Optical Font Recognition For Printed Korean Characters Using Serif Pattern of Strokes

Soo-Hyung Kim[1], Sam-Soo Kim[1], Hee-Kue Kwag[2], Guee-Sang Lee[1]

[1]Department of Computer Science, Chonnam National University, Korea
Tel : +82-62-530-0430, Fax : +82-62-530-3439
e-mail : {shkim,sskim}@iip.chonnam.ac.kr, gslee@chonnam.ac.kr
[2]Department of Electronic Engineering and Computer Science, KAIST
e-mail : hkkwag@ai.kaist.ac.kr

Abstract: This paper introduces the problem of typeface classification of Hangul characters and proposes features for typeface classification among Serif and Sans-serif classes. Serif classes have a small decorative stroke around the beginning of vertical strokes, while Sans-serif classes have no serif. Therefore, the serif part is first segmented from the vertical strokes, and the direction of the serif is computed as the feature for Hangul typeface identification. To evaluate the performance of the proposed system, we used 3,000 characters extracted from Korean documents – 1,500 from Serif fonts, other 1,500 from Sans-serif fonts.

1. Introduction

Today a lot of information has to be acquired from various kinds of printed documents, such as newspapers, magazines, books, forms, journals, technical reports, and so on. The amount of these paper-based documents is increasing day by day in spite of 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 volumes of printed documents. [3-19]

Two approaches for document indexing and retrieval have been developed. 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, the two approaches actually complement to each other and a hybrid approach is being developed as an alternative.

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 on Korean documents. Assuming that the document image is decomposed into characters, the system extracts some typographical features from the character strokes and then identifies two types of font typefaces – Serif and Sans-Serif fonts. These two fonts are the most popular font classes in Korean word processors, and further classifications can be performed after a rough classification by the proposed system. Figure 1 shows the overall diagram of proposed approach.

Beside the improvement in document indexing and retrieval technologies, optical font recognition has valuable applications:

- Improvement of OCR perfoemance using the font information of character [4, 5, 6, 7, 8, 9, 13]
- Generation of text summarization in the field of information retrieval using special attributes [3, 10, 11]
- Document structure analysis by the use of contextual information [3, 12]
- Document reproduction [14]

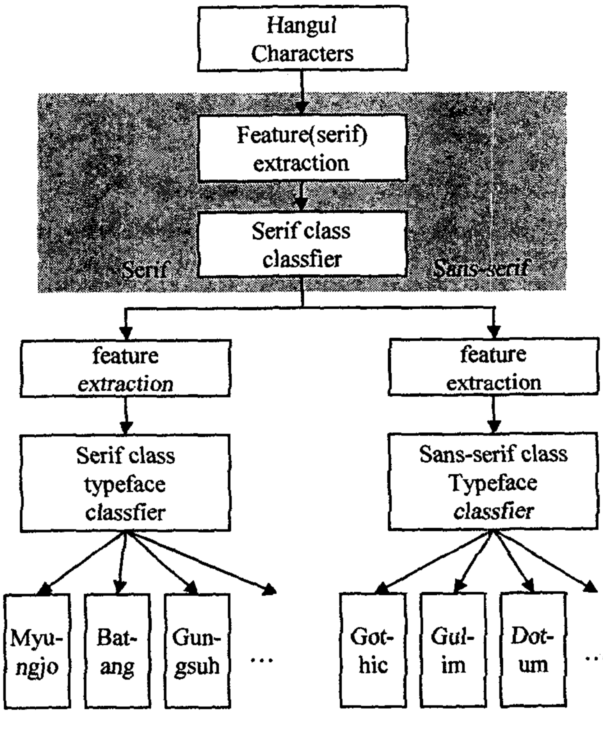

Figure 1. Overall structure of font recognition system

2. Proposed System

A remarkable distinction between Serif fonts (Myungjo, Batang, Gungsuh typefaces) and Sans-Serif fonts(Gothic, Gulim, Dotum typefaces) is in the serif. Serif is a kind of decoration around the end of vertical strokes in a character. Figure 2 illustrates some examples of serifs in Myung-jo typeface.

