투영 프로파일의 간략화 방법을 이용한 인쇄체 한글 문서 영상에서의 문자 분할

박 상 철[†] · 김 수 형[††]

요 약

본 논문에서는 한글 문서 영상에서의 문자 분할을 위한 2가지 알고리즘을 제안한다. 첫째는 투영 프로파일 기반 개선된 문자 분할 알고리즘이다. 이 알고리즘은 크게 문자수 추정, 분할 점 획득 및 문자 경계 탐색, 그리고 최적의 문자 분할 결과 선택으로 구성된다. 두 번째는 근접한 문자들이 서로 연결된 저 품질 문서 영상에 적합한 분할 알고리즘이다. 이 경우 연결요소를 제거하기 위해 투영 프로파일의 일부를 잘랐는데, 이를 $\alpha - cut$이라 한다. 그 후 전자의 방법을 변형하여 문자 분할을 수행한다. 다양한 폰트 속성을 갖고 품질이 낮은 43,572개의 한글 단어 영상을 대상으로 실험한 결과, 투영 프로파일 기반 개선된 문자 분할 알고리즘이 91.81%, 투영 프로파일에 $\alpha - cut$을 적용한 알고리즘이 99.57%의 문자 분할 성공률을 나타내어 저 품질 한글 문서 영상에서 $\alpha - cut$을 이용한 문자 분할 알고리즘이 효과적임을 입증하였다. [1]

키워드 : 문자 분할, 투영 프로파일, 저 품질 인쇄체 한글 단어 영상, 키워드 검출, 광학 문자 인식

Character Segmentation on Printed Korean Document Images Using a Simplification of Projection Profiles

Sang Cheol Park[†] · Soo Hyung Kim[††]

ABSTRACT

In this paper, we propose two approaches for the character segmentation on Korean document images. One is an improved version of a projection profile-based algorithm. It involves estimating the number of characters, obtaining the split points and then searching for each character's boundary, and selecting the best segmentation result. The other is developed for low quality document images where adjacent characters are connected. In this case, parts of the projection profile are cut to resolve the connection between the characters. This is called $\alpha - cut$. Afterwards, the revised former segmentation procedure is conducted. The two approaches have been tested with 43,572 low-quality Korean word images printed in various font styles. The segmentation accuracies of the former and the latter are 91.81% and 99.57%, respectively. This result shows that the proposed algorithm using a $\alpha - cut$ is effective for low-quality Korean document images.

Key Words : Character Segmentation, Projection Profile, Printed Korean Word Image, Keyword Spotting, OCR

1. 서 론

문자 분할은 단어 영상을 문자 단위로 분할하는 것을 의미하며, 주로 OCR(Optical Character Recognition, 광학 문자 인식) 시스템의 일부로서 구현되어 있다. 대부분의 OCR 시스템은 문서 영상을 문자 단위로 인식하기 때문에 문자 분할 과정에서 발생하는 에러는 OCR 시스템의 전체 성능을 저하시키는 주된 원인이다[1].

OCR을 이용한 문서 영상 검색 방법은 문서 영상의 문자를 인식하여 기계가 판독할 수 있는 텍스트 형태로 변형을 한 후 텍스트 매칭을 이용하여 검색한다. 이 방법의 문제점은 검색 문자가 오인식될 경우 검색이 불가능하다는 점이다. 특히 문서의 훼손이 심한 경우 검색 성능이 급격히 떨어지는 단점이 있다. 문서 영상 검색을 위한 새로운 방법으로는 키워드 검출(Keyword spotting)법이 있다[2]. 키워드 검출은 검색어 영상과 문서 내 단어 영상의 특징을 매칭하여 검색한다. 영문 문서 영상에서의 키워드 검출은 일반적으로 단어 영상을 일차원 특징 벡터로 표현한 후 단어 단위로 매칭을 수행 한다[3-7]. 이 경우 문자 분할은 필요하지 않지만 한문과 한글처럼 문자의 구성요소들을 모아쓰는 언어에서는 문자 단위의 2차원 특징 정보를 이용한 키워드 검출을 수행한다[8-10]. 이 경우 문자 분할은 검색 성능에 결정적인 요

※ 이 논문은 2005년도 한국학술진흥재단 선도연구과제의 지원에 의하여 연구되었음.
† 정 회 원 : 전남대학교 정보통신연구소 연구원
†† 정 회 원 : 전남대학교 전자컴퓨터정보통신공학부 교수
논문접수 : 2005년 9월 26일, 심사완료 : 2006년 2월 24일

인이 된다.

