

ANALYSIS SYSTEM OF MIXED DOCUMENTS CONSISTING OF HANDWRITTEN KOREAN/ALPHANUMERIC TEXTS AND GRAPHIC IMAGES

Chang Bum Lee*, Dong-Gyu Sim**, In Kwon Kim**, Young Kug Ham**, and Rae-Hong Park**

*Communication Processing Technology Section, ETRI, Yusong-gu Daejeon 305-350, Korea

**Dept. of Electronic Engineering, Sogang University, C. P. O. Box 1142, Seoul 100-611, Korea

ABSTRACT

This paper proposes an effective recognition system which recognizes the mixed document consisting of handwritten Korean/alphanumeric texts and graphic images. In the preprocessing step, an input image is binarized by the proposed thresholding scheme, then graphic and character regions are separated by using chain codes of connected components. In the character recognition step, to recognize Korean characters, we use the branch and bound algorithm based on DP matching. Also we recognize alphanumeric characters using several robust features. Finally, to validate recognition results, we use a dictionary and knowledge employed in a recognition step. Computer simulation with several test documents shows that the proposed algorithm recognizes effectively handwritten mixed texts.

INTRODUCTION

Data processing by computers has expanded its influence on every part of modern information society. It overcomes human limitations in several respects, e.g., in speed and performance of accumulation, retrieval, and management of information. It has taken an honorable part in leading modern information society, however, it has drawbacks. The computer cannot come up to man's capability in various applications such as information recognition, analysis, input/output data processing, and so on.

Generally, a mixed document contains graphics as well as text. In addition, in Korea we commonly deal with mixed texts consisting of different sets of characters, e.g., Korean characters, alphanumeric characters, and Chinese characters. Many papers on handwritten character recognition have been published, but most of them have dealt with only one set of characters.⁽¹⁾

This paper proposes an effective recognition system which recognizes the mixed document consisting of handwritten Korean/alphanumeric texts and graphic images. In the preprocessing step, an input image is binarized by the proposed thresholding scheme, then graphic and character regions are separated by using chain codes of connected components. Separated Korean characters are merged depending on partial recognition results based on their character types and sizes. To recognize Korean characters, we use the branch and

bound algorithm based on DP matching costs. Also we recognize alphanumeric characters using several robust features. Finally, to correct wrong recognition results, we use a dictionary and knowledge employed in a recognition step.

PREPROCESSING

The proposed algorithm consists of the preprocessing and recognition steps. In the preprocessing step, each character is extracted from an input document image. First, an input image is binarized by using an adaptive thresholding scheme. Then, graphic regions and isolated characters are extracted.

1. Binarization

Although binarization is important in the preprocessing, the common scheme supported by commercial scanners is far from satisfactory: merging and distortion arise by the simple binarization scheme employed.

In this paper, to find the valley of the gray level histogram of an image, we propose the binarization algorithm based on the water flow model. By analogy, an intensity image is regarded as a terrain. Pools or deep valleys are considered as the character regions, and other regions are considered as background regions. After it rains over the terrain, rain flows into the lower terrain. Then water in the shallow pool is evaporated. In this paper, using the water flow model, we apply a simple rule to binarization of an input document.

2. Extraction of graphic regions

In the binarized document, graphic regions must be extracted. Based on the size and geometry of the connected components, these regions are extracted. 8-directional chain codes of connected components are employed to track only edges (boundaries) of connected components. The size of the minimum bounding rectangle (MBR) encompassing the connected components is used to extract graphic regions, in which statistics of the size of connected components is used and the number of character regions in a document are assumed to be more than that of graphic regions.

3. Extraction of isolated characters

In the document with graphic regions separated, isolated characters are extracted by chain codes. In general, a character consists of a few connected components. In extracting isolated characters, these connected components have to be grouped based on relationship between them, their geometric forms and partial recognition results.

In Korean character recognition, if two connected components are positioned along the vertical direction, they are merged. To merge a vertical vowel, we use not only the positional relationship between phonemes but also the partial recognition results of a vertical vowel in order not to confuse the type of a character set: e.g., alphanumeric character or Korean vertical vowel.

