DOI: 10.3745/KIPSTB.2009.16B.5.377

지역적 엔트로피와 텍스처의 주성분 분석을 이용한 문서영상의 분할 및 구성요소 분류

김 보 람[†] · 오 준 택[††] · 김 욱 현[†††]

요 약

본 논문은 지역적 엔트로피 기반의 히스토그램을 이용한 문서영상의 분할과 텍스처 기반의 주성분 분석을 이용한 구성요소인 글자, 그림, 그래프 등의 구성요소 분류방안을 제안한다. 지역적 엔트로피와 히스토그램을 이용함으로써 문서영상의 다양한 변형이나 잡음에 강건하며 빠르고 손쉬운 이진화가 가능하다. 그리고 문서영상 내 존재하는 구성요소들이 각기 다른 텍스처 정보를 가지고 있다는 것에 착안하여 각 분할 영역의 텍스처 정보를 기반으로 주성분분석을 수행하였으며 이를 통해 사전에 구성요소들에 대한 구조정보를 설정할 필요가 없다는 장점을 가진다. 실험결과에서 다양한 문서영상의 분할 및 분류결과를 보였으며, 기존 방법보다 우수한 성능을 가져 그 유효함을 보였다.

키워드 : 문서영상분할, 컨텐츠 분류, 질감, 주성분 분석

Segmentation and Contents Classification of Document Images Using Local Entropy and Texture-based PCA Algorithm

Bo-Ram Kim[†] · Jun-Taek Oh[††] · Wook-Hyun Kim[†††]

ABSTRACT

A new algorithm in order to classify various contents in the image documents, such as text, figure, graph, table, etc. is proposed in this paper by classifying contents using texture-based PCA, and by segmenting document images using local entropy-based histogram. Local entropy and histogram made the binarization of image document not only robust to various transformation and noise, but also easy and less time-consuming. And texture-based PCA algorithm for each segmented region was taken notice of each content in the image documents having different texture information. Through this, it was not necessary to establish any pre-defined structural information, and advantages were found from the fact of fast and efficient classification. The result demonstrated that the proposed method had shown better performances of segmentation and classification for various images, and is also found superior to previous methods by its efficiency.

Keywords : Document Image Segmentation, Contents Classification, Texture, PCA(Principal Component Analysis)

1. 서 론

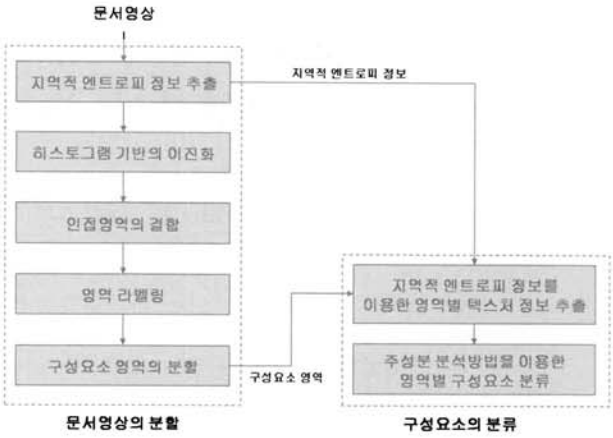
문서영상은 최근 컴퓨터를 이용한 멀티미디어 정보처리와 데이터베이스 시스템의 발전과 더불어 그 사용량이 급격히 증가하고 있으며, 사용자들에게 더 나은 서비스를 제공하기 위한 하나의 방안으로 구성요소의 인식에 따른 문서영상의 자동분류 및 인식은 필수적인 연구과제로 요구된다. 그러나 문서의 유형과 특징, 그리고 구성요소의 종류 등이 매우 다양하기 때문에 현재 자동분류는 어려운 실정이다. 이에 일정한 서식의 문서영상들에 대한 연구가 활발히 이루어지고 있으며, 광학문자인식 분야로 확장되어 글자 자동 판독, 페이지 인덱싱(page indexing), 멀티미디어 검색뿐만 아니라 필기인식 범위까지 활용되고 있다[1].

일반적인 문서영상 처리는 문서영상에 존재하는 구성요소의 분할과 분류로 구성된다. 문서영상의 분할은 문서영상 내 구성요소들의 분류 및 인식을 손쉽게 처리하기 위해 전경부분인 구성요소들과 배경부분들을 분할하는 과정이며, 연결요소 분석에 의한 방법[2], 투영프로파일 분석방법[3, 15] 및 조세(coarse-fine) 단계에 의한 방법[4] 등의 기존 연구들이 있다. 문서영상 내 구성요소의 분류는 분할된 영역들을 글자, 그림, 그래프, 표 등의 다양한 구성요소들을 분류하는 과정이며, 현재에는 구성요소에 대한 구조적 정보[5, 16], 통계적 정보[6], 구성요소의 반복성 등을 특징정보로 하여 신

※ 이 연구는 2008학년도 영남대학교 학술연구조성비에 의한 것임
† 준 회 원 : 영남대학교 컴퓨터공학과 박사과정
†† 준 회 원 : 한국조폐공사 기술연구원 연구원
††† 정 회 원 : 영남대학교 전자정보공학부 교수
논문접수 : 2008년 12월 9일
수 정 일 : 1차 2009년 6월 19일, 2차 2009년 7월 27일
심사완료 : 2009년 7월 28일