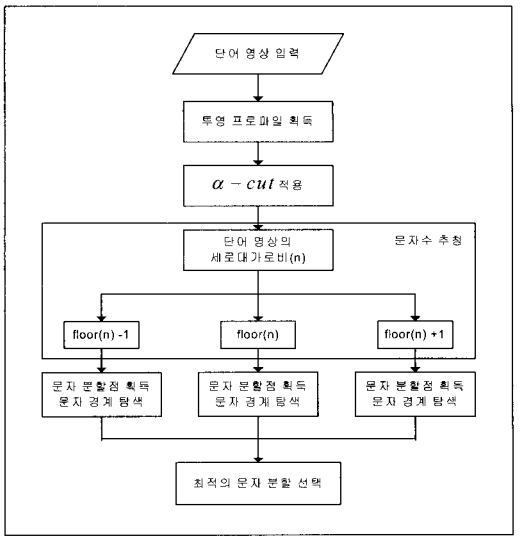
문자 분할 알고리즘은 다양한 형태로 분류[11-13]할 수 있지만 분할을 위한 특징 추출 방법에 근거하여 분류하면 투영에 기반한 문자 분할[14-17], 연결 화소 분석과 외곽선 추적에 의한 문자 분할[18-20], 구조적 특징 분석에 기반한 문자 분할[21] 등이 있고, 앞서 나열된 방법을 인식과 결합한 방법[15, 18, 21]이 있다. 다양한 폰트, 성능이 낮은 프린터, 이진화 등의 전처리는 하나의 문자를 여러 조각으로 분리하거나 여러 개의 문자를 연결시키는 요인이 되며, 이들은 문자 분할을 어렵게 한다[11].

투영에 기반한 문자 분할은 일반적으로 '문자 사이 칼럼의 투영 값이 0이면 이 칼럼은 문자 사이의 여백'이라는 특성을 이용한다. 한글은 자소를 모아쓰기 한다. 자소의 모아쓰기는 한 문자에서 자소 사이의 여백을 발생시키기 때문에, 투영 값을 이용하여 문자를 분할할 경우 문자 폭을 고려하여 분리된 자소를 합성시키는 과정이 필요하다. 이 방법은 문자가 기울어지지 않고 겹침이나 접촉이 심하지 않는 인쇄체 문자 분할에 효과적이다[13].

특히, [16]에서는 한글 문자열 영상에서 문자를 분할하였는데, 문자열 영상을 수직 투영하여 투영 값 0인 곳에서 문자를 분할하였다. 분할된 초기 문자 중 너비가 가장 큰 두 개를 선택, 평균하여 평균 문자 너비로 삼았다. 한글 문자 중에서 왼쪽 자음이 오른쪽 모음의 중간 정도에 존재 한다는 특성과 표준 문자 너비는 합성된 문자의 높이와 비율이 거의 1에 가깝다는 특성을 이용하여 문자를 합성하였다. 이 방법은 바탕체 계열의 폰트에 적합하도록 설계되었으며, 수직 획의 변형이 있을 때 합성 원칙에 적용할 수 없어 분할이 어렵다. 이와 비슷한 방법으로 [17]이 있는데, 후보 문자에서 휴리스틱을 적용하여 문자별로 분할하였다.

본 논문에서는 한글 문서 영상에서의 문자 분할을 위한 2가지 알고리즘을 제안한다. 첫째로 투영 프로파일 기반 개선된 문자 분할 알고리즘을 제안한다. 기존의 투영 프로파일 기반 문자 분할 방법들이 기준이 되는 하나의 문자 너비를 이용한 반면, 이 방법에서는 단어 영상의 문자 개수를 3가지로 추정하여 에러를 최소화하고, 최적의 분할 결과를 선택하기위해 분할된 문자 너비의 분산이 최소가 되는 분할 결과를 선택한다. 이 알고리즘은 크게 문자수 추정, 분할 점 획득 및 문자 경계 탐색, 그리고 최적의 문자 분할 결과 선택으로 구성된다. 두 번째는 전자의 알고리즘이 저 품질의 영상에 강인하도록 투영 프로파일의 일부를 변형한 투영 프로파일에 $\alpha - cut$을 적용한 분할 알고리즘이다.

첫 번째로 제안된 문자 분할 알고리즘은 상태가 양호한 한글 단어 영상에 대해서는 효과적이다. 하지만 저 품질 단어 영상에서는 문자와 문자 사이의 공백이 훼손되어 분할의 신뢰성이 매우 낮아진다. 따라서 두 문자간 분할점 탐색을 위해 투영 프로파일에 $\alpha - cut$을 적용한다. (그림 1)은 투영 프로파일에 $\alpha - cut$을 적용한 문자 분할 알고리즘의 블록 다이어그램이다. 입력 단어 영상은 정창부 등 [22]의 시스템을 이용하여 문서 영상으로부터 이미 분할되었다고 가정한다.



(그림 1) 투영 프로파일에 $\alpha - cut$을 적용한 문자 분할 알고리즘의 블록 다이어그램

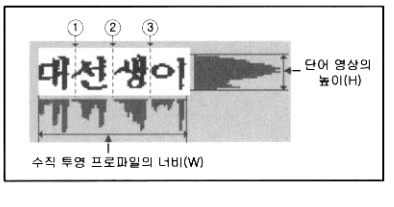
2. 투영 프로파일 기반 개선된 문자 분할

한글 폰트는 개개의 문자를 동일한 크기의 정사각형 안에 위치하도록 디자인(typesetting) 되어 있다. 따라서 각 문자의 너비와 높이의 크기가 같고 모든 문자의 크기 역시 동일하다. 또한 문자와 문자 사이의 공백으로 문자를 구분하여 가독성을 높이고 있다. 본 논문에서는 위 사실에 근거하여 투영 프로파일 기반 개선된 문자 분할 알고리즘을 제안한다.

투영 프로파일 기반 개선된 문자 분할 알고리즘은 3단계로 구성된다. 첫 번째 단계에서는 단어 영상이 몇 개의 문자로 구성되었는지 추정한다. 추정 값은 3가지로 구성된다. 두 번째 단계에서는 각각의 추정 문자수로 단어 영상을 나누어 분할 점을 얻고 그 점을 기준으로 문자의 경계를 탐색한다. 세 번째 단계에서는 3가지 문자 분할 결과 중에서 최적의 문자 분할 결과를 선택한다.

