한국음향학회지 제28권 제5호 pp. 441~446 (2009)

오디오 정보를 이용한 골프 동영상 자동 색인 알고리즘

Automatic Indexing Algorithm of Golf Video Using Audio Information

김 형 국*
(Hyoung-Gook Kim*)

*광운대학교 전파공학과
(접수일자: 2009년 4월 6일; 채택일자: 2009년 6월 23일)

본 논문에서는 오디오 정보 분석을 이용하여 골프 동영상을 자동 색인하는 알고리즘을 제안한다. 제안하는 알고리즘에서는 입력되는 골프 동영상을 비디오 신호와 오디오 신호로 분리한 후에, 연속적인 오디오 스트림을 Adaboost Cascade 분류방식을 통하여 스튜디오 환경에서의 아나운서의 음성구간, 선수이름이 TV 화면에 소개 될 때 수반되는 음악구간, 선수들의 플레이에 따라 반응하는 관중들의 박수 및 환호성 소리구간, 필드에서의 레포터의 음성구간, 바다나 바람 등의 필드환경 잡음 사운드구간 등의 5가지 구간으로 분류한다. 그리고 드라이브 샷, 아이런 샷과 퍼팅 샷 시에 발생하는 스윙 사운드는 onset 검출과 변조스펙트럼 검증 방법을 통해 검출되며, 관객의 박수 소리 구간과 결합하여 액션 및 하이라이트를 효율적으로 색인할 수 있게 한다. 제안된 알고리즘은 오디오 신호의 간단한 연산을 통해 의미를 지니고 있는 기본구간들을 검출하기 때문에 골프 동영상에서 사용자가 원하는 부분을 빠르게 브라우징하는 임베이디드 시스템에 적용가능하다.
핵심용어: Adaboost Cascade 분류방식, 오디오 분류, Onset 검출
투고분야: 뉴 미디어 분야 (13)

This paper proposes an automatic indexing algorithm of golf video using audio information. In the proposed algorithm, the input audio stream is demultiplexed into the stream of video and audio. By means of Adaboost-cascade classifier, the continuous audio stream is classified into announcer's speech segment recorded in studio, music segment accompanied with players' names on TV screen, reaction segment of audience according to the play, reporter's speech segment with field background, filed noise segment like wind or waves. And golf swing sound including drive shot, iron shot, and putting shot is detected by the method of impulse onset detection and modulation spectrum verification. The detected swing and applause are used effectively to index action or highlight unit. Compared with video based semantic analysis, main advantage of the proposed system is its small computation requirement so that it facilitates to apply the technology to embedded consumer electronic devices for fast browsing.
Keywords: Adaboost cascade classifier, Audio classification, Onset detection
ASK subject classification: New Media (13)

I. 서 론

최근 디지털 TV와 멀티미디어 핸드폰 같은 디바이스들은 대용량의 비디오가 존재하는 인터넷에 접속 할 수 있을 뿐만 아니라, Personal Video Recorder를 이용하여 방송되는 프로그램을 녹화할 수 있는 기능을 갖고 있다. 이로 인하여 디바이스를 통해 저장되는 대용량의 비디오를 사용자가 효과적으로 브라우징하기 위한 내용기반 컨텐츠 분석 및 자동색인 기술이 급속히 발달되고 있다. 특히 오랜 시간동안 지속되는 스포츠 동영상에서 발생한 세부사항들을 빠르게 브라우징하기 위한 실시간 하이라이트 검출 및 색인 기술이 필요 되고 있으며, 이러한 기능을 임베디드 기기에 적용하기 위해서 메모리의 사용량과 연산량을 줄이는 효과적인 방법이 연구되고 있다.

