2000년 한국정보처리학회 추계 학술발표논문집 제7권 제2호

오디오와 영상 정보를 이용한 비디오 세그먼테이션 및 크래시피케이션

정해준*, 정성환
창원대학교 전자계산학과
e-mail:hjjung@cosmos.changwon.ac.kr

Segmentation and Classification Using Audio and Image Information

Hae-Jun Jung*, Sung-Hwan Jung

요 약

본 논문에서는 효과적인 내용기반 비디오 검색을 위한 샷 경계 검출, 장면 경계 검출, 그리고 비디오 크래시피케이션 방법을 연구하였다. 먼저, 샷 경계 검출을 위해 칼라 히스토그램과 DCT 변환 계수를 통합하여 사용했다. 그리고 장면 경계 검출을 위해서는 영상 정보뿐만 아니라 오디오 정보를 함께 사용하여 장면 경계를 검출하였다. 또한 비디오 크래시피케이션에서는 장면 경계 검출시 추출한 오디오 정보를 이용해 비디오를 내용별로 분류하는 연구를 제안하였다.

뉴스, 광고, 스포츠 등 다양한 3개 분야의 TV 프로그램으로 구성된 약 8,500개 영상 프레임과 약 50,000개의 오디오 프레임을 가진 실험 비디오 데이터베이스를 구성하여 제안된 시스템을 실험하였다. 실험한 결과, 약 88%의 정확도(Precision)를 가지는 장면 경계 검출과 약 85%의 평균 분류율을 보였다.

1. 서 론

최근 초고속정보통신망의 구축과 H/W, S/W의 발달로 이미지, 비디오 등 멀티미디어 정보를 위한 응용이 많이 등장하고 있다. 멀티미디어 정보 중 비디오는 일반 문자 데이터와 달리 비구조적이며 영상, 음성, 문자 정보 등의 여러 가지 정보를 함께 포함하고 있기 때문에 효과적인 검색 방법이 필요하다.

비디오는 크게 4가지의 계층 즉, 프레임(Frame), 샷(Shot), 장면(Scene), 시이퀀스(Sequence)의 계층으로 구조화함으로써 원하는 부분을 볼 수 있을 뿐 아니라 검색을 위해 주석을 달거나 특징을 추출하는데 기본 단위로 활용이 가능하다. 여기서 프레임은 비디오를 구성하고 있는 계층 중 최하위 계층이며, 하나의 정지영상이다. 프레임의 상위 계층인 샷은 하나의 카메라로 기록된 연속적인 프레임들의 모임이며, 장면은 주제가 같은 내용을 가진 인접한 샷들의 모임이다. 그리고 시이퀀스는 비디오의 최상위 계층이며, 연관된 장면들의 모임이다.

일반적으로 비디오를 검색하기 위한 방법들을 크게 두 가지로 나누면, 주석기반 검색 방법(Caption-based retrieval method)과 내용기반 검색 방법(Content-based retrieval method)으로 나누어진다[1,2].

주석기반 검색 방법은 각 비디오에 대하여 직접 비디오를 보면서 내용을 텍스트로 입력하는 수동적인 방법으로서 비디오의 내용을 자세하게 묘사할 수 있으나, 노력과 시간이 요구되고 사람마다 비디오 내용에 대하여 주관적일 수 있는 단점이 있다. 이러한 단점을 해결하기 위해 내용기반 비디오 검색 방법이 제안되었다. 이 방법은 비디오자체의 시각적 특징정보 즉, 영상 특징 정보와 오디오 특징 정보를 추출해 자동적으로 비디오를 검색할 수 있기 때문에 기존의 주석기반 검색 방법의 단점을 해결할 수 있다.

본 논문에서는 영상 정보와 오디오 정보를 이용하여 효과적인 비디오의 샷(Shot), 장면(Scene) 세그먼테이션 그리고 크래시피케이션을 연구하였다.

정확한 샷 경계(Shot break)를 검출하기 위해 본 연구에서는 간단한 공간영역의 특징인 칼라 히스토그램[3,4]과 주파수 영역의 특징인 DC계수의 통합된 접근 방법을 사용하였다[5]. 그리고 장면 경계 검출(Scene break detection)은 좀더 효과적인 검출을 위해 영상 정보 외에 오디오 특징 중 Volume, Pitch, Silence의 특징 5가지를 함께 사용하였다. 마지막으로 크래시피케이션은 앞에서 추출한 장면의 오디오 특징들을 장면별로 평균해서 특징값들로 사용하였다. 특징값들 중 3가지 이상이 특정 TV 프로그램

형태를 나타낼 때, 이를 해당 TV 프로그램으로 분류하였다.

