Font Descriptor Construction for Printed Thai Character Recognition

Ungsumalee Suttapakti[1], Kuntpong Woraratpanya[2], and Kitsuchart Pasupa[3]
Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang,
Chalongkrung Rd. Ladkrabang, Bangkok, Thailand 10520
e-mail: un.ung@hotmail.com[1], kuntpong@gmail.com[2], and kitsuchart@it.kmitl.ac.th[3]

Abstract

The font evolution with various types is a great impact on a recognition performance of optical character recognition (OCR) systems. The more diversity of fonts leads to the less accuracy of recognition rate, particularly Thai-fonts. In order to overcome this obstacle, this paper proposes a font descriptor for printed Thai-character recognition. The role of such a descriptor is a representative of various fonts and sizes. The font descriptor construction is based on principal component analysis (PCA) in a combination with predefined patterns in multi-level processing. The proposed font descriptor is tested on Thai character image corpus consisting of consonants, vowels, and tones. The experimental results show that the proposed font descriptor is efficient and robust to font type and size variations.

1. Introduction

With constantly increasing variety of Thai-font types, this degrades recognition rate of optical character recognition (OCR), which is a process of converting document images to editable text. That is, the more diversity of fonts leads to the less accuracy of recognition rate [1]. This problem is one of the grand challenges to improve recognition rate.

Over the past decades, printed Thai-character recognitions have been continually researched. As the recognition performance is a key issue, many techniques have been introduced in order to improve the recognition rate. The important factors in achieving high recognition rate can be roughly divided into two categories which are (i) feature extraction and (ii) recognition algorithms.

The former category focuses on feature extractions. For instances, Tangwongsan et al. [2] presented stroke structural features and classification rules for the recognition system of printed Thai documents, and Kawtrakul et al. [3] introduced Thai character recognition based on multiple features and minimum Euclidean distance technique to classify unknown symbols. Both techniques improve higher recognition rate.

The latter category concentrates on recognition algorithms to improve performance. For examples, Tanprasert et al. [4] presented artificial neural network (ANN), Kohonen self-organizing feature map, and back propagation algorithms to perform a two-step classification of all characters consisting of two fonts and two resolutions. Kijsirikul et al. [5] introduced a combination of back propagation neural network and inductive logic programming (ILP). Thammano et al. [6] presented hierarchical cross-correlation ARTMAP neural network for recognizing printed Thai characters of without head fonts. Jawahar et al. [7] presented the character recognition process from printed documents containing Hindi and Telugu text. The bilingual recognizer was based on principal component analysis (PCA) and followed by support vector machine (SVM). Aradhya et al. [8] introduced an approach based on Fourier transform and PCA for printed South Indian scripts and English documents. These papers focus on recognition algorithms to improve the performance.

However, as mentioned in the previous paragraphs, most of the papers show achievements of improving recognition rate, but none of these approaches are tested with variant Thai-fonts and character sizes. A few papers studied on various fonts and sizes of Arabic [1], and South Indian and English [8].

Therefore, this paper proposes a method to generate a font descriptor by means of PCA in a combination with predefined pattern in multi-level processing. Such a font descriptor is invariant to different fonts and sizes of printed Thai-characters. This helps the Thai-OCR to perform on various font types with higher recognition rate.

This paper is organized as follows: section 2 reviews backgrounds of Thai characters and PCA. In section 3, a font descriptor construction is proposed. Section 4 shows experimental results and discussions. Finally, conclusions are presented in section 5.

2. Background

This section describes characteristics of Thai language and analyzes factors which have an impact on recognition rate.

2.1 Characteristics of Thai language

Thai typing starts from left to right and from top to bottom. It does not require spaces between words and sentences. Figure 1 shows a noun phrase “คิดกลยุทธ์” (thinking strategy) consisting of three-level characters [9], i.e., “ิ” and “์” are in an upper level, “ค ด ก ล ย ท and ธ” are in a middle level, and “ุ” is in a lower level. In addition, Thai alphabet is composed of 53 character images for middle level, 12 character images for upper level, and 2 character images for lower level. This paper uses characteristics of Thai language to analyze factors which have an impact on recognition rate.