2.1 문자수 추정

한글 문자의 바운드 박스(최소 외곽 사각형)는 일반적으로 정사각형의 형태를 띠고 그 크기가 일정하다. 그러므로 단어 영상의 높이로 수직 투영 프로파일의 너비를 나누면 단어영상을 구성하는 문자수를 추정할 수 있다. (그림 2)는 한글 단어 영상의 수직투영 프로파일과 수평 투영 프로파일을 도시하고 있다. 식 1은 한글 단어 영상에서 문자수를 추정하는 일반적인 방법을 나타낸다.



(그림 2) 4문자로 구성된 단어 영상의 투영 프로파일

$$n = W/H \quad \text{(식 1)}$$

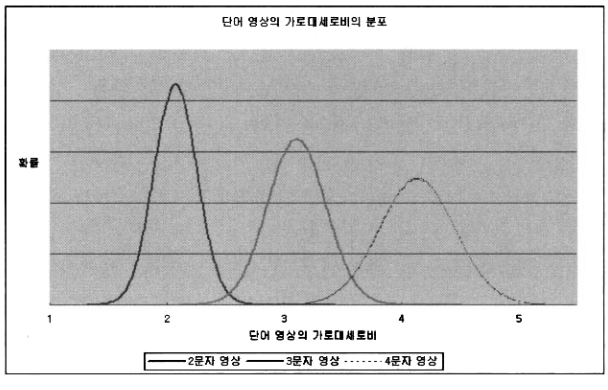
n: 가로대세로의 비

$\lfloor n \rfloor$: 추정 문자수

W:수직 투영 프로파일의 너비

H:단어 영상의 높이

하지만 이러한 방법으로 추정된 문자수는 항상 신뢰하기는 어렵다. 예를 들어 (그림 3)은 2문자 영상(2문자로 이루어진 단어 영상), 3문자 영상 및 4문자 영상들의 가로대세로비의 분포를 4.1절의 표 1에 근거하여 보여주는데, 단어 영상의 가로대세로비가 3.2라면 실제 문자수는 2자, 3자, 혹은 4자일 수 있다. 따라서 가로대세로비가 n이라면 실제 문자수는 $\lfloor n \rfloor$-1, $\lfloor n \rfloor$, $\lfloor n \rfloor$+1 중에 하나라고 할 수 있기 때문에 추정 문자수로 이 3가지 모두를 고려한다.

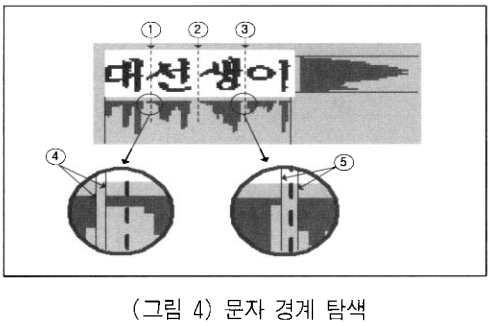


(그림 3) 문자수에 따른 단어 영상의 가로대세로비의 분포

2.2 분할 점 획득 및 문자 경계 탐색

앞서 언급한 한글 문자의 형태적인 특성을 이용하여 추정된 문자수로 단어 영상을 균등하게 나누어 분할 점을 획득하고 각 문자의 경계를 탐색한다. (그림 4)에서 ①, ②, ③을 추정 문자수 $\lfloor n \rfloor$으로 단어 영상을 균등하게 나누어 얻은 분할 점이라고 가정하자. 이때 ④와 ⑤는 분할 점 ①과 ③으로부터 각각 탐색된 문자의 경계이다.

$\lfloor n \rfloor$은 단어 영상의 문자수를 의미하기 때문에 $\lfloor n \rfloor$으로 나누어 얻어진 분할 점은 문자 사이의 공백에 위치해야한다. 따라서 탐색된 분할 점에서 투영 프로파일 값은 0이라고 추정할 수 있다. ②와 ③의 경우, 해당 지점에서 투영 프로파일의 값이 0이기 때문에 그 지점은 두 문자의 분할 점으로 합당하다. 문자의 경계는 문자 사이의 공백이 끝나는 지점에 위치한다는 사실에 근거하여 찾아진다. 분할 점에서 투영프로파일의 좌우로 이동하면서 그 값이 0이 아닌 지점이 관련된 문자의 경계이다.



(그림 4) 문자 경계 탐색

분할 대상 점 ①에서 투영프로파일의 값은 0이 아니므로, 분할 점이 문자 위에 있음을 알 수 있다. 이 경우 분할 대상 ①의 지점에서 투영프로파일의 좌우로 이동하면서 투영프로파일 값이 처음으로 0인 지점을 만나면 이 점을 문자의 경계로 결정한다. 여기서 탐색된 문자의 경계는 두 문자 사이를 구분 짓는 공백의 한쪽 끝이기도 하다. 앞서 추정된 문자수 $\lfloor n \rfloor$-1, $\lfloor n \rfloor$+1에 대해서도 위와 같은 방법으로 분할 점 및 문자 경계를 탐색한다.

