오디오와 영상 정보를 이용한 비디오 세그먼테이션

정해준° 정성환
창 원 대 학 교 전 자 계 산 학 과
hjjung@cosmos.changwon.ac.kr

Video Segmentation Using Audio and Image Information

Hae-Jun Jung° Sung-Hwan Jung
Dept. of Computer Science, Changwon National University

요 약

본 논문에서는 영상 정보뿐만 아니라 오디오 정보를 함께 사용한 비디오 세그먼테이션에 대해 연구하였다. 대용량의 정보를 가지고 있는 비디오에 대하여 장면 경계 검출(Scene Break Detection)을 할 경우, 카메라 팬이나 장면 내에 여러 가지 다른 샷(Shot)으로 인하여 영상 정보만으로는 효과적인 검출이 어렵다. 이러한 문제를 해결하기 위해 비디오 내의 오디오 정보도 함께 사용함으로써 문제를 개선했다.

뉴스, 광고, 스포츠 등 다양한 3개 분야의 TV 프로그램으로 구성된 약 4,000개 영상 프레임과 약 30,000개의 오디오 프레임으로 구성된 비디오 데이터베이스에 대하여 실험한 결과, 영상 정보만 사용한 경우보다 우수한 성능을 확인하였다. 영상 정보 특징값으로는 칼라 히스토그램과 DC계수를 사용했고, 오디오 특징값으로는 SR(Silence ratio), VSTD(Volume standard deviation), NPR(Non pitch ratio)을 사용했다.

1. 서 론

멀티미디어 기술의 발달로 최근 멀티미디어 정보 중 대용량의 데이터를 가지고 있는 동영상에 관한 연구가 활발히 진행되고 있다. 그러나 동영상은 일반 문자 데이터와 달리 비구조적이며 영상, 음성, 문자 정보 등의 여러 가지 정보를 함께 포함하고 있기 때문에 효과적인 검색 방법이 필요하다. 영상 정보를 검색하기 위한 방법들은 크게 두 가지로 나누면, 문자기반 검색(Text-based Retrieval Method)과 내용기반 검색 방법(Content-based Retrieval Method)으로 나누어진다[1,2].

문자기반 검색 방법은 각각의 영상에 대해 문자로 주석을 부여해야 하고 영상이라는 시각적 정보를 정확하게 문자로 표현하기가 어렵다는 단점이 있다. 이러한 단점을 해결하기 위해 현재 많이 연구되고 있는 내용기반 검색 방법은 영상자체의 시각적 특징, 즉 영상 특징 정보를 추출해 검색함으로써 기존의 문자기반 검색 방법의 단점을 해결할 수 있다. 그러나 현재 연구되고 있는 내용기반 검색 방법은 대부분 영상 정보만을 고려하고 있다. 따라서 동영상, 즉 비디오를 장면들로 분할하는 장면 경계 검출(Scene Break Detection)은 카메라 팬이나 장면 내에 여러 가지 다른 샷(Shot)으로 인하여 장면 경계 검출이 어렵다. 비디오는 영상 정보 뿐만 아니라 오디오 정보도 함께 가지고 있으므로 본 논문에서는 영상 정보만으로 해결하기 힘든 장면 경계 검출을 오디오 정보와 함께 사용함으로써 이러한 문제를 개선했다.

대용량의 정보를 가지고 있는 비디오에 대한 내용기반 검색은 3가지 단계로 나눌 수 있다.

첫 번째 단계는 샷(Shot)으로 분할하기 위해 각각의 프레임을 분석해 샷 경계(Shot Break)를 검출하는 비디오 세그먼테이션이고, 두 번째 단계는 검출된 샷의 특징을 이용해 주제와 내용이 유사한 장면들로 분할하기 위해서 장면 경계를 검출하는 비디오 세그먼테이션이다. 그리고 마지막 단계인 비디오 크래시피케이션(Classification)으로 이루어진다[3].

본 논문에서는 좀 더 효율적이며 정확한 샷(Shot) 경계를 검출하기 위해 간단한 공간영역의 특징인 칼라 히스토그램과 주파수 영역의 특징인 DC계수의 통합된 접근 방법을 사용했다[4]. 그리고 장면 경계 검출은 샷 경계 검출에서 추출한 값을 이용하여 이웃하는 여러 프레임간의 특징값들의 변화율을 사용하였다. 또한 좀 더 효과적인 장면 경계 검출을 위해 영상 정보 외에 오디오 특징 중 Volume, Pitch, Silence를 함께 사용하였다.