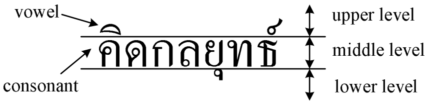


Figure 1. Thai characters in three levels.

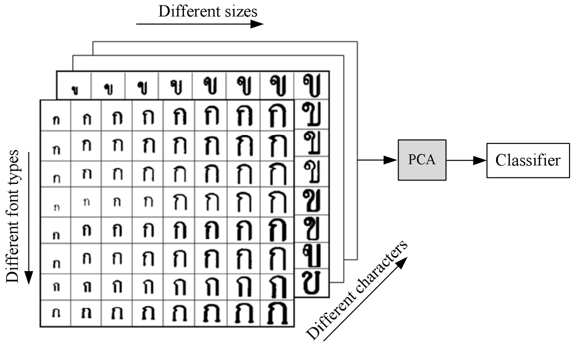


Figure 2. The first level of feature extraction and feature vector classification.

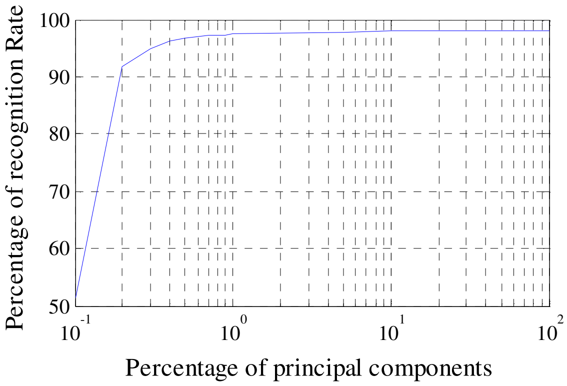


Figure 3. Tendency of percentage of recognition rate versus percentage of principal components.

2.2 Analysis of recognition rate by PCA

PCA is a well-known transformation used for dimensionality reduction in a large data set by retaining its characteristics that mostly contributes to its variance. It aims to keep the highest variance principal components, which often contain the most important aspects of the data, and ignore the lowest variance principal components.

In this subsection, PCA is used as a tool in OCR. Initially, all Thai-characters with different fonts and sizes are formed as shown in Figure 2. Then each character is resized to a bounding box with 32×32 pixels. Hence, the different image sizes are in a standard form. Then the PCA is applied to a matrix which is extracted the features in the form of principal components. The key factor directly related to recognition performance is the number of principal components. Thus, the tendency of recognition rate versus percentage of principal components is tested and graphically depicted in Figure 3. This graph shows that the recognition rate approaches to the highest accuracy when 10% of principal components are used. Therefore, this is a suitable number of principal components. In other words, the recognition rate is retained as high as possible, while the minimum number of components is selected. However, Table 1 illustrates misclassification characters (MC) when 10% of principal components are used to classify all characters by means of Euclidean distance. In middle characters, for example, ต is classified as ด character class (No. 20), and vice versa. Figure 4 plots three principal components of four characters, ต, ด, ฎ, and ฏ. The distribution of components clearly shows that those characters can be classified into two classes, ต-ด and ฎ-ฏ classes. In Table 1, the upper character ◌้ is classified as ◌ั character (No. 54). These cases are called misclassification which often occurs when the shape of characters is similar. In lower level, No. 66-67, there is no misclassification, since both characters have totally different shapes. Misclassification is a crucial problem in recognizing Thai alphabets. In order to solve this problem, a font descriptor is regenerated. The font descriptor construction is explained in the next section.

