Thai Handwritten Character Recognition by Genetic Algorithm (THCRGA)

Chomtip Pornpanomchai, Verachad Wongsawangtham, Satheanpong Jeungudomporn, and Nannaphat Chatsumpun

***Abstract*—The objective of this research is to develop computer software that can recognize the Thai handwritten characters by using the genetic algorithm technique (THCRGA). The system consists of 5 main modules, which are: 1) image acquisition module, 2) image preprocessing module, 3) feature extraction module, 4) character recognition module, and 5) display result module. Each module has the following details. First, the image acquisition module collects an unknown input character from a user. Second, the input image is transformed into a suitable image for the feature extraction module. Third, the system extracts character features from the image. There are 3 main features of Thai characters which are stroke, loop and location of loop and stroke connection. Fourth, the extracted character information is kept in the form of bits string chromosome in a genetic algorithm. Finally, the system displays the best fitness chromosome for the recognition result. The experiment was conducted on more than 10,000 Thai handwritten characters by using 8,160 for training characters and 2,040 for testing characters. The precision of the system is around 88.24 percent, with recognition speed of 0.42 second per character.**

***Index Terms*—Thai Handwritten Character Recognition, Thai Character Features Extraction, Genetic Algorithm**

I. Introduction

Handwritten character recognition is one of the most difficult tasks in the pattern recognition system. There are a lot of difficult things that need many image processing techniques to solve, for examples: 1) how to separate cursive characters into an individual character, 2) how to recognize unlimited character fonts and written styles, and 3) how to distinguish characters that have the same shape but different meaning such as the character o and number 0. Many researchers try to apply many techniques for breaking through the complex problems of handwritten character recognition. There are many applications that need to take advantage of the handwritten character recognition system, namely, 1) automatic reading machine, 2) non-keyboard computer system, and 3) automatic mailing classification system. [1] The objective of this research is to try to help researchers to recognize Thai handwritten characters by using the genetic algorithm technique.

The Thai alphabet has 44 consonant characters, 14 vowels, as well as 8 tonal and 10 numerical symbols, as shown in Figure 1 (a)-(d), respectively. Normally, Thai characters consist of small circles or loops, which are connected to circular zigzag lines and straight lines. Most Thai characters are written by using a single stroke. The structure of Thai words is written in a four-line level style, which is shown in Figure 2. The first-line level is a tonal line level and vowel characters can be written on the second-, third- and fourth-line levels. Consonant characters are written on the third-level line. [1]

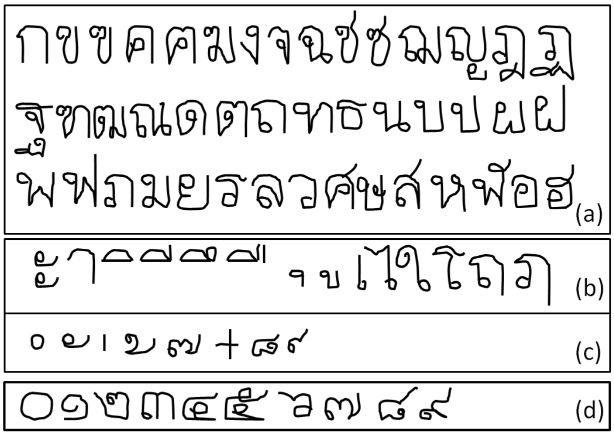


Figure 1. (a) Thai consonant (b) vowel (c) tonal and (d) nummerical character symbols.

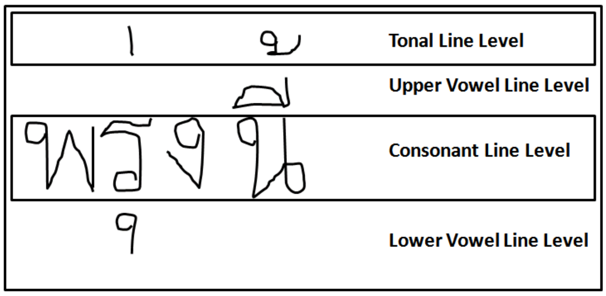


Figure 2. The Thai words in a four-line level style

II. Literature Reviews

Historically, handwritten character recognition applications used three major approaches; the statistical approach, the structural or syntactic approach, and the neural network-based approach. This section reviews handwritten character recognition applications based on these three approaches.

