論文 97-22-11-18

연결특성함수를 이용한 문서화상에서의 영역 분리와 문자열 추출

正會員 김 석 태*, 이 대 원*, 박 찬 용**, 남궁 재 찬***

Segmentation of Region & Extraction of Strings Using Connection-Characteristic function

Seok-Tae Kim*, Dae-Won Lee*, Chan-Yong Park**,
Jae-Chan Namgung*** *Regular Members*

※본 연구는 학술진흥재단 1996년도 연구비 지원(과제번호 : E-0178)에 의해 수행되었음

요 약

본 논문은 문서화상내의 연결성분의 구조적 특징을 이용해 사진 및 그래픽 등이 혼합되어 있는 문서에서 사진 및 그래픽영역과 문자영역을 분리해 내고, 문자영역에서 각 문자열을 찾아내는 방법을 제안한다. 사진 및 그래픽 영역의 분할에는 문서 화상내의 연결성분이 가지는 구조적 특징을 이용해 연결특성함수를 작성하고, 이에따라 분리처리 하였다. 문자열을 찾을 때는 입력 화상내 평균 연결성분의 길이의 4분의 1의 크기로 기본단위영역을 구성하고, 이를 다시 주어진 결합 임계값보다 작은 거리의 기본단위영역들을 결합해 단어를 생성한다. 그 후, 비슷한 크기와 기울기를 가진 단어들을 병합해 하나의 완성된 문자열을 추출해낸다. 이 방법은 문자의 크기가 다양하거나 문자열이 기울어진 경우에도 정확한 추출을 할 수가 있었다. 다양한 형식을 가진 문서를 대상으로 실험을 한 결과, 본 논문의 유용성을 확인하였다.

ABSTRACT

This paper describes a method for region segmentation and string extraction in documents which are mixed with text, graphic and picture images by the use of the structural characteristic of connected components.

*부경대학교 정보통신공학과
**시스템공학연구소
***광운대학교 컴퓨터공학과
論文番號:97174-0524
接受日字:1997年 5月 24日

In segmentation of non-text regions, with connection-characteristic functions which are made by structural characteristic of connected components, segmentation process is progressed. In the string extraction, first we organize basic-unit-region of which vertical and horizontal length are 1/4 of average length of connection components. Second, by merging the basic-unit-regions one other that have smaller values than a given connection intensity threshold. Third, by linking the word blocks with similar block angles, initial strings are created. Finally the whole strings are generated by merging remaining word blocks whose angles are not decided, if their height and position are similar to the initial strings. This method can extract strings that are neither horizontal nor of various character sizes.

Through computer experiments with different style documents, we have shown that the feasibility of our method successes.

Ⅰ. 서 론

최근 들어 정보통신 기술의 급속한 발달로 정보 전달의 매체가 기존의 종이를 이용한 인쇄매체에서 컴퓨터와 인터넷 등의 정보 통신망으로 바뀌어 가고 있다. 이같은 추세에 의해 지금까지 존재해 온 많은 양의 인쇄된 문서를 컴퓨터가 인식해 새로운 정보 전달 및 저장 매체로 전환하는 기술이 필요하게 되었다. 이와 같은 기술은 컴퓨터가 단순히 각 개별 문자를 인식하는 차원을 넘어 문서가 가지는 내용(그림의 크기 및 위치, 표의 인식, 제목, 저자 등의 인식등)까지 인식해 향후 문서 인식 시스템(Document Analysis System)의 개발에 핵심적인 기술이 된다. 문서인식 시스템이 개발되면 사용자가 원하는 부분만 파일로 복사하거나, 기존의 형식을 새로운 형식으로 재편집도 가능해져 방대한 양의 문서를 취급하는 도서관 등의 전산화에 유용하게 쓰일 수 있다[1-3].

문서 인식은 먼저 문서화상을 문자영역과 비문자영역으로 분할한다. 분할된 각 영역중 비문자영역은 영역의 성격에 맞게 압축 등의 처리를 하고, 문자영역에서 각 문자열을 찾고 개별 문자를 인식하는 단계를 거친다. 마지막으로 문서의 전체적인 구조(Layout)를 파악하는 분석이 이루어져 이들 정보를 모아 다시 재구성하는 단계를 가지게 된다[4]. 현재까지 관련된 연구는 각 개별 문자의 인식에 관한 연구에 집중되어 상대적으로 영역 분할 및 구조 분석에 대한 연구는 그 중요성에 비해 미약한 실정이다. 잡지나 신문처럼 현존하는 대부분의 문서는 매우 다양한 형식을 갖고, 그 양도 방대하기 때문에 이에 대한 적응력을 갖는 영역 분할법의 개발은 문서인식의 전체 성능을 좌우할 정도로 중요하며, 기본적이라 할 수 있다.

지금까지 문서의 영역 분할 및 문자열 추출은 함께 연구되어 왔는데, 대표적으로 투영 프로파일링(Projection Profiling), RASA(Run Length Smoothing Algorithm), Morphology, Hough transform, 그리고 문서의 배경영역을 이용하는 방법 등이 연구되어 왔다. 투영 프로파일링[5, 6]은 매우 빠른 속도로 분할이 가능하지만 문서내 문자열의 기울기가 수평 또는 수직으로 일정해야 하고, 그래픽 영역이 함께 존재하는 경우에는 그 추출이 어렵다. RLSA[7-9]는 문서에 대한 사전 지식을 필요로 하지는 않지만, 각 문서의 특징에 맞는 임계값을 찾는 것이 가장 어려운 작업이다. Morphology방법[10]은 연산이 매우 간단하고 병렬처리가 가능하나, 연산의 기본이 되는 구조요소와 연산의 횟수를 정의하는데 문서에 대한 많은 사전지식을 필요로 한다. 문서화상의 연결성분을 Hough transform하는 방법[11]은 그림 속의 문자열을 찾을 수 있으나, 처리시간이 길고 한글과 같이 여러개의 연결성분으로 구성되어 있는 경우에는 그 적용이 곤란하다. 문서의 배경 영역을 이용하는 방법[4, 12, 13]도 제안되고 있으나, 최소 문자열의 높이를 알 수 없을 경우에 적용이 어렵고, 문서가 기울어져 있는 경우에 이 방법 또한 추출이 거의 불가능하다는 단점이 있다. 그 외에도 연결화소에 Simulated annealing법을 적용한 코스트 최소화법[14]도 개발되어 곡선의 문자열까지 추출이 가능하나, 이 방법은 처리 시간이 늦고 알고리즘의 구조상 문자열 후보의 조건이 같은 경우 무한 루프에 빠질 가능성을 가진다.

