특집논문-07-12-4-02

다중모드 특징을 사용한 뉴스 동영상의 앵커 장면 검출 기법

유 성 열[a], 강 동 욱[a], 김 기 두[a], 정 경 훈[a]‡

Multi-modal Detection of Anchor Shot in News Video

Sung Yul Yoo[a], Dong Wook Kang[a], Ki Doo Kim[a], and Kyeong Hoon Jung[a]‡

요 약

본 논문에서는 뉴스 동영상 요약 정보의 생성을 위해 뉴스 단위의 기준이 되는 앵커 장면을 효과적으로 검출하는 기법을 제안한다. 우선 뉴스 동영상의 오디오 및 비디오 구성 요소에 대한 관찰을 통하여 앵커 장면 검출에 적합한 기본적인 특징들을 선택하였다. 제안 알고리듬에서는 색인의 정확도를 높이기 위해 몇몇 오디오 특징과 함께 비디오 특징으로서 움직임 특징을 함께 이용하였으며, 전체적인 구조는 '오디오 정지 구간 검출', '오디오 클러스터 분류', 그리고 '움직임 활동도와의 매칭'의 3단계로 구성된다. MPEG-2 방식으로 부호화된 뉴스 동영상에 대한 실험을 통해 제안 알고리듬의 성능이 만족스러움을 확인하였다.

ABSTRACT

In this paper, an efficient detection algorithm of an anchor shot in news video is presented. We observed the audio visual characteristics of news video and proposed several low level features which are appropriate for detecting an anchor shot in news video. The overall structure of the proposed algorithm is composed of 3 stages: the pause detection, the audio cluster classification, and the matching with motion activity stage. We used the audio features as well as the motion feature in order to improve the indexing accuracy and the simulation results show that the performance of the proposed algorithm is quite satisfactory.

Keyword : news video indexing, multi-modal feature, MFCC, and motion activity

I. 서 론

지난 십수년간 신호처리 및 통신기술의 발달을 통해 DTV(Digital Television), DMB(Digital Multimedia Broadcasting) 및 PMP(Portable Multimedia Player) 등 다양한 멀티미디어 서비스들이 상용화되면서 멀티미디어 콘텐트를 손쉽게 접할 수 있게 되었으며 데이터의 양도 비약적으로 증가하게 되었다. 이에 따라 사용자 입장에서 관심이 있는 멀티미디어 콘텐트를 어떻게 검색하고 활용할 것인가의 문제가 콘텐트를 생성하고 전송하는 문제 못지않게 중요한 의의를 가지게 되었다. 그리고 멀티미디어의 내용기반 검색 기법은 연구 차원에서 뿐만 아니라 상업적인 차원에서도 많은 관심의 대상이 되고 있다. 특히 멀티미디어 콘텐트의 핵심이라고 할 수 있는 동영상의 경우에는 전체적인 동영상 시퀀스를 몇 개의 클립으로 구분하여 각 클립마다 동일한 의미적 구성을 가지도록 하는 동영상 색인이 필수적인 기법으로 주목을 받고 있다.

한편 동영상 콘텐트를 제작함에 있어서 의미론적 정보 전달의 채널은 비디오, 오디오, 및 텍스트의 세 가지 모드로 생각할 수 있으며 다중 모드란 이들을 결합하여 활용하는

a) 국민대학교 전자공학부,
School of Electrical Engineering, Kookmin University
‡ 교신저자 : 정경훈 (khjung@kookmin.ac.kr)
※ 이 논문은 IWAIT2007에 발표된 논문임.
본 연구는 국민대학교 교내연구비 지원으로 수행되었음.

것을 말한다[1]. 동영상 색인을 위해서 단일 채널의 특징만을 사용하기 보다는 여러 채널의 특징을 함께 사용할 때 색인의 정확도를 향상시킬 수 있음은 당연하다. 동영상 색인 기법 연구에 있어서 고려해야 할 사항 가운데 하나는 각각의 알고리듬이 특정 장르에 종속될 수밖에 없어서 다른 장르의 동영상에는 활용되기 곤란하다는 점이다. 예를 들어 스포츠 동영상과 같은 경우에 중요하게 고려될 수 있는 '초록색 잔디가 깔린 운동장' 이나 '관중들의 함성'과 같은 특징은 뉴스나 드라마 동영상에서는 사용하기 곤란하다. 따라서 동영상의 장르 및 구체적인 응용분야에 따라 적합한 특징들을 선택하는 것이 중요한 문제이며, 본 논문에서의 관심의 대상인 뉴스 동영상에 적합한 색인 기법을 개발하기 위해서는 당연히 뉴스 동영상의 구별에 적합한 특징들을 고려해야 한다.

다중 모드 특징을 사용하여 뉴스 동영상을 분석하고 요약하기 위한 연구는 다양하게 진행되어왔다. Qi 등은 비디오와 오디오 정보를 사용한 뉴스의 구분 결과와 캡션 영역으로부터 추출한 텍스트 정보를 함께 사용하는 방법을 제안하였으며[2], Hsu 등은 움직임, 얼굴, 음악/음성 등 여러 가지 특징을 사용하면서 최대 엔트로피(Maximum Entropy)에 근거한 통계적 모델을 통해 뉴스를 구분하였다[3]. 또한 Chaisorn 등은 뉴스 장면을 카테고리 별로 구분한 뒤 HMM(Hidden Markov Model)을 통해 분석하였는데 여기에서도 색상 히스토그램(color histogram) 등의 비디오 특징과 함께 오디오 및 텍스트 특징을 사용하였으며[4], Wu 등은 비디오 정보와 텍스트 정보를 동시에 고려하는 클러스터링 기법을 통한 뉴스 동영상의 자동문서화 기법을 제안하였다[5].