경망을 이용한 분류 등이 연구되고 있다[7-9]. 이러한 방법들은 문서영상을 전경부분과 배경부분으로 분할하거나 텍스트 또는 비텍스트 영역만으로 분류하기 때문에 문서영상에 존재하는 다양한 구성요소들을 고려하지 못하는 문제점이 존재한다. 이에 Hang Wang 등[10]은 사전에 정의된 블록들의 분류, 이의 합병 및 분할에 의한 하향식 분할알고리즘과 분할된 구성요소들의 구조 및 통계적 정보를 이용한 알고리즘을 제안하였다. Laura Caponetti 등[11]은 다양한 해상도에서 구성요소들의 분할과 분류를 위해 신경-퍼지(neuro-fuzzy) 네트워크를 이용한 알고리즘을 제안하였으며, 화소 단위의 처리를 수행하기 때문에 문서영상에 대한 다양한 변형에도 비교적 나은 분할과 분류결과를 보였다. Zheru Chi 등[12]은 세선화 알고리즘에 의한 배경부분의 분할과 계층적 접근방식을 이용한 구성요소의 분류알고리즘을 제안하였다. 구성요소들의 분류를 위해 신경망, 상호상관(cross-correlation), 콜모고로프 복잡도(Kolmogorov complexity) 등을 이용한 다단계 분류과정을 거치기 때문에 비교적 높은 분류성능을 획득할 수 있으나 많은 처리과정을 요구한다. M-W Lin 등[13]은 사전에 정의된 블록들에 대해서 추출한 텍스처 정보를 기반으로 K-Mean 알고리즘에 의한 분할과 구성요소들의 사전정보를 이용한 분류알고리즘을 제안하였다. 그러나 여전히 문서영상에 존재하는 다양한 구성요소들의 특성에 따라 민감한 분류결과를 보였으며, 사전에 정의된 블록 단위의 클러스터링 분할을 수행하기 때문에 화소 단위의 방법보다 부정확한 분할결과를 보였다. 그리고 이러한 방법들은 글자, 그림, 그리고 배경영역에 대한 분류만을 수행하였기 때문에 보다 나은 문서영상의 이해와 분석을 위해서는 더 많은 구성요소들의 분할과 분류에 대한 연구가 필요하다.

본 논문은 화소 단위의 지역적 엔트로피를 이용한 문서영상의 분할과 텍스처 정보의 주성분 분석에 의한 구성요소 분류알고리즘을 제안하며, 전체적인 처리과정은 그림 1과 같이 문서영상 분할단계와 구성요소 분류단계로 크게 구성된다. 문서영상 분할단계에서는 지역적 엔트로피 정보를 이용한 N.Otsu의 방법[14]에 의해 문서영상을 전경부분과 배경부분으로 이진화한 후 전경부분에 대해서 연결 영역 라벨링 알고리즘을 이용하여 구성요소들에 대한 분할을 수행한다. 연결요소 및 투영프로파일 분석방법과 달리 지역적 엔트로피는 화소들간의 밝기 정보에 따른 특징으로, 영상 또는 영역이 다양한 밝기 값으로 구성되었거나 회전이 가미되어도 효율적인 분할이 가능하다. 또한, 문서를 영상화할 때 추가되어지는 잡음에도 지역적 엔트로피의 특성에 의해 강건한 분할이 가능하다. 그리고 구성요소 분류단계에서는 분할된 영역들의 지역적 엔트로피 정보를 기반으로 추출한 텍스처 정보에 대해서 주성분 분석을 이용하여 분할영역들을 글자, 그림, 그래프, 표 등의 구성요소들로 분류한다. 제안된 분류방법은 문서영상 내 구성요소들에 대한 구조정보를 사전에 설정할 필요가 없으며, 분류기로 이용되는 주성분 분석은 신경망과 달리 적은 수의 분류정보만으로도 간단한 처리과정에 의해 효과적으로 구성요소들을 분류할 수 있다.



(그림 1) 제안방법의 처리과정

또한, 화소 단위의 분할과 영역 단위의 분류를 수행함으로써 좀 더 정확한 분할과 분류성능을 기대할 수 있다.

본 논문의 구성은 다음과 같다. 2절에서는 본 논문에서 제안한 지역적 엔트로피를 이용해 문서영상을 분할하는 방법을 소개한다. 3절에서는 본 논문에서 제안한 분할된 영상을 영역별로 분류하는 방법을 소개하고, 4절에서는 실험 결과와 성능 평가를 분석한다. 마지막으로 5절에서 결론을 맺는다.

2. 지역적 엔트로피 기반의 히스토그램을 이용한 문서영상 분할

문서영상의 분할은 문서 내에 존재하는 구성요소들을 분류 및 인식하기 이전에 수행한다. 본 논문에서는 히스토그램 기반으로 임계치를 결정하여 이진화를 수행한 후, 전경부분에 대해서 연결 영역 라벨링 알고리즘을 이용하여 구성요소들을 분할한다.

2.1 화소 단위의 지역적 엔트로피 추출

문서영상의 이진화는 추후에 수행할 분할 및 분류과정의 정확도에 영향을 미치는 중요한 단계이다. 본 논문은 N.Otsu의 방법을 이용하여 문서영상을 이진화하며, 화소의 밝기 값이 아닌 지역적 엔트로피 정보를 이용함으로서 보다 개선된 결과를 얻을 수 있었다. 화소의 밝기 값은 문서를 영상화할 때 발생될 수 있는 빛의 영향에 민감하기 때문에 이진화 과정을 수행하는데 있어 오류를 야기할 수 있다. 그러나 지역적 엔트로피는 이웃화소들의 밝기값을 이용함으로써 빛의 밝기변화나 잡음에 강건한 특성을 가진다. 식 (1)은 엔트로피의 정의를 나타낸다.

$$entropy = -\sum_{i=0}^{255} P(i)\log(P(i)) \qquad P(i) = \frac{N_i}{N} \tag{1}$$

i는 밝기 값으로 회색 영상의 경우 0에서 255까지의 범위를 가지며, $P(i)$는 i번째 밝기 값에 대한 확률값이다.

