Proceedings of the 12th
KACC, October 1997

인식률을 향상한 한글 문서 인식 알고리즘 개발
Development of an Image Processing Algorithm for Korean document Recognition

김희식*, °김영재 **, 이평원 **
* 서울시립대학교 제어계측공학과(Tel: 210-2428; FAX: 213-8317)
** 서울시립대학교 대학원 (Tel:210-2569; FAX: 213-8317)

Abstracts : This paper proposes a new image processing algorithm to recognize korean documents. It take out the region of text area from input image, then it makes segmentation of lines, words and characters in the text. A precision segmentation is very important to recognize the input document. The input image has 8-bit gray scaled resolution. Not only the histogram but also brightness dispersion graph are used for segmentation. The result shows a higher accuracy of document recognition.

Keyword : Pattern Recognition, Korean Letter Recognition, Image Processing, OCR

1. 서론

정보화 시대가 시작되면서부터 각종 정보, 문서들을 데이터베이스화 하여 저장하고, 이용하게 되었다. 그러면서 , 기존의 문서를 컴퓨터에 입력 시킬 필요성이 발생하고 , 그 일을 주로 사람이 키보드로 입력 하여 왔다. 또한 최근에는 스캐너를 이용하여 문서 영상을 입력 받고, 그 영상에서 문자만을 추출하여 인식하는 연구가 활발하게 진행되어 있고, 일부 상용화되어 사용되고 있다.

기존의 영문, 숫자의 인식은 우리나라 뿐만 아니라, 외국에서도 오랜연구에 의해 상당한 기술적 진보가 있었지만, 한글 인식에 경우, 기존의 영문, 숫자 인식 알고리즘으로는 부족한 면이 있다. 영문, 숫자는 하나의 독립된 모양체이지만 한글의 경우 초성,중성 또는 초성,중성,종성 등으로 조합된 문자체계이다. 따라서, 문자 모양의 종류가 다른 언어에 비해 상당히 많다. 주로 사용하는 문자, 즉 완성형 한글 코드를 기준하여도 2,350자에 이른다. 실제 조합형 한글 코드의 경우는 수 만가지의 모양을 가진다. 기존의 영문, 숫자 인식 방법을 그대로 적용하면 , 상당한 인식처리 시간을 필요로 하고, 인식률도 떨어진다. 따라서, 한글 고유의 문서 영상 처리, 분리 및 인식 방법을 필요로 한다.

본 논문에서는 한글 문자 인식에 앞서서 올바른 문서 영상을 얻기 위해 문서의 기울기를 구하고 그 것을 교정하는 알고리즘과 문서영상에서 줄을 분리하고, 다시 어절을 분리하고 음절을 분리하는 알고리즘에 관한 연구를 다루었다.

기존의 분리 과정을 보면 줄에서부터 직접 각 음절을 분리하여 인식을 하였다. 하지만 음절의 분리 이전에 어절의 분리를 통해 각 어절에 관한 정보를 취하므로서 단순한 한글 음절 인식에 의한 오류 발생을 줄이는 데 이용할수 있도록 했다. 단순한 음절 단위의 인식의 경우 한글 단어가 아닌 이상한 단어로 잘못 인식 오류 발생하기 쉽다. 차후연구 내용에는 어절 즉 음절간의 연관성에 관한 정보를 이용하여 인식도를 향상시키고 오류를 감소시키는 연구를 수행하고자 한다.

2. 문서 기울기의 측정 및 교정

일반적인 문서의 경우 줄과 줄사이 여백에 의해 y축에 대한 히스토그램을 계산하여 보면 여백과 글자 부분이 구별되어진다. 그렇지만 기울어진 정도에 따라 y축에 대한 히스토그램의 모양이 틀려진다. 그점을 이용하여 문서 영상을 일정 범위에서 일정 범위 까지 회전 시킨후 각각의 문서 영상에 대해 y축으로 히스토그램을 구하여 보면 여백과 글자 부분의 차이가 가장 큰 영상의 기울기 값이 문서의 기울기가 되고 그값에 의해 문서를 회전 시키므로서 문서 인식을 보다 정확하게 할수 있다.

어렵다는 단점이 있다. 분할과 인식을 병행하는 방법은 문자의 인식 결과를 고려하여 적합한 문자 분할을 할 수 있는데 반하여 사전 분할 위치의 조합 과정에서 반복적으로 생성되는 후보 문자에 대한 인식을 병행해야 하기 때문에 분할 속도가 상대적으로 느리다는 단점을 갖고 있다.

