International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 18, No. 4 (2004) 541–561

World Scientific
www.worldscientific.com

WORD-LEVEL OPTICAL FONT RECOGNITION USING TYPOGRAPHICAL FEATURES

SOO H. KIM*

Department of Computer Science, Chonnam National University,
300 Yongbong-Dong, Buk-Gu, 500-757 Kwangju, Korea
shkim@chonnam.chonnam.ac.kr

HEE K. KWAG

Artificial Intelligence R&D Department, DongBang Media Co, Ltd, Korea
hkkwag@dbmedia.co.kr

CHING Y. SUEN

Center for Pattern Recognition and Machine Intelligence, Concordia University, Canada
suen@cenparmi.concordia.ca

Previous research efforts on optical font recognition have mostly limited applications since they deal with only a few types of font attributes and estimate them from a line or block of text. This paper proposes a word-level optical font recognition system for printed Korean and English documents. At the word-level, it has the advantages of obtaining more detailed font attributes including the following: script (Korean and English), font style (regular, bold, italic, and underlined), typeface (Myung-jo and Gothic), point size (10, 12, 14 pts), and word length (2, 3, 4, 5 for Korean, and 4 to 10 for English). A hierarchical classifier and several typographical features have been devised for the system, and their effectiveness are proven by an experiment with a database of 100 sets of 264 font categories.

Keywords: Optical font recognition; font attributes; typographical features; hierarchical classifier; script; font style; typeface; point size.

1. Introduction

Today considerable information has to be obtained from various kinds of printed documents, such as newspapers, magazines, books, forms, journals, technical reports, and so on. The amount of paper-based documents is increasing day by day despite the omnipresence of electronic documents. Therefore, there is a huge demand on document imaging technologies for the storage, processing, indexing, retrieval and reproduction of large volume of printed documents.[1]

*Author for correspondence.

Two approaches for document indexing and retrieval have been developed.[4] One reads the document image by an optical character recognition (OCR) system and converts it into an adequate electronic format, and then applies both indexing and retrieval with this format. The other approach is based on keyword spotting where the document image is first segmented into words, and the user keywords are located in the image by a word-to-word matching. Although some researchers have shown that the latter approach is superior in document indexing and retrieval,[12] the two approaches actually complement each other and a hybrid approach is being developed as an alternative.[9]

In this paper, we propose a system for optical font recognition (OFR) that can be used to improve the performance of OCR and keyword spotting technologies. Assuming that the document image is decomposed into words, the system extracts several typographical features from each word and then identifies five types of font attributes, such as script, font typeface, font style, point size, and word length, which are useful in many applications:

- A significant improvement of OCR can be achieved by utilizing the font information. It is reported that the use of a mono-font OCR, rather than a general purpose OCR, or de-italicization of italic style texts can improve the OCR performance drastically.[2,6,11]
- The search space of word-to-word matching in keyword spotting can be reduced by utilizing the script and word-length information.
- Reproduction or reprint of a scanned document can be more accurate with the font attributes as well as the character recognition.
- Generation of logical structure and creation of summary of the document are possible from the text having font attributes other than a normal one.[6]

Numerous studies on optical font recognition have been made, see e.g. Refs. 2, 3, 5, 7, 8, 10, 11, 13–16. Most of them deal with only a few types of font attributes, but the proposed system deals with five types of attributes derived from a survey of 600 English and Korean documents: script (Korean and English), font style (regular, bold, italic and underlined), font typeface (Myung-jo and Gothic), point size (10, 12, 14 pts), and word length (2, 3, 4, 5 for Korean and 4 to 10 for English). The five types of attributes are combined into 264 kinds of fonts for a word image.

In addition, most of the previous methods extract the font attributes from a large chunk of data such as a line or block of text, but our system works with word-level data. The use of a large-scale input enables it to obtain an accurate classification of fonts, but it restricts the application domain of the system because it assumes that all words and characters in the input text have the same font attributes. A word-level OFR is essential for the application of document indexing and retrieval since the keyword matching is performed by words.

A hierarchical classifier has been designed to handle the 264 combinations of fonts by a divide-and-conquer analysis. Also a number of features reflecting typographical properties of word images have been devised for each type of font

attributes, i.e. two features for script identification, three for the style, one for the point size, one for word length, and one (or two) for the Korean (or English) typeface, respectively. The effectiveness of each feature has been proven by an experiment with a database of 100 sets of 264 font categories.

This paper is organized as follows. Section 2 describes some related work on OFR, and Sec. 3 presents an overview of the proposed system. A number of typographical features and the methods of their extractions are shown in Sec. 4. Experiments and evaluations are discussed in Sec. 5. Finally, a conclusion and some comments on future work are presented in Sec. 6.

2. Related Work

Researchers on optical font recognition (OFR) have focused their attention on the extraction of four types of font attributes, such as scripts, font styles, font typefaces, and point sizes. Table 1 summarizes some representative algorithms reported in the last few years. These algorithms differ in font attributes considered, and required input against which the attributes are extracted.

2.1. *Script identification*

Worldwide, there are many different languages in common use and many different scripts in which these languages are typeset.[13] Script or language identification plays an essential role in implementing the capability of recognizing multilingual documents. This capability can be applied to postal automation for international mail, multilingual access to digital libraries, automatic document translation, and so on. In addition, most Asian countries require this capability because they use at

Table 1. Related work on optical font recognition.

	Number of classes				Required
	Script	Style	Typeface	Size	Input
Ref. 7	13				Document
Ref. 16		4	15	4	Line
Ref. 13	2				Line
Ref. 3	2				Line
Ref. 12		2	10	5	Block
Ref. 2		4			Word
Ref. 14	2				Block
Ref. 8			7		Word
Ref. 15		4	6		Block
Ref. 5	2				Line or Word
Ref. 10	5				Line
Ref. 17			100		Char.
Ref. 18		2			Word
Ref. 19		2			Word

least two kinds of scripts in their documents — their own language and other Roman scripts such as English, French, etc. Actually, most script identification methods listed in Table 1 deal with the classification of Oriental and European scripts.

