Improved Mixed Thai & English OCR using Two-step Neural Net Classification

Chularat Tanprasert[1], Wasin Sinthupinyo[1], Premnath Dubey[1], and Thitipong Tanprasert[2]

[1]Software Technology Laboratory,
the National Electronics and Computer Technology Center (NECTEC),
National Science and Technology Development Agency (NSTDA),
Ministry of Science, Technology and Environment,
108 Bangkok Thai Tower Building, Floor 11th, Rangnam Road,
Phayathai, Bangkok 10400 THAILAND

[2]Department of Computer Science,
Faculty of Science and Technology, Assumption University,
Hua Mark, Bangkok, 10240, THAILAND

ABSTRACT -- This paper presents a new technique in applying artificial neural networks to solve the problem of Thai and English Optical Character Recognition (Thai & English OCR). The Kohonen self-organization and back propagation models are applied for rough and fine classifications to the Thai & English typed fonts, respectively. The experimental results confirm that the proposed technique is a major improvement in Thai & English OCR technology.
KEY WORDS -- Thai & English OCR, Kohonen self-organization, Back propagation model

บทคัดย่อ – บทความนี้ได้เสนอแนวทางใหม่ในการประยุกต์ใช้โครงข่ายประสาทเทียมกับการรู้จำตัวอักษรไทยและอังกฤษ โดยใช้การจัดตัวเองแบบโคโฮเนนในการแบ่งกลุ่มตัวอักษรอย่างหยาบๆ และใช้รูปแบบการแพร่กระจายย้อนกลับในการรู้จำอย่างละเอียดอีกขั้นหนึ่ง ผลการทดลองแสดงให้เห็นถึงความสามารถในการรู้จำที่สูงขึ้นสำหรับตัวอักษรไทยและอังกฤษ
คำสำคัญ – การรู้จำตัวอักษรไทยและอังกฤษ, การจัดตัวเองแบบโคโฮเนน, รูปแบบการแพร่กระจายย้อนกลับ

1. Introduction

Has continually evolved since being developed about 700 years ago, Thai language has reached the level of being standardized for computer software. However, there has been no Thai Optical Character Recognition software that produces truly satisfying result. At present, there are three commercial Thai OCR products available in Thailand for Thai typed fonts, one of them having its engine developed at NECTEC. Nevertheless, the recognition rate of such software still ranges only from 90% to 95% on the average. This results in a lot of researches being conducted for Thai OCR because the software is now considered one of the most needed by a lot of users in Thailand.

Several different techniques have been proposed for implementing the Thai OCR problem. Some are the statistical pattern recognition techniques [3] but some are artificial neural networks [4]. The performance of those techniques is satisfactory only in the experimental level, not for using in the real life.

Most of Thai documents usually contain both Thai and English characters. Hence, OCR software in Thailand should be capable of recognizing both languages. In Thai language, there are 44 consonants, 32 vowels, 10 Thai digits, 4 tone marks, and 5 special symbols as shown in Figure

Reprint from Proceeding of the 1997 International Conference on Neural Information Processing and intelligent Information System, Vol. 2, pp. 1227-1230

1. In addition, there are 26 capital English letters, 26 small English letters, 10 Arabic digits, and 20 special symbols. So the total number of characters for the OCR is about 187 characters which is quite a large number for recognizing by the artificial neural networks.

44 Consonants : ก ข ฃ ค ฅ ฆ 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

32 Vowels : อะ อา อิ อี อึ อื อุ อู เอะ เอ
แอะ เอือ โอะโอ เอาะ ออ
เออะ เออ เอียะ เอีย เออะ
เอือ เวะ อัว ฤ ฤๅ ฦ ฦๅ
อำ ใอ ไอ เอา

10 Thai digits : ๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙

4 Tone Marks : อ่ อ้ อ๊ อ๋

5 Special Symbols : อ์ อ็ ๆ ฯ ฿

Figure 1. Illustrates all Thai Characters.

