# 멀티모달 특징을 이용한 비디오 장르 분류

*진성호, *배태면, *추진호, *노용만, **강경옥
*한국정보통신대학교 공학부 멀티미디어 그룹
**한국전자통신연구원, 전파방송연구소
wh966@icu.ac.kr

# Video genre classification using Multimodal features

*Sung Ho Jin, *Tea Meon Bea, *Jin Ho Choo, *Yong Man Ro, and **Kyeongok Kang
*Multimedia Group, Information and Communications University (ICU)
**Electronics and Telecommunications Research Institute (ETRI)

## 요약

본 논문에서는 멀티모달(multimodal) 특징을 이용한 비디오 장르 식별 방법을 제안한다. 비디오 장르 식별 기술은 방대한 양의 방송 컨텐츠를 보다 효율적으로 분류할 뿐 아니라 자동적인 비디오 요약을 위한 전처리 과정으로 활용될 수 있는 기술이다. 따라서, 그 필요성 및 중요성이 부각되고 있다. 본 논문에서 제안하고 있는 방법은 MPEG-7의 오디오 및 비주얼 서술자들을 적용하여 멀티모달 특징을 추출하고 여러 가지 방송 비디오 장르(genre)들로 구성된 데이터베이스에서 장르 분류를 위해 설계된 인식기(classifier)를 통한 성능을 평가한다.

## 1. 서론

디지털 방송이 소개된 이후 방송 서비스들은 컨텐츠 제공자들에 의해 시청자들에게 개인화 된 컨텐츠를 제공받을 수 있는 양방향 방송을 지향하고 있다. 방송 채널 및 컨텐츠의 양이 증가함에 따라 컨텐츠 제공자들은 시청자들의 기호에 적합한 컨텐츠를 제공하기 위해 그들의 방대한 컨텐츠를 일정한 기준에 의해서 분류하고 조작할 필요가 있다. 즉, 방송 컨텐츠를 여러 가지 장르 및 카테고리로 자동적으로 분류하는 것은 방대한 방송 컨텐츠들을 가진 데이터베이스를 효율적으로 분류하고 관리할 수 있는 기능을 제공한다. 또한, 장르 식별은 컨텐츠 저작도구의 전처리 기술로 활용될 수 있다. 비디오 요약 기술은 장르에 따라 서로 다른 알고리즘을 사용함에 따라 입력부에서 조작자에 의한 결정이나 저작 도구 자체에서 장르를 인식해야할 필요성을 가진다. 현재까지 비디오 요약이나 조작 기술들은 수동적인 장르 결정을 하고 있다[1]. 더불어, 현재까지의 저작도구들은 MPEG-7의 서술자들과 MDS (Multimedia Description Scheme)을 이용하는 추세이다.

지금까지 비디오 컨텐츠에서 하이레벨 의미들 추출하기 위한 여러 연구들이 진행되어 왔다. 대부분의 기존 연구들은 오디오나 비주얼 특징중 하나만을 사용하고 있다 [2-4]. 일반적으로 영점 교차, 비정적 비율 (non-silence ratio), 주파수 중심(centroid) 또는 영역(bandwidth), 칼라 및 모션 정보들을 이용하고 있다. 또한 중간레벨의 특징인 샷 길이 정보를 함께 사용하고 있는 연구들도 있다. 최근에는 비디오 분석을 위해 오디오 및 비주얼 정보를 함께 적용하는 연구들이 진행되고 있다. 일반적으로 멀티모달 특징들을 사용하는 방법들은 유니모달(unimodal)을 사용하는 방법들에 비해 더 좋은 결과들을 보여주고 있다[1] [5] [6]. 이와 같은 비디오 분석 연구들에서 중요한 요건 중의 하나가 처리 속도이다. 대부분의 기존 연구들에서 처리속도에 대해서는 아직까지 선행 과제로 남아 있다.

따라서, 본 논문에서는 다음과 같은 3가지 측면들에 대해 중점을 두고 있다. 1) 높은 분류 성능, 2) 빠른 처리속도, 3) 그리고 기존의 MPEG-7 프레임 웍과 일관성를 유지할 수 있는 멀티모달 특징들의 사용에 중점을 둔다. 제안하는 방법의 유효성을 검증하기 위해서 카툰(cartoon), 드라마, 뮤직비디오, 뉴스 그리고 스포츠로 구성된 장르들을 가진 MPEG-1 비디오 파일들을 가지고 실험한다.