Manuscript received November 04, 2010.

Dr. Chomtip Pornpanomchai is an Assistant Professor in the Faculty of Information and Communication Technology, Mahidol University, Member, IACSIT,Bangkok 10400, Thailand, e-mail: itcpp@mahidol.ac.th, Tel:662-354-4333.

Verachad Wongsawangtham, Satheanpong Jeungudomporn, and Nannaphat Chatsumpun are graduated from the Faculty of Information and Communication Technology, Mahidol University, Bangkok 10400, Thailand, e-mail:{u4988225,u4988055,u4988023}@student.mahidol.ac.th, Tel:662-354-4333.

A. Statistical Analysis Approach

Statistical Pattern Recognition uses statistical and/or probabilities functions for building a recognition algorithm. The input features are extracted from a set of characteristic pattern measurements. A limitation of this approach is the difficulty to express pattern classification in terms of structural information. [2, 3, 4, 5, 6, 7 and 8]

B. Structural or Syntactic Analysis Approach

Syntactic Pattern Recognition uses syntactic or structural information of patterns to generate knowledge that is related to patterns. This approach extracts the similarity of patterns and builds pattern syntax or structural rules. The information of pattern syntax rules is used to explain, classify and recognize unknown patterns. This approach is suitable for building a handwritten character recognition system because it uses a structural approach to build unlimited handwritten character patterns syntax. A limitation of this approach is the difficulty to build learning structural rules. [9, 10, 11, 12 and 13]

C. Neural Network Based Approach

Neural Pattern Recognition emulates knowledge of how a biological neural system stores and manipulates information. This artificial neural system is called “neural networks”. The notion is that an artificial neural network can solve all problems in automatic reasoning, including a pattern recognition problem. This approach classifies patterns through predictable properties of neural networks. A limitation of this approach is a little amount of semantic information from a network. [14, 15 and 16]

III. METHODOLOGY

In this section, we present all details of our system design. First, we start with the overall framework of the Thai handwritten character recognition system. Then, we give each component detail. Finally, we present the user interface.

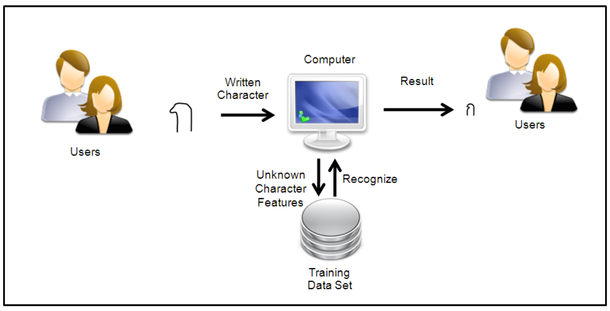


Figure 3. The framework of Thai handwritten character recognition system

A. System Architecture Overview

Based on the Thai handwritten character recognition system in the previous section, we design the THCRGA system framework as shown in Figure 3. The THCRGA has the following workflow. First, the system captures the Thai written character images and stores them in a computer system. Second, the system extracts several features from the character images such as the number of circles, number of lines and connection locations between line and circle in each character by using image processing techniques. Third, the system uses all features of a character to generate a genetic chromosome. Fourth, the system recognizes Thai characters by comparing genetic chromosome between unknown characters and the training character set in a database. Finally, the THCRGA displays the best fitness genetic chromosome for the output result.

