매개변수에 무관한 새로운 문서 구조 분석 방법

류 대석, 강 선미†, 이 성환
고려대학교 컴퓨터학과, †서경대학교 컴퓨터과학과
E-mail:{dsryu, swlee}@image.korea.ac.kr, smkang@bukak.seokyeong.ac.kr

A New Method for Nonparametric Document Layout Analysis

Dae-Seok Ryu, Sun-Mee Kang† and Seong-Whan Lee
Dept. of Computer Science and Engineering, Korea University
†Dept. of Computer Science, Seokyeong University

요 약

본 논문에서는 매개변수 없이 입력 문서 영상을 최대 동질 영역들로 분할한 다음, 각 동질 영역을 텍스트, 그림, 표 그리고 선으로 자동 분류하는 새로운 방법을 제안한다. 다단계 분석과 하향식 접근 방법을 사용하기 위하여 문서 영상을 피라미드 구조로 계층화하였으며, 어떤 영역을 분할할 지의 여부를 결정하기 위하여 그 영역의 주기성을 이용하여 판단하였다. 이러한 주기성 정보를 이용함으로써, 어떠한 매개변수 없이도 활자체 크기와 행간에 무관하게 텍스트 영역을 정확히 분석할 수 있었으며, 피라미드 구조를 만드는데 걸리는 시간이 질감 분석 접근 방법보다 빠른 방법으로 설계되었다. Washington 대학의 문서 영상 데이터베이스를 이용한 실험 결과, 제안된 방법이 기존의 방법들보다 더 정확하게 문서 영상을 분할 및 분류할 수 있음을 확인할 수 있었다.

1. 서론

정보기술의 발전과 산업 정보화로 인해서 정보를 담고 있는 문서의 사용량이 꾸준히 증가하고 있다. 전자 문서의 사용에도 불구하고 종이 문서가 끊임없이 사용되고 있는 이유는 신문, 잡지, 보고서, 책 등과 같은 간행물이 계속 출판되고 있으며, 사람이 보기에 전자 문서보다 종이 문서가 더 읽기 편하다는 사실에 기인한 것이다. 대용량의 종이 문서를 전자 문서로 자동 변환할 수 있다면 내용 검색이 쉬워질 뿐만 아니라 저장 공간 역시 획기적으로 줄일 수 있다. 그러나 종이 문서에서 전자 문서로의 자동 변환 작업은 쉬운 일이 아니다. 이 작업을 수행하기 위해서는 문서 구조 분석이라는 작업이 선행되어야 한다. 문서 구조 분석이란 문서 영상을 최대 동질 영역들로 분할하고, 각 영역을 텍스트, 그림, 표, 선 등과 같은 영역으로 분류하는 작업을 말한다. 이러한 문서 구조 분석에 의해 문서 영상을 최소 단위 영역으로 분류한 후에, 텍스트 영역은 OCR을 적용하여 ASCII 코드로 변환하고, 그림 영역은 압축하며, 표 영역은 재생성하고, 선과 같은 그래픽 영역은 벡터화하는 등의 효과적인 처리가 가능하게 된다.

문서 분할이나 문서 구조 분석에 대한 연구는 다양한 방법으로 많은 연구가 진행되어 왔다. 그 중 질감 분석 접근 방법[1, 2]에서의 가장 큰 문제점은 시간 복잡도가 크다는 것이다. 특정 질감 정보를 추출하기 위해 여러 가지 필터를 사용하는 방법[1]의 경우 마스크 크기를 작게 하면 큰 질감 정보를 뽑아낼 수 없게 되고, 마스크 크기를 크게 하면 계산 시간이 지수함수적으로 증가하게 되는 문제점을 지닌다. 이를 보완하기 위해 다단계에서 2개의 필터만을 사용하여 분석하는 방법[2]이 제안되었는데, 이 방법은 다단계 질감 정보를 얻기 위한 과정의 계산 부담이 크고, 다른 형태의 영역이지만 비슷한 질감을 갖고 있는 경우 혼동되거나 병합되는 문제점이 있다. 질감 정보를 사용하지 않는 방법[3, 4, 5]에서의 가장 큰 장점은 처리 시간이 빠르다는 것이다. 그러나 기울기에 민감[3]하며, 사람이 임의로 정한 매개변수나 임계치에 의존[4, 5]하는 단점이 있다. 또한 상향식 접근 방법[5]의 경우에는 초기에 작은 요소들을 큰 요소들로 잘못 결합한 경우 마지막 분할 결과도 잘못되는 결과를 초래한다.