## 2. 제안하는 방법

먼저, 하나의 비디오 컨텐츠로부터 60초 길이를 가지는 짧은 비디오 클립을 랜덤하게 추출한다. 짧은 길이의 입력 데이터는 비디오 분석의 처리시간을 빠르고 효율적으로 한다. 하지만, 비디오 길이가 짧아질수록 규칙기반의 분석(rule-based analysis)가 어렵다. 예를 들면, 장르들 중에서 가장 간단한 규칙을 가지는 뉴스를 임의적으로 60초 길이로 자를 경우 가지게 되는 구조는 앵커, 뉴스기사, 앵커+뉴스기사, 뉴스기사+앵커, 앵커+뉴스기사+앵커, 뉴스기사+앵커+뉴스기사 등과 같은 형태를 가진다. 뉴스에 비해 더욱 복잡한 구조를 가지는 그밖의 장르들은 랜덤하게 추출된 비디오 클립에서 더욱 더 복잡한 구조를 가지게 된다. 따라서, 본 논문에서는 확률기반 분석(statistical-based analysis)을 통해 비디오를 분석한다.

제안하는 방법에서 사용하는 특징들은 중간레벨의 특징값인 샷(shot) 개수와 로우레벨의 특징값인 오디오 및 비주얼 특징값을 사용하고 있다. 특징값, $F_set = \{F_1, F_2, F_3, \cdots, F_{15}\}$들은 1개의 중간레벨 특징과 14개의 로우레벨 특징들로 구성된다. 특징 추출을 위해 사용되는 MPEG-7 서술자들 중 Motion Activity와 Camera Motion은 추출된 샷 정보(샷 길이 및 샷 개수)에 기반하여 특징값들을 추출하게 된다. HSV 칼라 정보는 비디오 클립에서 일정한 간격마다 추출하여 칼라 흐름의 변화를 측정한다. 오디오 정보들은 각각 MPEG-7에서 추천하는 오디오 프레임 간격마다 추출하게 된다. 서술자들의 특징값들을 추출한 후 비디오 클립의 특성을 대표할 수 있는 특징셋으로 구성한다. 특징 추출에 관한 내용은 다음 절에서 자세히 설명한다. 구성된 특징셋을 가지고 장르를 분류하기 위해서 본 논문에서는 Gini 인덱스를 가지는 CART를 이용한 결정 트리 기반의 인식기를 사용한다 [7]. 제안하는 방법은 그림 1에서 보여준다.

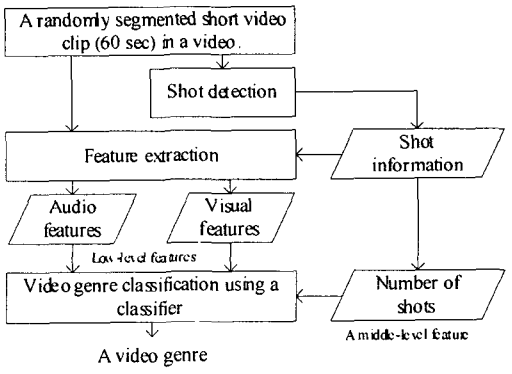


그림 1. 제안하는 장르 분류 방법의 흐름도

## 3. 샷 검출

비디오 분석에서 샷 길이 및 개수는 샷 템포를 나타내는 유용한 특징이 된다 [2]. 이 특징들은 로우레벨의 특징들에 의해 추출되는 중간레벨의 특징이 된다.

