Handwritten Myanmar Character Authentication between Exhibit and Specimen

YadanarKhaing, Zin Mar Win and Myint Myint Sein
University of Computer Studies, Yangon
yadanarkhaing@ucsy.edu.mm, zinmarwin@ucsy.edu.mm, myint@ucsy.edu.mm

Abstract

Handwriting authentication is considered as the most natural method of authenticating a person's identity compared to other biometric and cryptographic forms of authentication. It involves specialize human skills and it also can be used for forensic tasks, person authentication, crime suspect and financial activities. Handwriting authentication is not a simple task but useful for practical applications. The main aim of this paper is to evaluate the accuracy of Myanmar Character Authentication of Handwriting between Exhibit and Specimen, which is able to identify handwritten characters and words several different writers. The modified Euclidean distance is used for character authentication. Experimental results verified the effectiveness of this method.The development of this system for Myanmar Character identification is supported as a result to obtain good authentication accuracy rate. Therefore this system can be applied effectively and efficiently in Myanmar characters authentication of different writers.

1. Introduction

Units of meaningful handwriting can be categorized into the following four types: character, word, phrase and sentence. The dictionary definitions of the four units are: character: a graphic symbol as a hieroglyph or alphabet letter or digit used in writing or printing: word: a written or printed character or combination of characters representing a spoken word; phrase: a word or group of words forming a syntactic constituent with a single grammatical function; sentence: a word, clause, or phrase or a group of clauses or phrases forming a syntactic unit. Handwritten character, word, phrase and sentence are defined as those that are written by a human hand.

Handwriting is one of the easiest and natural ways of communication between humans and computers. Further, handwritings can be used for forensic tasks and for person authentication.

Zheng et al. [14] addressed different character recognition problems have their own specific characteristics. The state-of-art OCR technologies take different recognition approaches, which are most effective, to recognize different types of characters. How to identify characters type automatically, and then use specific recognition engines, has not brought enough attention among researchers. Most of the limited researches are based on the whole document image, a block of text or a text line. This paper addresses the problem of character type identification independent of its content, including handwritten or printed Chinese character identification, and printed Chinese/ English character identification, based on only one character. Exploiting some effective features, such as run-lengths histogram features and stroke density histogram features, they have got very

promising result. Theidentification correct rate is higher than 98% in their experiments.

Authors [13] presented a method for off-line writer identification, using the contours of fragmented connected components in mixed-style handwritten samples of limited size. The writer is considered to be characterized by a stochastic pattern generator, producing a family of character fragments (fraglets). Using a codebook of such fraglets from an independent training set, the probability distribution of fraglet contours was computed for and independent test set. Results revealed a high sensitivity of the fraglet histogram in identifying individual writers on the basis of a paragraph of text.

Bulacu et al. [10] evaluated the performance of edge-based directional probability distributions as features in writer identification in comparison to a number of non-angular features provide additional valuable information. Rank-combination was used to realize a sparse-parametric combination scheme based on nearest neighbor search. Limitations of the proposed methods pertain to the amount of handwritten material needed in order to obtain reliable distribution estimates. The global features treated in this study are sensitive to major style variation (upper- vs lower case), slant, and forged styles, which necessitates the use of other features in realistic forensic writer identification procedures.The rest of the paper is set up as follows: Section 2 provides an overview of Handwritten Myanmar Character Authentication System.Section 3 presents preprocessing steps for character extraction. Section 4 presents the steps of character extraction. Section 5 presents the techniques of feature extraction and matching and Section 6 describes the experimental results of this study.

2. Handwritten Myanmar Character Authentication System

Research into handwriting authentication has focused on two streams, off-line and on-line writer identification. This system focuses only on the off-line identification. Off-line systems are based on the use of computer image processing and pattern recognition techniques.

The system consists of six steps: namely scanning, preprocessing, character extraction, feature extraction, matching, authentication and finally the result produced.

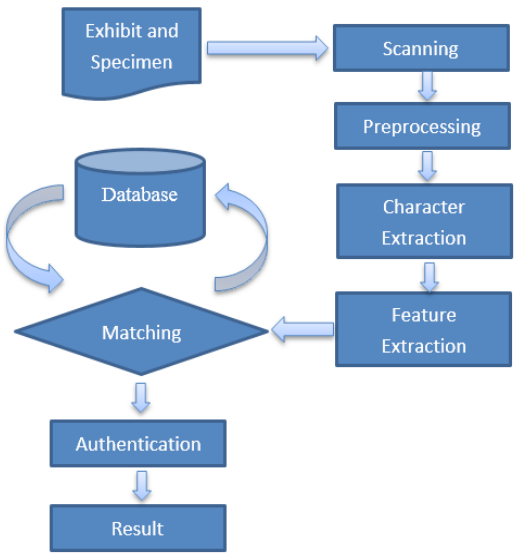


Figure 1: Handwritten Myanmar Character Authentication System

Step 1:Scan the exhibit and specimen handwriting documents.