이러한 연구들은 다중 모드 특징을 사용하면서도 주로 장면 검출 및 키프레임 추출 등의 비디오 신호처리기술에 기반을 두고 있다. 뉴스 동영상에서 비디오 정보는 가장 특징적이라는 점에서 유용하지만 계산의 복잡도 측면에서는 많은 데이터량으로 인해 오디오 정보보다 불리한 것이 사실이다. 본 논문에서는 '평균 에너지', '주파수 성분 차이', 그리고 'MFCC(Mel Frequency Cepstral Coefficients)' 등의 오디오 특징을 기반으로 하면서 비디오 특징으로서 '움직임 활동도'를 함께 고려하는 앵커장면 검출 기법을 제안하고자 한다.

논문의 구성은 다음과 같다. 2장에서는 뉴스 동영상에 포함된 다양한 오디오-비디오 요소들의 구성 방법에 대한 관찰을 통하여 선택된 기본적인 특징들을 정리한다. 그리고 3장 및 4장에서는 제안하는 앵커 장면 검출 알고리듬을 설명하고 실험 결과를 나타낸다. 이어서 5장에서 결론을 기술한다.

II. 뉴스 동영상에 대한 관찰 및 기본 특징

그림 1에서는 뉴스 동영상으로부터 획득한 전형적인 장면들을 나타내었다. 일반적으로 뉴스 비디오는 동일한 소재를 취급하는 방송 기사들이 모여 만들어지는데, 일반적으로 각각의 방송 기사는 그림 1 (a)와 같은 앵커 장면과 그림 1 (b)와 같은 보조설명 장면으로 구성된다. 그리고 뉴스 동영상이 진행되면서 앵커 장면과 보조설명 장면이 교

(a) (b)

그림 1. 뉴스 동영상의 전형적 장면. (a) 앵커 장면, (b) 보조설명 장면.
Fig. 1. The example pictures in news video. (a) anchor shot, (b) supplementary shot.

대로 번갈아 가면서 등장하는 것이 뉴스 동영상의 가장 두드러진 특징이다. 즉 하나의 방송 기사는 앵커 장면이 시작하면서부터 다음 앵커 장면이 등장하기 직전까지에 해당한다. 따라서 앵커 방면의 검출은 뉴스 동영상 색인을 위해서 가장 중요한 단계가 된다.

앵커 장면의 특징은 화면 내의 전체적인 움직임은 거의 없고 경우에 따라 화면 내의 일부분, 즉 소위 '어깨걸이' 영역에서만 움직임 정보가 있다는 것이다. 그러나 이와 같은 비디오 특징만 가지고는 앵커 장면을 검출하기에 불충분하다. 보조설명 장면에서도 앵커 장면과 유사한 장면이 수시로 관찰되기 때문이다. 그림 2에서는 보조설명 장면에 해당하면서도 앵커 장면과 유사한 예를 나타내었다. 더군다나 뉴스 제작에 있어서 시각적인 효과가 강조되면서 어깨걸이의 위치도 화면 우측 상단이라는 전통적인 자리를 벗어나는 경우가 증가하고 있고 컴퓨터 그래픽이 아닌 실제 촬영한 영상을 앵커의 배경으로 하는 등 다양한 화면 구성이 등장하고 있기 때문에 단순한 비디오 특징만으로는 앵커 장면의 검출의 정확도에 한계를 보일 수 밖에 없다.

이 문제의 해결을 위해 앵커 장면과 관련된 오디오 특성을 살펴볼 필요가 있다. 뉴스 동영상에서 새로운 뉴스 기사로 전환되면서 앵커가 등장하게 되면, 짧은 일정 시간 동안 오디오 신호가 없는 구간이 나타나며 이는 장면의 경계를 찾는 중요한 특징이 된다[3]. 또한 앵커 장면의 경우에는 앵커가 단독으로 말하고 있는 반면 보조설명 장면의 경우에는 주위의 잡음이나 배경음악이 함께 나타나는 것이 일반적이다. 즉 동영상 내의 오디오적인 요소가 앵커 장면을 검출하는 데에 있어서 중요한 특징이 된다. 이러한 관찰에 기초하여 뉴스 동영상의 색인을 위해서 다음과 같은 세 가지 오디오 특징을 선택하였다.

첫 번째 특징은 다음의 식 (1)과 같이 정의되는 평균에너지(Short time Average Energy) E_W 이다. 여기서 W_{size}는 윈도우의 크기이며 S_i는 오디오 신호의 표본값을 나타낸다.

$$E_W = \frac{1}{W_{size}} \sum_{i=0}^{W_{size}-1} |S_i|^2 \tag{1}$$

두 번째 특징으로서 주파수 성분차이(Delta Spectrum Magnitude) $\triangle F_i$를 사용한다. 이 특징은 오디오 신호를 FFT(Fast Fourier Transform)한 후 인접 성분과의 차이를 계산함으로써 얻어지는데 이를 식(2)에 나타내었다. 여기서 $f_i(u)$는 i-번째 프레임의 FFT 변환된 주파수 성분을 나타내고, M은 FFT의 크기에 해당된다.

$$\triangle F_i = \sum_{u=0}^{M-1} \| |f_i(u)| - |f_{i+1}(u)| \| \tag{2}$$

이 두가지 특징을 사용함으로써 뉴스 동영상을 'silence' 구간과 'non silence' 구간으로 구분하는 것이 가능하다.

