한국음향학회지 제28권 제2호 pp. 112~118 (2009)

뮤직비디오 브라우징을 위한 중요 구간 검출 알고리즘

Salient Region Detection Algorithm for Music Video Browsing

김 형 국*, 신 동*
(Hyoung-Gook Kim*, Dong Shin*)

*광운대학교 전파공학과

(접수일자: 2008년 10월 10일; 수정일자: 2008년 12월 3일; 채택일자: 2009년 2월 11일)

본 논문은 모바일 단말기, Digital Video Recorder (DVR) 등에 적용할 수 있는 뮤직비디오 브라우징 시스템을 위한 실시간 중요 구간 검출 알고리즘을 제안한다. 입력된 뮤직비디오는 음악 신호와 영상 신호로 분리되어 음악 신호에서는 에너지기반의 음악 특징값 최고점기반의 구조분석을 통해 음악의 후렴 구간을 포함하는 음악 하이라이트 구간을 검출하고, SVM AdaBoost 학습방식에서 생성된 모델을 이용해 음악신호를 분위기별로 자동 분류한다. 음악신호로부터 검출된 음악 하이라이트 구간과 영상신호로부터 검출된 가수, 주인공의 얼굴이 나오는 영상장면을 결합하여 최종적으로 중요구간이 결정된다. 제안된 방식을 통해 사용자는 모바일 단말기나 DVR에 저장되어 있는 다양한 뮤직비디오들을 분위기별로 선택한 후에 뮤직비디오의 30초 내외의 중요구간을 빠르게 브라우징하여 자신이 원하는 뮤직비디오를 선택할 수 있게 된다. 제안된 알고리즘의 성능을 측정하기 위해 200개의 뮤직비디오를 정해진 수동 뮤직비디오 구간과 비교하여 MOS 테스트를 실행한 결과 제안된 방식에서 검출된 중요 구간이 수동으로 정해진 구간보다 사용자 만족도 측면에서 우수한 결과를 나타내었다.

핵심용어: 중요 구간 검출, 하이라이트 검출, 음악 분위기 분류, 얼굴 검출, 뮤직비디오 브라우징
투고분야: 음악음향 및 음향심리 분야 (8)

This paper proposes a rapid detection algorithm of a salient region for music video browsing system, which can be applied to mobile device and digital video recorder (DVR). The input music video is decomposed into the music and video tracks. For the music track, the music highlight including musical chorus is detected based on structure analysis using energy-based peak position detection. Using the emotional models generated by SVM-AdaBoost learning algorithm, the music signal of the music videos is classified into one of the predefined emotional classes of the music automatically. For the video track, the face scene including the singer or actor/actress is detected based on a boosted cascade of simple features. Finally, the salient region is generated based on the alignment of boundaries of the music highlight and the visual face scene. First, the users select their favorite music videos from various music videos in the mobile devices or DVR with the information of a music video's emotion and thereafter they can browse the salient region with a length of 30-seconds using the proposed algorithm quickly. A mean opinion score (MOS) test with a database of 200 music videos is conducted to compare the detected salient region with the predefined manual part. The MOS test results show that the detected salient region using the proposed method performed much better than the predefined manual part without audiovisual processing.

Keywords: Salient region detection, Highlight detection, Automatic music emotion classification, Face detection, Music video browsing
ASK subject classification: Musical Acoustics and Psychoacoustics (8)

I. 서 론

방송·통신 기술의 발달로 인해 방송채널수가 급신적으로 증가됨에 따라 다양한 대용량의 콘텐츠로부터 사용자가 선호하는 영상 콘텐츠를 빠르고 편리하게 브라우징할 수 있는 기술이 필요 되고 있다. 이로 인해 스포츠 비디오, 뉴스 비디오, 영화 비디오를 대상으로 하는 자동 브라우징 시스템에 대한 많은 연구 및 개발이 진행되어 오고 있으며 점차적으로 DVR에 적용되고 있다. 그 반면에, 인터넷 사이트와 음악채널을 통해 제공되고 있는 뮤직비디오에 대한 브라우징 연구는 최근까지 많은 연구가 진

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

행되어 오지 않았다.