1	2	4
3	5	3
6	3	2

(a) 엔트로피 ≒ 0.7

1	3	7
1	7	3
7	3	1

(b) 엔트로피 ≒ 0.47

(그림 2) 밝기분포에 따른 엔트로피의 변화

그리고 N_i와 N은 각각 i번째 밝기 값을 가지는 화소수와 전체 화소수를 나타내며, 이진화를 위한 히스토그램 입력 값으로 이용하기 위해 0~255사이의 값으로 정규화하였다.

(그림 2)는 다른 밝기분포를 가지는 3×3 크기의 영상샘플에 대한 엔트로피를 보여준다. 그림 2(a)는 그림 2(b)과 비교하였을 때 다양한 밝기 값으로 구성되어 있고 비동질한 영역임을 알 수 있다. 또한, 각각의 측정된 엔트로피 값은 약 0.7, 0.47로서 그림2(a)의 값이 더 큰 것을 알 수 있다. 즉, 문서영상의 구성요소에 해당하는 그림, 표, 글자, 그래프 등의 전경영역은 비동질적이므로 비교적 큰 엔트로피 값을 가지며, 배경영역은 동질적인 영역이므로 비교적 작은 엔트로피 값을 가진다. 이러한 지역적 엔트로피의 특성을 이용하면 문서영상을 효과적으로 이진화할 수 있다. 즉, 문서영상의 회전이나 기울어짐에 불변하며, 문서를 영상화할 때 발생할 수 있는 잡음이나 다양한 밝기 값으로 구성된 영상이라도 보다 나은 이진화 결과를 획득할 수 있다.

2.2 문서영상의 이진화와 구성요소의 분할

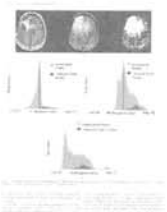
앞서 추출한 지역적 엔트로피 정보를 이용하여 문서영상의 이진화를 수행하며, 이때 N.Otsu의 방법을 이용한다. N.Otsu의 방법은 히스토그램을 이용하는 대표적인 이진화 방법 중에 하나로서, 임의의 임계값으로 전경부분과 배경부분을 분할한 후에 각 부분의 밝기 값에 대한 분산이나 전경부분과 배경 부분간의 밝기 값에 대한 분산을 이용하여 얻은 최적의 임계값으로 이진화를 수행한다[14]. 최적의 임계치는 전경부분과 배경부분간의 분산값인 외부거리가 최대이거나 전경부분의 분산값과 배경부분의 분산값의 합하여 나타내는 내부거리가 최소일 때 결정된다.

이진화 작업을 수행한 후, 전경부분으로 분류될 영역의 수를 줄여서 처리 시간을 줄이고 작업을 간소화하기 위해 형태학적 기법 중에 하나인 닫힘 연산(closing operation)을 수행한다. 그리고 전경부분의 최종 분할을 위해서 간단하고 구현하기 쉬운 연결 영역 라벨링 알고리즘을 이용한다.

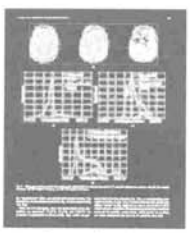
(그림 3)은 문서영상의 분할과정을 보여준다. 그림 3(a)는 원본영상이며, (그림 3) (b)는 엔트로피 변환영상을 나타낸다. (그림 3) (c)는 엔트로피 변환영상에 대해서 N.Otsu의 방법을 이용하여 생성한 이진영상이다. (그림 3) (c)를 통하여 다양한 밝기 값으로 이루어진 문서영상도 정확하게 이진화가 수행되는 것을 알 수 있다. (그림 3) (d)는 이진영상에 대해서 닫힘 연산을 수행한 영상으로, 글자 부분의 흰 공간이 메워지며 인접영역이 합쳐지는 것을 알 수 있다. (그림 3) (e)는 연결 영역 라벨링 알고리즘에 의해 분할된 영상이며, (그림 3) (f)는 분할된 영역의 위치정보를 이용하여, 원본영상에 경계상자(bounding box)를 생성한 영상이다. 경계상자는 분할영역의 텍스처 정보를 추출하기 위한 범위일 뿐만 아니라 분할 영역들 간의 공간적 관계를 고려하기 위한 수단으로 이용되며, 하나의 분할영역에 대한 경계상자 내부에 또 다른 분할영역의 경계상자가 포함될 경우 이것은 하나의 분할영역으로 처리된다.

3. 텍스처 기반의 주성분 분석을 이용한 구성요소 분류

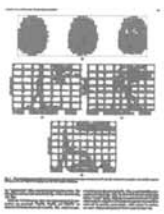
전경영역을 대상으로 분할된 영역들을 구성요소 별로 분류하기 위해서 지역적 엔트로피 변환영상을 이용하여 분할