스포츠 경기에서 중요구간 검출은 크게 오디오 기반 검출 방법 [1], 비디오 기반 검출 방법 [2], 자막 기반 검출 방법 [3], 멀티모달 기반 검출 방법 [4] 등으로 나뉘어 연구되어 왔다. 비디오 기반 검출 방법은 의미 있는 중요구간을 검출하는데 있어서 많은 연산량이 발생되어 임베디드 시스템에 적용하기에는 많은 어려움이 따르기 때문에 연산량이 작은 오디오 정보만을 이용하여 중요구간을 검출하거나, 검출된 오디오 중요구간과 연산량이 작은 비디오 기반 장면전환 검출 알고리즘을 결합하는 방식을 이

책임저자: 김 형 국 (hkim@kw.ac.kr)
서울시 노원구 월계동 447-1 광운대학교 전파공학과
(전화: 02-940-5574; 팩스: 02-913-0429)

용하여 중요구간을 검출하는 방식을 선호하고 있다.

본 논문에서는 축구나 야구경기보다 배경잡음이 상대적으로 조용한 골프 동영상에서 오디오 신호 분석을 통해 사용자가 찾고자 하는 구간 및 사용자가 선호하는 하이라이트와 같은 액션구간이 포함된 의미적 기본구간들을 검출하는 알고리즘을 제안한다.

전형적인 스포츠 채널의 골프 동영상 프로그램의 구성을 살펴보면 크게 스튜디오 구간, 음악 구간, 필드 구간 등으로 나뉘고 편집에 의해 특정한 형식을 유지한다는 사실을 알 수 있다. 이에 따라 오디오 신호에서 추출할 수 있는 5개의 의미적인 기본 구간을 다음과 같이 정의할 수 있다: 스튜디오 음성 (STD: 스튜디오에서 녹음된 음성), 음악 (SOM: 음악이나 필드와 아나운서의 음성과 함께 녹음된 음악), 관객들의 호응 (APP: 관객들의 호응과 박수소리), 아나운서와 해설의 중계 음성 (SPC: 필드배경에서 아나운서와 해설의 중계음성), 바다나 바람 등의 환경잡음 (SIL: 필드 배경). 이러한 5개의 기본구간에서 스튜디오 구간은 스튜디오 환경에서 아나운서가 경기를 소개하는 구간, 음악 구간은 TV 화면에 소개되는 출전 경기 선수의 이름이 나타나는 구간, 배경잡음이 섞여 있는 음성구간은 필드에서의 레포터의 해설구간, 선수들의 플레이에 따라 반응하는 관중들의 호응구간 등으로 각각 색인될 수 있다. 특히, 관중들의 호응구간은 드라이브 샷, 아이런 샷과 퍼팅 샷 시에 발생하는 스윙구간과 결합되어 골프 동영상에서의 하이라이트를 나타내는 액션구간으로 색인되고, 검출된 각 기본구간을 이용하여 사용자는 골프 동영상에서 자신이 찾고자 하는 부분을 빠르게 브라우징 할 수 있게 된다.

본 논문의 구성은 다음과 같다. 제 II장에서는 전체적인 시스템의 구성도와 스윙 검출방법 및 의미기반 기본구간 분할 방법을 기술한다. 제 III장에서는 제안된 시스템의 실험결과를 분석 및 고찰하며 제 IV장에서 결론과 향후 연구 방향을 기술한다.

II. 시스템 개요도

그림 1은 본 연구에서 실행한 전체적인 알고리즘의 구조를 나타낸다.

입력되는 골프 동영상 스트림은 신호분배기에 의해 영상신호와 오디오 신호로 각각 분리된다. 분리된 오디오 신호의 AC-3 Audio Encoder에 의해 추출된 MDCT (Modified Discrete Cosine Transform) 계수를 기반으로 오디오 특징값을 추출한다. MDCT는 부분 압축영역에서 추출되는 오디오 계수로서 MDCT 계수를 이용하여 사용하고자 하는 오디오 특징값을 추출함으로써 비 압축영역으로 전환하여 오디오 특징값을 추출하는 신호처리 시간을 단축할 수 있게 된다. 추출된 오디오 특징값을 이용하여 연속된 오디오 신호는 STD, SOM, APP, SPC, SIL등의 5가지 기본구간으로 자동 분할된다.