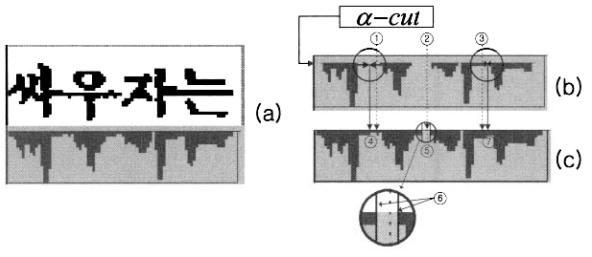
2.3 최적의 분할 결과 선택

문자 분할 결과가 3가지이기 때문에 이들 중 최적의 문자 분할 결과를 선택한다. 문자의 바운드 박스가 일정한 크기라고 가정하면, 추정 문자수에 따라 문자 분할이 올바르게 수행될 경우 문자 너비의 분산은 그렇지 않은 경우보다 적은 값을 갖는다. 따라서 3가지 추정 문자수로 문자 분할을 수행한 후, 그들의 문자 너비 값 분산이 최소가 되는 분할 결과를 선택한다.

3. 투영 프로파일에 $\alpha - cut$을 적용한 문자 분할

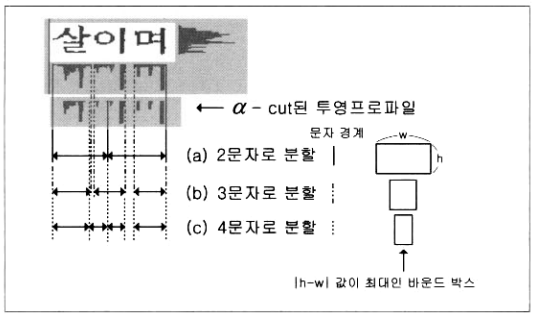
(그림 5)의 (a)와 같이 두 문자가 서로 연결되었을 경우 2절에서 제안된 투영 프로파일 기반 개선된 문자 분할은 이를 해결할 능력이 없다. 따라서 우리는 연결된 두 문자를 단절시키기 위해 (b)처럼 수직 투영 프로파일을 상위 α비율만큼 삭제($\alpha - cut$)하였다. (b)는 $\alpha - cut$을 적용한 투영 프로파일을 나타낸다. α는 실험에 의하여 수직 투영프로파일의 평균값의 7%로 결정하였다.

(그림 5) (b)의 ①, ②, 그리고 ③을 문자 수 $\lfloor n \rfloor$으로 나누어 얻어진 분할 점이라고 가정하자. 두 원안의 마주보는 기호 →←는 투영 프로파일에 $\alpha - cut$을 적용한 후, 두 문자 사이의 연결이 제거되어 생성된 공백을 나타낸다. 분할 점 ①과 같이 $\alpha - cut$된 투영 프로파일의 값이 0이고 원래 투영 프로파일의 값이 0이 아닌 경우, (b)의 첫 번째 원안에 있는 공백의 가운데 점에 대응하는 (c)의 ④를 두 문자의 공동 경계로 한다. 분할 점 ②의 경우와 같이 해당 지점에서 $\alpha - cut$된 투영 프로파일과 원래의 투영 프로파일 값 모두 0인 경우, 해당 분할 점에서 원래 투영 프로파일의 좌우로 이동하면서 그 값이 0이 아닌 지점 ⑥을 문자의 경계로 결정한다. 분할 점 ③과 같이 해당 프로파일 값이 모두 0이 아닌 경우, 분할 점에서 $\alpha - cut$된 투영 프로파일의 양쪽을 검색하여 먼저 나타난 공백인 (b)의 두 번째 원 안의 공백의 중간점에 대응하는 (c)의 ⑦을 두 문자의 경계로 판단한다.



(그림 5) 문자간 연결 성분 제거

투영 프로파일에 $\alpha-cut$을 적용하면 연결된 문자들을 분리하는데 효과적이다. 그렇지만 문자 획의 두께가 얇을 경우, 과다 분할되는 문제점이 있다. 그 결과 하나의 문자를 두개로 오 분리할 수 있다. 예를 들어 (그림 6)처럼 투영 프로파일에 $\alpha-cut$을 적용한 후, 세 개의 추정 문자수에 따라 (a), (b), (c)로 분할되었다고 가정하면, 투영 프로파일 기반 개선된 문자 분할에 적용하였던 최적의 문자 분할 결과 선택 방법은 문자 너비의 분산이 최소가 되는 (a)을 선택하게 된다. 이 결과는 문자 "이"를 둘로 분리했음에도 최적의 분할 결과로 선택한 오류이다.