이와 같이, 많은 연구가 진행되어 왔지만 다양한 자간 간격과 다양한 활자체 크기로 인해 범용적인 구조 분석 알고리즘을 개발하기가 어려웠으며[5], 이를 보완하기 위해 몇 개의 임계값과 매개변수 설정이 불가피하였다. 본 논문에서는 어떠한 매개변수도 사용하지 않고 다양한 활자체 크기와 행간에 무관하게 비교적 정확히 구조 분석을 수행하는 새로운 방법을 제안한다.

2. 제안된 문서 분할 및 영역 분류 방법

2.1 다단계 피라미드 구조 생성

인간의 시각은 어떤 물체를 볼 때 주위 영역은 저해상도에서 처리하고 관심 영역에 대해서만 선택적으로 주의 집중하여 고해상도로 처리한다. 본 논문에서는 이와 유사한 방법으로 구조 분석을 수행하기 위해 입력 문서 영상을 반복적으로 해상도를 줄여 다단계 피라미드 구조를 만든 후, 저해상도에서 고해상도 방향으로 이동하면서 점점 더 자세히 분할하도록 하였다. 피라미드 구조를 만드는데 드는 계산 비용을 줄이기 위해 간단한 방법을 사용하였는데, 만약 현재 단계의 주위 4개 화소 중 하나 이상이 검은 화소

이면 상위 단계의 해당 화소가 검은 화소가 되도록 하였다. 이렇게 함으로써 상위 단계로 올라갈수록 영상 크기가 줄어들며 인접 영역들이 붙게 되는 효과가 생긴다. 이와 같은 방법으로 그림 1과 같은 다단계 피라미드 구조를 생성한다.

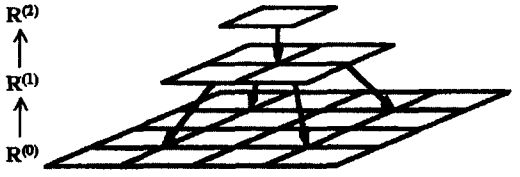


그림 1. 3 단계인 경우의 피라미드 구조

2.2 텍스트 영역의 주기성을 이용한 문서 분할

다단계 영상을 이용하여 최상위 단계에서 최하위 단계까지 반복적으로 본 문서 분할 방법을 적용한다. 먼저 최상위 단계에 대해서 연결 요소 분석을 통해 경계 사각형을 구한다. 그 하위 단계부터는 각 단계의 각 경계 사각형 영역에 대해 가로 방향과 세로 방향으로 그림 2와 같이 주기성을 찾고, 그 결과 주기적이지 않다고 판단되는 경우에 2개의 영역으로 분할 후 각 영역에 대해 이를 반복 적용시킨다.

일반적으로 텍스트 영역은 가로 또는 세로로 정렬되어 있다는 특성으로 인해 다른 영역과 쉽게 구별이 가능하다. 이 특성을 뽑아내기 위해 본 논문에서는 어떤 영역이 단일 주기로 이루어졌는지 그 주기성을 측정하였다.

