농구 비디오에서 특정 음성 특징 추출에 관한 연구

공현장, 김원필, *김판구
조선대학교 대학원 전자계산학과
*조선대학교 컴퓨터공학부

A Study on the Extraction of Specific Audio Feature In Basketball Video

Hyunjang Kong, Wonpil Kim, *Pankoo Kim
Dept. of Computer Science. Graduate School of Chosun Univ.
* Dept. of Computer Science and Engineering, Chosun Univ.
E-mail : kisofire@hotmail.com, * pkkim@mina.chosun.ac.kr

요 약

최근 멀티미디어 정보 시스템에서의 음성 및 시각적 내용의 분류에 관한 연구가 활발히 진행되고 있다. 이에 본 논문에서는 농구 경기의 비디오 데이터로부터 특정 음성 정보를 추출하는 방법과 이를 농구 게임의 중요 이벤트 검출에 이용하는 방법을 제안한다. MFCC 특징들과 LPC 엔트로피의 조합을 이용하여 검출된 관중들의 환호 소리로부터 중요한 이벤트의 위치를 예측할 수 있다. 농구 경기의 다양한 소리들 중에서 관중들의 환호 소리를 분류하여 이를 농구 비디오 데이터에서 중요한 이벤트들을 검출하는데 사용함으로써 매우 효과적 결과를 얻을 수 있었다.

1. 서론

멀티미디어 정보 시스템에서 음성 및 시각적 내용의 분류에 대한 연구가 활발히 진행되고 있다. 이러한 연구들은 내용기반 검색을 가능하게 하며, 자동적 객체 분류, 그리고 사용자로부터 피드백을 이용해 학습기반 자동 검색 등에 지원 할 수 있다. 이러한 기술의 발전은 텍스트 키워드를 통한 검색에 의존한 기존의 기술에 비해 괄목할만한 발전이다. 그렇지만, 이러한 기술들도 상위-레벨 정보 검색을 위한 다양한 요구에 만족할 수는 없다. 이러한 요구들을 지원하기 위하여, 이 시스템은 상위-레벨의 의미를 표현하기 위하여 " Intelligence" 개념을 도입해야 한다[2][3][7][12]. 이러한 영향으로, 인지 과학, 인공 지능, Semiotics 그리고 컴퓨터 비젼과 같은 다양한 연구들을 토대로 패턴분류 방법과 같은 처리 기술에 도움이 될 수 있다.

이 논문에서는 음성 정보를 사용하는 스포츠 비디오 중에서 농구 비디오로부터 음성 특징 추출 기법 및 중요한 이벤트를 추출하는 방법을 제안한다. 일반적으로, 칼라 히스토그램이나 움직임 같은 시각 정보나 음성 정보를 비디오 데이터로부터 의미적 정보를 검출하기 위하여 사용할 수 있다. 농구 경기는 축구나 테니스와 같은 실외 경기에 비해 음성 정보

가 풍부하기 때문에 중요한 이벤트들을 추출하는데 많은 도움을 받을 수 있다. 여기에서는 먼저, 간단한 샷 검출 알고리즘을 이용한 입력 비디오 데이터를 샷들로 분할하고, 각각의 샷들에 제안된 알고리즘을 적용한다.

2. 관련연구

최근에, 많은 연구들은 비디오 데이터로부터 의미 정보를 자동적으로 추출하기 위한 노력을 기울이고 있다. 특히, 음성적으로 풍부한 데이터를 가지고 있는 스포츠관련 비디오의 분류는 비디오 인덱싱 그리고 검색분야에서 매우 흥미있는 분야로 부각되어져 왔다.

