The Clustering Technique for Thai Handwritten Recognition

Ithipan Methasate, Sutat Sae-tang
Information Research and Development Division
National Electronics and Computer Technology Center
National Science and Technology Development Agency
112 Phahol Yothin Road, Klong Luang,
Pathumthani 12120, Thailand
ithipan.methasate@nectec.or.th, sutat.sae-tang@nectec.or.th

Abstract

This paper describes an algorithm for clustering free-style Thai handwritten character models. The algorithm groups the characters that have a similar structure. Firstly, the algorithm begins with the vertical stroke detection. The vertical stroke is an important Thai character structure. Secondly, the character area will be divided into 7x10 blocks by using the stroke information. Then, the pixel distribution feature is calculated from each block. The features will be trained using backpropagation neural network. Finally, the confusion matrix will be used to analyze the result in a clustering process. The characters are divided into 21 groups and the accuracy of the clustered model is 97.60 percent.

1. Introduction

In Thailand, the research on handwritten character recognition has been active for a decade. Most of them are focused on the feature extraction and recognition model. In the early days, most researches work involved techniques for detecting an important structure of the character[1,2,5] such as loops, end points, curls and etc. Choruengwiwat,P.[2] applied the stroke changing sequence (SCS) together with the structure features. The recognition with cavity features was proposed by Phokharatkul, P., Kimpan C.[4]. They also applied fourier descriptor technique with character edge information[3]. In regular writing, the structure informations are usually degraded and difficult to detect. So, there are some researches attempt to solve these problems. Methasate I., et al.[5] presented the fuzzy syntactic method. Theeramunkong, T., et al.[6] used Island-based projection with interpolated N-gram Models and Hidden Markov Models. However, these problems still need solution.

This paper describes a method for clustering Thai character by using global feature. The feature is extracted using vertical strokes of the character. The proposed method work as a rough classifier. In a complete system, the character will be classify the local structure features. But, it is out of the scope of this paper.

2. Thai Language

Thai character set consists of 44 consonants, 17 vowels, 4 tones and 2 punctuation marks, as shown in Table 1. From Fig. 1, Thai sentence structure is divided into 3 main levels as in Latin family language. Additionally, Thai language has 2 sub-levels in the upper level, when the upper level vowel belongs to the same word as tone. Thai language has some characteristics as Latin family style such as, the word is composed of characters, tones and vowels. The big difference between Thai and Latin is Thai does not have space between words. It is hard to segment Thai words in a sentence. Furthermore, Thai characters have more complex structure than Latin. Touching and cursive character problem is also complicated as it may occor in both vertical and horizontal directions.

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 อ ฮ
Vowels	ะ า ิ ี ึ ื ุ ู เ แ โ ใ ไ ั ็ ํ ์
Tones	่ ้ ๊ ๋
Punctuation Marks	ฯ ๆ

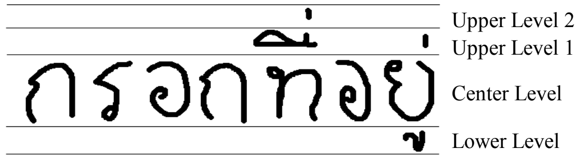


Figure 1. Thai sentence structure

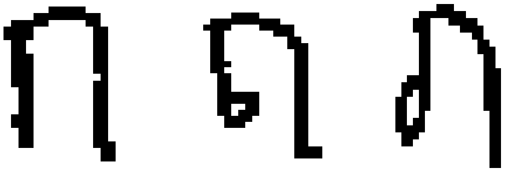
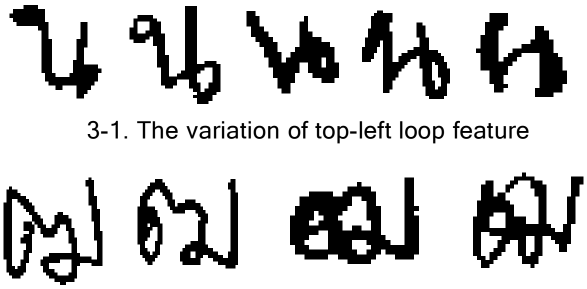


Figure 2. Sample of some similar characters

Some Thai characters have similar structures as shown in Fig 2. They can be identified by using local features, such as loop or curl. However, some character distortions are from the writing style and the writing environment. The variation of feature pattern may be the reason of failure in the feature extraction technique.

