# 비디오 자막 추출 기법에 관한 연구

김성섭° 문영식
한양대학교 컴퓨터공학과
{sskim, ysmoon}@cse.hanyang.ac.kr

# Extraction of open-caption from video

Sung Sub Kim° Young Shik Moon
Dept. of Computer Science and Engineering, Hanyang University

요 약

본 논문에서는 동영상으로부터 색상, 서체, 크기와 같은 사전 지식 없이도 글자/자막을 효율적으로 추출하는 방법을 제안한다. 해상도가 낮고 복잡한 배경을 포함할 수 있는 비디오에서 글자 인식률 향상을 위해 먼저 동일한 텍스트 영역이 존재하는 프레임들을 자동적으로 추출한 후 이들의 시간적 평균영상을 만들어 향상된 영상을 얻는다. 평균영상의 외각선 영상의 투영 값을 통해 문자영역을 찾고 각 텍스트 영역에 대해 1차 배경제거 과정인 region filling을 적용하여 글자의 배경들을 제거 함으로써 글자를 추출한다. 1차 배경제거의 결과를 검증하고 추가적으로 k-means를 이용한 color clustering을 적용하여 남아있는 배경들을 효율적으로 제거 함으로써 최종 글자영상을 추출한다.

## 1. 서론

최근 멀티미디어 데이터 처리기술의 급속한 성장과 고속 통신의 발전은 멀티미디어 데이터(video, audio, image) 서비스에 대한 높은 관심을 불러 일으키고 있다. 여러 종류의 멀티미디어 데이터들 중에서도 비디오는 동영상과 함께 오디오와 텍스트 정보를 포함하고 있는 복잡한 성격의 데이터로서 그 중요성이 점차 증가하고 있으며 오락, 교육, 멀티미디어 어플리케이션 등의 넓은 분야에서 중요하게 사용되어지고 있다. 비디오 영상에 포함되어 있는 텍스트는 비디오의 내용을 함축적으로 표현하고 있기 때문에 이 텍스트를 정확하게 인식할 수 있다면 비디오 색인 및 검색에 중요하게 사용될 수 있다. 예로서 뉴스 비디오에 삽입되어있는 자막정보는 보도되고 있는 내용을 정확히 나타내며 특히 하이라이트가 되어있는 제목들은 보도 내용 전체를 대표하는 정보이다[1][2]. 뉴스 자막정보를 인식하여 색인 정보로 사용하면 사용자는 찾고자 하는 뉴스를 손쉽게 검색 할 수 있다. 본 논문에서는 동영상으로부터 글자/자막을 효율적으로 추출/인식함으로써 이를 비디오 색인과 검출에 사용 할 수 있도록 하는 기법을 기술한다. 비디오에 존재하는 문자는 해상도가 낮고 복잡한 배경을 포함하기 때문에 기존의 문자인식기(OCR)에 곧바로 사용하여 인식하기는 곤란하다[1]-[4]. 비디오로부터 문자를 추출/인식하기 위해서는 크게 세 가지의 기본적인 과정을 필요로 한다. 비디오를 연속된 이미지 프레임으로 볼 때 먼저 비디오에서 문자의 존재여부를 확인하는 문자프레임의 검출과정과 문자가 존재하는 프레임들로부터 문자영역을 획득하는 영역추출과정 그리고 추출된 문자영역으로부터 글자를 인식하는 문자인식과정으로 나눌 수 있다. 본 논문에서는 비디오 텍스트를 인식하기 위해 연속된 비디오프레임으로부터 동일한 텍스트를 포함하고 있는 프레임들을 자동 검출하고 검출된 텍스트프레임으로부터 region filling과 color clustering을 사용하여 문자 영역을 추출한다.

## 2. 텍스트프레임 검출

동영상에서의 문자는 다양한 색상, 서체, 크기 등을 갖기 때문에 문자영역의 유무를 일반화 하기는 쉽지않다. 하지만 동영상에서의 문자는 정지 영상에 비할 때 여러 프레임에 걸쳐 나오기 때문에 이러한 특성은 문자 프레임 검출에 유용하게 사용된다. 본 논문에서는 텍스트프레임을 검출하기 위해 먼저 각 프레임으로부터 후보 문자영역을 추출하고 서로 인접한 프레임들을 비교하여 두 프레임에 존재하는 후보 문자영역들이 일정치 이상 유사 할 때 이들을 텍스트프레임으로 검출한다.