2.1 Serif Region Extraction

Serif fonts have a small decorative stroke at the beginning of vertical strokes, but Sans-Serif fonts have 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.

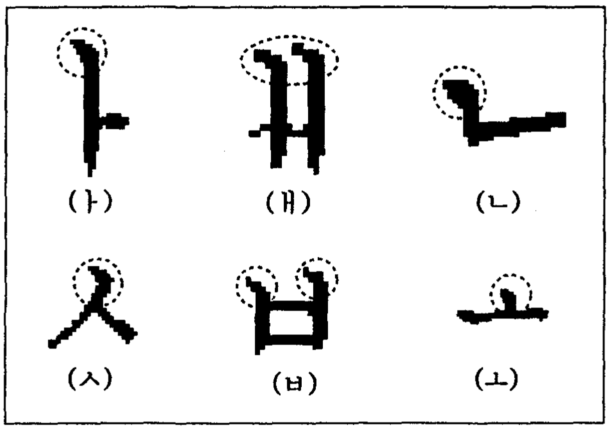

Figure 2. Various serif type of Hangul

First a thinning operation using vertical and horizontal run analysis is applied on the character image, and then segments for vertical (or vertical) strokes are gained by horizontal (or vertical) run analysis. Serif region exists around the beginning part of a vertical stroke segment which does not meet with any horizontal stroke segment. Figure 3 shows the extraction of serif region, where the left image is an input Hangul character and the right one is a thinned version of the character. When we consider two vertical segments s_1, s_2 in Figure 3, the beginning part of s_2 meets with a horizontal segment, but that of s_1 does not contacts any horizontal stroke. Therefore, we extract s_1 as a serif region.

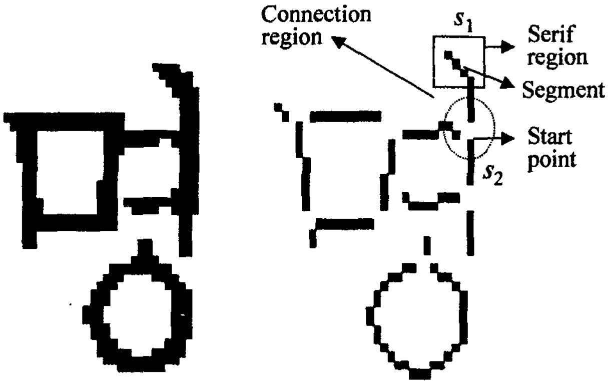

Figure 3. Extraction image for Hangul in Serif region.

2.2 Direction Vector of Serif

Extraction of serif parts from a Korea character is performed by an analysis of its skeleton - refer to Figure 4. Let a segment be a sequence of connected skeletal pixels, whose 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, P_0P_1, P_1P_2, P_2P_3 and P_3P_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.

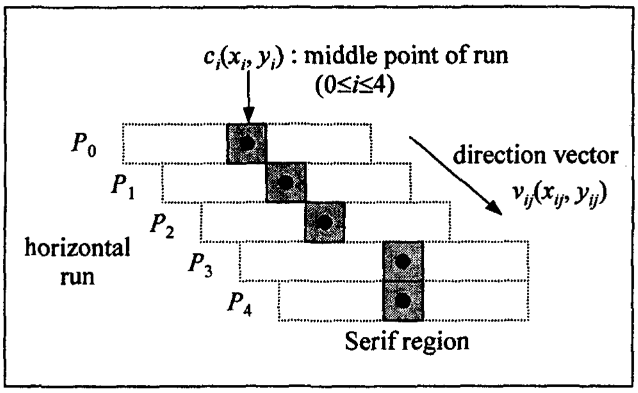

Figure 4. Extracted serif region of run

Direction vector v_{ij}, defined as the vector from *i*-th run to *j*-th run, is computed as following:

$$v_{ij} = (x_j - x_i, y_j - y_i), (j > i).$$

Assume that D_{ij} is the position in which is direction vector v_{ij} falls in within a 36 sector plane, the final feature *D* of the serif region is computed as an average:

$$D = (\sum_{i=0}^{3} D_{ij}) / (N-1), (j = i+1)$$

where *N* is the number of runs in serif region.

