2006년도 대한전자공학회 하계종합학술대회 제29권 제1호

텍스처 정보 기반의 PCA를 이용한 문서 영상의 분석

김보람*, 김욱현**
영남대학교 컴퓨터공학과

Texture-based PCA for Analyzing Document Image

Bo Ram Kim*, Wook Hyun Kim**
Department of Computer Engineering, Yeungnam University
E-mail : *coupstar@yumail.ac.kr, **whkim@yumail.ac.kr

Abstract

In this paper, we propose a novel segmentation and classification method using texture features for the document image. First, we extract the local entropy and then segment the document image to separate the background and the foreground using the Otsu's method. Finally, we classify the segmented regions into each component using PCA(principle component analysis) algorithm based on the texture features that are extracted from the co-occurrence matrix for the entropy image. The entropy-based segmentation is robust to not only noise and the change of light, but also skew and rotation. Texture features are not restricted from any form of the document image and have a superior discrimination for each component. In addition, PCA algorithm used for the classifier can classify the components more robustly than neural network.

I. 서론

효과적인 정보 처리를 위해서는 문서의 영상화 및 문서 영상으로부터 정보 추출과 관련된 연구가 요구된다.[1] 본 논문에서는 문서 영상의 분할 및 분류를 위해 텍스처 정보 기반의 분석 방법을 제안하며 제안한 방법의 전체 구성도는 그림 1과 같다. 분할 단계에서는 잡음, 회전, 빛의 변화 등에 강건한 특징을 가지는 지역적 엔트로피를 추출하며, 분류 단계에서는 각 영역의 텍스처 정보를 기반으로 PCA(principle component analysis) 알고리즘에 의해 4가지 구성요소(글자, 그림, 표, 그래프) 등으로 분류한다. 이와 같은 방법을 통해, 문서의 형식과 쓰여진 언어에 영향 받지 않고, 또한 적은 양의 데이터로도 효과적인 분류를 수행할 수 있다.

II. 텍스처 정보 기반의 영상 분석

제안된 시스템은 분할과 분류의 두 단계로 구성된다. 분할 단계에서는, 정보 이론의 관점에서 각기 다른 확률을 가지는 랜덤 변수들의 불확실을 측정하기 위한 수단으로 이용되어온 엔트로피 정보를 이용한다.[2] 문서 영상의 경우 배경 부분은 동질적인 영역으로 이루어 지며 전경 부분(구성요소 부분)은 비동질적인 영역으로 이루어져 있기 때문에 지역적 엔트로피를 이용함으로써 전경과 배경을 효과적으로 분할할 수 있다. 그리고 영상의 회전이나 기울어짐에 불변하며 문서를 영상화 할 때 발생되는 각종 작은 크기의 잡음에 강건하다는 장점도 가진다. 문서 영상의 분할을 수행한 후 분할된 각 영역들이 어느 구성 요소(글자, 그림, 표, 그래프 등)에 해당하는지를 판단하는 분류 작업이 요구된다. 제안한 분류 시스템의 구성도는 그림 2와 같다. 본 논문에서는

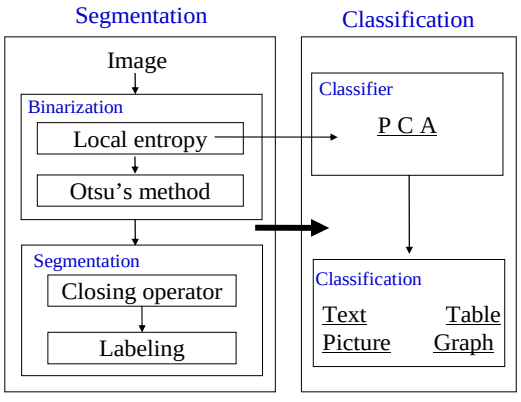


그림 1 전체 시스템 구성도

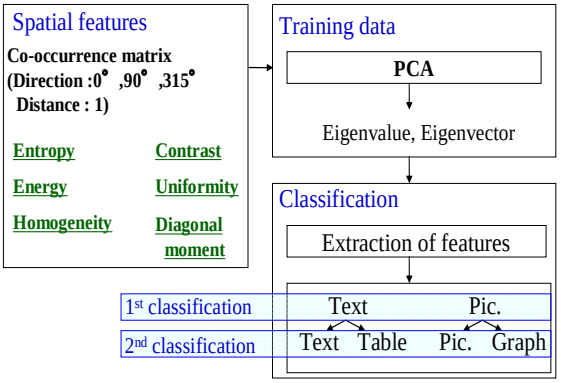
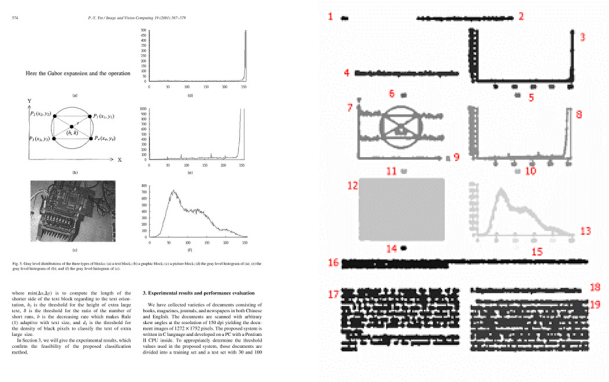