Hochberg *et al.*[7] classified the scripts into 13 kinds. Their system learns a set of templates for each script by clustering "textual symbols", defined as large and long components in the document images of the corresponding script. To identify a new document's script, the system compares a subset of textual symbols from the document to each script's templates, and chooses the script whose templates provide the best match.

The algorithm proposed by Spitz[13] discriminates Asian and Roman scripts — it can discriminate Asian script such as Korean, Chinese or Japanese from four kinds of Roman scripts such as English, French, German or Russian. The algorithm extracts a statistical feature, named "upward concavity" from a line of text, and classifies the script into one of the two based on the differences in the distributions of upward concavities with respect to baseline position.

Ding *et al.*[3] observed that Spitz's algorithm does not work well when they consider 20 kinds of European scripts including Cyrillic as well as other Roman scripts, and proposed another algorithm to distinguish between Oriental and European scripts. Their algorithm extracts three kinds of typographical features from a line of text, and identifies the script of the line using a decision tree. In their processing, however, the input text line should have more than 50 components.

Xi *et al.*[14] proposed a fractal feature, representing the textual structure complexities of a text block, to discriminate between Oriental and European scripts. They found that the fractal feature has small values (less than 1.0) for European text blocks and large values (larger than 1.0) for Oriental ones.

Elgammal *et al.*[5] proposed three statistical features, called histogram peaks, moments, and run-length histogram, for Arabic and English language identification. Also, they measured the performance of each feature both on line-level and word-level data with a back-propagation neural network. Generally, a line-level data contains more information for classification than a word-level data, but the capability of word-level script identification has more applications, including the handling of multilingual text lines.

Pal *et al.*[10] proposed shape-based features, statistical features and some features obtained from a concept of water reservoir for the discrimination of five most popular languages in the world — English, French, Arabic, Devnagari and Bangla. The features are extracted from a text line and a decision tree is used for the classification. In their processing, short lines containing few (10 or less) characters are appended to the previous line to reduce the error in feature extraction.

2.2. *Font style identification*

Certain parts of documents like section or chapter titles, paper abstracts, figure captions, etc., usually appear in a nontypical font style, such as bold, italic, bold-italic,

all-capital, and so on.[2] Identification of font styles has many benefits in document analysis and recognition: improving OCR accuracy by normalizing the variations among different styles, document reproduction and recognition of document structure by assigning subsets of the document with logical labels. Also it facilitates the generation of meta-information from the document such as keywords or summary.

Zramdini *et al.*[16] detected four kinds of font styles, such as regular, italic, bold and bold-italic, used often in English and French documents. Given a text line of about 6 cm, their algorithm extracts several typographical features and identifies the style of the line with a Bayesian classifier. They used 100 English text lines for training and 100 French text lines for testing.

Park *et al.*[11] dealt with italic and nonitalic (regular) styles in Korean documents. A scalar feature is extracted from the projection profile of a text line, and a nearest neighbor classifier using a Mahalanobis distance is applied for the identification.

Chaudhuri *et al.*[2,6] studied more than 6,000 document pages of technical papers extracted from different journals, proceedings, books, etc., and observed that four kinds of font styles, such as regular, italic, bold, and all-capital styles are used most frequently in such documents. Also they proposed a heuristic algorithm for detecting the font styles of English word images.

Zhu *et al.*[15] proposed a method for font style identification from Chinese text blocks. A Gabor filter is used to extract a 32-D texture feature from a block of 128×128 pixels, and a nearest neighbor classifier based on a weighted Euclidean distance is used for the classification of the block into one of the four font styles.

Doermann *et al.*[18] dealt with italic and boldface styles in English documents. To identify italic words, the minimum upright bounding parallelogram is constructed for each component and the slant is measured relative to the vertical axis. Words in which 50% of the characters have slants greater than a threshold are classified as italic. Boldface is also identified as the word level. A morphological opening transform is applied to eliminate nonboldface text, i.e. an erosion operation is applied until more than 80% of the pixels have been eliminated, at which point a dilation operation is applied for an equal number of steps.

Bloomberg[19] also identified italic and boldface words from an English document by using a morphological opening. Italic words are detected by opening with a 13×6 structuring element (SE) which is designed to respond to edges inclined at about 12°. Boldface words are detected similarly by the use of a vertical SE. This method can be applied to a variety of fonts and font sizes by adjusting the SE sizes.

2.3. *Font typeface identification*

Font typeface refers to the font family like "Times", "Helvetica", "Courier", etc. We use different typefaces in the same document when we emphasize some important terms or discriminate some sentences from the others. Similar to font styles, the typeface information can also be used for document reproduction, recognition of logical document structures, generation of meta-information for document indexing

and retrieval. But the most important use of typeface is to help the recognition of multifont characters both in segmentation and recognition.

Zramdini *et al.*[16] detected 15 categories of font typefaces used frequently in English and French documents. Similar to their method for font style identification, they used a Bayesian classifier trained with several typographical features extracted from English text lines, and the classification performance was measured with French text lines.

Park *et al.*[11] dealt with ten popular typefaces in Korean documents. Given an image of text block, they extracted a 64-D feature using a Fourier transformation of the block image, and they adapted a three-layer perceptron for the classification of typeface used in the block. In the process of training, they tried to optimize the number of hidden neurons.

Jung *et al.*[8] proposed a system which can classify the font typeface of a word image. They deal with seven commonly used typefaces for English documents. An input word image is normalized to 9 rows and 9 columns. A 3×3 grid is used to partition the image into 9 equal regions. The number of pixels with a particular slope, 0, 45, 90, or 135° in each region is determined. The resulting 36 values are fed into an MLP for the typeface classification.

Zhu *et al.*[15] proposed a method for typeface identification for Chinese text blocks. Similar to their method for font style identification, a Gabor filter is used to extract a global texture feature from a block of 128×128 pixels, and a nearest neighbor classifier based on a weighted Euclidean distance is used for typeface identification.

Baird *et al.*[17] dealt with 100 typefaces commonly used to print body-text in English documents in US. They extracted a 512-dimensional binary feature from the input alphabet (at the character-level), and applied a Bayesian classifier assuming that the feature vectors extracted from the samples in one typeface take a Bernoulli distribution.