세 번째 특징은 MFCC이다. MFCC는 인간의 청각인지 시스템에 근거한 오디오 특징으로서 음성인식에 매우 효율적인 것으로 알려져 있는데 스포츠 동영상의 색인을 위해 사용되기도 하였다[6]. 그림 3 (a)는 MFCC를 계산하는 과정을 나타내었으며 그림 3 (b)는 50Hz에서 16kHz 사이에서

그림 2. 앵커 장면과 유사한 보조설명 장면의 예.
Fig. 2 The example supplementary shots similar to anchor shot.

Mel 스케일 필터 뱅크(filter bank)의 예를 보였다.

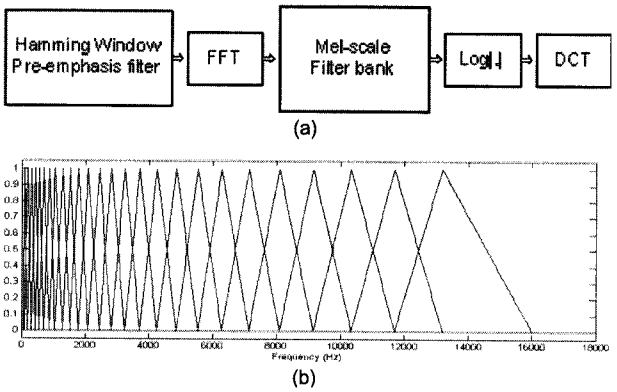


그림 3. MFCC 계산 및 Mel 스케일 필터 뱅크. (a) MFCC 계산과정, (b) Mel 스케일 필터뱅크의 예

Fig. 3. MFCC calculation and Mel-scale filter bank. (a) the block diagram of calculating MFCC, (b) the example of Mel-scale filter bank

한편 앞서 언급했듯이 앵커 장면과 보조설명 장면을 구별하는 특징으로서 움직임의 정도를 고려할 수 있다. 앵커 장면의 움직임 정도가 미미한 반면 보조설명 장면에서는 그렇지 않은 것이 일반적이기 때문이다. 본 논문에서는 움직임의 정도를 나타내기 위해 멀티미디어 콘텐트를 기술하는 표준 인터페이스인 MPEG-7에서의 움직임 기술자(motion descriptor)를 사용하였다[7]. MPEG-7에서는 'intensity of activity', 'direction of activity', 'spatial distribution of activity', 그리고 'temporal distribution of activity'와 같이 네 가지 종류의 기술자를 정의하고 있는데, 본 논문에서는 이 가운데 움직임 활동도의 세기를 나타내는 'intensity of activity'를 선택하였다. MPEG-7에서는 움직임 활동도의 세기를 움직임 벡터의 크기의 표준편차를 기준으로 1에서부터 5까지의 숫자로 구분하며 숫자가 높을수록 움직임의 활동도가 높은 경우에 해당한다.

III. 제안 알고리듬

제안 알고리듬의 전체적인 구조는 3단계로 구성된다. 첫 번째 단계는 뉴스 동영상을 'silence' 및 'non silence' 구간으로 나누는 '오디오 정지 구간 검출' 단계이고, 다음은 'non-silence' 구간을 더 세밀하게 분류하는 '오디오 클러스터 분류' 단계이며, 마지막은 두 번째 단계의 분류 결과를 움직임 활동도 특징과 함께 고려하는 '움직임 활동도와의 매칭' 단계이다. 각각의 단계를 살펴보면 다음과 같다.

1. 오디오 정지 구간 검출

뉴스 동영상의 분류를 위한 첫 번째 단계는 주어진 동영상의 오디오 신호 내에 정지 구간이 존재하는 지를 판단하는 것이다. 본 논문에서는 일반적인 오디오 신호의 내용기반 검색 기법을 위한 분류 기법을 제안한 Li 등의 방법[8]과 의사결정 트리(decision tree)를 사용한 검출 기법을 제안한 Sethi 등의 방법[9]을 활용하여 오디오 신호의 정지 구간을 결정하였다.

이 단계에서는 오디오 신호의 평균 에너지 및 주파수 성분 차이의 두가지 특징을 사용한다. 그림 4 (a)에서는 E_W를 x-축으로 그리고 $(\Delta F_i)^2/E_W$를 y-축으로 설정한 이차원 특징 공간을 나타내었다. 그림에서 S1으로 표시된 음영 영역이 'silence' 구간에 해당하고 S2로 표시된 나머지 영역이 'non-silence' 구간에 해당한다. 그리고 이에 해당하는 의사결정 트리를 그림 4 (b)에 보였다. 이 때 문턱값 T1, T2, T3 및 T4들은 실험을 통해 결정된다.

