지능형 PVR을 위한 축구 동영상 하이라이트 요약

*김형국 **신동

광운대학교

{hkim, shinloves}@kw.ac.kr

Soccer Video Highlight Summarization for Intelligent PVR

*Kim, Hyoung-Gook **Shin, Dong

Kwangwoon University

요약

본 논문에서는 MDCT기반의 오디오 특징과 영상 특징을 이용하여 축구 동영상의 하이라이트를 효과적으로 요약하는 방식을 제안한다. 제안하는 방식에서는 입력되는 축구 동영상을 비디오 신호와 오디오 신호로 분리한 후에, 분리된 연속적인 오디오 신호를 압축영역의 MDCT계수를 통해 이벤트 사운드별로 분류하여 오디오 이벤트 후보구간을 추출한다. 입력된 비디오 신호에서는 장면 전환점을 추출하고 추출된 장면 전환점으로부터 페널티 영역을 검출한다. 검출된 오디오 이벤트 후보구간과 검출된 페널티 영역장면을 함께 결합하여 축구 동영상의 이벤트 장면을 검출한다. 검출된 페널티 영역 장면을 통해 검출된 이벤트 구간을 다른 이벤트 구간보다 더 높은 우선순위를 갖는 하이라이트로 선정하여 요약본이 생성된다. 생성된 하이라이트 요약본의 평가는 precision과 recall을 통해 정확도를 평가하였다.

1. 서론

최근 디지털 TV와 셋탑박스, PVR 등의 보급으로 원하는 프로그램을 원하는 시간에 녹화하여 제공하는 서비스가 증가하고 있다. 이러한 TV 프로그램 시청방식은 사용자가 원하는 장면만을 볼 수 있도록 하이라이트로 요약하는 서비스를 요구한다. 이러한 하이라이트 요약 서비스 없이 사용자는 원하는 장면을 찾기란 어렵고 시간이 많이 소요된다. 특히, 장시간 지속되는 스포츠 동영상을 다양한 구간별로 분류 및 분할하고, 분할된 구간에서 중요한 이벤트와 관련된 장면들을 검출하여 하이라이트를 요약하는 다양한 연구들이 진행되어 오고 있다.

이러한 스포츠 동영상에 대한 내용기반 이벤트 검출 방식은 다음과 같다. 먼저 스포츠 동영상의 기본 플레이 구간을 추출하는 연구로는 고정된 템플릿을 사용하여 낮은 수준의 칼라 특징 기반의 플레이 구간과 나머지 구간을 분할하는 방법[1], 모델 학습 방법을 사용한 특정 도메인 정보 기반의 검출 방법[2] 등이 연구되고 있는데, 고정된 템플릿이나 고정된 모델을 사용한 방식은 다양한 방송 조건에 따라 기본 플레이 구간 검출성능이 일정하지 않다. 또한 중요 이벤트 장면 검출 방식에 대한 연구로는 오디오 기반 검출 방법[3], 비디오 기반 검출 방법[4], 자막 기반 검출 방법[5], 멀티모달 기반 검출 방법[6] 등으로 연구되고 있는데, 오디오 기반 검출 방법은 이벤트 시작-끝점의 검출이 부정확하다. 나머지 방식은 TV에 적용되기 어렵다.

본 논문에서는 압축영역의 MDCT 계수 기반의 오디오 특징과 영상 특징을 이용하여 이벤트 검출이 가능한 하이라이트 요약 생성 방식을 제안한다. 중요 이벤트 장면을 빠르고 정확하게 검출하기 위해 영상신호의 장면 전환점을 검출하여 키 프레임에 대한 페널티 영역 검출 및 압축영역의 오디오신호에서 오디오 이벤트 후보구간 검출 결과를 결합한다. 검출된 이벤트 장면은 오디오 이벤트 후보구간의 길이와 페널티 영역 검출 여부에 따라 우선순위를 결정하여 하이라이트로 사용자에게 제공된다.

