문서 영상의 그림 영역에서 효과적인 단어 영상 추출에 관한 연구

정창부*, 김수형**
*호남대학교 인터넷소프트웨어학과
**전남대학교 전산학과
e-mail:cbjeong@honam.ac.kr

A Study on an Efficient method of Word Decomposition from Document Images

Chang-Bu Jeong*, Soo-Hyung Kim**
*Dept of Internet Software, Honam University
**Dept of Computer Science, Chonnam University

요 약

본 논문에서는 그림 영역에서 단어 영상을 효과적으로 추출하는 방법을 제안한다. 제안 방법은 문자 성분과 그래픽 성분을 분류하기 위하여 구성 원소들의 통계값을 이용하는 상자그림 분석을 응용하고, 분류된 문자 성분들에 대하여 지역적 밀집도를 분석하여 문자 영역을 추출한다. 추출된 문자 영역에서 문자열 및 단어 영상을 추출하는 방법은 투영 히스토그램 분석 등을 적용한다. 제안 방법은 임계치 대신에 그림 영역의 통계값을 이용하였기 때문에 그림의 형태 변화에 민감하지 않으며, 지역적 밀집도 분석으로 보다 정확한 문자 영역을 추출하였다.

1. 서론

문자/그래픽이 혼합된 영상에서 문자열을 추출하는 연구는 일반적인 문서 영상보다는 공학 설계 도면이나 래스터 지도에 대하여 활발히 진행되어 왔으며, 최근에는 동영상의 자막이나 자연영상에서의 문자 추출 등의 연구 분야로 폭넓게 진행되고 있다. 이러한 연구들의 핵심은 영상으로부터 문자 성분과 그래픽 성분을 분리하는 연구와 추출된 문자 성분들을 적절한 단어나 문자열로 구성하기 위한 연구로 구분될 수 있다[1-6].

전자는 문서 영상의 연결 요소 분석으로서 구해진 연결 요소들의 정보(평균 면적, 평균 수직수평비, 밀집도 등)를 이용한 휴리스틱 필터링 방법이 일반적이었다. 반면에 후자는 문자 성분들의 의미있는 연결을 위하여 허프변환, 런-길이 smoothing, 모폴로지 등의 다양한 방법을 이용하여 연구되었으며, 그중에서도 문자열의 속성(문자열의 방향, 폰트 형태, 문자 크기 등)의 다양함에 영향을 덜 받고 좋은 성능을 보이는 허프변환이 많이 이용되었다.

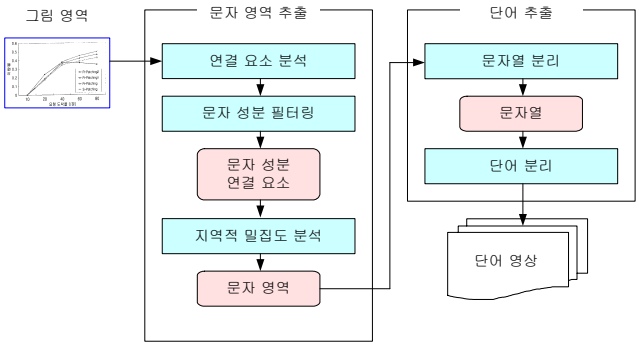
그러나 기존의 방법들은 문서 영상의 종류와 타입에 따라 민감하게 반응할 수밖에 없다. 우선 문자 성분과 그래픽 성분을 분리하는 방법에서 대부분의 연구가 테스트 영상에 적합한 임계값을 사용하고 있기 때문에 다른 종류의 영상에 대해서 문자 성분이 누락되거나 그래픽 성분이 문자 성분으로 잘못 분리되는 오류가 많았다. 또한 문서의 해상도나 획득 과정의 오류로 그래픽 성분의 단락이 발생할 수 있고, 이로 인하여 그래픽 성분이 문자 성분으로 분리되는 오류가 가능하다.

