2000年度 大韓電子工學會 夏季綜合學術大會 論文集 第23卷 第1號 2000/6

Coarse/fine 전략을 이용한 문서 구조 분석

박 동 열(朴 東 烈), 곽 희 규(郭 熹 珪), 김 수 형(金 壽 衡)
전남대학교 전산통계학과
전화 : (062) 530-0430 / 팩스 : (062) 530-3439

Document Layout Analysis Using Coarse/Fine Strategy

Dong Yeol Park, Hee Kue Kwag, Soo Hyung Kim
Dept. Computer Science & Statistics, Chonnam National University
E-mail : u9997464@chonnam.ac.kr, hkkwag@chonnam.ac.kr, shkim@chonnam.ac.kr

Abstract

We propose a method for analyzing the document structure. This method consists of two processes, segmentation and classification. The segmentation first divides a low resolution image, and then finely splits the original document image using projection profiles. The classification deterimines each segmented region as text, line, table or image. An experiment with 238 documents images shows that the segmentation accuracy is 99.1% and the classification accuracy is 97.3%.

I. 서론

전자문서(electronic document)의 사용이 보편화되고 있지만 인쇄된 문서(printed document)에 대한 사람의 기호 때문에 신문, 잡지, 보고서, 책 등과 같은 간행물이 계속해서 출판되고 있다[1]. 따라서 날로 발생량이 증가하고 있는 각종 문서를 효과적으로 저장, 가공, 검색, 재생산하기 위해 문서를 디지털 영상의 형태로 관리하는 방식이 일반화되고 있다. 이러한 시스템에서 사용하는 문서 영상 처리 기술에는 영상 전처리, 문서 구조 분석, 어절 분할 등이 있다. 스캐너를 통해 입력된 문서 영상은 전처리를 수행한 후에 의미있는 영역들로 분할하게 되는데, 각 영역들은 텍스트 또는 그래픽 정보(사진, 그림, 도표 등)를 포함할 수 있다. 지난 몇 년간 문서 구조 분석에 대한 연구가 꾸준히 진행되어 왔음에도 불구하고 문서 구조의 복잡성 때문에 아직도 많은 기술적인 문제점을 안고 있다.

문서 구조 분석 방법에는 크게 상향식(bottom-up) 방법[2,3]과 하향식(top-down) 방법[4]이 있다. 상향식 방법은 연결요소 분석 등을 통한 지역적인 정보를 기반으로 단어에서 행으로, 행에서 문단으로 결합하는 방법이고, 하향식은 상향식의 반대 방향으로 진행된다. 상향식 방법은 시간 복잡도(time complexity)가 크고, 초기의 잘못된 확장으로 인해 분할 오류를 가져올 수 있다. 반면, 하향식 방법은 상향식 방법에 비해 시간 복잡도가 작지만 사각형으로 나눌 수 없는 복잡한 문서 구조나 다양한 폰트가 존재할 때에는 비효율적인 단점을 가지고 있다. 지난 몇 년간 문서 구조 분석에 대한 연구가 꾸준히 진행되어 왔음에도 불구하고 문서 구조의 복잡성 때문에 아직도 많은 기술적인 문제점을 안고 있다.

본 논문에서는 이러한 두 방법의 단점을 보완하기 위해 입력 영상으로부터 다양한 해상도를 가지는 다단계 영상을 생성한 후, coarse/fine 전략을 이용하는 계층적 문서 구조 분할 방법을 사용한다. 제안 방법의 coarse 단계에서는 분할에 따른 시간 비용 및 분할 오류를 줄이기 위해 저해상도 영상에 상향식 방법을 사용하여 대략적인 영역을 추출한다. Fine 단계에서는 분할된 고해상도 입력 영상에 시간 복잡도가 작은 하향식 방법을 사용하여 세밀한 영역 분할을 수행한다.

따라서 전체적인 시간 비용 및 정확도가 향상 될 수 있다. 제안 방법의 성능 평가는 300dpi 해상도로 스캔한 238개 논문 영상을 사용하여 시간 및 정확도를 측정하였다.

Ⅱ. 제안 방법

1. 개 요

본 논문에서 제안하는 방법은 coarse/fine 전략을 이용한 계층적 문서 구조 분석 방법이다. 제안 방법은 다단계 영상 생성, coarse/fine 영역 분할, 영역 부류 결정의 3단계로 구성한다. 먼저, TIF(Tagged Image File)형식의 입력된 영상을 64×64 보다 작을 때까지 2×2단위로 크기를 줄인다. 다라서 0 레벨의 고해상도의 입력 영상으로부터 L 레벨의 저해상도 입력 영상까지 다단계 해상도의 영상을 생성한다. Coarse 영역 분할 단계에서는 저해상도의 L과 L-1 레벨의 영상에 연결요소(connected component) 분석을 사용하여 대략적인 영역 분할을 수행한다. Fine 단계에서는 분할된 고해상도의 각 영역에 대하여 수평 및 수직 투영을 구하여 보다 세밀한 영역 분할을 수행한다. 영역 부류 결정 단계는 분할된 각 영역들을 텍스트(text), 테이블(table), 그림(image), 선(line)중에 하나로 분류한다. 제안 방법의 전체적인 시스템의 구성은 그림 1과 같다.