4. Thinning

Because the proposed recognition algorithm is based on the stroke analysis, a thinning algorithm is needed. In the proposed character recognition system, each separated character is skeletonized by the safe-point thinning algorithm (SPTA).^[2] This algorithm does not deteriorate the geometric structure of strokes, and preserves the connectivity of strokes.

PROPOSED CHARACTER RECOGNITION

In this paper, we propose an efficient algorithm that analyzes the mixed document consisting of the handwritten Korean/alphanumeric texts and graphic images. As shown in Fig. 1, the proposed document analysis system for two different character sets is proposed based on the graph search algorithm minimizing the DP matching costs. Each stage of an algorithm is described briefly as follows.

1. Classification of different character sets

Because the mixed document considered in this paper consists of two different sets of characters, e.g., Korean and alphanumeric characters, we adopt suitable recognition algorithms for each character set. Thus the determination whether an input character is a Korean or alphanumeric character is important and required.

Using several features such as the numbers of end and branch points, we determine whether the separated character is a Korean or alphanumeric character. Except for 'i' and 'j', alphanumeric characters consist of a single connected component with the number of end points n_e less than 5. But due to distortion or variation of writing style, the redundant end points can be generated. In case of Korean characters, a character consisting of one connected component with n_e less than 3 does not exist. Thus this character is passed into an alphanumeric character recognition routine. If a character consists of a single connected component with n_e larger than 2, first of all, it is passed into the alphanumeric character recognition routine: if it is not recognized, then it is passed into a Korean character recognition routine. Otherwise, it is passed into a Korean character

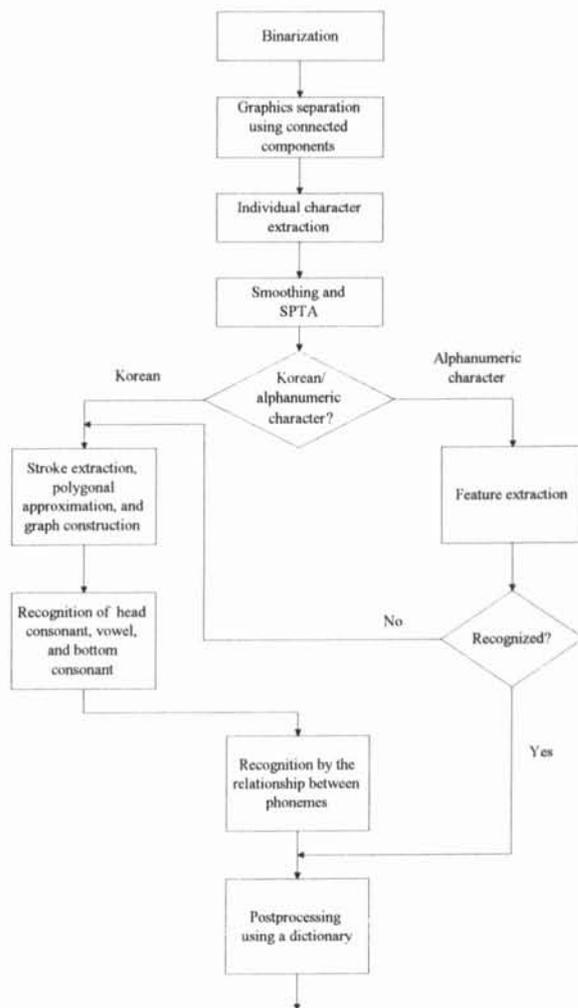


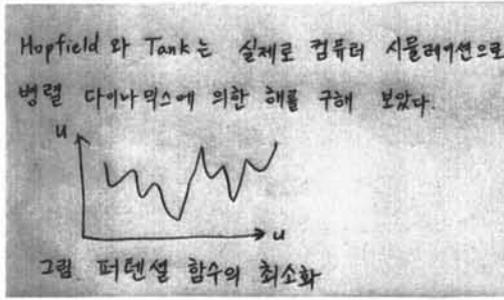
Fig. 1. Flowchart of the proposed document analysis system.

recognition routine. If an alphanumeric character other than 'i' and 'j' is split due to distortion, this algorithm can not recognize it. 'i' and 'j' are classified by the ratio of width to height of a character.

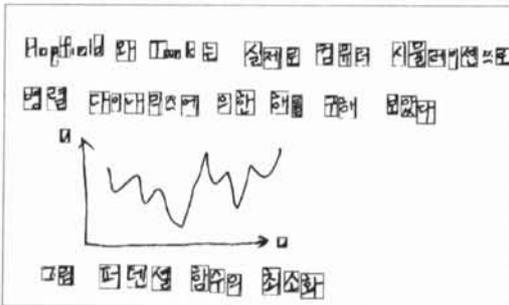
2. Recognition of Korean characters

In case of a Korean character, segments are extracted from character strokes after thinning. At first, the segment is defined as a set of black points between end, cross, or break points. Next, from these segments straight segments are generated by using polygonal approximation.^[3] For effective matching, a graph is constructed with these straight segments according to their positional relationships.

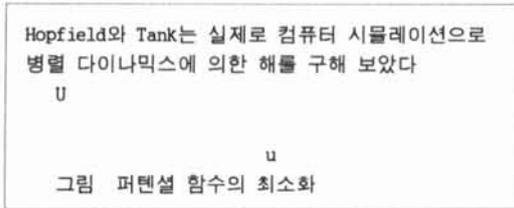
In Korean characters, structural analysis is important for the proposed structural algorithm. Korean characters consist structurally of a few phonemes, and each phoneme consists of primitive strokes and phonemes. The proposed algorithm uses attributes of segments and relationship between segments. As attributes, length, linearity, and direction of segments are adopted. As



(a) Input image.



(b) Binarization and character isolation.



(c) Recognition result.

Fig. 3. Simulation results for a handwritten mixed document 1 containing a graphic image.

CONCLUSION

We develop an efficient analysis system for mixed documents consisting of handwritten Korean/alphanumeric texts and graphic images. In this paper, we present the preprocessing algorithm that can effectively separate and extract texts from graphic images. The Korean character can be recognized by using a graph search algorithm based on DP matching. Also the syntactic method based on the phoneme matching results and positional relationships between phonemes are employed. In addition, the performance is further improved by using a dictionary. Computer simulations with several test documents show that the proposed document analysis system recognizes effectively handwritten texts as well as printed ones. Future work will focus on development of the effective postprocessing scheme employing high-level knowledge such as syntax analysis and semantics.

REFERENCES

1. T. Agui, M. Nakajima, T. K. Kim, and E. T. Takahashi, "A method of recognition and representation of Korean characters by tree grammars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 3, pp. 245-251, July 1979.
2. N. J. Naccache and R. Shinghal, "SPTA: A proposed algorithm for thinning binary patterns," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-19, no. 2, pp. 409-418, May/June 1984.
3. K. Wall and P.-E. Danielsson, "A fast sequential method for polygonal approximation of digitized curves," *Computer Vision, Graphics, Image Processing*, vol. 28, no. 2, pp. 220-227, Nov. 1984.
4. H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, no. 1, pp. 43-49, Feb. 1978.
5. C.-K. Lin, K.-C. Fan, and F. T.-P. Lee, "On-line recognition by deviation-expansion model and dynamic programming matching," *Pattern Recognition*, vol. 26, no. 2, pp. 259-268, Feb. 1993.
6. D.-G. Sim, Y. K. Ham, and R.-H. Park, "On-line recognition of cursive Korean characters using DP matching and fuzzy concept," in press, *Pattern Recognition*.
7. Y. K. Ham, H. K. Chung, I. K. Kim, R.-H. Park, C. B. Lee, S. J. Kim, and B. N. Yoon, "Hierarchical recognition of mixed documents consisting of the Korean/alphanumeric texts and graphic images," in *Proc. MVA'92 IAPR Workshop*, pp. 287-290, Tokyo, Japan, 1992.
8. Y. K. Ham, H. K. Chung, and R.-H. Park, "Automated analysis of mixed documents consisting of printed Korean/alphanumeric texts and graphic images," *Optical Engineering*, vol. 33, no. 6, pp. 1845-1853, June 1994.
9. Y. K. Ham, C. B. Lee, W. S. Kim, S. Y. Doh, R.-H. Park, and S. J. Kim, "A simple sequentially designed rule-based alphanumerics recognition for OCR document processing using a thinning process," in *Proc. SPIE Intelligent Robots Computer Vision X: Algorithms and Techniques*, vol. 1607, pp. 146-157, Boston, Mass., Nov. 1991.