신호 에너지의 갑작스런 증가로 나타나는 급격하게 짧게 발생하는 소리의 시작을 나타내는 Onset 검출과 변조 스펙트럼 방법을 이용하여 MDCT 계수로부터 스윙으로 인해 발생하는 임팩트 소리를 검출한다. 오디오 분류를 통해 검출된 APP 구간과 스윙 검출구간 (SWN)이 결합되어 중요구간인 액션구간 (ACT)을 구성하며 하이라이트의 흥미로움 정도에 따른 최종 결과는 랭킹화되어 자동적으로 리플레이를 위한 하이라이트의 집합을 이루게 된다. 각 검출된 STD, SOM, SPC, ACT 구간의 처음 시작점과 끝점은 골프동영상과 동기화되어 색인된다.

2.1. 오디오 분류를 통한 기본구간 분할

의미기반 기본구간 분할은 MDCT 기반 오디오 특징 추출, Adaboost cascade 학습모델을 이용한 구간별 오디오 분류 및 분할, 기본구간 조정 분할의 3 단계에 의해 수행된다.

오디오 특징은 AC-3 audio encoder에 의해 추출된 MDCT 계수로부터 1초 길이의 오디오 세그멘트 별로 1차의 RMS Energy, 23차의 Normalized Logarithmic MDCT (NLMDCT), 4차의 Delta NLMDCT, 2차의 SE_{23}를 결합하여 하나의 특징벡터를 추출하여 사용한다. 특히, Delta NLMDCT와 SE_{23}는 아나운서의 목소리가 주변잡음환경

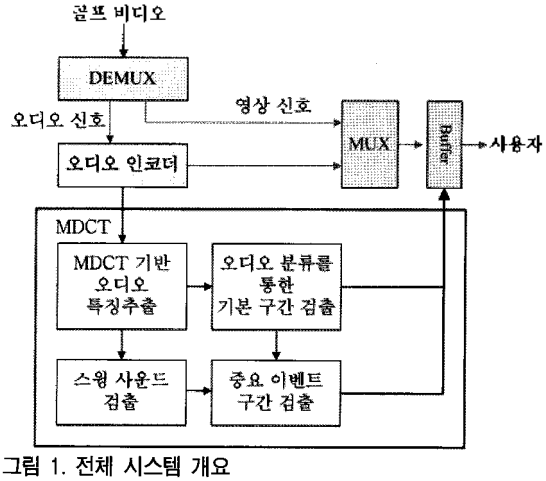

그림 1. 전체 시스템 개요
Fig. 1. System overview.

에 섞여있는 골프 필드의 주변 잡음환경소음으로부터 아나운서의 음성을 구분할 수 있게 한다.

SE_{xy}는 MDCT 계수를 0-630 hz, 630-1720 hz, 1720-4400 hz, 그리고 4400 hz 이상의 범위로 분할된 4개의 주파수 밴드의 두 번째와 세 번째 밴드로부터 추출된 특징값이다. RMS energy, NLMDCT, Delta NLMDCT의 상세한 특징 추출과정은 [5]에 나타나 있으며 특징 값의 차수와 계수는 실험을 통해 선택되었다.

구간별 오디오 분류는 Support Vector Machine (SVM) 기반의 AdaBoost Cascade 분류구조 [6]를 이용하여 오디오 신호를 STD, SOM, APP, SPC, SIL 등의 5개의 정의된 구간으로 각각 자동 분류한다. 본 논문에서는 Linear SVM 을 이용하여 학습된 모델을 생성하였다. 커널을 사용하는 형태의 다른 SVM의 경우에는 커널 적용 및 서포트 벡터와의 계산 때문에 연산량이 많아 실시간 검출이 불가능하다. Linear SVM은 식 (1)와 같이 검출을 원하는 클래스가 Hyperplane으로 표현이 되고, 입력되는 특징벡터(X)를 적용하여 $f(X)$ 값이 0보다 큰 값이 나오면 해당 클래스로 분류되게 된다. 그러므로 단순히 사용되는 특징벡터 차원만큼의 곱셈과 덧셈의 계산량 정도만 필요하게 되어 빠른 시간 내에 검출이 가능하다.

$$f(X) = W^T X + b;\ b: bias,\ X = \{x_1, \ldots, x_n\} \tag{1}$$

Cascade 각 층에서는 분류속도와 성능을 고려하여 정해진 하나의 기본구간 클래스에 대해 모델을 생성하고 다른 나머지 클래스로부터 또 다른 하나의 모델을 생성하여 입력된 1초 단위의 오디오 클립을 하나의 클래스로 분류하게 된다. 본 논문에서는 5개의 기본구간을 사용하기 때문에 5개의 층으로 이루어진 cascade 분류구조를 통해 입력된 오디오 클립을 5개의 기본 구간 중에 하나로 분류한다.