2.4. *Point size identification*

Font size is expressed in typographic points — 1 inch corresponds to 72.27 points. Therefore, one can determine the font size with the scanning resolution and the height (or vertical pixel distance) of the text. Here the most important task is the accurate estimation of text heights.[11,16]

Zramdini *et al.*[16] estimated three kinds of text heights from the vertical projection profile V_p of an English text line, namely height of the whole V_p, height of the upper part of V_p and height of the central part of V_p. A Bayesian classifier is used to classify the point size in the text line into 10, 11, 12 or 14 pts.

Park *et al.*[11] calculated the average of the vertical distances (or heights) of bounding boxes for the words in a text block. A nearest neighbor classifier using a Mahalanobis distance is applied to identify the point size as one of the most popular sizes in document images, ranging from 10 to 14 pts.

2.5. *Required input*

The algorithms listed in Table 1 differ in the required input against which the font attributes are extracted. Some of them regard the input as the entire document image or a block composed of a number of text lines,[7,11,14,15] and some others take the input as a line of text containing a number of words.[3,5,10,13,16] Only a few algorithms deal with word-level data.[5,8,18,19]

The use of a large input, such as the entire document or text block, makes it possible to obtain an accurate and fast estimation of the font attributes of the document. However, it is based on the assumption that the whole data is homogeneous, i.e. the font attributes should be the same in every text of the input image. This fact restricts the application domain of the algorithm.

In line-level font recognition, it is assumed that a text line in the document is written in the same font. It can be useful for postal automation applications that handle international mail where every line of address is usually written in the same language, typeface, size and style.

Generally, a text line in a document contains the words of different font attributes. As an example, a word in a text line is in Korean, but the next word in the same line can be English. In addition, most OCR systems maintain words in the document to apply a lexicon-driven semantic analysis of the character recognition results. Document indexing and retrieval are also performed by words since the keyword in user query or a thesaurus is matched against every word in the document image. A font recognition for the word-level data is necessary in these cases.

3. Proposed System

Proposed is a word-level optical font recognition (OFR) system for printed documents. The font attributes extracted by the system are classified into five types as shown in Table 2. Since the system has to deal with a variety of documents used in Korea, these attributes are deduced from a survey of 600 Korean and English documents including journal papers, company reports, tax forms, mail envelops, official documents, etc.

Most of the documents used in Korea are bilingual, i.e. they are composed of Korean and English words. Four kinds of font styles, two kinds of typefaces and three kinds of point sizes have been identified since they are the most popular in the documents. In terms of word length, there are four kinds for Korean and seven for English, respectively since most keywords in the documents have these numbers of characters. Combining the five types of attributes produces 264 font categories — 96 for Korean and 168 for English. Here the word length attribute has not been tried before, and is recognized first by the proposed system for use in document indexing and retrieval applications.

Recognition of word-level font attributes can be regarded as a problem of classifying the word images into 264 classes. Since this 264-class problem is so huge to be solved by a simple approach, we adopt a divide-and-conquer strategy which

Table 2. Five types of font attributes.

Script	Korean	English
Style	bold, italic regular, underline	bold, italic regular, underline
Typeface	Myung-jo, Gothic	Myung-jo, Gothic
Point size	10, 12, 14 pts	10, 12, 14 pts
Word length	2, 3, 4, 5	4, 5, 6, 7, 8, 9, 10
No. combinations	96 ($4 \times 2 \times 3 \times 4$)	168 ($4 \times 2 \times 3 \times 7$)

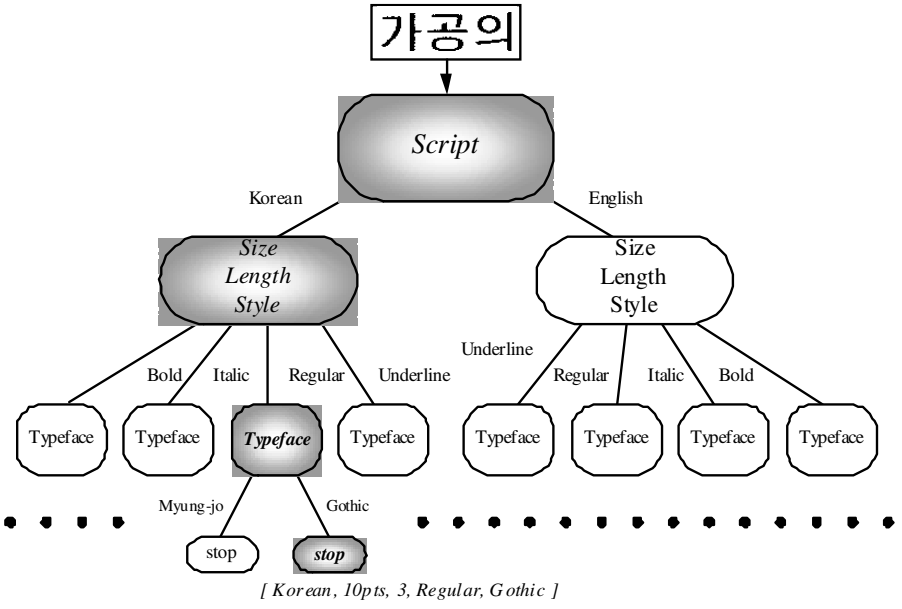


Fig. 1. A decision tree for the hierarchical classification.

splits the problem into a sequence of simpler subproblems, solves the subproblems separately, and then combines the results into a final solution. This approach is primarily a hierarchical classification scheme represented by a decision tree.

To solve the subproblems in the decision tree, three types of classification methodologies such as a multilayer perceptron (MLP), a quadratic discriminant function (QDF), and a linear discriminant function (LDF) are used in a fixed sequence determined empirically. An MLP is called first to discriminate between Korean and English. Next, QDF classifiers for style and point size identifications as

well as a measurement of the word length are applied. These classifiers have been trained for a specific type of script since the script information is given from the previous classification with an MLP. For typeface identification, an LDF classifier trained for a specific combination of script and font style is used. Since there are two kinds of scripts and four kinds of font styles, our system has eight different typeface classifiers. The hierarchical classifier with MLP, QDF and LDF is shown in Fig. 1.