본 논문에서는 문서 영상의 그림 영역에서 단어 영상을 효과적으로 추출하는 방법을 제안한다. 제안 방법은 문자 성분과 그래픽 성분을 분리하기 위하여 임계값을 이용하지 않고 연결요소들의 통계적 특징을 이용하기 때문에 그림 형태의 변화에 민감하지 않고, 지역적 밀집도 분석으로 보다 정확한 문자 영역을 추출이 가능하다.

2. 제안 방법

제안 방법은 문자 영역 추출과 단어영상 추출의 두 단계 과정으로 수행된다. 첫 번째 단계는 연결

요소의 특징에 대한 통계분석의 상자그림을 이용하여 문자 성분을 추출하고, 추출된 문자 성분의 연결 요소에 대한 지역적 밀집도를 분석하여 문자 영역을 결정한다. 두 번째 단계에서는 추출된 문자 영역의 연결 요소를 분석하여 테이블 영역에서의 방법과 동일하게 문자열 분리와 단어단위 분리를 수행한다. 그림 1은 제안 방법의 단계별 수행 과정을 도식화한 것이다.



(그림 1) 제안 방법의 단계별 수행 과정

(1) 문자 성분 추출

그림 영역에서 그래픽 성분에 해당하는 연결 요소는 문자 성분의 것보다 월등하게 크거나 작은 것으로서 연결 요소의 통계적 분석에서는 이상점(outlier) 또는 특이점으로 간주할 수 있다. 이와 같은 이상점들을 추출하기 위하여 탐색적 자료 분석 방법인 상자그림을 이용한다. 상자그림이란 다섯 숫자 요약을 그림으로 그린 것으로, 최소값, 제1사분위수(Q1), 중앙값, 제3사분위수(Q3), 자료의 최대값이 다섯 숫자 요약에 해당된다. 이는 자료의 분포상태를 나타내주며, 특히 Q3과 Q1의 차를 사분위범위(IQR)이라 한다. 즉 $IQR = Q3 - Q1$이다. 그리고 상자그림에서 상자의 좌측 끝과 우측 끝으로부터 각각 IQR의 1.5배 이내의 거리를 안울타리(IF : Inner Fence)라 하고, 3배 이내의 거리를 바깥울타리(OF : Outer Fence)라 한다. 이때 IF와 OF 사이에 있는 관측값을 이상점이라 하며 그 자리에 '*' 표시를 하고, OF 밖에 있는 자료를 특별한 이상점(special outlier)이라 하며 그 자리에 'o' 표시를 한다[7].

문서/그래픽이 혼합된 문서에서 자료의 특징 값을 연결 요소에 대한 BB(bounding box)의 대각선 길이라 하면, 문자들처럼 작은 성분에 비해 소수의 큰 그래픽 성분이나 아주 작은 그래픽 성분(도트나 잡음)은 상자그림의 이상점으로 표현될 것이다. 문자 성분의 추출은 이처럼 이상점으로 간주되는 성분을 제외함으로써, IF 안에 있는 성분들만을 문자 성분으로 간주하고 추출하는 것이다. 그러나 Q1과 Q3이 문자 성분과 같이 크기가 작고 다수인 자료를 표현하기 때문에, 크기가 아주 작은 그래픽 성분(잡음 포함)들도 IF 범위 안에 존재할 수 있다. 그러므로 하위 IF 값을 높이기 위하여

$$IF_{lower} = Q1 - \frac{IQR}{2},$$

으로 수정한다. 그림 2는 상단의 입력 영상에 대하여 문자 성분을 추출한 결과 영상(하)이다.