그림 1. 수평한 문서 영상

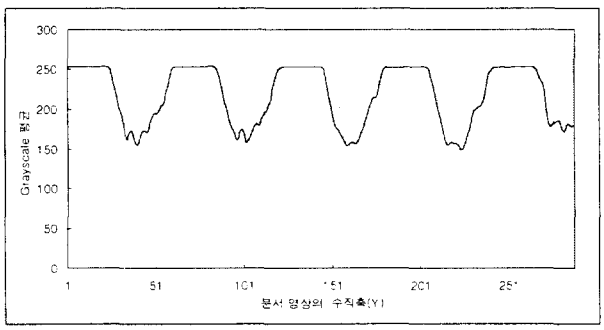

그림 2. 문서영상의 수직축에 대한 밝기 값의 히스토그램

$$y_i = \frac{1}{N}\sum_{i=1}^{n} g_i \qquad (1)$$

식(1)에서 g_i 는 각 픽셀의 그레이 레벨이고 N은 픽셀의 갯수이다.

어렵다는 단점이 있다. 분할과 인식을 병행하는 방법은 문자의 인식 결과를 고려하여 적합한 문자 분할을 할 수 있는데 반하여 사전 분할 위치의 조합 과정에서 반복적으로 생성되는 후보 문자에 대한 인식을 병행해야 하기 때문에 분할 속도가 상대적으로 느리다는 단점을 갖고 있다.

그림 3. 0.5 ° 기울어진 문서영상

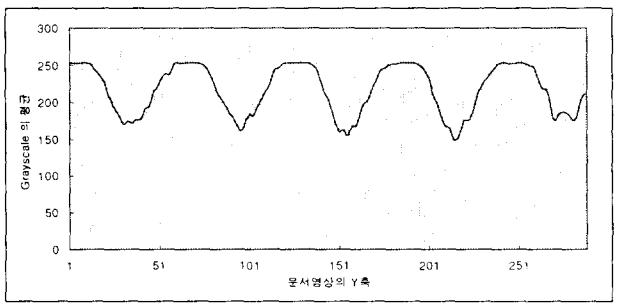

그림 4. 수직축에 대한 히스토그램 (0.5 °기울어진 문서영상)

어렵다는 단점이 있다. 분할과 인식을 병행하는 방법은 문자의 인식 결과를 고려하여 적합한 문자 분할을 할 수 있는데 반하여 사전 분할 위치의 조합 과정에서 반복적으로 생성되는 후보 문자에 대한 인식을 병행해야 하기 때문에 분할 속도가 상대적으로 느리다는 단점을 갖고 있다.

그림 5. 5 ° 기울어진 문서영상

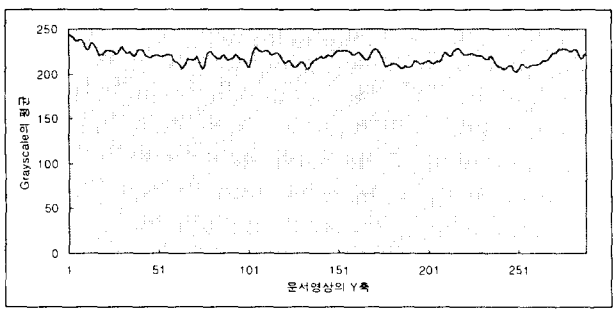

그림 6. 수직축에 대한 히스토그램 (5 °기울어진 문서영상)

입력문서영상이 위와 같이 기울어져 있을 때에는 문서 기울기 각도는 Try and Error 방법으로 최적합기울기를 찾아서 자동으로 각도가 계산 되어 진다.

3. 줄의 구별

줄의 구별의 경우, 문서 영상의 기울기가 올바르게 교정된 경우 기울기를 구할 때 이용한 y축 방향의 히스토그램을 보면 문자가 있는 어두운 부분과 밝은 여백부분을 정확히 구별할수 가 있다. 여기에서 문자가 있는 부분, 즉 Grayscale 값이 낮은 영역을 찾아 수평 문서줄을 분리하였다.