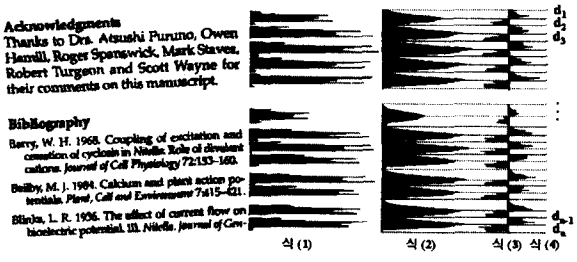


그림 2. 영역의 주기성을 찾는 순서(왼쪽에서 오른쪽 방향)

그림 2에 주기성을 찾는 순서를 나타내었는데, 첫째, 식 (1)과 같이 수평 (또는 수직) 투영 히스토그램을 구한다. 문서 기울기 검출 알고리즘[6]에 의해 얻은 문서의 기울기를 θ라 할 때, 그 기울기 방향에 따라서 투영시킨다. 여기서 $I(x,y)$는 크기가 $width \times height$인 문서 영상에서 (x,y)의 명도값을 나타내며, $P_H^{(n)}$는 n번째 단계에서의 수평 방향 투영 히스토그램을 나타낸다. 둘째, 식 (2)와 같이 투영된 결과에 대해 평활화를 수행한다. 여기서 s는 커널 크기를 나타내며 m_y는 정수형이다. 셋째, 식 (3)과 같이 평활화된 히스토그램에 대해 미분, 즉 기울기를 구한다. 여기서, $D(x,y) = \sqrt{y^2 - x^2}$ 이다. 넷째, 식 (4)와 같이 미분한 히스토그램에 대해 부호 교차점을 구한다. 이렇게 해서 얻게 된 부호 교차점은 평활화된 투영 히스토그램의 극소값이나 극대값에 해당하게 된다. 따라서 이웃 부호 교차점들의 차를 구하여 이들의 분산을 구하면 그 영역에 대한 주기의 변화율을 알 수 있다. 부호 교차점의 차에 대한 분산을 구하는 식은 식 (5)와 같다. 분산 값이 낮다는 것은 주기의 변화율이 거의 없다는 뜻이므로 단일 주기를 갖는 한 단으로 구성된 텍스트 영역으로 분류하여 더 이상의 분할을 하지 않도록 한다. 분산 값이 높다는 것은 주기의 변화율이 크다는 것이며 두 단 내지 다른 영역과 혼합되어 있는 상태이므로 어딘 가에 분할할 부분이 있다는 것을 의미한다.

$$P_H^{(n)} = \{p_y \mid p_y = \sum_{x=0}^{width-1} I(x, y + x\tan\theta),\ 0 \le y < height\} \quad (1)$$

$$M_H^{(n)} = \{m_y \mid m_y = \frac{1}{s}\sum_{i=y-s/2}^{y+s/2} p_i,\ 0 \le y < height,\ p_i \in P_H^{(n)}\} \quad (2)$$

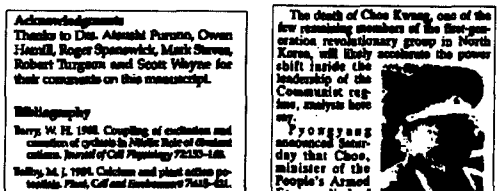
$$F_H^{(n)} = \{f_y | f_y = \frac{\partial D(m_y, m_{y+1})}{\partial x}, 0 \le y < height, m_y \in M_H^{(n)}\} \quad (3)$$

$$Z^{(n)} = \{z \mid (f_z < 0 \ and \ f_{z+1} \ge 0) \ or \ (f_z > 0 \ and \ f_{z+1} \le 0), f_z \in F_H^{(n)}\} \quad (4)$$

$$V = \frac{\sum_{i=1}^{n}(d_i - m)^2}{n} \quad \begin{cases} m = \frac{\sum_{i=1}^{n} d_i}{n} \\ d_i = z_{i+1} - z_i,\ z_i \in Z^{(n)} \end{cases} \quad (5)$$