다음 단계는 1초단위로 결정된 각 구간들의 연속성을 고려하여 시간에 따라 의미적인 세그먼트들의 단위로 기본 구간을 조정한다. 일반적으로 골프 동영상 프로그램의 구조에서 STD, SOM, APP는 SPC와 SIL보다 중요한 역할을 한다. STD와 SOM은 편집에서 추가되고 STD는 언제나 프로그램의 시작과 끝에 위치하며 때때로 프로그램의 중간에 STD의 한 두 세그먼트가 추가 될 수 있다. 스튜디오에서 촬영된 영상과 함께 있는 STD의 경우는 대부분 오랜 시간동안 지속되고 SOM의 경우도 편집자가 경기 결과나 중요한 문장을 보여줄 때 사용되며 오랜 시간동안 지속된다. 이렇게 편집에 의해서 추가되는 STD, SOM 부분은 불규칙적으로 자주 존재하는 SPC와 SIL과 달리 규칙적으로 존재한다. 그리고 대부분의 경우에 스윙 소리 다음에 존재하는 APP는 하이라이트 검출에 가장 중요한 역할을 한다. 이를 기반으로 의미기반 기본구간 조정은 식 (2)의 존재 비율 값 $R(*)$에 의해 수행된다.

$$R(*) = \frac{B(*)}{W(*)} \tag{2}$$

위의 식에서 *는 기본 구간 중의 하나의 STD, SOM, APP를 나타내고, $W(*)$는 주어진 기본 구간이 존재하는 전체적인 지속기간, $B(*)$는 $W(*)$ 내의 주어진 기본구간의 지속기간을 나타낸다. 이러한 방법은 적절하지 못한 시간 동안 지속되거나 불가능한 장소에서 발생하는 의미없는 기본구간을 제거할 수 있게 한다. 또한, 의미 없는 기본 구간을 합치면서 적절한 시간과 위치에 있는 하나의 완전한 기본구간 중의 하나로 결합하고, 하나의 기본구간을 시간과 위치에 어울리는 형태의 기본구간으로 바꿔준다.

기본구간의 길이 비율인 R (STD), R (SOM), R (APP)는 W (STD)가 30초, W (SOM)이 15초, W (APP)가 3초일 때 각각 계산된다. 3개의 비율이 계산되면 $W(*)$를 1초 구간 이동하여 각각의 STD, SOM, APP의 순서대로 의미기반 기본구간을 아래에 나타난 (Step1)-(Step3) 과정을 통해 찾게 된다.

(Step1) 먼저 R (STD)값을 계산한 후에 '의미기반 구간 찾기 알고리즘'에 의해서 STD의 위치와 지속 시간을 결정하는 R (STD)가 결정된다. 나머지 STD로 분할된 구간들은 SPC로 간주된다.

(Step2) Step1 수행이후의 나머지 오디오 스트림에서 R (SOM)값을 계산한 후에 '의미기반 구간 찾기 알고리즘'에 의해서 SOM의 위치와 지속 시간을 결정하는 R (SOM)가 결정된다. 다른 SOM 구간들은 SPC로 간주된다.

(Step3) Step2 수행이후의 나머지 오디오 스트림에서 R (APP)값을 계산한 후에 '의미기반 구간 찾기 알고리즘'에 의해 APP의 위치와 지속 시간을 결정하는 R (APP)가 결정되고 다른 APP들은 SIL로 간주된다. 다음의 그림 2는 '의미기반 구간 찾기 알고리즘'을 나타낸다.