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

2. 샷 경계 검출

2.1 칼라 히스토그램의 차이를 이용한 방법

칼라 히스토그램의 차이를 이용한 방법은 식(1)과 같이 공간영역에서 서로 인접하는 프레임간에 RGB칼라 채널별로 히스토그램을 구한 후, 각각의 채널별 차이의 합을 구하여 평균한다[5]. 그리고 연속되는 프레임간에 이 평균값 차를 이용함으로써 샷 경계를 검출한다. 이 특징값은 샷 경계, 즉 컷(Cut)이 있는 프레임간의 차는 크고 그렇지 않는 곳에서는 차가 작다.

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

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.

2.2 DCT를 이용한 방법

DCT을 이용한 방법은 영상 프레임에 대해 기존의 공간 영역에서 추출된 특징값과는 달리 변환영역, 즉 주파수 영역에서의 특징값을 얻는 방법[6] 중에 하나이다.

먼저, 각 프레임을 셀로 분할하여 식(2)와 같이 DCT(Discrete Cosine Transform)변환을 수행한다. 다음으로 변환된 계수로부터 DC정보을 추출하고 추출된 정보를 이용해 인접한 프레임에서 같은 위치에 있는 셀의 DC계수 차이를 구한다. 이렇게 구한 DC계수의 차이를 합하여 프레임의 전체 셀의 개수로 나눈 평균값을 특징값으로 사용한다.

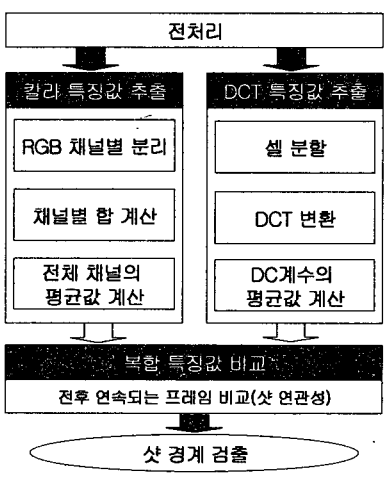
$$C(u,v) = \alpha(u)\alpha(v)\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} f(i,j) \cdot \cos\left(\frac{(2i+1)u\pi}{2N}\right) \cdot \cos\left(\frac{(2j+1)v\pi}{2N}\right) \quad (2)$$

단, $u,v,i,j = 0,1,2,\ldots,N-1$

2.3 칼라 히스토그램과 DC계수 통합 방법

샷은 카메라 조작이나 비디오에서 프레임을 추출시 또는 일반적으로 포함될 수 있는 잡음 때문에 같은 샷이지만 픽셀의 밝기값에 심한 변화를 일으킬 수 있다. 이런 변화에 좀 더 강인성을 가지게 하기 위하여, 본 연구에서는 (그림 1)과 같은 칼라 히스토그램과 DC계수 통합 방법을 사용하였다.

먼저, (그림 1)과 같이 비디오가 입력되면 비디오 전처리 과정을 거쳐 칼라 히스토그램 특징값과 DCT 특징값을 추출한다. 그리고 복합 특징값 비교 단계에서 추출된 샷 경계들이 칼라 히스토그램의 방법과 DC계수 방법에서 동일한 샷 경계를 나타낼 경우, 이 때 추출된 샷 경계는 정확한 샷 경계라고 할 수 있다. 그러나 일반적으로 비디오의 샷과 샷사이의 프레임의 개수는 적어도 2~5 프레임을 초과한다. 따라서 추출된 샷 경계에서 전후 연속되는 2~5 프레임내에 또다른 샷 경계가 나타나는 경우는 이를 최종 샷 경계에서 제외한다.



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

3. 장면(Scene) 경계 검출

3.1 칼라와 DC계수를 이용한 장면 경계 검출

장면 경계 검출을 위해 앞 절의 샷 경계 검출에 사용한 칼라 히스토그램과 DC계수 방법과 같은 특징을 이용한다. 그러나 장면 검출은 컷(Cut) 검출과 달리 카메라의 팬 또는 장면(Scene)내에서 서로 다른 칼라 분포를 가지고 있는 샷(Shot)이 있다. 이러한 문제를 해결하기 위해 식(3), 식(4)와 같이 서로 이웃하는 프레임들사이의 연관성을 이용해 장면 경계를 검출한다.

