A Keyword Matching for the Retrieval of Low-Quality Hangul Document Images*

In-Seop Na (나인섭)**
Sang-Cheol Park (박상철)***
Soo-Hyung Kim (김수형)****

Contents

ABSTRACT

It is a difficult problem to use keyword retrieval for low-quality Korean document images because these include adjacent characters that are connected. In addition, images that are created from various fonts are likely to be distorted during acquisition. In this paper, we propose and test a keyword retrieval system, using a support vector machine (SVM) for the retrieval of low-quality Korean document images. We propose a keyword retrieval method using an SVM to discriminate the similarity between two word images. We demonstrated that the proposed keyword retrieval method is more effective than the accumulated Optical Character Recognition (OCR)-based searching method. Moreover, using the SVM is better than Bayesian decision or artificial neural network for determining the similarity of two images.

Keywords: Low-Quality Korean Document Keyword Retrieval, SVM, OCR, Digital Library

---

* A preliminary version of this paper appeared in 'Character Segmentation and Keyword Matching for the Retrieval of Low Quality Korean Document Images', Ph.D. diss., 2006, Chonnam National University, Korea.

** 전남대학교 계약교수(ypencil@hanmail.net) (제1저자)
*** 삼성메디슨 책임연구원(park.sangc@gmail.com) (교신저자)
**** 전남대학교 전자컴퓨터공학부교수(shkim@chonnam.ac.kr)

논문접수일자: 2013년 1월 8일 최초심사일자: 2013년 1월 21일 게재확정일자: 2013년 2월 14일
한국문헌정보학회지, 47(1): 39-55, 2013. [http://dx.doi.org/10.4275/KSLIS.2013.47.1.039]

1. Introduction

Nowadays, many digital documents are in the image format. We can locate many digital documents on the internet in image format, including scanned journal/conference papers, student theses, handbooks, out of print books, etc. Moreover, many digital libraries and Web portals such as ACM, IEEE, MEDLINE, etc., keep scanned document images without their text equivalents. Many tools have been developed for retrieving information from text format documents, but they are not applicable to image format documents. There is, therefore, a need to study the strategies of retrieving information from document images. For example, a user facing a large number of imaged documents on the Internet has to download each one to see its contents before knowing whether the document is relevant to his/her interest, e.g., by looking for keywords. Obviously, it is of practical value to study a method that is capable of notifying the user, prior to downloading, whether a document image contains information of interest (Lu and Tan 2004).

In the last several decades, there has been much interest in the research area of Document Image Retrieval (DIR) (Doermann 1998; Strathy, Suen, and Krzyzak 1993; Mitra and Chaudhuri 2000) to search digitized document images. Methods of searching document images for a keyword can be classified as Optical Character Recognition (OCR)-based or image-based (keyword retrieval or keyword detection). An OCR-based method uses a text-to-text matching approach, that is, it transforms the document image into machine-readable codes by applying an appropriate recognition process and then examines the document with a keyword (Kim et al. 2005; Marukawa et al. 1997; Ohta, Takasu, and Adach 1997). Although the technology of OCR may be utilized to automatically transfer the digital images of these documents to their machine-readable text format, the OCR still has its inherent weaknesses in recognition ability, especially for poor quality document images. Generally speaking, manual correction or proofing of the OCR results is usually unavoidable, which is typically not cost effective for transferring a huge amount of paper documents to their text format (Lu and Tan 2002).

Some keyword retrieval approaches for English document images have been reported in the past few years (Chen, Wilcox, and Bloomberg 1995; Lu and Tan 2002; DeCurtins and Chen 1995; Tan et al. 2002). As a representative among them, Lu and Tan represents each word as a string of feature codes and therefore each document image can be represented as a series of strings. To match a query word with the words in a document, an inexact string matching technique is used. English is the string representation, but Korean is not much. There are some studies on the image-based keyword retrieval of Korean document images. Kim et al. (Kim et al. 2005) segments

a document image into word images by using Kwak's algorithm (Kwag 2001), and constructs a database of word images. To make query images, they generated a character set using a text-editing tool with the same font used in the document image. A two-stage retrieval scheme is used to reduce processing time, where a profile feature and a wavelet feature are used in each stage respectively. Furthermore the wavelet feature is obtained by selecting the largest 30 coefficients using the Haar wavelet transform. Korean(Hangul) shall be extracted from spatial space. The performance of features extracted from the spatial space is affected by three font shapes and font sizes.

In this study, we proposed and tested a keyword retrieval technique using character models and a support vector machine (SVM) for the retrieval of low-quality Hangul (Korean Character) document images in a digital library. The remainder of this paper is organized as follows. Section 2 describes the keyword matching technique. Section 3 describes the proposed methods, including how to segment words from document images, segment characters, normalize and extract features of reference word DB. Section 4 provides experimental results. Finally, section 5 discusses and summarizes the unique characteristics and limitations of this study.

