문서 영상의 그림 영역에서 통계적 분석을 이용한 단어 영상 추출

정 창 부[†] · 김 수 형[††]

요 약

본 논문에서는 문서 영상의 그림 영역에서 통계적 분석을 통한 단어 영상을 추출하는 방법을 제안한다. 제안 방법은 그림 영역의 구성 요소를 문자 성분과 그래픽 성분으로 분류하기 위하여 연결요소에 대한여 통계적 분석 방법인 상자그림 분석을 적용하고, 분류된 문자 성분들에 대하여 지역적 밀집도를 분석하여 문자 영역을 추출한다. 추출된 문자 영역에서 투영 히스토그램 분석을 통하여 문자열을 추출하고, 문자열을 단어단위 영상으로 분리하기 위하여 투영 히스토그램 분석과 갭 군집화, 특수 기호 검출 등을 수행한다. 제안 방법은 임계값의 사용 대신에 그림 영역의 구성 요소들에 대하여 통계적 분석을 수행하기 때문에 그림의 형태 변화에 민감하지 않으며, 지역적 밀집도 분석으로 보다 정확한 문자 영역을 추출하였다. 또한 제안 방법의 응용 분야인 주제어 검색을 위한 오프라인의 전처리에 해당하는 문서 영상의 단어단위 영상 추출에 적용하여 제안 방법에 대한 연구의 필요성을 제시하였다.

키워드 : 단어 영상 추출, 문서 영상, 그림 영역

Word Image Decomposition from Image Regions in Document Images using Statistical Analyses

Chang Bu Jeong[†] · Soo Hyung Kim[††]

ABSTRACT

This paper describes the development and implementation of a algorithm to decompose word images from image regions mixed text/graphics in document images using statistical analyses. To decompose word images from image regions, the character components need to be separated from graphic components. For this process, we propose a method to separate them with an analysis of box-plot using a statistics of structural components. An accuracy of this method is not sensitive to the changes of images because the criterion of separation is defined by the statistics of components. And then the character regions are determined by analyzing a local crowdedness of the separated character components. Finally, we devide the character regions into text lines and word images using projection profile analysis, gap clustering, special symbol detection, etc. The proposed system could reduce the influence resulted from the changes of images because it uses the criterion based on the statistics of image regions. Also, we made an experiment with the proposed method in document image processing system for keyword spotting and showed the necessity of studying for the proposed method.

Key Words :Word Image Decomposition, Document Image, Image Region

1. 서 론

문자/그래픽이 혼합된 영상에서 문자/단어를 추출하는 연구는 공학 설계 도면이나 래스터 지도의 영상에 대하여 지금까지도 활발히 진행되고 있고, 최근에는 동영상의 자막이나 자연영상에서의 문자를 추출하는 연구로 폭넓게 응용되고 있다. 또한 디지털 도서관이나 인터넷에서 원문 검색 서비스로 활용할 수 있는 주제어 검색(keyword spotting) 시스템은 문서 영상을 단어단위 영상으로 분할하는 오프라인상의 전처리 모듈에 응용하고 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].

전자에 속하는 연구들은 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 의미 있는 연결을 위하

※ 본 논문은 2006년 호남대학교 교내연구비로 지원되었음.
† 정 회 원 : 호남대학교 인터넷소프트웨어학과 전임강사
†† 정 회 원 : 전남대학교 전자컴퓨터공학부 부교수
논문접수 : 2006년 9월 19일, 심사완료 : 2006년 10월 31일

여 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. 하지만 이러한 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 통계적 분석을 통하여 단어 영상을 추출하는 방법을 제안한다. 제안 방법은 문자 성분과 그래픽 성분을 분리하기 위하여 임계값을 이용하지 않고 연결 요소들의 통계적 특징을 이용하기 때문에 그림 형태의 변화에 민감하지 않고, 지역적 밀집도 분석으로 보다 정확한 문자 영역을 추출이 가능하다. 그리고 문자 영역에서 투영 프로파일 분석과 갭 군집화 등의 방법을 적용하여 문자열 및 단어 영상을 추출한다.

2. 관련 연구

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. 전자는 문서 영상의 연결 요소 분석으로서 구해진 연결 요소들의 정보(평균 면적, 평균 수직수평 비, 밀집도 등)를 이용한 휴리스틱 필터링 방법이 일반적이었다. 반면에 후자는 문자 성분들의 의미 있는 연결을 위하여 HT(Hough Transform), RLSA(Run-length Smoothing Algorithm), 모폴로지 등의 다양한 방법을 이용하여 연구되었으며, 그중에서도 문자열의 속성(문자열의 방향, 폰트 형태, 문자 크기 등)의 다양함에 영향을 덜 받고 좋은 성능을 보이는 HT가 많이 이용되었다[1-6].

Fletcher[1]은 문자와 그래픽이 혼합된 문서 이미지에 대하여 문자 성분을 분리하고, HT를 이용하여 문자열을 추출하며 문자열을 단어나 구로 분리하였다. 먼저 문자 성분과 그래픽 성분의 분리는 8-방향 연결 요소 분석으로 구해진 연결 요소들의 면적과 수직수평 비율을 이용하였다. 그리고 분리된 문자 성분의 연결 요소 중심점(centroid)에 대한 HT를 실행하여 동일 직선상의 문자열을 구성하는 문자 성분을 찾았으며, 이들의 높이에 대한 특징을 이용하여 이들을 단어나 구로 분리하였다. 그러나 이 방법은 문자 성분과 그래픽 성분을 분리하기 위하여 단순한 연결 요소들의 특징만을 고려하기 때문에 글자의 크기, 문자열간의 간격, 해상도, 문자간이나 단어간의 간격 등의 여러 가지 제약조건을 제시하였다.

Tombre[2]는 그래픽 성분이 많은 공학도면 영상에 대해서 Fletcher[1]의 방법을 보완하여 적용하였다. 우선 두 가지 성분(문자 성분과 그래픽 성분)으로의 분리를 세 가지 성분(small components, large components, small enlongated components)으로 분리로 확장시켰으며, 문자 성분을 분리하기 위하여 보다 다양한 특징값을 이용하였다. 문자열 추출은 기존 방법인 HT를 적용하였으며, 그래픽 성분에 접촉한 문자 성분의 분리를 위해서 그래픽 성분에 대한 후처리를 수행하였다.

