2000년 한국정보처리학회 추계 학술발표논문집 제 7 권 제 2 호

웨이블렛 계수의 통계적 이산 분석을 이용한 문서 영상 분할

이인수*, 김민수**, 김우성*, 한광록*
*호서대학교 컴퓨터공학과
**호서대학교 벤처대학원
e-mail : wantpark@dreamwiz.com

Document Image Segmentation by the Statistical Distribution Analysis of Wavelet Coefficients

In-Sue Lee*
Min-Soo Kim**
Woo-Sung Kim*
Kwang-Rok Hahn*
*Dept. of Computer Engineering, Hoseo University
**Dept. of Computer Application Engineering, Hoseo University

요 약

본 논문은 문서 영상에 대해 투영을 사용하여 영역을 나누었고 각 영역에 대해 고주파 밴드의 웨이블렛 계수의 통계적 분산과 히스토그램을 기반으로 한 두 가지 특징을 사용하여 문자와 그림으로 분류하였다. 투영으로 나누어진 영역들에 대해 일정 크기의 블록으로 나누고 두 가지 특징에 따라 문자와 그림으로 분류하였다. 따라서 투영에 의해 나뉜 영역 중 문자와 그림이 혼합되어 의미가 모호한 영역에 대해 잘못 분류되는 가능성을 줄일 수 있었다.

1. 서론

문서를 구성하게 되는 구성 성분은 크게 두 가지 형태로 나누어 볼 수 있다. 문서의 주된 내용을 포함하는 문자 영역과 설명과 이해를 돕기 위해 쓰이는 그림 등의 비문자 영역이다. 일반적으로 문서 영상에서 영상을 구성하는 화소의 대부분을 차지하는 것이 문자 영역이다. 문자의 특징은 화소가 일정한 크기의 사각형 안에 표현이 된다는 것과 각 사각형의 중심부터 인접한 사각형의 중심까지의 거리가 비슷하다는 것이다. 비문자 영역은 문서의 구성 요소 중 문자 성분들이 포함하고 있는 내용의 이해를 돕기 위한 것이거나 설명을 쉽게 하기 위해 쓰인다[5].

분류 중에서 가장 많은 주목을 받고 있는 부분이 문자와 그림을 분리하는 것이다. 그림이란 사진과 같이 연속적인 색조를 갖는 이미지를 의미한다. 문서란 일반적인 문자들, 테이블 그리고 그래프를 의미한다. Wolrd Wide Web 의 점진적인 인기로 인해 웹 정보가 점점 더 자주 활용되고 있다. 그리고 많은 웹 페이지가 문자와 그림의 혼합으로 되어있기 때문에 문자와 그림 즉 비문자의 분류는 다양한 처리를 하는데 효율성을 증가시킨다.

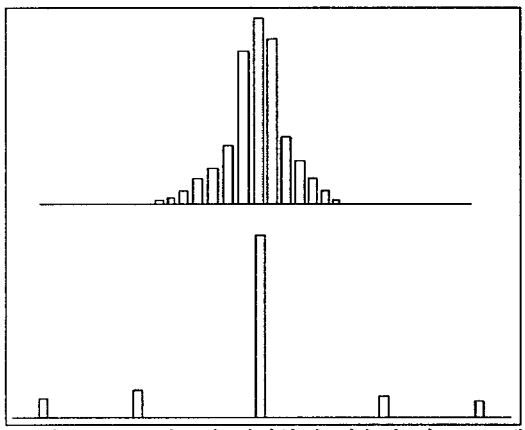
문서를 분할하는 방법에는 경계 추출을 이용하는 방법, 투영에 기반을 둔 분할 방법, 확대 반복에 의한 분할 방법, 연결화소를 이용한 분할 방법, 소블록 분할에 효과적인 RLSA 에 의한 방법 등 여러 알고리즘이 있다.

본 논문에서는 문서 영상에 대해 먼저 투영을 수

행하였다. 수평, 수직 방향으로 각각 한번씩 수행하여 영역을 나누고 각 영역에 대해 웨이블렛 변환 기법을 사용하여 표와 그래프를 포함하는 문자 영역과 사진과 같은 이미지를 의미하는 그림 영역 즉 비문자 영역으로 분류하였다.