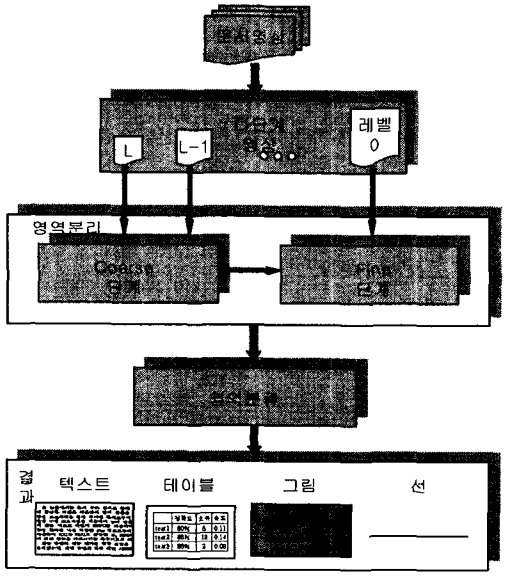
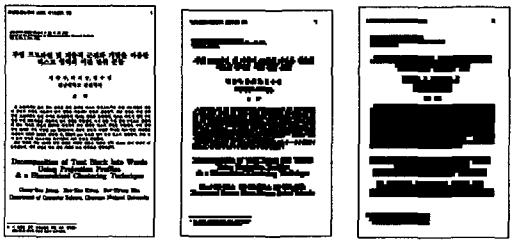


그림 1. 시스템 구성도

2. Coarse 단계 영역 분할

Coarse 단계에서의 영역 분할은 L-1과 L단계를 사용한다. L 단계의 영상은 그림 2에서 보는 바와 같이, 간격이 작은 행간 영역들은 합쳐져서 하나의 큰 검은 영역들을 생성한다. 따라서 L 단계의 영역에 대해 연결요소 분석[5]을 수행하면 영역 구분이 확실한 영역들만이 완전히 연결된 요소로 남게 된다. 그러나 문서의 구조가 보다 복잡한 경우 연결요소의 외곽 사각형들이 서로 접쳐지게 되는데, 이때 한 영역이 다른 한 영역에 완전히 포함되는 경우 하나의 영역으로 합치고 그 외의 경우에는 겹치는 영역의 화소들과 연결이 없는 영역을 수평과 수직방향으로의 투영을 하여 연결강도 작을 부분을 분할하여 두 영역이 서로 겹침이 발생하지 않도록 한다. 이처럼 coarse 단계에서는 작은 해상도의 영역에 대해 연결요소 분석을 수행하기 때문에 시간 복잡도를 줄일 수 있고 영역 구분이 확실한 영역들만을 분할함으로써 오류를 최소화 할 수 있다.



(a) 레벨 0 (b) 레벨 L-1 (c) 레벨 L

그림 2 다단계 문서 영상

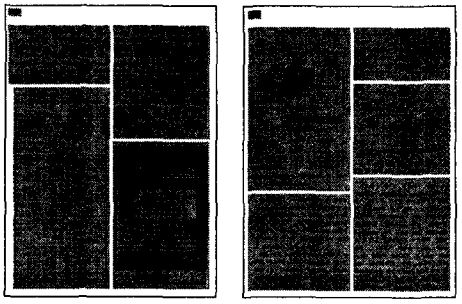


그림 3 Coarse 단계의 영역 분할 결과

3. Fine 단계 영역 분할

Coarse 단계에서 분할된 각 영역들은 동질적인 영역과 이질적인 영역으로 나누어진다. 동질 적인 영역은 동일한 성분들로 구성되어 더 이상 분할을 수행하지 않으며, 이질적인 영역은 서로 다른 성분들로 구성되어 추가적인 분할을 수행해야 한다. 이 단계의 분할

은 원 입력 영상인 0 단계 영상에 대하여 수행하는데, 하향식 접근 방법인 수평 및 수직 프로젝션 프로파일 (horizontal/vertical projection profile) 방법을 사용한다. 예를 들어, 한 영역이 동질의 텍스트만을 포함하는 경우, 수평 프로파일의 폭 및 출현 간격은 일정하다. 반면, 서로 다른 크기의 텍스트만을 포함하는 경우, 수평 프로파일의 폭 및 출현 간격은 일정하게 나타나지 않는다. 또한 한 영역이 다단의 텍스트 영역을 포함하는 경우 수직 프로파일에서 분할 지점을 찾을 수 있다.