뮤직비디오는 가장 대중적인 장르 중의 하나로 다양한 인터넷 음악사이트나 IPTV 음악채널들에서 끊임없이 제공될 뿐만 아니라 휴대용 멀티미디어 기기에 저장하여 사용자가 언제 어디서나 즐길 수 있는 콘텐트이다. 이러한 다양한 수많은 뮤직비디오들로부터 사용자가 원하는 뮤직비디오를 선별하기 위해서는 각 뮤직비디오를 대표하는 중요구간이 빠르게 자동으로 검출되어 사용자에게 제공되어야 한다.

본 논문에서는 이러한 대중적인 장르중의 하나인 뮤직비디오를 대상으로 뮤직비디오 브라우징을 위한 실시간 중요구간검출 알고리즘을 제안한다.

뮤직비디오 브라우징과 관련된 대표적인 연구 [1-2]를 살펴보면, 뮤직비디오의 음악신호에서 검출된 후렴구간, 영상신호로부터 검출된 장면분류와 가사검출을 통해 뮤직비디오를 요약하는 방법 [1]이 있으며 음악을 색으로 표현하는 컬러 맵을 사용하여 반복되는 구조를 찾아 후렴구간을 검출 [2]하는 방식 등이 있다. 이러한 방식들은 음악신호의 전체구간 구조 분석을 통해 후렴구간을 검출하기 때문에 계산량이 많아 긴 음악을 대표하는 후렴구간을 실시간으로 검출할 수 없으며, 또한 PMP, PDP와 같은 휴대용 기기에서 브라우징을 위한 실시간 신호처리가 어렵다.

본 논문에서는 휴대용 기기 및 DVR에 적용할 수 있는 뮤직비디오 브라우징을 위한 중요 구간 검출 방법을 제안한다. 제안하는 방법은 뮤직비디오에서 분할되는 음악신호에서 전체 구간이 아닌 음악 후렴구간이 존재하는 특정구간에 대해 후렴구간을 포함하는 음악 하이라이트 구간을 검출하고 영상신호에서도 음악의 후렴구간과 상응하는 특정구간에서 얼굴장면을 검출하기 때문에, 음악 및 영상신호 전체 구간에 대해 후렴구간 및 얼굴영상 장면을 검출하는 기존의 연구보다 빠른 연산 속도를 보인다.

본 논문의 구성은 다음과 같다. 제 II장에서는 전체적인 시스템의 구성도를 살펴보고 음악신호에서 후렴구간을 포함하는 하이라이트 검출 방법과 분위기 별 음악분류 방법, 영상신호에서 얼굴을 검출하는 방법, 그리고 음악신호에서 검출한 하이라이트와 영상신호에서 얼굴이 많이 검출되는 구간을 결합하여 중요구간을 결정하는 과정을 설명한다. 제 III 장에서는 제안된 시스템의 실험결과를 분석 및 고찰하며 제 IV장에서 결론과 향후 연구 방향을 기술한다.

II. 뮤직비디오 브라우징을 위한 중요 구간 검출 구성도

그림 1은 제안된 뮤직비디오의 중요구간 검출방법의 전체적인 구성도를 나타낸다.

입력된 뮤직비디오는 신호 분배기를 통해 음악신호와 영상신호로 분리된다. 분리된 음악신호는 특정구간으로부터 추출된 음악 특징 값을 이용하여 음악 하이라이트 구간을 검출하고, 검출된 음악 하이라이트 구간을 분위기 별로 자동 분류한다. 영상신호에서는 음악신호의 특징구간에 상응하는 영상신호로부터 영상 특징 값을 추출하여 사전에 학습된 얼굴 모델과의 유사도를 비교하여 얼굴장면을 검출한다. 이렇게 음악신호에서 검출된 음악 하이라이트 구간과 영상신호로부터 검출한 얼굴 출현 빈도수가 많은 구간을 결합하여 뮤직비디오의 중요구간을 최종적으로 결정한다. 구성된 시스템을 통해 사용자는 인차적으로 뮤직비디오를 분위기 별로 선택하고 이차적으로 선별된 뮤직비디오의 중요구간을 브라우징하여 선호하는 뮤직비디오를 빠르게 선별하게 된다.