Step 2: Remove some grains noise in the exhibit and specimen handwriting documentsand also remove straight underline.

Step 3: Randomly extract at least any ten characters from these exhibit and specimen handwriting documents. But comparison characters need to be the same character shape.

Step 4: Features points can be obtained from these extracted characters.

Step 5:Match the two characters by finding the shortest distance among these two feature points sets.

Step 6: Authenticate that these two handwritings are written by one writer (same writer) or not (different writers).

3. Preprocessing For Character Extraction

Preprocessing is one of the most important steps required for image processing. The goal is to reduce irrelevant information. In this system, the first step for processing is preprocessed both the exhibit and specimen handwriting documents. Noise removal and straight underline removal are included this step. For the purpose of next step, character extraction this preprocessing step is basically required. It was necessary to preprocess all handwriting images.

Therefore, first have to remove noise from these exhibit and specimen handwriting images. Preprocessing is to remove noise artifacts and then normalize the handwriting. E.g., underline removal and noise removal. Noise may be in the scanning context; this refers to random, incorrectly read pixel values, normally due to electrical interference or device instability. There are several ways that noise can be several ways of methods for noise removal.

3.1. Exhibit and Specimen Paper Documents

In this system, exhibits may be piece of paper or paper, which have some handwriting on it. This is also the evidence or the sample handwriting examiner examines this is written by he or she or someone. To examine this they compare this handwriting to the other handwriting. The other handwriting sample can get from suspect person. This other sample handwriting is specimen handwriting for this system.

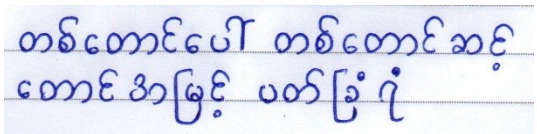
Therefore, exhibit and specimen are a paper that consists of processing images. Exhibit and specimen is also the input of off-line system. This off-line processing can be done with any type of document.

Exhibit is sample writing, compared to another sample of writing (specimen or exemplar).

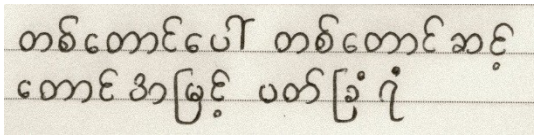
Specimen is a comparison sample could be writing submitted as having been written by a particular person or it could be another sample handwriting Specimen writings are the most common form of comparison samples.

First and foremost, the original document is captured through the scanning process. For the system Epson 210 model scanners are used. The documents (both exhibit and specimen) are scanned at a resolution of 300 dpi (dot per inch) and stored by grayscale images as shown in figure 2 (b).

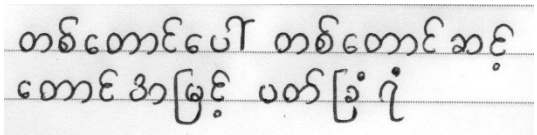
This image is converted to binary image by thresholding. Threshold is a value for text type. For pixel value above the threshold value, it is represented by 1 (i.e., white) and for pixel value below the threshold value it is represented by 0 (i.e., black) as shown in figure 2 (c). Some of the common operations performed prior to recognition are thresholding, the task of converting a gray-scale image into a binary black-white image; noise removal, the extraction of the foreground textual matter by removing such as textured background, salt and pepper noise. Figure 2 (d) is the image after preprocessing.



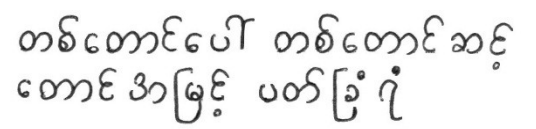
(a)



(b)



(c)



(d)

Figure 2. (a) Original Color Image (b) Gray Scale Image (c) Binary Image (d) Image after preprocessing

4. Character Extraction

Myanmar characters are very complex; they contain circles, curves, lines and dots. Handwriting is easy to obtain but different people have different handwriting styles and use different writing instruments. So the identification task of these characters is so difficult that in order to get optimum solution because of large variation in personal handwriting style, different handwriting instruments, segmentation problem and large vocabulary, etc.

