word Recognition Using the IFNENIT- Database*," Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003).