2. Keyword Matching

To match the keyword image and sub-image (target) in a word image, three matching methods were tested including the Bayesian decision rule, Neural Network and Support Vector Machine (SVM).

2.1 Bayesian Decision Rule

When two character images are composed of same letters, the distance between them in v-dimension features space is smaller than that of the features of different letters. Therefore, the similarity between *i*-th characters of the query and word images is defined in the equation (1). To improve the speed of the matching phase, city block distance requiring the minimum calculation among kwon distance function (Gose, Johnsonbaugh, and jost 1996) is in use. Here T_c is a threshold for character matching.

$$\begin{cases} \text{if} \quad Dist(C_i^q, C_i^t) < T_c, then\, C_i^q \equiv C_i^t \\ else \qquad\qquad\qquad\qquad\quad C_i^q \neq C_i^t \end{cases} \tag{1}$$

$$\text{where } Dist(C_i^q, C_i^t) = \sum_{j=1}^{v} |C_{i,j}^q - C_{i,j}^t| \quad (2)$$

T_c is selected based on the Bayesian decision rule using the training dataset. Given the similarities for 'k' character pairs are computed by the equation (1), the word-level similarity is calculated using the equation (3). Here T_W^B is a threshold for word matching.

$$\begin{cases} \text{if} \quad Dist(Q,T) < T_W^B, then \; Q \equiv T \\ else \qquad\qquad\qquad\quad Q \neq T \end{cases} \quad (3)$$

$$\text{where } Dist(Q,T) = \frac{1}{k} \sum_{i=1}^{k} Dist(C_i^q, C_i^t) \quad (4)$$

2.2 Artificial Neural Network

An artificial neural network (ANN) trained by back propagation involves three stages including (1) the feed forward of the input training pattern, (2) the back propagation of the associated error, and (3) the adjustment of the weights (Kim 1992). The ANN is a popular machine learning tool used in various pattern recognition applications because of its advantages in learning the function to optimally approximate the relationship between the input features and desired output classes using the relatively noisy or partially available training data. We trained and implemented a feedforward ANN to verify whether two character images are similar. The ANN was comprised of one input layer, one hidden layer, and one output layer. The input layer had 37 neurons for 36 features and one bias. The number of hidden-layer is three composed of 2-times of relevant nodes to one output node and one bias. The learning rate and bias were set at 0.05 and 0.95, respectively. Each neuron used a sigmoid function as the activation function. The two word images matching results by the ANN is based on equation 5 and 6.

$$\begin{cases} \text{if} \quad Score_{ANN}(Q,T) > T_W^A, then \; Q = T \\ else \qquad\qquad\qquad\quad Q \neq T \end{cases} \quad (5)$$

$$Score_{ANN}(Q,T) = \frac{1}{k} \sum_{j=1}^{k} Y_j \quad (6)$$

, where T_W^A and Y_j is a threshold for word matching and similarity score between two character images by the ANN, respectively.

2.3 Support Vector Machine

Support vector machines (SVMs) were well known methods for the supervised learning of high-level semantic concepts over visual features. SVMs have shown their superior abilities in pattern recognition and have been widely adapted for classification in content-based image retrieval (Lazebnik 2006). The basic concept of the SVMs is to construct a hyper-plane or set of hyper-planes in a high dimensional transformed feature space to find the best decision boundaries. The SVMs are trained by finding the separation with the largest distance to the nearest training data points of any class because the larger the margin the lower the generalization error of the classifier. Maximum margin classifiers are well regularized and the high dimensional transformed feature makes the problem more separative (Steinwart and Christmann 2008). In this study, we utilized Torch web site(http://www.torch.ch/) tools for our SVM to measure the similarity between two character images. Equation 7 and 9 are used to verify if a keyword image and the sub-image of word images are the same.

$$\begin{cases} \text{if} \quad Score_{SVM}(Q,T) > T_W^S, then\, Q = T \\ else \qquad\qquad\qquad\qquad\quad Q \neq T \end{cases} \tag{7}$$

$$Score_{SVM}(Q,T) = \frac{1}{k}\sum_{j=1}^{k} S_j \tag{8}$$

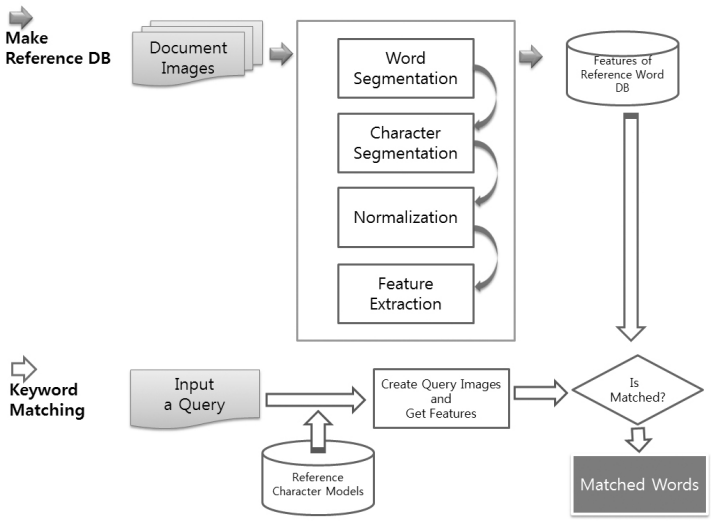
, where T_W^S and S_j is a threshold for word matching and similarity score between two character images by the SVM, respectively.