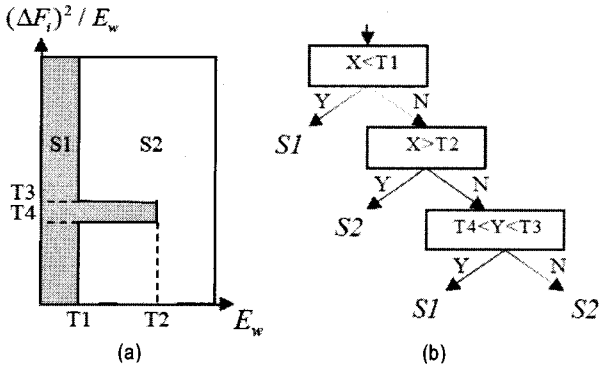


그림 4. 특징 공간 및 의사결정 트리. (a) 이차원 특징 공간, (b) 의사결정 트리

Fig. 4. The feature and decision tree. (a) the 2D feature space, (b) the decision tree

첫 번째 단계에서 오디오 신호의 정지구간이 일단 결정된 후에는 분류 결과의 타당성을 보장하기 위해 '채우기' 과정과 '버리기' 과정을 수행한다. '채우기' 과정에서는 'silence' 구간으로 분류된 프레임의 길이가 너무 짧은 경우에 이를 'non-silence' 구간으로 변경하며, '버리기' 과정에서는 'non-silence' 구간의 길이가 너무 짧은 경우 이를 'silence' 구간으로 변경한다.

2. 오디오 클러스터 분류

첫 번째 단계를 거쳐 뉴스 동영상에서 오디오 신호가 존재하는 구간과 존재하지 않는 구간이 구별되었다고 하더라도, 신호가 존재하는 'non-silence' 구간 내에는 앵커의 음성, 기자의 음성, 주위 잡음, 배경 음악 등등의 다양한 종류의 오디오 신호가 존재할 수 있다. 두 번째 단계에서는 이와 같이 다양한 오디오 신호를 카테고리별로 분류하고자 한다. 이는 앵커 장면에서는 일반적으로 잡음이나 음악이 섞이지 않고 앵커의 목소리만 단독으로 존재하는 반면, 보조설명 장면에서 기자의 질문이나 전문가의 인터뷰가 등장하는 경우에는 사람의 목소리와 더불어 취재 현장 주변의 잡음이나 배경 음악 등이 함께 존재하는 경우가 많다는 관찰 결과에 따른 것이다.

본 논문에서는 오디오 신호를 다음과 같이 5 가지의 카테고리로 분류한다. 실제로 앵커 장면의 검출만을 목적으로 한다면 이와 같이 여러 개의 카테고리가 필요하지 않을 수도 있으나 앵커장면 검출 이후의 후속 색인 작업을 고려하여 카테고리를 세분화하였다. 만일 첫 번째 단계에서 분리한 'silence'를 별도의 카테고리로 생각한다면 총 6개의 카테고리가 존재하는 셈이다.

카테고리 1 앵커 음성 (anchor speech)
카테고리 2 음악 (music)
카테고리 3 주변 잡음 (background noise)
카테고리 4 기자 및 인터뷰 음성 (reporter or interviewee speech)
카테고리 5 음악 및 음성 (music and speech)

두 번째 단계에서의 세부 분류는 MFCC 특징을 사용하여 이루어진다. 우선 다양한 오디오 클러스터들을 카테고리 별로 분류하고 각 카테고리 마다 MFCC의 평균값을 계산하여 이를 카테고리의 기준 특징으로 삼는다. 본 논문에서는 사용한 MFCC 필터뱅크의 개수는 20개이다. 그림 5에서는 실험에서 얻어진 오디오 카테고리 별 MFCC 특징을 나타내었다.

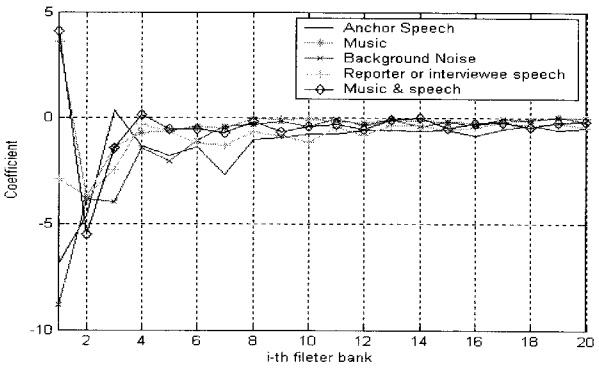


그림 5. 오디오 카테고리 별 MFCC 특징
Fig. 5. The MFCC characteristics of audio categories

이제 분류하고자 하는 동영상이 입력되면 프레임마다 MFCC를 계산하고 이를 각 카테고리의 기준 MFCC와 비교하여 가장 거리가 가까운 카테고리로 분류한다. 이 때 거리함수로는 식 (3)으로 주어지는 Mahalanobis distance를 사용하였다. 식 (3)에서 $\mu=(\mu_1,\mu_2,\mu_3,\cdots,\mu_n)$는 MFCC의 평균벡터이고 Σ^{-1} 는 $\boldsymbol{x}=(x_1,x_2,x_3,\cdots,x_n)$의 공분산 행렬의 역행렬을 의미한다.

$$D_M(x)=\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \quad (3)$$

한편 이 단계에서도 분류의 타당성을 높이기 위한 후처리 과정을 수행한다. 즉 오디오 신호를 카테고리 별로 분류한 결과, 주변의 카테고리와 어울리지 않고 고립된 카테고리로 분류된 프레임이 있을 경우 이를 변경해주는 작업이 필요하다. 본 논문에서는 이를 위해 메디안 필터를 사용하였다.