본 논문의 구성은 다음과 같다. 2 장에서는 전체 시스템 구성도에 대해 설명하고, 3장에서는 오디오신호를 이용한 오디오 이벤트 후보구간 추출 방법, 4장에서는 장면 전환점 추출방법에 대해 설명한다. 5장에서는 추출된 장면 전환점을 기준으로 페널티 영역을 검출방법에 대해 설명하며, 6장에서는 검출된 페널티 영역과 오디오 이벤트 후보구간을 결합한 이벤트 장면 결정방법 및 최종적인 하이라이트 요약 구성 방법에 대해 설명한다. 실험 결과 및 분석은 7장에서 설명하며 마지막으로 8장에서는 결론과 향후 연구 방향을 기술한다.

2. 전체 시스템 구성도

본 논문에서 제안하는 하이라이트 요약 시스템의 구성도는 그림 1과 같다.

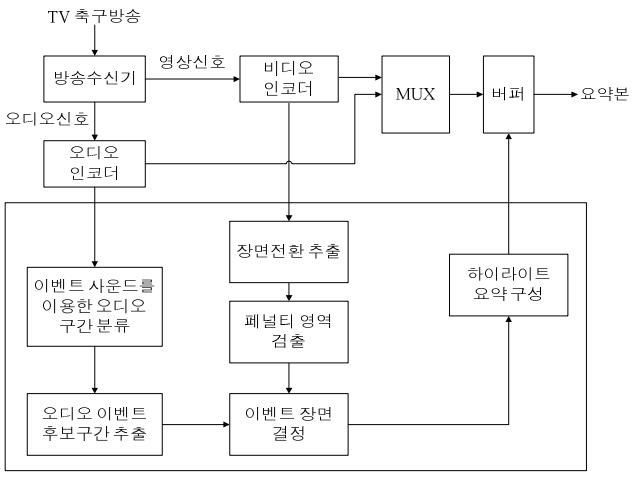

그림 1. 전체 시스템 구성도

입력된 TV 방송 신호는 방송수신기를 통해 영상신호와 오디오신호로 분리된다. 분리된 오디오신호로부터 오디오 인코더를 통해 압축 영역의 modified discrete cosine transform (MDCT)계수를 추출한다. MDCT 계수를 사용하여 오디오신호를 이벤트 사운드 구간별로 구분하게 된다. 분류된 오디오 이벤트 사운드 구간으로부터 오디오 이벤트 후보구간을 검출한다. 입력되는 영상신호는 비디오 인코더를 통해 추출된 RGB정보를 이용하여 장면 전환점을 추출하고 페널티 영역을 검출한다. 검출된 오디오 이벤트 후보구간과 검출된 페널티 영역장면을 함께 결합하여 축구에서의 이벤트 구간을 검출한다. 검출된 이벤트 구간은 장면 전환점과 검출된 페널티 영역과 관련되어 우선순위가 결정된다. 페널티 영역을 포함하는 하이라이트를 우선순위로 하는 요약본을 생성한다.

3. 오디오 이벤트 후보구간 추출

축구 동영상에서 골, 패스, 코너킥, 개인기 장면은 일반적으로 관중들의 함성소리를 동반한다. 이러한 중요한 이벤트 장면들은 오디오 신호를 통해 분류할 수 있다. 입력된 오디오 신호는 AC-3 오디오 인코더로부터 MDCT계수를 추출한다. 추출된 MDCT계수를 식(1)과 같은 log단위 21차 필터뱅크를 통해 스무딩시킨다.

$$H(p,f)=\begin{cases}0, & 1<f(p-1)\\ \dfrac{2l-f(p-1)}{f(p+1-f(p-1)f(p)-f(p-1)}, & f(p-1)\le l\le f(p)\\ \dfrac{2l-f(p-1)}{f(p+1-f(p-1)f(p)-f(p-1)}, & f(p-1)\le l\le f(p)\\ 0, & l<f(p+1)\end{cases} \quad (1)$$

식에서 p는 차수, I은 MDCT 서브밴드 개수를 나타낸다.