2.1. 음악 하이라이트 검출

일반적으로 음악의 후렴구간은 하나의 음악 곡에서 두 번 이상 반복되며 청취자에게 다른 음악 부분보다 강하게 영향을 주는 구간이기 때문에 대부분의 청취자는 음악을 들은 후에 후렴구간의 멜로디나 리듬을 기억하는 경우가

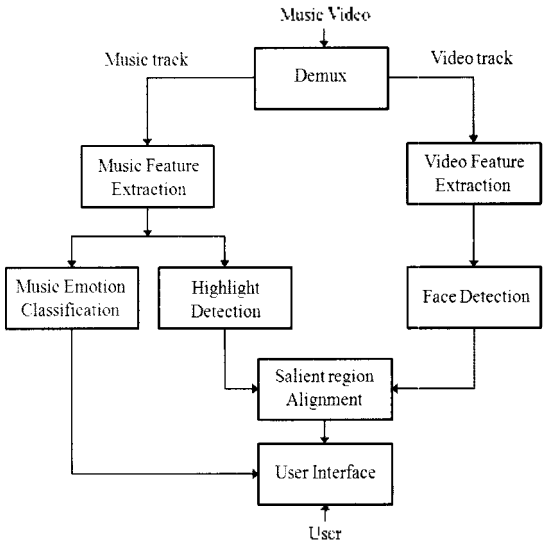

그림 1. 중요구간 검출 방법 구성도
Fig. 1. Block diagram of the salient region detection algorithm.

많다. 이로 인해 음악요약 시스템에서는 이러한 후렴구간을 음악을 대표하는 요약본으로 사용하며 후렴구간을 검출하는 다양한 방식들에 대한 연구를 진행해 오고 있다. 그 중에서도 입력된 음악신호를 decompose한 옥타브 서브밴드로부터 검출한 onset에서 autocorrelation을 이용하여 최고점 (peak)을 찾고 이 최고점을 기반으로 전체 음악신호를 세그멘트 별로 분리한 후에 각 세그멘트의 유사성을 측정하여 유사성이 있는 세그멘트를 후렴구간으로 결정하는 방식 [3]이 검출정확도가 높은 방식으로 알려져 있다. 그러나 이 방식은 음악신호의 전체 구간에 대해서 신호처리를 수행하고 유사성을 측정하기 때문에 계산속도가 느린 단점을 갖고 있다.

본 논문에서는 200곡의 음악신호에서 후렴구간을 관찰해 본 결과 첫 번째 후렴은 40초에서 130초 이내의 구간에 존재함을 확인하고 이 구간을 하이라이트 후보영역으로 정의하여 이 영역 안에서 음악의 후렴구간을 포함하고 있는 하이라이트 구간을 검출하여 검출속도를 향상시킨다. 검출되는 하이라이트 구간은 후렴구간의 처음과 끝 부분 모두를 정확하게 포함하고 있지는 않지만 60% 이상의 후렴구간 일부분과 함께 후렴구간 전이나 후렴구간 후 부분을 포함하기 때문에 사용자에게 음악 곡의 대표적인 전환적인 흐름을 제공하여 사용자가 원하는 뮤직비디오를 빠르게 선별할 수 있도록 한다.

그림 2는 음악 하이라이트 검출 과정을 나타낸다.

하이라이트 후보 영역의 연속적인 입력 음악신호는 1초 단위의 세그먼트로 균등하게 분할된다. 분할된 음악 세그멘트 신호로부터 10 ms 단위의 프레임에서 음악 특징값을 추출한다. 특징 값으로 normalized audio spectrum envelope (NASE) [4] $x(l,b)$를 사용하며 식 (1)과 같이 정의된다.

$$x(l,b)=\frac{10\log_{10}\left|\sum_{k=loK_b}^{hiK_b}P(l,k)\right|}{\sqrt{\sum_{b=1}^{B}\left|10\log_{10}\left(\sum_{k=loK_b}^{hiK_b}P(l,k)\right)\right|^2}} \tag{1}$$

여기서 $P(l,k)$는 파워스펙트럼, l은 프레임 인덱스, b는 서브밴드 인덱스, k는 주파수 인덱스, loK_b는 각 서브밴드의 가장 낮은 주파수, hiK_b는 각 서브밴드의 가장 높은 주파수, B는 서브밴드의 개수를 나타낸다.