Then, every word in turn is segmented and each character is normalized into a size of 36 by 36. A 36-D feature is extracted by applying a 6×6 mesh. The system has been implemented in a C/C++ language on a personal computer with a Pentium-4 2.80 GHz CPU and 1 GB RAM.

3. Proposed Method

In this paper, we propose and test a keyword retrieval system using a support vector machine (SVM) for the retrieval of low-quality Korean document images. We propose a keyword retrieval method using a SVM to discriminate the similarity between two word images as shown in Fig. 1. It is assumed that word images, which are segmented by using Jeong's system (Jeong et al.

2005) are already stored in a database. The keyword retrieval technique is composed of feature extraction, designing character models, and keyword retrieval by SVM. In the feature extraction procedure, we choose a mesh feature, which has the best performance in the experiment among well-known features such as mesh, profile, and wavelet, for describing a Korean character. Document images could be written in various fonts and distorted during the acquisition. In designing the character models, we propose character models to produce a robust keyword against different fonts and distortions. Lastly, we propose keyword retrieval using a SVM to discriminate the similarity between two word images. Fig. 1 describes a block-diagram for the proposed keyword retrieval system.



〈Fig. 1〉 A Flow-diagram of the proposed system

3.1 Data Set

We typed part of a Korean book, “Baek-Bum Il Ji”, to make a text file of 20 A4 pages. They were formatted by using a Microsoft word processor in 12 different combinations of fonts, two typefaces (Batang and Gullim), three types of point sizes (8, 10, and 12), and two kinds of font styles (bold and plain). They were printed by a SAMSUNG ML8065 printer, and then copied iteratively by 8 times using a XEROX Document Centre 285 PLUS G copier, and finally scanned by an EPSON GT-30000 scanner at 200 DPI. All the document images were partitioned into word images using the system of (Jeong et al. 2005). One half of the data is used for training,

and the other half is used for testing (Table 1).

〈Table 1〉 Experimental data set

	Typeface	Style	Size	No. Characters	Total
Train	Batang	Bold	8, 10, 12	12,927	51,708
		Plain	8, 10, 12	12,927	
	Gullim	Bold	8, 10, 12	12,927	
		Plain	8, 10, 12	12,927	
Test	Batang	Bold	8, 10, 12	12,756	51,024
		Plain	8, 10, 12	12,756	
	Gullim	Bold	8, 10, 12	12,756	
		Plain	8, 10, 12	12,756	

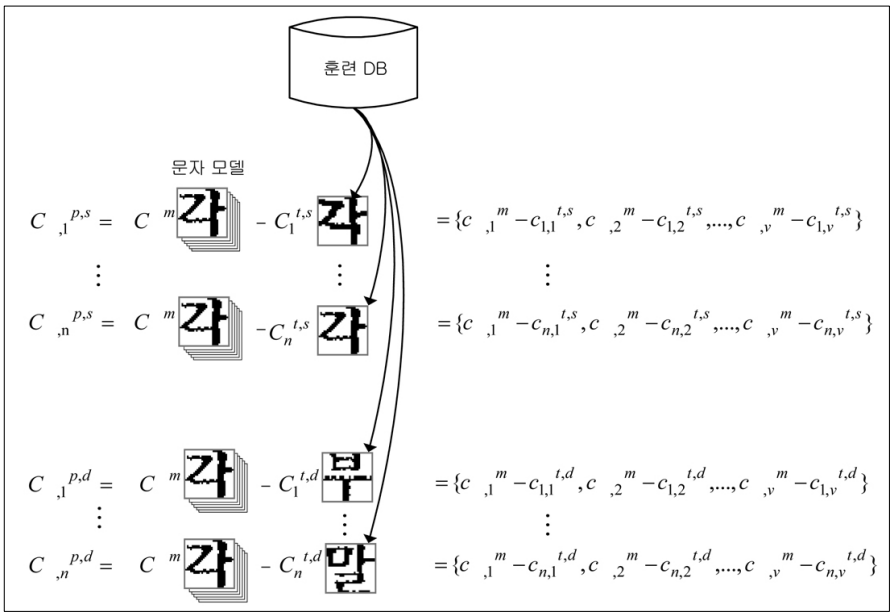
The keywords used in the experiment are selected by the following methods. In the Hangeul documents image, two-character words (words composed of two characters) and three-character words have the highest frequency. Therefore, 30 keywords of two-character words and three-character words are used to assess the searching capacity as shown in table 2. The selection method for 30 keywords is first to align two-character words and three-character words of train data in line with their frequency; then to select the top 20 two-character words and 10 three-character words. To execute keyword matching in Bayesian decision rules, neural network and SVM, the training of each descriptor should be preceded in advance.