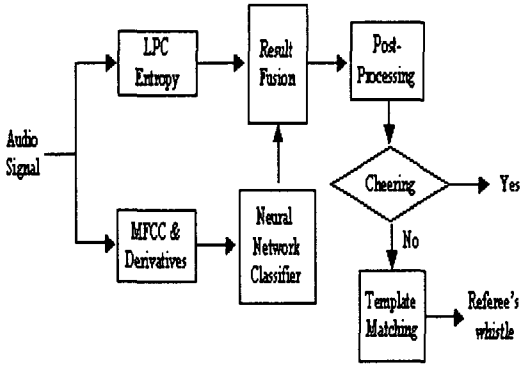
Babaguchi는 시각적, 음성적 그리고 텍스트를 포함한 정보 스트림들을 결합하여 이벤트에 기반한 비디오 인덱싱 방법을 제안했다[1]. Rui는 야구 TV 프로그램으로부터 주요한 장면들을 추출하는 시스템을 개발했다[9]. 그들은 음소 단위 특징들을 이용하여 방송의 마지막 부분을 추출하고, 중요한 이벤트를 찾도록 하고 있다. 이벤트를 표현하는 것 중에 아나운서의 방송을 사용하여 실험하였다. 그들은 특별한 이벤트를 검출하기 위하여 Template Matching 방법을 적용하였다. Zhou는 농구 경기의 비디오를 구조화하고 카테고리화하는 실험을 했다. 예를 들어, left-fast break, right-fast break, left dunk, right dunk, close-up shot 등, 규칙 기반의 분류에 의해 시각적 그리고 움직임의 특징 구별에 따른 분류에 관한 연구를 수행했다[11]. Nepal는 스포츠 장면의 수동적 관찰을 통한 개별적으로 학습되는 방법들을 조합하여 사용한 특징추출을 위한 자동화된 기술을 제안했다[8]. 그들은 농구 비디오에서 자동적으로 " 골" 에 관한 부분을 찾았다.

3. 오디오 특징을 이용한 주요 샷 검출

이 장에서는 먼저, 농구 비디오 데이터로부터 중요한 샷을 검출할 때, 환호 소리 그리고 심판의 호루라기 소리의 중요성에 대해서 설명한 후, 간단한 컷 검출 알고리즘에 대해 설명하고, 음성 정보로부터 기본적인 음성 추출에 대하여 설명한다.

환호 소리는 농구 게임을 분류하는데 매우 중요하다. 왜냐하면 환호 소리의 전, 후 상황에 덩크슛과 같은 흥미있는 이벤트가 일반적으로 야기되기 때문이다. 여기에서 그것들은 몇 개의 특징적인 특성들을 가지고 있다. 첫째로, 환호 소리는 일반적으로 꽤 오랜 시간동안 지속되고, 타임 영역안에서 스펙트럼의 변동이 없다. 둘째로, 환호 소리는 일반적으로 다른 소리에 비해 크다. 마지막으로, 스펙트럼안에 음원(Phoneme) 구조가 없다.

환호 소리 이외에도 심판의 호루라기 소리는 농구 경기에서 매우 중요하다. 그러므로, 오디오 처리의 마지막 단계에서 Template Matching을 통하여 심판의 호루라기 소리를 추출했다. 심판이 호루라기을 불었을 때, 게임의 상황이 어떻게 변하는지 살펴보았다. 첫째로, 만약 호루라기 소리가 샷의 뒷부분에서 검출되었다면, 그 다음 샷은 Close-up 샷이고, 게임은 멈출 것이다. 다음으로 호루라기 소리가 샷의 처음이나 중간 또는 평균적인 샷의 길이보다 훨씬 긴 샷에서 추출되었다면, 거기에는 다른 심판의 호루라기 소리가 포함되어 있거나 공격권의 변화가 될 가능성이 있다. 이러한 사전 영역 지식을 사용함으로써, 환호 소리를 포함한 영역을 제외한 그 외의 영역도 Template Matching을 통하여 검출이 가능하다.



[그림 1] 음성정보를 이용한 특별한 샷 검출

3.1 컷 검출

우리는 서로 다른 프레임들 사이에서 히스토그램을 이용한 간단한 컷의 검출에 대해서 설명한다. MPEG에서 Y Channel을 이용하여 DC 이미지를 추출하고, Y-Component의 프레임들간의 히스토그램의 차이를 얻는다. 여기에 사용되는 칼라 히스토그램의 비교식은 다음과 같다.