따라서 본 논문에는 그림과 도표 등이 혼합되어 있는 정형화된 문서뿐만 아니라 비정형화된 문서에서 아무런 사전 지식 없이 연결특성함수를 이용해 각 영역들을 정확히 분리해내고, 문자열의 추출시에는 연결특성함수의 각 특징량과 문자영역내의 각 연결성분의 기하학적 특징을 조사해 각 문자의 크기와 문자열의 기울어짐에 관계없이 문자열을 추출해내는 방법을 제안한다. 연결특성함수는 문서 화상내 각 연결성분이 가지는 4가지의 특징량을 변수로 사용하는 함수로서, 이 연결특성함수를 사용하여 문서화상의 비문자영역을 분리한다. 연결성분의 4가지 특징량은 흑화소와 백화소의 교차횟수, 흑화소의 밀도, 연속되는 연결화소의 길이와 연결성분의 크기를 0에서 10까지 표본화시킨 값이다. 사진영역과 그래픽영역의 특징이 구분되기 때문에 각 영역을 추출하는 2개의 연결특성함수를 만든다. 그후, 문서내에서 분리된 사진 및 그래픽 영역을 제외한 부분은 문자영역이 되므로 문자영역만 남은 문서에서 각 연결성분의 위치, 크기를 이용해서 문자열에 기초가 되는 단어를 생성하고, 그 단어의 기울기를 기초하여 문자열을 추출한다. 이 방법은 문서내에 문자열의 크기가 다양하거나 여러 방향의 문자열이 존재해도 문자열을 찾아낸다. 또한 그림 영역이나 표 영역내의 문자나 문자열도 정확히 찾아낼 수가 있다.

Ⅱ에서는 연결특성함수의 각 특징량을 설명하고, Ⅲ에서는 연결특성함수를 이용한 영역분리법과 문자열 추출법을 논한다. Ⅳ에서는 문서에 대한 실험과 고찰을 통해 본 방법의 유용성을 검증한다.

Ⅱ. 연결특성함수

1. 문서의 구획 및 기본단위영역의 구성

스캐너로 입력된 문서 화상은 연산 속도의 감소와 매우 긴 연결성분이 연산에 영향을 미치는 것을 막기 위해서 적당한 크기로 구획을 한다. 구획을 나누지 않을 경우 문서내의 전체 테두리와 같은 길이가 긴 연결성분이 연결성분의 평균 길이를 증가시켜 결과적으로 기본단위영역의 크기도 증가시켜 이후의 처리가 매우 어려워진다. 실험에서는 입력 해상도를 고려해 각 문서를 가로, 세로 각 1인치로 구획한다. 나누어진 i번째 구획을 B_i라고 하면, 전체 문서 D는 식 (1)과 같이 표현되어 진다. 그림 1의 (a)는 원화상을 나타내고 (b)는 구획된 화상을 나타낸다.

$$D=\sum_{i=0}^{n} B_i \quad \text{n은 구획의 갯수} \tag{1}$$

(a)

(b)

그림 1. 문서의 구획
Fig. 1 Partition of document
(a) 원화상 (b) 구획된 화상
(a) Original image (b) Partitioned image

사진 및 그래픽 영역의 분리와 단어생성을 위한 기본 단계로 문서화상의 각 연결성분에 대해 기본단위영역(basic-unit-region)을 구성한다. 이것은 연결성분내 흑화소의 연결된 형태 및 정도에 따라 이웃하는 영역간의 결합 여부를 결정하는 기본단위로 이용하기 위함이다. 기본단위영역의 구성은 문서화상내 연결성분의 평균길이를 기준으로 하는 정사각형 패턴을 이용한다. 연결성분은 각 구획의 왼쪽 상단에서부터 흑화소의 존재를 조사해 흑화소가 존재하면 그 흑화소를 중심으로 8연결된 흑화소의 집합을 말하며, 그 길이는 연결성분을 둘러싸는 사각형의 가로 및 세

로의 길이중 큰 값으로 한다. 연결성분의 평균길이(C_{avg})는 식 (2)를 통해 구해진다. 구획과 구획에 걸쳐지는 연결성분은 독립된 하나의 연결성분으로 간주한다. $C_j(long)$는 i번째 구획내 j번째 연결성분을 둘러싸는 사각형의 가로, 세로 길이 중 큰 값을 나타내고, k_i는 i번째 구획내 연결 성분의 개수이고, n은 전체 문서화상내 구획의 개수을 나타낸다. 기본단위영역의 흑화소의 개수는 기본단위영역에 속하는 흑화소의 개수에 4방향으로 인접한 기본단위영역내의 흑화소의 개수의 평균을 더한 값이다.

$$C_{avg}=\frac{\sum_{i=0}^{n}\sum_{j=0}^{k_i}C_j(long)}{\sum_{i=0}^{n}k_i} \quad (2)$$

기본단위영역은 그 크기가 매우 크면 연결성분이 가지고 있는 특징을 정확히 나타내기가 어렵게 되고, 반대로 너무 작으면 기본단위영역이 포함하는 데이터의 내용이 너무 작아 특징 파라메타의 추출이 어렵게 된다. 따라서 본 논문에서 실험적으로 연결성분 평균길이(C_{avg})의 1/4의 값으로 기본단위영역의 가로와 세로 길이로 정하였다. 그림 2는 문서화상 중 "탈피하고"의 기본단위영역(흑색 사각형)을 구성한 것이다.