〈Table 2〉 Examples of 30 keywords

	Keywords					
1	선생	마음	서울	아들	동학당	그동안
2	진사	모양	머리	접주	황해도	어머니
3	나라	동네	보고	결심	이야기	부모님
4	말씀	이름	다리	당시	저고리	충청도
5	해주	생각	형제	도유	아버지	호랑이

The training of each descriptor is done in the unit of character and when all characters of the search words are successful in matching, the critical value is decided for the decision of similarity regarding relevant words to each method. Therefore, this study obtains the training data for the descriptor by the following methods. For each explained character model, random extraction for same-shape and different-shape characters are made as 200, respectably. In 400 extracted character

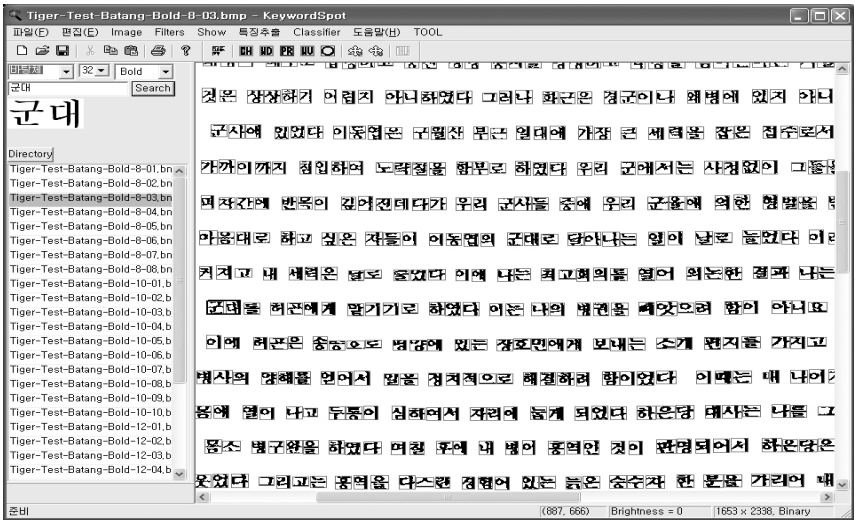
images, the 36-dimension mesh features vector is made. The difference vector among 400 feature vectors responding to the character model and relevant model is used in the training for neural network and SVM and the city block distance of difference vectors is used to decide the critical boundary of Bayes decision rules. In this paper, the kernel of SVM (Kernel) is a Gaussian kernel. Penalty for Sigma and error is decided as 100 and 100, respectably through experiments..



〈Fig. 2〉 Training data model to decide character similarity

Figure 2. illustrates the training data model to decide character similarity. Capital C is a mark for the features vector of a character. The small letter c is a mark for the small factor of a character while the upper index t is for training data; upper index s is for the character same-shape with the character model; upper index p is for the deference vector between two feature vectors. Down index ‘gak’ is a mark for related features vectors of ‘gak’ character and n is an expression for the number of data, set as 200 in the experiment. V is about the dimensions of feature vector, set as 36 in the experiment.

Figure 3. is a demonstration scene of the proposed keyword matching system.



〈Fig. 3〉 Demonstration scene of the proposed keyword matching system

3.2 Character Segmentation, Feature Extraction and Character Model For Query

We use that character segmentation method which is segmented by using Kim's algorithm for robustness to the low-quality Korean document images based on projection profile using *a*-cut (Salton et al. 1994). In the Korean document, there exists a small gap between two adjacent characters.

The keyword retrieval plays its role by inspecting the similarity of two images in use of character image's features. Therefore the extraction of representative features for a character is one of the major issues. As for features extraction, three well-known feature extraction methods are compared to select the features of favorable performance. Before the extraction of features, normalization of character image into 36×36 size is preceded. As for the normalization method, the features density equalization method as one of linear form normalization which is wildly used for character recognition is used. In the image of normalization, the feature extraction method for mesh, profile (Jung, Shin, and Srihari, 1999) and wavelet (Kim et al. 2001) extracts the representative feature vectors for relevant characters. Based on the three features, the most favorable features are selected as the representative features of low-quality Hangeul characters (Kim 2005).

Two groups of the typefaces, Batang (Ming) and Gullim (Gothic), are generally used in Korean documents, but the shapes of characters printed in the typefaces of the groups are quite different. There is no doubt that the image-based keyword retrieval system has poor performance on the

document in different typefaces. Thus classifying the typefaces in advance is more efficient to improve the retrieving performance. In this paper, we assume that the typeface of the document is already discriminated (Kwag 2001).

A number of photocopying and/or scanning of a document can produce some noises or distortions of character strokes. The skew correction might also give additional distortions to the image. Therefore we need a feature minimizing the effect of noises and distortions. In our system, a number of samples for each character class have been used to train the model (Kim 2005).

