Broken Characters Identification for Thai Character Recognition Systems

NUCHARIEE PREMCHAISWADI*, WICHIAN PREMCHAISWADI*
UBOLRAT PACHIYANUKUL**, SEINOSUKE NARITA***

*Faculty of Information Technology, Dhrakijpundit University, Bangkok, 10210, THAILAND
**Faculty of Information Technology, Mahanakorn University, Bangkok, THAILAND
***Department of Electrical, Electronics and Computer Engineering, School of Science and Engineering, Waseda University, Tokyo, JAPAN

Tel. (81)3-5286-3374 Fax. (81)3-3202-9940

Abstract: - This paper presents a scheme for solving the problem of broken characters in Thai character recognition systems. The proposed scheme consists of two main steps namely: broken character identification and connection of broken characters. There are two techniques used to identify broken characters as follows: 1) an overlapping area and 2) character code. The specific characteristics of Thai characters are also employed in the scheme such as position of the head, leg and endpoint of characters. The experimental results show that the scheme can identify and repair broken characters both in normal and bold font efficiently. The recognition rate of commercially available software is improved significantly.

Key-Words: - Broken characters, Thai character, Character recognition, Vertical broken, Character code

1 Introduction

Documents or forms are typically prepared only as originals. If additional copies are needed, they are made from the original. Each time copies are made, the quality of the image deteriorates, which can result in broken characters. The output quality from a typical facsimile is usually low in terms of clarity, because of low image resolution. In Thailand, broken characters are usually seen in fax-transmitted documents, especially from the transmission of copied (non-original) documents. When rather thin sheets of paper are fed into a printer, small folds are sometimes be found, which can result in broken characters in the folded areas.

Sheets which have already been used on one side are often re-used on the other side for economic reasons, and folding can occur easily due to changed characteristics caused by heat from the first use. This phenomenon can often be found in sheets re-used in laser printer or copy machines.

The completeness of the input character images affects the accuracy of a character recognition system significantly. Therefore, much research on character recognition systems has focused on pre-processing processes such as segmentation of touching characters [1], document image binarization [2]. Although much research has been done on Thai character recognition [3] [4], this problem still exists and causes erroneous results in recognition systems. Table 1 shows an example of a recognition result obtained by using commercially available Thai language OCR software for recognition of broken character images.

Table 1. Example of an OCR recognition result for broken characters.

<table>
<thead>
<tr>
<th>Test data</th>
<th>Correct result</th>
<th>ThaiOCR Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>ถูกอักขระ</td>
<td>ถูก</td>
<td>ถูกอักขระ</td>
</tr>
</tbody>
</table>

When a character image is broken into several segments, these segments cannot be treated as a single complete character. A segment is defined as a rectangular area that covers connected pixels. A simple merging of these small segments is not feasible, since it is not clear beforehand which segments belong to which character. Thus, a more advanced technique is required to solve this type of
problem: grouping segments of characters to form an entity representing a character. A method for reconstructing damaged characters was proposed in [6], but it can only be used for characters damaged by lines of a table. It does not address the problem of broken characters and cannot be applied in the case of Thai characters.

This paper is an extension of the method proposed by the authors in [5]. The proposed scheme can be applied to characters broken along the vertical line which can not be supported in [5].

2 Thai Character Set and their Characteristics
Thai character set consists of 44 consonants, 15 vowels and 8 voice tones as shown in Table 2 [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>
<thead>
<tr>
<th>Type</th>
<th>Member</th>
</tr>
</thead>
</table>
| Consonants | กขฃคฅฆงจจญฉชฌญฎฏฐฑฒณดตถทธนบปผฝพฟภมยรล
               | เชซสหฮอเขาซื้อกลุ่ม |
| Vowels     | ะาอิอีอึอือุอูเแไใอัอ็ |
| Voice tones| อ่เออสาม |

Table 2. Thai character set

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 Fig. 1. The multi-level structure of a Thai sentence makes 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 algorithms.

3 Pre-processing process
The objective of the pre-processing process is to determine the segment boundaries (SBs) and classify each SB into one of the three zones mentioned in the previous section. An SB is defined as the smallest rectangular area that covers all of a set of connected pixels. Each SB may consist of a single complete character, several connected characters, or a part of a broken character. The pre-processing process consists of two steps: segment boundary determination and zone classification [1]. An example of input and output of this process is the sequence of SBs, as shown in Fig. 2.

![Fig. 1. A sample of a Thai sentence.](image)

![Fig. 2. An input and output of the pre-processing process.](image)

4. Identification of Broken Characters
In this paper, only vertically broken characters are considered. The horizontally broken characters already reported by the authors in [5]. In this process, there are two techniques that used to identify broken characters namely: an overlapping area and character code as follows.

4.1 An Overlapping Area
The coordinate in x-axis of each SB is used to determine whether there is any overlapping area. The overall process of the step is shown in Fig. 3.
The process to find overlapping area

Normal character

Broken character

Fig. 3. The overall process of finding an overlapping area.

Table 3. Character code in 16 patterns.