서론에 이어, 2장에서는 샷 경계 검출에 대해 살펴보고, 3장에서는 영상 정보와 오디오 특징을 이용한 장면 경계 검출 방법에 대해 설명한다. 그리고 4장에서는 제안한 비디오 크래시피케이션에 대해 살펴보고, 5장에서는 실험 결과, 그리고 6장에서는 결론 및 향후 연구과제에 대해 기술한다.

2. 샷 경계 검출

샷 경계를 검출하기 위해 본 논문에서는 공간 영역상의 특징값과 주파수 영역상의 특징값을 함께 사용하였다. 공간영역상의 특징값으로는 식(1)과 같이 칼라 히스토그램을 특징값으로 이용했다. 그리고 픽셀 밝기값 변화에 민감한 칼라 히스토그램의 단점을 보완하기 위하여 주파수 영역상의 특징값들 중 식(2)와 같은 DCT변환 후, DC계수 평균값을 특징값으로 이용하였다[5].

$$\frac{\sum_{j=1}^{N}(|R_i(j)-R_{i+1}(j)|+|G_i(j)-G_{i+1}(j)|+|B_i(j)-B_{i+1}(j)|)}{X\times Y} > T \quad (1)$$

$$C(u,v)=\alpha(u)\alpha(v)\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}f(i,j)\cdot \cos(\frac{(2i+1)u\pi}{2N})\cdot\cos(\frac{(2j+1)v\pi}{2N}) \quad (2)$$

단 $u,v,i,j=0,1,2,\ldots,N-1$

식(1)에서 N은 칼라 레벨의 수이고, $X\times Y$는 프레임의 전체크기이다. 그리고 R_i, G_i, B_i는 i번째 프레임 각각의 RGB 채널별 히스토그램의 값이다.

(그림 1)은 본 논문에서 사용한 샷 경계(컷) 검출에 대한 전체적인 블록도이다.

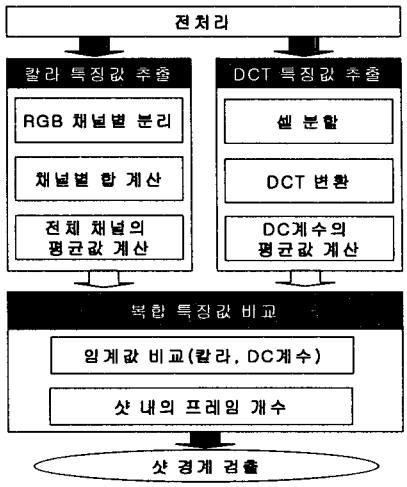
먼저, 전처리에서는 비디오를 초당 10프레임으로 나누고, 각각의 프레임을 128×128 크기로 정규화시킨다. 그리고 칼라 특징값 추출은 각 채널별로 처리한 다음, 전체 채널의 평균값을 계산한다. DCT 특징값 추출에서는 각 셀을 8×8로 분할한 다음, DC계수의 평균값을 구하여 특징값을 추출한다. 추출한 칼라와 DC계수의 특징값을 사용해 각각의 특징에 대한 연속한 프레임간의 차를 계산한다. .

복합 특징값 비교에서는 실제로 샷 경계 검출을 한다. 일반적으로 비디오의 샷 경계 검출에서 샷과 샷사이의 프레임의 개수는 적어도 2~5 프레임을 초과한다.

먼저, 칼라 히스토그램의 방법과 DC계수 방법에서 식(3)과 같이 임계값 비교를 통하여 동일한 샷 경계를 나타낼 경우, 이 때 추출된 샷 경계는 정확한 샷 경계라고 정의한다. 그리고 추가로 앞에서 언급한 바와 같이 추출된 샷 경계에서 전후 연속되는 2~5 프레임 개수 내에 또다른 샷 경계가 나타나는 경우는 정의된 샷 경계에서 제외하는 방식으로 최종적인 샷 경계를 검출하였다.