3.3 Keyword Retrieval and Evaluation

Let's assume that a keyword image (*Q*, query image) and target image(*T*) is composed of characters $C(q/1), C(q/2), ..., C(q/k)$ and $C(t/1), C(t/2), ..., C(t/k)$, respectively. Here, the upper index *q* and *t* means each keyword image and target image. C_i is the *i*-th character of a word image and denoted with the v dimensional features vector as $C_i = (C_{i,1}, C_{i,2}, ..., C_{i,v})$,. The character model is used to represent characters of the keyword image (Kim 2005). To match the keyword image and sub-image (target) in a word image, three matching methods were tested including the Bayesian decision rule (Fausett 1994), Neural Network (Fausett 1994) and Support Vector Machine (SVM) (Steinwart and Christmann 2008).

The recall rate refers to the ratio of searched words out of relevant words to searching words in the data base. The precision rate refers to the ratio of identical and searching words out of searched words. Generally, the recall rate and precision rate are used for the performance evaluation of a searching system; but when it comes to the evaluation of two systems, the recall rate and precision rate have their limits in the capacity comparison since they are not unified measurements. In this study, the harmonized mean-F(F-measure, harmonic mean F) as an integration of recall rate and precision rate is used as a unified measurement (Yates and Neto 1999). Equations are for recall rate, precision rate and harmonized mean-F.

$$Recall = \frac{|Ra|}{|R|} \times 100 \quad (9)$$

$$\mathrm{Precision} = \frac{|Ra|}{|A|} \times 100 \quad (10)$$

$$F-measure = \frac{2 \times Recall \times \mathrm{Precision}}{Recall + \mathrm{Precision}} \quad (11)$$

$|A|$is the sum of the number of relevant words to searching words in the searched words and

the number of other words. $|R|$is the number of relevant words to searching words in the data base words. $|R_d|$is the number of relevant words in the result of searching.

4. Experimental Results

4.1 Individual Character Capacity Comparison

As head of keyword retrieval, the comparison and analysis of two methods are made to assess the differences in matching capacity for individual characters and recognition rates in use of OCR. Table 3 is presents the accumulative recognition rate in use of OCR by size attribute data composed in three character sizes and two thicknesses for Batang and Gullim. The precision rate is assumed as 100%. The size mark is used as an integration of three symbols such as BB8 for Batang, Bold (thickness) and 8 (size 8) and GP10 for Gullim (Gullim), Plain (regular) and 12 (size 12). The capacity of the first ranked recognition rate and the fifth ranked accumulated recognition rate improve substantially in the small size characters of Batang and the capacity in the date relevant to Gullim changes constantly. The experiment confirms that ARMI 6.0 is a recognition system, fitting to data for Batang.

〈Table 3〉 Cumulative recognition rate using OCR

Unit (%)	1Rank		2Rank		3Rank		4Rank		5Rank	
	Recall	F-Measure	Recall	F-Measure	Recall	F-Measure	Recall	F-Measure	Recall	F-Measure
BB8	59.26	74.42	73.70	84.86	79.07	88.31	81.48	89.80	83.89	91.24
BB10	89.07	94.22	94.26	97.04	95.19	97.53	95.74	97.82	95.93	97.92
BB12	92.22	95.95	95.19	97.53	95.74	97.82	95.93	97.92	95.93	97.92
BP8	71.11	83.12	83.89	91.24	87.22	93.18	88.52	93.91	89.63	94.53
BP10	88.52	93.91	90.56	95.04	91.30	95.45	91.85	95.75	91.85	95.75
BP12	93.52	96.65	96.11	98.02	96.11	98.02	96.11	98.02	96.30	98.11
GB8	47.22	64.15	51.67	68.13	56.11	71.89	57.22	72.79	57.59	73.09
GB10	71.67	83.50	75.19	85.84	77.41	87.27	78.52	87.97	79.06	88.30
GB12	78.33	87.85	78.89	88.20	80.19	89.00	80.93	89.46	81.30	89.68
GP8	44.44	61.54	49.81	66.50	52.04	68.45	54.44	70.50	56.30	72.04
GP10	80.93	89.46	82.41	90.36	83.15	90.80	83.33	90.91	84.44	91.57
GP12	87.96	93.60	90.56	95.04	91.30	95.45	91.67	95.65	92.04	95.85

Table 4 and Table 5 show confusion matrices for the character verification of different data by differentiated attributes. MA is for the model of character A; A is for the same character sample with MA; and B is for character sample, different from MB. Therefore, the MA:A value means recall rate and the MB:A value is relevant to false positive error. Table 4 and Table 5 demonstrate that every classifier distinguished each character model exclusively, so our data set is well represented by each font model.