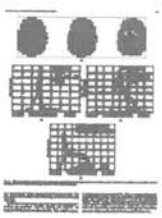
(a) 원본영상



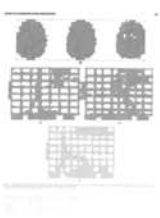
(b) 엔트로피 변환영상



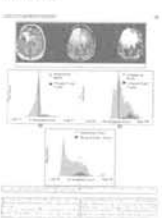
(c) 지역적 엔트로피를 이용한 이진영상



(d) 닫힘 연산에 의한 영상



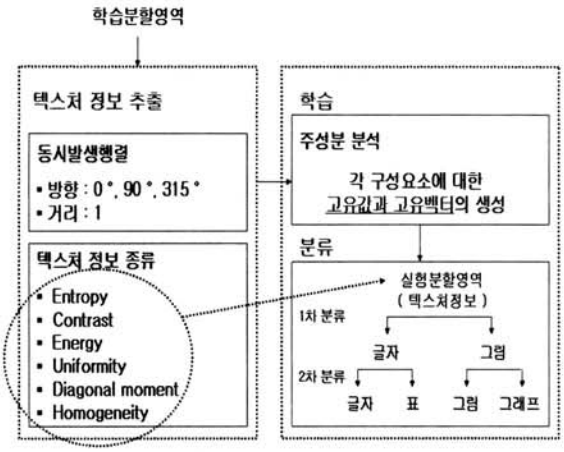
(e) 연결 영역 라벨링 알고리즘에 의한 분할영상



(f) 최종분할영상

(그림 3) 문서영상의 이진화와 전경영역 분할 과정

영역들에 대한 동시발생행렬을 생성한다. 동시발생행렬을 이용하여 추출한 텍스처 정보는 주성분 분석의 입력 데이터이며, 문서영상 내 구성요소(글자, 표, 그림, 그래프)들을 분류하기 위한 특징정보로 이용된다. (그림 4)는 분할영역들로부터 추출한 텍스처 정보를 이용하여 주성분 분석에 의해 구성요소들로 분류하는 과정을 보여준다.



(그림 4) 문서영상 내 구성요소들의 분류과정

3.1 분할된 영역별 텍스처 정보 추출

본 논문은 문서를 이루는 구성요소들의 밝기값 변화가 서로 다르다는 점에 착안하여 텍스처 정보를 분류특징으로 이용한다. 텍스처 정보를 추출하기 위한 수단으로 동시발생행렬을 이용하며, 분류특징으로는 식 (2)-식 (7)과 같이 엔트로피(entropy), 모멘트(diagonal moment), 동질성(homogeneity), 대비(contrast), 에너지(energy), 균일성(uniformity)등을 이용한다.

$$\text{Entropy} = -\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} P(i,j)\log(P(i,j)) \tag{2}$$

$$\text{Diagonal moment} = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \frac{|i-j|P(i,j)}{2} \tag{3}$$

$$\text{Homogeneity} = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \frac{P(i,j)}{1+(i-j)^2} \tag{4}$$

$$\text{Contrast} = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} (i-j)^2 P(i,j) \tag{5}$$

$$\text{Energy} = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} P(i,j)^2 \tag{6}$$

$$\text{Uniformity} = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \frac{P(i,j)}{1+|i-j|} \tag{7}$$

먼저, (그림 3) (f)와 같이 분할된 각 영역들을 감싸는 최외각 사각형 영역의 지역적 엔트로피 정보를 이용하여 거리 1의 변위벡터(0도, 90도, 315도)를 이용한 3개의 동시발생행렬을 생성한다. 그리고 동시발생행렬에 대해서 식 (2)-식 (7)을 이용하여 텍스처 정보를 추출한다. i와 j는 엔트로피 값을 이용한 동시발생행렬에서의 위치값으로써 0~255사이의 값을 가진다. $P(i,j)$는 분할영역의 동시발생행렬의 (i,j)에 존재하는 계수이다. N은 동시발생행렬의 크기이며, 엔트로피 변환영상이 회색조 영상이므로 N은 256이 된다.

3.2 주성분 분석에 의한 구성요소의 분류

앞서 분석한 텍스처 정보를 바탕으로 주성분 분석 알고리즘을 이용하여 분할된 영역들을 글자, 표, 그림, 그래프 등의 구성요소들로 분류한다. N차원을 갖는 학습 자료의 입력벡터 집합 $\overline{X}=[\overline{x^1}|\overline{x^2}|\overline{x^3}\cdots|\overline{x^P}]$ 을 바탕으로 식 (8)와 같이 입력데이터의 평균(m)를 구한다.

$$m=\frac{1}{P}\sum_{i=1}^{P} x^i \tag{8}$$

P는 구성요소별 학습에 사용된 정보의 개수를 의미하며, x^i는 6가지의 텍스처 정보를 가지는 6차원 벡터이다. 그리고 입력벡터에 대한 영의 평균 즉, 입력 벡터와 평균 벡터의 차$(\overline{x^i})$ 및 입력 벡터의 전체 영의 평균 집합인 $\overline{X}$의 전치 행렬을 바탕으로 한 공분산 행렬을 생성하며, 그 정의는 식 (9)과 같다.

$$\overline{x^i}=x^i-m,\ \Omega=\overline{X}\overline{X^T} \tag{9}$$

위의 과정을 통하여 글자, 그림, 표, 그래프, 각각의 정보로 만들어진 4개의 공분산 행렬과 2단계 분류를 위해 추가로 추출한 2개의 공분산 행렬인 글자 및 표에 대한 공분산 행렬과 그림 및 그래프에 대한 공분산 행렬을 생성한다. 그리고 공분산 행렬의 직교 정규화 고유 벡터 v를 이용하여 고유값 $\lambda_i=[\lambda_{i1},\lambda_{i2},\lambda_{i3}\cdots,\lambda_{iN}]$에 따른 고유벡터 v_i를 구할 수 있다. 이러한 고유값, 고유벡터, 그리고 공분산 행렬은 식 (10)과 같은 성질을 가진다.

