비디오 자막 추출 및 인식 기법에 관한 연구

김 종 렬, 김 성 섭, 문 영 식
한양대학교 컴퓨터공학과
전화 : 031-407-8991 / 핸드폰 : 017-270-5610

Study on video character extraction and recognition

Jong Ryul Kim, Sung Sub Kim, Young Shik Moon
Dept. of Computer Science and Engineering, Hanyang University
E-mail : jrkim@cse.hanyang.ac.kr

Abstract

In this paper, a new algorithm for extracting and recognizing characters from video, without pre-knowledge such as font, color, size of character, is proposed. To improve the recognition rate for videos with complex background at low resolution, continuous frames with identical text region are automatically detected to compose an average frame. Using boundary pixels of a text region as seeds, we apply region filling to remove background from the character. Then color clustering is applied to remove remaining backgrounds according to the verification of region filling process. Features such as white run and zero-one transition from the center, are extracted from unknown characters. These feature are compared with a pre-composed character feature set to recognize the characters.

I. 서론

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 내용전체를 대표하는 정보이다[1][2]. 뉴스 자막 정보를 인식하여 색인 정보로 사용하면 사용자는 찾고자 하는 뉴스를 손쉽게 검색할 수 있다. 본 논문에서는 동영상으로부터 글자/자막을 효율적으로 추출/인식함으로써 이를 비디오 색인과 검출에 사용 할 수 있도록 하는 기법을 기술한다. 비디오 텍스트를 인식하기 위해 본 논문에서는 연속된 비디오 프레임으로부터 동일한 텍스트를 포함하고 있는 프레임들을 자동 검출하고 검출된 텍스트 프레임으로부터 region filling과 color clustering을 사용하여 배경을 제거한다. 추출된 문자영역으로부터 각 글자들의 영역을 얻어와 이들의 특징값을 구하고 이를 비교함으로써 인식을 수행한다.

※ 본 연구는 한국과학재단 목적기초연구(과제번호 : 2000-2-303-005-3)지원으로 수행되었음.

II. 텍스트 프레임 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 텍스트 프레임을 검출하기 위해 먼저 각 프레임으로부터 후보 문자영역을 추출하고 서로 인접한 프레임들을 비교하여 두 프레임에 존재하는 후보 문자영역들이 일정치 이상 유사할 때 이들을 텍스트 프레임으로 검출한다.

2.1 후보 문자영역의 추출

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 문자영역 비교

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

III. 문자영역 추출

본 논문에서는 문자영역 추출을 위해 먼저 동일한 텍스트가 나타나는 프레임들의 시간적 평균을 통해 영상의 화질을 향상하고 영상에 존재하는 문자의 영역들을 2단계의 배경제거 과정을 거쳐 문자영역 추출을 수행한다. 첫 번째 과정은 문자영역의 외각선 상에 놓여있는 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. 두 번째 과정은 앞의 결과에 따라 color clustering을 적용한 추가적 배경 제거 과정이다. 마지막으로 크기가 작은 잡음 등을 제거하여 문자영역 추출을 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 텍스트 영역 추출 결과를 얻을 수 있다[3]. 비디오 텍스트의 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 1은 MPEG 비디오에서 하나의 텍스트 샷에 존재하는 모든 I 프레임들의 시간적 평균 프레임을 만들어 영상의 질을 향상시키고 프레임의 평균 영상에 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 수행함으로써 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.

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

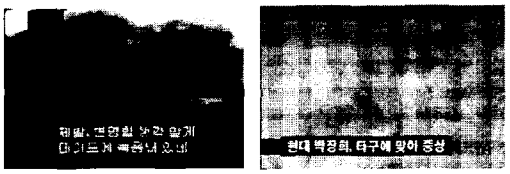


그림 2. Region filling을 사용하여 배경을 제거한 결과

3.3 1차 배경제거의 결과 검증

1차 배경제거 단계를 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 제거되지 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 잘 되어진 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. 그림3.(a)에서 동그라미가 쳐져 있는 부분의 획이 심하게 상해있는 것을 볼 수 있다. 3.(b)에서는 각 글자 영역별로 1차 배경제거 과정의 결과를 검증한 후 추가적이 과정을 필요로 하는 글자 영역에만 color clustering을 수행한 결과이다.

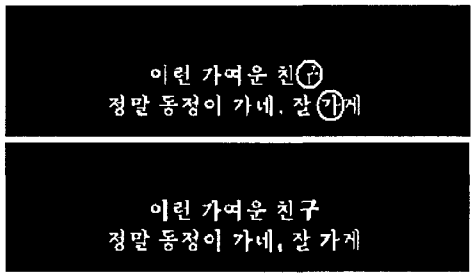


그림 3. (a)모든 글자에 대해 2차 배경제거 단계를 적용한 결과 (b) 2차 배경제거 단계를 필요로 하는 글자에만 적용한 결과

3.4 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 사용한다. 좋은 clustering을 위해 글자영역의 color histogram(8x8x8)에서 나타나는 두개의 local max color를 찾아 각 cluster의 center로 한다. 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.

