내용기반 검색을 위한 비디오텍스트 검출

곽동엽*, 김은이**, 장재식*, 김항준*
*경북대학교 컴퓨터공학과
**가야대학교 컴퓨터공학부
e-mail : jerryone@ailab.knu.ac.kr

Videotext Detection for Content - based Retrieval

Dong-Youp Kwak*, Eun-Yi Kim**, Jae-Sig Chang*, Hang-Joon Kim*
*Dept. of Computer Engineering, Kyungpook National University
**School of Computer Engineering, Ka Ya University

요 약

본 논문은 비디오 영상에서 내용 기반 검색을 위한 비디오 텍스트를 검출하는 방법을 제안한다. 영어와 달리 한글과 같이 다중 분할된 문자가 포함된 비디오 텍스트를 자동으로 검출하기 위해 형태와 크기 및 위치 정보를 이용하고 이러한 정보들은 K-mean 클러스트링 알고리즘을 이용해 얻어진 템플릿의 형태로 표현 된다. 연결 성분 분석(connected component analysis)방법을 통해 비디오 영상을 분할하고, 잡음을 제거한 후 정확한 문자 성분을 검출하기 위해 클러스터 기반의 템플릿 매칭을 한다. 제안된 방법은 정확도와 에러율에서 기존의 방법보다 효과적 이었다.

1. 서론

최근 비디오 영상 데이터는 가장 일반적인 멀티미디어 데이터의 하나로서 자리잡고 있다. 이러한 영상 데이터 중 일반적으로 다양한 칼라와 배경을 가진 비디오 프레임에서 문자의 내용은 영상의 핵심적인 정보를 나타낸다[1,2,3]. 예를 들면, TV 뉴스에서 각 비디오 영상에서 나오는 문자들 - 비디오 텍스트 - 는 각 영상의 의미적으로 핵심이 되는 문장을 사용하여 영상의 내용을 표현 할 뿐만 아니라, 프로그램 채널의 종류와 앵커의 이름 등과 같이 내용 기반 비디오 검색에 필요한 여러 가지 요소들을 포함하고 있다. 따라서, 최근의 내용기반 검색 시스템과 최근 표준화가 제정된 MPEG-7 의 Description Scheme 을 구성하기 위해 비디오 영상에서의 문자영역 검출 및 추출에 관한 연구가 활발히 진행되고 있다[2].

스캔한 문서에서의 문자영역 검출 및 추출은 동일한 배경색과 일정한 문자 크기 및 위치 분포를 가지고 있는 반면, 비디오 텍스트는 복잡한 배경을 가지고 있으며, 문자열의 크기 또한 다양하여 문자 영역의 검출이 어려운 문제로 여겨진다. 아울러, 비디오 영상에서의 문자영역 검출과 검출된 문자 영역에 대한 각각의 문자 인식은 별개의 영역으로 간주되고 있다[1,2]. 따라서, 검출된 문자 영역에 대한 각각의 문자 인식은 상용화된 문자 인식기(OCR)를 이용하기로 하고 본 논문에서는 제외시킨다[2].

기존의 연구들을 살펴보면 크게 텍스쳐(texture) 기반 방법과 연결 성분 분석(connected-component analysis)방법으로 나눌 수 있는데, 문자 영역의 텍스쳐 성질을 이용한 방법은 Garbor 필터, 웨이블릿(wavelet), 공간적 편차(spatial variance) 등을 이용하고, 연결 성분 분석(connected component analysis)방법은 상향식(bottom-up)방법으로써 작은 영역에서 점차 큰 영역으로 합쳐 가면서 문자 영역과 비-문자 영역으로 나누는 방법이다. 이는 구현 측면에서 용이하며 문자의 크기와 문자간의 거리와 같은 사전 지식을 필요로 한다[4]. 대부분의 사전 지식은 문자들의 배열 규칙과 관련이 있다 : 비디오 텍스트의 문자들은 대개 같은 크기와 선의 두께를 가지고 있다 ; 특히, 문자들은 거의 일직선으로 나열되어 있으며, 각각의 거리가 비교적 규칙적이다. 그러나, 이러한 특성에도 불구하고 한글과 같은 다중 분할된 문자들이 포함된 비디오 프레임에서 비디오 텍스트를 검출하는 것에는 한계가 있다.