프레임 별로 구해진 NASE는 한 세그먼트 단위로 NASE의 각 서브밴드별 평균값을 계산한다. 서브밴드별로 구해진 평균 NASE는 모든 서브밴드의 값을 합쳐서 세그먼트 단위의 평균 NASE의 합인 NASE 에너지 $X(s)$을 구한다. 이러한 평균 NASE의 합은 식 (2)와 같다.

$$X(s)=\sum_{b=1}^{B}\left[\frac{\sum_{l=l_s}^{l_e}x(l,b)}{l_e-l_s}\right] \tag{2}$$

여기서 l_s는 각 세그먼트의 시작 프레임이고 l_e는 각 세그먼트의 끝 프레임이다.

NASE 에너지 $X(s)$는 최고점 (PP) 검출, 하이라이트 시작 및 하이라이트 끝 후보 위치 (HSECP)를 검출하기 위해 사용된다. 대부분의 음악후렴구간은 악기음, 보컬 등이 어울려져 다른 구간보다 에너지가 높기 때문에 각 세그멘트 단위로 구한 NASE 에너지 $X(s)$의 크기를 비교하여 후렴구간에 위치하는 최고점을 찾을 수 있게 된다. 이 최고점이 후렴구간 안에 존재하기 때문에 최고점을 기준으로 최고점의 앞부분에서 에너지가 낮은 세그멘트를 하이라이트 시작점 (HSV)으로 추정할 수 있고 최고점의 뒷부분에서 에너지가 낮은 세그멘트를 하이라이트 끝점 (HEV)으로 추정할 수 있다. 에너지가 낮은 세그멘트

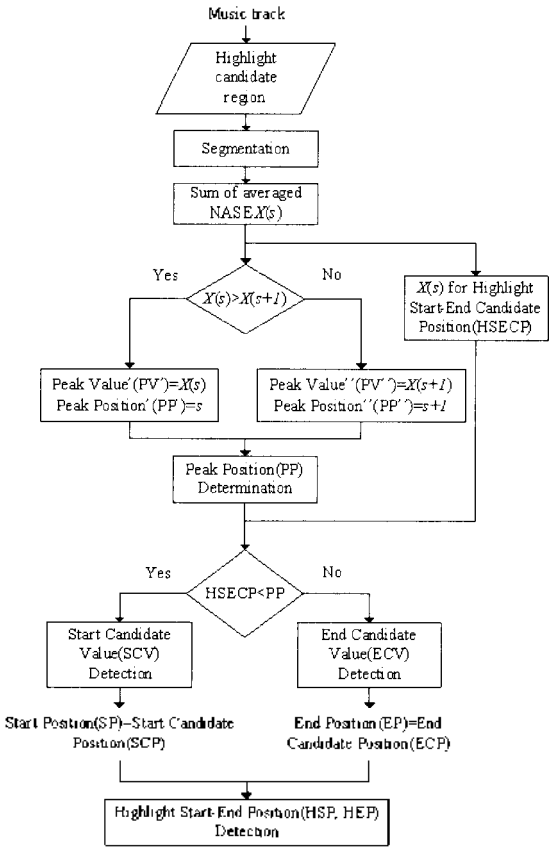

그림 2. 하이라이트 검출 알고리즘 구성도
Fig. 2. Block diagram of highlight detection algorithm.

는 스무딩한 음악하이라이트의 시작점과 끝점으로 사용된다.

최고점 검출 과정을 그림 2를 통해 살펴보면, $X(s)$와 $X(s+1)$을 비교하여 큰 값을 최고값 (PV)으로 정의하고 PV의 위치를 최고점 위치 (PP)로 정의한다. $X(s)$가 $X(s+1)$보다 클 경우, $X(s)$가 PV가 되고 s는 PP가 된다. 반대로 $X(s+1)$이 $X(s)$보다 클 경우, $X(s+1)$이 PV'가 되고 $s+1$은 PP'가 된다. 특정구간에서 각 1초 단위의 세그먼트로 연속적으로 계산되는 $X(s)$와 $X(s+1)$를 비교하여 최고점 PV를 최종적으로 찾아서 최고점 위치 PP를 검출한다. 하이라이트의 시작 위치 (HSP) 및 끝 위치 (HEP)는 구해진 PV와 PP를 이용하여 검출된다. 즉, HSECP가 PP보다 작고 HSECP가 PP로부터 25초 영역 내에 존재하며 하이라이트 시작 및 끝 후보 값 (HSECV)이 일정 문턱값 이상이 되면 HSECV는 하이라이트 시작 후보값 (SCV)이 되고, 이러한 시작 후보값 (SCV)들 중에서 가장 작은 값을 갖는 SCV에 대응되는 시작 후보점 SCP을 음악하이라이트 시작점 (SP)으로 결정한다. 반대로 PP보다 HSECP가 큰 경우, HSECP는 끝 후보 점 (ECP)이 되며 HSECV는 끝 후보값 (ECV)이 된다. SCV 검출과정과 마찬가지로, ECV 중에서 가장 작은 값을 갖는 ECV를 구하여 이에 대응되는 ECP는 음악 하이라이트 끝점 (EP)로 결정한다. 이와 같은 과정을 통해 구해진 HSP와 HEP가 구하고자 하는 하이라이트 구간이 된다.