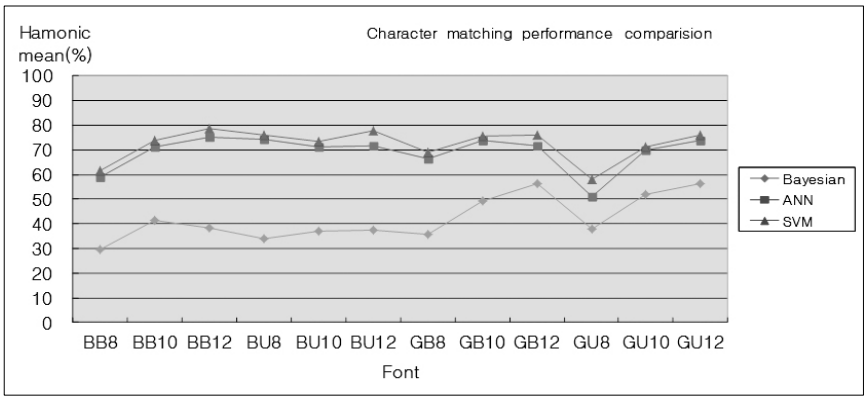
〈Table 4〉 Confusion matrices for Batang

Confusion Matrices (Unit: %)		BB8		BB10		BB12		BP8		BP10		BP12	
		A	B	A	B	A	B	A	B	A	B	A	B
Bayesian	MA	98.33	1.67	97.59	2.41	98.52	1.48	97.22	2.78	98.70	1.30	98.89	1.11
	MB	8.83	91.17	5.16	94.84	5.80	94.20	7.04	92.96	6.31	93.69	6.27	93.73
ANN	MA	91.48	8.52	92.04	7.96	93.15	6.85	87.04	12.96	92.96	7.04	95.74	4.26
	MB	2.26	97.74	1.26	98.74	1.04	98.96	0.89	99.11	1.28	98.72	1.35	98.65
SVM	MA	97.59	2.41	97.78	2.22	98.70	1.30	99.07	0.93	98.33	1.67	98.15	1.85
	MB	2.25	97.75	1.26	98.74	1.00	98.86	1.22	98.36	1.28	98.67	1.02	98.98

Table 5 shows the capacity of character matching for Gullim with 12 different sizes in the harmonized mean. Compared to the recall rate in Table 4 and Table 5, the harmonized mean is lower due to the huge impact from false positive error. The values in Table 4 and Table 5 are illustrated as graph in Fig. 4. Fig. 4 shows that the best performance for 3 kinds of classifiers is SVM.

〈Table 5〉 Confusion matrices for Gullim

Confusion Matrices (Unit: %)		GB8		GB10		GB12		GP8		GP10		GP12	
		A	B	A	B	A	B	A	B	A	B	A	B
Bayesian	MA	99.07	0.93	98.52	1.48	97.96	2.04	94.07	5.93	92.78	7.22	96.48	3.52
	MB	6.68	93.32	3.80	96.20	2.85	97.15	5.76	94.24	3.10	96.90	2.78	97.22
ANN	MA	91.85	8.15	94.44	5.56	92.96	7.04	91.30	8.70	87.96	12.04	90.56	9.44
	MB	1.59	98.41	1.15	98.85	1.25	98.75	3.12	96.88	1.21	98.79	1.04	98.96
SVM	MA	99.07	0.93	98.52	1.48	98.70	1.30	97.22	2.78	97.96	2.04	98.70	1.30
	MB	1.86	98.14	1.29	98.71	1.15	98.85	2.58	97.42	1.20	98.30	1.17	98.46



〈Fig. 4〉 Performance comparison for 3 kinds of classifier

4.2 Comparison Keyword Matching Methods

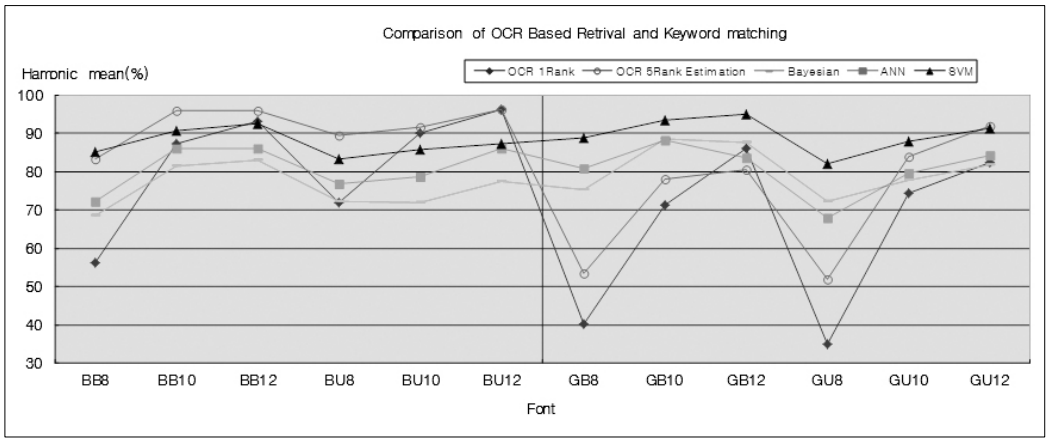
As expected, SVM has shown outstanding performance compared to ANN. Therefore, character matching using SVM is better than that of other methods with other descriptors. For the comprehensive comparison analysis in Table 6 of matching capacity by the OCR-based searching method and Bayesian decision rules, neural network and SVM, keyword retrieval shows 2.9% favorable performance better than that of Bayesian description rules. In Batang and Gullim, neural network is 5.19% and 0.23% respectively better than that of Bayesian description rules. Even though there is no capacity difference in Gullim, the capacity of neural network improves in Batang. It means that the critical boundary which is decided by the multi-layer neural network becomes flexible in Batang data. SVM-using keyword retrieval method illustrates 7.54% in performance improvement to the neural network-oriented matching method. As for Batang and Gullim data, the SVM-using method illustrates 6.50% and 9.06% in performance improvement, respectively.