3. 움직임 활동도와의 매칭

움직임 정보는 앵커 장면과 보조설명 장면을 구별하기 위해 매우 유용한 특징이다. 앵커 장면내의 수평 방향 및 수직 방향 움직임 벡터는 일반적으로 낮은 움직임 활동도를 보이

는 반면 보조설명 장면에서의 움직임 벡터는 앵커 장면에 비해 높은 움직임 활동도를 보이기 때문이다. 앞서 살펴보았듯이 움직임 활동도의 정도는 움직임 벡터 크기의 표준편차로 나타낼 수 있으며, MPEG-7 표준에서와 같이 활동도를 5 단계로 구분하여 가장 낮은 정도를 1로 가장 높은 정도를 5로 나타낸다. 이러한 움직임 특징을 이전 단계에서 얻어진 분류 결과와 함께 고려한다. 즉 두 번째 단계에서 특정 프레임이 앵커 음성에 해당하는 카테고리 1으로 분류되었다고 하더라도 움직임 활동도가 1보다 크면 앵커 장면이 아니라고 간주한다. 이와 같이 오디오 특징과 비디오 특징을 결합하여 사용함으로써 앵커 장면 검출의 정확도를 향상시키고 동영상 색인 알고리듬의 성능을 개선할 수 있다.

한편 뉴스의 제작 기법이 변하면서 그림 6의 (a)에서 보인 바와 같이 화면의 하단에 자막을 넣는 경우가 많이 발생하고 있다. 그리고 이 자막 영역 내의 텍스트는 오른쪽에서 왼쪽으로 이동하면서 그림 6의 (b)에 나타낸 바와 같이 상당히 큰 수평방향의 움직임을 보이는 것이 일반적이다. 따라서 만일 움직임 활동도를 구할 때 이러한 자막 영역을 포함하게 되면 실제로는 앵커 장면임에도 불구하고 자막의 수평 움직임 때문에 제대로 검출되지 않을 가능성이 있다. 이러한 현상을 방지하기 위해서 움직임 활동도를 계산하는 영역을 제한할 필요가 있다. 실제로 방송되는 뉴스 동영상의 경우를 살펴보면, 자막 영역의 높이가 고정되기 보다는 방송국마다 그리고 특정한 뉴스 프로그램 마다 달라진다. 제안하는 알고리듬에서는 자막 영역이 전체 화면의 20%를 차지한다고 가정하고 나머지 80%의 영역에서만 움직임 활동도를 계산한다.

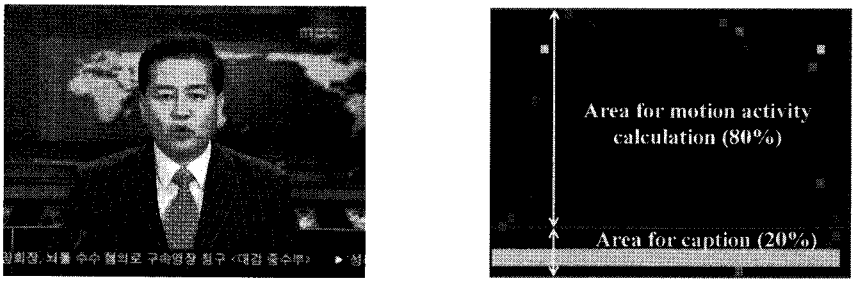
IV. 실험 결과

제안한 알고리듬의 성능을 평가하기 위해 한국방송공사의 뉴스 프로그램을 실험 대상으로 선택하였다. 선택한 뉴스 동영상은 디지털 방송 프로그램으로서 MPEG-2 비트스트림으로 부호화되었으며 총 방송시간은 약 60분 분량이다. 뉴스 동영상의 오디오 신호의 특징들을 계산하기에 위해서 본격적인 실험에 앞서서 먼저 스테레오 또는 5.1 채널로 주어지는 디지털 방송 프로그램의 오디오 신호를 단일 채널로 만들 필요가 있다. 따라서 스테레오 신호의 경우에는 좌, 우 2 채널의 오디오 신호를 평균하여 단일 채널의 오디오 신호를 얻고, 5.1 채널 신호의 경우에는 별도의 신호처리과정 없이 중앙 채널만을 사용할 수 있다.

제안 알고리듬의 첫 번째 단계 및 두 번째 단계의 실험에서 사용한 파라미터를 다음의 표 1에 정리하였다.

표 1. 실험 파라미터.
Table 1. The simulation parameters.