2.2. 자동 음악 분위기 분류

본 논문에서는 음악신호를 분위기 별로 분류하기 위해 검출된 음악하이라이트 구간으로부터 음색 (timbre)와 템포 (tempo) 특징을 추출하고 AdaBoost 학습기반의 Cascade 구조 [5]를 사용한다. 음색 특징은 입력된 음악신호의 옥타브 단위의 서브밴드로부터 구해진 audio spectrum envelope (ASE)

$$ASE(l,k)=\sum_{k=loK_l}^{hiK_l} P(l,k) \tag{2}$$

의 arithmetic mean, peak, valley, flatness, crest, band width, centroid, roll-off frequency와 spectral flux를 추출한다. 템포 특징은 각 옥타브 단위의 서브밴드의 ASE 값에 연속 웨이블렛 변환 (WT)을 적용하여 추출된다. 분류기는 AdaBoost learning 알고리즘 [5] 기반의 두 개의 층으로 구성된 분류구조를 사용하여 음악신호를 강렬한 곡, 쾌적한 곡, 조용한 곡 등의 3개의 분위기 클래스로 자동 분류한다. 첫 번째 층에서는 음색 특징을 이용해 음악을 분류하고, 두 번째 층에서는 템포 특징을 이용해 음악을 분류한다. 각 층에서는 support vector machine (SVM)기반의 AdaBoost cascade 분류구조를 이용하여 정해진 하나의 무드 클래스에 대해 모델을 생성하고 다른 2개의 클래스로부터 또 다른 하나의 모델을 생성하여 3개의 무드로 분류한다. 즉, 첫 번째 분류기에 의해 오류로 판단된 음악신호는 두 번째 분류기에 의해 다시 분위기를 판별하여 분위기분류의 정확성을 높인다.

2.3. 얼굴 영역 검출

뮤직비디오의 영상에서는 음악의 후렴구간에 해당하는 영상장면에 가수의 얼굴이나 주인공의 얼굴을 빈번하게 등장시켜 뮤직비디오의 감흥을 고조시킬 수 있도록 제작되기 때문에 얼굴영역 검출을 통해 음악신호의 후렴구간을 추적할 수 있다. 또한 뮤직비디오의 음악과 관계없이 가수의 얼굴이나 뮤직비디오에 출연한 유명 배우의 얼굴을 시청하고자 하는 사용자 선호도가 높기 때문에 얼굴영역 검출은 뮤직비디오 브라우징에 있어서 중요한 요소이다.

본 논문에서는 얼굴영역 검출을 위해 잘 알려진 Viola [6]의 방식을 사용하였다. 먼저, 얼굴과 얼굴을 포함하지 않은 사진 데이터베이스로부터 추출한 haar-like features를 AdaBoost 학습 알고리즘을 통해 얼굴 모델을 생성한다. 얼굴영역을 검출하기 위해서 입력된 영상신호는 1초 단위의 세그먼트로 분할되고, 분할된 세그멘트로부터 에지검출을 통해 찾은 얼굴후보 영역으로부터 haar-like features를 추출한다. 추출된 haar-like features와 생성된 얼굴 모델과의 유사도를 비교하여 Cascade 구조로 이루어진 분류기를 통해 얼굴의 존재 유무를 판단하게 된다. 본 논문에서는 2000개의 얼굴을 포함한 사진과 4000개의 얼굴을 포함하지 않은 사진을 데이터베이스로 구성하여 AdaBoost 학습기반의 얼굴 모델을 생성하였으며, 총 1785개의 HF 값을 사용하여 30단계로 이루어진 Cascade 구조의 분류기를 통해 얼굴 영역을 검출하였다.