B. System Structure Chart

Based on the system framework in the previous section, we convert the THCRGA framework to the system structure chart as depicted in Figure 4. The THCRGA system consists of five main modules, which are 1) image acquisition, 2) image preprocessing, 3) feature extraction, 4) character recognition, and 5) display result, as shown in Figure 4. The detail of each module is described as the following:

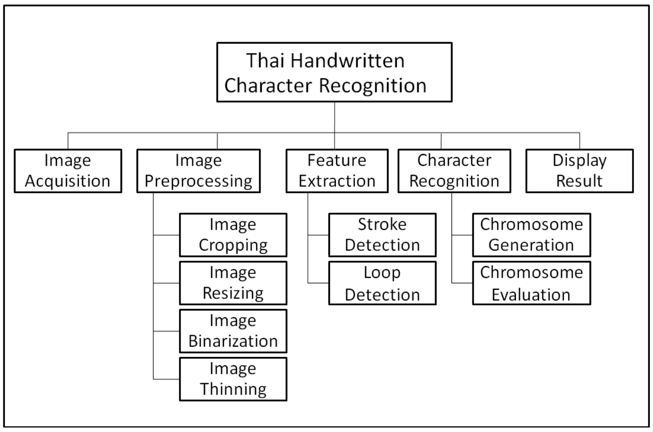


Figure 4. The structure chart of Thai handwritten character recognition by genetic algorithm system

1) Image Acquisition

In the first stage, the image acquisition module captures an unknown input character from the system interface. The THCRGA system provides a workspace and lets the user to draw an online Thai handwritten character on the workspace. After the user finishes drawing, all the details of an written character is saved into a bitmap image and the process passes to the next step.

2) Image Preprocessing

In the image preprocessing module, the system prepares a suitable handwritten character image for the feature extraction module. The image preprocessing stage consists of four sub-processes, which are: 1) image cropping, 2) image resizing, 3) image binarization, and 4) image thinning. Each sub-process has the following details.

a) Image Cropping Sub-process

The character image from the image acquisition stage has the white space that is not necessary in the recognition process. Moreover, the white space needs more CPU power for the recognition process and may cause an erroneous result. Therefore, the system needs to crop only the written character boundary. The example of the cropped character image is shown in Figure 5.

b) Image Resizing Sub-process

The input image may have different size, which will affect the recognition results. Therefore, every input image will be resized to 100 x 100 pixels image. The example of the resized character image is shown in Figure 6.

c) Image Binarization Sub-process

The binarization sub-process will change an input character image into a binary image (0s and 1s only). A binary image helps the feature extraction module to extract character features easily. The example of the binarized character image is shown in Figure 7.

d) Image Thinning Sub-process

The thinning sub-process reduces a written character of a thick line into a thin character or skeleton character. A thin character is easier to extract its feature than a thick character. The example of the thinned character image is shown in Figure 8.

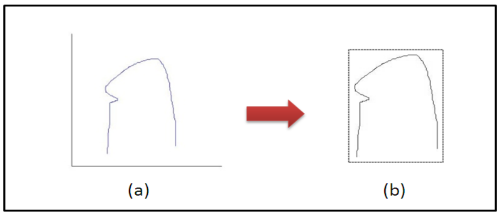


Figure 5. The example of cropping character (a) input image (b) cropped image

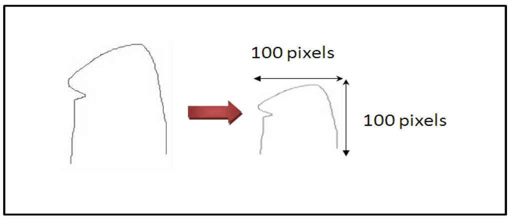


Figure 6. The example of resized character image

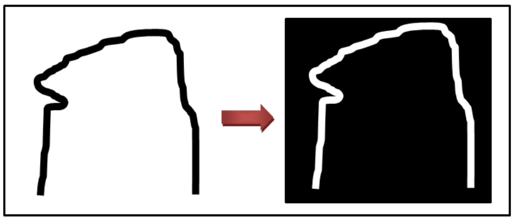


Figure 7. The example of binarized character image

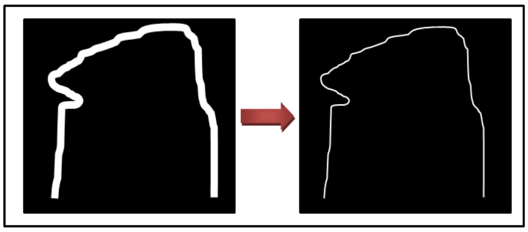


Figure 8. The example of thinned character image

3) Feature Extraction

The feature extraction module extracts the basic components of Thai characters, such as loops, straight lines, zigzag lines and the position of the connection between the loop and the straight line. Normally,a Thai character consists of none to three loops, none to three zigzag lines and none to five straight lines. After this module extracts the basic components from the Thai character image, then it translates all basic components into chromosome bits string in a genetic algorithm. There are two main basic components that the THCRGA extracts, which are: 1) a stroke extraction, and 2) a loop extraction. Both of them have the following details.