추출된 21차의 MDCT 특징에서 식 (2)와 같은 로그에너지인 logarithmic MDCT (*LMDCT*)를 계산한다.

$$LMDCT(p)=\ln\left(\sum_{l=0}^{576}|MDCT(l)|H(p,l)\right) \quad (2)$$

계산된 *LMDCT* 특징은 식 (3)과 같은 데시벨 단위로 변환된다.

$$LMDCT_{dB}(n,p)=10\log_{10}(LMDCT(n,p)) \quad (3)$$

식에서 n은 프레임 지수를 나타낸다.

각 데시벨 단위의 특징은 식 (4)와 같은 root-mean square (RMS) 에너지를 통해 정규화하여 식 (5)와 같은 normalized LMDCT (*NLMDCT*)를 추출한다.

$$R_n=\sqrt{\sum_{p=1}^{P}(LMDCT_{dB}(n,p))^2} \quad (4)$$

$$NLMDCT(n,p)=\frac{LMDCT_{dB}(n,p)}{R_n} \quad (5)$$

추출된 *NLMDCT* 계수에서 스펙트럼의 변화를 고려하기 위해 delta 계수를 사용한다. Delta *NLMDCT*는 식 (6)과 같이 인접 프레임 간의 차이로부터 생성된다.

$$\Delta c(l)=-c(l-2)-\frac{1}{2}c(l-1)+\frac{1}{2}c(l+1)+c(l+2) \quad (6)$$

사람의 인지특성을 고려하여 *LMDCT* 계수를 4개의 서브밴드로 분리한다. 4개의 서브밴드는 0-630Hz, 630-1720Hz, 1720-4400Hz, 4400Hz와 나머지 부분으로 구성된다. 사람의 음성 에너지는 대부분 중간의 두 서브밴드에 존재하기 때문에 이 두 서브밴드의 음성 에너지 SE_{23}로부터 에너지들을 추출한다. 음성 특징으로 *NLMDCT* 계수, delta *NLMDCT* 계수, RMS 에너지, 음성 에너지 SE_{23}를 사용하여 support vector machine (SVM)기반의 오디오 이벤트 모델을 생성한다. SVM기반의 오디오 분류 방식은 입력된 특징의 우도가 최대가 되는 오디오 이벤트 모델을 찾는다. 이렇게 이벤트 구간으로 분류된 오디오 구간들은 이벤트 후보구간으로 지정한다.

일반적으로 이벤트의 끝 부분에는 아나운서와 관중의 흥분된 소리가 발생되며, 이벤트는 이보다 오래 지속적으로 나타난다. 이러한 오디오 구간은 이벤트 후보구간으로 선택된다. 중요한 이벤트 일수 록 더 긴 구간에서 오디오 이벤트가 발생하게 된다. 실험적으로 최소 10초 길이의 구간을 이벤트 후보구간으로 결정한다. 이러한 이벤트 후보구간에서 실질적인 중요 이벤트 구간을 검출하기 위해 음성 에너지를 측정한다. 저 주파수 밴드의 에너지를 포함하지 않는 음성에너지 SE_{23}은 비 음성 신호의 에너지 레벨이 큰 경우와 음성 신호의 에너지 레벨이 큰 경우를 구분할 수 있다. 또한, 높은 분산을 갖는 배경잡음은 delta *NLMDCT*를 통해 분류할 수 있다.

4. 장면 전환점 추출

입력되는 영상신호에서는 MPEG-2 인코더를 통해 RGB정보가 추출된다. 효율적인 장면 전환점을 추출하기 위해 pixel 레벨 비교와 히스토그램 비교 방식을 결합하여 적용한다. 일반적인 pixel 레벨 비교 방식은 연속적인 두 프레임간의 동일한 위치의 픽셀간의 차이 값을 비교하여 장면 전환 여부를 결정한다. 이 방식은 계산이 간편하지만 사물의 변화, 음영의 변화에 민감하다는 단점이 있다. 이를 보완하기 위해 히스토그램 비교 방식을 적용하여 인접한 프레임 간의 히스토그램 차이를 통해 이러한 오류를 보완한다. 위 두 방식을 통해서도 검출하기 어려운 장면 전환점은 object 분할 및 추적방식[7]을 통해 결정한다. 추출된 object의 주 배경화면이 무엇인지에 따라 그림 2와 같이 장면 전환여부를 결정하게 된다.

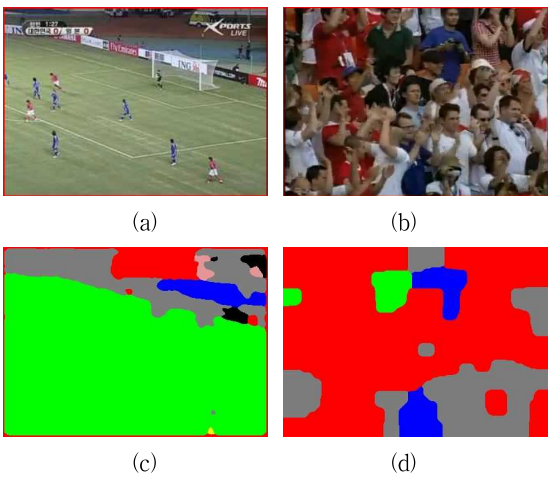

그림 2. (a) 골 장면 샘플; (b) 골 장면 후의 응원 장면 샘플; (c) 그림 2.(a)의 object 분할 결과; 그림 2.(b)의 object 분할 결과

그림 2.(a)는 골이 발생하기 직전의 장면이며 그림 2.(b)는 골 발생 후 응원석을 비추는 장면이다. 골 발생 후 장면 전환점이 발생하는 것을 object 분할을 통해 검출할 수 있다.

5. 페널티 영역 검출

일반적인 축구 경기에서 골 장면은 가장 중요한 이벤트이며 페널티 영역 장면을 추출함으로써 검출된다. 페널티 영역 검출은 경기장을 구성하는 잔디의 일반적인 색인 초록색부분을 영상에서 검출하고, 검출된 초록색 부분에서 페널티 영역 근처라고 판단되는 영상을 찾게 된다. Adaptive dominant green방식[8]을 통해 영상에서 초록색 영역과 그 이외의 영역으로 이진화 하여 분류한다. 검출된 초록색 영역은 coarse spatial representation (CSR)방식[8]을 통해 32 x 32 서브샘플 단위로 경기장 그라운드인지 아닌지를 판단하여 이진화를 수행한다. CSR을 통해 입력된 영상 신호의 대략적인 초록색 분포를 판단할 수 있으며, 그라운드의 중앙 부분은 사각형의 분포, 페널티 영역은 계단 형식의 분포를 보인다. 검출된 CSR의 초록색 영역에서 Hough 변환 선 검출[9]방식을 수행하여 에지 잡음을 감소 시켜 Hough 선을 검출한다. 검출된 Hough 선은 영상신호에서 비 초록색 영역 밑에 위치하게 되면 페널티 영역으로 판단된다. 검출된 페널티 영역에서 골라인이 좌측과 우측에 각각 존재하게 되는데 이 부분을 검출함으로써 페널티 영역을 결정한다. 왜냐하면 경기장 그라운드의 중앙부분에도 흰색 선이 위치하기 때문이다. 이를 구분하기 위해 좌측과 우측의 골라인은 중요한 구분 요소가 된다. 검출된 좌우측 골라인과 연결된 수평형태의 크로스바를 검출함으로써 그림 3과 같은 최종적인 페널티 영역을 결정하게 된다.

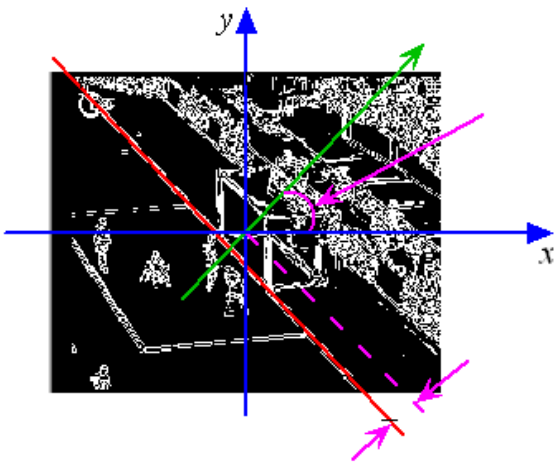

그림 3. 페널티 영역 검출

6. 이벤트 장면 검출 및 하이라이트 요약 구성

이벤트 장면 검출은 오디오신호에서 검출된 오디오 이벤트 후보 구간과 영상신호에서 검출된 페널티 영역장면을 함께 결합하게 된다. 그림 4는 이벤트 장면검출에 대한 예를 나타낸다.

입력된 축구 동영상의 분리된 영상신호에서 장면 전환점을 그림 4와 같이 추출한다. 추출된 장면 전환점을 기준으로 키 프레임에서 페널티 영역 검출을 수행한다. 입력된 오디오신호에서는 오디오 이벤트 구간 분류를 수행한다. 일정한 세그먼트 단위로 분류된 신호는 일정한 구간동안 연속적으로 이벤트로 분류된 구간을 오디오 이벤트 후보구간으로 선정한다. 검출된 페널티 영역 장면과 오디오 이벤트 후보구간을 결합하여 최종적인 이벤트 구간으로 선정한다. 이벤트 구간 중 오디오 신호가 가장 강한 에너지를 갖는 이벤트 구간을 첫 번째 하이라이트 구간으로 지정한다. 또한 페널티 영역이 포함된 하이라이트 구간을 우선순위로 하여 하이라이트 요약본이 생성된다.

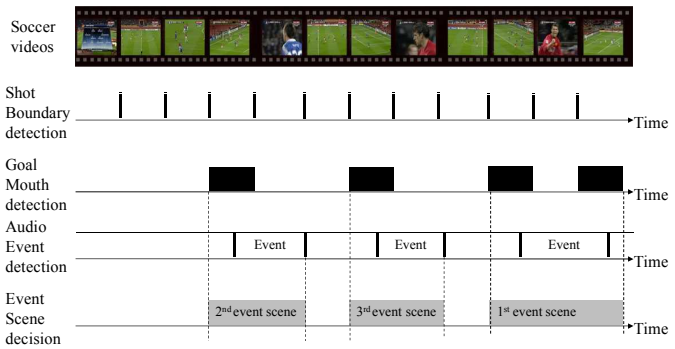

그림 4. 이벤트 장면 검출

7. 실험 결과 및 분석

본 논문에서는 제안하는 알고리즘의 정확성을 증명하기 위해 국내 축구 동영상 10경기, 해외 축구 동영상 20경기를 대상으로 실험을 수행하였다. 이벤트 검출 알고리즘의 정확도를 판단하기 위해 식 (7)과 같은 precision과 recall을 적용하여 성능을 측정하였다.

$$Precision = \frac{N_c}{N_c + N_f} \tag{7}$$

$$Recall = \frac{N_c}{N_c + N_m}$$

식에서 N_c는 정확하게 찾은 이벤트 개수, N_m은 검출이 되지 않은 이벤트 개수이며, N_f는 잘못 검출된 이벤트 개수를 나타낸다.