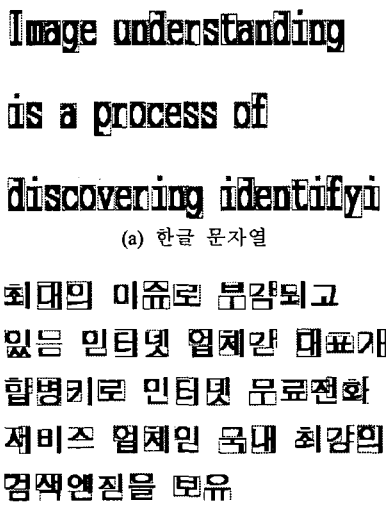
본 논문은 한글과 같이 다중 분할된 문자들이 포함된 비디오 영상에서 비디오 텍스트를 자동으로 검출하기 위한 방법을 제안하였다. 따라서, 비디오 텍스트

의 형태 정보 뿐만 아니라 크기와 위치 정보를 이용하며, 이러한 형태 정보를 표현하기 위해 K-means 클러스트링 알고리즘을 이용-생성된 클러스터 기반의 템플릿(cluster-based template)을 이용한다.

본 논문은 2 장에서 다중 분할된 문자의 특성을 살펴보고, 3 장에서 제안된 방법의 구성 및 단계를 설명한다. 4 장은 실험에 대한 결과를 비교, 마지막 5 장에서는 결론 및 향후 연구과제에 대해 서술한다.

2. 다중 분할된 문자의 특성

한글은 24 개의 단순한 음소와 26 개의 복합적 음소로 구성되어져 있다. 이들 중 10 개는 단순모음, 11 개는 복합 모음이고, 나머지는 자음이다. 한글은 2 차원 구조로 가지고 있으며, 자세한 내용은 [5]에 기술되어 있다. 아래 <그림 1>은 한글과 영어 문자열로부터 추출된 성분(component)들을 보여준다.

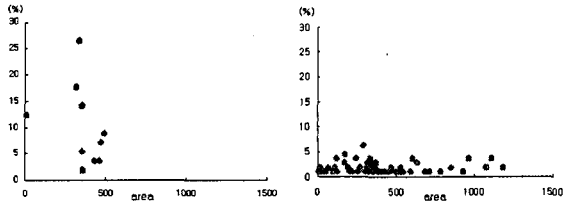


(a) 한글 문자열

(b) 영어 문자열

<그림 1> 추출된 연결 성분들

<그림 1>에서 보면 영어 문자의 경우 i 와 j 를 제외하고는 하나의 연결 성분으로 구성되어 있으나, 한글 문자의 경우 한 문자 안에 여러 개의 성분들이 존재한다. 따라서, 성분들의 크기 분포로 볼 때 한글 문자열 성분들의 크기는 영어 문자열 보다도 더 넓게 분포 되어 있다. 그림 2 는 성분들의 크기에 대한 히스토그램을 나타내는데, 영어 문자의 경우 수직 분포가 일정하지만 한글 문자의 경우 그 분포가 불규칙적이며 다양하다. 또한 각 문자 성분들의 거리가 영어와는 달리 다양하다. 결론적으로, 크기와 위치 정보만으로는 문자성분과 비-문자 성분을 구별하기 어려우며 자동으로 비디오 텍스트를 검출하기 위해서는 추가적인 정보가 필요하다.

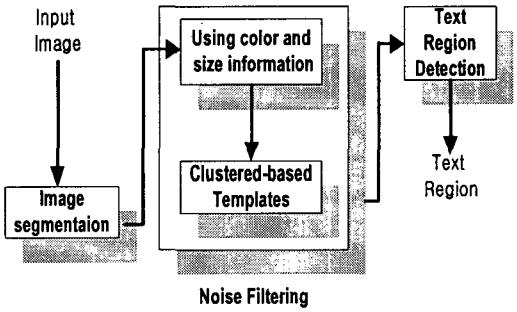


(a) 영어 문자열 (b) 한글 문자열

<그림 2> 연결 성분(connected component)들의 크기에 따른 분포

3. 제안된 방법

본 논문에서 제안된 방법은 <그림 3>과 같이 비디오 영상 분할, 잡음 제거, 템플릿 매칭, 비디오 텍스트 검증 등 4 개의 모듈로 이루어져 있다. 입력으로 들어오는 비디오 영상은 동일한 칼라를 가지는 연결 성분(connected component)로 분할 된다. 그 후 분할된 영상에서 잡음과 문자 성분이 아닌 것들을 제거한다. 그 후 클러스터 기반으로 만들어진 템플릿과 매칭함으로서 비디오 텍스트의 검증 및 위치를 확인하게 된다.