Table 1. Misclassification of characters

No.	char	MC	No.	char	MC	No.	char	MC	No.	char	MC	No.	char	MC
1	ก	ถ ภ	15	ฏ	ฎ	29	ฝ		43	ห		57	◌ี	◌ี ◌็
2	ข	ฃ ช	16	ฐ		30	พ		44	ฬ		58	◌ึ	◌ี
3	ฃ	ข ช	17	ฑ		31	ฟ		45	อ		59	◌ื	◌ิ ◌ี ◌ึ
4	ค	ฅ	18	ฒ		32	ภ	ถ ก	46	ฮ	ธ ช	60	◌็	◌ึ
5	ฅ	ค ศ	19	ณ		33	ม		47	ฯ		61	◌่	◌ำ
6	ฆ	ฃ	20	ด	ต	34	ย	ฆ	48	า		62	◌๊	◌ำ
7	ง		21	ต	ด	35	ร		49	เ		63	◌๋	
8	จ		22	ถ	ฉ	36	ฤ		50	โ		64	◌์	
9	ฉ		23	ท		37	ล		51	ใ	โ ไ	65	◌ำ	◌๋
10	ช	ซ ข	24	ธ		38	ฦ	ฤ	52	ไ		66	◌ุ	
11	ซ	ช	25	น	ม	39	ว		53	ๆ		67	◌ู	
12	ฌ	ณ	26	บ	น	40	ศ		54	◌ั	◌้			
13	ญ		27	ป		41	ษ		55	◌้				
14	ฎ	ฏ	28	ผ		42	ส		56	◌ิ	◌ำ			

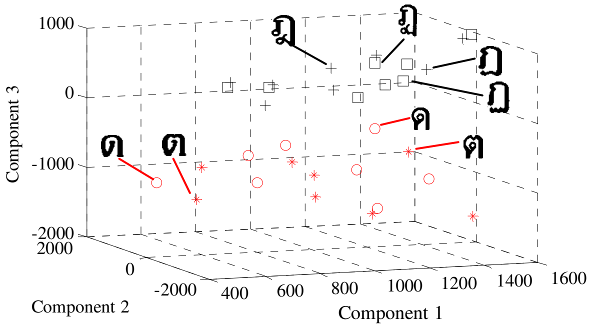


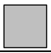
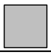
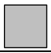
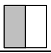
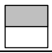
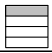
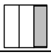
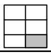
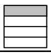
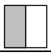
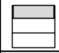
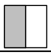
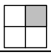
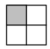
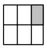
Figure 4. Misclassification occurred with ต, ด, ฎ, and ฏ characters and demonstrated by three principal components.

3. Font Descriptor Construction

In order to construct a font descriptor (FD) for printed Thai character recognition, all Thai-characters with different fonts and sizes are formed as shown in Figure 2. The PCA is applied to extract the features and 10% of principal components are selected to form feature vectors. These feature vectors are classified by means of Euclidean similarity measure. As a result, each class can represent one or more feature vectors. If one class contains only one feature vector, it implies a perfect recognition. Otherwise, each class is classified by predefined patterns as shown in Table 2.

A predefined pattern is a sub-region of a difference of characters in the same class. It is useful to identify and classify similar characters. Figure 5 shows an example of constructing the predefined pattern in the vertical segmentation. In Table 1, ก, ถ, and ภ, shown in Figures 5(a), 5(b),

Table 2. Predefined patterns of middle, upper, and lower level characters for 1st, 2nd, and 3rd PCA.

1st PCA			2nd PCA										3rd PCA		
middle	upper	lower	middle								upper		middle		upper
P1	P1	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14
															
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 ฤ ล ฦ ว ศ ษ ส ห ฬ อ ฮ ฯ า เ โ ใ ไ ๆ	◌ั ◌ั ◌ิ ◌ี ◌ื ◌็ ◌่ ◌๊ ◌๋ ◌์ ◌ำ	◌ุ ◌ู	ก ถ ภ	ข ฃ ช ซ	ค ฅ ศ	ฌ ณ	ฎ ฏ	ด ต	ภ ฤ	ใ โ ไ	◌ั ◌ั	◌ิ ◌ี ◌ึ ◌ื	ข ฃ	ช ซ	◌ึ ◌ื