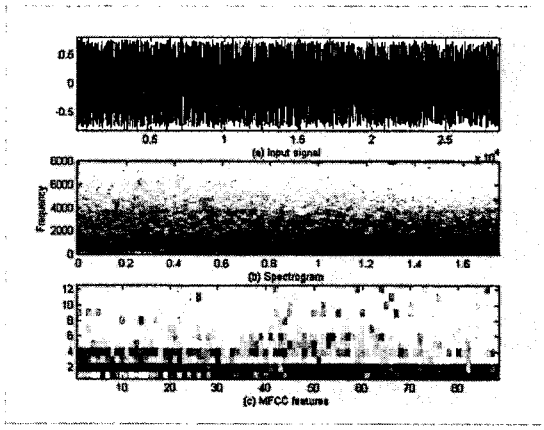
$$D(H_i, H_j) = \frac{1}{256}\sum_{1}^{256}\left|H_i(k) - H_j(k)\right| \qquad (1)$$

H_i 와 H_j 는 각 프레임 i 와 j 의 히스토그램을 나타내고, k 는 256 Gray Level중의 하나이다. 샷의 변화를 검출하기 위한 다음 단계에서, 우리는 주요한 프레임의 추출 방법에 기초한 각 적용 윈도우 범위의 복잡성에 따라 각각의 분할로부터 주요한 프레임을 추출하기 위해 자동적으로 시발점을 선택한다. 경험적으로 볼 때, 900개의 프레임으로 구성된 적용 윈도우 범위가 사용 되어진다.

3.2 오디오 정보로부터 사운드 검출

비디오 신호를 샷으로 분할한 후, 우리는 이벤트들에 대한 중요한 정보 중 환호 소리를 추출하기 위하여 음성 신호를 분류한다. 위의 사전 영역 지식들에 기초하여, 우리는 스펙트럼의 속성들을 특징짓기 위하여 MFCC(Mel-Frequency Cepstrum Coefficient) 특징들을, 그리고 긴 시간 윈도우에서 일시적인 특징들을 검출하기 위해 LPC(Linear Prediction Coefficient)의 엔트로피를 활용한다. MFCC 특징들은 음성인식에 사용되어져 온 기술이다. MFCC는 스펙트럼내에서 치우친 수치의 cepstral을 계산하고, 이것은 전형적인 cepstral 계수와 다르다. 특히, 이것은 높은 빈도의 분해 대신에 낮은 빈도의 분해를 탐지하는 기능을 지니고 있다. 이 스펙트럼의 수치는 음성 분할 스펙트럼의 음원 지역에 확대된다. 다양한 특징들은 음성 검출, 인식을 위해 개발 되어진 스펙트럼의 수치들에 의해 정의 되어진다.

기본적으로, MFCC는 이러한 특징들 중에 하나이다. 우리는 초당 50개의 프레임에서 12개의 MFCC 특징들을 검출하고 그것들 중에 10개의 특징들을 사용한다. 또한 인접한 프레임들 사이에서 스펙트럼 관련 5개의 MFCC 파생물이 더 검출된다.



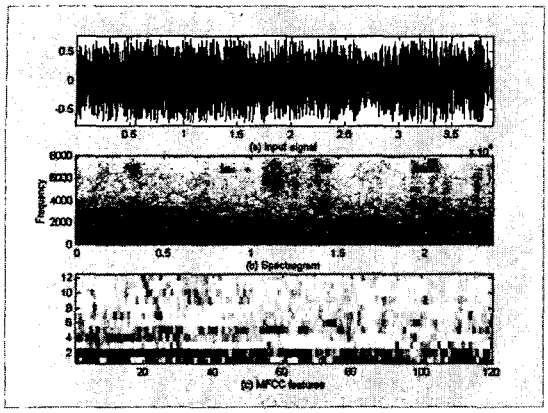
[그림 2] 환호 소리

[그림 2]는 대부분의 환호 소리의 오디오 클립의 타임 영역, 스펙트럼 사진 그리고 MFCC 특징들 사이에서 원래의 소리를 보여준다. 우리는 스펙트럼 사진의 낮은 빈도 지역에서 꽤 오랜 시간 동안 높은 음량을 찾을 수가 있다. [그림 3] 역시 대부분의 말소리의 오디오 클립의 타임영역, 스펙트럼 사진 그리고 MFCC 특징들에서 원래의 소리를 보여준다. 우리는 또한 스펙트럼의 특정 부분에 특정의 주파수가 집중하는 것을 볼 수 있다.

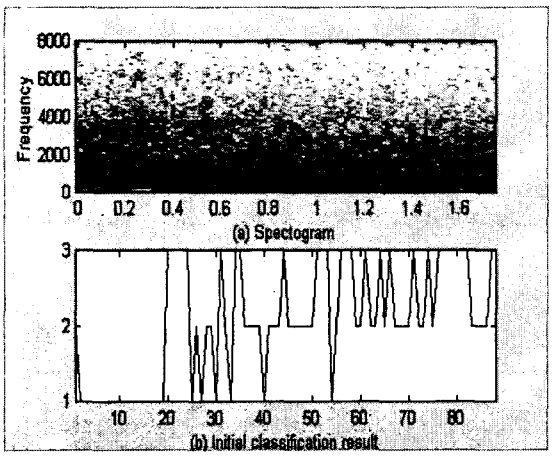
우리는 처음으로 MFCC 윈도우를 분류하기 위해 신경 조직에 기반한 분류법을 활용한다. 이 신경 조직 분류법은 고차원의 분류법의 한 종류이다. 입력, 출력 그리고 숨겨진 층의 노드 수는 기대했던 것처럼 각각 15, 5, 3이다. 우리의 연구에서, MFCC 윈도우는 환호소리, 말소리 그리고 다른 이외의 소리의 세 클래스로 분류된다. [그림 4]는 각 처리중인 샷의 Scoring 샷의 오디오 클립 분류결과 그리고 스펙트럼 사진을 보여준다.

실제적으로, 대부분의 환호 소리는 그림의 왼쪽 사이드에 위치해 있다. [그림 4]의 아래 그래프에서 숫자 1, 2, 3은 기대 되어지는 환호 소리, 말 소리 그리고 다른 그 이외의 소리들을 나타낸다. 비록 분

류된 결과에서 말 소리는 다른 소리들과 혼동되기는 하지만, 환호 소리는 거의 혼동되지 않음을 볼 수 있다.



[그림 3] 말 소리



[그림 4] 신경 조직 분류법에 의한 분류

MFCC 특징들은 오직 짧은 윈도우에서 스펙트럼 특징의 이점만을 가져다 준다. 그렇지만 우리는 환호 소리의 스펙트럼이 거의 일정할 때 긴 윈도우에서 소리의 경과를 고려하기 위하여 타임 영역 안에서 정지 신호를 정확히 측정한다. 그 속성은 말 소리 그리고 다른 이외의 소리에서는 요인이 아니다. 우리는 타임 영역 내에서 스펙트럼의 변화를 측정하기 위하여 엔트로피 특징을 이용한다. 엔트로피는 변화가 없는 신호의 측정 시 사용 가능하다. LPC 계수는 신호 스펙트럼의 다항식 근사값 중 하나이다. 그러므로, 이것은 오디오 신호에서 소음의 영향을 평가하는데 효과적이다. 우리는 각 LPC 요소의 엔트로피 평균으로 타임 영역 내에서 정지신호를 측정한다.