Oth(*)와 Nth(*)는 각각 기본구간이 존재하는 영역을 찾기 위한 점유비율 임계값 및 기본구간의 시작점과 끝점을 찾기 위한 임계값으로서 본 논문에서 구현된 STD, SOM, APP의 점임계값은 실험을 통해서 획득된 Oth (STD) =0.8, Nth (STD)=0, Oth (SOM)=0.5, Nth (SOM)=0.2, Oth (APP)=0.5, Nth (APP)=0을 사용하였다.

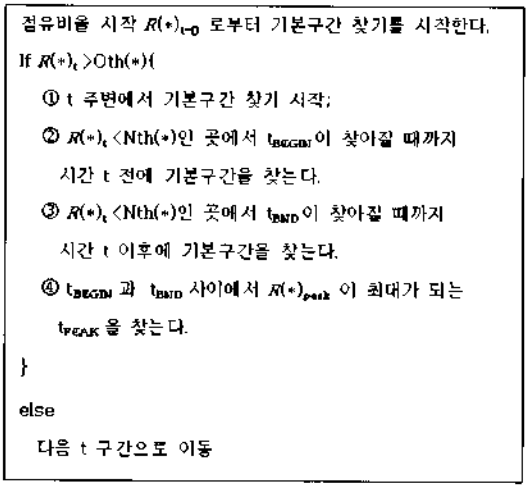

그림 2. 의미 기반의 검색 알고리즘
Fig. 2. Flowchart of the semantic search algorithm.

긴 시간동안 지속되는 골프 프로그램에서 하이라이트가 포함되는 액션구간은 APP의 지속시간을 기반으로 Exciting Hit (EH), Good Hit (GH), Ordinary Hit (OH) 등의 세 가지로 구분된다. 즉, APP의 지속기간이 8초 이상 일 경우는 EH, APP의 지속기간이 3초에서 8초 미만일 경우는 GH, APP의 지속기간이 3초 미만일 경우는 OH로 구분한다.

2.2. 스윙 검출

스윙 사운드는 신호 에너지의 갑작스런 증가로 나타나는 급격하게 짧게 발생하는 소리로서 5개의 기본구간 들 중에서 골프 플래이어의 스윙이 존재하는 APP, SPC, SIL 등의 세 구간에 대해서 Impulsive Onset Detection 방식을 이용하여 골프 스윙을 검출하게 된다.

본 논문에서는 onset 검출과 변조스펙트럼 방법을 이용하여 스윙으로 인해 발생하는 임펙트 소리를 검출한다. 이러한 과정은 두 단계로 수행된다. 첫 번째 단계는 AC-3 Audio Encoder에서 검출된 MDCT 계수로부터 Banked-Power Increasing Ratio (*IRb*), Logarithmic Power (log_{en}), Delta Logarithmic Power ($\triangle log_{en}$)를 다음과 같이 구한다.

$$IRb = \frac{2}{N}\sum_{n=0}^{\frac{N}{2}-1} s(\log|MDCT_2(n)|^2 - \log|MDCT_1(n)|^2 - th) \tag{3}$$

여기서 $MDCT_1$과 $MDCT_2$는 N샘플 창의 $MDCT$ 계수의 처음과 끝의 반을 각각 나타내며, IRb는 갑작스런 신호 에너지의 민감한 증가변화 시작점인 onset를 효과적으로 추적한다. $MDCT_1$과 $MDCT_2$사이의 변화를 구분하기 위해 실험을 통해 설정된 th는 0보다 큰 임계값을 의미하고, $s(t)$는 단계함수로서 다음과 같이 정의되어 있다.

$$s(t) = \begin{cases} 1, & t > 0 \\ 0, & otherwise \end{cases} \tag{4}$$

$$\log_{en} = \log\left(\sum_{n=0}^{N-1} MDCT^2(n)\right) \tag{5}$$

$$\Delta\log_{en} = \log\left(\frac{\sum_{n=N/2}^{N-1} MDCT^2(n)}{\sum_{n=0}^{N/2-1} MDCT^2(n)}\right) \tag{6}$$