4. Feature Design

4.1. *Script identification*

The objective of script identification is to classify a word image into Korean or English. Two features are extracted from the image based on some typical differences between the two. In an English text, a character is composed of one connected component, except for "i" and "j". In Korean text, however, a character is composed of consonants and vowels, and therefore two or more connected components constitute a character.

Figure 2(a) shows an example. Here the Korean word contains four characters composed of 11 connected components. But these two numbers in the English word are equal to 5. In general, the number of connected components in a Korean word is greater than that of characters, while the two are almost the same in English words.

Based on this observation, the first feature is defined as

$$F_1^c = C/W_{\text{length}}, \text{ where } W_{\text{length}} = W_w/W_h\,.$$

Here C is the number of connected components, W_w and W_h are the width and height of word image in pixels, respectively. Here the value F_1^c of an English word depends on the presence of upper zone and/or lower zone. To measure the height of an English word precisely, we have to extract the exact location of baseline and meanline, but many research results report that this problem is not trivial. Fortunately, we have observed that the variance of W_H of English words does not affect the classification performance.

Since the connected components in Korean overlap each other vertically, the second feature is defined as follows — see Fig. 2(b).

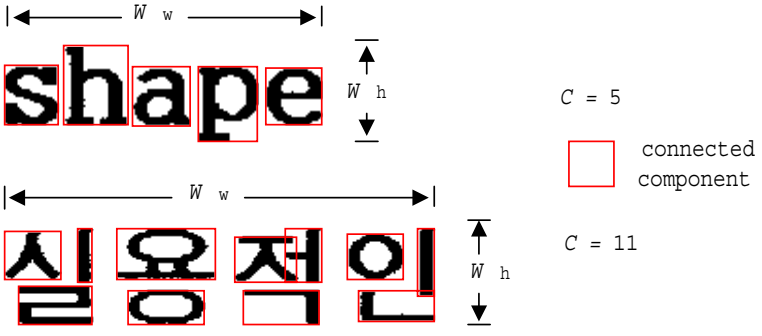
$$F_2^c = C_{\text{overlapped}}/W_{\text{length}}, \text{ where}$$

$$C_{\text{overlapped}} = \sum_i \sum_{j>i} f_s(\mathit{right}_{ij} - \mathit{left}_{ij})\,,$$

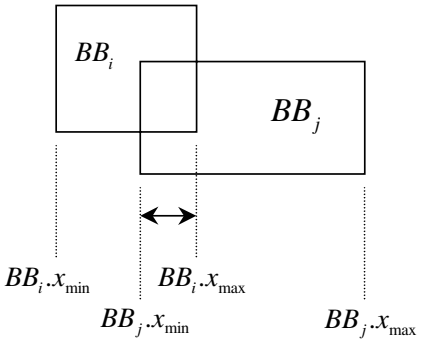
$$\mathit{right}_{ij} = \mathit{MIN}\{BB_i.x_{\max}, BB_j.x_{\max}\},\ \mathit{left}_{ij} = \mathit{MAX}\{BB_i.x_{\min}, BB_j.x_{\min}\}, \text{ and}$$

$$f_s(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

In the above equation, BB_i and BB_j are bounding boxes for ith and jth components, and $BB_i.x_{\max}$ and $BB_i.x_{\min}$ denote the rightmost and leftmost x-coordinates



(a)



(b)

Fig. 2. Feature extraction for script identification: (a) connected components for English and Korean words; (b) computation of the amount of overlapping.

of BB_i, respectively. In the case of Fig. 2(b), the following two equations hold:

$$right_{ij} = MIN\{BB_i.x_{\max}, BB_j.x_{\max}\} = BB_i.x_{\max}, \text{ and}$$

$$left_{ij} = MAX\{BB_i.x_{\min}, BB_j.x_{\min}\} = BB_j.x_{\min}\,.$$

So, $right_{ij} - left_{ij} = BB_i.x_{\max} - BB_j.x_{\min}$ (or the portion marked by $\leftrightarrow$ in the figure) is accumulated to compute $C_{\text{overlapped}}$. By the use of f_s, $(right_{ij} - left_{ij})$ is not accumulated for $C_{\text{overlapped}}$ when the value is less than 0. The limiting expression $(j > i)$ means that jth component is on the right side of ith component, in which the components are ordered from left to right by their leftmost x-coordinates.

4.2. *Font style identification*

Three features are extracted to deal with the three kinds of styles. The first feature is an average width of vertical strokes for bold style detection, the second is the maximum horizontal run length for detecting underline styles, and the third is the

maximum difference between the neighboring vertical projection profiles for italic styles.

To compute the average width of vertical strokes, an array $H[i]$ is constructed so that it stores the number of horizontal runs whose length is i. If we set most_freq as

$$\text{most\_freq} = \arg\max_i H[i]\,,$$

then this value represents the majority of run-lengths. Three values in $H[i]$ including $H[\text{most\_freq}]$ are chosen to compute the average width of vertical strokes as follows:

$$F_1^s = \frac{1}{N} \sum_{i=\text{most\_freq}-1}^{\text{most\_freq}+1} H[i] \times i, \text{ where } N = \sum_{i=\text{most\_freq}-1}^{\text{most\_freq}+1} H[i]\,.$$

The second feature, the maximum horizontal run-length, is computed from the length of longest run in the word image, i.e.

$$F_2^s = i_{\max}/W_w\,.$$

Here $i_{\max}$ is the length of longest run, and W_w is the width of the word image, respectively. The value of this feature for an underlined word is far larger than that of a normal word.

To obtain the third feature, a vertical projection profile is first constructed. Next the differences between neighboring values in the profile are computed. Figure 3 shows a vertical projection profile and the difference values for normal and italicized words. As can be seen from the figure, the maximum difference, $DVP_{\max}$, is an adequate feature for distinguishing italic words from normal ones:

$$F_3^s = DVP_{\max} = \max\{|VP[i] - VP[i+1]|\}\,,$$

where $VP[i]$ is the vertical projection profile at the ith column of the word image.