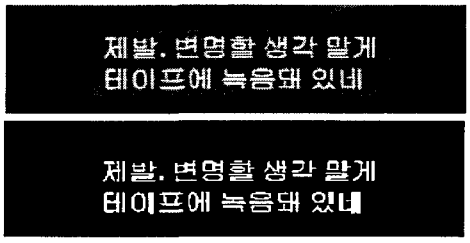


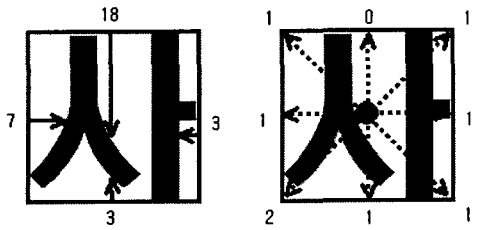
그림 4. 2차 배경제거 단계를 수행한 결과

IV. 문자인식

비디오는 다양한 서체의 문자들을 포함하고 있고 또 비디오 문자영역 분할 시 서체의 형태의 변화가 일어나기 때문에 일반적인 문자인식 기법으로 이를 인식하기는 매우 힘들다. 본 논문에서는 다양한 서체의 문자를 인식하기 위해 여러 서체의 문자들로부터 특징값을 추출하고 인식하고자 하는 문자의 특징값과 비교를 통해 인식을 수행한다.

4.1 특징값 추출

투영을 이용하여 배경이 모두 제거된 문자영상으로부터 각 글자의 영역을 얻어온다. 비디오에는 다양한 크기의 문자가 나타나게 때문에 인식 과정에 앞서 일정한 크기(30×30)로 정규화 시킨다. 인식을 위한 특징값은 각 글자들을 특성을 독자적으로 표현하여야 하며 많은 수의 문자와 서체를 표현해야 하기 때문에 가능한 적은 수의 특징값을 갖는 것이 바람직하다. 본 논문에서는 글자의 상, 하, 좌, 우로부터 외각 정보를 나타내는 white run(그림 5.(a))과 글자영역 중심으로부터 외각 쪽으로 획의 분포를 나타내는 zero-one transition(그림 5.(b))을 특징값으로 사용한다.



(a) (b)

그림 5. (a)글자의 외각정보를 나타내는 특징값 white run (b) 글자 획의 분포를 나타내는 특징값 zero-one transition

4.2 특징값의 비교

앞에서 설명한 특징값들을 사용하여 모든 문자에 대한 특징값의 집합을 만든다. 문자의 인식은 모든 문자의 특징값의 집합과 인식하고자 하는 문자와의 비교 통해 가장 유사한 특징값을 갖는 글자를 찾아냄으로써 인식을 수행한다. 비디오에서는 다양한 서체의 문자들이 나타나기 때문에 어느 특정 서체의 문자들의 특징값만을 사용하여 인식을 수행하기는 힘들다. 따라서 가능한 많은 수의 서체를 사용하여 특징값의 집합을 구성하면 높은 인식률을 기대할 수 있다. 하지만 너무 많은 수의 서체를 사용하게 되면 특징값 비교를 위해 많은 시간이 소요됨으로 적절한 서체와 글자의 수를 정하는 것이 중요하다. 식(2)는 인식하고자 하는 글자와 특징값의 집합들과의 비교를 위한 식이다. α는 특징값 집합에 있는 서체의 수이고 β는 각 서체별 글자의 수이다. N은 각 문자를 인식하기 위해 사용된 특징값의 수이고 ω_i는 각 특징값에 대한 가중치이다. $d_{\alpha\beta i}$와 u_i는 각각 특징값 집합에 있는 α 서체의 β 글의 i번째 특징값과 인식하고자 하는 문자의 i번째 특징값이다.

$$best\ \ match = \min_{\alpha}(\min_{\beta}(\sum_{i=0}^{N}\omega_i(d_{\alpha\beta\ i}-u_i)^2)) \quad \text{식(2)}$$

V. 실험 결과 및 분석

실험을 위해 약 1~5 분의 352×240 크기의 뉴스, 영화, 애니메이션 등의 MPEG-1 대상으로 제안한 문자영역 분리와 인식 기법을 평가하였다. 문자인식을 위해 한글97로부터 견고딕, 굴림, 둥근고딕, 휴먼고딕, 견명조의 5개 서체당 588개의 글자들 추출하여 문자집합을 만들었다. 이는 588자 만으로도 일반적으로 비디오에 나타나는 자막의 98퍼센트 이상을 포함하기 때문이다[5]. 표 1은 제안된 방법으로 문자영역을 분리하고 인식을 시도했을 때의 인식률을 보여주고 있다. 전체적으로 약 58.3%의 인식률을 보이고 있다. 하지만 실험에서 사용된 비디오는 우리가 일상에서 일반적을 사용하는 MPEG-1 압축의 저화질 영상이었기 때문에 잡음이 많이 포함 되어있었고 글자의 크기도 10~25pixel의 작은 글자들이 대부분이었다. 좀더 좋은 화질의 비디오를 사용하였을 경우 보다 높은 인식률을 기대할 수 있을 것이다.

표 1 동영상 자막 인식률 (애니 : 애니메이션)

	영화1	영화2	애니1	애니2	뉴스
인식률(%)	61.8	62.7	53.3	55.2	68.3
전체	58.3 %				

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 시작프레임과 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 1차 배경 제거 단계의 검증시 사용되는 분산값을 자동으로 구하는 것도 문자인식기의 성능을 향상시켜 인식률을 올리는 것이다.

참고문헌

[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

[5] 나지훈, "뉴스 영상에서의 자막영역 추출 및 문자 인식," 석사학위논문, 1999