그림 2 분류 시스템 구성도

문서 영상의 구성 요소들이 가지는 밝기값의 변화 정도가 서로 다르다는 사실에 착안하여 6가지 텍스처 정보를 분류 특징으로 이용한다.[3] 이 때 텍스처 정보는 분할 단계에서 사용한 지역적 엔트로피의 변환 영상을 기반으로 추출하며, 추출된 정보를 기반으로 주성분 분석 알고리즘에 의해 구성요소를 분류한다. 주성분 분석은 다변량 분석이나 입력 데이터의 특징 추출뿐만 아니라 영상 처리 분야에서 영상 압축, 얼굴 인식 등에 널리 이용되고 있는 방법 중에 하나로서 신경 회로망과 달리 학습을 위한 표준 데이터의 양이나 특징 정보의 수가 적더라도 좋은 효과를 얻을 수 있다.[4]

Ⅲ. 실험 결과 및 결론

본 논문에서 제안한 시스템은 Pentium Ⅳ 2.8 GHz에서 Visual C++ 6.0을 이용하여 구현하였으며 실험 영상으로 624 × 512 크기의 영문 및 국문으로 된 회색조의 문서 영상을 이용하였다. 그림 3은 글자, 그림, 그래프, 표로 구성된 영문 문서를 이용하여 분할 및 분류를 수행한 결과를 보여준다. 이를 통해 문서 영상의 분할 및 각 영역의 분류 작업이 효과적으로 수행된 것을 알 수 있다. 또한 제안한 방법을 사용한 실험 결과 지역적 엔트로피에 의한 이진 영상을 기반으로 분할하기 때문에 빛의 변화나 밝기 값의 변화에 강건하다는 장점을 가지며 더불어 기울어지거나 회전이 가미된 문서 영상에 대해서도 정확한 분할결과를 얻을 수 있었다. 그리고 영어와 한글로 작성된 문서 영상들, 서로 다른 글자의 크기로 구성된 다양한 형식의 문서영상들에 대해서도 적절한 분할 및 분류 결과를 얻을 수 있었다. 이는 서로 다른 언어로 작성된 문서라 하더라도 각 구성요소들이 가지는 특성은 유사하기 때문에 나타나는 결



(a) 원본 영상 (b) 영역 분할 결과 영상

No	text	pic	graph	table	tt	pg	Result
1	0.588047	0.967305	0.961777	1.066094	0.253142	0.674947	TEXT
2	0.154816	0.730987	1.257515	1.140972	0.172517	0.569808	TEXT
3	1.442926	1.211215	0.555670	0.981488	0.571019	0.250925	GRAPH
4	0.551559	0.194301	1.040303	0.841913	0.089182	0.504741	TEXT
5	0.351676	0.889244	1.220346	1.174387	0.270365	0.682443	TEXT
6	0.283964	0.855174	1.137421	1.099280	0.251453	0.664974	TEXT
7	1.173848	0.817670	0.188745	0.315366	0.316638	0.167041	GRAPH
8	1.398550	1.150993	0.461682	0.788594	0.550781	0.184434	GRAPH
9	0.184424	0.418686	0.967303	0.891651	0.358872	0.724883	TEXT
10	0.319509	0.364662	0.946992	0.857917	0.278497	0.673940	TEXT
11	0.016057	0.654726	1.064001	1.021710	0.362495	0.699922	TEXT
12	0.837920	0.182257	0.412832	0.179180	0.686135	0.441143	PIC
13	1.343449	1.064899	0.253754	0.643967	0.466548	0.071385	GRAPH
14	0.372508	0.868091	1.126781	1.091173	0.248452	0.657543	TEXT
15	0.032907	0.662665	1.180382	1.047767	0.372171	0.721638	TEXT
16	0.648601	0.097776	1.059411	0.773487	0.009809	0.357996	TEXT
17	0.525178	0.183860	1.078133	0.980197	0.125146	0.561955	TEXT
18	0.148852	0.548381	1.190946	1.021212	0.117869	0.482812	TEXT
19	0.150454	0.728753	1.285070	1.178314	0.181768	0.586431	TEXT

(c) 분류 결과

그림 3 문서 영상의 분할 및 분류 결과

과로서 본 논문에서 사용한 분할 및 분류 알고리즘이 일반적인 다양한 종류의 문서에 대해서 강건한 특성을 가진다는 것을 의미한다.

참고문헌

[1] Y.Y Tang, C.D. Yan, and C.Y. Suen, "Document Processing for Automatic Knowledge Acquition," IEEE Trans. on Knowledge and Data Engineering, Vol. 6, No. 1, pp.3-21, Feb. 1994

[2] C. Yan, N. Sang and T. Zhang,. “Local entropy-based transition region extraction and thresholding” Pattern Recognition Letters 24(2003) 2935-2941

[3] 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

[4] Lindsay I Smith, "A tutorial on Principal Components Analysis," Feb. 26. 2003