Lu[3]은 그래픽 성분의 접촉한 문자 성분도 문자 영역으로 포함시키는 방법을 제안하였다. 이 방법은 기존의 방법들이 우선적으로 그래픽 성분의 특징(큰 면적, 수직수평 비율)을 이용하여 문자 성분을 분리한 것과는 달리, 라인 성분을 제거한 다음에 문자 성분을 분리한다. 여러 방향의 라인 성분을 제거하기 위하여 원 영상을 여러 각도로 스트레칭 변환하여 수평 방향의 라인 성분을 제거한 후, 반대로 스트레칭 변환을 수행한다. 이와 같이 라인 성분이 제거된 영상에 대하여 연결 요소 분석을 수행하고 연결 요소의 특징(밀집도, 수직수평 비 등)을 이용하여 일차적으로 문자 성분을 분리하였다. 다음으로 브러싱(brushing) 연산과 모폴로지의 열림(openning) 연산을 적용한 후에 다시 연결 요소를 분석하여 최종적인 문자 성분을 추출하였다.

Shiku[4]는 문자 영역을 추출하기 위하여 세그멘트(연결 요소)의 지역적 밀집도를 고려하고, 문자 영역에서 문자열을 추출하기 위하여 세그멘트의 전역적 밀집도를 이용하였다. 세그멘트의 지역적 밀집도는 영상의 한 픽셀에 대하여 주변 세그멘트의 개수와 거리를 고려한 특징으로 세그멘트의 픽셀이나 주변 픽셀의 밀집도는 높게 형성된다. 세그멘트의 전역적 밀집도는 어느 직선($\rho = x\cos\theta + y\sin\theta$)과 만나는 세그멘트의 개수를 고려한 것으로 세그멘트의 동일선상 유무를 결정한다.

Tan[5]는 피라미드 방식을 이용하여 연결 요소 추출에 기반한 문자열을 추출하는 새로운 방법을 제안하였다. 이 방법은 연결 요소의 특징(크기, 수직수평 비, 밀집도)을 이용하여 상대적으로 큰 그래픽 성분을 우선적으로 제거한 후, 수정된 영상을 2×2단위로 크기를 줄여가는 피라미드 영상을 생성한다. 그리고 피라미드 영상의 낮은 레벨에서 연결 요소의 면적, 폭, 높이 등을 고려하여 문자열을 추출한다. 이 파라미드 방식은 영상에서 단어를 효과적이고 빠르게 추출할 수 있으나, 각각의 글자들을 하나의 의미 있는 단어로 연결할 때 결과의 성능이 떨어지는 단점이 있다.

김석태[6]은 문서 영상내의 연결 성분의 구조적 특징을 이용해 사진 및 그래픽 등이 혼합되어 있는 문서에서 그래픽 영역과 문자 영역을 분리하고, 문자 영역에서 각 문자열을 찾아내는 방법을 제안하였다. 문자 영역과 그래픽 영역의 분리에서 이용한 연결 성분의 구조적 특징으로 흑백화소

의 교차횟수, 연결화소 길이, 기본단위영역의 밀도 등을 이용한다. 문자열을 찾을 때는 임계값보다 작은 거리의 기본 단위영역들을 결합해 단어를 생성하고, 비슷한 기울기를 가진 단어들을 병합해 최종적인 문자열을 추출한다.

Park[7]은 신문이나 잡지의 광고 등의 문서에서 양각의 문자뿐만 아니라 음각의 문자들도 추출해 단어로 연결하는 시스템을 제안하였다. 이 시스템은 run-length code를 이용하여 문자나 이미지의 경계선(bound) 모양의 특징을 추출하여 음각 문자와 이미지, 양각 문자를 구분한다. 그리고 추출된 문자들을 3차원 공간상에 매핑한 후, 3차원 면적 가중치 그래프를 이용하여 관련된 단어들로 연결한다.

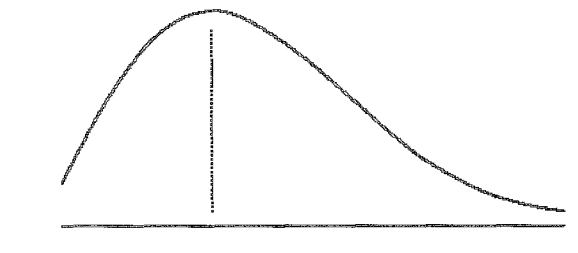
하지만, 이러한 많은 방법들은 문서 영상의 종류와 타입에 따라 민감하게 반응할 수밖에 없다. 우선 문자 성분과 그래픽 성분을 분리하는 방법에서 대부분의 연구가 테스트 영상에 적합한 임계값을 사용하고 있기 때문에 다른 종류의 영상에 대해서 문자 성분이 누락되거나 그래픽 성분이 문자 성분으로 잘못 분리될 수 있다. 또한 문서의 해상도나 획득 과정의 오류로 그래픽 성분의 단락이 발생할 수 있고, 이로 인하여 그래픽 성분이 문자 성분으로 분리되는 오류가 가능하다.

3. 제안 방법

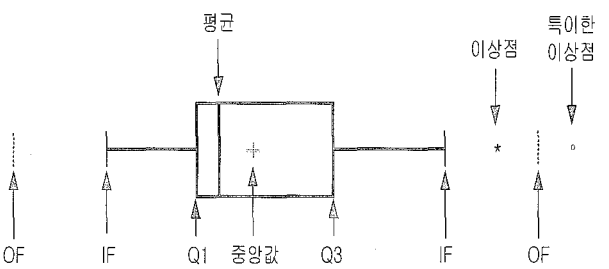
그림 영역에서의 단어 분리는 문자 성분과 그래픽 성분을 분리하여 문자 영역을 추출하고, 추출된 문자 영역에서 문자열을 찾고 문자열을 단어단위로 분리하는 두 단계의 과정으로 수행된다. 첫 번째 단계는 연결 요소의 특징에 대한 통계분석의 상자그림(box plot)을 이용하여 문자 성분을 추출하고, 추출된 문자 성분의 연결 요소에 대한 지역적 밀집도(local crowdedness)를 분석하여 문자 영역을 결정한다. 두 번째 단계에서는 추출된 문자 영역의 연결 요소를 분석하여 정창부[9]의 알고리즘을 적용하여 문자열 분리와 단어 단위 분리를 수행한다. (그림 3.1)은 제안 방법의 단계별 수행 과정을 도식화한 것이다.