분할이 더 필요하다고 판단되는 경우, 즉 식 (5)의 값이 높은 경우에는 그림 3과 같이 두 가지로 나누어 생각할 수 있는데, 그림 3 (a)는 수평 투영을 했을 때 흰 공간이 다른 흰 공간보다 커서 그 곳을 분할해야 하는 경우이며, 그림 3 (b)는 수평 투영을 했을 때 검은 공간이 다른 검은 공간보다 커서 그 옆의 흰 공간을 분할해야 하는 경우이다. 이 두 경우에 대해 분할 위치를 찾는 방법은 아래와 같으며, 분할된 두 개의 각 영역에 대해 본 문서 분할 과정을 반복 적용시킨다.

Let W denote the set of the white space of regions, w_i
Let B denote the set of the black space of regions, b_i
Sort the set W(or B) in the increasing order of the magnitude of w_i(or b_i)
$w_{med} \leftarrow$ the $\frac{n}{2}th$ element of W
$b_{med} \leftarrow$ the $\frac{n}{2}th$ element of B
$w_{max} \leftarrow$ the last element of W
$b_{max} \leftarrow$ the last element of B
if $w_i >> w_{med}$ and $w_i = w_{max}$, split w_i
if $b_i >> b_{med}$ and $b_i = b_{max}$, split w_{i-1}



(a) 경우 1 (b) 경우 2

그림 3. 수평 방향으로 분할이 필요한 두 가지 경우

2.3 영역 분류

최소 단위의 최대 동질 영역으로 분할된 각 영역을 분류하는 방법은 아래와 같은 순서로 수행한다.

- 텍스트 영역 분류
  수평 방향으로 주기적 성분을 갖고 있으면 한 단으로 구성된 글자 영역으로 분류하고, 수직 방향으로 주기적 성분을 갖고 있으면 제목이나 부제목, 범례 또는 캡션과 같이 한 행으로 구성된 글자 영역으로 분류한다.

- 선 영역 분류
  끊어진 선을 이어주기 위해 RLSA[6]를 적용한 후에, 그 영역이 하나의 연결 요소로 구성되어 있고 가로와 세로의 비율이 5배 이상이면 선으로 분류한다.

- 표 영역 분류
  어떤 영역의 윗 선과 아랫 선의 길이가 그 영역의 넓이와 비슷하고, 그 영역 안에 윗 선과 아랫 선을 제외한 하나 이상의 선이 존재하면 표 영역으로 분류한다.

- 그림 영역 분류
  위에서 분류되지 않은 나머지 영역들은 모두 그림 영역으로 분류한다.

3. 실험 및 결과 분석

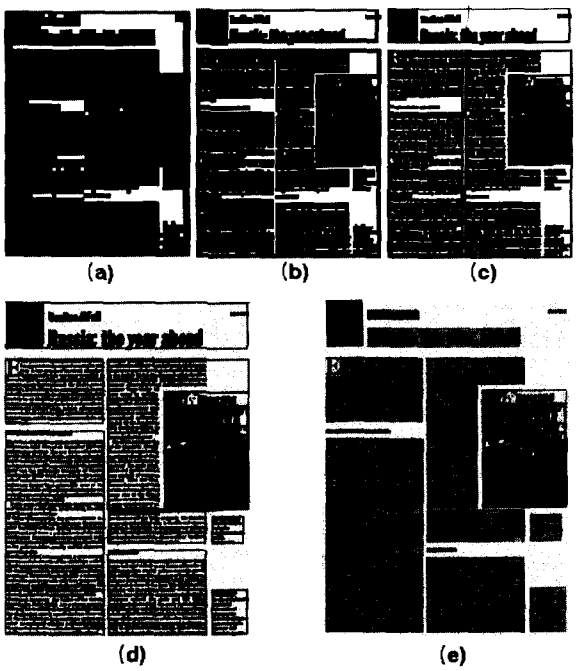


그림 4. 실험 결과 영상의 예: (a) 단계 4 (48x65) (b) 단계 3 (97x130) (c) 단계 2 (194x260) (d) 단계 1 (389x521) (e) 단계 0 (778x1043)