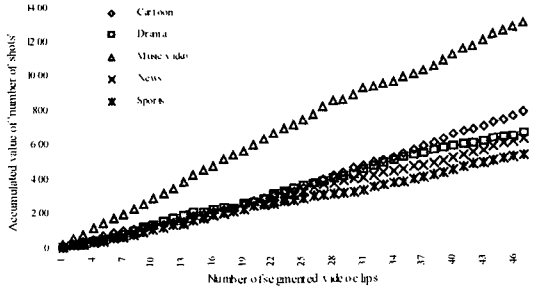


그림 2. 비디오 클립 수에 따른 축적된 $F_2$

비디오 분석을 통해 여러가지 편집효과들에 의해 생성된 대부분의 뮤직비디오들은 짧은 길이의 샷들을 가지고 있음을 알 수 있다. 그와 반대로 스포츠와 뉴스 장르들은 긴 샷들을 가지고 있다. 그림 2는 비디오 클립의 수가 증가함에 따라 축적된 샷 개수, 즉, $F_1$의 값들을 보여준다. 샷 길이 및 개수는 본 논문에서 제안하는 샷 검출 알고리즘에 의해 추출된다. 샷 검출 알고리즘은 MPEG-7 비주얼 서술자들 중 HT(Homogeneous Texture), SC(Scalable Color) 그리고 EH(Edge Histogram)을 사용하여 생성된다. 식 (1)은 샷 변이를 결정하는 방법을 보여준다.

$$Shot\_Transition_i = \begin{cases} True, if\ Frame_i\_diff > Th\_1\ and \\ \quad |Frame_{i-1}\_diff - Frame_i\_diff| > Th\_2 \\ False, Otherwise \end{cases} \quad (1)$$

여기서, $Frame_i\_diff$ 은 (i-10)번째 프레임과 i 번째 프레임 사이의 프레임 차이이다. 그리고, $Th\_1$ 과 $Th\_2$ 는 경험적인 임계값들이다. 제안하는 샷 검출 알고리즘은, 현재 비디오 프레임과 10번째 앞의 프레임을 계산하도록 10 프레임의 간격을 갖는 마스크를 사용하여 프레임간의 차이를 계산한다. $Shot_Transition_i$이 $True$값을 가지면, i번째 프레임에서 샷 변이를 검출된다. $Frame_i\_diff$ 은 다음의 식들을 통해 계산된다.

$$Frame_i\_diff = Diff\_SC_i + Diff\_EH_i + Diff\_HT_i \quad (2)$$

$$Diff\_SC_i = (SC\_frame_{i-10} - SC\_frame_i)^2 / Max\_Diff\_SC \quad (3)$$

$$Diff\_EH_i = (EH\_frame_{i-10} - EH\_frame_i)^2 / Max\_Diff\_EH \quad (4)$$

$$Diff\_HT_i = (HT\_frame_{i-10} - HT\_frame_i)^2 / Max\_Diff\_HT \quad (5)$$

여기서 $Diff\_SC_i$, $Diff_EH_i$과 $Diff_HT_i$은 (i-10)번째 프레임과 i프레임에서 각각 SC, EH와 HT의 정규화 된 차이를 나타낸다. 그리고, $SC\_frame_i$, $EH_frame_i$과 $HT\_frame_i$ 은 i번째 프레임에서 각 서술자들의 값들을 말한다. Max_Diff_SC, Max_Diff_EH과 Max_Diff_HT 은 $Diff\_SC_i$, $Diff_EH_i$과 $Diff_HT_i$의 최대값이다. 샷 변이가 추출되면 $F_1$은 비디오 클립에서 나타나는 샷 변이의 빈도수로 정의된다.

## 4. 특징 추출

### 4.1 오디오 특징들

스포츠, 액션, 대화 및 앵커 장면등 여러 장면들은 각각의 고유한 오디오 특성을 가지고 있다. 예를 들면, 드라마의 대화 장면의 경우 낮은 오디오 파워, 플랫한 오디오 스펙트럼 및 낮은 하모닉 중심값(Harmonic centroid)를 보여준다. 따라서, 장르의 특성을 측정하기 위해 오디오 특징들은 도움을 주는 요소임을 확인할 수 있다. 본 논문에서 제안하는 오디오 특징들은 오디오 프레임 전체 개수인 $N$을 기준으로 추출된다.

#### 4.1.1 AudioPower

뮤직비디오는 배경음악 및 가수의 노래소리에 의해 다른 장르에 비해 높은 오디오 파워를 가진다. 반면, 스포츠는 가끔씩 순간적으로 높은 오디오 파워를 보이지만 비디오 클립 전체를 보면 뮤직비디오에 비해 낮은 값을 보인다. 제안하는 방법에서는 AudioPower 서술자의 평균값을 $F_2$라고 나타내며, 이 값은 비디오 클립에서 얻어지는 AudioPower의 전체 합에 의해서 계산 될 수 있다. $F_2$는 다음의 식을 통해서 구할 수 있다.

$$F_2 \underset{=}{\Delta} Mean\_AP = \frac{1}{N}\sum_{i=1}^{N}\sum_{n=1}^{m}|s(n)_i|^2 \quad (6)$$

여기서 $s(n)$ 는 오디오 신호를 가르킨다. $n$ 은 홉사이즈(hopsize)에 해당되는 시간 간격에서 샘플 수를 나타낸다.