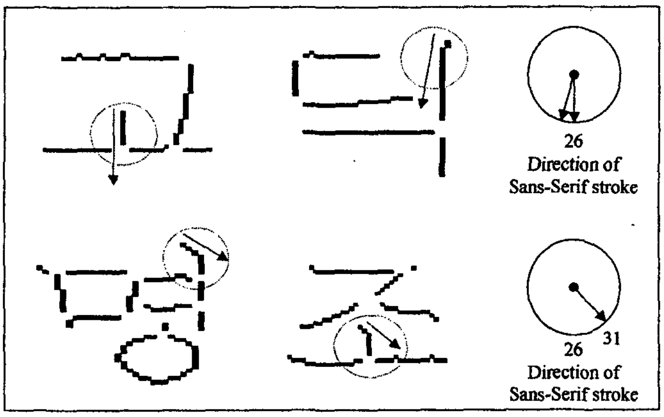

Figure 5. The direction of Serif and Sans-serif fonts in a 36 sector plane

Figure 5 illustrates some examples of serifs in Myung-jo and Gothic typefaces for Korean characters. As can be seen from the figure, the presence of serif is quite different. 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 a Serif font is greater than 27 and that of Sans-serif font is less than or equal to 27.

3. Experimental Result

To evaluate the performance of the proposed system, we used 3,000 characters extracted from Korean documents – 1,500 from Serif fonts, other 1,500 from Sans-serif fonts. Table 1 and Figure 6 show the distribution of serif strokes for 6 different font typefaces – three of them are Serif fonts, and the other three are Sans-serif fonts, respectively. It is obvious from the figure that Serif and Sans-serif fonts have quite different distributions in the feature space of stroke directions. This fact proves that the proposed feature for Serif and Sans-serif classification is effective and can be used for a optical font recognition successfully.

Table 1. The direction vector distribution of Serif regions

D	Gulim	Dotum	Gothic	Batang	Gung -suh	Myung -jo
20	8	0	0	0	0	0
21	3	0	0	0	0	
22	2	0	0	0	0	0
23	40	75	3	0	0	0
24	0	1	18	0	0	0
25	0	0	199	0	0	0
26	447	424	275	0	0	4
27	0	0	5	0	0	19
28	0	0	0	33	1	68
29	0	0	0	180	43	191
30	0	0	0	120	263	174
31	0	0	0	118	81	44
32	0	0	0	20	73	0
33	0	0	0	29	38	0
34	0	0	0	0	1	0
35	0	0	0	0	0	0

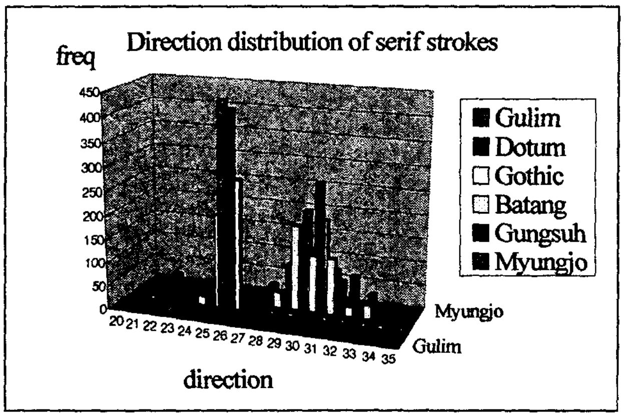

Figure 6. Distribution of Serif stroke direction

As can be seen from the diagram in Figure 6, distribution of Serif (Batang, Gungsuh, Myungjo) fonts are more sparse than that of Sans-serif ones. Most of the errors has been observed from the characters which have no vertical strokes ('으', '고', etc) .