a) Stroke Detection

The stroke detection sub-process extracts the type of a line by separating a character into 5 X 5 pixels image blocks and finding the slope of a line in each block. Finally, it analyses and classifies a line type of a character in four categories, which are 1) vertical line, 2) horizontal line, 3) zigzag line, and 4) the tail line. Each line category has the following details.

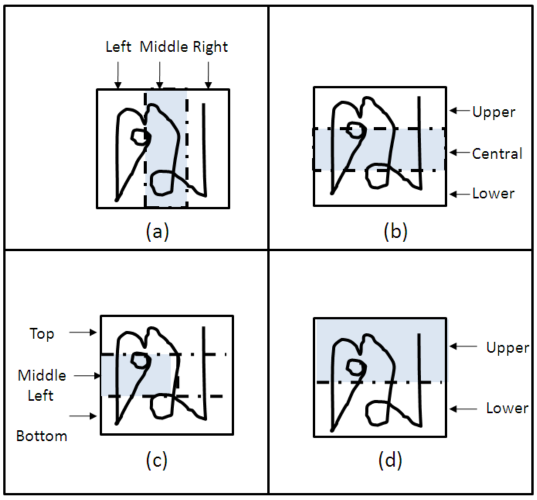


Figure 9. The Thai character stroke analysis region

(1) Vertical Stroke Analysis

The vertical stroke analysis is a function that extracts the vertical stroke lines in the Thai character. The THCRGA separates one character into three vertical regions, namely: 1) Left Region, 2) Middle Region, and 3) Right Region, as shown in Figure 9 (a). The output of this function is produced in 4-bit chromosome. The first three chromosome bits string represents the vertical line in left, middle and right regions, respectively. And the fourth chromosome bit string represents the second vertical line in the right region.

(2) Horizontal Stroke Analysis

The horizontal stroke analysis is a function for extracting the horizontal stroke line in the Thai character. The THCRGA separates a character into three regions, which are 1) Upper Region, 2) Central Region, and 3) Lower Region, as shown in Figure 9 (b). The output of this function generates 3-bit chromosome, which are 1) Upper Region, 2) Central Region and 3) Lower Region.

(3) Zigzag Stroke Analysis

The zigzag stroke analysis in THCRGA means the stroke that contains a turning point. The zigzag line in a Thai character is divided into 3 zones, namely; 1) Top Zone, Middle Left Zone, and Bottom Zone, as shown in Figure 9 (c). The output of this function gives 3-bit chromosome, which are 1) Top zigzag line, 2) Middle left zigzag line, and 3) Bottom zigzag line.

(4) Tail Stroke Analysis

The tail stroke analysis is a function to extract the tail stroke of the Thai character. Some Thai characters have a long tail stroke. THCRGA focuses on the location of a tail.

There are 2 positions, which are Upper tail and Lower tail, as shown in Figure 9 (d). The output of this function produces 2-bit string chromosome.

b) Loop Detection

The loop detection sub-process extracts three characteristics of the loop, which are 1) number of loops in a character image, 2) position of each loop in a character image, and 3) type of the loop. The loop detection sub-process has the following details.

(1) Number of Loops Analysis

Based on the observation, every Thai character consists of zero to three loops, for example, the Thai character "ก" has no loop, character "ง" has one loop, character "น" has two loops and character "ฐ" has three loops. The THCRGA applies a color filling algorithm to find the loops in a character image. The concept of a color filling algorithm is filling a white color pixel in an image with a black-color background. A background will change to a white-color if there is no completed loop in the image. The color filling algorithm process is shown in Figure 10. The output of this process generates 3-bit string chromosome to flag number of loops in a character.