### 2.1 후보 문자영역을 추출

본 논문에서는 후보 문자영역을 추출하기 위하여 각 프레임을 휘도 영상으로 변환하고 sobel 연산자를 사용하여 에지 영상을 만든다. 에지 영상로부터 수평 projection을 시켜 분포가 조밀한 부분의 시작과 끝을 찾아 각 문자열들의 높이를 구하고 각 문자 열에 대하여 수직으로 projection시켜 같은 방법으로 문자열의 폭을 찾아 후보 문자영역을 추출한다.

### 2.2 인접한 프레임의 후보 문자영역의 비교

후보 문자영역이 결정되면 현재 프레임과 이전 프레임간의 문자영역의 수, 위치, 크기, 분포 등을 비교하여 일정치 이상 유사하면 동일한 텍스트 프레임으로 간주한다.

## 3. 문자영역추출

본 논문에서는 먼저 동일한 텍스트가 나타나는 프레임들의 시간적 평균을 통해 영상의 화질을 향상하고 영상에 존재하는

문자의 영역들을 2단계의 배경제거 과정을 거쳐 문자영역 추출을 수행한다. 첫번째 과정은 문자영역의 외각선 상에 놓여 있는 pixel들의 color값을 seed로 한 region filling을 수행하여 1차 배경제거를 한다. 배경이 어느 정도 제거된 글자 영상으로부터 각 글자 영역의 분산 값을 구하고 이를 토대로 1차 배경제거의 결과를 검증하여 추가적인 color clustering의 적용 여부를 결정한다. 두번째 과정은 앞의 결과에 따라 color clustering을 적용한 추가적 배경 제거 과정이다. 마지막으로 크기가 작은 잡음 등을 제거하여 문자영역추출을 완료한다.

### 3.1 문자영상 향상

비디오에서 동일한 텍스트는 여러 프레임에 걸쳐 나타난다. 많은 잡음을 포함하거나 복잡한 배경이 있는 비디오 영상에서는 하나의 프레임으로부터 텍스트를 추출하는 것보다는 동일한 텍스트를 갖는 모든 프레임을 사용한다면 보다 좋은 텍스트영역추출 결과를 얻을 수 있다[3]. 연속된 비디오 프레임에서 화면의 급격한 전환이 이루어지는 부분을 컷(cut)이라고 한다. 유사성이 있는 연속된 프레임들의 집합을 샷(shot)이라고 하며 화면의 급격한 전환이 되는 컷은 새로운 샷의 시작 또는 끝을 이룬다. 마찬가지로 비디오텍스트의 변환이 일어나는 부분의 시작과 끝을 찾으면 유사성이 있는 연속된 텍스트 프레임들의 집합을 텍스트 샷(text shot)을 알 수 있으며, 텍스트 샷 사이에 있는 모든 프레임들의 정보를 텍스트영역 추출 과정에 사용 할 수 있게 된다. 2.2에서 설명한 방법을 사용하여 동일한 텍스트의 처음과 마지막 프레임을 찾는다. 비디오에서 배경은 대부분 움직이지만 동일한 텍스트는 여러 프레임에 걸쳐 변화가 없다는 특징을 이용하여 텍스트 샷에 존재하는 모든 프레임의 시간적 평균프레임을 구한다. 시간적 평균프레임에서 배경부분은 대부분 변화하기 때문에 배경의 움직임이 많을수록 컬러에 변화가 많이 일어나는 반면에 텍스트영역의 컬러는 적은 변화만 일어나게 된다. 시간적 평균프레임을 앞에서 설명한 문자 후보영역 추출 방법을 사용하여 프레임에 존재하는 문자 영역들을 찾는다. 그림 1 은 MPEG 비디오에서 하나의 텍스트 샷에 존재하는 모든 I 프레임들의 시간적 평균프레임을 만들어 영상의 질을 향상시키고 프레임에 나타나는 문자영역을 찾은 결과이다.