검출된 프레임은 식(5)와 같이 칼라의 임계값과, DC계수의 임계값이 모두 클 경우에만 최종 장면 경계로 검출한다.

$$R_c = \frac{D_k + c}{\min(\mu_-, \mu_+) + c} \quad (3)$$

$$R_d = \frac{D_c + c}{\min(\upsilon_-, \upsilon_+) + c} \quad (4)$$

$$findSceneV = \{R_c > T_c \text{ and } R_d > T_d\} \quad (5)$$

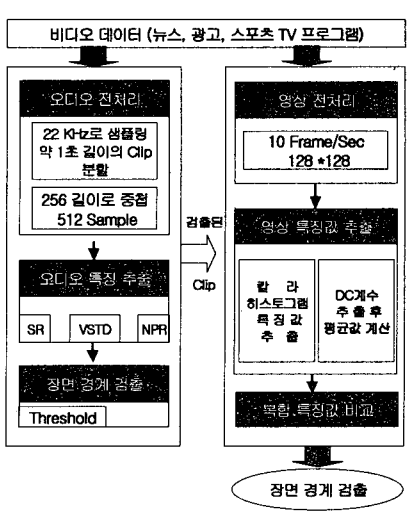
식(3)에서 μ_-, μ_+는 각각 전후 연속되는 N개 프레임의 히스토그램 평균값 차이의 합이고, D_k는 현재 프레임에서 인접하는 프레임의 히스토그램 평균값 차이다. 그리고 c는 임의의 상수이다. 식(4)에서 υ_-, υ_+은 각각 전후 연속되는 N개 프레임의 DC계수 차이의 평균값 합이며, D_c는 현재 프레임에서 인접하는 프레임의 DC계수 차이의 평균값이다. 식(5)에서와 같이 칼라 히스토그램의 R_c와 DC계수의 R_d가 T_c, T_d 임계값보다 각각 크면 장면 경계로 검출한다.

3.2 오디오 정보와 영상 정보를 이용한 장면 경계 검출

비디오에는 영상뿐만 아니라 오디오 정보도 함께 포함되어 있다. 일반적으로 뉴스와 같이 스튜디오에서 방송되는 오디오와 광고나 스포츠와 같이 백그라운드 사운드 및 잡음이 많은 오디오는 여러면에서 큰 차이가 있다.

그러므로 본 논문에서는 영상 정보와 함께 오디오 정보를 이용해 주제가 서로 유사한 장면들로 분할하기 위한 장면 경계 검출을 하였다[3,7].

(그림 2)는 제안한 장면 경계 검출 방법에 대한 전체적인 블록도이다. 먼저, 오디오 정보를 이용하여 개략적인 장면 경계 주변을 검출한 후, 영상 정보를 이용하여 최종적인 장면 경계를 검출한다.



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

(그림 2)에서 먼저, 오디오 신호를 분석하기 위해 오디오 시이퀀스(Sequence)를 약 1초(22KHz/sec) 길이의 클립(Clip)으로 분할 후, 각 클립의 정보 손실을 막기 위하여 256길이로 중첩하여, 512샘플(Sample)길이의 프레임들로 나눈다. 그리고 본 논문에서는 다음 식(6), (7), (8)과 같이 전체 클립에서 Silence

Ratio(SR), Volume Standard Deviation(VSTD), 그리고 Non-Pitch Ratio(NPR)를 오디오 특징들로 추출하였다.

$$SR = \frac{N_{sf}}{N_f} \quad (6)$$

$$VSTD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2} \quad (7)$$

$$NPR = \frac{NP_f}{N_f} \quad (8)$$

식(6)에서 N_f는 전체 클립에서의 프레임의 수이며, N_{sf}는 Silence 프레임의 수이다. 식(7)에서 n은 전체 클립의 Volume 수이며, x_i는 Volume의 값, m은 전체 클립에서의 Volume의 평균이다. 그리고 식(8)에서 NP_f는 Pitch가 없는 프레임의 개수를 의미한다.

오디오 특징들을 추출한 후, 장면 경계 검출을 위해서 식(9)를 이용했다.

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

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

(그림 2)의 왼쪽의 오디오 처리의 마지막 단계에서 장면 경계를 가진 것으로 검출된 클립은 실제 많은 영상 프레임들을 가지고 있다. 따라서 정확한 장면 경계를 찾기 위하여 (그림 2)의 오른쪽과 같이 영상 정보를 이용한 샷 경계 검출 기법을 추가로 적용하여 최종적인 장면 경계를 검출한다.