추출된 세 가지의 IRb, log_{en}, $\triangle log_{en}$에 식 (7)과 같은 Onset Filter $h(n)$를 적용하여 세 영역에서 모두 추출되는 급격히 증가하는 상승부분의 시작을 Onset로 검출한다.

$$h(n) = (1-e^{1/\tau_e})e^{n/\tau_e} - (1-e^{1/\tau_i})e^{n/\tau_i},\ n = 1,2,3,... \tag{7}$$

Onset Filter는 2차의 필터로서 긍정적인 결과일 때 활성화 되고 부정적인 결과일 때 억제되는 두 개의 1차 지연-추가 (delay-and-add) 필터로 구성되어 있다. 식 (7)에서 사용된 τ_e와 τ_i는 1차 성분들의 시정수이며 $\tau_e \leq \tau_i$ 관계를 유지한다. 이러한 필터는 입력신호의 느린 변화보다 급격히 증가하거나 감소하는 빠른 변화에 더욱 민감하므로 스윙과 같은 임펄스를 검출해내기에 적합하다. 특정한 임계값 이상의 트레이스들이 onset 후보들로 지정되고, onset 필터의 동일한 간격에 위치하는 후보들은 하나의 onset로 결합된다.

스윙 검출의 두 번째 단계는 검출된 onset 후보들을 대상으로 수행된다. 스윙소리와 같은 갑자기 발생하는 이벤트는 charge-release 구조 특성을 갖고 있다. Charge 부분은 onset로부터 power의 최고점까지의 간격을 포함하며 외부의 힘으로부터 얻어지는 갑작스러운 에너지로 발진하는 과정을 반영하는 반면에, release 부분은 소음레벨까지의 떨어지는 과정을 포함한다. Charge-release 부분은 평균적인 두 세그먼트로 나뉘고, 각 세그먼트에 대해 12 ms 단위의 윈도우 별로 6 ms 중첩과정을 겹쳐 웨이브랫 변환 (wavelet transform)을 수행한다. 스펙트럼들은 주파수 축에서 대수적인 4개의 octave bank 형태로 샘플링 되고, 각 세그먼트의 스펙트럼 진폭은 시간에 따라 4개의 octave bank에 대한 4개의 계수그룹으로 분

할되어 평균화 된다. 하나의 갑작스러운 이벤트의 4개의 세그먼트들은 4×4 계수 행렬로 나타나고 행렬의 최고값으로 나뉘어 일반화 시킨 후 변조스펙트럼 스케일 특징 행렬 값이라 정의한다.

최종적으로 추출된 변조스펙트럼 스케일 특징 행렬값에 통계적인 방법을 적용하여 스윙 이벤트를 검출한다. 본 논문에서는 통계적인 방법 적용에 있어서 H1 (스윙으로 발생한 이벤트)와 H0 (스윙으로 발생하지 않은 이벤트)의 두 가지의 이벤트 발생을 가정하고, H1, H0의 파라미터를 추정하기 위해 Standard Gaussian Mixtures를 사용하였다. H1 가정의 모델은 학습 데이터 내의 스윙 이벤트의 특징 값을 바탕으로 학습되고 H0 가정의 모델은 스윙 이벤트 검출에 실패한 특징 값으로 생성된다. 스윙 발생 결정과정은 onset 후보로부터 GMM (Gaussian Mixture Models)에 의해 계산된 H1과 H0의 우도비의 차이와 사전에 정의된 임계값과 비교하여 이벤트 발생을 1과 0을 통해 스윙 검출을 결정하게 된다.

검출된 스윙 구간과 중요 이벤트와 관련되는 관중들의 APP에 의해 추출된 Exciting Hit, Good Hit, Ordinary Hit 등의 세 가지의 Action Unit를 결합하여 하이라이트 요약본을 자동으로 색인한다. 때때로 검출된 APP의 위치를 보면 스윙구간에 근접하지 않게 발생하는 2가지 경우가 있다. 첫 번째 경우는 스윙이 너무 약해서 검출되지 않는 경우이고, 두 번째는 스윙과 APP가 너무 근접하게 발생하여 APP가 스윙을 포함하는 경우이다. 두 경우 모두 다 APP가 액션 구간을 포함한다.

