2001年度 第14回 信號處理合同學術大會 論文集 第14卷 1號

압축 MPEG 비디오 상에서의 자막 검출 및 추출

전승수, 김정림, 오상욱, 설상훈

고려대학교 전자공학과

Video Caption Extraction in MPEG compressed video

Seong Soo Chun, Jung-Rim Kim, SangWook Oh, Sanghoon Sull

School of Electrical Engineering, Korea Univ

{sschun,jrkim,sull}@mpeg.korea.ac.kr

요 약

본 논문은 DCT를 기반으로 하여 비디오 내에서 자막을 I-frame들로부터 추출하였다. 본 논문에서 제안하는 자막 검출 및 추출 방법은 자막이 주위 배경 화면과 그 대비 값이 크다는 점과 화면상에 일정한 시간동안 유지된다는 점을 이용하였다. 먼저 비디오 내에서 I-frame들의 DCT 값들로부터 주위 배경화면과 비교하여 그 대비 값이 큰 영역들을 표시하였다. 이로부터 자막의 시간적 특성과 공간적 특성을 이용하여 자막을 포함하는 프레임을 검출하여, 그 내에 있는 자막 영역을 추출하였다.

I. 서 론

자막은 비디오에 자주 등장한다. 자막이 비디오 내에 있을 경우 비디오 내용을 함축적으로 표현하고 있기 때문에 비디오의 색인 및 검색에 있어서 상당히 중요하다. 그렇기 때문에 비디오 프레임에서 자막을 추출하기 위한 많은 연구가 과거로부터 많이 연구 되어왔다. Zhong[1] 은 수평적인 공간적 편차(spatial variance)를 이용함으로써 이미지 내에서 자막 영역을 추출하는 방법을 이용하였다. Manmatha[2]는 이미지 내의 질감을 분석함으로써 자막의 고유한 질감을 갖는 영역을 자막영역으로 추출하였다. 그들은 피라미드 기법과 텍스트의 특성을 이용함으로써 다양한 크기와 모양을 가지는 자막을 추출했다. Miller[3]는 반면에 색상 클러스터링과 Connected Component를 이용함으로써 자막을 추출하였다. 또한 Linehart[4]는 분할(segmentation)을 수행하기 위해서 진보된 분할 및 병합(split and merge) 방법을 통해 자막을 비디오로부터 추출하였다. 하지만 멀티미디어에 대한 현대인들의 수요가 급증함으로써 그에 따른 멀티미디어 데이터의 용량 또한 상당히 방대해졌기 때문에 대다수의 비디오들은 그 용량을 줄이기 위해 압축된 형태로 저장되어 있다. 그렇기 때문에 자막을 압축된 상태에서 바로 추출할 수 있는 빠르고 효율적인 알고리즘이 필요하게 되었다. 최근 이러한 압축 상태에서 자막을 추출하기 위한 다양한 연구가 진행되어 왔다. Zhong[5]는 DCT 블록의 수평적 특성을 띄는 AC계수를 이용함으로써 자막을 추출하였다. 그 후에 수직적 특성을 띄는 AC계수를 이용함으로써 프레임 내에서 자막이 존재하는 지를 결정하였다.

본 논문은 압축 비디오에서 DCT 계수를 활용하여 효과적인 자막 추출 방법을 보인다. 이 연구의 목적은 자막이 있는 프레임 내에서 자막을 찾고 OCR 입력을 위한 각각의 글자들을 추출하기 위해 실시간으로 자막이 있는 프레임을 찾아내는 것이다. 자막의 밝기, 이웃 화소들간 차이, 자막의 형태 및 시간 정보들을 이용하여 축소된 영상 내에서 자막이 있는 프레임을 검출 및 추출하는 DCT 압축 기반 알고리즘을 제안하고자 한다.

II. 제안 알고리즘

MPEG 압축기술은 공간적 및 시간적 과잉 정보들을 줄임으로써 비디오를 압축하는 기술이다. MPEG 압축 비디오는 I-프레임(intracoded)과 그 사이에 B-와 P-프레임으

로 이루어져 있다. I-프레임은 이산 여현 변환(DCT)을 통해 8x8 블록 단위로 압축된다.

본 논문에서 제안하고자 하는 알고리즘은 MPEG 비디오의 I-프레임의 DCT 계수 값들을 이용한다. 제안하고자 하는 자막 추출 알고리즘은 기본적으로 세 가지의 단계를 거친다 : 우선 자막 프레임을 식별하게 되고 그 후에 자막 프레임으로부터 DCT 압축 상태에서 자막 후보 영역을 검출하고 후보 자막 영역들 중에서 자막에 부합하는 영역들만 고르게 된다.