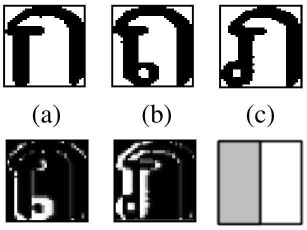
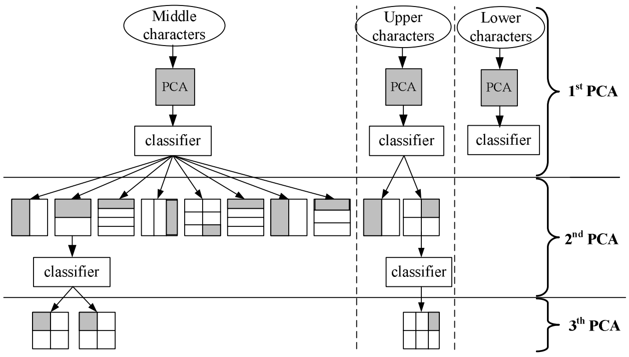

(a) (b) (c)
(d) (e) (f)

Figure 5. An example of a predefined pattern construction.



Note: ▨ is a feature extraction sub-region

Figure 6. Font descriptor construction by using PCA in a combination with predefined patterns in multi-level processing for middle, upper, and lower level characters.

and 5(c), respectively, are recognized as the same character class. Figure 5(d) represents a difference between ก and ถ characters, while Figure 5(e) represents a difference between ก and ภ characters. The difference is represented with white color regions useful for pattern constructions. In this case, the predefined pattern in Figure 5(f) can be generated by a vertical symmetric segmentation according to the difference of Figures 5(d) and 5(e). Therefore, the predefined pattern of this class can be generated as illustrated in Figure 5(f). The gray-color area of the pattern is extracted features by the 2nd PCA. The remaining predefined patterns are constructed by the same criteria as previously mentioned and are followed by 2nd or 3rd PCA, as shown in Figure 6. In this way, the predefined patterns help increase the precision of features extracted by PCA method. Finally, the font descriptor is formed as feature vectors which are independent to font and size variations.

Table 3. A result of testing robustness on different fonts.

Different fonts	Recognition rate (%)					
	Middle		Upper		Lower	
	PCA1	FD1	PCA1	FD1	PCA1	FD1
AngsanaUPC	79.25	**90.09**	87.50	**95.83**	100.00	100.00
BrowalliaUPC	91.75	**94.58**	92.71	**96.88**	100.00	100.00
CordiaUPC	86.56	**91.98**	84.38	**94.79**	100.00	100.00
DilleniaUPC	92.45	**95.99**	88.54	**94.79**	100.00	100.00
EucrosiaUPC	88.68	**91.27**	88.54	**96.88**	100.00	100.00
FreesiaUPC	89.86	**91.04**	78.13	**90.63**	100.00	100.00
IrisUPC	73.82	**84.20**	70.83	**85.42**	100.00	100.00
JasmineUPC	45.05	**58.96**	79.17	**84.38**	100.00	100.00
Average	80.93	**87.26**	83.72	**92.45**	100.00	100.00

Table 4. A result of testing robustness on different sizes.

Different sizes	Recognition rate (%)					
	Middle		Upper		Lower	
	PCA1	FD1	PCA1	FD1	PCA1	FD1
8	91.04	**96.93**	86.46	**89.58**	100.00	100.00
10	92.22	**93.63**	93.75	**96.88**	100.00	100.00
12	94.81	**98.11**	95.83	95.83	100.00	100.00
14	95.99	**99.06**	91.67	**98.96**	100.00	100.00
16	96.46	**99.29**	95.83	**97.92**	100.00	100.00
18	96.93	**99.06**	95.83	**97.92**	100.00	100.00
20	95.75	**98.82**	92.71	**96.88**	100.00	100.00
22	95.75	**98.58**	89.58	**94.79**	100.00	100.00
Average	94.87	**97.94**	92.71	**96.09**	100.00	100.00