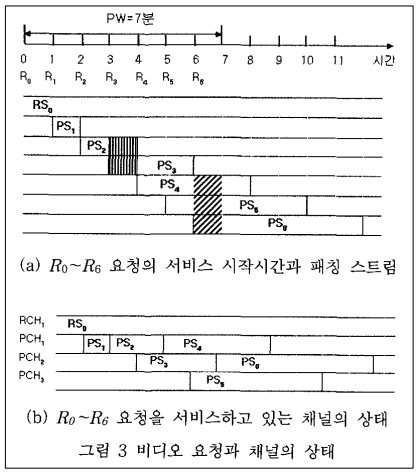


그림 3 비디오 요청과 채널의 상태

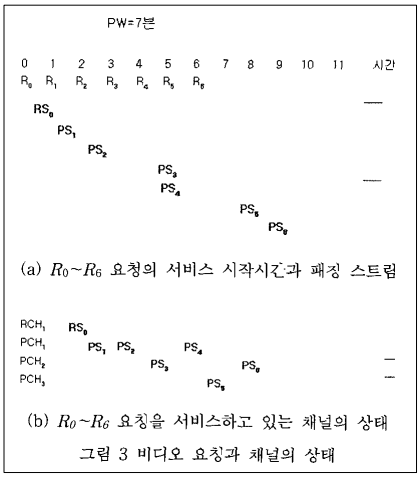


그림 3 비디오 요청과 채널의 상태

(그림 2) 입력 영상(상)과 문자 성분 추출 결과(하)

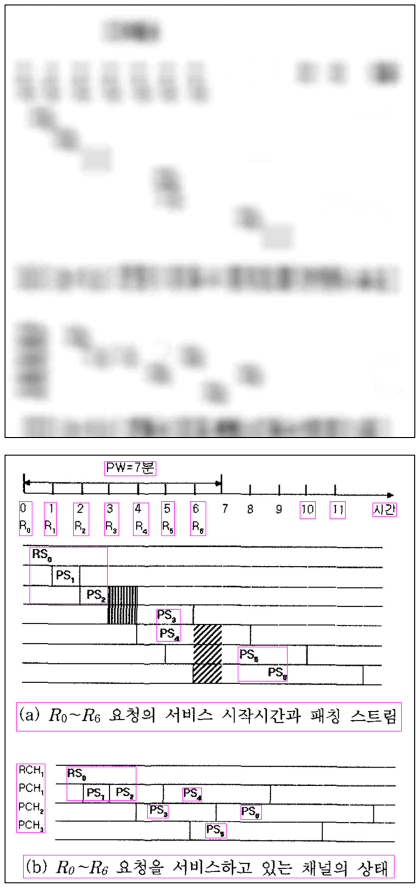
(2) 문자 영역 추출

추출된 문자 성분들로부터 문자 영역의 추출은 [4]에서 제안한 세그멘트의 지역적 밀집도 개념을 응용한다. [4]에서는 문자 성분을 세그멘트로 변환하여 지역적 밀집도를 계산하였지만, 본 논문에서는 문자 성분의 BB를 그대로 이용한다. 영상의 (x,y)에 대한 지역적 밀집도 $D(x,y)$의 계산은 다음과

같다.

$$D(x,y)=\sum_{i=1}^{N} e^{-\frac{di^2}{2\sigma^2}}$$

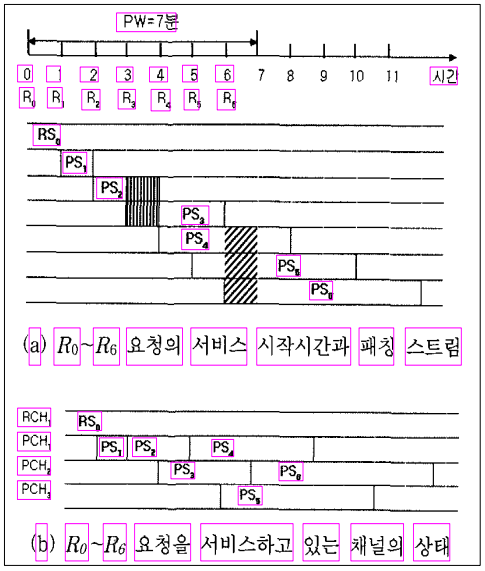
위 식에서 N은 중심이 점 (x,y), 반지름이 T_2 ($T_2=3\sigma$)인 원과 만나는 문자 성분의 BB 개수이며, $e^{-di^2/2\sigma^2}(0<e^{-di^2/2\sigma^2}\leq 1)$은 i번째 만나는 문자 성분의 BB에 대한 가중치이고 di는 점 (x,y)와 BB의 최소 거리를 의미한다. 즉, 가중치는 점 (x,y)와 문자 성분의 BB의 거리가 멀수록 작아진다. 이와 같은 지역적 밀집도 $D(x,y)$은 일정 거리 안에 있는 문자 성분의 BB 개수와 거리가 반영된 것으로서, 픽셀 (x,y)이 문자 영역에 포함되는 지를 결정할 값이다. 그림 3의 상단 영상은 그림 2의 문자 성분에 대하여 지역적 밀집도를 계산한 결과로 문자 영역에 해당되는 픽셀은 문자 영역과는 멀리 떨어진 픽셀에 비하여 검은 색을 보여주고, 하단 영상은 좌측 영상의 지역적 밀집도를 이진화한 후, 그룹핑하여 결정된 문자 영역의 결과이다.