표 1. 이벤트 장면 검출 실험 결과

	Recall	Precision
페널티 장면검출	95.6%	98.7%
골인과 같은 하이라이트 검출	98.7%	96.5%
골인이외의 하이라이트 검출	97.8%	78.7%

표 1에 나타난 바와 같이 축구 경기에서 가장 많은 이벤트가 발생하는 페널티 장면 검출 성능은 95.6%의 높은 검출성능을 보이고 있으며, 검출하지 못한 경우를 살펴보면 페널티 영역 검출 시 사용되는 좌우측 골라인 중 하나의 골라인이 선수에 의해 완전히 가려진 경우 에러가 발생하게 된다. 골인과 같은 중요장면 검출 성능은 98.7%의 높은 것을 알 수 있으며, 오디오 이벤트 후보 구간과 페널티 영역 검출 결과의 결합을 통해 검출 가능한 것으로 판단된다. 골인이외의 중요장면 검출 같은 경우 정확도에서 상당히 낮은 것을 알 수 있다. 오디오 이벤트 후보 구간 검출 시 검출된 오디오 이벤트로부터 일정 시간 지속될 경

우 오디오 이벤트 후보 구간으로 결정하기 때문에 관중의 환호성이 골과 상관없이 큰 경우 오류가 발생할 수 있다. 예를 들면, 선수 교체, 특정 선수의 비신사적 행동 등의 중요장면과는 관계가 적은 장면들에서 오디오 이벤트 후보 구간으로 발생한다.

8. 결론 및 향후 연구 방향

본 논문에서는 축구 동영상에서 압축영역의 MDCT 계수 기반의 오디오 특징과 영상특징을 이용하여 검출된 이벤트 장면으로부터 하이라이트 요약 생성 방식을 제안한다. 제안된 방식은 오디오신호와 영상신호를 사용하여 각각 오디오 이벤트 후보 구간과 페널티 영역을 검출하기 때문에 골인과 같은 하이라이트 검출 성능이 98.7%의 높은 성능을 보여주었다. 제안된 방식의 검출 성능은 다양한 축구 경기에서 골인과 같은 하이라이트 검출에 적용가능하다는 것을 보여준다.

향후 계획으로는 제안된 방식을 휴대용 멀티미디어 기기에 효과적으로 적용하는 방법을 연구할 것이다.

[참고문헌]

[1] L. Baoxin and M. I. Sezan, "Event detection and summarization in sports video," IEEE Workshop on Content based Access of Image and Video Libraries, pp. 132-138, 2001.

[2] D. Zhong and S. F. Chang, "Structure analysis of sports video using domain models," Proceedings of International Conference on Multimedia and Expo, pp. 713-716, 2001.

[3] A. Divakaran, A. Betro, K. Asai, and H. Nishikawa, "Video browsing system based on compressed domain feature extraction," IEEE Trans. on Consumer Electronics, Vol. 36, No. 3, pp-637-644, 2000.

[4] L. Xie, P. Xu, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with domain knowledge and hidden Markov models," Science direct Pattern recognition letters, Vol. 25, pp. 767-775, 2004.

[5] D. Zhang, and S. F. Chang, "Event detection in baseball video using superimposed caption recognition," Proc. of the 10th ACM international Conf. on Multimedia, France, pp. 315-318, Dec. 2002.

[6] W. Qi, L. Gu, H. Jiang, X.-R. Chen, and H.-J. Zhang, "Integrating visual, audio and text, analysis for news video," IEEE International Conf. on Image Processing, Canada, Vol. 3, pp. 520-523, 2000.

[7] S.-C. Chen, M.-L. Shyu, C. Zhang, L. Luo and M. Chen, "Detection of soccer goal shots using joint multimedia features and classification rules," Procceding of the 4th International Workshop on Multimedia Data Mining, USA, 2003

[8] K. Wan, X, Yan, X, Yu and C. Xu, "Real-time goal-mouth detection in MPEG soccer video," Procceding of the 11th ACM international Conf. on Multimedia, pp. 311-314, USA, 2003.

[9] http://www.intel.com/research/mrl/research/opencv/overview.htm