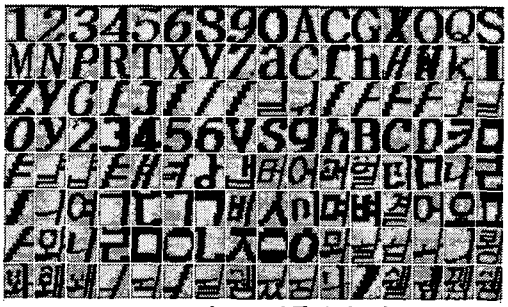


<그림 3> 제안된 방법의 구성

입력으로 들어온 비디오 영상을 연결 성분 분석(connected component analysis)방법을 통해 분할한다. 이는 대부분의 영상 분할 방법이 많은 계산량을 요구한다. 이를 위해 본 논문에서는 입력된 비디오 이미지의 칼라를 256 의 그레이 레벨로 변환 하고, [1]의 방법을 통해 양자화한다. 이후에 영상을 8-방향의 연결 성분으로 분할하면, 각 연결 성분은 동일한 값을 가지게 된다. 그러나, 연결 성분 분석으로 분할된 영상에는 많은 비-문자 성분이 포함되어 있기 때문에 이를 제거하기 위해 잡음 제거 과정을 거치게 되는데 연결 성분의 칼라와 크기 정보를 이용해 너무 크거나 아주 작은 문자 성분들은 잡음으로 간주하여 제거한다.

정확한 문자 성분을 선택하기 위해, 각 문자 성분들(components)은 클러스터 기반의 템플릿과 비교한다. 이를 위해 사용된 템플릿은 다음과 같은 과정을 거쳐 만들어 졌다. 먼저 다양한 크기와 폰트를 가진 문자가 포함된 흑백의 신문, 잡지 등의 스캔한 영상을 트레이닝 하였다. 모든 심볼들은 트레이닝 되는 문서에서 8-

방향 연결 성분(8-connected component)로 추출하고, 30×30 의 크기로 평활화 하였다. K-means 알고리즘은 트레이닝 데이터를 가장 잘 표현하는 클러스터들의 집합을 인증하는데 이용된다. 이 결과로, 총 22,108 의 문자들로부터 총 550 개의 템플릿들을 생성하였다. 각 템플릿은 1 과 0 의 시퀀스를 가지고, 아래 <그림 4> 에서 그 일부분의 예를 보인다.



<그림 4> 템플릿의 예

이전 단계에서 분할 및 잡음이 제거된 문자 성분을 템플릿에 매칭하기 위해 폰트 크기는 간과하고, 문자 성분들의 크기를 30×30 크기로 평활화 한다. 본 연구에서는 템플릿 자체가 다양한 폰트 스타일로 구성되어 있기 때문에, 실험에 쓰인 비디오 영상에서의 비디오 텍스트 중 여러 가지 형태의 문자 성분들을 검출한다. 평활화 한 후, Tanimoto 유사도에 기반하여 가장 적합한 템플릿을 찾는데 만약, 각각 가장 적합한 이진 문자열 X_C 와 X_T 의 유사도 $s(X_C, X_T)$가 임계치 θ_{sim} 보다 크면 문자 성분이다.

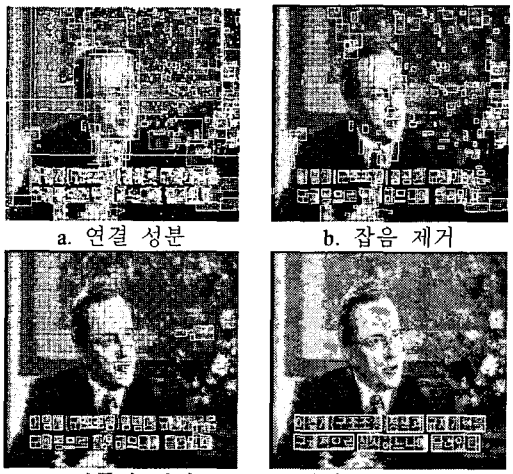
$$s(X_C, X_T) = \frac{X_C{}' X_T}{X_C{}' X_C + X_T{}' X_T - X_C{}' X_T} > \theta_{sim} \quad (1)$$

식 (1)에서 보는바와 같이, 본 연구에서는 다양한 실험을 통해 임계치를 0.64 로 두었다.