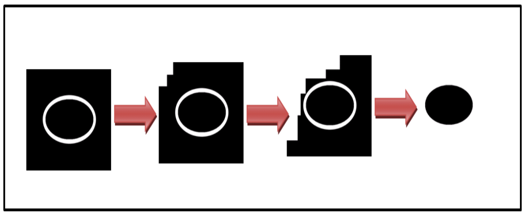


Figure 10. The color filling algorithm to find a loop

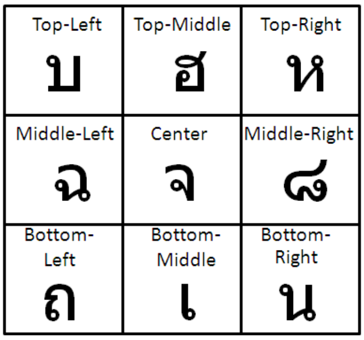


Figure 11. The locations of loop(s) in Thai characters

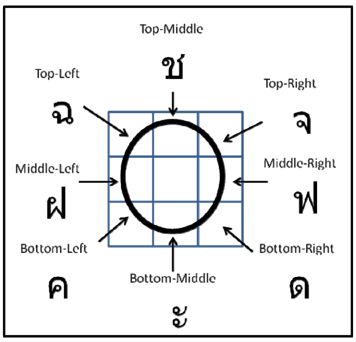


Figure 12. The positions of loop and line connection

(2) Loop Location Analysis

Each loop in a Thai character is located in nine locations, which are 1) Top-Left, for example, a character "บ", 2) Top-Middle, for example, a character "ฮ", 3) Top-Right, for example, a character "ห", 4) Middle-Left, for example, a character "ฉ", 5) Center, for example, a character "จ", 6) Middle-Right, for example, a character "๘", 7) Bottom-Left, for example, a character "ถ", 8) Bottom-Middle, for example, a character "โ"and 9) Bottom-Right, for example, a character "น", as shown in Figure 11. Each character has maximum 3 loops and each loop has 9 locations; therefore, there are 3X9 equal to a 27-bit chromosome.

(3) Loop Type Analysis

There are eight points in a loop that can connect to a line which are 1) Top-Left, for example, a character "ฉ", 2) Top-Middle, for example, a character "ช", 3) Top-Right, for example, a character "จ", 4) Middle-Left, for example, a character "ฝ", 5) Middle-Right, for example, a character "ฟ", 6) Bottom-Left, for example, a character "ค", 7) Bottom-Right, for example, a character "ด" and 8) Bottom-Middle, for example, a character "ะ" as shown in Figure 12. Each character has maximum 3 loops and each loop has 8 connection points to a straight line; therefore, there are 3X8 equal to a 24-bit chromosome.

4) Character Recognition

The image recognition module uses information from feature extraction to form a genetic chromosome, after that it uses chromosome string to recognize the Thai character image. This module consists of 2 functions, namely, 1) chromosome generation function, and 2) chromosome evaluation function. Each function has the following details.

(1) Chromosome Generation Function

The chromosome generation function produces the Thai character chromosome by combining all features extracted in the previous module together, namely:

- 3-bit for number of loops
- 27-bit for location of each loop
- 24-bit for loop connected with a straight line
- 12-bit for location of the lines

There are 66-bit chromosomes in the Thai character. The sample of a Thai character "ฒ" chromosome bits string are "11100 00100 00110 00010 00000 00000 01000 00000 00000 00001 00101 00000 00000 0", where 0s is none of the features are shown in a character, and 1s represent a character image shows that feature. The meaning of each chromosome bit string is shown in Table 1.