Parameter	Value (unit)
The size of window, W_{size}	40 (ms)
The hopping time	20 (ms)
The FFT size, M	2048
The frequency range	50 - 16,000 (Hz)
The number of filter bank	26
The number of used MFCC	20
The size of median filter	100 (ms)



(a) (b)

그림 6. 자막 영역내 수평방향 움직임 벡터의 분포. (a) 자막이 존재하는 앵커 장면, (b) 자막 영역 내 수평 움직임성분
Fig. 6. The horizontal motion activity in caption area. (a) A anchor shot with caption, (b) H motion vector component in the caption area

그림 7은 실험에서 사용한 뉴스 동영상의 일부분에 대한 실험결과를 나타낸다. 스테레오 오디오 신호를 평균하여 얻은 오디오 파형을 해당 뉴스 장면과 함께 그림 7 (a)에 나타내었다. 시간상으로 519초 부근에서 앵커가 등장하기 직전에 그리고 530초 부근에서 앵커 장면에 이어지는 보조 설명 장면의 시작하는 부분에 오디오 신호의 정지구간이 있음을 볼 수 있다. 이어서 오디오 정지구간을 검출한 결과를 그림 7 (b)에 나타내었다. 그림 7 (b)에서 출력값 1은 'silence'를, 출력값 0은 'non-silence'를 의미한다. 그림 7 (b)를 통해 2개의 'silence' 구간이 검출되며 이 결과가 그림 7 (a)에서 오디오 신호가 존재하지 않는 구간과 일치함을 확인 할 수 있다. 오디오 신호의 특성상 여러 개의 짧은 정

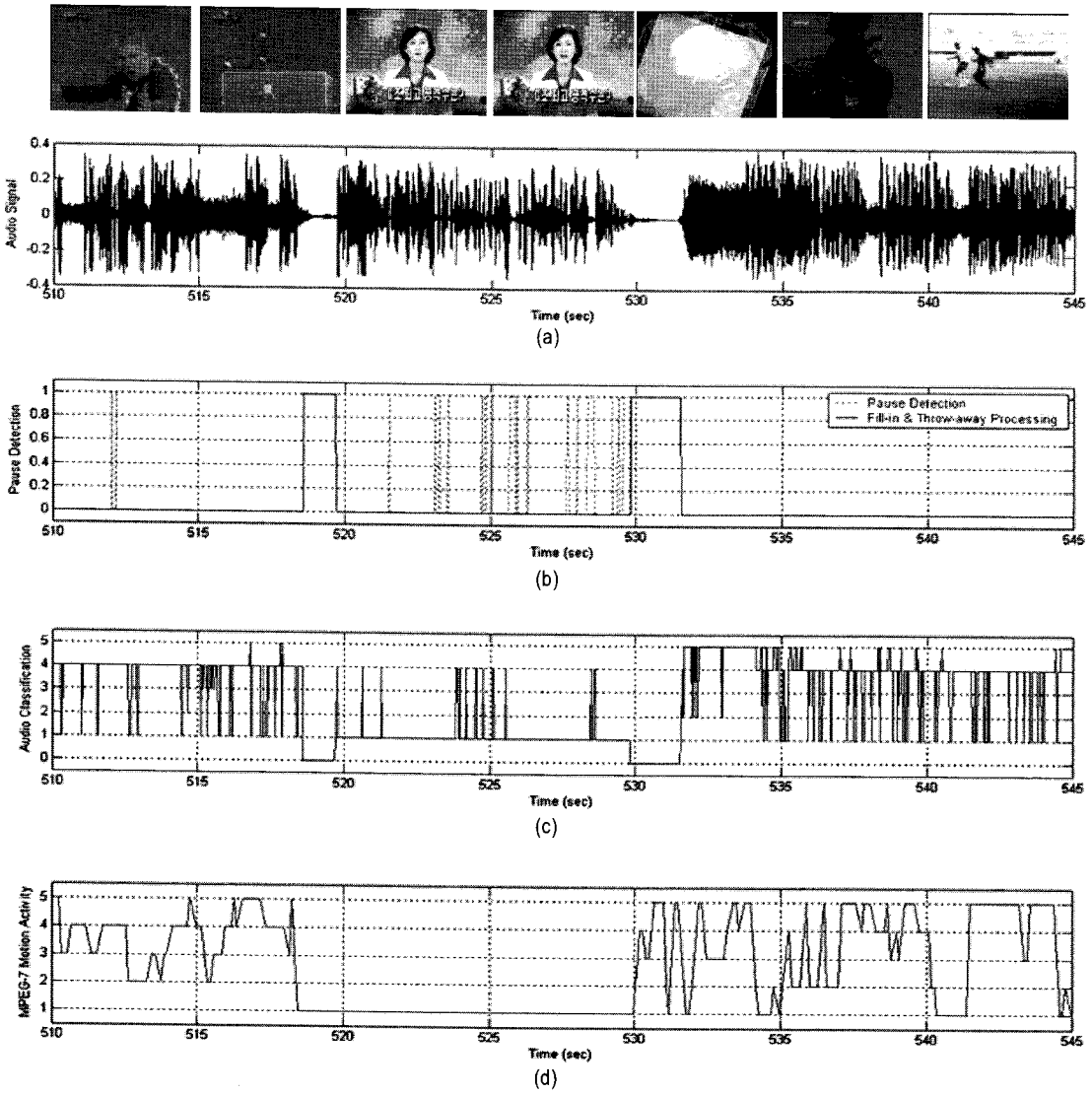


그림 7. 실험 결과. (a) 오디오 신호 파형, (b) 정지 구간 검출 결과, (c) 오디오 클러스터 분류 결과, (d) 움직임 활동도

Fig.7. The simulation results. (a) the waveform of audio signal, (b) the result of pause detection stage, (c) the result of audio cluster classification stage, (d) the motion activity

지 구간이 나타났지만 후처리 작업을 통해서 오디오 신호가 존재하는 구간으로 간주되었다.