4. Experimental Results

In order to evaluate the efficiency of a font descriptor (FD) for printed Thai character recognition, two experiments are set up. The first experiment aims to test the robustness of the proposed font descriptor. The font descriptor 1 (FD1) is constructed from features—extracted by PCA in a combination with predefined patterns in multi-level processing—using a leave one out method. That is, test font-types and font-sizes are not included in training sets. The second experiment aims to test the recognition rate. The font descriptor 2 (FD2) is constructed from features, extracted from all font types and font sizes. The PCA1 and PCA2, baseline methods, are constructed by the same criteria as FD1 and FD2, respectively. The raw materials are Thai character image corpus consisting of consonants, vowels, and tones. A resolution of such images is a 400 dpi. Regular font types are divided into AngsanaUPC, BrowalliaUPC, CordiaUPC, DilleniaUPC, EucrosiaUPC, FreesiaUPC, IrisUPC, and JasmineUPC, and font sizes are divided into 8, 10, 12, 14, 16, 18, 20, and 22. There are totally 8,576 samples.

For robustness testing, the experimental results show that the proposed FD1 outperforms the PCA1 in all cases of middle and upper level characters as illustrated in Tables 3 and 4. In case of lower level character, both me-

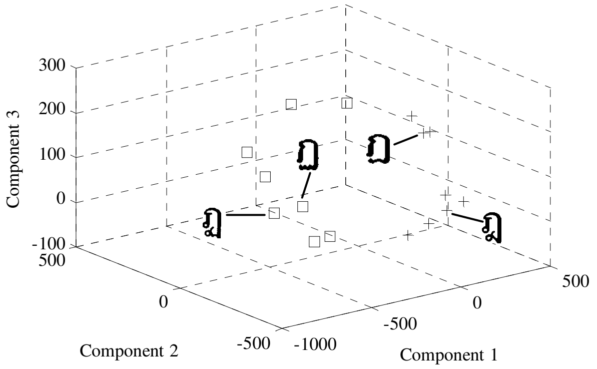


Figure 7. A ฎ-ฏ class classified into two classes by using PCA in a combination with the predefined pattern, P6 in Table 2, in multi-level processing.

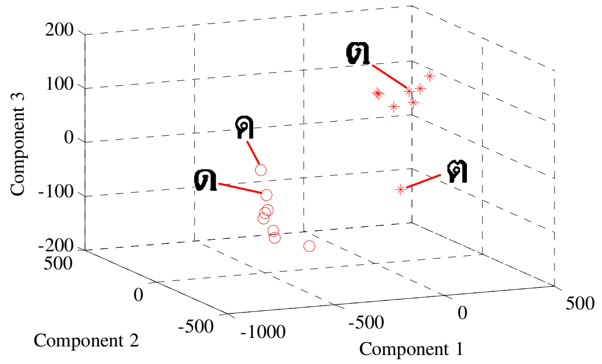


Figure 8. A ต-ด class classified into two classes by using PCA in a combination with the predefined pattern, P7 in Table 2, in multi-level processing.

Table 5. A comparison of recognition rates of PCA1, PCA2, proposed FD1, and proposed FD2.

Level of characters	Recognition rate (%)					
	Robustness test				Recognition test	
	Different fonts		Different sizes		Diff-fonts-sizes	
	PCA1	FD1	PCA1	FD1	PCA2	FD2
Middle	80.93	**87.26**	94.87	**97.94**	97.61	**99.32**
Upper	83.72	**92.45**	92.71	**96.09**	95.31	**97.01**
Lower	100.00	100.00	100.00	100.00	100.00	100.00
Average	88.22	**93.24**	95.86	**98.01**	97.64	**98.78**

thods yield the same results, 100%. In Table 3, the highest recognition rate of middle level is DilleniaUPC font and upper level is BrowalliaUPC and EucrosiaUPC fonts. The lowest recognition rate of middle and upper levels is JasmineUPC, since this font has no character heads, making it difficult to provide a good representative. FD1 can improve the recognition rates for different fonts in average 7.82% and 10.43%, for middle and upper levels, respectively, when compared with PCA1. In the same way, Table 4 also shows the improvement of the recognition rate; FD1 is better than PCA1. This implies that the FD1 is more robust than PCA1.