그림 1. 시간적 평균프레임과 찾아진 문자영역

### 3.2 Region filling을 이용한 1차 배경제거

찾아진 문자영역의 외각선(boundary)상에 놓여있는 pixel들의 컬러값을 seed로 하여 region filling을 수행 함으로써 boundary와 유사한 색상을 갖는 부분들을 제거한다. 식(1)은 region filling을 위한 두 컬러의 거리를 결정하는 식이며 R1,G1,B1은 seed의 컬러값 R2,G2,B2는 문자영역내의 임의의 화소의 컬러 값이다. 그림 2는 1차 배경제거를 수행한 과이다.

$$dist = (R1-R2)^2 + (G1-G2)^2 + (B1-B2)^2 \quad (1)$$

그림 2. Region filling을 사용하여 배경을 제거한 결과

#### 3.1.2 1차 배경제거의 결과 검증

1차 배경제거단계를 거쳐 어느 정도 분리된 각각의 글자영역의 분산값를 구하여 1차 배경제거의 결과를 검증한다. 먼저 한 글자 영역의 전체 분산값을 구한다. 만약 1차 단계에서 글자의 분리가 잘 되었다면 동일한 글자영역에서의 분산값은 작은 값을 갖지만 분리가 잘 되지 않았을 경우에는 큰 분산값을 갖게 된다. 1차 배경 제거 단계인 region filling 만으로도 대략적인 배경 제거를 할 수 있지만 글자 주위의 제거되지 않은 배경들이 남아 있거나' ㅁ,ㅇ,ㅂ' 등과 같은 글자에서 나타나는 고립된 영역은 제거 되지 않는 결과가 발생한다. 따라서 이들을 제거하는 추가적인 과정을 필요로 한다. 본 논문에서는 k-means color clustering을 통해 글자 영역을 두개의 cluster로 나눔으로써 글자와 배경을 최종 분리한다. 하지만 1차 배경제거 단계에서 이미 배경제거가 잘되어진 글자들에 대해 color clustering을 적용하면 오히려 글씨의 획이 사라지는 등의 좋지 않은 결과를 초래하기 때문에 color clustering에 앞서 1차 배경제거 결과에 따라 2차 배경제거 과정 적용의 필요성을 검증하여야 한다. 그림 3의 (a)는 region filling을 수행한 후 모든 글자 영역에 대해 color clustering을 했을 때의 결과이다. 그림(a)에서 동그라미가 쳐져 있는 부분의 획이 심하게 상해 있는 것을 볼 수 있다. (b)에서는 각 글자 영역별로 1차 배경제거 과정의 결과를 검증한 후 추가적인 과정을 필요로 하는 글자 영역에만 color clustering을 수행한 결과이다

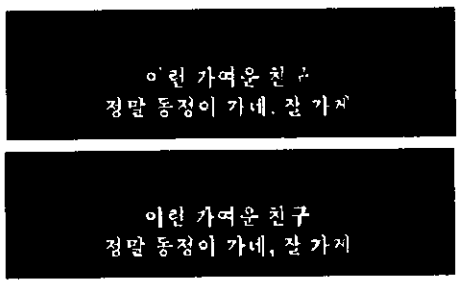


그림 3. (a)모든 글자에 대해 2차 배경제거 단계를 적용한 결과 (b) 2차 배경제거 단계를 필요로 하는 글자에만 적용한 결과

### 3.2 Color clustering을 이용한 2차 배경의 제거

1차 배경제거 과정에서 추출된 문자 영역이 높은 분산값을 갖을 경우 color clustering을 통해 글자영역과 남아있는 배경영역을 분리한다. Clustering의 입력벡터는 각 pixel의 컬러값(RGB)이고 글자와 배경의 2개의 cluster를 갖는 k-means algorithm을 사용하였다. 좋은 clustering을 위해 글자영역의 color histogram(8x8x8)에서 나타나는 두개의 local max color를 찾아 각 cluster의 center로 하였다. Clustering을 마친 후 두개의 cluster중 많은 수의 요소(element)를 갖으며 상대적으로 밝은 밝기값을 갖는 cluster를 선택하여 글씨영역으로 하고 적은 수의 요소(element)와 상대적으로 어두운 부분을 배경으로 선택한다. 그림 4는 2차 배경제거 단계를 수행한 결과이다.