웨이블렛 변환을 사용하는 일반적인 에플리케이션은 단순히 웨이블렛 계수의 통계만을 사용하지만 본 논문은 웨이블렛 계수의 분산 패턴과 히스토그램을 사용하였다. 또한 분할된 영역에 대해 일정 크기의 블록으로 나누고 각각의 블록들을 독립적으로 문자와 비문자 영역으로 분류시켰다.

본 논문은 2 장에서 투영과 문자와 그림 영역을 분류하기 위해 사용한 두 가지 특징을 정의한다. 3 장에서는 분류 알고리즘을 설명한다. 4 장에서는 실험 및 결과 순으로 논문을 기술한다.



(그림 1) LH 밴드의 웨이블렛 계수의 히스토그램. 위 : 그림 영상, 아래 : 문서 영상

2. 영역 분류에 사용한 투영과 웨이블렛 특징

2.1 투영

문서 영상을 투영 알고리즘을 사용하여 후보 영역을 추출하고, 추출된 영역은 일정 크기의 블록으로 나누고 각 블록에 대해 독립적으로 배경, 문자 그리고 그림으로 분류하였다.

투영에 의한 방법은 주로 사각형의 블록으로 이루어진 경우에 처리가 빠른 장점을 가지고 있으며, 문서 분석 방법들 중에서도 계층적 영역 분할 방법에 속하며, 하향식으로 처리된다. 처리 방법은 블록과 블록 사이의 여백을 기준으로 한 번은 수평 분할, 한 번은 수직 분할하는 방식이다. 본 논문에서는 수평, 수직 투영을 각각 한 번만 수행하여 후보 영역을 추출하였다[4].

2.2. Goodness of Fit to the Laplacian Distribution

그림 영상에 대한 고주파 밴드 (LH, HL 그리고 HH 밴드) 내의 웨이블렛 계수값을 살펴보면 Laplacian distribution 을 따르는 경향을 알 수 있다. 이런 추정은 몇몇 에플리케이션에서는 문제가 될 수 있지만 goodness of fit 이란 의미를 갖는 문자로부터 연속 색조 화상을 구분할 수 있다는 것은 아주 효율적이다. 그림 1 은 그림과 문자 영상의 LH 영역의 계수들의 히스토그램이다.

또 다른 중요한 특징은 관찰된 분포(observed distribution)의 연속성이다. 문자 영상의 웨이블렛 계수 히스토그램은 몇 개의 이산된 값이 있지만 그림 영상의 웨이블렛 계수 히스토그램은 조금 더 많은 연속적인 분포를 이룬다.

Observed distribution 과 Laplacian distribution 과 일치 정도를 측정하기 위해서 실험 크기인 N 으로 x^2 를 정규화시킨 $\bar{x}^2$ 를 구하였다. 데이터 범위를 k 로 나누고 fi 는 i 번째 상대 도수 그리고 Fi 는 Laplacian distribution 에 따른 i 번째 확률이다. $\bar{x}^2$ 를 계산하면

$$\bar{x}^2 = x^2 / N = \sum_{i=1}^{k} (fi - Fi)^2 / Fi$$

이다.

그림 영상은 문자 영상과 비교해볼 때 $\bar{x}^2$ 값이 매우 낮다는 것을 알 수 있다[1][2].

2.3. Likelihood of Being a Highly Discrete Distribution

매우 집중된 값들에 의해서 구성된 웨이블렛 계수들의 밀집도를 L 로 표기한다. L 를 계산하기 위해서 zone 들의 집합으로 데이터 범위를 나눈다. 데이터 범위의 양끝의 사소한 차이를 제거한 zone 은 $[t_i, t_{i+1}]$, i = 0, 1, ..., r-1 로 표기하며 zone 의 집중 레벨을 β_i 로 표기한다. Zone 내의 최대값은 t*로 가정하고 t* 주위의 가로 영역을 w 라 한다. 제한된 데이터 범위를 설정하기 위해 t*를 중심으로 양끝 위치를 τ_1 , τ_2 라 한다.

$$\tau_1 = t^* - w, \ t_i < t^* - w \ \ or \ \ \tau_1 = t_i, \ otherwise$$

$\tau_2 = t^* + w, \quad t^* + w < t_{i+1} \quad or \quad \tau_2 = t_{i+1}, \quad otherwise$

$[\tau_1, \tau_2]$ 내의 확률 $p_i = \int_{\tau_1}^{\tau_2} h(t)dt$

zone $[t_i, t_{i+1}]$ 내의 확률 $p_i' = \int_{t_i}^{t_{i+1}} h(t)dt$

집중 레벨 β_i 는 $\beta_i = p_i / p_i'$

마지막으로 L 을 계산하면

$$L = \sum_{i=0}^{r-1} p_i \cdot \beta_i \quad (0 \le L \le 1)$$

이다.