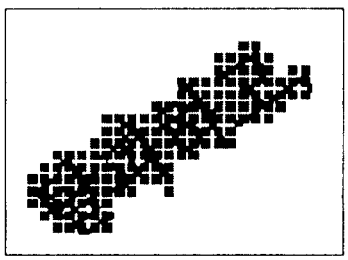

그림 2. 기본단위영역
Fig. 2 Basic unit region

2. 연결특성함수의 각 변수

2.1 교차횟수 특징(Chcross)

교차횟수란 하나의 연결성분에서 가로와 세로 방향으로 각 화소를 조사하여 그 방향으로 백화소에서 흑화소로 또는 백화소에서 흑화소로 바뀐 횟수를 말한다. 교차횟수의 특징(Ch_{cross})이란 한 연결성분내에서 각 방향의 교차횟수의 합을 구해, 그 중에서 큰 값을 0에서 10까지 정규화시킨 값이다. 교차횟수의 특징은 연결성분의 복잡성을 나타내는데, 그래픽 영역의 경우에는 교차횟수가 매우 적은 값으로 나타나지만, 문자나 사진 영역의 경우에는 그 값이 일정하게 나타나지 않는다.

2.2 연결화소 길이 특징(Ch_{run})

연결화소는 연결성분의 가로 및 세로 방향으로 각 화소를 조사해 흑화소가 존재하면 그 방향으로 연속적으로 연결된 흑화소를 말한다. 연결화소의 길이 특징(Ch_{run})은 한 연결성분내의 전체 연결화소에서 그 길이가 4픽셀보다 작은 연결화소의 비율을 의미한다. 문자영역의 경우 연결화소의 길이가 거의 대부분 4픽셀보다 작은 것으로 조사되어 실험적으로 임계값을 4픽셀로 설정하였다. 그림 3에 "大"자의 연결화소의 길이를 나타낸다. ①은 왼쪽에서 오른쪽으로, 위에서 아래로 주사할 때 처음으로 만나는 흑화소에서 동일방향으로 연결된 흑화소의 개수를 나타내고 ②는 두번째로 백에서 흑으로 바뀌는 흑화소에서 동일방향으로 연결된 흑화소의 개수를 나타낸다. 그림 3의 방향별 화소에 개수는 각각 42, 22, 12, 12이다. 연결화소의 길이 특징값은 식 (3)을 이용해 구한다. 그림 3에서의 연결화소의 길이 특징값은 전체 88개의 연결화소의 길이중 4픽셀보다 작은 연결화소가 19개이기에 2.16의 값을 갖는다.

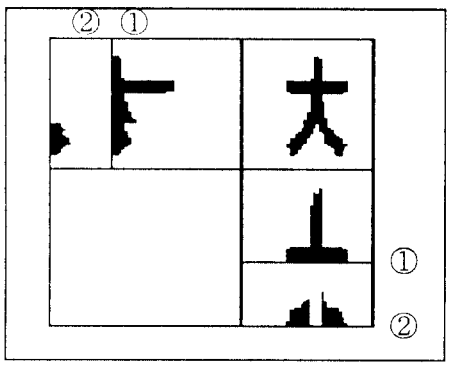

그림 3. 연결화소의 길이
Fig. 3 Run length of connected pixel

$$Ch^{k}_{run} = \frac{\sum Run^{k}_{4}}{\sum Run^{\max}_{k}} \times 10 \quad (3)$$

Ch^{k}_{run}는 하나의 구획내 k번째 연결성분의 연결 화소의 길이 특징을 나타낸다. $\sum Rum^{4}_{k}$는 k번째 연결성분중 연결 화소의 길이가 4픽셀보다 작은 연결 화소의 수를 나타내고, $\sum Run^{ma}_{k}$은 k번째 연결 성분 중 전체 연결화소의 수를 나타낸다. 이 연산은 하나의 연결성분을 기준으로 수평과 수직 방향으로 2번 행해진다. 표나 그래픽 영역인 경우 짧은 연결화소가 다수 존재하기 때문에 그 특징값이 상대적으로 높이 나타난다. 사진이나 문자의 경우에는 흑화소의 길이가 균일하지 않기 때문에 그 특징값도 일정하지가 않다.

2.3 기본단위영역 밀도 특징(Ch_{denB1})

기본단위영역 밀도 특징(Ch_{denB1})은 하나의 연결성분의 가로와 세로의 길이중 큰 길이를 기준으로 만든 정사각형 내에 기본단위영역의 개수이다. 정사각형으로 나타내면 한 방향으로 긴 연결성을 이루는 선성분인 경우도 밀도의 정도를 나타낼 수 있고, 기울기를 가지고 있는 선성분인 경우에도 기울기를 찾기가 용이해진다. 또한 한 변의 길이가 매우 작은 경우 상대적으로 기본단위영역의 밀도 특징이 매우 커지는 것을 방지한다. 따라서 선성분인 경우에 기본단위영역의 밀도는 낮게 나타나고 밀도 특징은 커진다. 식 (4)를 이용해 기본단위영역의 밀도 특징을 구한다.

$$Ch_{denB1} = \frac{\sum_{x,y} |BUR^{x,y}_{black}|}{(CC_j(len))^2} \times 10 \quad (4)$$

$BUR^{x,y}_{black}$는 j번째 연결성분에서 각 기본단위영역의 흑화소의 양을 나타내며, $|BUR^{x,y}_{black}|$은 $BUR^{x,y}_{black}$이 1보다 크면 1의 값을 가지고, 그렇지 않으면 0의 값을 가진다. 그러므로 $\sum_{x,y} |BUR^{x,y}_{black}|$은 연결 성분내의 기본단위 영역의 개수의 합을 나타낸다. $CC_j(len)$은 j번째 연결성분의 가로 세로 길이 중 큰값을 나타낸다.