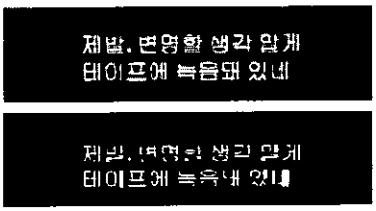


그림 4. 2차 배경제거 단계를 수행한 결과

### 4 실험 결과 및 분석

실험 환경은 Pentium III 800MHz PC에서 Visual C++ 6.0을 이용하였고 약 1~5 분량의 352 x 240 크기의 뉴스, 영화, 애니메이션 등을 대상으로 제안한 문자영역 분리 방법을 평가하였다. 본 실험에서 찾아진 텍스트 샷의 검출율(DR: Detection Rate)는 아래의 식과 같이 수작업으로 검출한 총 텍스트 샷(동일한 텍스트가 나타나는 프레임을 모두 하나로 묶어 카운트했음)과 본 논문에서 제안한 방법을 사용하여 검출한 프레임 샷의 비율로 구하였다.

$$DR = \frac{\text{Automatically detected text shot number - false detection}}{\text{Manually detected text shot number}}$$

실험 비디오에 대한 텍스트 샷의 검출율은 표 1 과 같다.

표 1. 텍스트 샷(shot)의 검출율

| | 뉴스 | 영화 I | 영화 II | 애니메이션 |
|---|---|---|---|---|
| Total Number of caption appearances | 21 | 59 | 69 | 92 |
| Correct detection | 21 | 59 | 67 | 91 |
| Miss detection | 0 | 0 | 2 | 1 |
| False detection | 2 | 6 | 5 | 14 |
| Detection Rate (%) | 90.4% | 89.8% | 89.8% | 83.7% |

그림 5는 다른 비디오에서의 문자영역 추출 결과를 보여 주고 있다.

### 5 결 론

본 논문에서는 글자의 색상, 크기, 서체 등의 사전 지식 없이도 비디오로부터 문자영역을 추출하는 방법을 제안하였다. 기존의 동일한 자막 프레임을 판별하는 방법을 보완하여 시작 프레임과 끝 프레임을 찾았고 이들 사이에 존재하는 모든 프레임을 이용하여 문자영역 분할에 사용하였다. 2단계에 걸친 배경제거 과정을 통해 뛰어난 문자 영역 이진화를 수행 할 수 있었다. 향후 연구 과제로는 본 논문에서 제안한 방법을 통해 이진화 된 자막 영상의 인식 실험을 통해 제안한 방법의 성능을 정량적으로 평가하는 것이다.

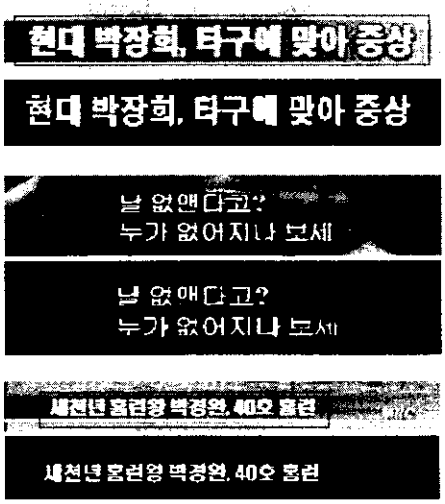


그림 5. 문자영역 추출의 결과

### 참고 문헌

[1] 전병태, 배영래, 김태윤, "일반화된 문자 및 비디오 자막 영역 추출 방법," 정보과학회 논문지 : 소프트웨어 및 응용 제27권 제6호, pp.632-641, 2000

[2] 곽신상, 김소명, 최영우, 정규식, "효율적인 비디오 자막 인식을 위한 영상 향상 방법," 제 12회 영상처리 및 이해에 관한 워크샵 발표 논문집, pp. 342-436, 2000

[3] 김소명, 최영우, 정규식, "비디오 자막 추출 및 이미지 향상에 관한 연구," 제 27회 정보과학회 가을 학술발표논문집, vol. 3

[4] 최경주, 변혜란, 이일병, "이진화를 위한 영상 강화 기법에 관한 연구," 제 10회 영상처리 및 이해에 관한 워크샵 발표 논문집, pp. 176-181, 1998