For recognition rate testing, all different fonts and sizes are used to construct FD2. The result proves that the FD2 provides the good representative. Figures 7 and 8 evidently show that using PCA in a combination with the predefined pattern in multi-level processing can solve the misclassification problem. In comparison of results as indicated in Table 5, the highest recognition rate is Diff-fonts-sizes, 98.78%. The recognition rate of different sizes and fonts are slightly reduced to 98.01% and 93.24%, respectively. In case of different font test, the recognition rate of PCA1 is reduced by 9.65% when compared to PCA2, whereas FD1 is reduced by 5.61% when compared to FD2. This proves that the proposed method is more robust to font type variations.

5. Conclusions

In this paper, the font descriptor construction for printed Thai characters is proposed. The contribution of this paper is constructing a font descriptor from features of different fonts and sizes by using PCA in a combination with predefined patterns in multi-level processing. The experimental results illustrate that the proposed font descriptor is invariant to font and size variations of printed Thai characters. Furthermore, it is robust to unknown fonts when tested with leave one out method. The proposed method achieves the efficiency and robustness. This helps improve the OCR efficiency when new fonts are available in the future.

References

[1] F. Slimane, S. Kanoun, J. Hennebert, A. Alimi, and R. Ingold: "A Study on Font-family and Font-size Recognition Applied to Arabic Word Images at Ultra-low Resolution," *Pattern Recognition Letters*, pp.209–218, 2013.

[2] S. Tangwongsan and O. Jungthanawong: "A Refinement of Stroke Structure for Printed Thai Character Recognition," *Proceedings of the 9th International Conference on Signal Processing*, pp.1504–1507, 2008.

[3] A. Kawtrakul and P. Waewsawangwong: "Multi-feature Extraction for Printed Thai Character Recognition," *The Fourth Symposium on Natural Language Processing*, 2000.

[4] C. Tanprasert, W. Sinthupinyo, P. Dubey, and T. Tanprasert: "Improved Mixed Thai & English OCR using Two-step Neural Net Classification", *NECTEC Technical Journal*, vol.1, no.1, pp.41–46, 1999.

[5] B. Kijsirikul, S. Sinthupinyo, and A. Supanwansa: "Thai Printed Character Recognition by Combining Inductive Logic Programming with Backpropagation Neural Network", *The 1998 IEEE Asia-Pacific Conference on Circuit and Systems*, pp.539–542, 1998.

[6] A. Thammano and P. Duangphasuk: "Printed Thai Character Recognition using Hierarchical Cross-correlation ARTMAP," *Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence*, pp. 695–698, 2005.

[7] C. V. Jawahar, M. N. S. S. K. Pavan Kumar, and S. S. Ravi Kiran: "A Bilingual OCR for Hindi-Telugu Documents and Its Applications," *International Conference on Document Analysis and Recognition*, pp.656–660, 2003.

[8] V. N. Manjunath Aradhya, G. H. Kumar, and S. Noushath: "Multilingual OCR System for South Indian Scripts and English Documents: An Approach Based on Fourier Transform and Principal Component Analysis," *Engineering Applications of Artificial Intelligence*, vol.21, pp.658–668, 2008.

[9] R. Chamchong and C. Fung: "Text Line Extraction using Adaptive Partial Projection for Palm Leaf Manuscripts from Thailand," *International Conference on Frontiers in Handwriting Recognition*, pp.586–591, 2012.