$$\Omega v_i=\lambda_i v_i \quad i=1,2,...N \tag{10}$$

각각의 공분산 행렬로부터 추출한 가장 큰 고유치의 고유벡터는 각 구성요소에 대한 주성분을 나타내며, 이를 바탕으로 분류를 수행한다. 본 논문에서 제안하는 분할영역들의 분류과정을 간략히 살펴보면, 다음과 같다. 먼저, 글자, 표, 그림, 그래프에 해당하는 각 구성요소들의 텍스처 정보에 대해서 위의 식을 이용하여 주성분 분석을 수행하고, 각 구성요소에 해당하는 고유벡터를 추출한다. 그리고 식 (11)과 같이 분류하고자 하는 분할된 영역의 텍스처 정보(X)와 학습 시 추출한 구성요소들의 평균(M_j)간의 차를 각 구성요소들에 대한 고유벡터(V_j)에 투영한 후 그 결과(Y_j)에 따라 정의된 각 구성요소에 얼마나 인접한지를 알 수 있다. 본

논문에서는 각 구성요소들마다 3개의 동시발생행렬을 이용하므로 3개의 고유벡터를 생성한다. 그러므로 각 구성요소마다 3개의 투영 결과값을 얻을 수 있으며, 최종적인 투영결과(Y)는 유클리디언 거리측정방법(Euclidean distance)에 의해 계산된다.

$$Y_j = V_j(X - M_j) \tag{11}$$

그리고 최종 투영결과(Y)를 이용하여 분할영역들을 구성요소로 분류한다. 본 논문에서는 두 단계의 분류과정을 수행한다. 첫 번째 단계에서는 글자와 표의 특징을 모은 정보, 그림과 그래프의 특징을 모은 2가지 정보를 이용함으로서 해당 영역이 글자 영역인지, 그림영역인지를 판단한다. 그리고 2번째 단계를 통해 글자 영역으로 분류된 영역은 글자 영역이나 표 영역으로 분류하고, 그림 영역으로 분류된 영역은 그래프 혹은 그림 영역으로 분류한다. 이를 통해 복잡한 그래프 영역이 표 영역으로 오분류되거나 복잡한 그림 영역이 글자 영역으로 오분류되는 문제점 등을 해결 할 수 있다.

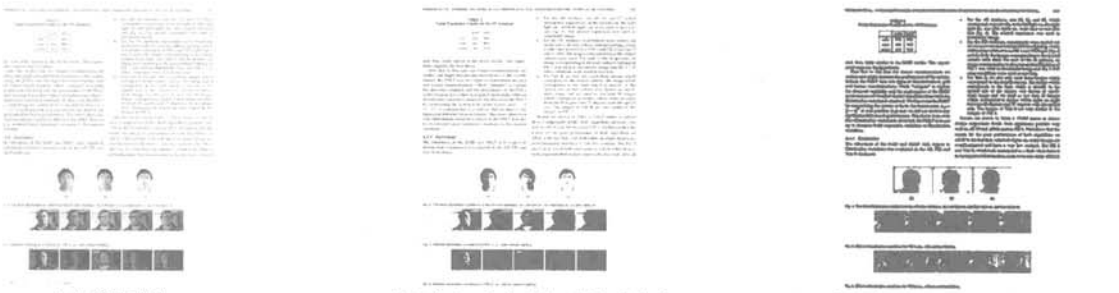
4. 실험 및 결과

본 장에서는 다양한 실험영상들을 대상으로 실험을 수행하고 결과를 분석한다. 제안한 방안들은 Pentium Ⅳ 2.8GHz 시스템 상에서 Visual C++ 6.0을 이용하여 구현하였으며, 실험영상은 다양한 구성요소들을 포함하는 PDF 문서파일을 캡쳐 변환한 720 × 880 크기의 150여개 영상들로 구성되며, 주성분 분석을 위한 학습데이터로는 글자, 그림, 표, 그래프 등의 구성요소들에 대해서 각 50여개를 이용하였다.

(그림 5)는 화소의 밝기 값과 지역적 엔트로피를 대상으로 N.Otsu의 방법에 의해서 이진화된 영상을 보여준다. (그림 5) (a)는 원본영상이며, (그림 5) (b)와 (그림 5) (c)는 화소의 밝기 값과 지역적 엔트로피를 이용하여 이진화된 영상이다. 기존의 화소 밝기 값을 이용한 (그림 5) (b)의 이진영상은 빛의 변화로 인하여 전경부분의 밝기 값이 왜곡되었으며, 이에 밝기 값이 다양하게 존재할 경우 잘못된 이진화의 결과를 초래하였다. 그러나 밝기 값의 변화에 따른 지역적 엔트로피를 이용한 이진 영상인 (그림 5) (c)는 밝기 값의 변화에도 무관하게 이진화가 보다 정확한 결과를 보였다. 또한, 밝기 값에 무관하게 그림을 둘러싸고 있는 선이나 표의 테두리도 보다 정확하게 이진화되는 것을 알 수 있었다.

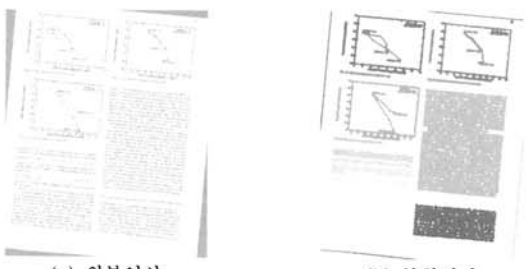
(그림 6)은 회전된 문서영상에 대해서 지역적 엔트로피를 이용한 영상분할의 결과를 보여준다. (그림 6) (a)는 5도 회전된 원본영상이며, (그림 6) (b)는 분할된 결과영상이다. 기존의 구조적 방식을 이용한 방법들과 달리 지역적 엔트로피에 의한 이진영상과 연결 영역 라벨링 알고리즘을 이용함으로써 기울어지거나 회전이 가미된 영상이라도 정확하게 분할되었다.