조합과 인식을 수행하여 그 결과로 표현된 그래프 상에서 최소 거리 탐색 기법으로 최적의 분할 위치를 결정한다.
본 논문의 구성은 다음과 같다. II장에서는 문자행 추출 및 문자 분할에 관한 기존의 관련 연구들을 살펴보고, III장에서는 한글 문서 영상에서 연결 요소 분석을 이용하

그림 7. 줄간격 분리 후 영상

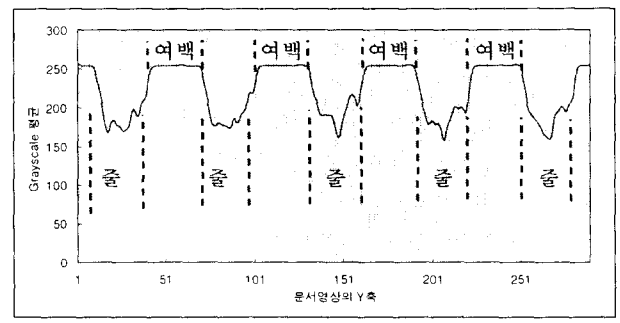

그림 8. 줄과 여백의 구분

4. 어절의 구별

입력문서에서 띄어쓰기와 붙여쓰는 단어와 조사를 합하여 한 어절이라고 한다. 이 어절단위는 한줄로 구성된 영상으로부터 음절단위의 분리 이전에 단어로 분리 하여 문서 인식시에 그 정보를 이용 할 수 있다. 문서 인식 과정에서 음절 단위의 인식일 경우 전후 음절과 관계를 고려하지 않는 경우 일반적으로 사용되는 단어가 아닌 이상한 단어가 나오는 경우가 많다 . 따라서 음절을 인식할 때 그 음절이 속한 어절내에서 그 음절 전후의 음절과의 관계를 이용하면 보다 정확한 인식을 할수 있다.

어절의 구별하는 방법으로는 글자 높이, 음절 간격, 어절간격의 관계를 이용하였다.

문서 종류	글자 높이(mm)	붙여쓴 음절간격(mm)	띄어쓴 어절간격(mm)
신문	3	0.8	2.2
잡지	3	0.8	2.2
논문	3	0.7	2.5
소설	4	0.7	3
일반문서	4	1	3.5

표 1 문서 영상에서 글자높이, 음절간격, 어절간격

위의 표에 보듯이 어절간격은 글자 높이에 70% 정도 이고 음절간격 은 약 27 %정도이다. 이점을 이용하면 음절 단위의 분리 이전에 어절단위로 분리할수 있다.

그림 10 에서 보면 평균 밝기값이 250 이상의 데이터가 일정한 크기로 연속적으로 보인다. 그 부분이 어절 간격 부분으로서 그폭은 약 글자 크기의 60 %에서 70 % 정도 된다.
따라서 이러한 특징을 가지고 분리하면 다음과 같다.

지금까지 연구되어 온 문자 분할 방법은 분할과 인식

그림 9. 어절 분리 전 영상

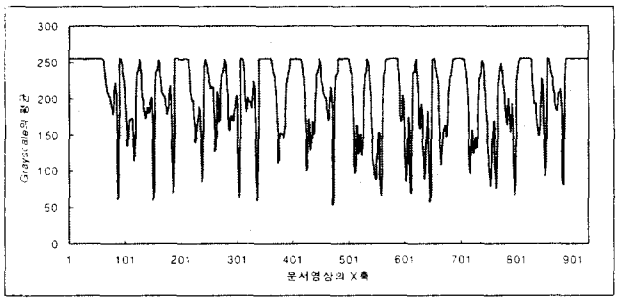

그림 10. 문서영상 한줄에 대한 수직축 평균값

그림 11. 어절 분리 하기 위한 알고리즘 적용결과

5. 음절의 분리

분리된 어절에서 음절별로 분리하여 각각의 음절을 인식하게 된다. 음절의 분리의 경우 어절의 분리와 같이 X축에 대한 히스토그램과 분산을 기본적으로 이용했다. 음절과 음절사이의 히스토그램은 그림13과 같이 250 이상의 값이 6~8 픽셀정도 연속되는 부분이다. 하지만 그림 14처럼 '인식이' 중에서 '이'의 부분을 보면 'ㅇ' 과 'ㅣ' 사이의 간격 부분도 음절 사이 간격으로 인식될 수 있다. 따라서 또 다른 조건을 추가 하므로서 분리할수 있었다. 어절의 분리와는 달리 분리 된 음절의 영역(크기)가 거의 일정하다. 그림 15처럼 그 점을 이용하여 히스토그램에 의해 영역을 분리하고 그 영역의 크기가 다른 영역 크기에 비해 50 % 이하일 경우 '이'에서 'ㅣ' 영역이라 판단하고 'ㅣ' 영역 이전 영역'ㅇ' 과 그 영역을 합쳐서 처리하였다.