2.1 특징 선정(Feature Selection)

어느 특정한 객체를 추출하기 위한 가장 중요한 점은 그 객체를 표현하는 특징을 얼마나 잘 선택하였는가 이다. 이러한 객체를 표현하는 특징을 선택함에 있어서, 그 특징치를 쉽게 얻을 수 있어야 하며, 그 차원 또한 작아야 한다. 자막은 다양한 색상과 명도를 가지고 있기 때문에, 색상 정보만을 이용하여 자막의 특징을 얻어 오는 것은 많은 한계점을 가지고 있는 것이 현실이다. 하지만 자막이 비디오 시청자들로부터 인식이 되기 위해서는, 자막이 그 자막의 배경 화소들과 뚜렷한 대비 차이를 가지고 있어야 한다. 그러므로 대비(constrast)가 텍스트를 추출함에 있어서 가장 중요한 특성이라고 할 수 있다.

텍스트 자막과 배경 화소들과의 고대비(high constrast)는 8x8 DCT 내의 특정한 블록에서 아주 큰 계수 값을 갖는다. 자막이 위치하는 8x8 DCT 블록들에 많은 시행을 거쳐 DCT 내의 블록들 중에서 그림 1에서 표시된 특정한 블록들이 큰 계수 값을 가짐을 실험적으로 알 수 있었다.

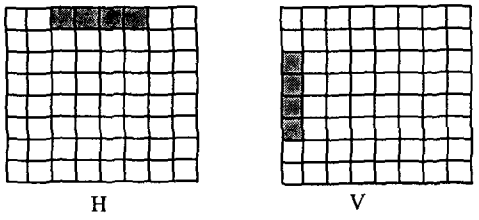


그림 1. 자막 추출을 위해 선택된 DCT 계수

곧 DCT 블록내의 계수 값들은 자막 추출을 위한 적절한 특징이다.

2.2 고대비 값을 가지는 DCT 블록 추출

고대비 값을 가지는 DCT 블록을 추출하기 위해서 다음과 같이 DCT블록 (m,n)의 에너지 $E_{m,n}$ 을 다음과 같이 정의하였다.

$$E_{m,n} = E^H_{m,n} + E^V_{m,n}$$

$$\text{where}$$

$$E^H_{m,n} = \sum_{m,n} \{DCT(i)\}^2, \quad i \in \{3, 8, 10, 20\} \tag{1}$$

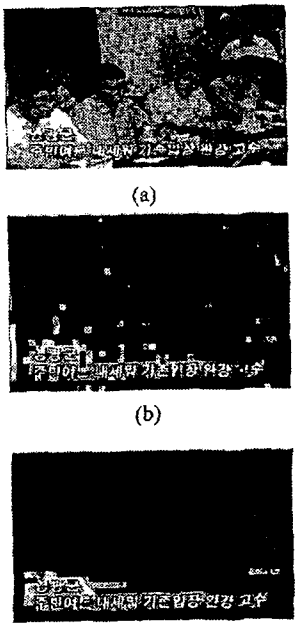
$$E^H_{m,n} = \sum_{m,n} \{DCT(i)\}^2, \quad i \in \{5, 6, 14, 15\}$$

여기서 $DCT(i)$란 지그재그 스캔에 입각한 i번째 AC 계수 값을 의미한다.

임의의 I-프레임 F에 대해서 각각의 DCT블록들의 에너지를 계산 한 후에 고대비 영역을 가지는 DCT 블록들을 추출하기 위해서 다음 정의와 같이 임계 값을 정하였다.

$$B^F_{m,n} = \begin{pmatrix} 1 & if\ E_{m,n} > \tau \\ 0 & otherwise \end{pmatrix} \tag{2}$$

그림 2 는 자막을 포함하는 프레임(a)의 고대비 DCT 블록들(b) 과 1x3의 구조적 요소를 통해 closing 연산을 한 후에 opening 연산을 한 후의 결과이다(c). 그림 2(c)에서 볼 수 있듯이 형상학적 연산을 통하여 대부분의 noise를 제거하였다. 이는 자막이 아닌 DCT 블록들이 자막인 DCT 블록처럼 형상학적 연산을 통해 세로축으로 병합하지 않는데 기인한다.



(a)

(b)

(c)

그림 2. 비디오 프레임에 대한 고대비 블록

2.3 자막 프레임 검출

어느 특정한 비디오 프레임에 대해서 자막을 추출하기 전에 우선 자막이 프레임 내에 존재 하는지를 다음과 같은 간단한 연산 과정을 통해 검출하게 된다.