As mention above, the variation of handwritten character directly affects to the local features. So, the local feature extraction can not work efficiently with ambiguous handwriting. On the other hand, the global feature,such as the projection or pixel distribution, still has the characteristic of the structure, but this is not enough to recognise them completely.



3-1. The variation of top-left loop feature

3-2. The touch of internal stroke may cause the variation of edge feature

Figure 3. Some structure features variation

3. Character Structure Feature

In the proposed modeling method, the character model is represented by a set of global features. The distribution of the pixels in spatial domain is used as a global feature. The pixel distribution feature plays an important role in printed Thai OCR, but it could not work efficiently when applied in a various problems as handwritten domain. Fortunately, Thai consonants have a remarkable vertical strokes. With this feature, the adaptive block can be applied. Then the image area is divided into 10x7 blocks. The size of the blocks are determined by considering from the vertical stroke and they may not be assigned equally. The pixels in each block are averaged and used as an input feature to the neural network.

3.1 Vertical Stroke Detection

From Fig 3-2, the same character "ฒ" are shown in various writing styles. The local features are distorted by touching problem, but the 3 main vertical strokes are still detectable. And the local features are in the same relative position with the strokes. So, the vertical strokes are used as reference points. The procedure for the vertical stroke detection is as follows.

- Project the pixel in the vertical direction.
- Smooth the projection by averaging technique.
- Find projection peaks and local minimum around the peaks.
- Find the projection hill by using the Triangle method [7]. Fig 4 depicts the method to locate the base of hill peak. A line is constructed beweeen the peak A and the reference point, image boundary, B. For each x-axis value from A and B, the distance(d) from the top of the projection to the connecting line AB was calculated. The x-axis value with the maximum distance is the base of the hill. The hill base values are calculated both in left and right side.
- Find the connectivity of the stroke within the base area. The shortest path between the highest pixels and lowest pixel y coordinate is found.
- Validate the hill area by considering from connected stroke path. If the path is within the hill bases more than 0.75 of the length of the path, it illustrates that there is a vertical stroke in this hill area base.

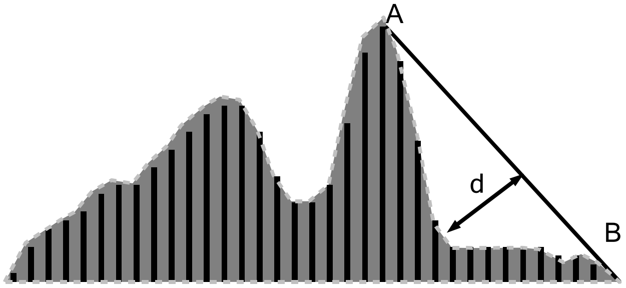
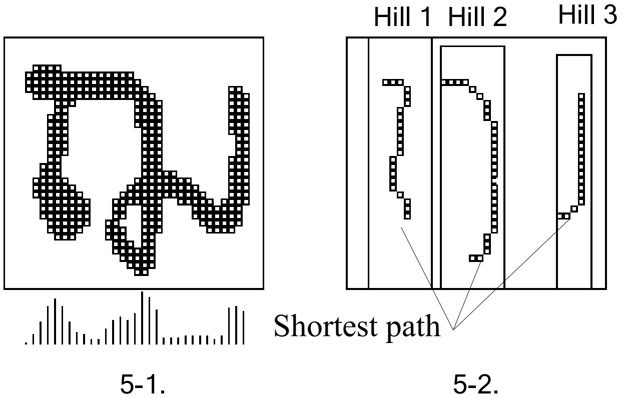


Figure 4. triangle method method.



5-1.

5-2.

Figure 5. Vertical stroke detecting method.
5-1. The image and its projection.
5-2. Hill boundary and the shortest stroke path.

3.2 Block Adjustment

In a nonlinear normalization, the image is divided according to the density of the pixel, unlikely, our method devides the image from the vertical stroke structure. In the research of Theeramunkong, [6] it was showed that Thai characters have more information in horizontal direction than in vertical direction. So, we assign a largher number of blocks for horizontal direction. The character is divided into 7 x 10 blocks. The character is divided in to 7 vertical parts equally. The character is also divided into 10 horizontal parts following this algorithm.