3.1 문자 영역 추출

3.1.1 상자그림(Box plot)



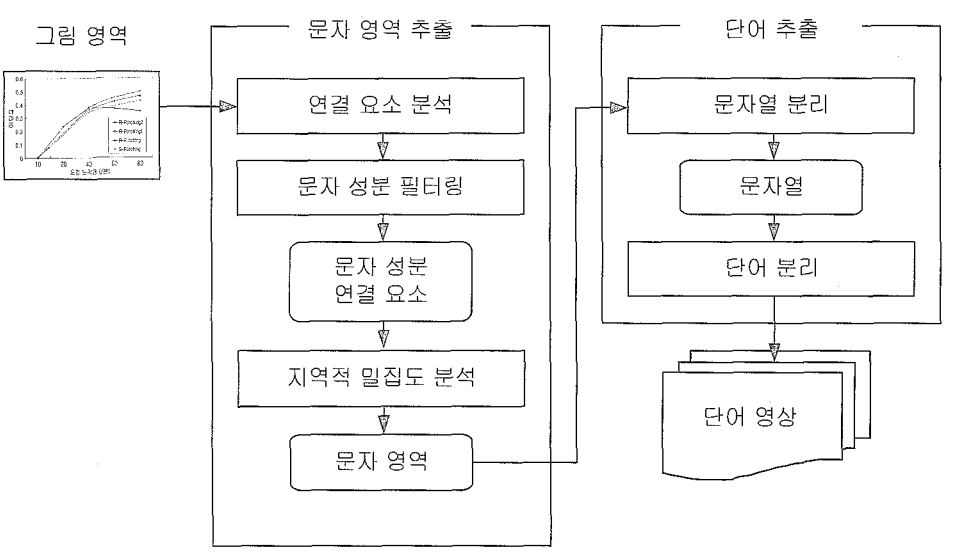
(a) 자료의 분포도



(b) (a)에 대한 상자그림

(그림 3.2) 상자그림 예

때로는 자료들 가운데에서, 대부분의 다른 관측값들보다 월등하게 크거나 작은 값들이 있을 수 있는데, 그런 값들을 이상점(outlier) 또는 특이점이라 한다. 이상점들은 전체자료에 크게 영향을 미치기 때문에 Tukey는 이상점을 식별하는 방법으로 상자그림을 제시하였다. 상자그림이란 탐색적 자료 분석 방법으로 다섯 숫자 요약(five number summary)을 그림으로 그린 것이며 "상자와 수염 그림(box-and-whisker diagram)"이라고도 한다. 최소값, 제1사분위수(Q1), 중앙값



(그림 3.1) 그림 영역에서 단어 영상 추출의 단계별 수행 과정

(median), 제3사분위수(Q3), 자료의 최대값을 다섯 숫자 요약이라 하는데, 이는 자료의 분포상태를 나타내주며, 특히 Q3과 Q1의 차를 사분위범위(interquartile range) IQR이라 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. (그림 3.2)는 상자그림의 예를 보여주고 있다[8].

문서/그래픽이 혼합된 문서에서 연결 요소들의 크기를 자료의 특징값으로 보고 위의 상자그림을 구성하면, 문자들처럼 작은 성분에 비해 소수의 큰 그래픽 성분이나 아주 작은 그래픽 성분(도트나 잡음)은 상자그림의 이상점으로 표현될 것이다. 문자 성분의 추출은 이러한 이상점으로 간주되는 성분을 제외함으로써, 즉 IF 안에 있는 성분들만을 문자 성분으로 간주하고 추출하는 것이다.

3.1.2 문자 성분 추출

문서/그래픽이 혼합된 영상에서 문자 성분의 추출은 연결 요소 분석 후, 일차적으로 상자그림을 이용하여 큰 그래픽 성분을 제거하고, 연결 요소의 형태적 특징을 고려하여 최종적인 문자 성분을 추출하는 것이다.

입력 영상에 대하여 8-방향 연결 요소 분석을 수행하고, 분석된 연결 요소의 BB(bounding box)를 구한다. 상자그림을 구성하는 관측값으로 연결요소의 BB 대각선 길이를 이용하고, Q1과 Q3, IQR를 순차적으로 계산하고 문자 성분의 조건인 IF를 구한다. (그림 3.4)는 (그림 3.3)에 대한 상자그림으로서, IF의 범위가 [6, 48]이므로 BB의 대각선 길이가 IF의 범위 안에 있으면 문자 성분으로 간주하는 것이다.

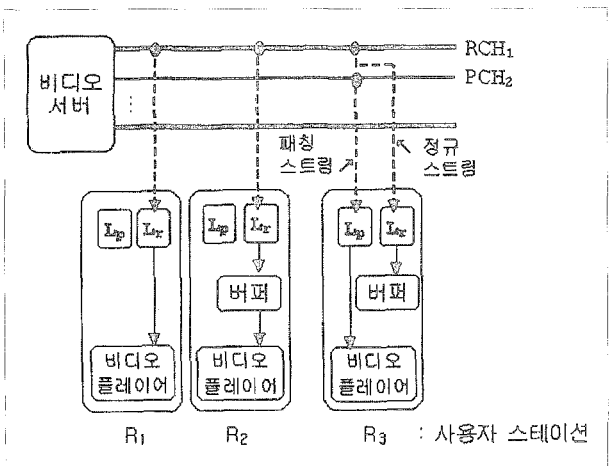
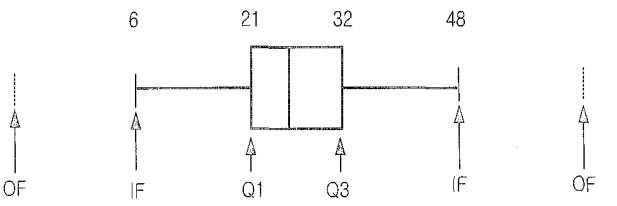