대부분의 자막이 직사각형의 형태를 가지고 여러 프레임에 걸쳐 유지된다는 전제 하에, 여러 프레임에 걸쳐 유지되는 고대비 블록($B^F_{m,n}=1$) 이 수직방향으로 얼마나 잘 연결되어 있는지를 측정한 *수직 연결성*을 정의 하게 된다. 다음 식은 임의의 프레임 F에 대해서 평준화된 고대비 블록들의 수직 연결성을 측정하게 된다.

$$R^F_i = \sum_{j=1}^{N-1}\left[\frac{(B^F_{i,j} * B^{F-1}_{i,j}) * (B^F_{i,j-1} * B^{F-1}_{i,j-1})}{N-1}\right]^2 \qquad (3)$$

그 후에, 수직 연결성 R^F_i 를 통하여 수평으로 이들이 얼마나 잘 연결되어 있는지를 특정하게 된다. 이를 측정하기 프레임 I 에 대하여 다름과 같은 식을 유도할 수 있다.

$$T_F = \frac{\sum_{j=1}^{M-1} R^F_j * L^F_{j-1}}{(M-1) * N_t - 1} \qquad (4)$$

위의 식에서 N_t 는 프레임 F의 고대비 블록들 중에서 이전 프레임 F-1에서도 고대비 값을 가지는 블록들의 개수와 프레임 F에 존재하는 총 블록들의 비율이다. 식 (4)는 다음과 같이 표현될 수 있다.

$$T_F = \frac{\sum_{i=1}^{M-1}\left\{\sum_{j=1}^{N-1}\left[(B^F_{i,j} * B^{F-1}_{i,j}) * (B^F_{i,j-1} * B^{F-1}_{i,j-1})\right]^2 * \sum_{j=1}^{N-1}\left[(B^F_{i-1,j} * B^{F-1}_{i-1,j}) * (B^F_{i-1,j-1} * B^{F-1}_{i-1,j-1})\right]^2\right\}}{[(M-1) * N_t - 1]^*(N-1)^4} \qquad (5)$$

식 (5)은 어느 임의의 프레임 F에 자막이 있을 확률을 측정하게 된다. 식 (5)은 매우 빠르게 계산될 수 있는데, 이는 분모가 미리 계산될 수 있는 상수이고,분자는 오직 0과 1 의 값을 가지는 $B^F_{m,n}$ 들의 곱과 합으로 이루어져 있기 때문이다.

2.4 압축 상태에서의 자막 영역 추출

압축 상태에서 자막 영역을 추출하기 위해서는 다음과 같이 I-프레임에서 8x8 블록 단위로 영역을 추출하게 된다.

2.4.1 자막 영역 분할(Segmentation of Text Region)

고대비 블록들로부터 자막을 추출하기 위해서 이번 장에서는 자막의 중요한 2가지 특징을 이용한다: (a) 자막은 특정한 형태는 가 .며, 그 형태는 보통 사각형 모양으로써 그 가로 세로 비율기 국한되어 있다 (b) 텍스트 자막은 여러 I-frame을 거쳐 유지된다. 본 절에서는 위에 명시된 자막의 제약조건을 바탕으로 자막 영역을 추출함에 있어서 필요로 하는 3 과정을 다루고자 한다

과정.1 자막 영역 제약 조건

앞에서 다루어진 일련의 과정을 통해 얻어진 고대비 블록들은 4-CC(Component Connectivity) 알고리즘을 통하여 클러스터링 되어, 자막 영역을 형성하고 이 영역을 포함하는 최소한의 경계 영역(Bounding Box)을 얻는다. 이렇게 얻어진 각각의 경계 영역들에 대해서 다음과 같은 조건들을 검사해본다. 우선, 경계영역의 가로 세로 비율에 임계 값을 적용한다.

$$\frac{Width}{Height} > \tau_a \qquad (6)$$

그 다음으로 경계 영역 내에 고대비 블록의 비율이 다음과 같이 주어진 임계 값보다 높은지를 확인해본다.

$$\frac{A_{HightConst\ \ rast}}{A_{TotalArea}} > \tau_b \qquad (7)$$

위의 식에서 $A_{HighContrast}$ 와 $A_{TotalArea}$ 는 각각 경계영역내의 고대비 블록의 개수와 경계영역내의 총 블록의 개수이다. 자막영역은 위에서 주어진 두 가지의 조건들을 모두 만족하므로 이들 조건을 만족하지 않는 영역은 자막 영역으로 간주되지 않는다.

과정.2 자막 유지 시간 제약 조건

그 후에 w의 크기를 가지는 윈도우를 이용함으로써 자막영역이 프레임 내에 일정한 시간 또는 w개의 연속적인 I -프레임동안 존재하는지 확인한다. 여러 비디오를 분석한 결과 사람이 자막을 인지하기 위한 최소한의 자막유지 시간은 약 2초이기 때문에, 본 논문에서는 w=4로 놓았다.