#### 4.1.2 FundamentalFrequency

기본 주파수(Fundamental frequency)는 음악의 피치와 음성 억양의 좋은 예측자가 되는 오디오 신호의 기본 정보이다[9]. 그림 3에서 보여지는 것과 같이 FundamentalFrequency 서술자의 평균값은 뮤직비디오에서 높은 값을 가지고 뉴스에서 낮은 값을 가진다. 나머지 장르들은 유사한 값들을 가진다. FundamentalFrequency의 평균, 즉, $F_3$ 은 다음과 같다.

$$F_3 \underset{=}{\Delta} Mean\_FF = \frac{1}{N}\sum_{i=1}^{N} FF_i \quad (7)$$

여기서, $FF_i$ 는 i 번째 프레임에서 Fundamental Frequency 서술자의 값이다.

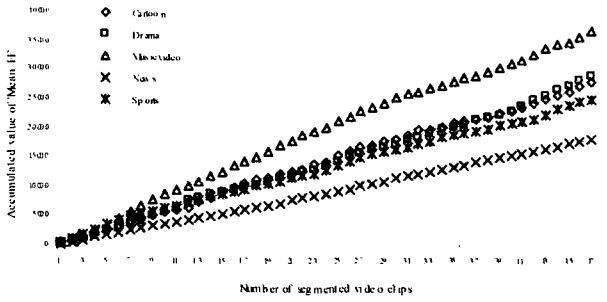


그림 3. 비디오 클립 수에 따른 축적된 $F_3$

#### 4.1.3 HarmonicSpectralCentroid

오디오 신호의 하모닉 정보들은 하모닉 스펙트럼을 갖는 사운드들 (음악 사운드, 유성음의 음성 등)과 하모닉 성분을 갖지 않는 사운드들 (노이즈, 무성음의 음성, 기구들의 집적된 혼합음등)을 구별해 준다. MPEG-7의 HarmonicSpectralCentroid 서술자는 사운드 세그번트(Segment)내에서의 순간 (instantaneous) HarmonicSpectralCetnroid의 평균이 된다[9]. 본 논문 에서는 HarmonicSpectralCentroid의 평균값을 $F_4$ 로 이용한다. 이 특징을 이용함으로써 뮤직 사운드와 앵커의 목소리등을 구별해 준다. $F_4$는 다음과 같다.

$$F_4 \underset{=}{\Delta} Mean\_HSC = \frac{1}{N}\sum_{i=1}^{N} HSC_i \quad (8)$$

여기서, $HSC_i$ 는 i번째 오디오 프레임에서 Harmonic SpectralCentroid 의 값이다.

#### 4.1.1 HarmonicSpectralDeviation

HarmonicSpectralDeviation 서술자는 글로벌 스펙트럼 편차와 로그진폭(log-amplitude)의 스펙트럼의 편차이다[9]. 이 서술자의 평균, 즉, $F_5$ 를 제안하는 방법에 적용한다. 이 특징들은 실험 장르들을 3개의 클래스로 구분한다. 뮤직비디오는 여러 개의 음악사운드의 영향으로 가장 큰 값을 갖는다. 반대로, 드라마와 뉴스는 하모닉 피크와 로컬 스펙트럼 엔빌롭(Envelope)사이의 작은 차이를 보이는 사람의 목소리 특성에 의하여 낮은 값을 보여준다. 나머지 장르들은 중간 클래스를 형성한다. $F_5$는 다음과 같다.

$$F_5 \underset{=}{\Delta} Mean\_HSD = \frac{1}{N}\sum_{i=1}^{N} HSD_i \quad (9)$$

여기서, $HSD_i$는 i번째 오디오 프레임에서의 Harmonic SpectralDeviation값이다.