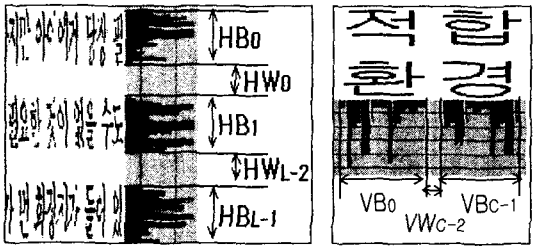


그림 4 수평/수직 방향의 투영

영역 분할을 수행을 수행하기 전에 각 영역에 대하여 분할을 해야되는 영역인지를 검사한다. 그 기준은 먼저 수평과 수직 방향의 프로젝션 프로파일 구한 후, 각 라인의 높이와 각 라인의 간격의 높이를 구한다.

$$HPP_r = \sum_{c=0}^{Width-1} IMG(r,c),\ 0 \le r < Height \quad (1)$$

$$VPP_c = \sum_{r=0}^{Height-1} IMG(r,c),\ 0 \le c < Width \quad (2)$$

$$HB_{avg} = \frac{1}{L}\sum_{r=0}^{L-1} HB_r,\ HW_{avg} = \frac{1}{L-1}\sum_{r=0}^{L-2} HW_r \quad (3)$$

$$HB_{avgdev} = \frac{1}{L}\sum_{r=0}^{L-1}(HB_r - HB_{avg}) \quad (4)$$

$$HW_{avgdev} = \frac{1}{L-1}\sum_{r=0}^{L-2}(HW_r - HW_{avg}) \quad (5)$$

라인의 수가 3이상인 영역에 대하여 HB_{avedev}가 4이상이거나 HW_{avedev}가 4이상이면 이질 영역이 포함된 영역으로 판단하고 분할을 수행하고 그렇지 않을 경우 영역 부류 결정 단계로 넘어간다. 영역 분할을 수행하기 위하여 HB와 HW 에서 각각의 다음 값들과의 차이값인 $\|HB_r - HB_{r+1}\|/3$과 $\|HW_r - HW_{r+1}\|/3$의 각각의 값들 중에서 최빈수인 $HB_{MaxFreq}$와 $HW_{MaxFreq}$을 구한다. $\|HB_r - HB_{r+1}\| < HB_{MaxFreq} + TH_{HB}$인 HB를 군집하고, 마찬가지로 $\|HW_r - HW_{r+1}\| < HW_{MaxFreq} + TH_{HW}$인 HW를 군집한다. 제안 방법에서 사용한 TH_{HB}와 TH_{HW}는 7과 4를 사용하였다. HB와 HW의 군집 결과에 대하여 $HB \cap HW$인 부분을 군집하고 나머지 군집이 안된 부분에서 서로 인접한 부분을 군집한다. 최종적으로 군집된 결과가 2 이상이면 영역 분할을 수행하고 그렇지 않으면 영역 결정 단계를 수행한다.

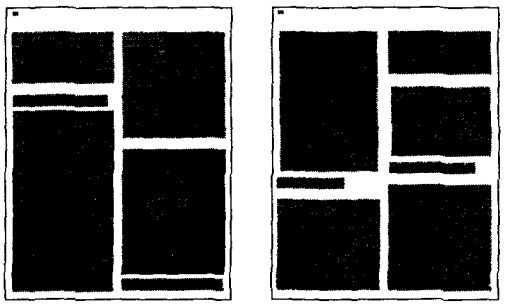


그림 5 Coarse 단계의 영역 분할 결과

4. 영역 부류 결정

Coarse/fine 전략을 사용하여 영역 분할을 수행한 후, 각 영역은 네 부류-텍스트, 그림, 표, 선 중 하나로 분류한다.

표 1 용어 설명

용어	설 명
라인성분	프로젝션 프로파일을 값을 영역의 폭으로 나누었을 때 0.9이상인 값
C	VB의 개수(문자의 개수)
L	HB의 개수(라인의 개수)
H	영역의 높이
W	영역의 폭
$Cross$	영역의 한 라인을 수평으로 스캔했을 때 검은 화소와 흰 화소의 교차회수

① 텍스트 영역으로 분류하는 경우: L이 3 이상이고 HB_{avedev} 가 4보가 작은 경우와 L이 2 이고 한 라인의 높이가 다른 한 라인 높이의 2배 보다 작고 각 라인의 프로젝션 프로파일의 값의 처음과 끝 부분에 라인 성분 없는 경우 그리고 L이 1이고 C가 H/W*0.7보다 크고 HB의 높이의 3분의 1과 3분의 2지점에서의 두 개의 $Cross$값 중에서 하나 이상이 H/W*2.5배보다 큰 경우