2.4 기본단위영역내 흑화소 밀도 특징(Ch_{denB2})

기본단위영역내 흑화소 밀도 특징(Ch_{denB2})은 기본단위영역이 포함하는 흑화소의 개수를 말한다. 이 특징값은 흑화소의 밀집 정도를 나타내기 때문에 사진영역에서는 다른 영역보다 상대적으로 큰 값을 나타낸다. 그래픽 및 문자영역에서는 그 특징값이 매우 다양하게 분포된다.

3. 연결특성함수의 작성

표 1은 신문, 잡지, 논문, 지도 등의 100개의 서로 다른 형식을 가진 문서를 대상으로 하여 앞서 계산한 연결특성함수의 변수로 사용되는 각 특징량의 평균과 분산을 나타낸다. 표 1과 같이 연결화소의 길이 특징, 기본단위영역의 밀도 특징은 그래픽 영역에서 다른 영역에 비해 상대적으로 큰 값을 보이고, 기본단위영역내의 흑화소 밀도 특징은 사진영역에서 큰 값을 나타내었다. 즉 다양한 형태의 문서에서 각 문자영역과 사진 및 그래픽 영역들의 연결특성함수의 값이 영역별로 비슷하게 분포됨을 알 수 있다. 그러므로 각 영역을 분리를 위해 식 (5)와 (6)과 같이 연결특성함수를 작성하였다. 식 (5)는 그래픽영역의 분리를 위한 연결특성함수이고, 식 (6)은 사진영역의 분리

표 1. 각 영역별 특징값의 비교
Table 1. Comparison of each region for value of characteristic

	문자		그래픽영역		사진영역	
	평균	분산	평균	분산	평균	분산
교차횟수 특징	6.67	1.12	9.30	0.18	7.99	2.60
연결화소의 길이 특징	8.81	2.85	9.49	0.24	2.98	6.53
기본단위영역의 밀도 특징	7.01	2.59	9.36	0.28	6.67	3.39
기본단위영역내의 흑화소의 밀도 특징	4.5	1.23	5.37	1.86	8.90	0.41

를 위한 연결특성함수를 나타낸다.

그래픽영역은 문서내의 사진을 제외한 모든 그림 이미지와 내부의 문자를 제외한 표영역을 말한다. 사진영역은 문서내 흑화소의 밀도가 높은 사진 등과 같은 영역을 말한다. 그래픽영역은 거의 대부분이 어느 한 방향으로 매우 긴 연결특징과 이웃 화소와의 일정한 연결성을 가지며, 사진영역은 영역내 흑화소의 밀도가 매우 크게 나타나고, 이웃 픽셀과 여러 방향으로 연결된 특징을 나타낸다.

$$F_{Grap} = Ch_{cross} + Ch_{run} + Ch_{denB1} \quad (5)$$

$$F_{Photo} = Ch_{denB2} \quad (6)$$

Ⅲ. 영역 분리와 문자열의 추출

1. 영역 분리

그래픽영역의 분리는 교차횟수 특징, 연결화소의 길이 특징, 기본단위영역의 밀도 특징량을 변수로 사용한 식 (5)와 같은 연결특성함수(F_{Grap})를 이용하여 그 함수 값이 25이상인 연결성분을 원화상에서 분리한다. 사진영역의 분리는 흑화소의 밀도특징량을 이용해 식 (6)과 같은 연결특성함수(F_{Photo})를 만들어 그 값이 8이상인 연결성분을 대상으로 분리한다.

일반적인 문서에서 연결성분의 평균길이는 문자영역의 연결성분에 의해 좌우되고, 사진 및 그래픽 영역의 연결성분의 길이는 문자영역의 연결성분의 평균길이 보다 크게 나타난다. 따라서 문자영역이나 기호, 잡음등이 사진 및 그래픽 영역으로 분리되는 것을 막기 위해 각 연결성분의 가로,세로의 길이가 연결성분의 평균길이(C_{avg})의 4배보다 큰 경우에 연결특성함수(F_{Grap}, F_{Photo})를 적용한다. 그림 1의 경우 그래픽영역으로 분리된 연결성분의 F_{Grap}의 평균값은 26.45로 나타났고, 사진영역으로 분리된 연결성분의 F_{Photo} 값의 평균은 9.14로 나타났다.

2. 문자열 추출

문자열 추출은 사진 및 그래픽영역이 분리되어 문자영역만 남게된 문서를 대상으로 각 문자열을 추출한다. 그림 4는 문자열 추출의 전체적인 순서를 나타낸다. 문자열 추출의 순서는 기본단위영역을 토대로 단어를 생성하고, 생성된 단어의 기울기와 간격을 이용하여 단어를 결합해 초기 문자열을 구성한다. 초기 문자열이 구성되면 이를 기준으로 기울기가 결정되지 않은 단어들을 해당되는 각 문자열로 병합해 최종적인 문자열을 구성하게 된다. 이 과정에서 각 단어 사이의 간격을 조사해 기준값 이상의 간격이 발견되면 두 개의 문자열로 나눈다.

2.1 단어의 생성

단어의 생성은 4방향으로 최대 결합 임계값 이하의 거리에서 기본단위영역의 결합으로 이루어진다. 결합 임계값은 각 연결성분 사이의 간격중 연결성분의 평균길이 (C_{avg})보다 큰 간격들의 평균을 구해 0에서부터 16단계로 정규화 시킨 값이다. 결합 임계값을 점차 증가시키면서 기본단위영역을 결합하기 때문에 문자열간의 결합이 생기기 전에 단어 생성이 된다. 실험에서 문자간의 간격이 작을 때는 결합 임계값이 0단계에서 7단계까지에서 대부분의 단어 생성이 되고, 문자간의 간격이 클 때는 8부터 15단계 사이에서 대부분의 단어가 생성이 된다.