III. 실험 결과 및 분석

본 논문에서 제안된 시스템의 성능을 평가하기 위해 스포츠 채널에서 녹화한 총 64시간 이상의 40개 골프 동영상들을 사용하였다. 그 중에서 20개의 동영상이 학습에 사용되었고 나머지 20개의 동영상들을 대상으로 성능을 평가하였다. 실험을 위해 수집된 16 kHz로 녹화된 골프동영상의 오디오 스트림은 AC-3 오디오 인코더를 위해 44.1 kHz 샘플링 신호로 변환하여 STD, SOM, SPC, APP, SIL의 5개의 기본 구간으로 라벨화 되었으며, 기본 구간을 기반으로 EH, GH, OH 구간에 대해서도 라벨화를 수행하였다. 본 논문에서는 구간별 오디오 분류, 기본구간 조정 분할, 하이라이트와 같은 액션구간 검출을 통해 제안한 알고리즘의 성능을 측정하였으며, 시스템의 성능 평가를 위해서는 널리 알려진 precision과 recall 방식을 사용하였다. 실험결과는 다음과 같다.

〈실험1〉 구간별 오디오 분류

연속된 오디오 신호를 1초 단위의 오디오 클립으로 나누어 STD, SOM, APP, SPC, SIL 등의 5개의 정의된 구간으로 분류하기 위해 본 논문에서 사용된 SVM 기반의 cascade 오디오 분류구조를 GMM 기반의 cascade 구조와 [1]에서 사용된 방식과 비교하여 분류성능을 측정하였다. 10개의 골프 동영상에 포함된 연속적인 오디오 신호를 1초 단위의 오디오 클립으로 나누어 실험을 수행한 결과는 표 1과 같다.

표 1. 구간별 오디오 분류 정확성
Table 1. Precision of audio classification.

분류방법	SVM-Cascade	GMM-Cascade	[3]에 의한 방식
분류 정확도	92.6%	88.5%	82.3%

표 1에 나타난 바와 같이 Adaboost cascade 학습기반을 이용한 SVM과 GMM 방식이 [1]의 GMM만을 사용한 방식보다 분류정확도가 높았으며, 같은 cascade 구조에서는 SVM 방식이 GMM 방식보다 분류성능이 4% 높음을 알 수 있었다.

〈실험2〉 기본구간 조정 분할

1초 단위의 오디오 클립을 분류한 결과는 다시 STD, SOM, APP, SPC, SIL의 5개의 의미구간으로 조정 분할된다. 조정 분할에 있어서 SOM은 대부분 10초 이상 지속되기 때문에 지속시간에 의존하여 측정되고, 명확한 시작과 끝점을 보유하고 있어 측정이 용이하다. APP는 1초에서 15초 내에 지속시간이 복합되어 있으며, 기본 구간의 전후에 자주 위치하고, 특정한 규칙 없이 산발적으로 분포하여 시작과 끝점이 명확하지 않으므로 박수 회수를 측정하였다. 표 2는 의미기반 기본구간 조정분할의 성능을 나타낸다.

표 2. 기본구간 조정분할 성능평가
Table 2. Performance of semantic segmentation.

분류구간	Recall (%)	Precision (%)
STD	100%	96.5%
SOM	88.6%	94.3%
APP	96.8%	98.2%

결과를 살펴보면 STD는 매우 높은 검출 성능을 나타냈

으며 SOM은 11.4%의 오류를 보였고, 반면에 하이라이트 검출에 많은 영향을 미치는 APP의 검출성능은 우수하게 나타남을 알 수 있다.

〈실험3〉 액션구간 검출

액션구간은 SWN과 APP 검출에 영향을 많이 받는다. 그러므로 표 3에서는 액션구간 검출과 함께 SWN 검출성능을 함께 표기하였다.

표 3. 액션구간 검출 성능
Table 3. Performance of action unit detection.