$$findCutH = H_i > T_h$$
$$findCut = findCutH \text{ and } D_i > T_d \quad (3)$$

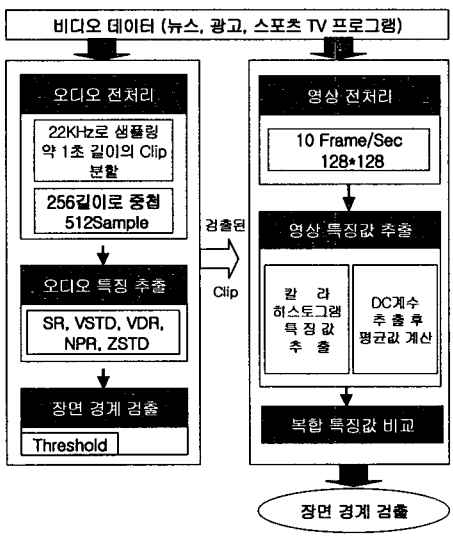
식(3)에서 H_i와 D_i는 각각 칼라 히스토그램와 DC계수의 특징값이다. 그리고 T_h와 T_d는 각각 칼라 히스토그램과 DC계수에 대한 임계값을 나타낸다.



(Fig. 1) The block diagram of the proposed cut detection method

3. 장면 경계 검출

본 논문에서는 효과적인 장면 경계 검출을 하기 위해 비디오에 포함되어 있는 여러 가지 정보 중 오디오 정보와 영상 정보를 함께 사용하였다[6-8]. 먼저, (그림 2)의 왼쪽 부분과 같이 오디오 정보를 사용하여 장면 경계로 검출된 클립(Clip)에 대하여 그림의 오른쪽과 같이 영상 정보를 사용하여 최종적인 장면 경계를 검출하였다. 자세한 각 처리단계를 다음에 기술한다.



(Fig. 2) The block diagram of the proposed scene detection method

(그림 2)의 왼쪽 오디오 처리 부분에서 전처리 과정은 초당 22KHz로 샘플링한 후, 약 1초 길이의 클립으로 분할한다. 그리고 분할된 클립에서 정보의 손실을 줄이기 위해 256길이로 중첩시킨 후 512길이의 프레임으로 정규화 시킨다. 전처리 과정을 거친 클립과 프레임에 대하여 식(4), (5), (6), (7), (8)와 같이 SR(Silence Ratio), VSTD(Volume Standard Deviation), VDR(Volume Dynamic Ratio), NPR(Non-Pitch Ratio), ZSTD(Zero Standard Deviation) 특징값들을 추출한다.

$$SR = \frac{N_{sf}}{N_f} \tag{4}$$

$$VSTD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_i - m)^2} \tag{5}$$

$$VDR = \frac{\max(v) - \min(v)}{\max(v)} \tag{6}$$

$$NPR = \frac{NP_f}{N_f} \tag{7}$$

$$ZSTD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(ZCR_i - m)^2} \tag{8}$$

식(4)에서 N_f는 전체 클립에서 프레임의 수이며, N_{sf}는 Silence 프레임의 수이다. 식(5)에서 n은 전체 클립의 Volume 수이며, v_i는 Volume의 값, m은 전체 클립에서의 Volume의 평균이다. 그리고 식(7)에서 NP_f는 Pitch가 없는 프레임의 개수를 의미하고, 식(8)에서 ZCR는 영교차율(Zero Crossing Ratio)을 의미한다.

앞에서 추출한 특징값들에 대하여 다음 식(9)를 사용하여 (그림 2)의 왼쪽 오디오 처리 블록에서 장면 경계를 포함하고 있는 클립을 검출한다.

$$findSceneA = \left| \frac{1}{N}\sum_{i=-N}^{-1} f(i) - \frac{1}{N}\sum_{i=1}^{N} f(i) \right| \tag{9}$$

식(9)에서 N은 이웃하는 클립의 수이며, $f(i)$는 현재 클립에서 i번째 이웃하는 클립의 특징 벡터이다.

그러나 검출된 클립은 실제 많은 영상 프레임들을 가지고 있으므로 정확한 장면 경계를 찾기 위하여 추가적인 처리가 필요하다. 이를 위하여 (그림 2)의 오른쪽과 같이 검출된 클립에 대하여 칼라 히스토그램과 DC계수의 복합 특징을 이용하여 최종적인 장면 경계를 검출한다.