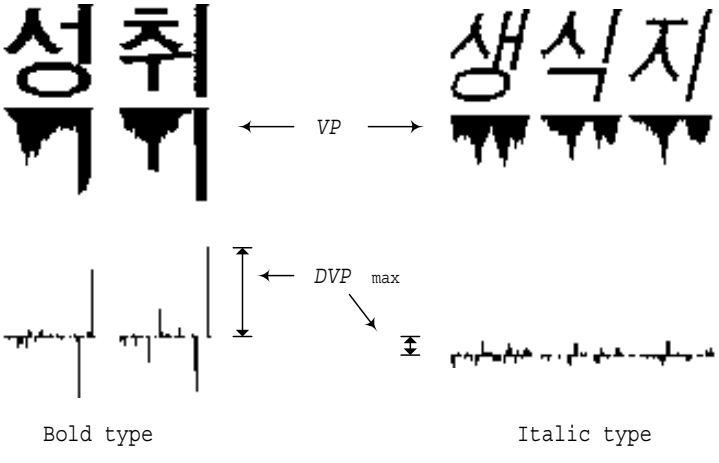


Fig. 3. Vertical projection profiles and their differences.

Fig. 4. Vertical distances of Korean and English words.

4.3. *Point size identification*

To identify the point size of a word image, an estimated value of vertical distance of the word is used. Characters in Korean words have almost the same height, but the heights of English characters vary considerably. Therefore, the vertical distance is estimated differently depending on the type of script — refer to Fig. 4.

In a Korean word, the vertical distance is computed as the height of the bounding rectangle. In an English word, on the other hand, the image is partitioned vertically into three zones such as upper, middle and lower zones by an analysis of peak positions in the horizontal projection profile, and the height of the middle zone is measured as the vertical distance. This is because some English words do not have upper or lower zone at all.

4.4. *Word length identification*

To identify the number of characters in a word, a language-dependent feature is also used. In the case of Korean, it is computed by a ratio between the width and height of the word image, since the length and breadth of each character is almost the same in a Korean text.

In the case of English words, on the other hand, the number of connected components is used to count the number of characters, since a connected component corresponds to a character in general. In counting the components in a word, small components in "i" or "j" are ignored. In addition, if the width of a connected component is bigger than its height, it is regarded as touching characters. Let the width and height of a connected component be C_w and C_h, respectively. If $C_w/C_h > 1$, the number of characters contained in the connected component is calculated as

$$\lfloor C_w/C_h + \alpha \rfloor ,$$

where α is an empirical constant, 0.7.

4.5. *Font typeface identification*

A remarkable distinction between Myung-jo and Gothic typefaces is the serif. Serif is a kind of decoration around the end of vertical strokes in a character. Figure 5 illustrates some examples of serifs in Myung-jo and Gothic typefaces for both Korean and English characters. As can be seen from the figure, the presence of serif is quite

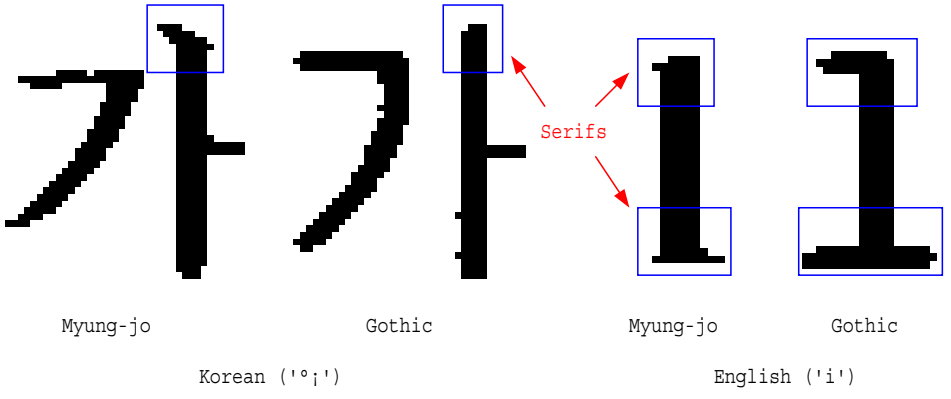


Fig. 5. Serifs of Myung-jo and Gothic typefaces in Korean and English characters.

different for Korean and English. It is also different from one font style to another. Therefore, the feature extraction for typeface identification is performed differently according to the script and font style.

In the case of Korean script, one feature value trained for a specific font style is used for typeface identification. Korean Myung-jo is a serif font that has a small decorative stroke at the end of vertical strokes, but Gothic is a sans-serif font which has no serif. Therefore, the serif part is first segmented from the vertical stroke, and the direction of the serif is computed as the feature for typeface identification.

Extraction of serif parts from a Korean character is performed by an analysis of its skeleton — refer to Fig. 6. Let a segment be a sequence of connected skeletal pixels, in which the degree, defined as the number of neighboring skeletal pixels, is two in every pixel except the starting and ending ones. Then the serif is defined as a set of the first five pixels in a vertical segment, p_0, p_1, p_2, p_3 and p_4, in which the degree of the starting pixel p_0 is one. Once a serif is detected, the four line segments, $\overline{p_0, p_1}$, $\overline{p_1, p_2}$, $\overline{p_2, p_3}$ and $\overline{p_3, p_4}$, are formed and the direction of each line is computed. Here the direction falls into one of the 36 sectors, and an average direction of the four lines is determined as the feature of serif. In fact, it is observed that the feature value in Myung-jo typeface is greater than 27 and that of Gothic is less than or equal to 27.

In an English script, on the other hand, both Myung-jo and Gothic have serifs. But the shapes of these serifs are different as can be seen from Fig. 5 — the serif in Myung-jo is thin and narrow but that of Gothic is thick and wide. Usually, one or more serifs exist around the top or bottom of a vertical stroke.

