DICTA2002: Digital Image Computing Techniques and Applications, 21—22 January 2002, Melbourne, Australia

FORM PROCESSING FOR THAI DOCUMENTS

Wichian Premchaiswadi*, Nucharee Premchaiswadi**,
Kridsada Vilailuk***, and Seinosuke Narita****
*Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand, E-mail:wichian@it.kmitl.ac.th
**Faculty of Information Technology, Dhurakijpundit University, Bangkok 10210, Thailand
***Faculty of Engineering, Vongchouvaritkul University, Thailand
****School of Science & Engineering, Waseda University, Tokyo, Japan

Abstract

Useful information is usually collected from the filled forms. This paper describes a scheme to recognize handprinted Thai characters on form and automatically stores them into database. The system consists of two main components: form management and handprinted Thai character recognition. The form management is used to create a template for a form being process. A user identifies the positions of the interested information together with their attribute names being stored in the database. Then, a database is generated to keep the contents of this information. The generated template can be stored for later use of the same form as well. The character recognition process needs to recognize characters only in the fixed area specified in the template. The handprinted Thai character recognition algorithm based on specific features of Thai characters is also proposed. The number of components and contact points when a character is horizontally segmented at its central of gravity are used to classify some class of characters in the recognition process. The effectiveness of the system is implemented and tested by using Visual Basic and Microsoft Access database.

1. Introduction

The usefulness of a system to convert documents into text has been so apparent. The literally many of research papers have touched on the subject. The most obvious and common source of structure text is the paper "form" [1]. Forms represent a large percentage of paperwork generated today. Pre-specified fields serve as expectations in order to allow a goal-directed processing of the different imprints of interest. The degree of formal specification varies depending on the domain and types of document. It ranges from documents with many preprinted fields like, for instance, life insurance applications to quasi free-format documents, e.g. business letter with complicate the prediction where the regions of interest are located. Form information is most commonly information destined for database applications. It could be of great benefit to be able to read forms directly into database.

This paper presents a method of form processing which can be used with any specific preprinted form. And a handprinted Thai character recognition algorithm is proposed for convert the image in these specific areas into text and automatically store them into a database.

2. Thai Character Set and Characteristics of a Thai sentence

Thai character set consists of 44 consonants, 15 vowels and 8 voice tones as shown in Table 1 [2]. The character "อ" in vowels and voice tones is used only to represent the location of vowels and voice tones characters. In writing Thai words, it will be replaced by one of a consonant character. There is no spacing between words and no special mark to identify the end of a sentence. The Thai vowel forms do not follow initial consonants; some are placed before the initial consonants, some after the consonants, some above the consonants, and some underneath the consonants. The vowels that are "complex" forms (i.e. composed of more than one part) can be placed around the consonants.

Table 1: Thai character set.

Type	Member
Consonants	ก ข ค ง จ ฉ ช ซ ฌ ญ ฎ ฏ ฐ ฑ ฒ ณ ด ต ถ ท ธ น บ ป ผ ฝ พ ฟ ภ ม ย ร ล ว ศ ษ ส ห ฬ อ ฮ
Vowels	ะ า อิ อี อึ อื อุ อู เ แ โ ไ ใ อั อี๋
Voice tones	อ่ อ้ อ๊ อ๋ อ์ ฯ ๆ

A Thai sentence is consists of up to the maximum of three zones. A character may occupy in one of the three zones namely: the upper zone (UZ), central zone (CZ) and lower zone (LZ). The upper zone is also classified into two sub-zones namely: the upper zone 1 (UZ1) and upper zone 2 (UZ2) as shown in Figure 1. The multi-level structure of a Thai sentence made it looks very complicate and difficult to recognize. However, on the opposite side, if the zone information is obtained, it will be very useful to classify characters into groups with a smaller number of members. The zone information is obtained by using histogram and each character block is obtained by using edge detection algorithm.

Figure 1. A sample of a Thai sentence.

3. Form Management

Forms are usually defined for the specific proposed as printed document with fields for insertion of requested information as an example shown in Fig. 2. In any form, there are only some specific areas used to input data while the content of the other parts are already known before. The objective of this process is to get the information of how the form appears and specify areas used for the recognition process. Each specific area is named and used as a name of attribute in the generated database.

The information of a form can be obtained by scanning an empty form with the scanner. The settings are adjusted in the same procedures, as they would be for the filled forms. After the empty image is acquired, a user specifies the interested areas on a form. The designed form management has functions as follows:

1. scrolling of input forms
2. position marking – to specify the area of interest
3. store and loading – used for storing specified positions with the name of attribute to be store in database and indicate the area to be recognized in the recognition process.

Figure 2. A sample of a form.

4. Handprinted Thai Character Recognition

Although many algorithms have been proposed for the recognition of Thai handprinted characters [3][4]. However, it still has some limitation to put in practical work. In this section, the algorithm for the handprinted Thai character recognition algorithm based on specific features of Thai characters is described [5]. The advantage of this method is the independent of size, style, thickness and some degree of slant characters.

4.1 Important features of Thai characters

In the proposed scheme, the center of gravity is used to separate each character into two parts namely: the upper part and lower part as shown in Figure 3. Each part will have dominant characteristics that can be used as the criteria in classification as the following.

a) The number of connected pixels

In the upper half, a number of connected pixels can be classified into 3 groups due to the number of connected pixels as shown in Figures 3.