생성된 단어는 그림 5에서 보는 바와 같이 θ_1과 θ_2의 평균을 구해 단어의 기울기 W_θ^i를 결정하게 된다. 단어 생성과정에서 단어의 기울기를 결정할 수 없는 경우가 생기기도 하는데, 이는 기본단위영역이 제대로 결합되지 않아 완전한 단어로 생성되지 않거나, 다른 연결성분보다 그 크기가 상대적으로 크거나, 연결성분을 나타내는 가로, 세로 선분이 그것을 둘러싸

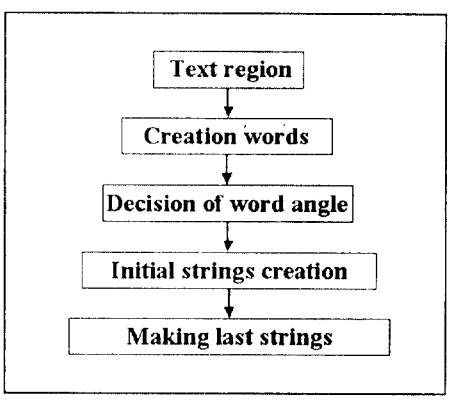

그림 4. 문자열 추출 블록다이어 그램
Fig. 4 Blockdiagram of string extraction

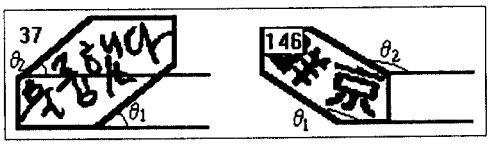

그림 5. 단어의 기울기 결정
Fig. 5 Decision of angle of words

는 사각형과 거의 일치할 때이다. 따라서 본 논문에서는 생성된 단어의 크기가 연결성분의 평균길이(C_{avg})의 3배보다 작거나 생성된 단어가 3개 미만의 연결성분으로 구성되거나, 연결성분을 둘러싸는 사각형이 연결성분을 나타내는 사각형과 같을 때를 단어의 기울기를 결정할 수 없는 것으로 보았다. 이런 경우의 단어들에 대해서는 단어의 결합이 해체되고, 다시 결합 임계값을 증가시켜 단어를 생성한 후, 동일한 방법으로 기울기를 결정하는 작업이 반복된다. 이 작업은 결합 임계값이 최대 결합 임계값보다 작을 때까지 반복된다.

2.2 초기 문자열 구성

문서내의 각각의 문자열은 거의 대부분 직선적인 선형함수 특성을 가지고 있으므로 하나의 문자열을 이루는 각 단어들의 기울기 차이가 매우 작고, 같은 문자열에 속하는 각 단어들의 높이는 대체로 일정한 특징을 가진다. 그러므로 기울기가 결정된 단어를 기준으로 하여 다음 세 가지 조건들을 이용해 초기 문자열을 생성한다. 기울기가 결정된 단어를 W_i, W_j라고 하면,

① $|W_i(angle) - W_j(angle)| < 8°$
② $|W_i(word\_height) - W_j(word\_height)| < C_{avg}$
③어느 한 단어가 같은 문자열에 속하는 단어의 중심을 잇는 직선과의 거리가 두 단어의 높이의 평균보다 작을 때 이 단어는 같은 문자열에 속하게 된다.

각 조건들의 각 파라메타들은 실험을 통하여 얻어진 값이며, 이들 각 조건을 만족하는 단어들은 같은 문자열을 이루는 초기 문자열이 된다. 그림 6은 그림 1의 원화상 중 일부분으로 위의 조건으로 생성된 초기 문자열을 나타내고 있다. 단어를 나타내는 사각형 왼쪽 상단의 숫자는 그 단어의 기울기를 나타낸다. "매직포토 동화면도 정지화면도"은 위의 조건 ①② 에 의해 연결되었고, "선명하게"는 조건③에 의해 초기 문자열로 구성된 모습을 보여준다.

그림 6. 초기 문자열의 생성
Fig. 6 Initial string

2.3 최종 문자열 구성

초기 문자열을 구성한 후 기울기가 결정되지 않은 단어들을 각 문자열에 병합하는 작업을 한다. 이와 같은 단어의 기울기가 결정되지 않은 단어는 단어와 단어 사이에 존재하기도 하고, 단어의 양끝 즉, 문자열의 가장자리에 존재하기 때문에 이들 모두 문자열에 병합할 수 있는 작업이 필요하다. 또한 원래 두 개의 문자열인데 초기 문자열 병합 과정에서 하나의 문자열로 된 문자열을 두 개의 문자열로 분리하는 작업을 거치게 되면 최종적인 문자열을 추출할 수 있게 된다. 알고리즘의 구조상 일단 하나의 문자열 사이에 존재하는 기울기가 결정되지 않은 단어를 먼저 문자열로 병합하고, 생성된 문자열을 분석해 두 개 이상의 문자열로 나뉘어야 할 문자열을 분리시킨 후, 각 문자열의 가장자리에 있는 단어들을 문자열로 병합한다. 최종 문자열의 구성은 다음의 3단계를 거치게 된다.

단계 1:한 문자열을 이루는 단어와 단어 사이에 기울기가 결정되지 않은 단어가 존재할 경우 그 단어의 높이가 양쪽 단어의 높이의 평균보다 작고, 양쪽 단어의 중심과 기울기가 결정되지 않은 단어의 중심과의 거리가 각 단어의 높이의 평균보다 작을 때 그 단어는 하나의 문자열에 속하게 된다.

단계 2:초기 문자열을 구성하는 과정에서 원화상에서는 다른 문자열이 하나의 문자열로 잘못 구성되어지는 경우가 생길 수 있다. 이를 분리하기 위해 문서의 각 문자열내의 각 단어들의 간격은 거의 일정하므로 이 특징을 이용해 각 단어와의 간격을 조사하여

기준값 이상의 간격을 보이는 단어와 단어사이를 두 개의 문자열로 나눈다. 본 논문에서는 이 기준값을 실험적으로 구해 단어 높이의 1.7배로 하였다.

단계 3:문자열 양끝에 기울기가 결정되지 않은 단어를 결합한다. 결합 방법은 초기문자열 생성시에 쓰는 방법을 그대로 적용하여 문자열을 결합해서 최종적인 문자열을 완성하게 된다.