#### 4.1.1 AudioSpectrumFlatness

AudioSpectrumFlatness 서술자는 주어진 주파수 밴드 내에서 오디오 신호의 숏 텀(Shot-term) 파워 스펙트럼의 Flatness 특성을 나타낸다[9]. 이 서술자의 평균, 즉, $F_6$은 스포츠와 카툰을 구별하는 데 사용가능하다. 다음식은 $F_6$을 나타낸다.

$$F_6 \underset{=}{\Delta} Mean\_ASF = \frac{1}{N}\sum_{i=1}^{N} ASF_i \quad (10)$$

여기서, $ASF_i$는 i번째 오디오 프레임에서 AudioSpectrumFlatness의 평균값이다.

### 4.2 비주얼 특징들

비디오 분석에서 비주얼 정보는 매우 유용한 단서를 제공한다. 비디오는 모션의 양, 야외와 실내에 따른 칼라정보, 카메라의 움직임등과 같은 비주얼 특징을 가지고 있다. 따라서, 비디오 장르에 따라 모션 활동량, 카메라 모션 그리고 칼라정보들이 대표적인 특징들이 된다. 제안하

는 방법에서는 MPEG-7 비주얼 서술자들로부터 기반한 비주얼 특징들을 사용한다.

4.2.1 Motion Activity

Motion Activity 서술자는 모션의 양이 작은 드라마와 모션의 양이 큰 뮤직비디오를 구분하는 데 유용하다. 서술자의 파라미터중 샷 활동성 히스토그램(shot activity histogram)을 나타내는 TemporalParameter는 전체 비디오 클립의 특성을 나타내는 데 유용하다[8]. 이 특징을 통해 일반적으로 뮤직비디오와 스포츠는 모션의 강도가 크고 다른 장르들은 몇몇 장면을 제외하고 낮은 강도의 모션을 가지고 있는 것을 보여준다. $F_7$ 은 다음과 같이 표현된다.

$$F_7 \triangleq Mean\_Centroid\_TP = \frac{1}{N}\sum_{i=1}^{N} Centroid\_TemporalParameter_i \quad (11)$$

$$Centroid\_TemporalParameter_i = \frac{\sum_{activity\_level=1}^{5}(activity\_level \times TemporalParameter[activity\_level])}{\sum_{activity\_level=1}^{5} TemporalParameter[activity\_level]} \quad (12)$$

여기서, $N$은 비디오 클립에서 전체 샷 개수 이다. $Centroid\_TemporalParameter_i$, $activity\_level$ 그리고 $TemporalParameter_i$ 는 각각 i번째 샷에서의 $Centroid\_TemporalParameter$, 개별적인 프레임들의 모션 활동량(motion acivity)의 레벨 및 TemporalParameter의 값들이다.

4.2.2 Camera Motion

Camera Motion 서술자들은 3-D 카메라 모션의 파라미터들을 나타낸다 [8]. 제안하는 방법은 이 파라미터들중 몇가지를 이용한다. 고정(fixed), 팬 (pan), 틸트(tilt), 줌(zoom), 그리고, 카메라 모션의 합이 이용된다. 특히, 팬 성분은 스포츠 장르를 구별하는 탁월한 특징이된다. 그 이유는 대부분의 스포츠 경기는 카메라가 좌우로 움직이는 모션을 많이 포함하고 있기 때문이다. 반면에, 카툰은 만들어지는 장면인 관계로 카메라 모션의 양이 매우 작다.

$$F_8 \triangleq Mean\_Panning = \frac{1}{N}\sum_{i=1}^{N}(Panning\_left_i + Panning\_right_i) \quad (13)$$

$$F_9 \triangleq Mean\_Tilting = \frac{1}{N}\sum_{i=1}^{N}(Tilting\_left_i + Tilting\_right_i) \quad (14)$$

$$F_{10} \triangleq Mean\_Zooming = \frac{1}{N}\sum_{i=1}^{N}(Zooming\_left_i + Zooming\_right_i) \quad (15)$$

$$F_{11} \triangleq Mean\_Fixed = \frac{1}{N}\sum_{i=1}^{N} Fixed\_Cameramotion_i \quad (16)$$

$$F_{12} \triangleq Sum\_CM = Mean\_Panning + Mean\_Tilting + Mean\_Zooming \quad (17)$$

여기서, $N$은 비디오 클립에서 전체 샷 개수이고 $Panning\_left_i$, $Panning\_right_i$, $Tilting\_left_i$, $Tilting\_right_i$, $Zooming\_left_i$, $Zooming_right_i$과 $Fixed_Cameramotion_i$들은 i번째 샷에서의 Camera Motion 서술자의 각각 컴포넌트들의 값들이다.