그림 2 그림 1의 t_4시점에서 사용자 스테이션의 상태

(그림 3.3) 그림 영역의 예제 영상

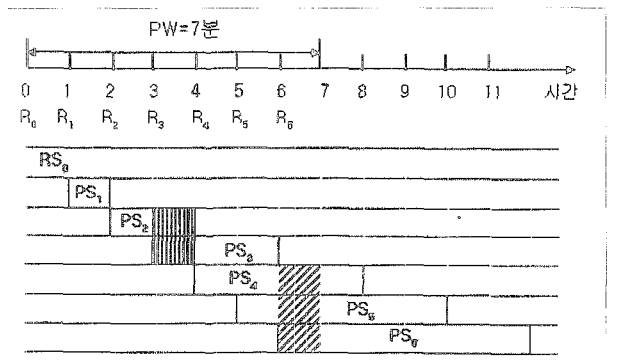


(그림 3.4) (그림 3.3)에 대한 상자그림

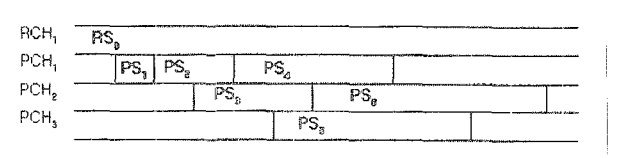
그러나 영상의 해상도가 낮거나 영상 획득 시의 오류로 인하여 큰 그래픽의 성분이 끊어지는 경우에는 위의 방법으로 제외되지 않는 그래픽 성분이 발생한다. (그림 3.5)의 (a)가 그런 경우의 영상으로서, 상단에 위치하는 그래픽 성분의 라인이 끊어져서 별개의 라인이 발생하였고 이런 라인은 (b)와 같이 상자그림 방법에서 문자 성분으로 간주된다. 그래서 이런 라인과 같은 문자 성분에서 제외하기 위하여 다음과 같은 조건

$$\frac{1}{T_1} \leq \frac{BB\text{의 폭}}{BB\text{의 높이}} \leq T_1 ,$$

을 문자 성분의 조건으로 추가한다. 위의 조건에서 T_1가 너무 작으면 라인들은 제외되겠지만 한글의 모음 'ㅡ'나 숫자 '1', 영문자 'I' 등도 제외될 수 있기 때문에 T_1는 다소 큰 값으로 결정해야 한다. (그림 3.6)은 (그림 3.5)의 (b)에 위의 문자 성분의 수직수평 비율에 대한 조건을 추가하여 문자 성



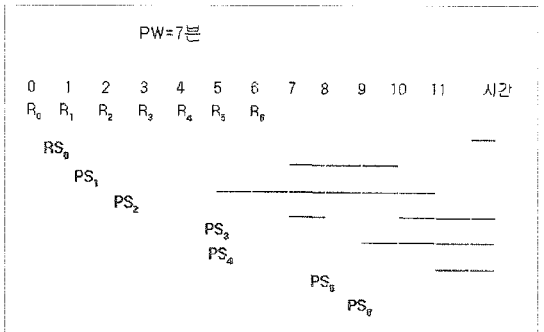
(a) R_0~R_6 요청의 서비스 시작시간과 패칭 스트림



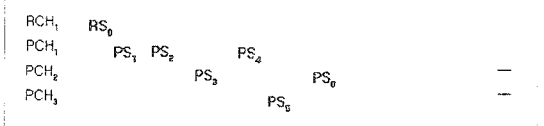
(b) R_0~R_6 요청을 서비스하고 있는 채널의 상태

그림 3 비디오 요청과 채널의 상태

(a) 그림 영역의 입력 영상



(a) R_0~R_6 요청의 서비스 시작시간과 패징 스트림



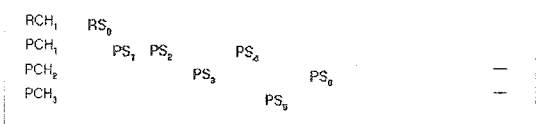
(b) R_0~R_6 요청을 서비스하고 있는 채널의 상태

그림 3 비디오 요청과 채널의 상태

(b) 상자그림을 이용한 문자 성분 추출

(그림 3.5) 상자그림을 이용한 문자 성분의 추출 예

(a) R_0~R_6 요청의 서비스 시작시간과 패징 스트림



(b) R_0~R_6 요청을 서비스하고 있는 채널의 상태

그림 3 비디오 요청과 채널의 상태

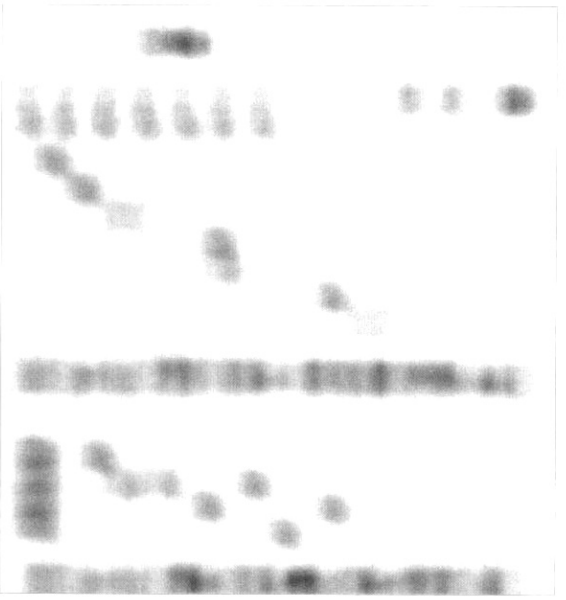
(그림 3.6) BB의 수직수평 비율에 대한 조건 추가 후의 문자 성분 추출 결과

분을 추출한 결과로 일부 라인의 문자 성분으로 추출됨을 허용하였다.