각 단계를 거쳐 최종적인 문자열을 그림 7에서 나타내었다. 그림 7도 그림 1에서 본 원화상의 일부로 (a)는 기울기가 결정된 단어를 하나의 문자열로 나타낸 초기 문자열을 나타내었다. 그림 6과 같이 기울기가 결정된 단어는 좌측 상단에 기울기가 표시되었다. "다채널" 이나 "멀티"는 문자열 가장자리에 위치한 기울기가 결정되지 않은 단어이다. 이들 단어들은 단어생성 단계의 3가지의 조건에 모두 만족하지 않아 단어생성이 되지 않았다. (b)는 이들 기울기가 결정되지 않은 단어를 한 문자열로 최종 문자열로 구성된 모습을 보여준다. 중앙의 직선은 각 단어의 중심을 이은 선으로 한 문자열임을 나타낸다.

(a)

(b)

그림 7. 최종 문자열 구성
Fig. 7 Making last string
(a) 단어의 생성 (b) 최종문자열
(a) Creation words, (b) last string

Ⅳ. 실험 및 고찰

그림과 표가 혼합되어 있고, 여러 크기의 문자열이 존재하는 신문, 잡지, 논문, 주간지, 광고지, 지도 그래프 등의 100여개의 문서를 대상으로 실험을 하였다. 하나의 문서에 평균 20개 정도의 문자열과 150개 정도의 문자가 존재하고 문서에 따라 그림과 표등이 다양한 위치와 기울기를 가지고 함께 존재하였다. 그 중 일부는 스캐너로 입력을 받을 때 임의로 기울기를 주어 기울기에 대한 유용성도 함께 검토하였다. 각 문서는 250dpi와 300dpi로 입력을 받고 pentium 133㎒를 이용하여 Windows95 환경에서 Boland C^{++}를 이용하였다. 전체 100개의 대상 화상을 이용해 실험을 한 결과, 문자열 추출율은 97.34%의 추출율을 보였고 문자열 추출 정확도는 약 94%를 보였다. 문자열 추출 정확도는 실제 하나의 문자열이 두 개의 문자열로 추출되거나 두 개의 문자열이 하나의 문자열로 추출되는 경우에 대해 실제 문자열과 비교 계산한 것이다. 추출율과 추출 정확도는 목시(目視)에 의해 평가하였다. 한 화상을 처리하는데 걸리는 시간은 1026 × 718 크기의 입력화상을 기준으로 하였을 때, 평균 18.6초였다.

그림 8의 각 그림은 잡지(Ⅰ), 광고(Ⅱ), 지도(Ⅲ), 논문(Ⅳ), 그래프(Ⅴ, Ⅵ)등을 대상으로 한 영역분리 및 문자열 추출의 결과를 나타낸다. (a)는 원화상을 나타내고, (b)는 사진 및 그래픽 영역이 분리된 화상이다. (c)는 문자열이 추출된 결과를 나타낸다. 이러한 문서는 단순한 사영특징이나 수학적 연산을 이용할 경우 복잡한 추출과정과 경험적인 방법이 필요하고 범용성에 대한 한계를 갖는다. 그러나 본 방법은 임계값의 계산이 단순한 4칙연산과 비교 연산이고 분리 결합에 쓰이는 파라메타는 대상화상에서 얻기 때문에 범용성을 갖는다. 특히 광고 전단의 예에서는 문자열의 방향이 불규칙적이고 짧은 문자열이 그림 속에 혼재되어 있음에도 불구하고 양호한 분리와 추출 결과를 얻고 있다. 또한 그래프 화상의 예와 같이 그래프 영역과 문자 영역의 분리와 향후 그래프를 자동으로 재구성하는데 유용하게 사용될 수 있을 것으로 사료된다.

그림 9는 사진 및 그래픽영역의 분리와 문자열의 추출에 실패한 전형적인 예를 나타내었다. 각 그림은 전체 문서에서 오류가 생긴 부분만 선별해서 그림으로 나타내었다. 추출의 오류는 사진 및 그래픽영역이 성확히 분리가 되지 않은 경우와 문자열의 추출이 정확히 되지 않은 경우가 있다.

그림 9의 (a)는 사진 및 그래픽영역의 분리시에 생기는 오류인데, "랜드로바" 앞의 그래픽영역의 크기가 사진영역의 크기 보다 문자영역과 크기가 비슷하고, 연결성분에 따른 화소의 연결특성함수의 특징값