그림 12. 분리전영상

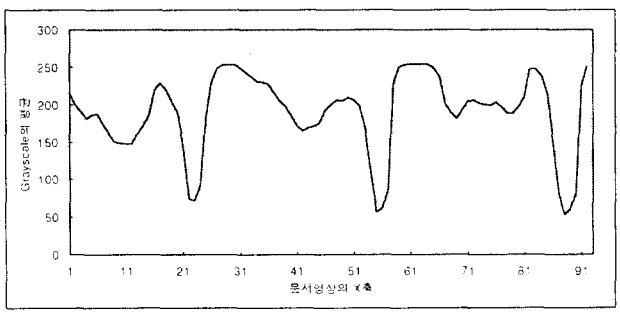

그림 13. 수직축 평균 밝기값
(음절과 그 여백간격의 구분하여 가능)

그림 14. 잘못된 분리

그림 15. 분리후 영상

그리고, 그림17 처럼 단순한 히스토그램만 이용할 경우 에는 '니다'의 '니' 처럼 글자의 두께가 얇은 부분을 여백으로 간주되는 경우가 있다. 그래서 분산을 이용 하여 그림 19 처럼 분리할수 있다. 분산 그래프의 경우 히스토그램보다 정확하게 여백 부분을 찾을 수있다.

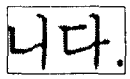

그림 16. 기본입력문서영상

그림 17. 단순한 수직축 평균 밝기 값만 적용한 결과

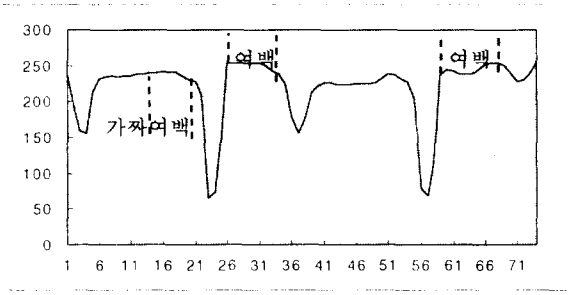

그림 18. 히스토그램

$$y_i = \frac{1}{N}\sum_{i=1}^{n}|m - g_i| \times 200 \qquad (2)$$

y_i : i 번째 위치의 밝기 분산값
i : 가로축 위치
N : 세로축 픽셀수
m : 픽셀들의 평균값
g_i : 각픽셀의 그레이레벨

식(2)는 픽셀의 분포도로 이 값이 높을 수 록 여백과 문자영역의 비율이 같은 경우 이고 여백만 있는 경우에는 그 값이 0에 가깝다. 문자만가 있는 경우는 문자 영역의 그레이레벨이 일정하지 않으므로 그 차이가 나타난다.

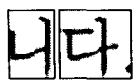

그림 19. 밝기 분포이용

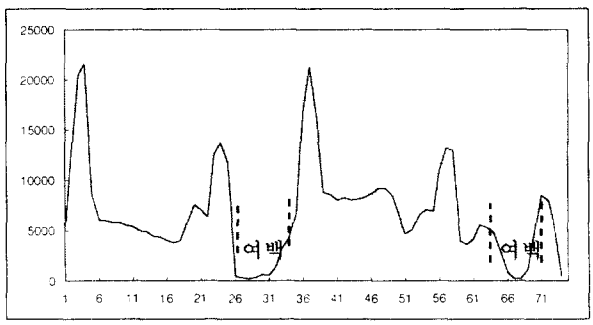

그림 20. 픽셀의 밝기 분포그래프

6. 실험 및 결과

실험 환경은 AMD K6-200 PC에서 Visual C++를 사용하였다. 인식 대상 문서의 문자 크기는 2mm x 3mm 이고, 문서 영상은

UMAX-8 로 8bit gray scan 하여 사용하였다.