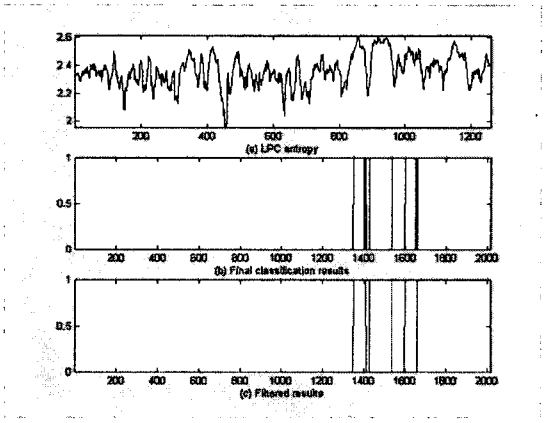
$$\text{Entropy} = -\frac{1}{D}\sum_{d=1}^{D}\sum_{n=1}^{w} P_{dn}\log P_{dn} \qquad (2)$$

P_{dn} 은 $\frac{|a(n,\ d)|^2}{\sum_{n=1}^{w}|a(n,\ d)|^2}$ 이고, a 는 LPC의 계수이고, w 는 타임 윈도우의 크기이다.

우리는 512개의 오디오 표본에서 각각의 LPC 계수가 추출되어진 한 윈도우에서 LPC 계수를 적당히 선택한다. 여기에서 D 는 LPC 순서이다. 이 방정식은 각각의 LPC 요소들이 긴 타임 윈도우 동안에 변하지 않을 때, 그것의 최대값을 포함한다. 우리는 환호 소리의 엔트로피 측정이 거의 최대값임을 볼 수 있다. LPC 계수의 변동으로 말 소리는 낮은 값을 갖는다.

마지막 단계에서, 우리는 이러한 두 개의 결과물을 결합한다. 모든 신호들은 true 또는 false로 이진화 되어지고, 이러한 이진 신호들은 조합되어진다. 그렇지만, 신경 조직 분류 의하여 발생 되어지는 소음 때문에 전처리 역시 이러한 과정에 꼭 필요하다. 우리의 시스템에서, 그 결합된 결과물은 1 아니면 0 이다. 만약 1 이라면 그 사운드 윈도우는 환호 소리이고, 그러지 않으면 0 이다. 소음 제거를 위하여 hysteresis 임계치를 사용하는데, 신호를 이진 값으로 하기 위해 두 개의 임계치를 사용한다. 첫 임계치는 초기 1 윈도우를 얻기 위한 것이고, 두 번째 임계치는 신호값이 제 2 임계치로 떨어질때까지 1 윈도우를 2 이웃 영역에 확장하게 하는데 이용한다. 여기서 0 윈도우를 좁힘으로써 갭을 줄이기 위해 후 처리 과정으로써 형태론적 필터를 이용한다.

[그림 5]는 오디오 표본의 결과값을 필터링하고 결합한 결과값과 마지막 분류 결과 그리고 LPC 엔트로피를 보여준다.



[그림 5] First example of audio clip (40 sec)

4. 성능 평가

실험 결과는 [표 1]에 나타나있다. [표 1]에서, ' false detection ratio' 는 환호 소리가 아닌 소리를 환호 소리나 호루라기 소리처럼 인식한 비율이다. ' False rejection ratio' 는 일반 소리 중에 잘못 환호 소리이거나 호루라기 소리라고 인식된 비율이다. 각각의 비율은 전체 처리된 윈도우에 대한 비율을 나타낸다.

Metrics / Sound	False detection ratio	False rejection ratio
Cheering Detection	230/950 ≅ 24 %	310/10050 ≅ 3.1 %
Referee's whistle Detection	119/519≅ 23 %	76/850 ≅ 9 %

[표 1] Experimental results

5. 결론

우리는 농구 비디오 데이터로부터 오디오 데이터를 이용하여 이벤트를 추출하기 위한 연구를 제안했다. 입력 비디오 데이터로 TV 농구 경기를 선택했고, 스포츠 비디오로부터 중요한 이벤트를 추출하기 위하여 시각적 분류보다 오디오 신호처리를 제안했다. 여기에서 중요한 이벤트는 Scoring shot, Dunk shot 등을 의미하고 또한 심판의 호루라기 소리 역시 농구 경기에서 중요하다고 가정한다. 입력 비디오 데이터를 샷으로 분할한 후, MFCC 특징들과 LPC 엔트로피를 결합해서 오디오 신호를 분할함으로써 환호 소리의 시작 지점을 추출한다. 그리고 Template Matching을 통해서 심판의 호루라기 소리를 검출한다. 우리는 또한 전 처리를 통하여 소음을 제거한다.