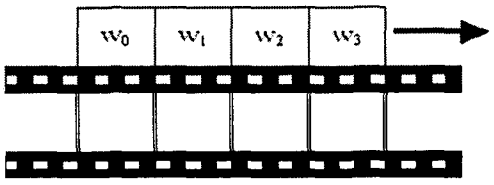


그림 3. 움직이는 윈도우를 이용한 자막 영역 추출

위의 그림 3 과 같이 윈도우를 취하고 윈도영역이 한 프레임 씩 움직이도록 한다. 이때 각각의 윈도우 영역 내에 있는 인접한 두 프레임에서 똑같은 경계 영역이 존재하는지를 확인한다. 앞선 프레임에서의 경계 영역이 다음 프레임의 경계 영역과 60%이상 겹치면 이를 자막으로 간주한다. 그렇지 않을 경우에는 이를 자막으로 처리하지 않는다. 이러한 일련의 과정을 거친 후에 우리는 다음과 같이 표현되는 자막 영역을 결국 얻게 된다.

$$Region_k[\, t^k_{start}, t^k_{end}\,], \quad k = 1,\ldots,N_{Regions} \quad (8)$$

위의 식에서 t^k_{start}, t^k_{end} 은 자막 영역 k 가 존재하는 시점의 시작점과 마지막 시점이다.

III. 실험결과

본 논문에서 제안된 알고리즘의 성능을 평가하기 위해서 총 13,305 I-프레임 분량의 비디오를 이용하였다. 다양한 자막을 포함하고 있는 KBS 뉴스 방송 클립(6802 I-프레임), 및 비디오 골프 레슨 클립(6503 I-프레임)을 이용하였다. 이 중 하나 또는 그 이상의 지문을 가지고 있는 총 프레임의 개수는 172개 였으며, 총 312개의 고유한 지문을 가지고 있다. 알고리즘의 성능을 평가하기 위해서는Precision 과 Recall을 이용하였다. 표 1은 자막 프레임 검출에 대한 결과이다. 실험 결과 자막 영역에 대한 조건만 검사할 때의 recall율은 비록 높게 나왔지만, precision은 자막이 아님에도 불구하고 자막으로 오인된 경우가 많아 그 값이 낮음을 볼 수 있다. 하지만 자막 유지 시간에 대한 조건을 동시에 검사했을 때 자막으로 오인되는 경우가 적어짐을 64%에서 84% 높아진 precision 값을 통해서 확인할 수 있었다. 그와 동시에 recall율은 비슷함을 볼 수 있다.

표 2 는 자막 영역 분할에 대한 결과로써, 여기서 또한 자막 영역 뿐만 아니라 자막 유지시간에 제약을 동시에 둠으로써 precision값이 월등히 높아졌음을 볼 수 있다.

제약조건	Recall	Precision
자막영역	0.92	0.64
자막 영역 및 자막 유지 시간	0.91	0.84

표 1. 자막 프레임 검출에 대한 결과
(총 172개 자막 프레임)

제약 조건	Recall	Precision
자막 영역	0.93	0.6
자막 영역 및 자막 유지 시간	0.92	0.79

표 2. 자막 영역 분할에 대한 결과 (총 312)자막

IV. 결 론

본 논문은 DCT 기반 MPEG 비디오의 I-frame을 대상으로 한 자막 검출 및 추출에 대한 알고리즘을 제안하였다. 자막 영역을 검출 및 추출하기 위해 본 논문에서 제시한 자막의 영역에 조건, 자막 유지시간에 대한 조건 및 자막과 인접한 화소들과의 큰 대비차는 만족할 만한 성능을 보였다. 하지만 8x8 block 단위로 자막 영역을 추출하게 되므로 자막의 일부분이 잘려 나가는 현상과 비교적 큰 크기를 가지는 자막에 대해서 취약점을 보였다. 하지만 압축 상태에서 실시간으로 만족할 만한 수준으로 자막을 추출 할 수 있기 때문에 실시간으로 자막을 추출해야 하는 응용 기술에 이용 될 수 있다.

참고문헌

[1] Y. Zhong, K. Karu and AK. Jain, "Locating text in complex color images," *Proc. of Int. Conf on Document Analysis and Recognition*, Aug. 146-149, 1995.

[2] V. Wu, R. Manmatha and EM. Riseman, "Finding text in images," *Proc. of ACM Int. Conf. on Digital Libraries*, 3-12, 1997.

[3] JE. Miller, SD. Roy, "A generalized algorithm for text detection in digital grayscale images," *IEEE Trans of the 1997 Region I Conference*, 1-6, 1997.

[4] R. Lienhart, "Automatic text recognition for video indexing," *Proc. of ACM MM*, 11-20, 1996.

[5] Y. Zhong, HJ. Zhang and AK Jain, "Automatic caption localization in compressed video," *IEEE Trans on PAMI*, 22(4), 385-392, 2000.