표 1. 영역 위치 판정 및 분류 실험 수행 결과

	N_{total}	$N_{correct}$	정확률 (%)
영역 위치 판정	1559	1523	97.7
텍스트 영역 분류	1473	1469	99.7
그림 영역 분류	139	135	97.1
표 영역 분류	4	4	100
선 영역 분류	101	101	100

본 논문에서 제안된 방법을 Washington 대학의 문서 영상 데이터베이스(UWDB)에 있는 125개의 영상에 대하여 실험한 결과는 표 1과 같다. 그림 4에 각 단계마다 제안된 방법을 적용한 중간 단계를 나타내었으며, 기존의 방법들과 비교 실험한 결과는 그림 5와 같다. 실험은 Pentium 200MHz PC, 64MB 메모리 상에서 수행했으며, 피라미드 구조를 만드는데 평균 0.22초, 구조 분석하는데 평균 1.70초로 총 수행 시간은 평균 1.92초 소요되었다.

만약 어떤 영역이 하나의 글자 행과 몇 개의 단어만으로 구성되어 있으면 그 주기성을 찾기 어렵기 때문에 잘못된 분할 결과를 가져왔다. 또한 이웃하는 몇 개의 영역이 우연히 단일 주기로 이루어져 있으면 하나의 영역으로 결합되는 분할 오류가 발생하였다. 본 논문은 기하학적 구조 분석에 한한 것으로 문서 영상에서 의미있는 전자 문서로 자동 변환하기 위해서는 논리적 구조 분석이 요구된다.

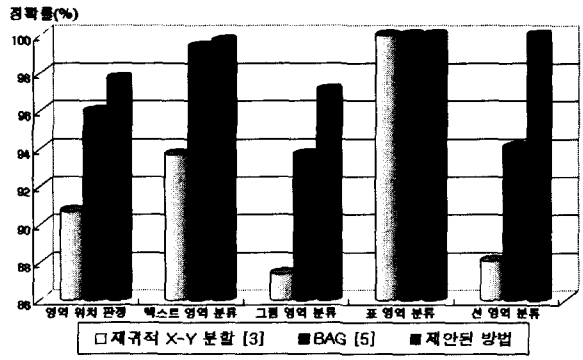


그림 5. 기존 방법들과 제안된 방법의 비교

감사의 글

본 연구는 부분적으로 과학기술부 창의적연구진흥사업의 연구비 지원을 받았음.

참고 문헌

[1] A. Jain and Y. Zhong, "Page Segmentation Using Texture Analysis," *Pattern Recognition*, Vol. 29, pp. 743-770, 1996.

[2] K. Etemad, D. Doermann, and R. Chellappa, "Multiscale Document Page Segmentation Using Soft Decision Integration," *IEEE Trans. on Pattern Anal. and Machine Intell.*, Vol. 19, pp. 92-96, 1997.

[3] J. Ha, R. M. Haralick and I. T. Phillips, "Recursive X-Y Cut using Bounding Boxes of Connected Components," *Proc. of the 3rd Int. Conf. on Document Analysis and Recognition*, Montreal, pp. 952-955, 1995.

[4] A. Antonacopoulos, "Page Segmentation Using the Description of the Background," *Computer Vision and Image Understanding*, Vol. 70, pp. 350-369, 1998.

[5] A. Jain and B. Yu, "Document Representation and Its Application to Page Decomposition," *IEEE Trans. on Pattern Anal. and Machine Intell.*, Vol. 20, pp. 294-308, 1998.

[6] B. Gatos, N. Papamarkos and C. Chamzas, "Skew Detection and Text Line Position Determination in Digitized Documents," *Pattern Recognition*, Vol. 30, pp. 1505-1519, 1997.