As for sizes of 8, 10 and 12, the improvement is made in 10.49%, 6.36 and 6.50%, respectively. As for attributes of thickness like bold and normal, the capacity improves 8.14% and 7.43%. It means that the optimal hyper-plane, acquired after the data becomes high-dimension space which is linearly separable improves the capacity of keyword retrieval.

Fig. 5 shows the capacity for each of the three keyword retrieval methods and the OCR-based searching method. What should be looked into here is the huge improvement of capacity which is sluggish in character matching. This is because the similarity decision of words with several characteristics improves the precision rate rather than deciding the similarity only with one character similarity.

〈Table 6〉 comparison keyword retrieval methods

Font	Retrival based OCR (%)	Keyword retrieval Accuracy (%)			Comparison					
		Baye's	ANN	SVM						
Type	F-measure (A)	F-measure (B)	F-measure (C)	F-measure (D)	B-A	C-A	C-B	D-A	D-B	D-C
BB8	56.33	68.53	72.09	85.18	12.20	15.76	3.56	28.85	16.65	13.09
BB10	87.34	81.52	85.9	90.75	-5.82	-1.44	4.38	3.41	9.23	4.85
BB12	93.17	83.01	86.17	92.50	-10.16	-7.00	3.16	-0.67	9.49	6.33
BP8	71.75	72.12	76.66	83.34	0.37	4.91	4.54	11.59	11.22	6.68
BP10	90.07	71.89	78.75	85.71	-18.18	-11.32	6.86	-4.36	13.82	6.96
BP12	96.11	77.45	86.1	87.21	-18.66	-10.01	8.65	-8.90	9.76	1.11
GB8	40.14	75.23	80.82	88.71	35.09	40.68	5.59	48.57	13.48	7.89
GB10	71.39	88.39	88.14	93.50	17.00	16.75	-0.25	22.11	5.11	5.36
GB12	85.93	87.53	83.55	94.85	1.60	-2.38	-3.98	8.92	7.32	11.30
GP8	34.92	72.31	67.85	82.15	37.39	32.93	-4.46	47.23	9.84	14.30
GP10	74.24	77.7	79.64	87.92	3.46	5.40	1.94	13.68	10.22	8.28
GP12	82.4	81.61	84.14	91.38	-0.79	1.74	2.53	8.98	9.77	7.24
Average	76.32	78.22	81.12	88.66	1.90	4.80	2.90	12.34	10.44	7.54



〈Fig. 5〉 Comparison of OCR based and keyword matching retrieval

The other important one is that SVM-using keyword retrieval shows 2% lower capacity than the harmonized mean OCR 5th rank recognition-based searching. The recall rate should be taken into account carefully. SVM-using keyword retrieval has about a 9% higher recall rate than average OCR 5th rank recognition-based searching. Regardless of the quality of Hangeul documents image, it would be useful in applications for all documents searching with keywords.

5. Conclusion

We demonstrated a keyword retrieval system for low-quality Korean document images even characters including adjacent and connected. In addition, we demonstrated that the propose and test images which are created from various fonts are likely to be distorted during the acquisition. We test that commercial OCR is fitted to data for Batang and demonstrated that, regardless of fonts, the proposed keyword retrieval method is more effective than the accumulated Optical Character Recognition (OCR)-based searching method. Moreover, using the SVM is better than Bayesian decision or artificial neural network for determining the similarity of two images. Our system can be further improved by incorporating the capabilities for font recognition, the discrimination of Korean and English texts, and the generation of similar-looking keyword images. The quality of a documents image differs according to the quality of original documents and equipment to obtain the documents image. In particular, it is difficult to obtain image from old documents and low-quality original documents. In the low-quality documents image, OCR recognition capacity is low. The keyword retrieval proposed in this paper can be used as a specialized search method for low-quality scanned document images. Considering the fact that digital libraries provide documents in the form of an image, users are able to tell if the documents are befitting to their interests through keyword retrieval before downloading the documents images. In addition, it can be used as a supportive measure for various systems which use document images as data. From now on, practical data such as the documents images of faxes, old papers and low-quality specified exclusive papers which are difficult to recognize with OCR shall be subject to keyword retrieval application and relevant issues shall be solved.