그림 4는 얼굴 검출 결과를 나타낸 그림이다.

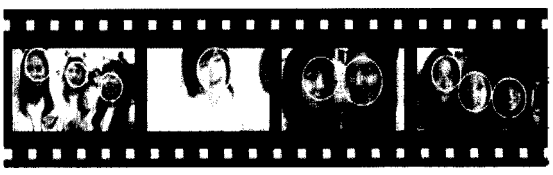

그림 4. 얼굴 검출 과정의 결과
Fig. 4. Results of the face detection.

2.4. 뮤직비디오 중요구간 결정

뮤직비디오의 중요 구간은 음악 신호로부터 검출된 하이라이트와 영상 신호로부터 검출된 가수의 얼굴을 포함하고 있는 영상 장면을 정련하여 결정된다. 그림 5는 중요구간 결정 과정을 나타낸다.

음악신호로부터 2.1절에 기술된 방식을 통해 검출된 음악하이라이트는 다음과 같이

음악 하이라이트 (HP)={하이라이트 시작점 (HSP), 하이라이트 끝점 (HEP)}

로 정의하고, 2.2절에 기술된 방식을 통해 영상신호로부터 검출한 얼굴 영상장면 세그멘트를 i-th VP라 정의하면 중요구간을 다음과 같이 결정할 수 있다.

음악 하이라이트 구간 위치에 대응되는 영상신호의 얼굴이 검출된 얼굴영상 세그멘트 i-th VP의 연속적인 빈도수가 δ보다 크면 δ보다 큰 얼굴영상장면이 존재하는 구간을 중요구간으로 결정하고

중요구간 (SR)={VSP, VEP}

로 나타낸다. VSP, VEP는 각각 음악 하이라이트 구간에 상응하는 영상신호에서 정의된 문턱값 보다 높은 얼굴을 포함한 영상장면의 시작점과 끝점의 위치를 나타낸다.

반면에 i-th VP의 연속적인 빈도수가 δ보다 적다면, 중요 구간은 얼굴영상장면과 관계없이 음악하이라이트 구간에 따라 결정되어

중요구간 (SR)={HSP, HEP}

으로 나타나게 된다.

그림 6은 중요 구간 결정의 예를 나타낸다. 그림 6의 아래 그림은 음악신호로부터 검출된 최고점 (-), 음악하이라이트 구간 (--), 실제 후렴구간 (-)을 표시하였다. 검출된 음악 하이라이트 구간은 전체 후렴구간을 포함하고 있으며 후렴구간 전의 곡의 일부분을 포함하고 있다. 그림 6의 위의 그림에서는 영상신호로부터 검출된 얼굴영역 세그멘트가 검은색으로 표시되어 있다. 음악 하이라이트 구간 위치에 존재하는 얼굴영역 세그멘트의 연속적 빈도수가 많은 구간을 중요구간으로 선정하면 그림6에서는 중요구간과 후렴구간이 거의 일치하게 됨을 알 수 있다.

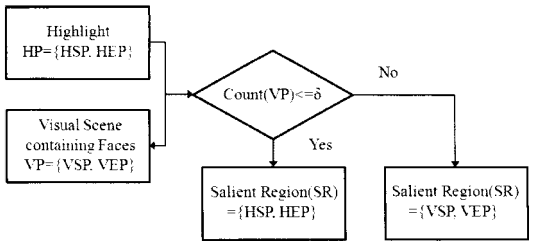

그림 5. 중요 구간 조정 구성도
Fig. 5. Block diagram of salient region alignment.

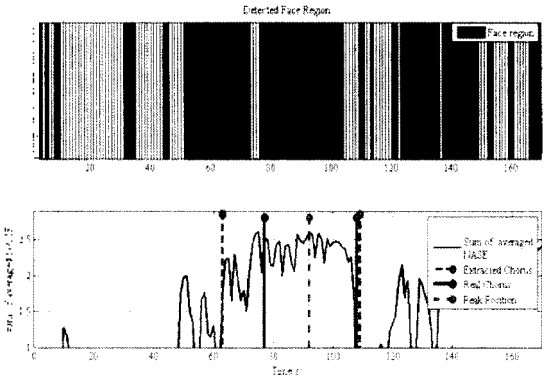

그림 6. 중요 구간 조정의 예
Fig. 6. The example of salient region alignment.