(그림 7)은 다양한 형식의 문서영상들의 분할 및 분류결과를 보여준다. (그림 7) (a)는 원본영상을 나타내며, (그림 7) (b)는 지역적 엔트로피를 이용한 분할영상이며 각각의 구성요소별로 화소값을 구분하여 나타낸다. 그림(Graphics)영역은 0, 글자(Text) 영역은 180, 배경(Space)영역은 255, 그래프영역(Graph)과 표영역(Table)은 각각 60과 120의 회색조 색상으로 표현되며, 이는 (그림 8)과 (그림 9)에서도 동일하게 적용되었다. 또한, (그림 7) (b)의 작은 영상들은 각 분할영역에 대해서 경계 상자를 표시하여 보여준다. 분할영역은 작은 유사도를 가지는 구성요소로 분류된다. 제안방법은 언어에 상관없이 정확한 분할 및 분류가 이루어졌으며,



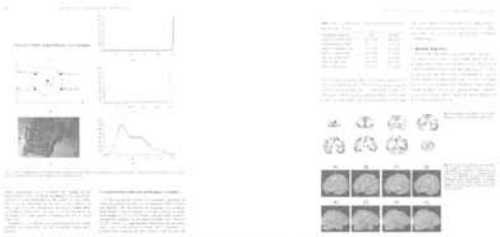
(a) 원본영상 (b) 화소 밝기값에 의한 이진화 (c) 지역적 엔트로피에 의한 이진화