- *if there is a space between left boundary of the imag e and left boundary of the first hill, then assign the first b lock for that space.*
- *if there is a space between right boundary of the image and right boundary of the last hill, then assign the last block for that space.*
- *if there is a hill in the left side,then assign 1-2 blocks from left for the hill.*
- *if there is a hill in the right side,then assign 1-2 blocks from right for the hill.*
- *if there is a hill in the middle, then assign 1-2 block s in the central*
- *For others, divided it equally.*

The blocks are assigned to fit the position of the hills and gaps. The algorithm begins with the detection of the hill and the gap on the left side of the image. The size of the blocks are depend on the size of the hill. If the hill has a large size, the algorithm will be assigned more than one block for that hill.

The pixels in each block are averaged. Then the 70 dimensions feature is trained by a neural network model.

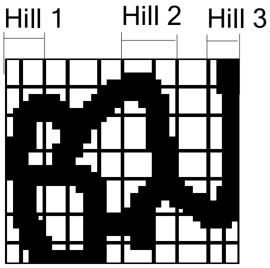


Figure 6. The character is divided in 7x10 blocks.

3. Clustering Method

The confusion matrix is used to analyze the result from neural network model. The characters that get low recognition rate will be grouped with the another character that have a highest error score. Table 2 shows the confusion matix. The values in the column i and row j position is the number of the j^th character datas that answer as i^th character. For example, the second group of characters, “ข”, are answer as “ข” for 52 characters at the 39.31 percent recognition rate. ,The recognition rate of character “ข” is lower than the threshold, so it is grouped with the character “บ” that have a highest error score in the second row. All of the characters are applied this condition. Then the grouped datas are trained again. These processes are iterated until all of the character groups reach the threshold. The threshold is computed by the following equation:

$$Threshold = \begin{cases} 90 & ; Threshold > 90 \\ Average\ Accuracy & ; 70 < Threshold < 90 \\ 70 & ; Threshold < 70 \end{cases}$$

The clustering method is shown in Fig 7. From the result, the characters that have a similar structure will be merged into the same group. Moreover, some characters, that extremely difference in printed, are grouped because they look similar in a writing style.

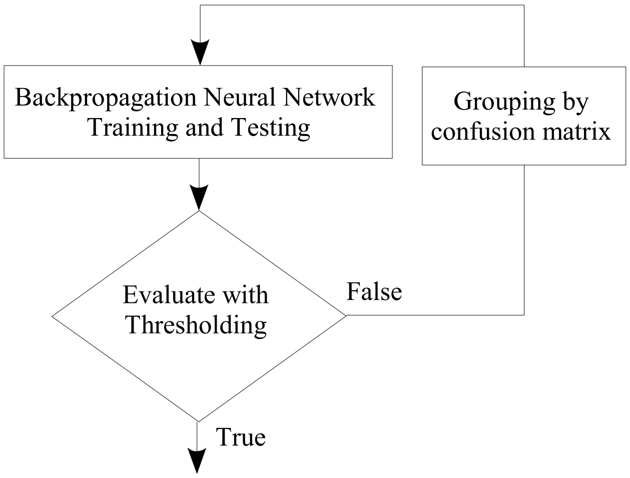

Backpropagation Neural Network Training and Testing
Grouping by confusion matrix
Evaluate with Thresholding
False
True

Figure 7. The clustering method

Table 3. The train/test result

Iteration (cycles)	*Accuracy Rate (%)*	*Number of Class*
1	64.17	60
2	87.42	47
3	93.23	29
4	95.38	22
5	97.60	21

4. Experimental Result

The database, that is used in this experimental, is from isolate character set, Nectec Corpus. It collected from 68 writers with each person written twice and scanned with 200 dpi. The characters(totally 9,020 characters) are devided into a training(5,940 characters), and testing sets (3,080 characters).