<table>
<thead>
<tr>
<th>No.</th>
<th>Structure</th>
<th>Character code</th>
<th>Character member</th>
<th>No.</th>
<th>Structure</th>
<th>Character code</th>
<th>Character member</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1110</td>
<td>1110</td>
<td>กินข้าวตัดกินข้าว</td>
<td>9</td>
<td>0011</td>
<td>0011</td>
<td>Broken character</td>
</tr>
<tr>
<td>2</td>
<td>1011</td>
<td>1011</td>
<td>ข้าวข้าวข้าว</td>
<td>10</td>
<td>1000</td>
<td>1000</td>
<td>Broken character</td>
</tr>
<tr>
<td>3</td>
<td>1111</td>
<td>1111</td>
<td>งงงงงงงงงง</td>
<td>11</td>
<td>0010</td>
<td>0010</td>
<td>Broken character</td>
</tr>
<tr>
<td>4</td>
<td>0110</td>
<td>0110</td>
<td>Broken character</td>
<td>12</td>
<td>0100</td>
<td>0100</td>
<td>Broken character</td>
</tr>
<tr>
<td>5</td>
<td>1001</td>
<td>1001</td>
<td>ไว้</td>
<td>13</td>
<td>0001</td>
<td>0001</td>
<td>Broken character</td>
</tr>
<tr>
<td>6</td>
<td>0111</td>
<td>0111</td>
<td>บูรณะ</td>
<td>14</td>
<td>0101</td>
<td>0101</td>
<td>Broken character</td>
</tr>
<tr>
<td>7</td>
<td>1101</td>
<td>1101</td>
<td>อี</td>
<td>15</td>
<td>1010</td>
<td>1010</td>
<td>No character</td>
</tr>
<tr>
<td>8</td>
<td>1100</td>
<td>1100</td>
<td>Broken character</td>
<td>16</td>
<td>0000</td>
<td>0000</td>
<td>No character</td>
</tr>
</tbody>
</table>
4.2 Specific feature and Character Code

The process is used to identify broken characters that can not be found by using an overlapping area described previously. In this process, the specific feature of Thai character such as head of a character, the circle used as the starting point of writing a character as shown in Fig.4.

![Fig. 4. Head position of a character](image)

The character code is constructed by finding the boundary from center of the character in four directions namely: top, bottom, left and right. The example of finding a character code is shown in Fig. 5.

![Fig. 5. Examples of character code.](image)

The character code of Thai character set can be classified into 16 patterns as shown in Table 3. The character code is not only used to identify Thai broken characters but also determined the area that characters could be connected together.

5. Connection of Broken Characters

The process is used to connect the broken pieces of character together. The process consists of two steps namely; connection by considering the overlapping area and connection by using character code.

5.1 Connection by considering the overlapping area

In this process, the position of broken character already known from previous processes. Therefore, the shortest distance is selected for connecting these broken characters.

5.2 Connection by using Character Code

All characters that passed to step 5.1 are also passed to the process to check by using character code again. The character code and head position of a character are used to determine which part, the next or previous block, should be used to form a complete character as examples shown in Fig.6.

![Fig. 6. Connection by character code](image)

6. Results and Conclusion

The method was implemented by using Visual Basic Version 6.0. The test documents are obtained from image scanner at resolution 300 dpi. Fig. 7 shows examples of broken character images and the reconstruction result of these images are shown in Fig. 8.

![Figure 7. Examples of broken character images.](image)
Figure 8. Result of reconstruction.

The effectiveness of the proposed scheme can be found by using the commercially available Thai character recognition software. The recognition results of character images shown in Fig. 7 and Fig. 8 by using the commercially available Thai character recognition software are presented in Table 4.

Table 4. Recognition results of character images in Fig. 7 and Fig. 8.

<table>
<thead>
<tr>
<th>Recognition with ThaiOCR Version 1.5</th>
<th>Broken characters</th>
<th>Repaired characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>นางมณีตาสาระ 80สี อก</td>
<td>เพิ่มเติมจักษาฮอร์โมน</td>
<td>เพิ่มเติมจังจังฮอร์โมน</td>
</tr>
<tr>
<td>ยา อาหารยา</td>
<td>ใจจริงไม่ยากมาก</td>
<td></td>
</tr>
<tr>
<td>[เรื่องาน] ปุณฑริก</td>
<td>ฉันต้องทำเรื่องกับคุณ</td>
<td></td>
</tr>
<tr>
<td>4เรื่องสำคัญ[เรื่องงาน]</td>
<td>ฉันต้องทำเรื่องกับคุณ</td>
<td></td>
</tr>
<tr>
<td>ผู้ที่มีสิ่งของยิ่งใหญ่</td>
<td>การดื่มของเมื่อนเสริม</td>
<td></td>
</tr>
</tbody>
</table>

The results shown in Table 4 also are in the form as they are in the output of the software. It can be seen that the software can not recognize almost of these broken characters if they are not repaired.

6. Conclusion

A scheme for detection and reconstruction of broken printed Thai characters has been proposed. Broken characters are identified by using two techniques namely: an overlapping area and character code. The use of a projection profile and the overlapping areas of broken segments make the reconstructed character images very similar to the original characters. It can be seen that the correctness of the recognition results of repaired characters is much higher than that of the results of broken characters. However, there are still some erroneous recognition results, although the reconstructed characters are very similar to the original characters. The rate of recognition errors also depends on the efficiency of the recognition software itself. From the experimental results, it can be concluded that the proposed scheme can be used to reconstruct broken characters efficiently and will be useful in significantly improving the recognition rate of Thai character recognition systems.

References: