MVA '92 IAPR Workshop on Machine Vision Applications Dec. 7-9,1992, Tokyo

Hierarchical Recognition of Mixed Documents Consisting of the Korean/Alphanumeric Texts and Graphic Images

Young Kug Ham, Hong Kyu Chung, In Kwon Kim, Rae-Hong Park
Chang Bum Lee*, Sang Joong Kim*, and Byeong Nam Yoon*

Dept. of Electronic Eng., Sogang Univ., C.P.O. Box 1142, Seoul 100-611, Korea
* Human Interface Section, Electronics and Telecommunications Research Institute
P.O. Box 8, Daeduk Science Town, Daejeon 305-606, Korea

ABSTRACT

In this paper, we propose an efficient algorithm which recognizes the mixed document consisting of the Korean/alphanumeric texts and graphic images.

In the preprocessing step, we separate graphic image parts from the text parts by considering chain codes of connected components. In the recognition step of the Korean/alphanumeric characters, we recognize the characters hierarchically using several features such as end points, branch points, cross points, partial projections, and distance features.

Computer simulation shows that the proposed algorithm recognizes the mixed document effectively.

INTRODUCTION

Data processing by computers has been expanded its influences on every part in modern society. It has disadvantages even though it has taken honorable part in leading the modern information society. The computer system makes it possible to overcome men's ability in several aspects such as fast processing, accumulation, retrieval, and management of information. On the other hand, the computer falls short men's ability in these points: information recognition, analysis, and input/output data processing.

To improve the efficiency of an input/output data processing part, computer vision which tries to implement the human visual capability to a man-made machine has been developed as an attempt to solve these problems. As one of the major areas of computer vision, character recognition consists of pattern representation part, and the decision part based on the presented pattern description.

Along with the recognition of Korean characters, their structural complexity and diverse font variations have been the topics of research. Many different techniques to solve these problems have been proposed for the past decades[1-3].

Recognition of Korean characters is based on the efficient segmentation of fundamental alphabetic symbols from a composite character pattern. Thus the efficiency of a recognition method of Korean characters depends largely on the reliable partitioning of characters into the subpatterns.

In this paper, we propose an efficient algorithm which recognizes the mixed document consisting of the Korean/alphanumeric texts and graphic images. The proposed system consists of three parts: prepocessing, feature extraction, and character recognition steps.

PREPROCESSING

The binary image sometimes shows the random noisy dots and blobs, which are independent of the input data. In the preprocessing step, the isolated noisy spots are removed and the input document is aligned, if necessary, by rotating it. We obtain the rotation angle using the Hough transform and align the input document horizontally.

From the input data, we segment graphic image parts and text parts according to the positions of lines and words on which the characters exist, by considering chain codes of connected components[4]. We further separate each character through a top-down algorithm and pass it to the next character recognition step. The separated graphic parts and the recognized characters are to be transmitted to generate an output file. In this section, we describe the preprocessing step briefly.

1. Separation of graphic parts

In this paper, we assume that the graphic region is larger than that of a single character. Therefore if the region of the connected components obtained using their chain codes is larger than that of a character, it is separated into the graphic region. The size of a character is determined by the projection values along the horizontal direction.

2. Extraction of characters

From documents classified into the text part, the individual character is extracted by using the horizontal and vertical projections. The broken character is to be combined with the adjacent character whereas the touched characters are to be divided, based on the projection value of each character and the character size predetermined. The length and width of a character are determined by the horizontal and vertical projections, respectively. For the mixed document consisting of Korean characters and alphanumerics, two alphanumerics may be extracted together since the size of the Korean character is assumed to be larger than that of alphanumerics. In the main processing, each block is classified into the Korean character or alphanumeric character based on the partial projection and distance feature, and then each character is recognized hierarchically. Also a blank and new line are recognized in the text parts. Special characters such as commas, periods, and quotation marks are recognized in the preprocessing part.

CHARACTER RECOGNITION

English is expressed by a sequence of basic alphabets, while Korean characters are generated by the structural composition of basic alphabets, i.e. 14 consonants and 10 vowels. The positions of each consonant and vowel in a composite character are determined definitely in accordance with the six structural composition rules[1]. In this paper, we first separate effectively the Korean characters from the alphanumeric characters, in which a partial projection and a distance measure are used. We assume that the size of the Korean character is twice the size of alphanumerics. In this section, the character recognition step is briefly described.

1. Feature extraction

The character may often be classified incorrectly, if the feature points are extracted wrong. To overcome the drawbacks of the thinning process, some other features such as the partial projection, cross points, and a distance measure are used. Several features used are briefly described.

(1) End point and branch point

The extracted individual character is thinned for the feature extraction[5]. At this time, small noise spurs are removed, then end points and branch points are extracted from the thinned characters.

The algorithm described in this paper uses the segments and positions of end points and branch points, as the basic features to recognize Korean or alphanumeric characters.

(2) Cross point[6]

A number of horizontal or vertical cross points (at which gray level changes from white to black or vice versa horizontally or vertically) are used to classify characters.

(3) Partial projection[6]

The partial projection is used to classify and extract vowels in the Korean characters. We first project vertically on the left and right sides, and then the ratio of the projection value to the character height is used. Similarly, we project horizontally in the top and bottom sides, and then the ratio of the horizontal projection value to the character width is used.

(4) Distance feature[6]

The distance feature is defined by the distance from the enclosed rectangular window to the black points of the character at several locations.

2. Recognition of Korean characters

(1) Vertical vowel

In a Korean character, all the long vertical vowels locate in the right. For the extraction of vertical vowels, the partial projection is used to test the existence of a vertical vowel. At this time, to remove the effect of the last consonants, the vertical projection without the last consonant is used. Then using the distance feature and cross points, we recognize the ' ㅏ' vowel class and ' ㅣ' vowel class.

(2) Horizontal vowel

Long horizontal vowels are divided into two classes. The bottom horizontal vowels are first identified, because the horizontal consonants may be taken for the horizontal vowels. For example, though The Korean character such as '표' does not have the central horizontal vowel, the part of a consonant may be taken for the central horizontal vowel incorrectly. Secondly, the central horizontal vowels are recognized.

(3) Consonant

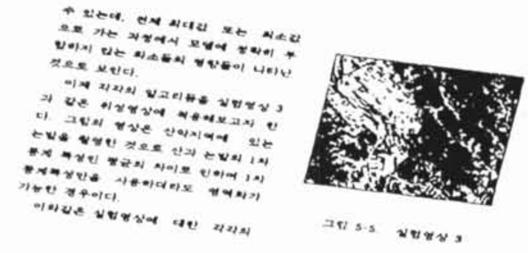
In this paper, we recognize first and last consonants separately according to the vowel class. By the distance measure, the existence of the last consonant is tested.

3. Recognition of alphanumeric characters

We also use the segments and positions of end points and branch points as the basic features to recognize alphanumerics. In the broken characters, the thinning method usually suffers the drawback of wrong feature extraction. The characters are often classified incorrectly since many feature points are extracted wrong. To overcome this problem, we add some features such as the cross point and the distance measure.

EXPERIMENTAL RESULTS AND DISCUSSIONS

Fig. 1 shows the skew input document and Fig. 2 shows the result of rotation by the angle detected by the Hough transform method. Fig. 3, with the image size of 1607 × 600, shows the mixed document having 256 gray levels, obtained from an image scanner with the resolution of 200 dots/inch. The graphic image parts are separated from the text parts by considering chain codes of connected components. Fig. 4 shows the extracted graphic parts from Fig. 3.



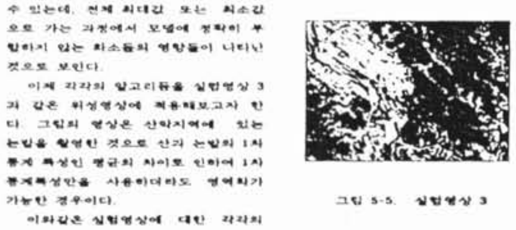
수 있는데, 전체 최대값 또는 최소값으로 가는 과정에서 모델에 정확히 부합하지 않는 화소들의 영향들이 나타난 것으로 보인다.

이제 각각의 알고리듬을 실험영상 3과 같은 위성영상에 적용해보고자 한다. 그림의 영상은 산악지역에 있는 논밭을 촬영한 것으로 산과 논밭의 1차 통계 특성인 평균의 차이로 인하여 1차 통계특성만을 사용하더라도 영역화가 가능한 경우이다.

이와같은 실험영상에 대한 각각의

그림 5-5. 실험영상 3

Fig. 1. Skew input document.



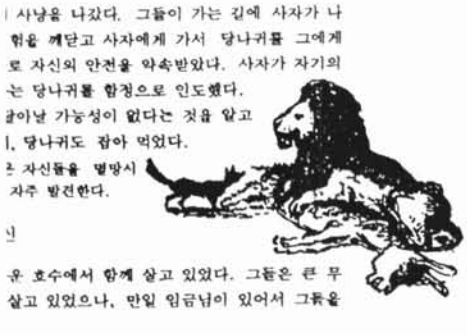
수 있는데, 전체 최대값 또는 최소값으로 가는 과정에서 모델에 정확히 부합하지 않는 화소들의 영향들이 나타난 것으로 보인다.

이제 각각의 알고리듬을 실험영상 3과 같은 위성영상에 적용해보고자 한다. 그림의 영상은 산악지역에 있는 논밭을 촬영한 것으로 산과 논밭의 1차 통계 특성인 평균의 차이로 인하여 1차 통계특성만을 사용하더라도 영역화가 가능한 경우이다.

이와같은 실험영상에 대한 각각의

그림 5-5. 실험영상 3

Fig. 2. Rotated document of Fig. 1.



| 사냥을 나갔다. 그들이 가는 길에 사자가 나
험을 깨닫고 사자에게 가서 당나귀를 그에게
로 자신의 안전을 약속받았다. 사자가 자기의
는 당나귀를 함정으로 인도했다.
갈아날 가능성이 없다는 것을 알고
I. 당나귀도 잡아 먹었다.
는 자신들을 별방시
자주 발견한다.

닌

유 호수에서 함께 살고 있었다. 그들은 큰 무
살고 있었으나, 만일 임금님이 있어서 그들을

Fig. 3. Mixed document 1.

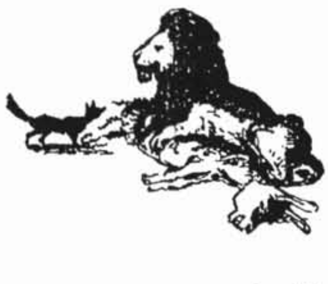
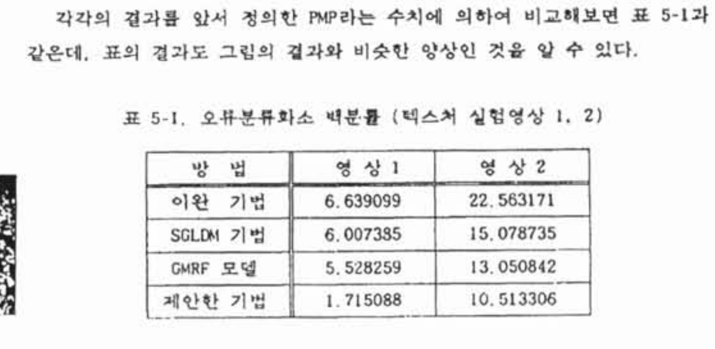


Fig. 4. Extraction of graphic parts from Fig. 3.

Fig. 5 shows the a mixed document consisting of the Korean/alphanumeric texts and a table. Fig. 6 shows the recognition result of the binary image shown in Fig. 5. It is observed that '숫' is incorrectly recognized as '숫', because the last consonant touches with the central vowel.

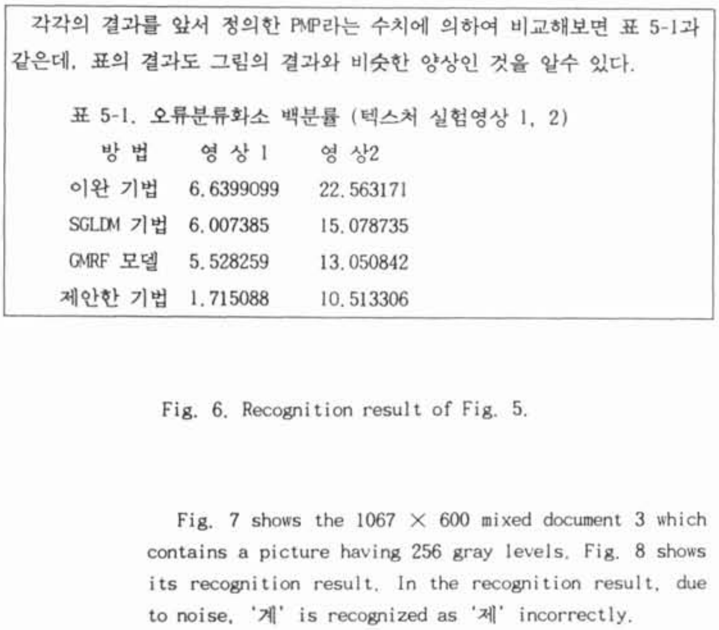


각각의 결과를 앞서 정의한 PMP라는 수치에 의하여 비교해보면 표 5-1과 같은데, 표의 결과도 그림의 결과와 비슷한 양상인 것을 알 수 있다.

표 5-1. 오류분류화소 백분률 (텍스쳐 실험영상 1, 2)

방 법	영 상 1	영 상 2
이완 기법	6.639099	22.563171
SGLDM 기법	6.007385	15.078735
GMRF 모델	5.528259	13.050842
제안한 기법	1.715088	10.513306

Fig. 5. Mixed document 2.



각각의 결과를 앞서 정의한 PMP라는 수치에 의하여 비교해보면 표 5-1과 같은데, 표의 결과도 그림의 결과와 비슷한 양상인 것을 알수 있다.

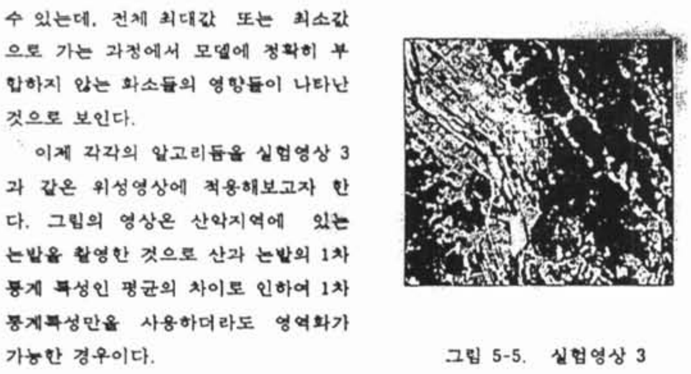
표 5-1. 오류분류화소 백분률 (텍스쳐 실험영상 1, 2)

방 법	영 상 1	영 상2
이완 기법	6.6399099	22.563171
SGLDM 기법	6.007385	15.078735
GMRF 모델	5.528259	13.050842
제안한 기법	1.715088	10.513306

Fig. 6. Recognition result of Fig. 5.

Fig. 7 shows the 1067 × 600 mixed document 3 which contains a picture having 256 gray levels. Fig. 8 shows its recognition result. In the recognition result, due to noise, '계' is recognized as '제' incorrectly.

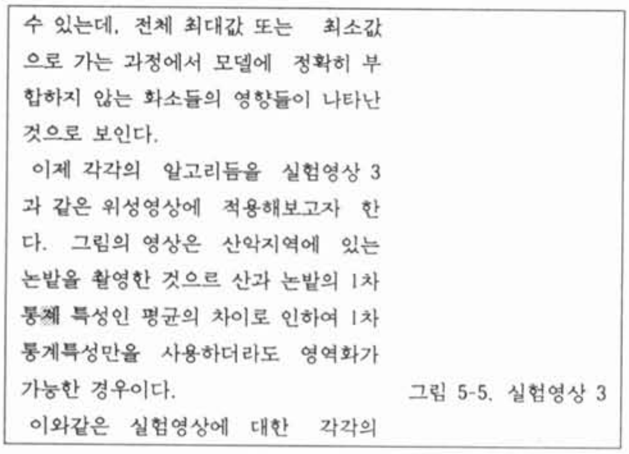
Fig. 9 shows the 1257 × 524 mixed document 4. Fig. 10 shows the recognition result of the mixed document 4. In the recognition result, all characters are recognized correctly.



수 있는데, 전체 최대값 또는 최소값
으로 가는 과정에서 모델에 정확히 부
합하지 않는 화소들의 영향들이 나타난
것으로 보인다.
이제 각각의 알고리듬을 실험영상 3
과 같은 위성영상에 적용해보고자 한
다. 그림의 영상은 산악지역에 있는
논밭을 촬영한 것으로 산과 논밭의 1차
통계 특성인 평균의 차이로 인하여 1차
통계특성만을 사용하더라도 영역화가
가능한 경우이다.
이와같은 실험영상에 대한 각각의

그림 5-5. 실험영상 3

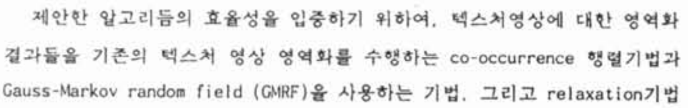
Fig. 7. Mixed document 3.



수 있는데, 전체 최대값 또는 최소값
으로 가는 과정에서 모델에 정확히 부
합하지 않는 화소들의 영향들이 나타난
것으로 보인다.
이제 각각의 알고리듬을 실험영상 3
과 같은 위성영상에 적용해보고자 한
다. 그림의 영상은 산악지역에 있는
논밭을 촬영한 것으르 산과 논밭의 1차
통계 특성인 평균의 차이로 인하여 1차
통계특성만을 사용하더라도 영역화가
가능한 경우이다.
이와같은 실험영상에 대한 각각의

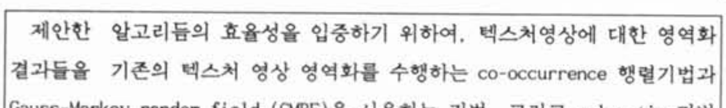
그림 5-5. 실험영상 3

Fig. 8. Recognition result of Fig. 7.



제안한 알고리듬의 효율성을 입증하기 위하여, 텍스처영상에 대한 영역화
결과들을 기존의 텍스처 영상 영역화를 수행하는 co-occurrence 행렬기법과
Gauss-Markov random field (GMRF)을 사용하는 기법, 그리고 relaxation기법

Fig. 9. Mixed document 4.



제안한 알고리듬의 효율성을 입증하기 위하여, 텍스처영상에 대한 영역화
결과들을 기존의 텍스처 영상 영역화를 수행하는 co-occurrence 행렬기법과
Gauss-Markov random field (GMRF)을 사용하는 기법, 그리고 relaxation기법

Fig. 10. Recognition result of Fig. 9.

CONCLUSION

In this paper, we propose an efficient algorithm which recognizes the mixed document consisting of the Korean/alphanumeric texts and graphic images. We separate graphic image parts from the text parts by considering chain codes of connected components. In the character recognition step we recognize separately the Korean and alphanumeric characters using several features pertinent to each class.

Computer simulation shows that the proposed algorithm recognizes the mixed document effectively. Further research will be focused on the recognition of the mixed document having various fonts and sizes.

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 Analysis Machine Intelligence*, vol. PAMI-1, pp. 245-251, July 1979.

[2] J. K. Lee, J. C. Namkung, and Y. K. Kim, "A study on the partial separation for subpatterns and recognition of the Hangul patterns," Journ. Korean Institute of Telematics and Electronics, vol. 18, no. 3, pp. 1-9, June 1981 (in Korean).

[3] H. S. Lee, T. Y. Choi, Y. K. Kim, and J. W. Kim, "A study on the text recognition using artificial intelligence technique," Journ. Korean Institute of Telematics and Electronics, vol. 26, no. 11, pp. 153-164, Nov. 1989 (in Korean).

[4] L. A. Fletcher and R. Kasturi, "A robust algorithm for text string separation from mixed text/graphics images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. PAMI-10, pp. 910-918, Nov. 1988.

[5] T. Pavlidis, "A thinning algorithm for discrete binary images," *Comput. Vision, Graphics, Image Processing*, vol. 13, pp. 142-157, June 1980.

[6] 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 algorithm for OCR document processing using a thinning process," in *SPIE Proc. Intelligent Robots and Computer Vision X: Algorithms and Techniques*, vol. 1607, pp. 146-157, Boston, MA, Nov. 1991.