3. 알고리즘

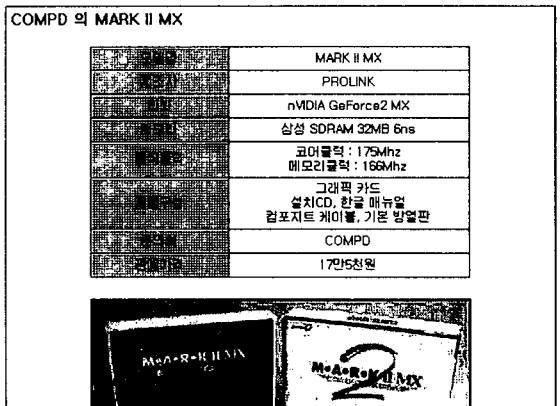
문서 영상에 대해 Haar 웨이블렛 변환을 사용하였고 블록을 분류하기 위해 LH 와 HL 밴드의 웨이블렛 계수를 사용하였다. 만일 분산이 0(zero)라면 그 해당 블록은 배경으로 분류된다. 그렇지 않다면 $\bar{x}^2$ 와 L 을 계산한다. $\bar{x}^2$ 와 L 이 일정 조건을 만족시키면 해당 블록을 문자 또는 그림 영역으로 분류하였다. 그래도 분류되지 않으면 비결정 영역으로 설정하고 해당 블록을 사등분한 서브블록으로 나누고 각 서브블록에 대해 다시 독립적으로 분류하였다. 서브블록의 특징값 $\bar{x}^2$ 와 L 이 일정 조건을 만족시키면 문자 또는 그림 영역으로 분류하였다. 그래도 분류가 되지 않으면 해당 서브블록을 분류하기 위해 내용 정보를 활용하여 분류하였다. 블록이 문자 블록과 그림 블록에 인접하다면 그 블록은 두 분류의 경계부분에 존재하는 것이므로 서브블록의 특징값 $\bar{x}^2$ 와 L 만을 고려하여 분류하였다. 서브블록이 문자 블록 내지 그림 블록과 인접하지 않으면 그림으로 분류하였다 [3][6]. 더욱 정확한 분류를 위해 [1]에서 제시한 임계값을 변형한 X1, X2, L1 그리고 L2 값을 사용하였다.

1. (a) 이미지에 대해 Haar 웨이블렛 변환
   (b) 이미지를 32 x 32 크기로 나눈다. 각 블록은 B_j, j = 1, ..., J 로 표현한다. 데이터 v_j = { x_1, ..., $x_{16x16x2}$ }에서 x_1, ..., x_{16x16} 은 해당 블록의 LH 밴드 웨이블렛 계수이고 $x_{16x16+1}$, ..., $x_{16x16x2}$ 는 해당 블록의 HL 밴드 웨이블렛 계수이다.
   (c) 1 → j
2. 블록 B_j에 대해 분산 σ_j^2 을 구한다.
3. 만일 σ_j^2 가 0 이면
   블록 B_j는 배경으로 분류, j + 1 → j, goto 2
   그렇지 않으면,
   (a) v_j에 대한 $\bar{x}^2$ 와 L 를 계산한다.
   (b) 만일 $\bar{x}^2$ < X1, L < L1 이면, X1 과 L1 은 임계 값, 블록 B_j 는 그림으로 분류된다. goto 3e.
   (c) 만일 L >= L1 이면 블록 B_j는 문자로 분류된다. goto 3e.
   (d) 블록이 분류되지 않으면
      i. B_j 를 B_{j1}, B_{j2}, B_{j3} 그리고 B_{j4} 네개의 서브블록으로 나눈다. 데이터 v_j 에 대해서도 $v_{j,1}$, $v_{j,2}$, $v_{j,3}$ 그리고 $v_{j,4}$로 나눈다.
      ii. 각 서브블록 $v_{i,l}$ 에 대해 $\bar{x}^2$ 와 L 를 계산한다. l = 1, ..., 4
      만일 $\bar{x}^2$ < X2, L < L2 이면, X2 와 L2 는 임계값, 블록 B_{jl}는 그림으로 분류한다. goto 3e.
      만일 L > (L1 + L2) / 2 이면 블록 B_{jl}는 문자로 분류한다. goto 3e.
      블록이 분류되지 않으면
      A. 블록 B_j 가 문자 블록과도 인접하고 그림 블록과도 인접하다면, $\bar{x}^2$ < X2 을 만족하면 B_{jl} 은 그림으로 분류한다. $\bar{x}^2$ < X2 을 만족하지 않으면 문자로 분류한다. goto 3e.
      B. 블록 B_j 가 문자 블록과도 그림 블록과도 인접하지 않다면 B_{jl} 을 그림으로 분류한다.
   (e) j + 1 → j, goto 2.