As shown in Table 3, the accuracy of the networks are increased in each iteration. At the 5th iteration, all of clustered model have a accuracy rate higher than the threshold, so the clustering is stopped. The characters that have the same structure are clustered in the same group. Furthermore, the experiment shows that there are some characters that have a similar stucture due to the writing variation. In some characters, only global feature is enough to separate itself with the others. Table 4 shows the clustering result of Thai character set. Our method yield a total recognition rate at 97.60 percent of 21 groups of character.With the smaller group of characters, it will simplify the overall problem.

Table 4 the total recognition rate of 23 character groups

Character Group	***Recognition Rate (%)***
ก ถ ภ	98.92
ข ฃ ฆ บ ม ย ษ	98.69
ค ต ด ฅ	98.99
ง จ	97.66
ฉ ล ว อ า	91.23
ช ซ ธ ฐ ร ฮ	94.38
ฌ ฒ ณ ญ	96.29
ฎ ฏ	98.28
ฑ ท ห	94.17
น พ	98.61
ป	99.98
ผ	94.52
ฝ	93.65
พ	99.98
ฟ	97.26
ศ ส ฬ	95.9
อุ อู	99.98
อ๋ อ่	93.32
อิ อี อึ อื อั	96.65
อ็ อ์ อ๊ อั	98.23
โ ไ ใ	97.45
total	97.33

5. Conclusion and Future work

This paper has described an algorithm for clustering Thai character models. The algorithm begins with the vertical stroke detection that is one of the important feature in Thai characters. The character area is divided in to 7x10 blocks using the stroke information. The average of black pixels is calculated for each block. The calculated value is used as a pixel distribution feature. Then the feature will be trained by neural network. The confusion matrix is used as a tool for analyse the trained result in a clustering process. The result shows that the clustering yield a 97.60 percent recognition rate for 21 groups. This clustered model, can be used as a rough classify model or used as a global feature that reflects the structure meaning of the characters.

As mentioned, a single global or local features are not enough but they would be combined together. So, the study on the variation of the structure models and the robust feature extraction techniques are our current focus.

10. References

[1] Airphaiboon S., Sangworasil M., Kondo S., “Off-line Handwritten Thai characters from Word Script” Proceedings of the 12th IAPR International, vol.2, 1994, pp. 445-449.

[2] Choruengwiwat P., Jitapunkul S., Wuttisittikulkij L., Seehapan P., “Distintive Feature Analysis for Thai Handwritten Charactwer Recognition Based on Modified Stroke Changeing Sequence”, Proceedings of the IEEE Asia-Pacific Conference on Circuit and Systems, November, 1998.

[3] Phokharatkul P., Kimpan C., “Recognition of Handprinted Thai Characters Using the Cavity Features of Character Based on Neural Network.” Proceedings of the IEEE Asia-Pacific Conference on Circuit and Systems, November, 1998.

[4] Phokharatkul P., Kimpan C., “Handwritten Thai Character Recognition Using Fourier Descriptors and Genetic Neural Networks.” Proceedings of the Symposium on Natural Language Processing, May, 2000.

[5] Methasate I., Jitapunkul S., Kiratiratanaphrug K., Unsiam W., “Fuzzy Feature Extraction for Thai Handwritten Character Recognition” Proceedings of the Symposium on Natural Language Processing, May, 2000.

[6] Theeramunkong T., Wongtapan C., Sinthupinyo S., "Offline Isolated Handwritten Thai OCR Using Island-Based Projection with N-Gram Models and Hidden Markov Model", Proceedings of the International Conference on Asian Digital Libraries , December, 2002.

[7] Tsai, C.M., Lee H., "Binarization of Color Document Images via Luminance and Saturation Color Features.", IEEE Transactions on Image Processing, no 4, vol 11 , April , 2002, pp. 434-451.

Table 2 confusion matrix

Character	'ก'	'ข'	'ฃ'	'ค'	'ฅ'	'...'	'บ'	Recognition rate (%)
'ก'	84	0	0	5	0		0	65.08
'ข'	0	52	27	0	0		32	39.31
'ฃ'	0	27	102	0	0		11	79.79
'ค'	5	0	0	93	7		0	74.31
'ฅ'	0	0	0	7	109		0	85.27
...								...
'บ'	0	32	11	0	0		97	77.67
Total								64.17

Grouped together