4. Conclusion

This paper introduces the problem of typeface classification of Hangul characters and proposes features for the classification among Serif (Myungjo, Batang, Gungsuh) and Sans-serif(Gothic, Gulim, Dotum). The proposed method extracts serif regions around the beginning of the vertical strokes and calculates the direction of the serif strokes within a 36 sector plane. We have proved the effectiveness of the proposed feature with 3,000 character patterns from Korean documents.

One of further studies is to recognize various fonts within each category of Serif and Sans-serif and to construct a complete typeface classification system of Hangul characters.

Acknowledgement

This work was supported by Brain Korea 21 Project.

References

[1] AIIM'96 Conference Handbooks, Association for Imaging and Information Methodologies, 1996.

[2] D. Doermann, "The Indexing and Retrieval of Document Images: A Survey," *Computer Vision and Image Understanding*, Vol. 70, No. 3, pp. 287-298, 1998.

[3] U. Garain and B.B. Chaudhuri, "Extraction of Type Style Based Meta-Information from Imaged Documents," *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, pp. 341-344, 1999.

[4] H.S. Baird, G. Nagy, "A Self-Correcting 100 Font Classifier," *Proc. of SPIE Conference on Document Recognition*, pp. 106-115, 1994.

[5] A. Zramdini, *Study of optical font recognition based on global typographical features,* PhD thesis, University of Fribourg, 1995.

[6] S. Kahan, T. Pavlidis and H.S. Baird, "On the Recognition of Printed Characters of Any Font and Size," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 9, No. 2, pp. 274-288, 1987.

[7] M.C. Jung, Y.C. Shin and S.N. Srihari, "Multifont Classification Using Typographical Attributes," *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, pp. 353-356, 1999.

[8] T.K. Ho, J.J. Hull and S.N. Srihari, "A Computational Model for Recognition of Multi-Font Images," *Machine Vision and Applications*, Vol. 5, No. 1, pp. 157-168, 1992.

[9] S. Zhao and S.N.. Srihari, "A Word Recognition Algorithm for Machine-Printed Word Images of

Multiple Fonts and Varying Qualities," *Proc. 3rd Int. Conf. Document Analysis and Recognition*, Montreal, pp. 351-354, 1995.

[10] B.B. Chaudhuri and U. Garain, "Automatic Detection of Italic, Bold and All-Capital Words in Document Images," *Proc. 14th Int. Conf. Pattern Recognition*, Brisbane, pp. 610-612, 1998.

[11] T.K. Ho, "Font Identification of Stop Words for Font Learning and Keyword Spotting", *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, pp. 333-336, 1999.

[12] Y. Zhu, T. Tan and Y. Wang, "Font Recognition Based on Global Texture Analysis," *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, pp. 349-352, 1999.

[13] H. Shi, T. Pavlidis, "Font Recognition and Contextual Processing for More Accurate Text Recognition," *Proc. 4th Int. Conf. Document Analysis and Recognition*, Ulm, pp. 39-44, 1997.

[14] M.H. Park, Y.W. Shon, S.T. Kim and J.C. Namkung, "The Font Recognition of Printed Hangul Documents," *Trans. Korea Information Processing Society*, Vol. 4, No. 8, pp. 2017-2024, 1997.

[15] H.K. Kwag, *A Study on word Segmentation and Attribute Extraction from Document Images*, Ph.D. Thesis, Chonnam National University, 2001.

[16] J. Hochberg, L. Kerns, P. Kelly, and T. Thomas, "Automatic Script Identification from Images Using Cluster-based Templates," *Proc. 3rd Int. Conf. Document Analysis and Recognition*, Montreal, pp. 378-381, 1995.

[17] J. Ding, L. Lam, and C.Y. Suen, "Classification of Oriental and European Scripts by Using Characteristic Features," *Proc. 4th Int. Conf. Document Analysis and Recognition*, Ulm, pp. 353-356, 1997.

[18] A.L. Spitz, "Determination of the Script and Language Content of Document Images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19, No. 3, pp. 235-245, 1997.

[19] D. Xi, S.W. Lee and Y.Y. Tang, "A Novel Method for Discriminating Between Oriental and European Languages by Fractal Features," *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, pp. 345-348, 1999.