(그림 5) 지역적 엔트로피와 화소의 밝기 값을 이용한 이진화

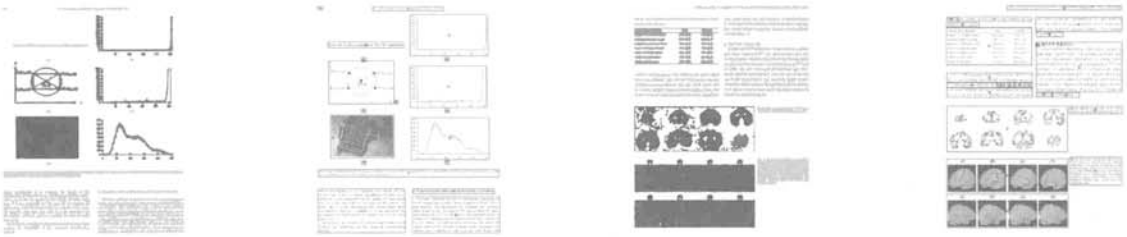


(a) 원본영상 (b) 분할영상

(그림 6) 지역적 엔트로피를 이용한 회전된 문서영상의 분할

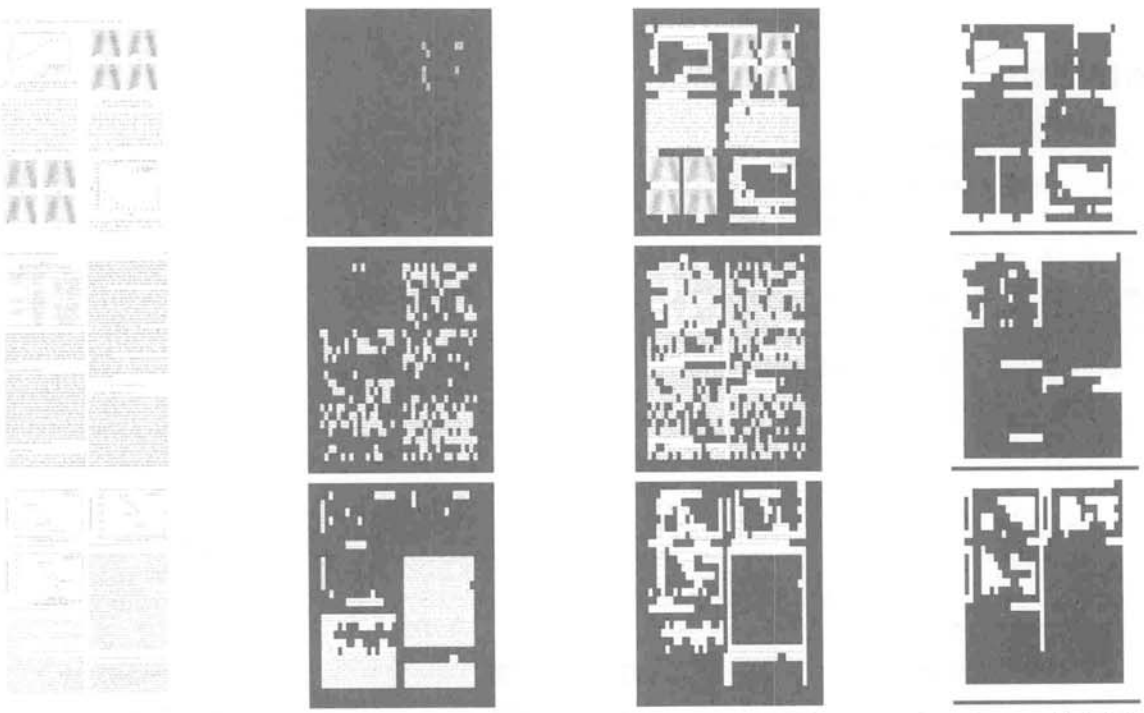


(a) 원본영상



(b) 분할영상

(그림 7) 문서영상의 분할결과



(a) 원본영상 (b) Graphics 분류영상 (c) Text 분류영상 (d) Space 분류영상

M-W Lin의 방법에 의한 결과

(e) 제안방법에 의한 결과

(그림 8) 문서영상의 분할 및 분류에 대한 성능평가비교

일반적인 문서에서 사용하는 글자의 크기에 대해서도 정확한 분류가 가능하였다. 다양한 실험영상들을 대상으로 문서영상에 존재하는 구성요소들에 대한 분할정확도는 약 96.01%를 보였으며, 대부분의 분할오류는 독립적으로 존재하는 글자영역에 의해 발생되어졌다. 그러나 (그림 7) (b)의 오른쪽 실험결과와 같이 분할된 블록의 수는 문서의 형식에 따른 구성요소들 간의 거리에 민감하게 작용하였으며, 글자 사이의 간격이 넓을 경우 분할된 블록의 수는 증가하였다. 그리고 작은 영역을 제거하기 위한 목적으로 이용한 형태처리기법에 의해서, 그림과 그림의 캡션의 여백이 좁을 경우 하나의 영역으로 합병되어져 하나의 그림으로 분류되어졌으며, 이러한 문제점을 해결하고 보다 나은 결과를 획득하기 위한 추가적인 연구가 요구된다.

(그림 8)은 제안방법과 기존방법간의 성능평가비교를 보여준다. (그림 8) (a)는 원본영상이며, (그림 8) (b)~(d)는 M-W Lin의 방법에 의해 분할 및 분류된 결과이다. M-W Lin은 사전에 정의된 블록들에 대해서 텍스처 정보를 추출한 후 클러스터링 알고리즘에 의해 문서영상을 분할하고, 각 구성요소의 특성에 기반하여 설정된 분류기준치를 이용하여 분할된 구성요소들을 그림(Graphics), 글자(Text), 그리고 배경(Space)으로 분류한다. (그림 8) (b)~(d)에서 나타나듯이 문서영상이 정확하게 분할 및 분류되어지지 않음을 알 수 있었다. 그리고 (그림 8) (e)는 (그림 8) (a)의 원본영상을 대상으로 제안방법에 의해 분할 및 분류된 결과를 보여준다. 제안한 방법은 화소 단위로 분할과정이 수행되어지기 때문에 블록 단위의 처리되는 M-W Lin 방법보다 훨씬 정교한 분할결과를 보였다. 또한, M-W Lin 방법의 경우에는 분할작업의 특징으로 이용되는 텍스처 정보가 정의된 블록의 위치 및 크기에 민감하기 때문에 제안한 방법에 비해서 더 나은 분할결과를 기대할 수 없었다. 그리고 분류실험에서도 알 수 있듯이 M-W Lin의 방법에서 이용한 분류기준치는 문서영상에 포함된 다양한 구성요소들의 특성을 적절히 반영하지 못하였다. <표 1>은 100여개의 구성요소별 분할영역들의 분류 정확도를 나타낸다. 분류 정확도는 각 구성요소별 실제화소수와 정확히 분류된 화소수의 상대 비율값이며, 모든 구성요소들에 대해서 제안한 방법이 M-W Lin 방법보다 더 나은 분류정확도를 가진다. 그리고 제안한 방법에서 배경 구성요소("Space")는 다른 구성요소에 비해 낮은 분류 정확도를 보였으며, 이는 분할단계의 닫힘 연산에서 배경 영역의 일부 화소들이 글자 및 그림 영역 등으로 분할되었기 때문이다.

(그림 9)는 잘못된 분할 결과를 보여준다. (그림 9) (a)는 원본영상이며, (그림 9) (b)는 연결 영역 라벨링 알고리즘에 의해 분할된 영상이다. (그림 9) (c)는 분할된 영역에 대해서 경계 상자를 생성한 분할영상이다. (그림 9) (b)에서는 글자 영역들은 임에도 불구하고 테이블 영역으로 잘못 분류된 사례를 볼 수 있다. 만약에 해당 영역이 두 개의 영역으로 분할된다면 글자 영역으로 분류되지만 하나의 영역으로 분할되어짐에 따라서 그 영역을 둘러싸고 있는 최외각 사각형의 흰색 공간이 상대적으로 커지고, 이에 분할영역을 테이블로 잘못 분류하는 오류를 유발하였다. 이와 마찬가지로 일반적인 문서에서 사용하는 글자의 크기보다 훨씬 큰 글자로 구성된 영역의 경우 테이블로 잘못 분류되는 문제점이 존재하였다. 그리고 닫힘 연산에 의해 그림영역을 둘러싸고 있는 테두리와 아래의 글자 영역이 하나의 영역으로 분할되었으며, 이는 글자영역의 경계 상자가 그림영역의 경계 상자를 포함하고 있어 이를 하나의 그림영역으로 잘못 분류된 사례이다. 또한, (그림 9) (c)의 왼쪽 하단에서 볼 수 있듯이 페이지 번호와 같이 독립적이며 작은 영역의 경우에는 정확히 분할되지 않은 문제점도 존재하였다.