(그림 6) $\alpha-cut$을 적용한 후 실패한 문자 분할 결과

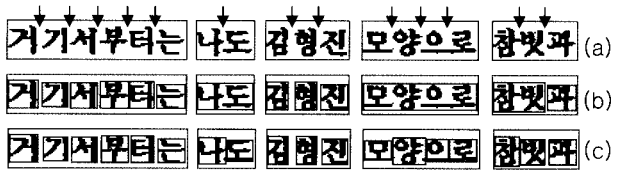
따라서 최적의 문자 분할 결과를 얻기 위해서 앞서 언급된 한글의 특성 '문자는 정사각형 안에 디자인 된다' 에 근거하여 올바르게 분할된 결과는 그렇지 않은 경우보다 문자의 높이와 너비의 차가 적음을 이용한다. 먼저 각각의 분할 결과에서 문자의 높이와 문자의 너비의 차가 최대가 되는 값을 찾는다. (그림 6)의 |h-w| 값이 최대인 바운드 박스를 참고하라. 이들 값 중에서 가장 적은 값을 갖는 문자 분할 결과를 최종 분할 결과로 선택한다. 식 2는 투영 프로파일에 $\alpha-cut$을 적용한 문자 분할 알고리즘을 위한 최적의 문자 분할 결과를 선택하는 과정을 나타낸다.

$$Best\,i = arg \min_{i=\lfloor n \rfloor, \lfloor n \rfloor \pm 1} \left(\max_{1 \le j \le i} | h_{ij} - w_{ij} | \right) \quad \text{(식 2)}$$

$Best\,i$: 3가지 문자 분할 방법 중 최적의 분할 결과를 갖는 인덱스

h_{ij}: i 번째 분할 방법으로 분리된 j 번째 문자의 높이

w_{ij}: i 번째 분할 방법으로 분리된 j 번째 문자의 너비



(그림 7) 문자 분할 결과

(그림 7)의 (a)는 입력 영상이다. 화살표는 문자의 분할 목표점이다. (b)는 2절의 제안 방법만을 적용한 문자 분할 결과이다. (c)는 $\alpha-cut$을 적용한 후의 문자 분할 결과이다. 문자들이 서로 연결된 경우 $\alpha-cut$은 문자 분할 목표점의 위치에서 문자가 분할되도록 돕는다.

4. 실험 결과

4.1 실험 환경 및 실험 데이터

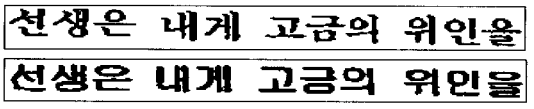
"백범일지" 일부를 마이크로소프트 워드를 이용하여 A4 10쪽, 4,252단어 분량의 문서 파일로 만들었다. 이를 서로 다른 12가지 폰트 속성-2가지 종류의 서체(바탕체, 굴림체), 3가지 종류의 문자 크기(8, 10, 12) 및 2가지 종류의 두께(bold, regular)]의 조합-으로 편집하였다. 이 문서 파일을 삼성 ML-8065 프린터로 출력한 후, 제록스 Document Centre 285 PLUS G 복사기로 복사하되, 복사 결과물을 다시 복사하는 방식으로 8회 복사하였고, EPSON GT-30000 스캐너를 사용하여 200DPI로 스캔하여 저장하였다. 이 문서 영상을 [21]의 시스템을 이용하여 단어 단위 영상으로 분할하여 총 51,024개의 단어 영상을 획득하였다. 실험에 사용된 기자재는 Pentium-4 CPU 2.80GHz와 1GB RAM 자원을 갖는 개인용 PC이다. 표 1은 실험 데이터에 출현한 단어 영상이 포함하고 있는 문자의 개수에 따라 단어의 세로대가로비의 분포를 나타낸다. (그림 8)의 (a)는 바탕체 속성을 갖고, (b)는 굴림체 속성을 갖는 한글 문자열 영상들이다. 두 영상은 모두 10포인트, 굵기 속성을 갖고 있다. 굴림체는 바탕체에 비해 상대적으로 문자와 문자의 위·아래 겹침이 적다.

〈표 1〉 단어 영상의 세로대가로비의 분포

실제 문자수	실험 단어 영상수	세로대 가로비 최소값	세로대 가로비 최대값	세로대 가로비 평균	세로대 가로비 표준편차
1	3,948	0.68	2.17	1.08	0.21
2	16,584	1.31	3.41	2.12	0.20
3	18,396	2.10	4.52	3.15	0.25
4	8,592	3.10	5.76	4.20	0.33
5	2,640	3.67	8.56	5.20	0.40
6	696	5.03	8.61	6.19	0.45
7	96	7.64	9.25	7.18	0.55
8	12	8.17	9.95	7.93	0.51
9	12	7.57	11.12	8.98	0.40
10	24	8.13	12.04	9.94	0.95
11	12	9.61	12.65	11.39	0.69
12	12	10.20	12.36	11.41	0.55
전체	51,024				