4. 비디오 크래시피케이션

비디오를 내용별로 분류하기 위해 먼저, (그림 3)과 같이 앞절의 장면 경계 검출시에 사용한 오디오 특징들 즉, SR, VSTD, VDR, NPR, ZSTD을 사용하여 장면별로 각각의 특징값들을 평균한다. 그리고 그 평균한 특징값들이 식(10)과 같이 3가지 이상이 특정 TV 프로그램 형태를 나타낼 때, 이를 해당 TV 프로그램으로 분류한다.

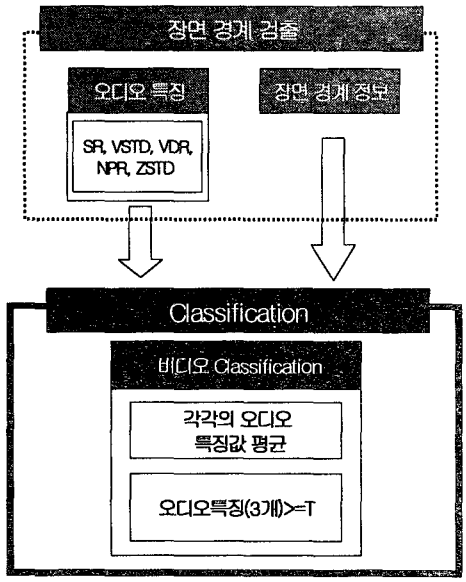
$$F = \{SR, VSTD, VDR, NPR, ZSTD\}, A_f \in F$$

$$I = \{\text{뉴스}, \text{광고}, \text{스포츠}\}, i \in I$$

$$C_f = \begin{cases} 1 & \text{if } TL_{i,f} < A_f < TH_{i,f} \\ 0 & otherwise \end{cases}$$

$$Class(i) \leftarrow \text{if } \sum_{f} C_f \geq 3 \tag{10}$$

식(10)에서 F는 본 논문에서 제안한 오디오 특징값들의 집합이고, I는 실험의 분류 대상인 TV 프로그램 즉, 뉴스, 광고, 스포츠로 구성된 집합이다. 그리고 $TL_{i,f}$, $TH_{i,f}$는 i형태의 분류를 위한 각각의 오디오 특징벡터(A_f)의 최저 임계값과 최대 임계값을 의미한다. $Class(i)$는 실험 데이터가 주어진 조건에 만족하여 i형태로 분류된 것을 말한다.



(Fig. 3) The block diagram of the proposed video classification

5. 실험 결과

본 논문에서는 실험 비디오 데이터베이스를 사용하여 오디오와 영상 정보를 이용한 비디오 세그먼테이션과 크래시피케이션에 대하여 실험하였다.

본 논문에서 사용한 비디오 데이터베이스는 <표 1>과 같이 뉴스(N), 광고(A), 스포츠(S)로 구성된 다양한 9개의 시이퀀스로 구성되어 있고, 전체 8,500개 영상 프레임과 약 850개의 클립을 가진 약 50,000개의 오디오 프레임으로 저장되어 있다.

<표 2>는 장면 경계 검출에 대한 실험 결과이다. 실험에서 28개의 전체 장면 경계 중에서 미 검출된 장면 경계는 없으며, 잘못 검출된 장면 경계는 4개로 나타났다. 따라서 장면 경계 검출에 대한 본 시스템의 정확도(Precision)는 약 88% 이다. 그리고 표에서 잘못된 장면 경계 검출들은 주로 광고와 스포츠를 많이 포함한 시이퀀스들에서 발생한 것을 알 수 있었다. 이는 광고와 스포츠는 서로 유사한 오디오 특성을 가지고 있기 때문인 것으로 판단된다.

<Table 1>Experimental video Database

시이퀀스 형 태	시이퀀스 길 이(초)	영상 프레임 개 수	오디오 클립 개 수
ANS	74	743	74
NAS	70	696	69
SAN	37	377	37
ANAN	100	1,003	100
NAN	53	534	53
SANA	103	1,037	103
ANSAN	133	1,337	133
NSASA	121	1,211	121
SANANSA	152	1,522	152
합 계	847	8,460	846

<Table 2>Scene segmentation results using Color, DC and Audio information

검출 / 주제별	실제 장면 경계	미 검출	잘못 검출된 장면 경계
ANS	2	0	0
NAS	2	0	0
SAN	2	0	0
ANAN	3	0	0
NAN	2	0	0
SANA	3	0	1
ANSAN	4	0	1
NSASA	4	0	0
SANANSA	6	0	2
합 계	28	0	4