References

[1] Chen, F. R., Wilcox, L.D., & Bloomberg, D.S. 1995. “A comparison of discrete and continuous hidden Markov models for phrase spotting in text images.” Proc. *International Conference on Document Analysis and Recognition*, 1: 398-402.

[2] DeCurtins, J., & Chen, E. 1995. “Keyword spotting via word shape recognition.” Proc. *SPIE Document Recognition II*, 270-277.

[3] Doermann, D. 1998. “The indexing and retrieval of document images.” a survey. *Computer Vision and Image Understanding*, 70(3): 287-298.

[4] Fausett, L. 1994. *Fundamentals of Neural Networks*. Prentice Hall.

[5] Gose, E., Johnsonbaugh, R., & Jost, S. 1996. *Pattern Recognition and Image Analysis*. Prentice Hall.

[6] Jeong, C.B., Park, S.C., Son, H.J., & Kim, S.H. 2005. "Word Extraction from Table Regions in Document Images for Keyword Spotting." *Lecture Notes in Computer Science*, 214-223.

[7] Jung, M. C., Shin, Y. C., & Srihari, S. N. 1999. "Machine printed character segmentation method using side profiles," in Proc. *IEEE Int. Conf. Systems, Man, Cybernetics (SMC)*, 6: 863-867.

[8] Kwag, H. K. 2001. *A Study on Word Segmentation and Attribute Extraction from Document Images*. Ph.D. dissertation, Chonnam National University, Korea.

[9] Kim, Dae Su, 1992. Neural Network Theory and Applications 1. Ha-Tech jeongbo Press.

[10] Kim, H. G., Yang, J. H., Lee, J. S., & Oh, I. S. 2001. "Image-based retrieval of printed Korean words using wavelets." *Journal of Korea Information Science Society*, 28(2): 91-103.

[11] Kim, Soo H., Park, S.C., Jeong, C.B., Kim, J.S., Park, H.R., & Lee, G.S. 2005. "Keyword Spotting on Korean Document Images by Matching the Keyword Image." *Lecture Notes in Computer Science*, 158-166.

[12] Lazebnik, S., Schmid, C., & Ponce, J. 2006. "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories." *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2169-2178.

[13] Liang, Y., Fairhurst, M.C., & Guest, R.M. 2012. "A synthesised word approach to word retrieval in handwritten documents." *Pattern Recognition*, 45(12): 4225-4236.

[14] Lu, Y., & Tan, C. L. 2002. "Word searching in document images using word portion matching." Fifth IAPR International Workshop on *Document Analysis Systems*, USA, 319-328.

[15] Lu, Y., & Tan, C. L. 2004. "Information Retrieval in Document Image Databases." *IEEE Transactions on Knowledge and Data Engineering*, 16(11): 1398-1410.

[16] Marukawa, K., Hu, T., Fujisawa, H., & Shima, Y. 1997. "Document retrieval tolerating character recognition errors-evaluation and application." *Pattern Recogn*, 30: 1361-1371.

[17] Mitra, M., & Chaudhuri, B.B. 2000. "Information Retrieval from Documents," A Survey, *Information Retrieval*, 2: 141-163.

[18] Ohta, M., Takasu, A., & Adach, J. 1997. "Retrieval methods for English-text width misrecognized OCR characters." *Proceedings of 4th International Conference on Document Analysis and Recognition*, 2: 950-955.

[19] Park, S.C., Son, H.J., Jeong, C.B., & Kim, Soo H. 2005. "Keyword Spotting on Hangul

Document Images Using Two-level Image-to-Image Matching." *Lecture Notes in Artificial Intelligence*, 79-81.

[20] Park, S.C., Son, H.J., Jeong, C.B., & Kim, Soo H. 2006. *Character Segmentation and Keyword Matching for the Retrieval of Low Quality Korean Document Images*. Ph.D. diss., Chonnam National University. Gwangju. Korea.

[21] Rodriguez-Serrano, Jose A. Perronnin, Florent. 2012. "Synthesizing queries for handwritten word image retrieval." *Pattern Recognition*, 45(9): 3270-3276.

[22] Salton, G., Allan, J., Buckley, C., & Singhal, A. 1994. "Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Text." *Science*, 264: 1421-1426.

[23] Strathy, N. W., Suen, C. Y., & Krzyzak, A. 1993. "Segmentation of handwritten digits using contour features." *Document Analysis and Recognition, Proceedings of the Second International Conference*, 577-580.

[24] Steinwart, Ingo, & Christmann, Andreas. 2008. *Support Vector Machines*. New York: Springer-Verlag.

[25] Tan, C. L., Huang, W., Yu, Z., & Xu, Y. 2002. "Image document text retrieval without OCR." *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24(7): 838-844.

[26] Yates, R. B., & Neto, B. R. 1999. *Modern Information Retrieval*. 75-82. ACM press.