〈표 2〉 폰트 별 문자 분할 결과

서체	굵기	크기	문자수	단어 개수	$\alpha - cut$ 적용 안함		$\alpha - cut$ 적용	
					성공 개수	성공률(%)	성공 개수	성공률(%)
바탕	굵게	8	2	1,382	1,229	88.93	1,381	99.93
			3	1,533	807	52.64	1,521	99.22
			4	716	292	40.78	711	99.30
			전체	3,631	2,328	64.11	3,613	99.50
		10	2	1,382	1,318	95.37	1,378	99.71
			3	1,533	1,138	74.23	1,524	99.41
			4	716	467	65.22	705	98.46
			전체	3,631	2,923	80.50	3,607	99.34
		12	2	1,382	1,314	95.08	1,380	99.86
			3	1,533	1,218	79.45	1,532	99.93
			4	716	500	69.83	695	97.07
			전체	3,631	3,032	83.50	3,607	99.34
	보통	8	2	1,382	1,348	97.54	1,370	99.13
			3	1,533	1,313	85.65	1,516	98.89
			4	716	575	80.31	704	98.32
			전체	3,631	3,236	89.12	3,590	98.87
		10	2	1,382	1,367	98.91	1,375	99.49
			3	1,533	1,493	97.39	1,525	99.48
			4	716	675	94.27	705	98.46
			전체	3,631	3,535	97.36	3,605	99.28
		12	2	1,382	1,377	99.64	1,381	99.93
			3	1,533	1,516	98.89	1,529	99.74
			4	716	694	96.93	705	98.46
			전체	3,631	3,587	98.79	3,615	99.56
굴림	굵게	8	2	1,382	1,374	99.42	1,380	99.86
			3	1,533	1,318	85.98	1,529	99.74
			4	716	570	79.61	714	99.72
			전체	3,631	3,262	89.84	3,623	99.78
		10	2	1,382	1,378	99.71	1,378	99.71
			3	1,533	1,524	99.41	1,530	99.80
			4	716	711	99.30	715	99.86
			전체	3,631	3,613	99.5	3,623	99.78
		12	2	1,382	1,381	99.93	1,381	99.93
			3	1,533	1,526	99.54	1,528	99.67
			4	716	714	99.72	715	99.86
			전체	3,631	3,621	99.72	3,624	99.81
	보통	8	2	1,382	1,382	100	1,382	100
			3	1,533	1,523	99.35	1,531	99.87
			4	716	710	99.16	715	99.86
			전체	3,631	3,615	99.56	3,628	99.92
		10	2	1,382	1,382	100	1,382	100
			3	1,533	1,525	99.48	1,526	99.54
			4	716	714	99.72	714	99.72
			전체	3,631	3,621	99.72	3,622	99.75
		12	2	1,382	1,382	100	1,382	100
			3	1,533	1,531	99.87	1,531	99.87
			4	716	716	100	716	100
			전체	3,631	3,629	99.94	3,629	99.94
전체				43,572	40,002	91.81	43,386	99.57

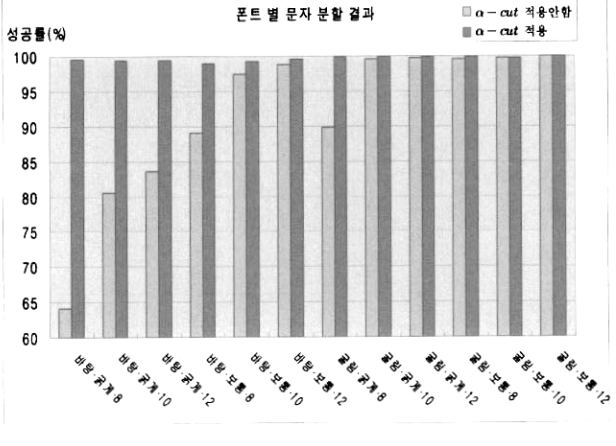


(그림 8) 바탕체 및 굴림체의 한글 영상의 예

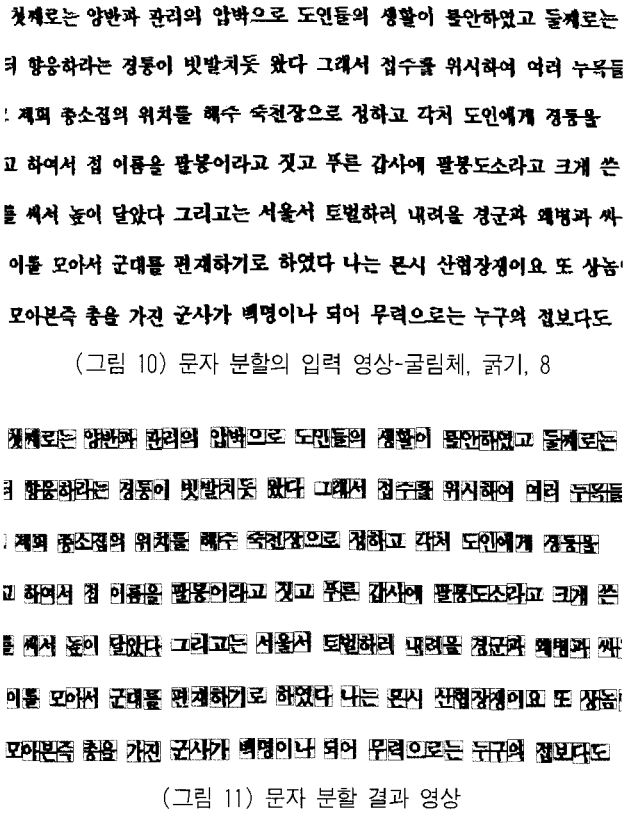
4.2 문자 분할 결과

본 논문에서는 실험 데이터 51,024개의 단어 영상에서 출현 빈도가 높은 2문자, 3문자 그리고 4문자로 이루어진 43,572개의 단어 영상을 대상으로 실험한 결과, 투영 프로파일 기반 개선된 문자 분할 알고리즘이 91.81%, 투영 프로파일에 $\alpha - cut$을 적용한 알고리즘이 99.57%의 문자 분할 성공률을 나타내었다. <표 2>는 문자 분할 결과를 12가지 폰트 및 단어 내 구성 문자수로 구분하여 나타낸다. (그림 10)은 <표 2>의 $\alpha - cut$을 적용한 경우와 그렇지 않은 경우의 문자 분할 성공률을 그래프로 비교하여 나타낸다.