4. 실험 결과

본 논문에서 사용한 실험 데이터는 <표 1>과 같다. <표 1>에서 A는 광고, N는 뉴스, 그리고 S는 스포츠 TV 프로그램을 나타낸다. 전체 실험 데이터를 살펴보면, 다양한 3개 분야의 TV 프로그램으로 구성된 약 4,000개 영상프레임과 약 30,000개의 오디오 프레임으로 구성된 비디오 데이터베이스에 대하여 실험하였다. 본 실험의 전처리 과정에서는 영상 프레임을 128×128 크기로 정규화하여 사용하였다. 그리고 DC계수를 이용하는 방법에서의 셀의 크기는 8×8 로 나누어 실험하였다.

<Table 1> Classification of experimental video data

시이퀀스 형 태	시이퀀스 길 이(초)	영상 프레임 개 수	오디오 클립 개 수
ANT	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
합 계	437	4,390	436

<표 2>는 영상 정보만을 사용해 세그먼테이션한 실험 결과이다. 표에서 보는 바와 같이 영상 정보만을 사용함으로써 생기는 잘못 검출된 장면 경계의 수는 전체 9개가 검출되었다.

<표 3>은 제안한 오디오와 영상 정보를 함께 사용한 세그먼테이션 방법의 실험 결과이다. <표 3>에서 보여 주듯이 오디오 정보를 함께 사용함으로써 전체 잘못 검출된 장면 경계는 3개로서, <표 2>에 비하여 6개가 줄어들었다. 그러나 광고(A)와 스포츠(S)는 오디오 특징 중 SR에서 서로 비슷한 특징값을 가지고 있기 때문에 여전히 잘못 검출된 장면 경계가 일부 나타났다.

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

검출 / 주제별	실제 장면 경계	미 검출	잘못 검출된 장면 경계
ANS	2	0	0
NAS	2	0	2
SAN	2	0	1
ANAN	3	0	2
NAN	2	0	1
SANA	3	1	3
합 계	14	1	9

<Table 3> Scene segmentation results using Color, DC and Audio information

검출 / 주제별	실제 장면 경계	미 검출	잘못 검출된 장면 경계
ANT	2	0	0
NAS	2	0	1
SAN	2	0	1
ANAN	3	0	0
NAN	2	0	0
SANA	3	0	1
합 계	14	0	3

5. 결론 및 향후과제

본 논문에서는 효율적인 비디오 세그먼테이션을 위해 영상 정보 뿐만 아니라 오디오 특징도 함께 이용하는 비디오 세그먼테이션을 제안하였다.

본 논문에서 제안한 방법은 오디오 정보를 이용해 먼저 개략적인 장면 경계 주변을 검출한다. 그리고 검출된 장면 경계 주변에 대하여 좀 더 세밀하게 접근할 수 있는 영상 정보를 이용해 최종적인 장면 경계 검출을 하였다.

실험 데이터는 뉴스, 광고, 스포츠로 구성된 6개의 다양한 시이퀀스를 만들어 실험하였다. 실험결과, 영상 정보만 사용한 방법보다 제안한 방법이 잘못 검출된 장면 경계 수를 1/3로 줄일 수 있었다. 그러나 오디오 특징 중 SR에서 서로 유사한 특징값을 가지는 광고(A)와 스포츠(S)간에서는 잘못 검출된 장면 경계가 일부 나타났다.

향후 연구 과제로는 효과적인 장면 경계 검출을 위해 좀 더 향상된 오디오 특징을 연구하는 것이다.

참고 문헌

[1] H. J. Zhang, C. Y. Low, S. W. Smoliar and J. H. Wu, "Video Parsing Retrieval and Browsing: An Integrated and Content-Based Solution," Proc. ACM Multimedia '95. pp.15-24, 1995.

[2] N. Dimitrova and M. A. Mottalab, "Content-based video retrieval by example video clip," SPIE Storage and Retrieval for Image and Video Database, Vol. 3022, pp.59-70, 1996.

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

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

[5] N. V. Patel and I. K. Sethi, "Video Shot Detection and Characterization for Video Databases," Pattern Recognition: Special issue on multimedia, 1996.

[6] 홍수민, "시그너처를 이용한 내용기반 비디오 세그먼테이션," 창원대학교 석사학위논문, 1999.

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