오디오 클러스터 분류 단계에서는 신호가 존재하는 구간 내에서 각 프레임의 MFCC를 계산하여 각 프레임을 다섯 개의 카테고리로 분류하였다. 이 가운데 '앵커 음성' 카테고리는 출력값 1로 나타난다. 클러스터 분류 이후의 후처리 과정으로서 카테고리 분류 결과의 오류 가능성을 줄이기 위해 메디안 필터를 적용하였다. 그림 7 (c)에서 이를 나타내었는데 그림에서 나타난 뉴스 동영상의 앞부분과 뒷부분에서는 카테고리 분류의 결과가 빠르게 변화하고, 519초에서 530사이의 상대적으로 긴 구간에서 대부분 '앵커 음성'으로 분류되었음을 알 수 있다. 따라서 이 사이의 프레임들이 'non-silence' 구간에 해당하면서 '앵커 음성'으로 분류되었으므로 앵커 장면에 해당한다. 실험에서는 하나의 클러스터 내에 앵커 프레임으로 분류된 비율이 90%가 넘는 경우에 이를 앵커 장면으로 검출하였다.

앞서 언급했듯이 검출 알고리듬의 성능은 오디오 특징뿐 만 아니라 움직임 활동도와 같은 비디오 특징을 결합함으로써 향상될 수 있다. 뉴스 동영상 내에는 앵커가 말하고 있는 동안에도 앵커를 보여주는 대신에 자료화면을 내보내는 경우가 있기 때문이다. 그림 7 (d)에서는 움직임 활동도를 계산한 결과를 나타내었다. 앞서 설명하였듯이 자막 영역을 제외한 움직임 활동도가 앵커 장면에서 낮게 나타남을 확인할 수 있다. 그림 8에서 오디오 특징만으로는 앵커 장면으로 판단되었으나 움직임 활동도를 결합함으로써 보조설명 장면으로 재분류된 예를 나타내었다.

제안한 알고리듬을 실험 대상 뉴스 동영상 전체에 대해 적용하여 성능을 살펴보았다. 동영상내 앵커 장면의 개수는 총 70개이다. 오디오 특징만을 사용한 경우에는 71개의 장면을 앵커장면으로 판정하였고 이 가운데 68개의 장면을 제대로 검출하였으며 3개 장면에서는 잘못된 검출 결과를 보였다. 따라서 recall 및 precision은 다음과 같이 각각 97.14% 및 95.77%를 나타내었다.

$$\text{Recall} = \text{correct}/(\text{correct} + \text{missed}) = 68/(68+2) = 97.14\ \%$$
$$\text{Precision} = \text{correct}/(\text{correct} + \text{incorrect}) = 68/(68+3) = 95.77\ \%$$

한편 움직임 정보까지 결합한 경우에는 68개 앵커 장면은 모두 그대로 판정하였고 앞서 잘못 검출된 3개 장면 가운데 2개 장면을 앵커 장면이 아닌 것으로 수정하였다. 따라서 recall은 동일하지만 precision은 98.55%로 향상되었다.

$$\text{Recall} = \text{correct}/(\text{correct} + \text{missed}) = 68/(68+2) = 97.14\ \%$$
$$\text{Precision} = \text{correct}/(\text{correct} + \text{incorrect}) = 68/(68+1) = 98.55\ \%$$

실험 결과에서 오류가 발생한 경우를 분석해보면, 첫 번째 단계에서의 대부분 오류는 앵커의 목소리가 이전 클러스터에 섞여 있을 때 발생하였다. 또한 뉴스 프로그램의 시작부분이나 마지막 부분에서 앵커가 말하면서 배경 음악이 함께 나오는 경우가 있는데 이 부분에서도 두 번째 분류단계의 결과에 오류가 나타나는 현상이 발생하였다.

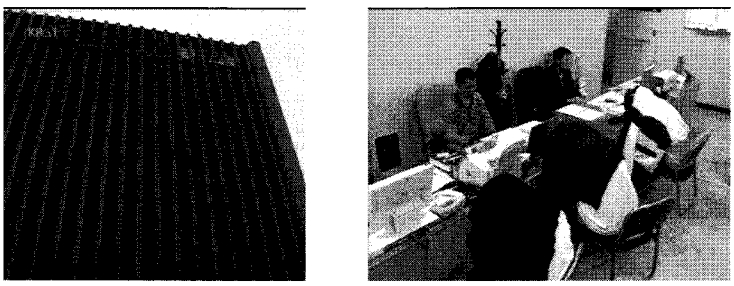


그림 8. 움직임 활동도 특징을 사용함으로써 재분류된 장면 예.
Fig. 8. The example pictures filtered out by use of the motion activity.

V. 결 론

본 논문에서는 뉴스 동영상에서 앵커 장면을 검출하는 알고리듬을 제안하였다. 뉴스 동영상에 대한 관찰을 통해 앵커 장면 검출에 적합한 오디오 및 비디오 특징들을 선택하였다. 오디오 특징 가운데 평균에너지 및 주파수 성분차이는 주어진 동영상을 오디오 신호가 존재하는 구간과 존재하지 않는 구간으로 구분하는 데에 사용되었고, MFCC 특징은 오디오 신호가 존재하는 프레임을 5개의 카테고리 가운데 하나로 세분화하는 데에 사용되었다. 또한 검출 성능의 향상을 위해 움직임 활동도라는 비디오 특징을 오디오 특징들과 결합하여 사용하였다. 실제 방송된 뉴스 동영상에 대한 실험을 통해 제안 알고리듬이 앵커 장면과 보조설명 장면을 성공적으로 구분함을 확인하였고 뉴스 동영상의 분석 및 색인 작업에 효과적으로 활용될 수 있음을 알 수 있었다.