지금까지 연구되어 온 문자 분할 방법은 분할과 인식이 분리된 방법과 분할과 인식을 병행하는 방법으로 크게 구분할 수 있다. 분할과 인식이 분리된 방법은 인식 결과의 도움 없이 문자 분할을 수행하므로 문자 분할 속도가 빠른 반면에 한글과 영·숫자가 혼용된 문서와 같이 문자간의 다양한 접촉 유형을 갖는 문서 영상에는 적용하기 어렵다는 단점이 있다. 분할과 인식을 병행하는 방법은 문자의 인식 결과를 고려하여 적합한 문자 분할을 할 수 있는데 반하여 사전 분할 위치의 조합 과정에서 반복적으로 생성되는 후보 문자에 대한 인식을 병행해야 하기 때문에

그림 21. 어절 분리 알고리즘 적용 결과


지금까지 연구되어 온 문자 분할 방법은 분할과 인식이 분리된 방법과 분할과 인식을 병행하는 방법으로 크게 구분할 수 있다. 분할과 인식이 분리된 방법은 인식 결과의 도움 없이 문자 분할을 수행하므로 문자 분할 속도가 빠른 반면에 한글과 영·숫자가 혼용된 문서와 같이 문자간의 다양한 접촉 유형을 갖는 문서 영상에는 적용하기 어렵다는 단점이 있다. 분할과 인식을 병행하는 방법은 문자의 인식 결과를 고려하여 적합한 문자 분할을 할 수 있는데 반하여 사전 분할 위치의 조합 과정에서 반복적으로 생성되는 후보 문자에 대한 인식을 병행해야 하기 때문에

그림 22. 음절분리 알고리즘 적용 결과

어절분리 알고리즘과 음절분리 알고리즘 적용한 결과는 각각 99% 및 98%은 구분 인식률을 보였다.

7. 결론 및 연구 방향

본 논문은 한글 인쇄체 문서 인식을 위한 회전 교정과 분할 과정에 관한 연구 이다. 줄을 구별하는 과정에서는 문서의 기울기를 정확히 찾아 낸 경우, 단순한 히스토그램의 분석 만으로도 줄을 분리 할수 있다. 하지만 문서의 기울기를 찾는 과정에서 문서가 다단인 경우, 첫 번째 단과 두 번째 단의 줄간격이 서로 엇갈려 있거나 그림이 있는 경우 그 값의 오류가 많았다.

어절의 구별은 일반적 문자 간격의 통계적 수치에 의해 하므로 일반적으로 음절의 구별에 비해 오류가 적었다. 음절의 구별은 일반적인 히스토그램과 픽셀의 분포도를 이용하여 분리하였다. 앞으로의 연구에서는 각 음절에서 자음, 모음을 다시 분리하여 각각을 인식하는 연구와 분리되어진 어절의 정보를 실제 문자 인식에 적용하는 방법에 대한 연구가 이루어져야 겠다.
연구 개발한 알고리즘을 좀 더 다양한 입력 문서 영상에 대하여 적용실험을 하여 인식률을 검정할 계획이다.

8. 참고 문헌

[1] Craig A.Lindly 지음, 류성렬 옮김," C 이미지 프로세싱",1991 , 동일출판사

[2] Edward R. Dougherty," Digital Image Processing Methods ", 1994, Marcel Dekker, pp. 43-102

[3] Rafael C. Gonzalez, Richard E. Woods, "Digital Image Processing", 1992, Addison-Wesley pp. 413-478

[4] Robert J.Schalkoff, "Digital Image Processing and Computer Vision", 1989, Jonh Wiley & Sons pp.130-178

[5] 김두식, 이성환,"한글과 영.숫자가 혼용된 문서를 위한 효과적인 문자 분할방법", 8회 영상처리 및 이해에 관한 워크샵 발표논문집, 1996년 1월 pp.19-26

[6] 김희승," 영상인식,- 영상처리 , 컴퓨터비전, 패턴인식, 신경망-",1993, 생능출판사

[7] 남궁재찬, "화상공학의 기초",1989, 기전연구사

[8] 이 성환, "문자인식 이론과 실제", 홍릉과학 출판사, 1권, 1994년 3월 pp 89-113

[9] 이 성환, "문자인식 이론과 실제", 홍릉과학 출판사, 2권, 1994년 3월

[10] 이 성환," 패턴인식의 원리", 홍릉과학 출판사, 1권 1994년 3월

[11] 일본공업기술센터편," 컴퓨터화상처리 입문", 1993, 기전연구사