<표 3>은 본 논문에서 제안한 방법으로 비디오 크래시피케이션한 결과이다. 실험을 위하여 본 비디오 데이터베이스의 다양한 시이퀀스를 조합하여 사용하였다. 구성된 시이퀀스의 전체 길이는 약 800초 가량의 비디오 데이터이다. <표 3>의 결과가 보여주듯이 뉴스가 입력될 때 100%의 분류율을 나타내고 있지만, 광고와 스포츠는 각각 약 79%, 약 75%의 분류율을 나타내었다. 이는 뉴스는 정형화된 스튜디오에서 방송되기 때문에 스포츠와 같이 백그라운드 사운드 및 잡음이 많은 오디오와 차이가 뚜렷하여 잘 분류되는 것을 확인 할 수 있었다. 그러나 광고와 스포츠는 서로 유사한 오디오 특성을 가지므로 부분적인 오인식이 발생하였다.

<표 3>Classification results using Audio features

In / Out	뉴스	광고	스포츠
뉴스	100.00	0.00	0.00
광고	0.00	78.58	21.42
스포츠	0.00	25.00	75.00

6. 결론 및 향후 연구 과제

본 논문에서는 오디오와 영상 정보를 이용한 비디오 세그먼테이션과 크래시피케이션에 대한 연구를 수행하였다. 본 연구에서는 비디오의 계층적 구조 중 두 번째 계층인 샷 세그먼테이션을 위해 공간 영역상의 특징값 중 칼라 히스토그램과 주파수 영역상의 특징값 중 DC계수를 함께 사용하는 통합 방법을 사용했다. 그리고 세 번째 계층인 장면 세그먼테이션을 위해 오디오와 영상정보를 사용했으며, 마지막으로 비디오 크래시피케이션을 위해 장면 세그먼테이션에서 사용한 오디오 특징을 이용하여 크래시피케이션을 수행하였다.

본 논문에서 제안된 세그먼테이션 및 크래시피케이션 방법을 다양한 시이퀀스를 가진 실제 비디오 데이터베이스에서 실험한 결과, 장면 경계 검출은 약 88%, 그리고 크래시피케이션에서는 뉴스, 광고, 스포츠에서 각각 100%, 79%, 75%의 분류율을 보였다.

본 논문에서 제안한 장면 경계 검출 및 비디오 크래시피케이션에서 광고와 스포츠는 오디오 특징들이 서로 비슷한 특징값을 가지고 있어서 일부 경계 검출과 분류에 영향을 미치는 것으로 확인되었다.

향후 연구 과제로는 현재 비디오 크래시피케이션 단계에서 오디오 특성만 사용하기 때문에 생기는 광고와 스포츠에 대한 분류 문제점을 개선하기 위해서 이 단계에도 영상 정보를 함께 고려함으로써 좀 더 효과적인 크래시피케이션을 연구하는 것이다.

참고 문헌

[1] Rune Hjelsvold, "Video Information Contents and Architecture," 4th International Conference on Extending Database Technology, Mar 1994.

[2] Rune Hjelsvold and Rogar Midtstratum, "Modeling and Querying Video Data," Proc of 20th International Conference on VLDB, pp 686-694, Oct 1994.

[3] Nilesh V. Patel and Ishwar K. Sethi, "Video shot detection and characterization for video databases," Pattern Recognition, Vol.30, No.4, pp.583-592, Apr 1997.

[4] 염성주, 김우생, "동적 영역 히스토그램을 사용한 동영상 데이터의 컷검출 방법", 정보과학회논문지(A) Volume 25, No.3 pp.221-230 Mar 1998.

[5] 정해준, 이우선, 정성환, "칼라 히스토그램과 DC 계수를 이용한 비디오 세그먼테이션," 정보처리학회, CD-paper#43, Oct 1999.

[6] Zhu Liu, Jincheng Huang Yao Wang, Tsuhan Chen, "Audio Feature Extraction & Analysis for Scene Classification," http://vision.poly.edu.

[7] Zhu Liu, Jincheng Huang, Yao Wang, and Tsuhan Chen, "Audio feature extraction & analysis scene classification," http://vistion.poly.edu

[8] Zhu Liu and Qian Huang, "Classification of Audio Events in Broadcast News," http://vision.poly.edu.