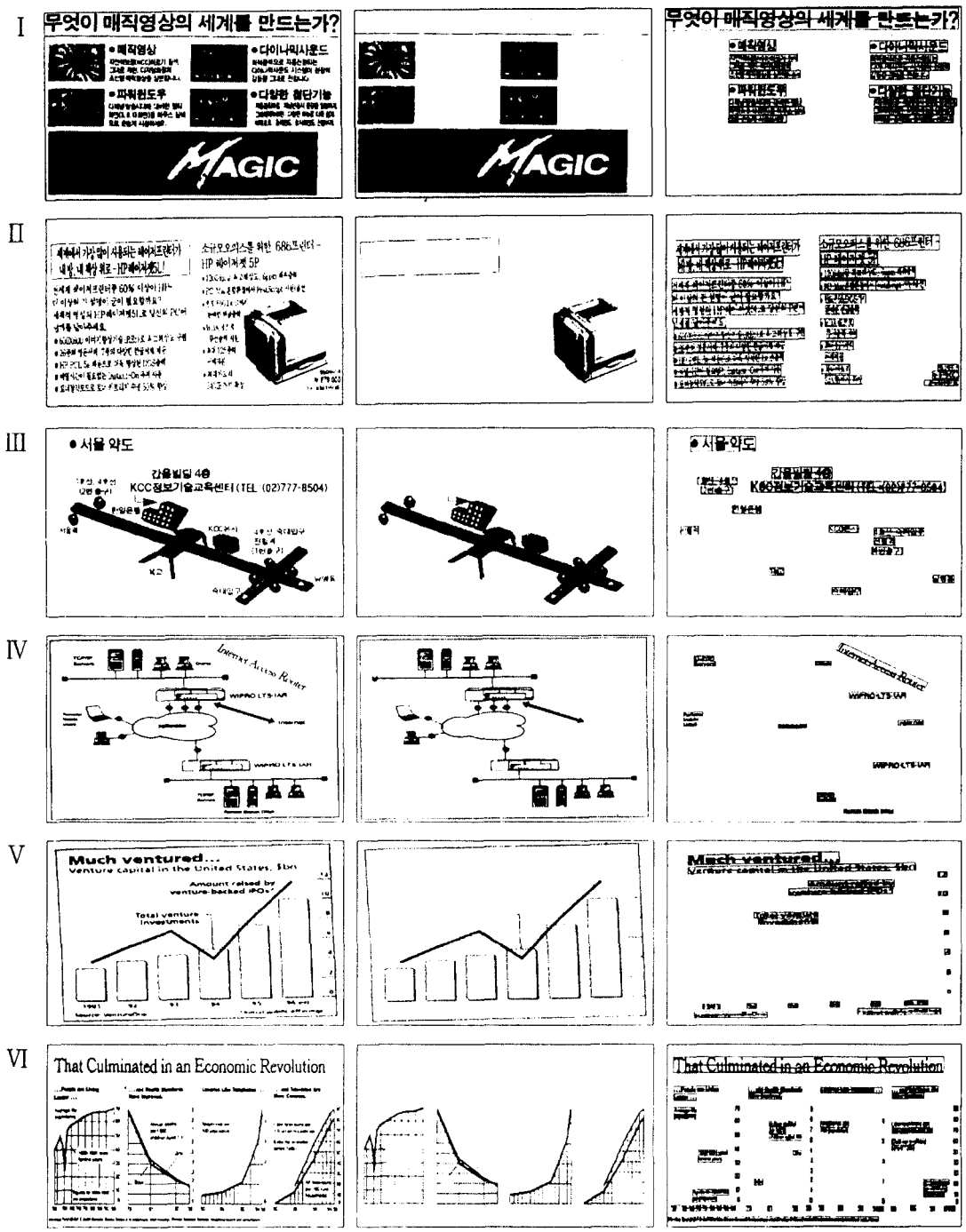

그림 8. 실험 대상 문서

(a) 원화상 (b) 사진 및 그래픽 이미지 분리 (c) 문자열 구성

Fig. 8 Document image for experiment

(a) Original image (b) Segmentation of picture and graphic region (c) Making Strings

이 임계값인 25보다 작아 문자영역으로 남게 된 경우이다. 이와 같은 오류는 문자의 크기와 비슷한 기호나 그래픽영역이 문자열과 인접한 경우 자주 발생하였다. 오류는 일단 하나의 문자열로 추출한 뒤, 문자의 인식 과정에서 인식할 수 없는 문자는 다시 사진 및 그래픽영역으로 전환한 후, 처리하면 대처가 가능하리라 생각된다.

그림 9의 (b)는 (a)와 반대되는 현상으로 단어의 크기가 사진 및 그래픽영역과 비슷하고 문자의 두께가 두꺼워 그림 상단의 "어"와 "위"자의 "ㅟ"가 사진 및 그래픽영역으로 잘못 추출된 경우이다. 사진 및 그래픽 영역의 분리 임계값을 살펴보면 "어"의 경우 기본단위영역내의 흑화소의 밀도 특징값이 9.25이고, "ㅟ"의 경우도 8.34로 이들 연결성분이 사진영역으로 분리되었다. 이러한 오류의 수정은 "ㅟ"와 같은 문자를 분리해 낼 수 있는 새로운 특징값을 찾거나, 문자열에서 각 문자를 분리시 좌우 문자의 위치와, 크기, 형태를 이용해 분리 할 수 있는 방법을 찾아야 한다.

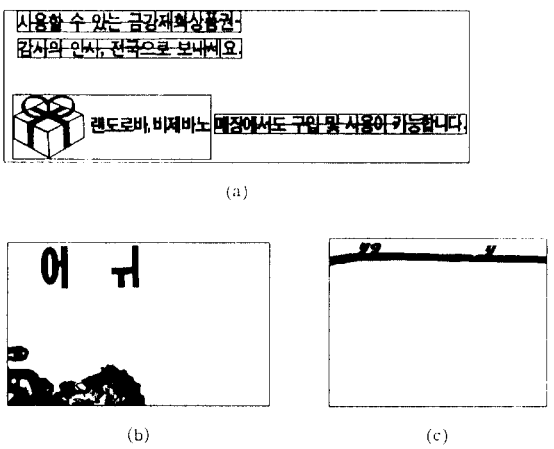

그림 9. 전형적 에러 패턴
Fig. 9 Typical error patterns
(a) 그래픽 영역이 문자영역에 속한 경우
(b) 문자의 크기가 사진 및 그래픽 영역과 비슷한 경우
(c) 문자가 사진 및 그래픽 영역과 붙은 경우
(a) Graphic region belongs to words
(b) Characters's sizes are similar to picture and graphic region
(c) Characters are attached to picture and graphic regions

그림 9의 (c) 경우는 문자가 사진 및 그래픽 영역에 붙어서 사진 및 그래픽 영역으로 함께 분리된 오류로 본 실험에서 가장 많이 나타나는 전형적인 오류이다. 이런 경우에서는 개별 문자 인식후 문맥을 인식하는 단계에서 보충해내는 후처리 작업이 필요하다.

V. 결 론

본 논문에서는 그림과 표가 혼합되어 있는 일반적인 문서에서 그림과 표등이 포함된 사진 및 그래픽 영역을 분리해내고, 나머지 문자영역에서는 각 문자열을 찾아내는 보다 효과적인 방법을 제안하였다. 이 방법은 단순한 연결성분의 위치와 모양, 크기, 흑화소의 밀도 등과 같은 변수를 가지는 연결특성함수를 정의하고, 이에 따라 일반적인 문서에서 사진 및 그래픽영역을 분리하였다. 또한, 문자열을 찾을 때에는 기본단위영역을 적당하게 결합하여 단어를 구성하고 구성된 단어는 문자열이 가지는 특징정보에 따라 결합해 문자열을 찾는다. 그래픽 영역의 분리에 쓰인 분리함수의 결합 임계값은 25이고, 사진 영역의 분리에 쓰인 결합 임계값은 8로 설정하였다. 문자열의 추출은 기본단위영역을 연결성분의 평균 간격을 16단계로 정규화한 결합 임계값을 증가시키면서 단어를 생성하고, 단어의 기울기를 중심으로 문자열을 추출하였다.