Thai characters are composed of circles, lines, curves, and zigzag as can be illustrated in Figure 1. In addition, some of them look very similar to each other such as ก ถ ภ, ข ฃ ช ซ, and ค ศ ฅ ด ต. This is one of the reasons that make Thai OCR very difficult to develop. Another problem is that - unlike English, locations of characters in a Thai sentence are organized into four levels as shown in Figure 2. This causes the difficulty for the pasting algorithm. The last point is that there is no space between words in Thai sentence, so the word cutting algorithm is required to be intelligent if the acceptable performance is to be obtained.

We have decided to apply artificial neural networks (ANN) to the problem since their capabilities match the problem, especially the ability to learn. However, due to the large total number of characters to be recognized and inherently complexity of the problem nature, the recognition rate from a straightforward application of standard back propagation learning have seemed to be limited to a lower than user-expected level. We have experimented applying the Kohonen self-organizing feature map [2] and the back propagation algorithm [1] to perform a two-step classification of all characters. Experiments are performed to approve the technique and a remarkable improvement in recognition rate has been achieved.

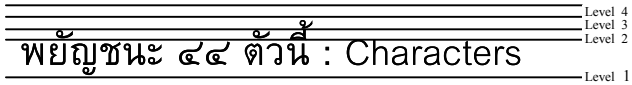

Figure 2. Shows Four-Level of Thai Sentence.

Following this introduction, the recognition concept for solving the specified problem is described in section 2. Next, the experimental results are discussed in section 3. Finally, the conclusion is given in section 4.

2. Thai Character Recognition Concept

Since there are several neural network approaches that have been applied successfully for characters recognition problems for other languages, we have experimented applying ANN to solve the problem as described in [4] and have pushed the recognition rate on unseen data up to around 95% for the Thai documents. That research has lead to the development of a recognition engine of a commercial Thai-and-English OCR software [6] in Thailand which have recognition rate at about 90% for mixed Thai & English documents in the trained fonts. Following such attempt, we have further investigated ANN in order to gain even better recognition rate of Thai OCR.

As mentioned above, there are 187 characters to recognize. When only the back propagation learning algorithm with multilayer perceptron is used to learn and memorize all characters, extremely large amount of character samples are required for the network to develop accurate character pattern classification. Furthermore, there are many similar characters in Thai language. This can

Reprint from Proceeding of the 1997 International Conference on Neural Information Processing and intelligent Information System, Vol. 2, pp. 1227-1230

easily guide the network to convert a local minimum which may not relate well to the unseen data. The proposed technique aims at reducing problem size for back propagation network by breaking the recognition process into rough and fine classifications. The performance is expected to be better if we can categorize all characters into groups of similar characters and then use back propagation networks to learn to classify characters in each group. Figure 3 illustrates the concept of our recognition engine.

While the biological neural network of a human being can categorize characters based on their similarity in shape, such technique requires a high level of intelligence which can be used even with handwritten characters. However, our problem only concerns the typed characters which have rather fixed forms. It could be more supportive for the back propagation networks if the characters are categorized by another technique. Such a technique should look at each character image in the way the back propagation networks look, that is - an *n*-dimensional vector whose each element describes each image pixel. We have determined to use an unsupervised learning algorithm. Since the Kohonen self-organizing feature map (SOFM) is one of the most popular unsupervised model [2], the technique is applied to categorize the characters into groups. With this technique, we have to specify the number of output clusters of the input patterns which is quite difficult to determine the optimal one. If there are too few clusters, many characters will be packed into the same group which will not effectively reduce workload for the back propagation network. However, with too many clusters, while higher accuracy of each back propagation network can be expected due to the fewer characters per cluster, accuracy of clustering phase may also reduce, resulting in lower overall recognition rate.

For fine classification, the back propagation algorithm with multi-layered perceptron [1] is applied. In [4] multi-layered perceptron with back propagation training was applied to recognize the 78 Thai characters directly and the technique has shown to be very promising. In this paper, the technique concerns mixed Thai & English data which is a substantially larger data set (187 characters). However, with characters being clustered, the complexity of the problem for each back propagation network should reduce and a higher recognition rate can be expected.