4.2.3 HSV 칼라

Scalable Color 서술자는 Haar 변환에 의한 인코딩되는 HSV 칼라 공간에서의 칼라 히스토그램이다[9]. 이 서술자는 장르에서 칼라의 변화 및 우세한 칼라를 검출하는 데 유용하다. 예를 들어, 골프나 축구의 녹색의 칼라 변화는 뮤직비디오의 칼라 변화에 비해서 훨씬 느리다. HSV 칼라 공간은 장르에 따라 2가지의 유용한 특성을 보인다. 대개 카툰의 평균 밝기(Brightness)는 다른 장르에 비해서 매우 높으며, 야외의 조명을 활용하는 스포츠의 경우 평균 채도 (Saturation)는 드라마나 뉴스와 비교했을 때 높은 수치를 보인다. 뮤직비디오의 경우 여러 편집 효과에 의해 칼라정보의 변동이 크다. 다음 식들에서 보여지는 것처럼 본 논문에서는 Scalable Color 서술자에서 추출된 Hue, Saturation과 Value의 값들의 평균값을 각각 $F_{13}$, $F_{14}$과 $F_{15}$으로 사용한다.

$$F_{13} \triangleq Mean\_Hue = \frac{1}{M}\sum_{i=1}^{M} Hue_i \quad (18)$$

$$F_{14} \triangleq Mean\_Saturation = \frac{1}{M}\sum_{i=1}^{M} Saturation_i \quad (19)$$

$$F_{15} \triangleq Mean\_Value = \frac{1}{M}\sum_{i=1}^{M} Value_i \quad (20)$$

$$M = N / frame\_duration$$

여기서, $N$ 은 비디오 클립내의 전체 샷 개수이다. $frame_duration$는 칼라 특징을 추출하기 위해 고정된 프레임 간격이다. $Hue_i$, $Saturation_i$ 및 $Value_i$ 들은 각각 i번째 비디오 프레임에서의 $Hue$, $Saturation$과 $Value$ 값들이다.

## 5. 비디오 장르 분류

비디오 장르를 분류하기 위해서, CART에 의해 훈련되는 결정트리 기반 인식기를 사용하였다. CART는 비디오 장르 식별을 위해 안정적인 성능과 신뢰성있는 결과를 보여주기 위해 신뢰성 있는 가지치기(pruning) 정책, 정교한 이분 분할(binary split) 탐색 접근 과 자동적인 유효성 체크(automatic self validation)등과 같은 특성들을 제공한다 [7]. 분할 규칙으로 특징셋에서 가장 큰 클래스를 찾고 다른 클래스들로부터 분리시키는 Gini 인덱스를 사용한다. 또한, 비용(cost)을 줄이고 안정적인 성능을 확보하기 위한 가치치기에 15-교차 검증(cross validation)을 적용하였다.

## 6. 실험

본 논문의 실험에서는 MPEG-7 XM(eXperimentation Model)를 사용하여 구현된 방법의 유효성을 증명한다. 데이터베이스는 MPEG-7의 실험데이터를 위주로 구성하였고 부족한 비디오들은 상업 방송 비디오들로 구성하였다. 5개의 장르들은 카툰, 드라마, 뮤직비디오, 뉴스 그리고 스포츠로 구성되었으며, 각각 장르마다 60초 길이와 352 x 240의 크기를 가지는 72개의 비디오 클립이 사용되었다. 전체 데이터베이스의 개수는 360개가 되고 이 중 235 (5 x 47)의 클립들은 인식기를 훈련하기 위해 사용하였다. 나머지 비디오 클립들은 분류 성능을 테스트하기 위한 목적으로 사용되었다. 카툰 클립들은 한국, 일본, 미국 그리고 유럽의 서로 다른 카툰 비디오로부터 표본화하였다. 드라마 클립들은 10개의 서로 다른 비디오로부터, 뮤직비디오 클립들은 30개의 비디오 소스로부터 획득하였다. 3개의 스포츠 기사 클립을 포함하는 뉴스 클립들은 6개의 다른 비디오들로부터 추출되고, 스포츠 클립들은 축구, 농구, 테니스, 수영, 유도, 골프, 펜드(달리기) 그리고 배구경기를 포함하고 있는 12개의 서로 다른 비디오 소스를 갖는다. 오디오 특징을 추출하기 위해서 입력되는 오디오 스트림을 16 KHz로 샘플링한다. MPEG-7 오디오 서술자들에서 사용하고 있는 각각의 홉사이즈와 윈도우사이즈들은 MPEG-7에서 추천하는 것들을 사용한다. 실험에 사용되는 컴퓨터는 펜티엄 IV 2.4 GHz의 CPU와 512 MByte를 가진다.