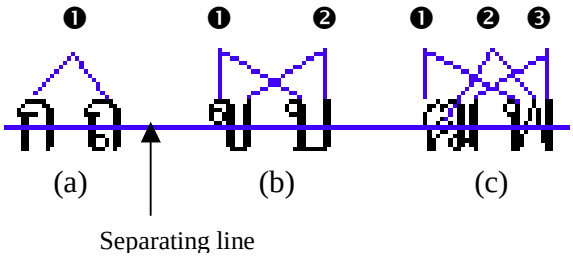

Figure 3. The number of connected pixels in the upper half of a character.

In the same way, the number of connected pixels in the lower half of a character can be classified into 2 groups due to the number of connected pixels as shown in Figure 4.

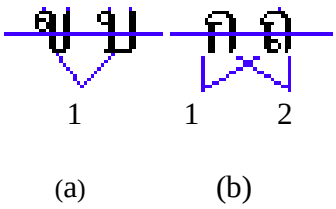

Figure 4. The number of connected pixels in the lower half.

b) The number of touching points at the separating line

With the number of touching points at the separating line, characters can be classified into 4 groups: one touching point, two touching points, three touching points and four touching points as shown in Figure 5.

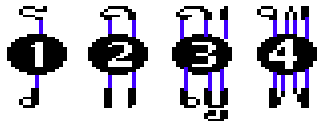

Figure 5. The number of touching points at the separating line.

c) Head of a character

A head of a character is defined as the circle portion of a Thai character. It is usually used as the starting point in writing Thai characters. When considering the writing Thai characters from starting from a head of character, we can classify Thai characters into 5 categories as follows:

1) Writing in clockwise and go down like q
2) Writing in clockwise and go up like b
3) Writing in anti-clockwise and go down like p
4) Writing in anti-clockwise and go up like d
5) No head

d) Width and height of characters

In Thai character set, there are some characters having width or height different from that of other characters. Therefore the characteristics can be used to classify characters into one of the three groups as follows.

1) Group 1: Characters that have width more than the width of average characters such as characters ฌ ญ ฅ ณ
2) Group 2: Characters that occupy more than one zone, in both the upper zone and central zone such as characters ป ฝ ฟ ศ ส พ ฮ ไ ใ โ
3) Group 3: Characters that occupy more than one zone, in both the central zone and lower zone such as characters ฎ ฏ ฐ ฤ

If we consider only the character images in central zone having normal width, it can give useful information for a recognition process too. For example, if we cut the rare part of characters such as **ฌ, ญ, ณ**, they will all look like a character "**ถ** " as shown in Figure 6(a). This method is also automatically helping in segmenting some touching character as shown in Figure 6(b).

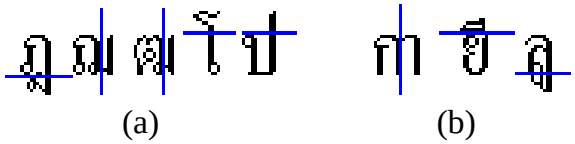

Figure 6. (a) cutting some parts of character (b) touching character segmentation.

4. Experimental results and conclusion

The scheme is implemented by using Visual Basic and Microsoft Access Database. The method was tested with data in forms by using handprinted Thai characters of 10 persons of age between 7 to 35 years. The data was scanned with 300 dpi, half-tone and scaling factor equal to 50 percents. The total number of characters used in the test is 2580 characters and the proposed method can recognize correctly 2250 characters, approximately 87 percents. An example of experimental result is shown in Fig. 7 to Fig. 9.

Fig. 7: The defined recognition areas

Fig. 8: Images and their recognition results

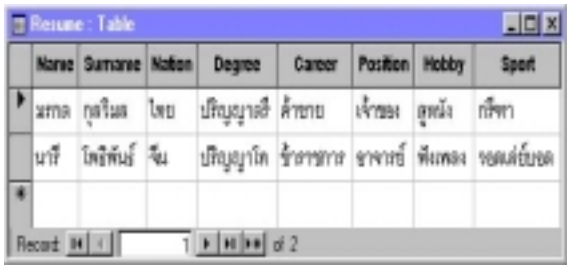

Figure 9 A database that used to store the recognition results.

The method of form processing and handprinted Thai characters recognition are described. It also demonstrates the usefulness of the application. Users can define the areas of recognition. There are still many works to do to improve the system such as the automatic form selection and the recognition algorithms to increase the recognition rate of the system. The use of the context-based recognition could also improve the recognition rate of the scheme because the scope of context in form is usually known in advance.

References

[1] Kelly Anderson and William Barrett "Context Specification for Text Recognition in Forms" SPIE Vol. 1384 High-Speed Inspection Architectures, Barcoding, and Character Recognition . 1990

[2] Nucharee Premchaiswadi, Wichian Premchaiswadi and Seinosuke Narita, "Segmentation of Horizontal and Vertical Touching Thai Characters", IEICE Trans. Fundamental, VOL-E83-A., No.6, June 2000, pp. 987-995.

[3] S. Manus, "The reconition of hand-writing by considering head of character" EECON-11, 1988.

[4] K. Limpanon, "The method of hand-writing recognition by using standard frame," EECON-16, 1993.

[5] K. Vilailuk, W. Premchaiswadi and N. Premchaiswadi "Thai Character Recognition by using Specific Feature," EECON-21, pp.90-93, 1998.