뉴스 제작에서 시각적인 효과가 강조되면서 다양한 화면 구성이 시도되고 있다는 측면에서 뉴스 동영상의 색인을 위해서는 본 논문에서와 같이 오디오 특징을 비디오 특징과 결합하여 사용하는 편이 바람직하다. 한편 뉴스 동영상의 경우에도 끊임없이 새로운 양식이 시도되고 있기 때문에 특정 프로그램에 대한 색인 기법을 개발하기 위해서는 어떤 특징을 선택할 것인가가 중요한 문제가 된다. 그러나 앵커가 뉴스를 진행하는 현재의 방식에서는 오디오적 특성은 크게 변하지 않음을 고려할 때, 기본적으로는 본 논문에서 사용한 오디오 신호의 특징을 사용하면서 특정한 화면 구성에 따라 적합한 비디오 특징을 결합하는 것이 바람직하다고 판단된다.

참고문헌

[1] C.G.M. Snoek and M. Worring, "Multimodal Video Indexing: A Review of the State-of-the-art," Multimedia Tools and Applications, vol.25, no.1, pp.5-35, 2005.

[2] W. Qi, L. Gu. H. Jiang, X.-R. Chen, and H.-J. Zhang, "Integrating Visual, Audio and Text Analysis for News Video," Proc. IEEE International Conference on Image Processing, vol.3, pp.520-523, 2000.

[3] W. Hsu, L. Kennedy, C-W. Huang, S.-F. Chang, C.-Y. Lin, and G. Iyengar, "News Video Story Segmentation using Fusion of Multi-level Multi-modal Features in TRECVID 2003," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol.3, pp.645-648, 2004.

[4] L. Chaisorn, T.-S. Chua, and C.-H. Lee, "A Multi-Modal Approach to Story Segmentation for News Video," World Wide Web, vol. 6, no.2, pp.187-208, 2003.

[5] X. Wu, C.-W. Ngo, and Q. Li, "Threading and Autodocumenting News Videos," IEEE Signal Processing Magazine, vol.23, no.3, pp.59-68, 2006.

[6] S. Quadri, S. Krishnan, and L. Guan, "Indexing of NFL Video using MPEG 7 Descriptors and MFCC Features," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol.2, pp.429-432, 2005.

[7] P. Salembier B.S. Manjunath and T. Sikora, "Introduction to MPEG 7: Multimedia Content Description Interface," John Wiley and Sons, England, UK, 2002.

[8] D. Li, I.K. Sethi, N. Dimitrova, and T. McGee, "Classification of General Audio Data for Content based Retrieval," Pattern Recognition Letters, vol.22, no.5, pp.533-544, 2005.

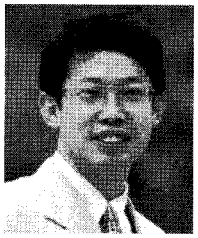
[9] I.K. Sethi, and G.P.R. Sarvarayudu, "Hierarchical Classifier Design using Mutual Information," IEEE Transactions on Pattern Recognition Machine Intelligence, vol. 4, no.4, pp.441-445, 1982.

저 자 소 개

유 성 열

· 2006년 2월 : 한동대학교 전산전자공학부 졸업
· 2006년 3월~현재 : 국민대학교 대학원 전자공학과 석사과정
· 주관심분야 : MPEG-7, Video Retrieval, SVC(Scalable Video Coding)

저 자 소 개

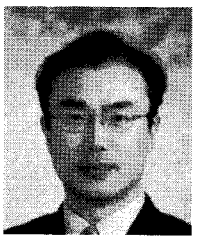


강 동 욱

- 1986년 2월 : 서울대학교 전자공학 졸업
- 1988년 2월 : 서울대학교 전자공학과 석사
- 1995년 2월 : 서울대학교 전자공학과 박사
- 2000년 9월~2001년 8월 : Lucent Technology MTS
- 1995년~현재 : 국민대학교 전자공학부 교수
- 주관심분야 : 비디오 코딩, 영상통신

김 기 두

- 1980년 2월 : 서강대학교 전자공학과 졸업
- 1988년 8월 : The Pennsylvania State University, MS(Electrical Eng.)
- 1990년 12월 : The Pennsylvania State University, MS(Electrical Eng.)
- 1980년 3월~1985년 12월 : 국방과학연구소 연구원
- 1998년 3월~1999년 2월 : 미국 UCSD, Visiting Scholar
- 1991년 3월~현재 : 국민대학교 전자공학부 교수
- 주관심분야 : 디지털통신, 디지털신호처리



정 경 훈

- 1987년 2월 : 서울대학교 전자공학 졸업
- 1989년 2월 : 서울대학교 전자공학과 석사
- 1996년 2월 : 서울대학교 전자공학과 박사
- 1991년 12월~1997년 2월 : 한국영상산업진흥원 선임연구원
- 1997년 3월~2005년 2월 : 한동대학교 전산전자공학부 교수
- 2005년 3월~현재: 국민대학교 전자공학부 교수
- 주관심분야 : 멀티미디어신호처리, 디지털 방송