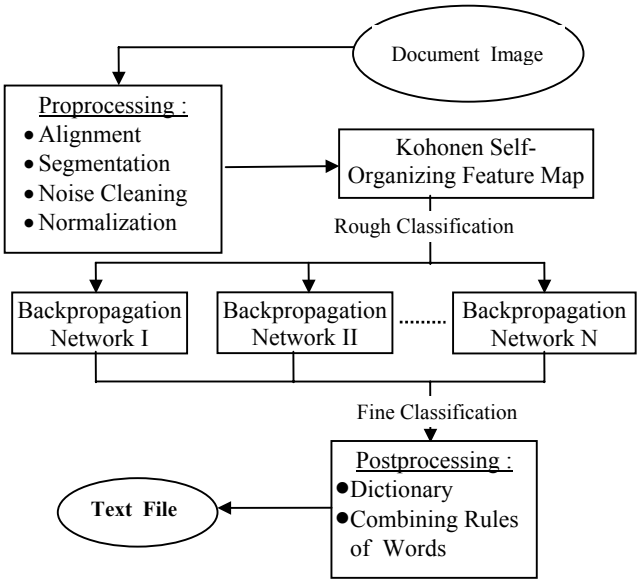

Figure. 3. Illustrates the diagram of the proposed Thai & English OCR system.

With the proposed technique, there are several parameters whose values need to be fine-tuned for example, the number of input nodes, the number of clusters of SOFM network, the number of hidden neurons for each back propagation network, the suitable values for the training parameters of each network.

3. Experimental Results

Experiments are conducted for the two most popular Thai & English typed characters fonts on the Microsoft Windows system, namely AngsanaUPC(AS) and BrowalliaUPC(BW). Example of alphabets in each font is shown in Figure 4. The total number of distinct characters of each font and each size is 180 characters. It is not 187 as mentioned previously because there are several Thai vowels that are combinations of other vowels as you can

Reprint from Proceeding of the 1997 International Conference on Neural Information Processing and intelligent Information System, Vol. 2, pp. 1227-1230

see in Figure 1. Therefore, we can cut such composite vowels out of our OCR system. Each training data set includes characters from each font in 7 different sizes, in 4 styles (normal, italics, bold, italics & bold). So the total number of patterns in the training set 7*4*180 = 5,040. However, the prepared document has been scanned with different levels of dot per inch (dpi) for obtaining several looks of each character. We have scanned the document with nine distinct dpi settings, resulting in the number of training patterns =5,040*9 =45,360 patterns. It means there are 252 distinct patterns for each character in the training set. These patterns may be clustered into different group in the rough classification step. The super-computer of High Performance Supercomputing Center of NECTEC (http://www.nectec.or.th/bureaux/hpcc/) is used for simulating both SOFM and back propagation networks. The problem of memory and time complexity has been handled successfully.

AngsanaUPC: จงภูมิใจเถิดที่เกิดเป็นไทย

A quick brown fog jumps over the lazy dog.

BrowalliaUPC: จงภูมิใจเถิดที่เกิดเป็นไทย

A quick brown fog jumps over the lazy dog.

Figure 4. Examples of the experimented typed fonts.

Each input document image is passed through the preprocessing techniques which include alignment (in case that the image is not in the upright position), noise cleaning, line & character segmentation, and each character's image normalization. After all preprocessing technique, image of each character is encoded in an array of size 8*8. This number is obtained from several experiments of several different sizes of input of Thai typed OCR in [4].

In rough classification training with the Kohonen SOFM model, the learning rate is gradually decreased and the learning process is stopped when no more weight changes occur. The first experiment is for determining the proper number of clusters. According to the probability concepts, the more the number of clusters is, the less accurate the rough classification phase will be, but that will result in the fewer characters per cluster on the average as well. An optimal number of clusters should produce the highest final recognition rate, that is - the *maximum product of rough classification accuracy and fine classification accuracy*. In determining the number of clusters, we examine the number of *distinct* characters per cluster by experimenting on four different number of clusters: 3, 5, 10, 20 clusters. The results on AS and BW fonts are shown in Table 1.1 and 1.2, respectively. The SOFM network with smaller number of clusters always result in higher number of distinct characters per cluster on the average.