(그림 3) 지역적 밀집도 분석 결과 영상(상)과 문자 영역을 추출한 결과 영상(하)

(3) 문자열 및 단어 분리

추출된 문자 영역을 문자열과 단어로 분리하는 것은 기본적으로 [7]의 방법을 적용하였고, 한 문자로만 구성된 문자열이나 단어에 대한 후처리가 추가하였다. 과다 분리된 문자열 및 단어를 병합하는 후처리의 기준으로 문자 영역에 포함되는 문자 성분에 대한 높이와 폭의 중앙값을 이용하였다. 그림 4는 문자 영역을 문자열 및 단어로 분리한 결과이다.



(그림 4) 문자열 및 단어 분리 결과

3. 실험 결과

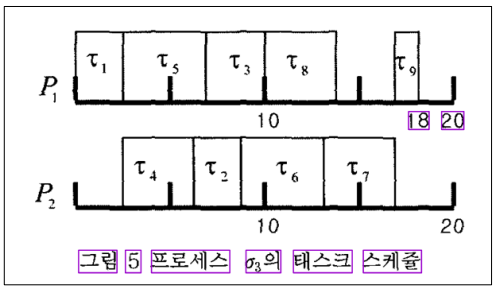
제안 방법의 성능 평가를 위하여 총 55개의 그림 영상에 대하여 실험하였으며, 실험 영상은 300dpi의 이진 영상이다. 표 1은 55개의 그림 영상에서 단어를 추출한 결과로서, 총 1,712개의 단어 중에 1,332개의 단어를 성공적으로 추출하여 74.53%의 단어 추출 성공률을 보여준다.

<표 1> 그림 영역에서의 단어 추출 결과

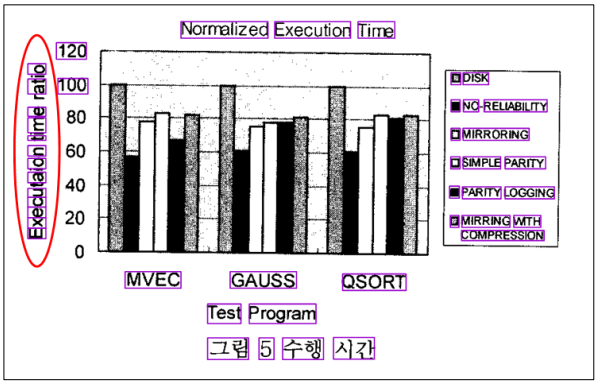
영상 개수	단어 개수	추출 성공한 단어	추출 실패한 단어
55	1,712	1,332개 (77.80%)	380개 (22.20%)

단어 추출의 실패 원인은 5가지 유형으로 분석되었다. 첫 번째 유형은 2개 이하의 연결요소로 구성된 단어 중에서 다른 단어들과 멀리 떨어져 고립된 경우로서, 이는 낮은 지역적 밀집도로 인하여 그래픽 성분으로 오분류되는 경우이다(그림 5). 이에 해당하는 단어는 196개이고, 대부분 문자의 수가 2개 이하로 구성된 단어였다. 두 번째는 단어의 작성 방