투영 프로파일 기반 개선된 문자 분할 알고리즘을 분석해 보면, 두 문자의 위·아래 겹침이 상대적으로 많은 바탕체에서 분할 성공률이 저조하고(바탕체: 85.56%, 굴림체: 98.05%), 문자의 획이 굵은 영상에서도 그렇지 않은 경우보다 분할의 성능이 낮다(굵게: 86.20%, 보통: 97.42%). 또한 문자의 크기 속성이 작을수록 상대적으로 낮은 문자 분할 결과를 보인다(8: 85.66%, 10: 94.27%, 12: 95.49%). 이는 문자 사이의 공백이 적고, 문자의 획이 굵고, 작은 크기의 폰트로 작성되고, 낮은 해상도(200DPI)로 스캔된 저 품질의 단어 영상들은 획득 과정에서 획에 변형이 가해져 두 문자가 붙거나 위·아래로 겹치게 되어 투영 프로파일을 이용한 문자 분할을 어렵게 하기 때문이다. 반면 투영 프로파일에 $\alpha - cut$을 적용한 문자 분할 알고리즘은 두 문자의 연결 요소를 효과적으로 제거하기 때문에 어느 폰트에서라도 일정한 문자 분할 성공률을 보인다. 결과적으로 투영 프로파일에 $\alpha - cut$을 적용하면 저 품질의 한글 문서 영상의 문자 분할에 큰 효과가 있다. (그림 10)은 실험 영상에서 일부 발췌된 부분 영상으로써 문자 분할을 위한 입력 영상이며, (그림 11)은 문자 분할 후 그 결과를 표시한 그림이다.



(그림 9) $\alpha - cut$ 적용 여부에 따른 문자 분할 성공률 비교

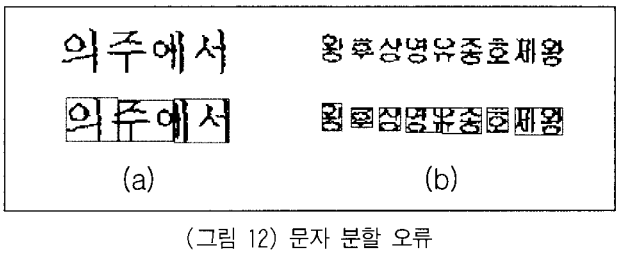


(그림 10) 문자 분할의 입력 영상-굴림체, 굵기, 8

(그림 11) 문자 분할 결과 영상

4.3 문자 분할 오류 분석

본 논문에서 제안한 투영 프로파일에 $\alpha - cut$을 적용한 문자 분할 알고리즘에 의해 발생하는 문자 오류는 크게 두 가지로 분류된다. (그림 10)은 이 두 가지 분할 오류를 예시하고 있다. (a)의 경우 4문자로 분할되어야 하지만, 마지막 문자 "서"가 세로로 길게 표현되어 문자의 높이와 문자의 너비의 차가 3문자로 분할되었을 때의 그것보다 크기 때문에, 단어 영상을 3문자로 잘 못 분할한 오류이다. (b)의 경우 단어 영상이 9개의 문자를 포함하고 있지만 단어의 세로 대가로비 n이 7.91이다. 따라서 3가지 추정 문자수는 6, 7, 8이기 때문에 8문자로 분할되어 발생한 오류이다.



(그림 12) 문자 분할 오류

5. 결 론

본 논문에서는 한글 문서 영상에서의 문자 분할을 위한 2가지 알고리즘을 제안하였다. 첫째는 투영 프로파일 기반 개선된 문자 분할 알고리즘이고, 둘째는 전자의 알고리즘이 저해상도 영상에 강인하도록 투영 프로파일에 $\alpha - cut$을 적

용한 분할 알고리즘이다. 실험결과 전자는 고해상도 한글 단어 영상에 대해서는 효과적이었다. 하지만 저해상도 단어 영상에서는 잡음으로 인해 문자와 문자 사이의 분할 점을 잃어버려 분할에 실패하였다. 이에 두 문자간 연결 성분을 제거하기 위해 투영 프로파일에 $\alpha - cut$을 적용하였다. 그러나 $\alpha - cut$은 두 문자간 연결 성분을 제거하는데 효율적이기도 하지만 하나의 문자를 둘로 분리하는 경향이 있기 때문에, 한글 특성에 근거하여 한 문자가 둘로 분리되는 오류를 방지하였다.

다양한 폰트 속성을 갖고 품질이 낮은 43,572개의 한글 단어 영상을 대상으로 실험 비교한 결과, 투영 프로파일에 $\alpha - cut$을 적용한 알고리즘이 99.57%, 투영 프로파일 기반 개선된 문자 분할 알고리즘이 91.81%의 문자 분할 성공률을 나타내어 저해상도 한글 단어 영상에서 제안된 문자 분할 알고리즘의 우수함을 입증하였다. 따라서 제안된 문자 분할 알고리즘은 저해상도 한글 문서 영상에서 키워드 검출 시스템 구현 및 OCR 시스템의 성능 개선 등에 기여할 것으로 사료된다. 향후에는 문자 획의 두께에 따라 α값이 자동으로 선택되도록 하는 α값 추정에 대한 연구와 한글이 영문 및 숫자 등과 혼용되어 있는 경우와 같이 문자의 세로대가로비가 일정하지 않은 문자열을 처리하기 위한 연구를 수행할 예정이다.