3.1.3 문자 영역 추출

추출된 문자 성분들로부터 문자 영역의 추출은 Shiku[4]에서 제안한 세그멘트의 지역적 밀집도 개념을 응용한다. Shiku[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

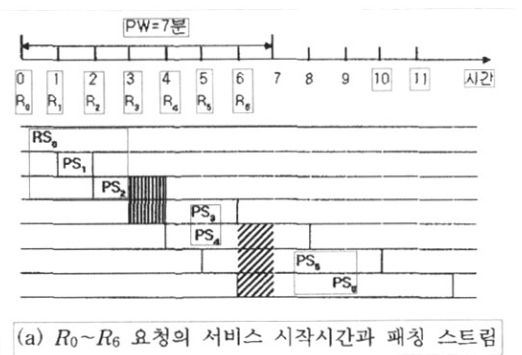


(그림 3.7) (그림 3.6)에 대하여 지역적 밀집도를 계산한 결과

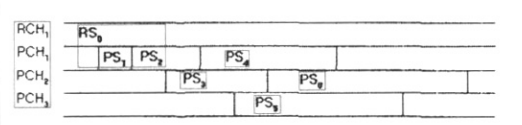
다음과 같다.

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

위 식에서 N은 중심이 점 (x,y), 반지름이 $T_2(T_2=3\sigma)$인 원과 만나는 문자 성분의 BB 개수이며, $e^{-di^2/2\sigma^2}$ $(0<e^{-di^2/2\sigma^2}\leqq 1)$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



(a) R_0~R_6 요청의 서비스 시작시간과 패징 스트림



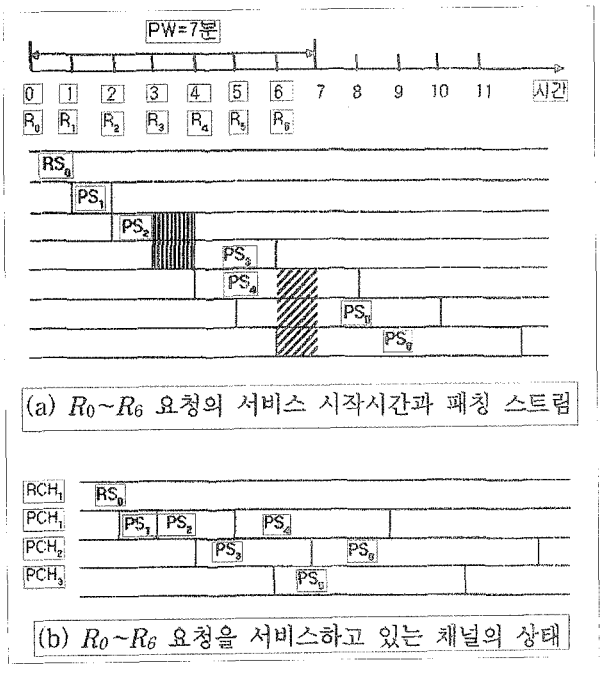
(b) R_0~R_6 요청을 서비스하고 있는 채널의 상태

(그림 3.8) 지역적 밀집도를 이용하여 문자 영역을 추출한 결과

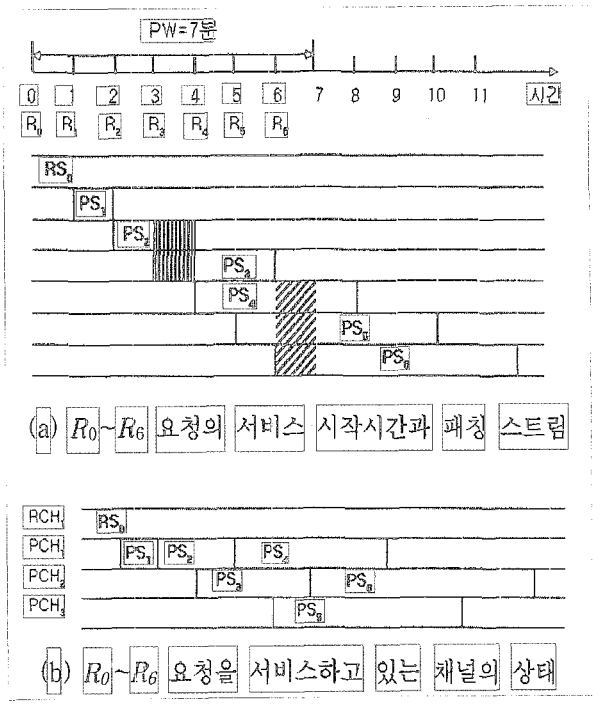
안에 있는 문자 성분의 BB 개수와 거리가 반영된 것으로서, 픽셀 (x,y)이 문자 영역에 포함되는 지를 결정할 값이다. (그림 3.7)은 (그림 3.6)의 문자 성분에 대하여 지역적 밀집도를 계산한 결과로 문자 영역에 해당되는 픽셀은 문자 영역과는 멀리 떨어진 픽셀에 비하여 검은 색을 보여준다. (그림 3.8)은 (그림 3.7)의 지역적 밀집도를 이진화한 결과를 이용하여 결정된 문자 영역의 결과이다.

3.2 문자열 및 단어 분리

그림 영역의 문자 영역을 문자열과 단어로 분리하는 것은 정창부[9]의 테이블 영역에 대한 단어 추출방법을 적용한다. 추출된 문자 영역의 문자열 분리는 문자 영역의 연결요소에 대한 수평 투영 프로파일을 분석하여 수행하고, 분리된 문자열은 수직 투영 프로파일 분석을 통하여 갭(gap)을 구한다. 단어 분리는 문자열에서 구해진 갭을 단어간의 갭(IWG: Inter-Words Gap)과 단어속의 갭(WWG: Within-Words Gap)으로 분류하여 IWG을 이용하여 단어 분리를 실행한다. 갭을 IWG와 WWG로 분류하기 위하여 임계치를 사용하지 않고 군집화 알고리즘에 적용함으로써 문서 영상의 변화에 적응적으로 수행한다. 또한 추가적인 단어 분리는 단어와 단어 사이에 위치하여 IWG을 대신할 수 있는 '~', '-', '(', '[', '{' 등과 같은 특수 기호를 인식 과정 없이 형태적 특징을 분석하여 검출함으로써 정확한 단어 영상의 추출을 가능하게 한다. (그림 3.9)는 (그림 3.8)의 문자 영역을 문자열 및 단어로 분리한 결과이다.