② 라인 영역으로 분류하는 경우: L이 1이고 라인 성분이 하나인 경우

③ 테이블 영역으로 분리하는 경우: L이 3이상이고 상단과 하단에 라인 성분이 있을 경우
④ 그림 영역으로 분리하는 경우: 위에 세 가지 이외의 경우

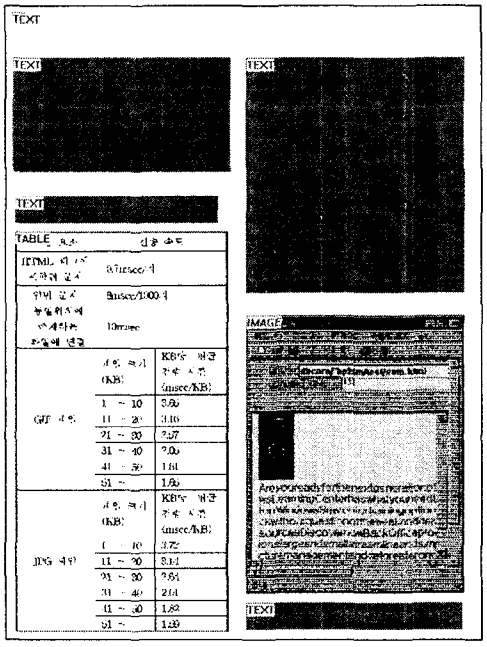


그림 6 Fine 단계의 영역 분할 결과

Ⅲ. 실험 결과 및 분석

문서 구조 분석단계에서 사용되는 데이터베이스는 2단으로 편집된 논문집 238페이지를 300dpi로 스캔하여 자체 구축하였다. 구축된 문서 영상은 기울어짐이 거의 없도록 하였고, 크기는 가로/세로 1432×2024 pixel로 되어있다. 문서 영상은 대부분 한글로 되어있으며, 텍스트, 그림, 표 및 선 등을 포함하고 있다.

표 2 분할 및 분류 실험 결과

		Total	Correct	정확률(%)
영역 분할	Coarse	238	236	99.2
	Fine	2362	2341	99.1
영역 분류	텍스트	2019	1968	97.4
	선	13	13	100
	그림	289	279	96.5
	테이블	41	39	95.1

영영 분할 의 정확률은 99.1%이고 영역 분류의 정확률은 97.3%이다. 그리고 처리시간은 장 당 평균 0.55초 소요되었다.

Ⅳ. 결론 및 향후 연구

본 논문에서는 coarse/fine 전략을 이용한 계층적 문서 구조 분석 방법을 제안하였다. 제안 방법은 전체적인 시간 비용 및 정확도를 향상시키기 위하여 다양한 해상도를 가진 다단계 영상을 생성한 후, coarse 단계에서는 저해상도 전체 영상에 상향식 방법을 사용하고, fine 단계에서는 분할된 고해상도 영역에 하향식 방법을 사용하였다. 또한 분할된 각 영역들은 그 부류를 결정하여 텍스트, 도표, 그림, 선으로 분류하였다. 제안 방법의 성능 평가는 구축한 238개 영상에 대해 99.1%영역 분할과 97.3% 영역 분류의 정확률을 보였다. 향후에는 보다 다양하고 복잡한 영상에 대해 실험하고, 기존 상품화된 방법들과의 성능 비교를 수행할 것이다.

참고문헌

[1] 류대석, 강선미, 이성환, "매개변수에 무관한 새로운 문서구조 분석 방법," 한국정보과학회 가을 학술발표논문집, Vol. 26, No. 2, pp. 482-484, 1999.

[2] D. Driva and A. Amin, "Page Segmentation and Classification Utilizing Bottom-Up Approach," Proc. 3rd Int. Conf. Document Analysis and Recognition, Montreal, Canada, pp. 610-614, 1995.

[3] A. Simon, J. Pret, and A. Johnson, "A Fast Algorithm for Bottom-Up Document Layout Analysis," IEEE Trans. Pattern Analysis. and Machine Intelligence., Vol. 19, pp. 273-276, 1997.

[4] J. Ha, R. Haralick, and I. Philips, "Recursive X-Y Cut Using Bounding Boxes of Connected Components," Proc. 3rd Int. Conf. Document Analysis and Recognition, Montreal, pp. 952-955, 1995.

[5] Anil K. Jain and Bin Yu, "Document Representation and Its Application to Page Decomposition", IEEE Trans. Pattern Analysis and Machine Intelligence. Vol. 20, No. 3, 294-308, 1998.