〈표 1〉 M-W Lin의 방법과 제안한 방법의 분류 정확도

(단위: %)

구성요소	제안한 방법	M-W Lin 방법
Graphics	92.06	61.73
Text	93.23	56.55
Space	87.27	69.30

5. 결 론

본 논문은 문서영상의 분할과 분할영역들의 분류에 대한 것으로, 지역적 엔트로피 기반의 히스토그램 임계치 결정방법에 의한 문서영상의 분할과 분할영역들의 텍스처 정보를 이용한 주성분 분석 기반의 구성요소 분류방법을 제안한다. 먼저, 문서영상 분할에서 지역적 엔트로피는 다양한 밝기값

(a) 원본영상

(b) 연결 영역 라벨링 알고리즘에 의한 분할영상

(c) 경계 상자 분할영상

(그림 9) 잘못된 분할 결과

이나 잡음에 강건하며, 문서의 회전 및 기울어짐과 같은 변형에 대해서도 불변한 특성을 가진다. 그리고 히스토그램 기반의 임계치를 이용한 이진화는 빠른 수행이 가능하다. 그리고 문서영상 내에 존재하는 글자, 그림, 표, 그래프 등의 구성요소들이 각기 다른 텍스처 정보를 가진다는 사실에 착안하여 제안한 텍스처 정보 기반의 주성분 분석은 간단한 처리과정과 적은 양의 학습으로도 효과적인 분류결과를 획득할 수 있었다. 제안한 방법은 다양한 문서영상들에 대해서 글자, 그림, 표, 그래프 등의 구성요소들로 효과적으로 분할 및 분류함을 보였다. 향후에는 문서영상 분할과정에서 각기 다른 특성을 가지는 영역들이 하나의 영역으로 분할되어지는 문제점을 해결하기 위한 연구가 기대된다.

참 고 문 헌

[1] J. Toyouda, Y. Noguchi and Y. Nishimura, "Study of Extracting Japanese Newspaper Article," Proc. 6th Int'l conf. Pattern Recognition, pp.744-747, 1998.

[2] A.K Jain. B. Yu, "Document Representation and Its Application to Page Decomposition," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.29, No.3, pp.294-308, 1998.

[3] X. Hao, J.T.L. Wang and P.A. Ng, "Nested Segmentation : An Approach for layout Analysis in Document Classification." Proc. 2nd Int. conf. Document Analysis and Recognition, pp.319-322, 1993.

[4] 박동열, "Coarse/fine 전략을 이용한 문서영상의 구조 분석에 관한 연구," 전남대학교 대학원 전산통계학과 학위논문, 2001.

[5] 곽희규, "문서영상의 단어 단위 분할 및 단어 영상의 속성 추출에 관한 연구," 전남대학교 대학원 전산통계학과 학위논문, 2001.

[6] F. M. Wahi K. Y. Wong, and R. G. Casey, "Block segmentation and text extraction in mixed text/image documents," Computer Graphics and Image Processing, Vol.22, pp.375-390, Feb., 1982.

[7] P.D Wasserman, "Neural Computing," Van Nostrand Reinhold, New York, 1989.

[8] S. Imade, S. Tatsuta. and T. Wada, "Segmentation and classification for text/image documents using neural network," Proc. of the Second International Conference on Document Analysis and Recognition, Tsukuba, Japan. pp.930-934, Oct., 1993.

[9] S. B. Park, J. W. Lee, S. K. Kim, "Content-based image classification using a neural network," Pattern Recognition Letter 25, pp.287-300, 2004.

[10] Han Wang, Stan Z Li, S Ragupathi, "A fast and robust approach for document segmentation and classification," MVA'96, pp.333-336, Nov., 1996.

[11] Laura Caponetti, Ciro Castiello, Przemyslaw Gorechki, "Document page segmentation using neuro-fuzzy approach," Applied Soft Computing, Vol.8, pp.118-126, 2008.

[12] Zheru Chi, Qing Wang, Wan-Chi Siu, "Hierarchical content classification and script determination for automatic document image processing," Pattern Recognition, Vol.36, pp.2483-2500, 2003.

[13] M-W Lin, J-R Tapamo, B Ndovie, "A texture-based method for document segmentation and classification," ARIMA/SACJ, Vol.36, pp.49-56, 2006.

[14] N. Otsu, "A threshold selection method from gray level histograms," IEEE Trans. on Syst. Man Cybern. Vol.9, No.1, pp.62-66, 1979

[15] 박상철, 김수형, "투영 프로파일의 간략화 방법을 이용한 인쇄체 한글 문서 영상에서의 문자 분할", 정보처리학회 논문지B, Vol.13, No.2, pp.89-96, 2006.

[16] 김병기, "연결요소와 색상정보를 이용한 실제적 문서영상 분할", 정보처리학회 논문지, Vol.7, No.1, pp.273-285, 2000.

김 보 람

e-mail : coupstar@ynu.ac.kr

2003년 경희대학교 컴퓨터공학과 졸업(학사)

2005년 영남대학교 대학원 컴퓨터공학과 (공학석사)

2005년~현 재 영남대학교 컴퓨터공학과 박사과정

관심분야 : 영상분할, 문서영상, 의료영상

오 준 택

e-mail : ohjuntaek@ynu.ac.kr

1999년 영남대 컴퓨터공학과(공학사)

2001년 영남대 컴퓨터공학과(공학석사)

2006년 영남대 컴퓨터공학과(공학박사)

2007년~현 재 한국조폐공사 기술연구원 연구원

관심분야 : 스마트카드, 영상처리, 패턴인식

김 욱 현

e-mail : whkim@yu.ac.kr

1981년 경북대 전자공학과(공학사)

1983년 경북대 전자공학과(공학석사)

1993년 일본 쓰꾸바대학 공학연구과(공학박사)

1983년~1993년 한국전자통신연구원 선임 연구원

1994년~현재 영남대학교 전자정보공학부 교수

관심분야 : 시각정보처리, 영상처리