(a) 문자열 분리 결과



(b) 단어 분리 결과

(그림 3.9) 문자열 분리와 단어 분리 결과

4. 실험 및 결과

4.1 실험 데이터

그림 영역에서의 단어 추출 방법의 성능 평가를 위하여 총 55개의 그림 영상을 사용하였다. 실험 대상인 그림 영상은 정보과학회에서 제공하는 원문 검색 서비스를 이용하여 다운받은 논문 영상에 대하여 기울어짐 교정과 문서 구조 분석의 전처리로 추출된 그림 영역만을 저장하였다. 또한 그림 영상은 (그림 4.1)과 같이 다양한 형태의 다이어그램으

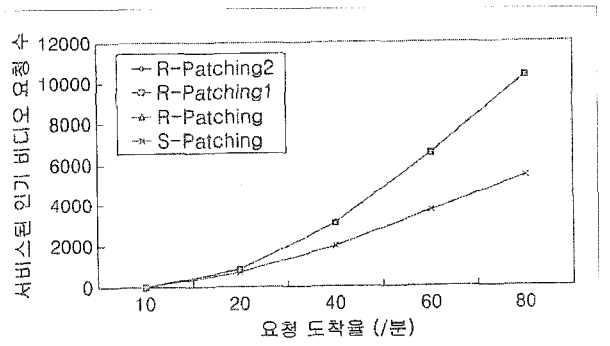


그림 9 서비스된 인기 비디오 요청 수

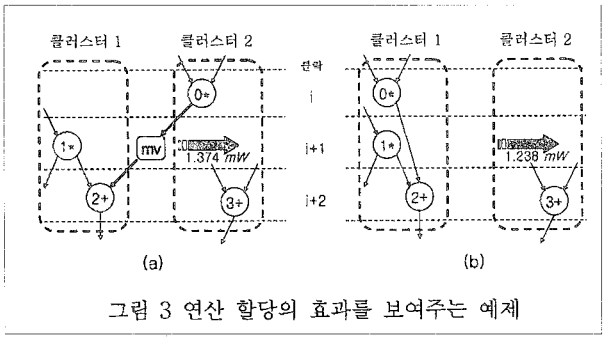


그림 3 연산 할당의 효과를 보여주는 예제

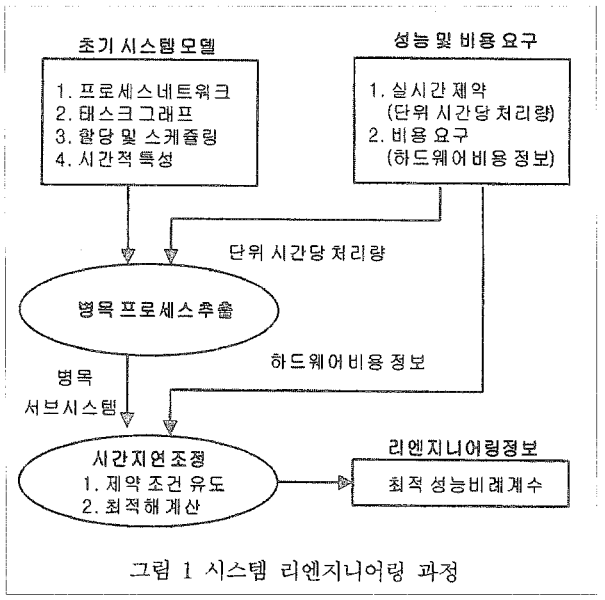


그림 1 시스템 리엔지니어링 과정

(그림 4.1) 그림 영역에서의 단어 분리 방법에 대한 실험 영상 예

로 구성되어 있지만, 문자가 포함되지 않은 순수 그림으로만 구성되는 것도 가능하다. 그러나 본 논문의 실험 영상은 문자가 포함된 영상으로서 300dpi의 이진 영상이다.

4.2 성능 평가

<표 4.1>은 55개의 그림 영상에서 단어를 추출한 결과로서, 총 1,712개의 단어 중에 1,332개의 단어를 성공적으로 추출하여 77.80%의 단어 추출 성공률을 보여준다.

단어 추출의 실패 원인은 <표 4.2>에서와 같이 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 4.2). 이에 해당하는 단어는 196개이고, 대부분 문자의 수가 2개 이하로 구성된 단어였다. 두 번째는 단어의 작성 방향이 가로가 아닌 세로나 대각선이여서 문자열 분리가 실패하는 오류로서, 총 80개의 단어가 이에 해당하였다(그림 4.3). 세 번째는 단어의 일부가 그래픽 성분과 접촉하여 그래픽 성분의 일부로 처리되는 오류로서, 총 45개의 단어가 이에 해당하였다(그림 4.4). 네

〈표 4.1〉 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%)

〈표 4.2〉 그림 영역에서의 단어 추출 결과

오류 유형	단어 추출의 실패 원인	해당 단어 개수
1	2개 이하의 연결요소로 구성된 고립된 단어	196
2	세로나 대각선으로 작성된 단어	80
3	문자 성분이 그래픽 성분과 붙은 경우	45
4	WWG가 IWG로 분류되는 군집화 오류	24
기타	음영이 있거나 밑줄로 연결된 단어 등	35
계		380