Extraction of the serif part is performed by utilizing two general characteristics of the serifs — refer to Fig. 7. Serif region of an English word is represented by a set of five runs. Search of a serif region starts from topmost run (or the most bottom run) in every connected component. Including this starting run, the algorithm constructs a set of five consecutive runs. The set is regarded as a serif region if it

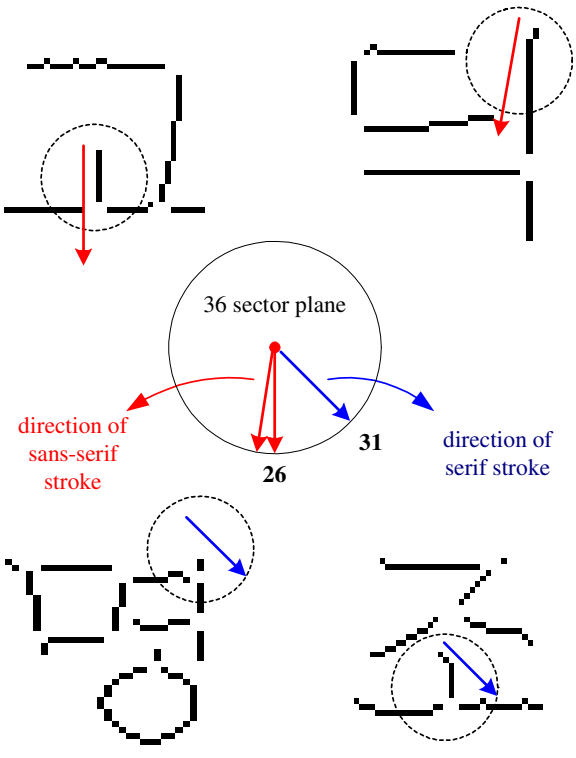


Fig. 6. The directions of serif and sans-serif strokes in a 36 sector plane.

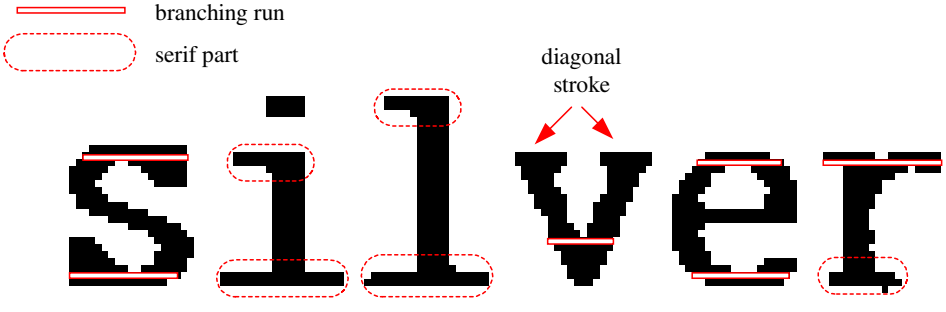


Fig. 7. Extraction of serif parts from English characters.

satisfies the two conditions: (1) there should be no branching run in the set, and (2) the five runs can be aligned vertically to form a part of vertical stroke.

Given an English word, two features are derived from all the serif parts within the word to classify the typeface. The first one is an average ratio between the length of second longest run and the length of most frequent run in every serif region, and the second one is an average of differences between the length of longest run and

that of second longest run in serifs. The first feature is computed as

$$F_1^t = \frac{1}{M}\sum_{j=1}^{M}(R^j_{\text{2nd\_max}}/\text{most\_freq\_width})\,,$$

where M is the number of serifs in the word image, and $R^j_{\text{2nd\_max}}$ is the length of second longest run in the jth serif, respectively. The second feature is computed similarly as

$$F_2^t = \frac{1}{M}\sum_{j=1}^{M}(R^j_{\text{max}} - R^j_{\text{2nd\_max}})$$

where M is the number of serifs in the word image, and R^j_{max} is the length of longest run in the jth serif, respectively.

5. Experiments

The proposed system for word-level font recognition has been evaluated with a database of 100 sets of word images of 264 different kinds of font combinations — 96 combinations for Korean and 168 combinations for English, respectively. The 26,400 words in the database have been prepared by a Korean word processor and scanned at 300 DPI (dots per inch) by a Sharp ScanJX scanner. Some word images in the database, along with their labels for script, style, size, length and typeface, are shown in Fig. 8. Fifty sets, or 13,200 word images, are used for testing, while the other 50 sets are used for training.

According to the hierarchical classification scheme, the script identification with an MLP is performed first, and the script-dependent classifications for size, style and word length are performed next. Finally the script- and style-dependent typeface recognition is performed. The recognition performances for individual font attributes are summarized in Fig. 9.

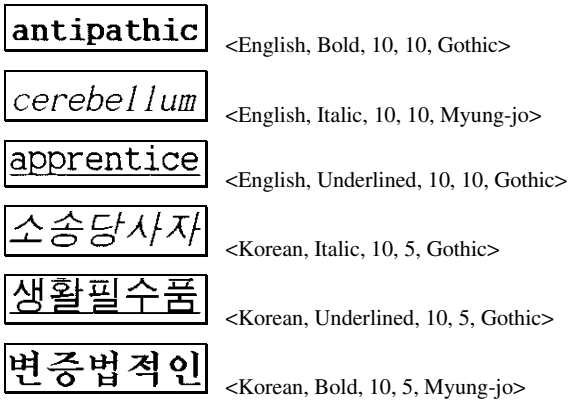


Fig. 8. Word images in the database.