분류구간	PER* (%)	Recall (%)	Precision (%)
EH	25%	100%	97.1%
GH	33%	98.2%	98.4%
OH	42%	93.8%	62.7%
SWN		94.5%	61.5%

PER*은 전체적인 액션 구간에서 각 종류의 분포도를 나타내며 확률의 결과가 1이 넘는 이유는 3 가지의 액션 구간이 중복되어 정의되었기 때문이다.

표 3에서 보듯이 GH와 EH의 검출 성능은 매우 우수하다. 단지 OH의 검출에서 6.2%의 검색 실패와 37.3%의 오검출율이 에러로서 발생하였다. 대부분의 SWN이 검출되었으나 절반정도가 잘못 검출된 이유를 살펴보면 아나운서의 감탄사나 음악에서의 강한비트, 갑자기 발생한 박수소리, 클럽이 바닥에 떨어지는 소리, 골프공이 바닥에 떨어지는 소리 녹화하는 채널의 클릭 소리 등의 갑자기 발생하는 강한 사운드가 SWN 검출에 많은 영향을 미쳤다고 판단된다. 그리하여 향후 연구방향은 SWN의 검출 성능을 높이는 방법에 집중할 계획이다. 그러나 다행스럽게도 하이라이트 검출에 많은 영향을 미치는 APP의 검출성능은 우수하게 나타났으며 OH보다 보다 GH와 EH의 검색 성능이 우수함을 확인 할 수 있었다. 그렇기 때문에 사용자가 good hit인 GH와 exciting hit인 EH를 시청하고자 한다면 충분히 만족할 만한 결과를 얻을 수 있을 것이다.

IV. 결 론

본 논문에서는 오디오 정보를 기반으로 골프 동영상을 여러 개의 의미를 지니고 있는 기본구간 별로 자동 색인하는 알고리즘을 제안한다. 골프 프로그램의 구조는 스튜디오 음성, 필드 음성, 음악, 관객의 박수, 배경 필드 등의 여러 개의 의미를 지니고 있는 기본 구간으로 구성되어 있다. 이 중에서 스윙과 관객의 박수 소리는 하나의 액션 구간을 구성한다. 스윙은 갑작스러운 onset 검출과 변조스펙트럼 검증 과정의 평균에 의해서 검출되며 나머지는 기본 구간의 통계적인 모델에 의해서 인식된다. 최종적으로 하이라이트 검출에 많은 영향을 미치는 관객의 박수소리의 검출 성능이 높게 나타나면서 전체적으로 신뢰할 만한 액션구간 검출 결과를 얻을 수 있었다. 전체적으로 good hit와 exciting hit은 98% 이상의 precision과 recall의 우수한 성능을 얻을 수 있었다. 또한 제안된 시스템은 오디오 정보만을 사용하여 하이라이트를 검출하기 때문에 연산량을 크게 감소 시켜 향후 임베디드 가전제품에 적용될 수 있다.

참 고 문 헌

1. I. Otsuka, R. Radhakrishnan, M. Siracusa, A. Divakaran, and H. Mishima, "An enhanced video summarization system using audio features for a personal video recorder," *IEEE Transactions on Consumer Electronics*, vol. 52, no. 1, pp. 168- 172, 2006.
2. A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Transactions on Image Processing*, vol. 12, no. 7, pp. 796-807, 2003.
3. D. Zhang, and S. F. Chang, "Event detection in baseball video using superimposed caption recognition," *Proc. of 10th ACM international Conf. on Multimedia*, Juan-les-Pins, France, pp. 315-318, Dec. 2002.
4. D. A. Sadlier, and N. E. O'Connor, "Event detection in field sports video using audio-visual features and a support vector machine," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 10, pp. 1225-1233, 2005.
5. H.-G. Kim, J. Jeong, J.-H. Kim, and J. Kim, "Real-time highlight detection in baseball video for TVs with time-shift function," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 2, pp. 831-838, 2008.
6. S. Ravindran, D. V. Anderson, and J. Rehg, "Cascade jump support vector machine classifiers," *IEEE Workshop on Machine Learning for Signal Processing*, pp. 135-139, Sep. 2005.

저자 약력

•김 형 국 (Hyoung-Gook Kim)

The Journal of the Acoustical society of Korea, Vol.26, No.2E, 2007