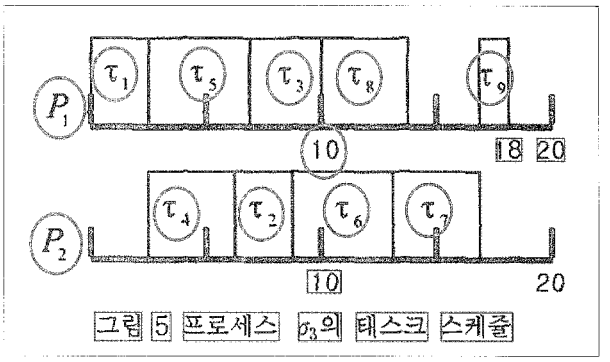


그림 5 프로세스 σ_3의 태스크 스케쥴

(그림 4.2) 그림 영역에서의 단어 추출 실패 1
- 소수 문자들로 구성된 고립된 단어

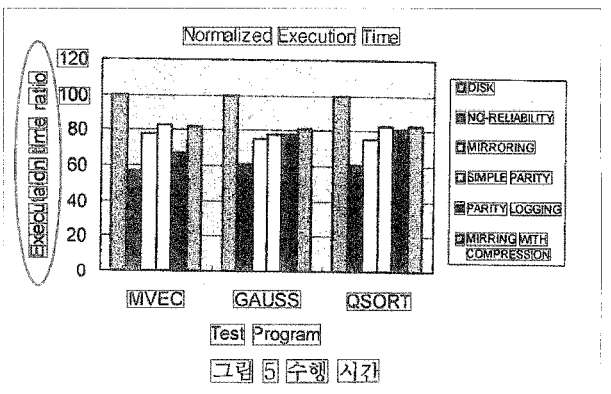


그림 5 수행 시간

(그림 4.3) 그림 영역에서의 단어 추출 실패 2 - 세로형 단어

번째는 소수점의 제거 등으로 WWG로 분류될 갭의 크기가 IWG에 더 유사하여 IWG로 오분류되는 군집화의 오류로서, 총 24개의 단어가 이에 해당하였다(그림 4.5). 그리고 기타 오류의 원인은 음영이 있는 단어(그림 4.6)나 밑줄로 연결된 단어들 등이 있었다.

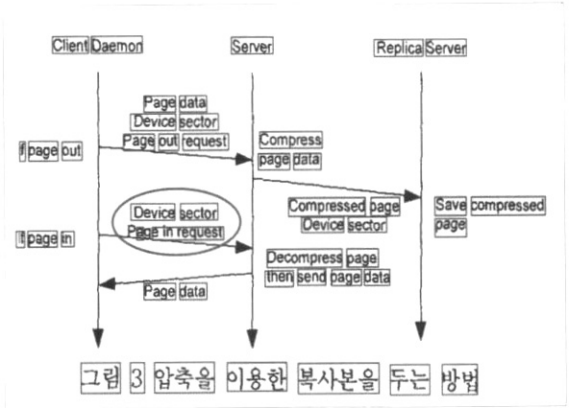


그림 3 압축을 이용한 복사본을 두는 방법

- 문자 성분과 그래픽 성분이 붙은 경우

(그림 4.4) 그림 영역에서의 단어 추출 실패 3

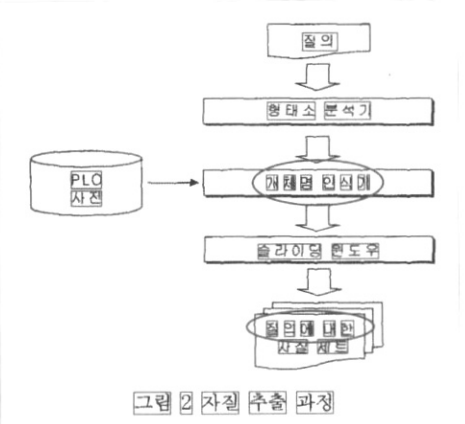


그림 2 자질 추출 과정

(그림 4.5) 그림 영역에서의 단어 추출 실패 4
- 군집화 에러

그림 영역에서의 단어 추출 오류에는 단어를 제대로 추출하지 못하는 오류뿐만 아니라, 문자 성분이 아닌 그래픽 성분을 단어로 추출하는 오류도 있었다. 이런 오류로 추출된 단어는 52개가 있었고, 이를 고려한 그림 영역에서의 단어 추출

〈표 4.3〉 각 영역에 대한 결과 및 전체 시스템 결과

영상 번호	실제 단어 수(개)				추출된 단어 수(개)					성능(%)	
	텍스트 영역 ⓐ	테이블 영역 ⓑ	그림 영역 ⓒ	전체영역 ⓓ =ⓐ+ⓑ+ⓒ	텍스트 영역 ⓐ′	테이블 영역 ⓑ′	그림 영역 ⓒ′	ⓔ =ⓐ′+ⓑ′	ⓕ =ⓔ+ⓒ′	=ⓔ/ⓓ	=ⓕ/ⓓ
1	474	0	0	474	472	0	0	472	472	99.58	99.58
2	581	0	35	616	581	0	29	581	610	94.32	99.03
3	570	0	27	597	569	0	24	569	593	95.31	99.33
4	707	0	0	707	707	0	0	707	707	100.00	100.00
5	501	0	38	539	498	0	29	498	527	92.39	97.77
6	500	30	26	556	497	30	8	527	535	94.78	96.22
7	304	0	114	418	290	0	72	290	362	69.38	86.60
8	588	0	0	588	588	0	0	588	588	100.00	100.00
9	463	0	0	463	463	0	0	463	463	100.00	100.00
10	602	0	0	602	602	0	0	602	602	100.00	100.00
...											
41	390	0	30	420	369	0	30	369	399	87.86	95.00
42	485	31	4	520	479	30	0	509	509	97.88	97.88
43	404	0	0	404	403	0	0	403	403	99.75	99.75
44	451	21	63	535	439	21	62	460	522	85.98	97.57
45	648	0	0	648	646	0	0	646	646	99.69	99.69
46	432	43	54	529	420	43	54	463	517	87.52	97.73
47	491	0	7	498	486	0	7	486	493	97.59	99.00
48	491	0	10	501	486	0	50	486	501	97.01	98.30
49	477	0	52	529	470	0	20	470	492	88.85	90.20
50	400	169	0	569	397	169	0	556	556	99.47	97.72
계	23,639	1,263	992	25,894	23,405	1,177	752	24,582	25,334		
성능	-				98.5% =ⓐ′/ⓐ	97.39% =ⓑ′/ⓑ	75.81% =ⓒ′/ⓒ	94.93% =ⓔ/ⓓ	97.84% =ⓕ/ⓓ		