Table 1.1. Shows the number of characters in each cluster of SOFM network on AS font data set.

Cluster = 3		Cluster = 5		Cluster = 10		Cluster = 20	
CR	NO	CR	NO	CR	NO	CR	NO
1	111	1	70	1	32	1	35
2	160	2	61	2	69	2	22
3	145	3	112	3	39	3	33
		4	93	4	47	4	20
		5	135	5	54	5	25
				6	66	6	5
				7	74	7	36
				8	66	8	20
				9	44	9	19
				10	31	10	16
						11	20
						12	60
						13	10
						14	25
						15	49
						16	44
						17	37
						18	53
						19	22
						20	29

Note :
CR : Cluster No.
NO : Number of distinct characters in each cluster.

From the result in Table 1.1 and 1.2, we have ruled out the use of 3 clusters simply because of two of its clusters contain too many distinct characters. Preliminary experiments by taking 5, 10, and 20 clusters for fine classification have consistently revealed that the highest final recognition rate is obtained with using 5 clusters.

Reprint from Proceeding of the 1997 International Conference on Neural Information Processing and intelligent Information System, Vol. 2, pp. 1227-1230

Table 1.2. Shows the number of characters in each cluster of SOFM network on BW font data set.

Cluster = 3		Cluster = 5		Cluster = 10		Cluster = 20	
CR	NO	CR	NO	CR	NO	CR	NO
1	93	1	74	1	70	1	49
2	163	2	131	2	49	2	48
3	172	3	109	3	76	3	17
		4	112	4	46	4	25
		5	128	5	58	5	4
				6	68	6	30
				7	42	7	21
				8	74	8	15
				9	54	9	30
				10	54	10	32
						11	22
						12	*26*
						13	*19*
						14	58
						15	19
						16	19
						17	44
						18	53
						19	28
						20	34

Note :
CR : Cluster No.
No : Number of distinct characters in each cluster.

The recognition rate of the full Thai & English OCR system which roughly classify input characters into one of the five clusters on the unseen training font documents are summarized in Table 2. Each back propagation network contains 64 input bits, 64 hidden neurons, and the number of output nodes being equal to the number of distinct characters in each cluster. The experimental data are real two-page documents of five different point sizes, two of which having untrained point sizes. Documents are generated in both of the trained fonts. Each document is scanned with the resolution of 200 dpi (Black and White) and 300 dpi (Sharp Black and White).

As shown in Table 2, the minimum recognition rate is 95.83%, already outperforms the approximately 90% recognition rate of the technique used in [4] which is implemented in a commercial Thai & English software [6]. The average recognition rate is 97%, a new high level of Thai & English OCR performance.

Table 2. The final recognition rate based on rough classification with 5 clusters.

Font	Type of Testing Data	Recognition Rate
AS	trained font sizes, 300 dpi	97.59 %
	trained font sizes, 200 dpi	96.93 %
	untrained font sizes, 300 dpi	95.83 %
	untrained font sizes, 200 dpi	98.38 %
BW	trained font sizes, 300 dpi	97.13 %
	trained font sizes, 200 dpi	95.91 %
	untrained font sizes, 300 dpi	97.54 %
	untrained font sizes, 200 dpi	96.62 %

Since future attempt will be focused on recognizing documents which contain mixed fonts. It is interesting to see how similar the characters of various fonts are clustered by SOFM networks. Table 3 summarizes the result of using SOFM weights trained by either AngsanaUPC (1st row) or BrowalliaUPC (2nd row) font to cluster characters from other Microsoft Windows system fonts. They are CordiaUPC (CD), DilleniaUPC (DL), EucrosiaUPC (EC), FreesiaUPC (FS), IrisUPC (IR), and JasmineUPC (JM). The high average percentage of common SOFM clustering among various fonts suggests that training our two-step classifying neural networks with character sets from combined fonts could yield better recognition rate for mixed-font documents.