II. 실험 결과 및 분석

본 논문에서는 제안된 시스템의 하이라이트 검출 성능을 측정하기 위해 락, 팝, 댄스, 발라드, 랩과 같은 다양한 장르를 포함하는 200개의 AVI 파일 포맷을 가진 뮤직비디오 (초당 30프레임, 해상도 400×300)를 실험에 사용하였다. 각 뮤직비디오의 길이는 3분에서 4분 30초 사이로 구성되어 있으며 모든 음악 신호는 샘플 당 16 bits, 44.1 kHz sampling rate를 사용하였다.

제안된 방식의 성능은 후렴구간 검출 정확도, 중요구간 MOS 테스트, 분위기별 음악분류 정확도를 통해 측정되었다.

검출된 음악하이라이트의 결과를 평가하기 위한 후렴구간 검출 정확도 (CDA)는 식 (3)과 같이 정의한다.

$$CDA = \frac{\text{검출된 후렴구간}}{\text{후렴구간의 총 개수}} \tag{3}$$

검출된 후렴구간은 후렴 구간 안에 최고값이 존재하고 실제 후렴구간의 60% 이상을 포함하는 것을 의미한다.

표 1에는 제안된 방식에서의 후렴구간 검출결과를 2.1절에서 언급한 Goto [3]의 방식과 비교하여 나타내었다.

표 1. 후렴구간 검출 정확도 결과
Table 1. Chorus detection accuracy results.

	Accuracy
제안된 방식	83%
Goto [2]의 방식	86%

제안된 방식에서 후렴구간 검출결과가 비교된 [3]의 방식에 비해 3% 검출성능이 낮은 이유는 랩 음악 곡에서는 에너지 최고점이 후렴구간에 존재하기 보다는 곡 중의 강한 억양 악센트와 비트가 혼합된 영역에서 검출되는 경우가 발생하기 때문이다. 반면에 [3]의 방식에서는 전체 곡에 대한 유사성을 측정하여 제안된 방식 보다 나은 후렴구간을 검출할 수 있었다. 그러나 제안된 방식과 [3]의 방식을 700 MHz Pentinum IV CPU를 가진 PC (2.53 GHz, 512 MB RAM memory)를 통해 3분 길이의 뮤직비디오로부터 후렴구간 검출 수행 시간을 측정한 결과 제안된 방식에서는 2초, [3]의 방식에서는 8초가 소요되었다. 그렇기 때문에 제안된 방식이 [3]의 방식에 비해 3% 정도의 낮은 검출정확도를 갖고 있더라도 후렴구간 검출 속도가 4 배정도 빠르기 때문에 모바일 단말기나 DVR의 실시간 신호처리에 적용가능하다.

뮤직비디오의 중요 구간에 대한 사용자 만족도를 평가하기 위한 MOS 테스트는 18살에서 26살로 구성된 20명의 참가자를 대상으로 수행되었다. 참가자들은 실험측정자가 보여주는 각 뮤직비디오의 처음부터 끝까지를 두 번 시청하고, 그 뮤직비디오에 해당되는 정해진 수동 구간 (Manual region)과 제안된 시스템에 의해 검출된 중요구간 (Salient region)을 각각 세 번 시청한 후, 각 구간에 대한 만족도를 1-5 점 (1: worst, 2: bad, 3: moderate, 4: good, 5: best)사이에서 정하였다. 1부터 5까지 점수에 대한 평가기준을 위해 피실험자들에게 제시된 지문은 다음과 같다.

- 제시된 구간을 통해 전체 뮤직비디오가 제시하는 내용을 이해 할 수 있는가?
- 제시된 구간의 음악을 통해 전체 곡에 대한 분위기를 파악할 수 있는가?
- 제시된 구간의 영상과 음악이 뮤직비디오를 대표할 수 있는 하이라이트라고 판단되는가?

위의 실험에서 사용된 뮤직비디오의 수동구간은 신호처리 수행 없이 뮤직비디오가 재생되는 처음 시작부터 30초까지의 정해진 고정구간이며, 중요구간은 제안된 방식을 통해 신호처리가 수행되어 검출된 30초에서 40초의 길이를 가진 구간이다.

표 2. MOS 테스트 결과
Table 2. Results of the MOS test.

	Quality
Salient region	3.25 ± 0.59
Manual region	1.9 ± 0.35

표 2는 실험을 통해 구해진 MOS 테스트의 결과를 보여주고 있다.

표 2는 제안된 시스템을 통해 검출된 중요구간이 사용자 만족도 측면에서 정해진 수동 구간보다 우수함을 제시한다. 이를 통해 음악 후렴구와 가수나 배우의 얼굴을 포함한 중요구간이 정해진 수동 구간보다 사용자 측면에서 뮤직비디오를 빠르게 선별하는 기준을 제공함을 알 수 있었다.

200개의 뮤직비디오에서 검출된 음악하이라이트 구간에 해당하는 음악신호를 상렬함, 쾌적함, 조용함의 세 개의 클래스에 대한 분류정확도는 83% 로서 전체 곡의 음악신호로부터 음악 특징값을 추출하여 분류한 결과보다 10% 정도 낮은 성능을 보였다. 그러나 전체 곡을 대상으로 할 때보다 6배 정도 빠르게 무드별로 분류하기 때문에 83%의 분위기별 분류성능은 모바일 단말기에 큰 문제가 되지 않는다고 판단된다.

IV. 결 론

본 논문에서는 뮤직비디오에서 후렴 구간을 포함하는 하이라이트와 얼굴 장면의 결합을 통해 실시간으로 중요구간을 검출하는 알고리즘을 제안하였다. 제안된 방식은 뮤직비디오의 특정 구간 안에 존재하는 음악의 하이라이트와 얼굴영상 장면을 검출하기 때문에 전체 음악신호에서 음악의 구조를 분석 및 전체 영상신호에서 얼굴장면을 검출하는 기존의 방법들보다 계산량을 크게 줄여 실시간 처리가 가능케 하였다. 또한, MOS 테스트의 결과를 통해 제안된 시스템을 통해 검출된 중요구간이 정해진 처음 구간보다 향상된 사용자 만족도를 나타내기 때문에 실시간 신호처리를 필요로 하는 모바일 단말기나 DVR에 효과적으로 적용될 수 있으리라 판단된다. 또한 사진과 음악을 결합하여 음악 비디오를 자동으로 생성하는 장치에 효과적으로 적용될 수 있다.

향후 계획은 다음과 같다. 첫째, 하이라이트 검출과 얼굴 검출을 위하여 보다 나은 특징 값을 사용하고 둘째, 실험 데이터베이스를 늘려 제안된 방식의 성능을 확인할

것이다. 마지막으로 제안된 방식을 실제 휴대용 기기에 적용하여 그 성능을 검증 하고자 한다.

감사의 글

이 논문은 2008년도 정부재원 (교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음 (KRF-2008-331-D00421).

참 고 문 헌

1. C. Xu, X. Shao, N.C. Maddage and M.S. Kankanhalli, "Automatic music video summarization based on audio-visual-text analysis and alignment," *Proc. 28th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil*, pp. 361-368, 2005.
2. C. H. Yeh and H. H. Lin, "The extraction of popular music chorus structural content analysis," *Proc. Industrial Electronics Society (IECON): 33rd Annual Conference IEEE*, Taipei, Taiwan, pp. 2532-2536, 2007.
3. M. A. Goto, "Chorus-section detecting method for music audio signals," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New York, U.S.A., pp. 437-440, Apr. 2003.
4. H.-G. Kim, N. Moreau and T. Sikora, "Audio classification based on MPEG-7 spectral basis representations," *IEEE Transaction Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 716-725, May 2004.
5. X. Zhu, Y.Y. Shi, H.-G. Kim and K.-W. Eom, "An integrated music recommendation system," *IEEE Transaction on Consumer Electronics*, vol. 52, no. 3, pp. 917-925, Aug. 2006.
6. P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proc. Computer Vision and Pattern Recognition (CVPR)*, Netherlands, pp. 511-518, 2001.

저자 약력

•김 형 국 (Hyoung-Gook Kim)

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

•신 동 (Dong Shin)

2009년 2월 광운대학교 전파공학과 (공학사)
2009년 2월~ 현재: 광운대학교 전파공학과 (석사과정)