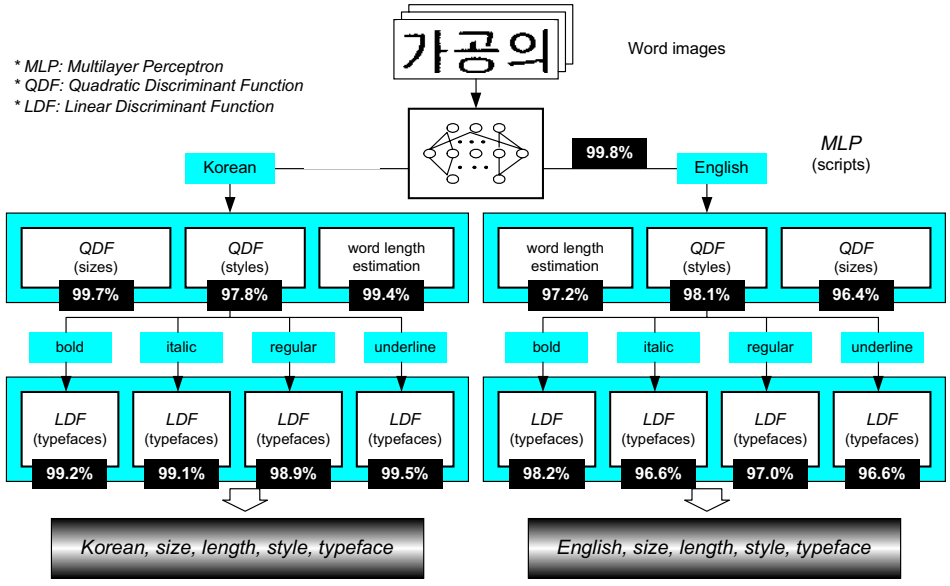


Fig. 9. The recognition results using the hierarchical classifier.

Table 3. Confusion matrix for script identification.

	Korean	English	Sum	Accuracy (%)
Korean	4,646	154	4,800	96.8
English	17	8,383	8,400	99.8
Total			13,200	98.7

5.1. *Script identification*

The accuracy of classifying all the 13,200 test images into Korean or English is 98.7% as a total — refer to Table 3. Errors in the classification of Korean scripts are due to the touching between consonants and vowels in Korean characters. The more the touching happens in a Korean word, the more the two values of F_1^c and F_2^c become similar to those of English words. However judging from the overall performance, the two features reflect well the significant difference between Korean and English words.

5.2. *Font style identification*

Once the script of an input word is identified, font style is classified next depending on the script. The confusion matrices for style recognition for Korean and English word images are shown in Tables 4 and 5, respectively. According to these tables, a nontrivial number of confusions occur between bold and regular styles as well as italic and regular styles. Confusion between bold and regular styles happens

Table 4. Confusion matrix for style classification of Korean words.

	Bold	Italic	Regular	Underline	Sum	Accuracy (%)
Bold	1,163		37		1,200	96.9
Italic		1,200			1,200	100
Regular	60	5	1,133	2	1,200	94.4
Underlined			1	1,199	1,200	99.9
Total					4,800	97.8

Table 5. Confusion matrix for style classification of English words.

	Bold	Italic	Regular	Underline	Sum	Accuracy (%)
Bold	2,060		39	1	2,100	98.1
Italic		2,020	78	2	2,100	96.2
Regular	13	26	2,061		2,100	98.1
Underlined				2,100	2,100	100
Total					8,400	98.1

Table 6. Confusion matrix for size classification of Korean words.

	10 pts	12 pts	14 pts	Sum	Accuracy (%)
10 pts	1,600			1,600	100
12 pts	8	1,592		1,600	99.5
14 pts	2	6	1,592	1,600	99.5
Total				4,800	99.7

when there is no vertical stroke in the input word and therefore the average width of vertical stroke F_1^s is not discriminative. Confusion between italic and regular styles, especially in English words, happens also when there is no vertical stroke in the word. If an English word consists of characters of curvilinear strokes, such as "a", "c", "e", "o", "s", etc., the maximum difference F_3^s has a low value both in regular and italicized words.

5.3. *Point size identification*

The recognition accuracies for point sizes are 99.7% for Korean and 96.4% for English words, respectively — refer to Tables 6 and 7. As shown in these tables, most misclassifications are in English words. This is due to the error in extracting the middle zone from an English word: the shapes of projection profile are quite complicated, and neither the upper nor the lower zone may exist in some words. Even worse, the peak position analysis makes more mistakes when the font size becomes smaller — one can see in Table 7 that the confusion between 14 and 10 pts (97 cases) is higher than that between 14 and 12 pts (57 cases).

Table 7. Confusion matrix for size classification of English words.

Size	10 pts	12 pts	14 pts	Sum	Accuracy (%)
10 pts	2,795	2	3	2,800	99.8
12 pts	128	2,658	14	2,800	94.9
14 pts	97	57	2,646	2,800	94.5
Total				8,400	96.4

Table 8. Recognition rates for typefaces of Korean words (%).

	Bold	Italic	Regular	Underlined	Average
Myung-jo	98.3	99.1	97.7	98.9	98.5
Gothic	100	99.1	100	100	99.8
Total	99.2	99.1	98.9	99.5	99.1

Table 9. Recognition rates for typefaces of English words (%).

	Bold	Italic	Regular	Underlined	Average
Myung-jo	99.1	97.3	98.2	98.0	98.1
Gothic	97.4	96.0	95.8	96.6	96.4
Total	98.2	96.6	97.0	97.3	97.25

5.4. *Word length identification*

The recognition accuracies for word length are 99.4% for Korean words (classification into 2, 3, 4, and 5) and 97.2% for English (classification into 7 classes from 4 to 10), respectively. In English words, the fragmentation of a character and touching of two or more characters affect significantly the feature value estimated from the number of connected components.

5.5. *Typeface identification*

Once the script and font style of a word image have been identified, the typeface is classified into Myung-jo and Gothic for the respective script and style combinations. The classification accuracies for the four styles in Korean words are shown in Table 8, and those for English words in Table 9. One can see from Table 8 that the difference between Myung-jo and Gothic for Korean words is well represented by the serif-based feature. In the case of English, on the contrary, the overall accuracy is lower than Korean. Most of the misclassifications are from Gothic words — refer to Table 9.

6. Conclusion

We have proposed an optical font recognition system. Differing from existing methods, the system can recognize word-level font attributes rather than line-level or block-level data. This property is useful in document indexing and retrieval as well as OCR applications. In addition, it extracts a large amount of font information, i.e. 264 font categories from five types of attributes which are quite common in Korean documents. All these distinguishing characteristics of the system have been obtained by combining a hierarchical classification scheme and several typographical features.