4. 실험 및 결과

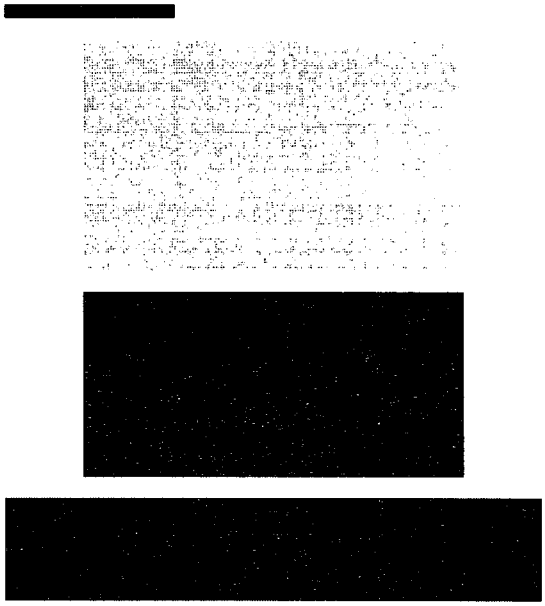
문서 영상 그림 2 에 투영을 수행하여 문서를 분할하고(그림 3), 분할된 영역에 대해 웨이블렛 계수의 통계적 분산을 이용하여 문서 영상을 분류하였다(그림 4). 그림 2 를 보면 테이블 영역 내의 바탕색이 흰색이 아니어서 일반적인 문서 영상 전처리 과정에서는 그림으로 오인식할 수 있으나 본 논문에서는 웨이블렛 계수를 이용하였기 때문에 문자 영역으로 올바르게 인식하였다. 그림 4 의 문서 영상 분할 결과

문자, 테이블 그리고 그림 영역으로 분할되었다. 또한 바탕색이 문서를 분류하는데 있어 영향을 끼치지 못한다는 것을 알 수 있다. 향후 이 결과를 토대로 컬러 문서 영상에 대한 OCR 분야에 확대 적용할 예정이다.



(그림 2) 테스트 문서 영상

(그림 3) 투영을 수행한 영상



(그림 4) 문서 영상 분할 결과

참고문헌

[1] Jia Li and Robert M. Gray. "Text and Picture Segmentation by the Distribution Analysis of Wavelet Coefficients", Volume 3 of 3. ICIP, 1998.

[2] Jia Li and Robert M. Gray. "Context Based Multiscale Classification of Images", Volume 3 of 3. ICIP, 1988.

[3] M. K. Mandal, T.Aboulnasr, "Fast Wavelet Histogram Techniques for Image Indexing", Computer Vision and Image Understanding, Vol. 75, 1999.

[4] 정미영, 김동근, 황치정, "투영 윤곽을 이용한 문서의 효율적인 블록 분할 기법", 정보처리학회 논문지 제 4 권 제 2 호, 1997.

[5] 문승원, "정보검색 시스템을 위한 자동 문서인식기에 대한 연구", 호서대학교 컴퓨터공학과, 석사학위 논문. 1998.

[6] Raghuveer M. Rao, Ajit S. Bopardikar, "Wavelet Transforms Introduction to Theory and Applications", Addison-Wesley, 1998