We developed the Fast Fourier Transform method (FFT) for character extraction. Our proposed method is an example of a character shape analysis. Therefore, we propose to extract the distinctive features of hand written characters on Myanmar handwriting documents of exhibit and specimen.

4.1. Cropping the Character from Exhibit and Specimen

The crop method objective is to locate a rectangular box around an object of interest and remove other objects outside this area. If the isolated character is the object of interest in the exhibit and specimen, this could be easily done.

Since exhibit and specimen of (Handwritten Myanmar Characters) documents have several types of characters and paper style (with/without line). Among these characters and words, choose some characters for matching.

From these several types of characters we choose particular types of character for matching. Therefore we choose randomly any eight character or more according to the size of paper. If we choose first character from the exhibit, we also choose like this character from the specimen. In this system we choose not only isolated character but also some words according to the handwriting.

First manually crop any character as shown in figure 3 and then auto crop this manually crop character as shown in figure 4. This cropping does until choosing final character. After cropping all character both from exhibit handwriting document and specimen handwriting document, we get the set of these characters feature points for matching.

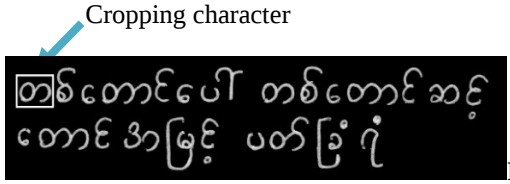


Figure 3. Extracted character

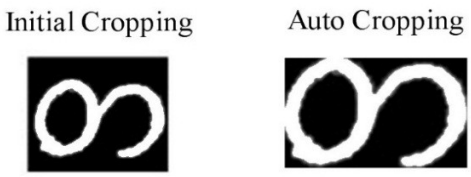


Figure 4. Cropping Character

5. Feature Extraction and Matching

The objective of feature extraction is to capture the essential characteristic of the symbols, and it is generally accepted that this is one of the most difficult problems of pattern recognition. The techniques for extraction of such features are often divided into three main groups, where the features are found from: the distribution points, transformation and series expansions.Different feature extraction methods were designed for different representations of the characters, such as solid binary characters, character contours and skeleton, or gray level images of each individual character.

In this system, the Euclidean distance is used for feature points matching and identification. Experimental result verified the effectiveness of the proposed method. The two extracted characters feature points are matching at the shortest/minimum distance at D. If pixel point P1 is stand for (x1,y1) and pixel point P2 is stand for (x2,y2). The next phase, the Euclidean Distance is

$$D = \sqrt{(x2-x1)^2+(y2-y1)^2} \qquad (1)$$

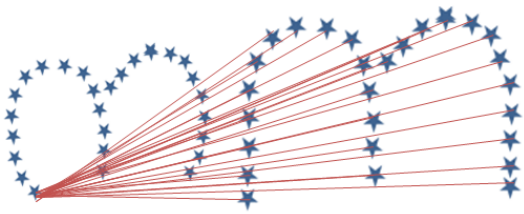


Figure 5. Shortest Distance Matching

Extracted character's feature points of exhibit handwriting is identical to the character's feature points of specimen handwriting if the Euclidean distance is minimum/shortest at D.

6. Experimental Results

The experiment used 76 handwriting documents images obtained from different writers of 38 persons. Each person wrote with only pen. Sample handwriting were gathering from Ph.D-9 Batch at University of Computer Studies, Yangon, Myanmar. The handwritten data include different characters and words. The size and position of the exhibit handwriting does not need same to the specimen handwriting images as shown in figure 6.

The system finally authenticates that the two handwriting are written by same writer or not after the steps of preprocessing, character extraction, feature extraction, matching.

(a) **(b)**

Figure 6. (a) Exhibit handwriting (b) Specimen handwriting

On the average, 98.7% of the characters authentication. Table 1 shows the correct authentication percentage for all exhibit and specimen handwriting images by the following formula:

$$\text{Accuracy rate} = \frac{\textit{Correct Images}}{\textit{Total Images}} \times 100$$

Table 1. Character Authentication Percentage

	Correct Authentication rate (%)	Incorrect Authentication rate (%)
Total 76 images	(75/76)*100=98.7	(1/76)*100=1.3

The percentage of correct rate and incorrect rate for character authentication is shown in figure 7.