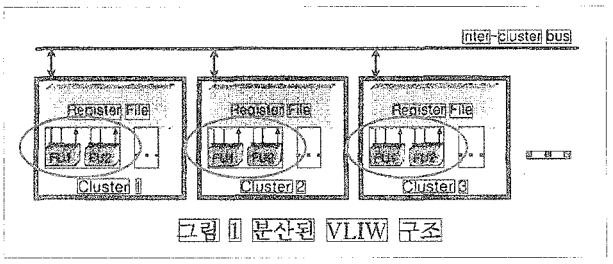


그림 1 분산된 VLIW 구조

(그림 4.6) 그림 영역에서의 단어 추출 실패 5 - 배경이 있는 단어

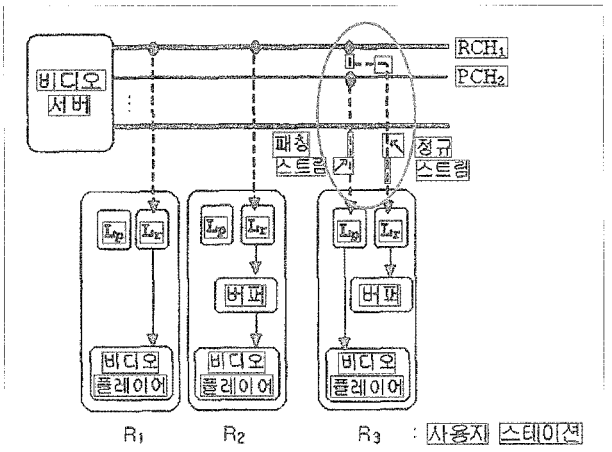


그림 2 그림 1의 t_4시점에서 사용자 스테이션의 상태

(그림 4.7) 그림 영역에서의 단어 추출 실패 6 - 그래픽 성분들로 구성된 단어

성공률은 75.81%가 된다. (그림 4.7)은 화살표나 점선이 문자열의 일부로 포함되어 단어로 추출되는 오류를 보여준다.

4.3 문서 영상의 전체 영역에 대한 성능 평가

제안 방법이 주제어 검색을 위한 영상 전처리 시스템[9]에서 효율적으로 적용됨을 증명하기 위하여 추가적인 성능 평가를 실시하였다. 실험 영상은 정보과학회에서 제공하는 원문 검색 서비스를 이용하여 다운받은 논문 영상 50개로서, 300dpi의 이진 영상이며, 크기는 A4 (210×297㎜)의 크기이다.

<표 4.3>은 50개의 실험 영상에 대하여 영상 전처리 시스템의 성능을 보여준다. 제안 방법을 추가하기 이전의 시스템은 텍스트 영역과 테이블 영역에 있는 단어만 추출하기 때문에 실험 영상의 25,894개의 단어 영상 중 94.93%에 해당하는 24,582개의 단어 영상을 추출하였다. 반면에 제안 방법을 추가한 시스템에서는 그림 영역에 있는 단어 영상 752개를 추가 추출하여 97.84%의 향상된 성능을 보였다. 그렇지만 이와 같은 성능 결과는 4번 영상과 7번 영상에 대한 실험 결과의 비교 분석에 알 수 있듯이 제안한 방법의 성능은 실험 영상의 형태에 상당히 의존적이다. 즉, 실험 영상에 단어 추출 성능이 다른 영역에 비해 저조한 그림 영역이 많이 있다면 성능은 낮아질 수 있다는 것이다. 그러나 시스템의 성능이 그림 영역에 있는 단어 영상을 추출함으로써 보다 더 향상될 수 있음을 확인하였다.

5. 결 론

본 논문에서는 주제어 검색 시스템의 오프라인 상의 전처리 모듈에 해당하는 영상 전처리 시스템의 성능을 향상시킬 수 있는 문서 영상의 단어 추출 알고리즘을 제안하였다. 제안한 단어 추출 알고리즘은 문서 영상의 그림 영역에서 문자 성분과 그래픽 성분의 분류를 위하여 임계값을 사용하는 대신에 통계적 분석 방법인 상자그림을 응용하였다. 그리고 문자 성분을 문자 영역으로 그룹핑하기 위하여 지역적 밀집도 분석을 이용하고, 문자 영역의 단어 추출은 테이블 영상에 대한 단어 추출 알고리즘을 적용하였다. 그러나 문자열 및 단어 분리가 수평으로 제한된 테이블 영상에 대한 알고리즘을 적용하였기 때문에 수평이외의 방향으로 구성된 단어는 추출할 수 없는 문제점을 보였다.

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.

참 고 문 헌

[1] L.A. Fletcher and R. Kasturi, "A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 10, No. 6, pp. 910-918, 1988.

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

[3] 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.

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

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

[7] 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.

[8] 심정욱, 손영숙, 백장선, 수리통계학, 자유아카데미, 제4

판, 2003.
[9] 정창부, 김수형, "투영 프로파일, Gap 및 특수 기호를 이용한 텍스트 영역의 어절 단위 분할," 정보과학회논문지: 소프트웨어 및 응용, 제31권, 제9호, pp. 1121-1130, 2004.

정 창 부

e-mail : cbjeong@honam.ac.kr
1999년 호남대학교 컴퓨터공학과(공학사)
2001년 전남대학교 일반대학원 전산통계학과(이학석사)
2002년 4월~10월 캐나다 Concordia 대학 CENPARMI연구소(방문연구원)
2006년 전남대학교 일반대학원 전산학과(이학박사)
2005년~현재 호남대학교 인터넷소프트웨어학과 전임강사
관심분야 : 문서영상전처리, 패턴인식, 의료영상

김 수 형

e-mail : shkim@chonnam.ac.kr
1986년 서울대학교 컴퓨터공학과(공학사)
1988년 한국과학기술원 전산학과 (공학석사)
1993년 한국과학기술원 전산학과 (공학박사)
1990년 9월~1996년 12월 삼성전자 멀티미디어 연구소 (선임연구원)
2000년 12월~2002년 1월 캐나다 Concordia 대학 CENPARMI 연구소(방문교수)
1997년~현재 전남대학교 전자컴퓨터공학부 교수
관심분야 : 인공지능, 패턴인식, 문서영상 정보검색, 유비쿼터스컴퓨팅