(2) Chromosome Evaluation Function

The chromosome bit string from the chromosome generation function is used to recognize a character by comparing the fitness value of an unknown character with all Thai characters in the database. The highest fitness value is the recognition result. The fitness value is calculated by using Equation 1 as the following:

$$\text{Fitness Value} = \sum_{i=1}^{66} |(S_i+1.0) - (L_i+1.0)| * W_i \qquad (1)$$

Where S is a chromosome bit string in database
L is a chromosome bit string of an unknown character
W is weight of each chromosome bit string

TABLE 1 THE BIT STRING CHROMOSOME REPRESENTATION

Bit No.	Description	Bit No.	Description
1	Left vertical line is presented	34	The third loop is located at Top-Left
2	Middle vertical line is presented	35	The third loop is located at Top-Middle
3	Right vertical line 1 is presented	36	The third loop is located at Top-Right
4	Right vertical line 2 is presented	37	The third loop is located at Middle-Left
5	Upper horizontal line is presented	38	The third loop is located at Center
6	Middle horizontal line is presented	39	The third loop is located at Middle-Right
7	lower horizontal line is presented	40	The third loop is located at Bottom-Left
8	Upper zigzag line is presented	41	The third loop is located at Bottom-Middle
9	Bottom zigzag line is presented	42	The third loop is located at Bottom-Right
10	Middle left zigzag line is presented	43	The first line is connected to Top-Left of loop
11	Upper tail is presented	44	The first line is connected to Top-Middle of loop
12	Lower tail is presented	45	The first line is connected to Top-Right of loop
13	The first loop is presented	46	The first line is connected to Middle-Left of loop
14	The second loop is presented	47	The first line is connected to Middle-Right of loop
15	The third loop is presented	48	The first line is connected to Bottom-Left of loop
16	The first loop is located at Top-Left	49	The first line is connected to Bottom-Middle of loop
17	The first loop is located at Top-Middle	50	The first line is connected to Bottom-Right of loop
18	The first loop is located at Top-Right	51	The second line is connected to Top-Left of loop
19	The first loop is located at Middle-Left	52	The second line is connected to Top-Middle of loop
20	The first loop is located at Center	53	The second line is connected to Top-Right of loop
21	The first loop is located at Middle-Right	54	The second line is connected to Middle-Left of loop
22	The first loop is located at Bottom-Left	55	The second line is connected to Middle-Right of loop
23	The first loop is located at Bottom-Middle	56	The second line is connected to Bottom-Left of loop
24	The first loop is located at Bottom-Right	57	The second line is connected to Bottom-Middle of loop
25	The second loop is located at Top-Left	58	The second line is connected to Bottom-Right of loop
26	The second loop is located at Top-Middle	59	The third line is connected to Top-Left of loop
27	The second loop is located at Top-Right	60	The third line is connected to Top-Middle of loop
28	The second loop is located at Middle-Left	61	The third line is connected to Top-Right of loop
29	The second loop is located at Center	62	The third line is connected to Middle-Left of loop
30	The second loop is located at Middle-Right	63	The third line is connected to Middle-Right of loop
31	The second loop is located at Bottom-Left	64	The third line is connected to Bottom-Left of loop
32	The second loop is located at Bottom-Middle	65	The third line is connected to Bottom-Middle of loop
33	The second loop is located at Bottom-Right	66	The third line is connected to Bottom-Right of loop

5) *Design Results*

The THCRGA system interface is shown in Figure 13. The interface has 2 panels. The first one is a workspace (label number 1). This space is provided for drawing a Thai character to test the system. The next one is a result panel (label number 2). This panel is used to show the picture of a character which is a result of recognition. Moreover, there is a label that shows the recognition process time (label number 3) and fitness value of the input character (label number 4).

There are 2 types of tools on the THCRGA interface. There are drawing and recognition tools. In a drawing tool, there are write button (label number 5) and clear button (label number 6). The write button will call the drawing function that lets the user draw on the workspace. The clear button is used for clearing workspace. For a recognition tool, there is a recognition button (label number 7). The recognition button accesses a recognition method to identify an image from the workspace and shows the result in the result panel. There is a group of radio boxes that are used to specify the line level of written characters. A user need to tell the system whether the input character is consonant (label number 8), ordinary vowel (label number 9), vowel in upper level (label number 10), or vowel in lower level (label number 11).