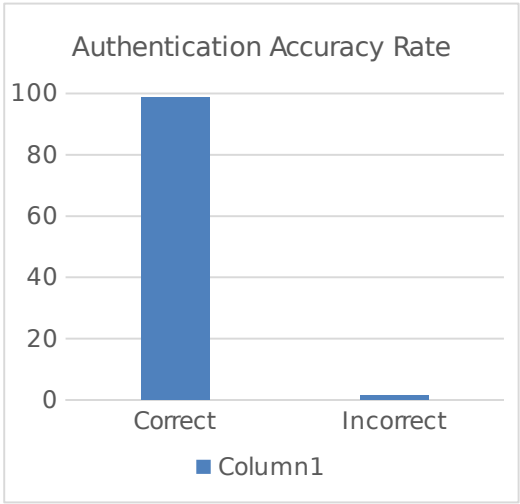


Figure 7. Correct and Incorrect Rate for Character Authentication

7. Conclusion

In this system, we presented the authentication of the handwritten Myanmar characters on Exhibit and specimen documents based on the shortest distance based method.

We well separated handwritten character on the exhibit and specimen handwriting documents. The performance of the system is good to authenticate.

This system identifies seventy-six different handwriting documents in 98.7% correct rate and 1.3% incorrect rate. Different characters were extracted from manually crop from exhibit and specimen handwriting images, which is written by Myanmar Handwritten character with different writing style and different size of different writers. For this authentication accuracy of different handwriting with different writers is between 85% and 95%

For this system, we use Epson 210 series scanner and JPEG images. The system use the programming language is MATLAB and the program running for this system is MATLAB 2015 version.

References

[1] SoeHae Mar, “A Matching Process of Handwriting between Exhibit and Specimen”, Proceedings of 2nd International Conference on Computer Application 04, January 8, 2004, Myanmar, pp.25-27.

[2] SoeHae Mar, “Feature Extraction and Identification of Handwriting between Exhibit and Specimen”, Proceedings of 3rd International Conference on Cjomputer Application 05, March 9-10, Myanmar, pp.471-475.

[3] SoeHae Mar, “Myanmar Character Identification of Handwriting between Exhibit and Specimen”, Proceedings of 6th International Conference on APSITT 2005, November 9-10, Myanmar, pp.95-98.

[4] N.A.A.H, “Automatic Extraction of Payee name and Legal Amount from Bank cheque”, In Proceeding of Fifth International Conference on Computer Application, February 2007.

[5] H.H.Thaung, “Recognition on Handwritten Date on Bank Checks using Zoning Methods,” PhD Dissertation University of Computer Studies, Yangon, 2004.

[6] K.L.Thaw, “Recognition for Myanmar Handwriting digits and Characters”, In Proceeding of Fifth International Conference on Computer Application, pp.415-419 February 2007.

[7] K.Sandar, “Off-line Myanmar Handwriting recognition using Hidden Markov Models”. In Proceedings of Third International Conference on Computer Application, pp.463-469 March 2005.

[8] M.M. Nge, “ Automatic Segmentation of Handwriting Myanmar Digit and Script on the Boucher Sheet”, In Proceedings of Fourth International Conference on Computer Application , pp.499-505 March 2005.

[9] A. Ensefia, T.Paquet and L.Hettte, “Information Retrieval Based Writer Identification”, in 11th Conference of the International Graphonomics Society, Scottsdale, Arizona. IGS’2003, pp. 320-323.

[10] M. Bulacu, L. Schomaker and L. Vuurpij, “Writer Identification using Edge-Based Directional Features”, Proc. Of 7th Int. Conf. on Document Analysis and Recognition (ICDAR 2003), IEEE Press, August, Edinburgh, Scotland 2003, vol. II, 3-6, pp. 937-941.
[11] M. Blumenstein, C.K. Cheng and X. Y. Liu, “New Preprocessing Techniques For Handwritten Word Recognition”, School of Information Technology, Queensland, Australia.
[12] B. Zhang Sargur and N. SrihariSangjik Lee,” Individuality of Handwritten Characters”, Proceedings of the Seventh International Conference on Document Analysis and Recognition IEEE (ICDAR’ 03) 2003.
[13] L. Schomaker, M. Bulacu and K.FRane, “Automatic Writer Identification Using Fragmented Connected-Component Contours”, Proceedings of the 9th Int’l Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004).
[14] Y. Zheng, C. Liu and X. Ding, “Single Character Type Identification”, Department of Electronic Engineering, Tsinghua University Beijing 100084, P.R. China.