실험 결과는 말 소리 또는 농구 경기의 특별한 소리 같은 다른 소리들로부터 환호 소리를 분류는 매우 효과적이라는 것을 보여준다. 앞으로는 음성적 데이터에 대한 더욱더 자세한 분류를 위한 방법의 개발을 할 것이다. 또 다양한 소음의 환경에서 강력한 알고리즘을 개발을 필요로 한다. 본 기술이 다른 종류의 스포츠 비디오의 분류와 요약을 할 때 유용하게 사용될 것이라고 생각한다. 또한 연구 결과는 멀티미디어 검색 그리고 인덱싱의 응용분야에서 효과적으로 사용될 수 있을 것이다.

[참고문헌]

[1] Babaguchi, N., Kawai, Y., Kitahashi, T.: Event Based Video Indexing by Intermodal Collaboration, Proc. of the First International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM'99) in conjunction with ACM Multimedia Conference, (1999) 1-9

[2] Benitez, A. B. and Smith, J. R.: New Frontiers for intelligent Content-Based Retrieval, Proc. of the SPIE 2001 Conference on Storage and Retrieval for Media Databases (IS&T/SPIE-2001), Vol. 4315, Jan. (2001)

[3] Benitez, A. B., Smith, J.R., Chang, S. F.: MediaNet: A Multimedia Information Network for Knowledge Representation, Proc. of the SPIE 2000 Conference on Internet Multimedia Management Systems (IS&T/SPIE-2000), Vol. 4210, Nov. (2000)

[4] Chang, S.F. and Messerschmitt, D.G.: Manipulation and compositing of mc-dct compressed video, IEEE Journal on Selected Areas in Communications, vol.13, (1995) 1:1-11

[5] Chang, Y., Zeng, W., Kamel, I., Aonso, R.: Integrated Image and Speech Analysis for Content-Based Video Indexing, Proc. of the Third IEEE International Conference on Multimedia Computing and Systems, (1996) 306-313

[6] Kittler, J, Messer, K., Christmas, W. J., Obadia, B.L., Koubarroulis, D.: Generation of Semantic Cues for Sports Video Annotation, Proc. of the ICIP2001, Oct. (2001)

[7] Naphade, M. R. and Huang, T. S.: Semantic Filtering of Video Content, Proc. of the SPIE Conference on Storage and Retrieval for Media Databases, Jan. (2001)

[8] Nepal, S., Srinivasan, U., Reynolds, G.: Automatic Detection of 'Goal' Segments in Basketball Videos, ACM Multimedia Sept. (2001)

[9] Rui, Y. and Gupta, A., Acero, A.: Automatically Extracting Highlights for TV Baseball Programs, Proc. of the ACM Multimedia, Oct. (2000) 105-115

[10] Zhong, D. and Chang, S.F.: Structure Analysis of Sports video Using Domain Models, IEEE Conference on Multimedia and Exhibition, Aug. (2001)

[11] Zhou, W, Vellaikal, A., Kuo, C.-C. J.: Rule-based Video Classification System for Basketball Video Indexing, ACM Multimedia Oct. (2000)

[12] Zhuang, Y., Rui, Y., Huang, T.S., Mehrotra, S.: Applying Semantic Association to Support Content-Based Video Retrieval, Proc. of IEEE VLBV98 Workshop, (1998) 45-48

[13] Xu, P., Xie, L., Chang, S.F.: Algorithms and Systems for Segmentation and Structure Analysisin Soccer Video, IEEE International Conference on Multimedia and Expo, Aug. (2001)