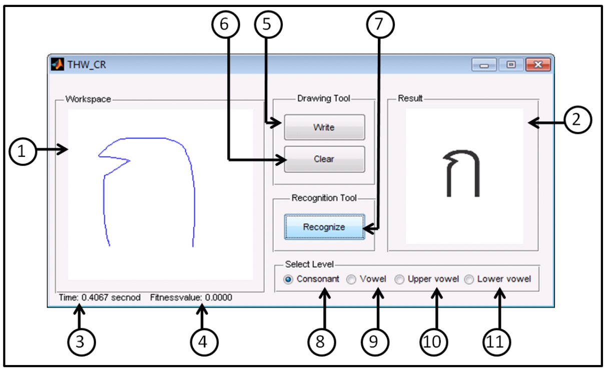


Figure 13. User Interface of THCRGA system

IV. TESTING AND EVALUATION

The experiments are conducted on more than 10,000 Thai handwritten characters. The testing characters are separated into two data sets, the training data set and the testing data set. The training data set contains 8,160 characters, and the testing data set contains another 2,040 characters. Both data sets do not overlap and some samples of matched and rejected handwritten characters are shown in Figure 14(a) and (b), respectively. The precision of the THCRGA system is 88.24 percent match, 10.10 percent mismatch and 1.66 percent rejection with the recognition speed 0.4192 second per character.

This research also tested with 312 English handwritten characters, with the precision of 1.06 percent match, 82.05 percent mismatch and 16.35 percent rejection.

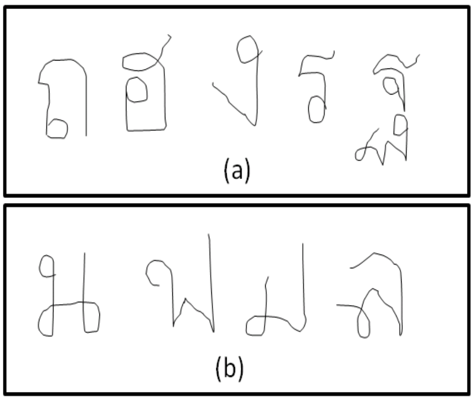


Figure 14. (a) Samples of matched characters (b) samples of rejected characters

V. CONCLUSIONS

In this paper, we fulfill our research objective by applying the genetic algorithm technique for recognizing Thai handwritten characters, based on the basic features of handwritten characters namely, 1) loop, 2) line, and 3) location of loop and line connection. The system generates

66-bit string chromosome to represent a handwritten character. Then the system uses the 66-bit string chromosome to identify each handwritten character.

The Thai handwritten character recognition system still has many difficult problems that need more complex algorithms to solve. For example, to recognize a cursive handwritten character, the recognition system needs a more effective character segmentation algorithm to separate a cursive character into individual characters and needs a more efficient character recognition algorithm to identify uncertain handwritten shapes. Because every handwritten character has unlimited shapes, patterns and styles even when it is written by the same or a different person. So it is difficult to define a standard structure, for a general algorithm to explain uncertain handwritten characters.

References

[1] Chomtip Pornpanomchai, Dentcho N. Batanov and Nicholas Dimmitt, "Recognizing Thai handwritten characters and words for human-computer interaction", International Journal of Human-Computer Studies, pp. 259-279, (2001)

[2] Chomtip Pornpanomchai, Pattara Panyasrivarom, Nuttakit Pisitviroj and Piyaphume Prutkraiwat, "Thai Handwritten Character Recognition by Euclidean Distance", The 2nd International Conference on Digital Image Processing (ICDIP 2010), pp. 53-58' (2010)

[3] Chomtip Pornpanomchai and Montri Daveloh, "Printed Thai Character Recognition by Genetic Algorithm", The International Conference on Machine Learning and Cybernetics, pp. 3354-3359, (2007)

[4] Boontee Kruatrachue, Nattachat Pantrakarn and Kritawan Siriboon, "State Machine Induction with Positive and Negative for Thai Character Recognition", The International Conference on Communications, Circuits and Systems, pp. 971-975, (2007)

[5] Parinya Sanguansat, Widhyakorn Asdornwised and Somchai Jitapunkul, "Online Thai Handwritten Character Recognition Using Hidden Markov Models and Support Vector Machines", The International Symposium on Communications and Information Technologies, pp.492-497, (2004)