4. Conclusion

At present, Thai & English OCR software is highly required by Thai users in all fields. The application of SOFM network to distribute the workload for back propagation learning has made the multi-layered perceptron, trained with back propagation algorithm, able to classify Thai & English characters more accurately. Finally, our two-step classification system produces significantly higher recognition rate over any existing Thai & English OCR system.

Reprint from Proceeding of the 1997 International Conference on Neural Information Processing and intelligent Information System, Vol. 2, pp. 1227-1230

Hence, a new highly improved Thai & English OCR software could be expected in the very near future

Table 3. The percentage of SOFM clustering applied to various other fonts.

font	AS	BW
AS	100	95.0
BW	96.6	100
CD	94.2	98.0
DL	95.2	95.0
EC	98.7	96.8
FS	96.3	99.2
IR	96.8	97.0
JM	96.7	95.0

References

[1] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning Internal Representations by Error Propagation, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition,* Vol. 1, Foundations, A Bradford Book, MIT Press, 1986.

[2] HNC Neurosoftware, Self Organizing Map, *HNC Hecht-Nielsen Neuro-computers ANZA Plus,* pp. 11-1 - 11-10, 1989.

[3] C. Kimpan and S. Walairacht, Thai Characters Recognition, *Proceedings of the Symposium on Natural Language Processing* in Thailand, pp. 196-276, March 1993.

[4] C. Tanprasert and T. Koanantakool, Thai OCR : A Neural Network Application, *Proceedings of 1996 IEEE Region Ten Conference*, Vol. I, pp. 90-95, November 1996.

[5] [SNNS, 1994] *SNNS (Stuttgart Neural Network Simulator) User Mannaul, Version 4.1,* University of Stuttgart, Institute for Parallel and Distributed High Performance Systems (IPVR), Report No. 3/94.

[6] [ArnThai v1.0, 1997] *ArnThai Version 1.0*, the Thai & English OCR software, cooperation between the National Electronics and Computer Technology Center (NECTEC) and ThaiSoft Company, January, 1997.

Chularat Tanprasert received her BS (Mathematics, 1st honor, 1989) from Chulalongkorn University, and her MS (1991) and PhD (1994) degrees in computer science from University of Southwestern Louisiana, USA. After her graduation, she has joined the Software and Language Engineering Laboratory (SLL) of NECTEC and worked in Thai OCR project since then. She has also been involved in a number of practical projects such as Thai word processor, web 13, and speaker identification. Her areas of interest include neural networks, pattern recognition, database systems, artificial intelligence, data mining, and knowledge base.

Wasin Sinthupinyo graduated B.Sc.(Math) from Khonkean University and M.Sc.(Computer) from Chulalongkorn University. He passed the Third Country Training Programme in "Intelligent System for Mannagement Information System Managers" course from JAPAN-SINGAPORE AI Center.Singapore. He presently is responsible for two main projects: Thai OCR Project and Web 13. He interested in Object Oriented Database Design, AI, Software Engineering and CBR.

Premnath Dubey (Prem), graduated in Computer Science from Ramkhamhaeng University, have been working at NECTEC in SLL (previously known as SWL) since April,1995. He is one in the team of Thai OCR Project. His mainly responsibility is research and development technologies in OCR for Thai and English language. His topics of interest include OOA/OOD , Component Base Development, Digital Signal Processing, Computer graphics, Distributed Computing.

Thitipong Tanprasert received his B.Engr (Electrical Engineering, 1987) from Chulalongkorn University, and his MS (1989) and PhD (1993) degrees in computer engineering from University of Southwestern Louisiana, USA. He is currently working at Assumption University as Head of computer science depratment and Director of master degreee in computer science program. His areas of interest include neural networks, pattern recognition, VLSI, and computer architecture.

Reprint from Proceeding of the 1997 International Conference on Neural Information Processing and intelligent Information System, Vol. 2, pp. 1227-1230