4. 실험결과

실험은 955 장의 MBC 및 KBS, SBS 의 TV 뉴스 비디오 영상을 가지고 펜티엄 II □ 233 PC 에서 실행되었다. 사용된 비디오 영상의 크기는 344 × 265 이고, 총 계산 시간은 1.95 초였다. 각 세부적인 시간 비용은 영상 분할 시간은 0.8 초, 잡음 제거 및 템플릿 매칭 1.13 초, 비디오 텍스트 검출 0.02 초 였다.

<그림 5>는 제안된 방법의 실험결과를 나타내는데, (a)는 입력된 비디오 영상을 연결 성분 분석(connected component analysis)방식으로 분할한 것이고, (b)는 분할 영상에서 잡음을 제거한 것이다. 잡음 제거 과정 후에도 남아 있는 비-문자 성분을 제거하기 위해 템플릿 매칭을 한 결과가 (c)와 같다. 마지막으로, 비디오 텍스트에 대한 검출 결과가 (d)와 같이 나타난다.



a. 연결 성분 b. 잡음 제거

c. 템플릿 매칭 d. 비디오 텍스트 검출

<그림 5> 실험 결과

실험 결과들을 평가하기 위해 각각의 입력 영상들을 수동으로 비디오 텍스트 영역에 대해 사각형의 박스를 그리고, 이후 실험된 영상과 비교한 결과가 <표 1>과 같다. 기존의 연결 성능 분석(connected component analysis)과 제안된 방법의 성능 분석 결과, 본 연구가 더 신뢰성 있는 결과를 나타내었다. 에러율은 실험에 사용된 영상에서의 총 비디오 텍스트 영역의 수와 비디오 텍스트를 검출하지 못하거나 잘못 검출한 결과를 비교한 것이다. 에러는 주로 비디오 영상의 해상도가 낮거나, 비디오 텍스트의 사이즈가 너무 작은 경우에 주로 발생 하였다.

	기존의 연결 성분 분석 방법	제안된 방법
정확도 (%)	90.05	96.4
에러 (%)	13.7	6.6

<표 1> 성능 분석 결과

5. 결론 및 향후과제

본 논문에서는 내용기반 검색을 위해 복잡한 배경을 가지는 비디오 영상에서 비디오 텍스트를 자동으로 검출하는 방법을 제안하였다. 제안된 시스템은 문자의 폰트와 크기를 알지 못한 상태에서의 복잡한 영상에서 다중 분할된 한글과 같은 문자의 영역을 자동으로 검출한다. 각 단계는 먼저, 칼라 정보를 연결 성분 분석(connected component analysis)방법을 가지고 성분(component)들을 분할하고, 사전 지식을 통해 비-문자 성분 및 잡음들을 제거한다. 잡음 속에 포함된 문자 성분을 정확히 분리하기 위해 다시 두 단계로 나뉘게 되는데, 크기와 위치 정보를 이용해 초기 잡음제거를 하고 클러스터 기반의 템플릿 매칭을 통해 최종 잡음 및 비-문자 성분을 제거한다. 마지막으로 MPEG-7 의 Description Scheme 을 구성하기 위한 비디오 텍스

트를 검출하게 된다.

실험은 수평방향의 문자열이 나열되어 있는 TV 뉴스 비디오 프레임에서 실험한 결과 96.4%의 정확도를 나타내었다. 향후 시스템의 정확도를 높이면서, 다양한 비디오 영상을 적용, 효율적인 내용 기반 검색시스템에 필요한 요소를 생성하는 것이 향후 연구 과제로서 남아있다.

참고문헌

[1] Anil K.Jain and Bin Yu. "Automatic Text Location in Images and Video Frames", *Pattern Recognition*, Vol.31, No.12, PP. 2055 ~2076, 1998.

[2] N. Dimitrova, L. Agnihotri, C.Dorai, R.Bolle, "MPEG-7 Video description scheme for superimposed text in image and video", *Signal Processing : IMAGE COMMUNICATION*, Vol.16, PP.137~155, 2000.

[3] D. Y. Kwak, E. Y. Kim, C. W. Lee, and H. J. Kim, "Text Location in Complex Image using cluster-based Templates " *Computer Applications in Industry and Engineering(CAINE 2000)*, pp. 357 ~ 359, 2000

[4] 정기철, 김항준, "텍스춰를 이용한 영상내의 문자 추출", 제 4 회 문자인식 워크샵, pp.49~58, 2000.

[5] K.Jung and H.J.Kim, "On-line Recognition of Cursive Korean Characters on Graph Representation",Vol.33, pp.399-412, 2000.