[6] Rud Budsayaplakorn, Widhayakorn Asdornwised and Somchai Jitapunkul, "On-line Thai handwritten character recognition using hidden Markov model and fuzzy logic", The IEEE 13th Workshop on Neural Networks for Signal Processing, pp. 537-546, (2003)

[7] Kritawan Siriboon, Apirak Jirayusakul and Boontee Kruatrachue, "HMM topology selection for on-line Thai handwriting recognition", The First International Symposium on Cyber Worlds, pp. 142-145, (2002)

[8] Arrak Pornchaikajornsak and Arit Thammano, "Handwritten Thai character recognition using fuzzy membership function and fuzzy ARTMAP", The IEEE International Symposium on Computational Intelligence in Robotics and Automation, pp. 40-44, (2003)

[9] Supachai Tangwongsan and Orawan Jungthanawong, "Refinement of stroke structure for printed Thai character recognition", The 9th International Conference on Signal Processing, pp. 1504-1507, (2008)

[10] Khampheth Bounnady, Boontee Kruatrachue and Takenobu Matsuura, "Online Unconstrained Handwritten Thai Character Recognition Using Multiple Representations", The International Symposium on Communications and Information Technologies, pp. 135-140, (2008)

[11] Jarernsri L. Mitrpanont and Urairat Limkonglap, "Using Contour Analysis to Improve Feature Extraction in Thai Handwritten Character Recognition Systems", The 7th IEEE International Conference on Computer and Information Technology, pp.668-673, (2007)

[12] Ithipan Methasate, Sanparith Marukatat, Sutat Sae-tang and Thanarak Theeramunkong, "The feature combination technique for off-line Thai character recognition system", The 8th International Conference on Document Analysis and Recognition, pp.1006-1009, (2005)

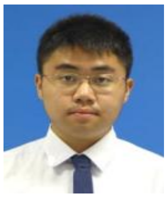
[13] Thirapiroon Thongkamwitoon, Widhyakorn Asdornwised, Supavadee Aramvith and Somchai Jitapunkul, "On-line Thai-English handwritten character recognition using distinctive features", The Asia-Pacific Conference on Circuits and Systems, pp. 259-264, (2002)

[14] Arit Thammano and Pruegsa Duangphasuk, "Printed Thai character recognition using the hierarchical cross-correlation ARTMAP", The 17th IEEE International Conference on Tools with Artificial Intelligence, pp. 695-698, (2005)

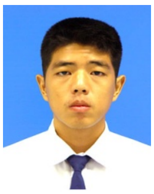
[15] Watchara Chatwiriya, Powsiri Klinkhachorn and Norman Lass, "Thai handwriting legal amounts recognition", The 35th Southeastern Symposium on System Theory, pp. 381-385, (2003)

[16] Arit Thammano and Patcharawadee Poolsamran, "Applying a Weighting Matrix to the Hierarchical Neural Network Model for Handwritten Thai Character Recognition", The International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, pp. 85-89 (2006)

Chomtip Pornpanomchai received his B.S. in general science from Kasetsart University, M.S. in computer science from Chulalongkorn University and Ph.D. in computer science from Asian Institute of Technology. He is currently an assistant professor in the faculty of Information and Communication Technology, Mahidol University, Bangkok, Thailand. His research interests include artificial intelligence, pattern recognition and object-oriented systems. Email: itcpp@mahidol.ac.th



Verachad Wongsawangtham was born in Bangkok Thailand. He received his B.S. in Information and Communication Technology, Mahidol University, Bangkok, Thailand. Email: u4988225@student.mahidol.ac.th



Satheanpong Jeungudomporn was born in Bangkok Thailand. He received his B.S. in Information and Communication Technology, Mahidol University, Bangkok, Thailand. Email: u4988055@student.mahidol.ac.th

Nannaphat Chatsumpun was born in Bangkok Thailand. She received her B.S. in Information and Communication Technology, Mahidol University, Bangkok, Thailand. Email: u4988023@student.mahidol.ac.th