본 논문에서는 모든 분리 및 결합에 쓰이는 파라메타의 값이 원화상에서 얻어지며 또 다른 문서가 입력되면 그 문서에 맞게 다시 파라메타를 수정하여 그 값을 정규회 시켜 분리, 추출 작업한다. 그러므로 여러 다양한 형식의 문서에 적응력을 가지고 높은 추출율을 얻게 된다. 이런 처리는 문서의 기울기에도 적응력을 가지기 때문에 별도의 기울기 보정 작업이 필요없고, 알고리즘이 고정적이여서 문서의 형태, 글자의 크기 등이 다양한 형식의 문서에는 적응력을 가지지 못하는 단점을 개선한 것이다.

본 연구에서 3가지의 전형적인 오류 패턴이 발견되었는데, 그래픽영역과 문자의 크기가 비슷하거나 문자영역과 인접해 있는 경우, 문자의 두께가 매우 두꺼워 사진 및 그래픽영역으로 잘못 분리되는 경우, 문자가 사진 및 그래픽 영역에 붙어 존재하는 경우 등이다. 이와 같은 오류는 문서내 주위의 영역과 비

교하거나 문자 인식과정에서 오류를 수정할 수 있다.

앞으로는 문자가 사진 및 그래픽 영역으로 잘못 추출된 경우, 각 문자의 크기와 위치를 이용해 문자열로 수정 보완하는 작업과 각 개별 문자를 인식 단계에서 수정하는 방법과, 화소의 구조적 특징을 좀더 체계적으로 나타내는 연결특성함수에 대한 더 연구가 필요하다. 또한 표의 각 문자와 문자열을 인식하여 새로운 형식으로 표를 생성해내는 등의 연구가 계속될 수 있다. 또한 기본단위영역을 좀 더 종합적이고, 상호보완적인 관계를 고려할 수 있는 새로운 방법의 연구가 필요하다.

참 고 문 헌

1. 이정한, "문자인식:이론과 실제," 서울, 홍릉과학출판사, 1993.
2. H. Ogawa, "패턴인식이해의 새로운 전개," 동경, 전자정보통신학회, 1994.
3. K. Y. Wong et al., "Document analysis system," IBM, J. Res. Develope, Vol. 26, No. 8, pp. 647-656, 1982.
4. S. Tsujimoto and H. Asada, "Major compoments of a complete text reading system," Proc. of the IEEE, Vol. 80, No. 7, pp. 1133-1149, 1992.
5. G. nagy,S Seth, and M. Viswanathan, "A prototype document image analysis system for technical journal," IEEE Computer, Special issue on Documenet Image Analysis System, pp. 10-22, 1992.
6. 이동준, 이성환, "명암 문자열 영상의 지형적 특징을 이용한 비선형 문자 분할 및 인식," 정보과학회 논문지(B) 제22권 11호, pp. 1581-1589, 1995.
7. D. Wang and S N. Srihari, "Classification of newspaper image block using texture analysis," Computer Vision Graphics And Image Processing 47, pp. 327-352, 1989.
8. Fisher J. L., Hinds S. C. and D'Amato D. P., "A rule-based system for document image segmentation," Processing of 10th International Conference on Pattern Recognition, pp. 567-572, 1990.
9. 박영석, "일반적인 문서화상의 영역식별법," 한국정보과학회논문지, 제21권 제5호, pp. 757-767, 1994.
10. 장희돈, 김석태, 남궁재찬, "Morphology를 이용한 문서화상내의 문자열 추출에 관한 연구," 한국통신학회논문지, 제8권 1호, pp. 123-132, 1993.
11. L. A. Fletcher and R. Kasturi, "A robust algorithm for text string separation from mixed text/graphics images," IEEE Trans. Patt. Anal. And Mach. Intell., Vol. 10, No. 6, pp. 910-918, 1988.
12. T. Pavlidis and J. Zhou, "Page segmentation by white streams," International Conference on Document Analysis and Recognition(ICDAR), pp. 945-953, 1991.
13. M. Ozaki, P. Alto, "Column segmenation by White Space Pattern Matching," 3th Proc. International Conference on Document Analysis and Recognition(ICDAR), pp. 134-137, 1995.
14. 김석태, "코스트 최소화법에 의한 문자영역의 추출," 한국정보처리학회 논문지, 제3권 2호, pp. 348-358, 1996.

김 석 태(Seok-tae Kim) 정회원

1983년:光云大學校 電子工學科 卒業(工學士)

1988년:京都工藝纖維大學 電子工學科 卒業(工學碩士)

1991년:大阪大學 通信工學科 卒業(工學博士)

1991년~1996년:釜山水產大學校 情報通信科 助教授

1996년~현재:釜慶大學校 情報通信工學科 副教授

※관심분야:화상처리, 패턴인식, 멀티미디어통신, 지적CAI등

이 대 원(Dae-wom Lee) 정회원

1995년:釜慶大學校 電子工學科 卒業(工學士)

1997년:釜慶大學校 電子工學科 卒業(工學碩士)

※관심분야:화상처리,패턴인식, 문서인식

박 찬 용(Chan-yong Park) 정회원
1994년 : 光云大學校 컴퓨터工學科 卒業(工學士)
1996년 : 光云大學校 컴퓨터工學科 卒業(工學碩士)
1996년~현재 : 시스템 공학연구소 연구원
※관심분야 : 문서인식, 컴퓨터 그래픽, 가상현실

남 궁 재 찬(Jae-chan Namgung) 정회원
한국통신학회논문지 제21권 12호 참조 (1996년)