Although it is very difficult to compare the proposed system to existing methods, we can conclude that an encouraging accuracy has been produced. Our future work is to verify how the proposed system contributes to document indexing and retrieval, as well as OCR applications.

Acknowledgment

This work was supported by grant number R05-2003-000-10396-0 from the Program for Regional Scientists of the Korea Science and Engineering Foundation (KOSEF).

References

1. AIIM Conference Handbooks, *Association for Imaging and Information Methodologies* (AIIM International Co., 1996).
2. B. B. Chaudhuri and U. Garain, Automatic detection of italic, bold and all-capital words in document images, in *Proc. 14th Int. Conf. Pattern Recognition*, Brisbane, Australia, 1998, pp. 610–612.
3. J. Ding, L. Lam and C. Y. Suen, Classification of oriental and European scripts by using characteristic features, in *Proc. 4th Int. Conf. Document Analysis and Recognition*, Ulm, Germany, 1997, pp. 353–356.
4. D. Doermann, The indexing and retrieval of document images: A survey, *Comput. Vis. Imag. Underst.* **70** (1998) 287–298.
5. A. M. Elgammal and M. A. Ismail, Techniques for language identification for hybrid Arabic-English document images, in *Proc. 6th Int. Conf. Document Analysis and Recognition*, Seattle, USA, 2001, pp. 1100–1104.
6. U. Garain and B. B. Chaudhuri, Extraction of type style based meta-information from imaged documents, in *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, India, 1999, pp. 341–344.
7. J. Hochberg, L. Kerns, P. Kelly and T. Thomas, Automatic script identification from images using cluster-based templates, in *Proc. 3rd Int. Conf. Document Analysis and Recognition*, Montreal, Canada, 1995, pp. 378–381.
8. M. C. Jung, Y. C. Shin and S. N. Srihari, Multifont classification using typographical attributes, in *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, India, 1999, pp. 353–356.
9. I. S. Oh and S. H. Kim, Document image processing technologies and information retrieval, submitted to *14th Workshop on Image Processing and Understanding*, Cheju, Korea, January 2002.
10. U. Pal and B. B. Chaudhuri, Automatic identification of English, Chinese, Arabic,

Devnagari and Bangla script line, in *Proc. 6th Int. Conf. Document Analysis and Recognition*, Seattle, USA, 2001, pp. 790–794.
11. M. H. Park, Y. W. Shon, S. T. Kim and J. C. Namkung, The font recognition of printed Hangul documents, *Trans. Korea Inform. Process. Soc.* **4** (1997) 2017–2024.
12. *SCRIBBLE: SRI's Keyword Spotting System*, http://www.erg.sri.com/projects/scribble, SRI International Co. (1998).
13. A. L. Spitz, Determination of the script and language content of document images, *IEEE Trans. Patt. Anal. Mach. Intell.* **19** (1997) 235–245.
14. D. Xi, S. W. Lee and Y. Y. Tang, A novel method for discriminating between oriental and European languages by fractal features, in *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, India, 1999, pp. 345–348.
15. Y. Zhu, T. Tan and Y. Wang, Font recognition based on global texture analysis, *Ibid*, 1999, pp. 349–352.
16. A. Zramdini and R. Ingold, *ApOFIS*: an a priori optical font identification system, in *Proc. 8th Int. Conf. Image Analysis and Processing*, Sanremo, Italy, 1995.
17. H. S. Baird and G. Nagy, A self-correction 100-font classifier, in *Proc. SPIE Conf. Document Recognition*, 1994, pp. 106–115.
18. D. S. Bloomberg, Multiresolution morphology analysis of document images, in *Proc. SPIE Conf. 1818, Visual Communications and Image Processing*, 1992, pp. 648–662.
19. D. Doermann, A. Rosenfeld and E. Rivlin, The function of documents, in *Proc. Int. Conf. Document Analysis and Recognition*, Vol. 2, Germany, 1997, pp. 1077–1081.

Soo Hyung Kim received his B.S. degree in computer engineering from Seoul National University in 1986, and his M.S. and Ph.D degrees in computer science from Korea Advanced Institute of Science and Technology in 1988 and 1993, respectively. From 1990 to 1996, he was a senior member of research staff in Multimedia Research Center of Samsung Electronics Co., Korea. Since 1997, he has been a professor in the Department of Computer Science, Chonnam National University, Korea.

His research interests are in handwritten character recognition, document image processing and neural network modeling.

Hee Kue Kwag received the B.S., M.S. and Ph.D. degrees in computer science from Chonnam National University, Korea, in 1996, 1998 and 2001, respectively. Currently, he is a chief researcher with DongBang SnC Co., Ltd., Korea.

Ching Y. Suen received an M.Sc.(Eng.) degree from the University of Hong Kong and a Ph.D. degree from the University of British Columbia, Canada. In 1972, he joined the Department of Computer Science of Concordia University where he became Professor in 1979 and served as Chairman from 1980 to 1984, and as Associate Dean for Research of the Faculty of Engineering and Computer Science from 1993 to 1997. He has guided/hosted 65 visiting scientists and professors, and supervised 60 doctoral and master's graduates. Currently he holds the distinguished Concordia Research Chair of Artificial Intelligence and Pattern Recognition, and is the Director of CENPARMI, the Centre for PR & MI.

Prof. Suen is the author/editor of 12 books and more than 300 papers on subjects ranging from computer vision and handwriting recognition, to expert systems and computational linguistics. He is the founder and Editor-in-Chief of a journal and an Associate Editor of several journals related to pattern recognition.

A Fellow of the IEEE, IAPR, and the Academy of Sciences of the Royal Society of Canada, he has served several professional societies as President, Vice-President, or Governor. He is also the founder and chair of several conference series including ICDAR, IWFHR, and VI. He was the General Chair of numerous international conferences, including the International Conference on Document Analysis and Recognition held in Montreal in August 1995 and the International Conference on Pattern Recognition held in Quebec City in August 2002.

Dr. Suen is the recipient of numerous awards, including the ITAC/NSERC Award (Information Technology Association of Canada and the Natural Sciences and Engineering Research Council of Canada) in 1992 and the Concordia "Research Fellow" award in 1998.