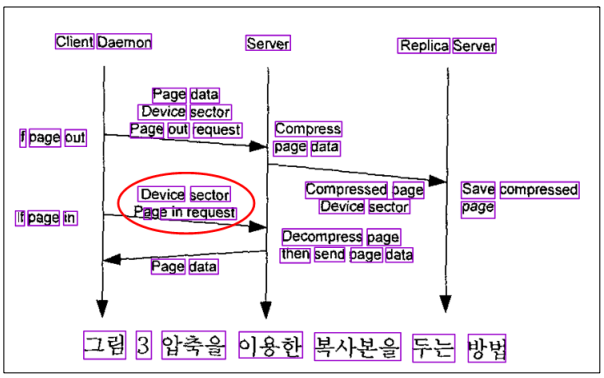
향이 가로가 아닌 세로나 대각선이여서 문자열 분리가 실패하는 오류로서, 총 80개의 단어가 이에 해당하였다(그림 6). 세 번째는 단어의 일부가 그래픽 성분과 접촉하여 그래픽 성분의 일부로 처리되는 오류로서, 총 45개의 단어가 이에 해당하였다(그림 7). 네 번째는 소수점의 제거 등으로 ICG(Inter-Character Gap)로 분류될 갭의 크기가 IWG(Inter-Word Gap)에 더 유사하여 IWG로 오분류되는 군집화의 오류로서, 총 24개의 단어가 이에 해당하였다. 그리고 기타 오류의 원인은 음영이 있는 단어나 밑줄로 연결된 단어 등이 있었다.



(그림 5) 실패 유형 1
- 소수 문자들로 구성된 고립된 단어



(그림 6) 실패 유형 2 - 세로형 단어



(그림 7) 실패 유형 3
- 문자 성분과 그래픽 성분이 붙은 경우

4. 결론

본 논문에서는 그림 영역에서 효과적으로 단어 영상을 추출하기 위하여 통계적 분석 방법인 상자그림을 응용하였고, 지역적 밀집도 분석을 이용하였다. 그러나 문자열 및 단어 분리가 수평으로 제한하였기 때문에 성능이 저하됨을 볼 수 있었다.

향후에 문자열 및 단어의 방향과 무관하게 추출할 수 있는 알고리즘이 개발되어서 문서 영상뿐만 아니라 자연 영상에서의 단어 영상 추출 등에 활용하도록 지속적인 연구가 계속되어야 할 것이다.

참고문헌

[1] 김석태, 이대원, 박찬용, 남궁재찬, “연결특성함수를 이용한 문서화상에서의 영역 분리와 문자열 추출,” 한국통신학회 논문지, Vol. 22, No. 11, pp. 2531-2542, 1997.

[2] Z. Lu, “Detection of Text Regions From Digital Engineering Drawings,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 4, pp. 910-918, April 1998.

[3] C.L. Tan and P.O. Ng, “Text Extraction using Pyramid,” Pattern Recognition, Vol. 31, No. 1, pp. 63-72, 1998.

[4] O. Shiku, K. Kawasue, and A. Nakamura, “A Method for Character String Extraction Using Local and Global Segment Crowdedness,” Proc. International Conference on Pattern Recognition, Vol. 2, pp. 1077-1080, 1998.

[5] K. Tombre, S. Tabbone, L. Pelissier, B. Lamiroy, and P. Dosch, “Text/Graphics Separation Revisited,” LNCS Vol. 2423, pp. 200-211, 2002.

[6] H.C. Park, S.Y. Ok, Y.J. Yu, and H.G. Cho, “A word extraction algorithm for machine-printed documents using a 3D neighborhood graph model,” International Journal of Document Analysis and Recognition, Vol. 4, pp. 115-130, 2001.

[7] 심정욱, 손영숙, 백장선, 수리통계학, 자유아카데미, 제4판, 2003.

[8] 정창부, 김수형, “투영 프로파일, Gap 및 특수기호를 이용한 텍스트 영역의 어절 단위 분할,” 정보과학회논문지: 소프트웨어 및 응용, 제31권, 제9호, pp. 1121-1130, 2004.