참 고 문 헌

[1] R. G. Casey and G. Nagy, "Recursive segmentation and classification of composite character patterns," 6th International Joint Conference on Pattern Recognition, pp. 1023-1026, 1982.

[2] D. Doermann, "The retrieval of document images: a brief survey," Proc. ICDAR 97, Ulm, pp.945-949, 1997.

[3] Y. Lu, and C. L. Tan, "Word searching in document images using word portion matching," 5th IAPR International Workshop on Document Analysis Systems, USA, pp. 319-328, 2002.

[4] Y. Lu, L. Zhang, and C. L. Tan, "A search engine for imaged documents in PDF files," 27th Annual International ACM SIGIR Conference, UK, pp.536-537, 2004.

[5] J. DeCurtins and E. Chen, "Keyword spotting via word shape recognition," Proc. SPIE Document Recognition II, pp. 270-277, 1995.

[6] F. R. Chen, L. D. Wilcox, and D. S. Bloomberg, "A comparison of discrete and continuous hidden Markov models for phrase spotting in text images," Proc. Document Analysis and Recognition, Vol.1, pp.398-402, 1995.

[7] C. L. Tan, W. Huang, Z. Yu, and Y. Xu, "Image document text retrieval without OCR," IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol.24, No.7, pp.838-844, July, 2002.

[8] Y. Lu and C.L. Tan, "Chinese word searching in imaged documents," International Journal of Pattern Recognition and Artificial Intelligence, Vol.18, No.2, pp.229-246, 2004.

[9] 김혜금, 양진호, 이진선, 오일석, "웨이브렛을 이용한 영상기반 인쇄 한글 단어 검색", 한국정보과학회 논문지, 제28권 제2호, pp.91-103, 2001.

[10] I. S. Oh, Y. S. Choi, J. H. Yang, and S. H. Kim, "A keyword spotting system of Korean document images," Proc. 5th International Conference on Asian Digital Libraries, Singapore, p.530, Dec., 2002.

[11] Y. Lu, "Machine printed character segmentation-An overview," Pattern Recognition, Vol.28, No.1, pp.67-80, 1995.

[12] R. G. Casey and E. Lecolinet, "A survey of methods and strategies in character segmentation," IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol.18, No.7, pp. 690-706, July, 1996.

[13] 김우성, 이기돈, 문승원, 유신재, 최명구, 김민수, "오프라인 인쇄체 문자 인식기 개발", 한국과학기술정보연구원, 1997년 12월.

[14] Y. Lu, B. Haist, L. Harmon, J. Trenkle, and R. Vogt, "An accurate and efficient system for segmenting machine-printed text," Postal Service 5th Advanced Technology Conference, Washington D. C, Nov., Vol.3, pp.A-93 to A-105, 1992.

[15] S. Liang, M. Ahmadi, and M. Shridhard, "Segmentation of touching characters in printed document recognition," Proc. 2nd International Conference on Document Analysis and Recognition, pp.569-572, Oct., 1993.

[16] 이근수, "퍼지 추론을 이용한 인쇄체 한글 인식," 숭실대학교 전자계산학과 박사학위논문, 1993.

[17] 구건서, "비디오 영상 정보 검색을 위한 문자 추출 및 인식", 컴퓨터산업교육기술학회논문지, Vol.3, No.7, pp.901-914, 2002.

[18] H. H. Kuo and J. F. Wang, "A new method for the segmentation of mixed handprinted Chinese/English characters," Proc. 2nd International Conference on Document Analysis and Recognition, pp.810-813, Oct., 1993.

[19] 김광백, 김영주, "다해상도 영상과 개선된 RBF 네트워크를 이용한 계층적 영문 명함 인식", 정보처리학회논문지B, Vol. 10, No.4, pp.443-450, 2003.

[20] N. W. Strathy, C. Y. Suen, and A. Krzyzak, "Segmentation of handwritten digits using contour features," Proc. 2nd International Conference on Document Analysis and Recognition, pp.577-580, Oct., 1993.

[21] M. C. Jung, Y. C. Shin, and S. N. Srihari, "Machine printed character segmentation method using side profiles," Proc. IEEE International Conference on Systems, Man, Cybernetics, Vol.6, pp.863-867, 1999.

[22] C. B. Jeong and S. H. Kim, "A document image preprocessing system for keyword spotting," Proc. International Conference on Asian Digital Libraries, China, pp.440-443, Dec., 2004.

박 상 철

e-mail : sanchun@iip.chonnam.ac.kr

1999년 조선대학교 전자계산학과(학사)

2001년 조선대학교 전자계산학과 (이학석사)

2006년 전남대학교 전산학과(이학박사)

2006년~현재 전남대학교 정보통신연구소 연구원

관심분야 : 패턴인식, 문서영상 정보검색, 의료영상

김 수 형

e-mail : shkim@chonnam.ac.kr

1986년 서울대학교 컴퓨터공학과(학사)

1988년 한국과학기술원 전산학과(공학석사)

1993년 한국과학기술원 전산학과(공학박사)

1993년~1996년 삼성전자 멀티미디어연구소 선임연구원

1997년~현재 전남대학교 전자컴퓨터정보통신공학부 교수

관심분야 : 인공지능, 패턴인식, 문서영상 정보검색, 유비쿼터스 컴퓨팅