장르 분류를 위해 제안하는 특징들의 유효성을 비교하기 위해 다음과 같은 3가지 경우에 대해서 실험하였다. 1) 오디오 특징 및 샷 개수, 2) 비주얼 특징 및 샷길이, 3) 오디오 및 비주얼 특징(샷길이 포함).

표 1은 오디오 특징 및 샷 개수, 즉 $F_set_{audio} = \{F_1, F_2, F_3, F_4, F_5, F_6\}$를 사용했을 때의 분류 결과를 나타낸다. 표에서 보여주는 것처럼, 오디오 특징들은 평범한 결과를 가지며 뮤직비디오, 뉴스, 스포츠 장르를 다른 장르들과 구별시킨다. 일반적으로 이들 장르들은 다른 장르들과 구별되는 독특한 오디오 특성을 가진다. 예를 들면, 음악 사운드는 높은 오디오 주파수와 파워를, 그리고 연속적인 앵커의 목소리는 플랫한 스펙트럼과 낮은 오디오 파워를 갖는다. 반면에, 카툰과 드라마는 사람 목소리와 음악, 배경 소리들이 섞여 있는 오디오 특성을 가짐으로써 분류가 잘 되지 않는다. 표 2는 비주얼 특징 및 샷 개수, 즉, $F_set_{visual} = \{F_7, F_8, F_9, F_{10}, F_{11}, F_{12}, F_{13}, F_{14}, F_{15}\}$를 사용했을때의 결과를 보여준다. 비주얼 정보는 뮤직비디오를 검색할 때 도움이 되며, 카툰과 스포츠를 구별하는 데 영향을 미친다. 뮤직비디오의 경우, 샷 템포가 다른 장르에 비해서 월등히 빠르다. 카툰의 경우 칼라의 변화가 중요한 요소가 되고, 스포츠는 모션 정보 와 카메라 모션에 의해 구별된다. 나머지 장르들은 액션 장면, 대화 장면, 러닝 장면등 여러가지 모션 정보와 샷 템포를 가지는 장면들이 섞여 있는 복잡한 구조를 가짐으로써 비주얼 정보에 의해 구별되기가 어렵다. 표 3에서는 오디오/비주얼 특징, 즉, $F_set_{multimodal} = \{F_1, F_2, F_3, F_4, F_5, F_7, F_{13}, F_{14}, F_{15}\}$를 적용한 경우가 나머지 2개의 경우에 비해 좋은 결과를 보여주고 있다. 125 테스트 비디오 클립을 통한 전체 분류 결과는 88.8%이다. 장르들 중 드라마는 다른 장르들에 비해 훨씬 낮은 결과를 보인다. 이는 이 장르가

다른 장르들에 비해 훨씬 복잡한 오디오/비주얼 구조를 가짐으로써 나타나는 결과이다.

표 1. 오디오 및 샷 개수를 사용한 장르 분류

| 예측 / 실제 | 카툰 | 드라마 | 뮤직 비디오 | 뉴스 | 스포츠 |
|---|---|---|---|---|---|
| 카툰 | 56.0% | 12.0% | 12.0% | 12.0% | 8.0% |
| 드라마 | 16.0% | 64.0% | 0.0% | 0.0% | 20.0% |
| 뮤직비디오 | 16.0% | 0.0% | 80.0% | 0.0% | 4.0% |
| 뉴스 | 0.0% | 12.0% | 0.0% | 84.0% | 4.0% |
| 스포츠 | 4.0% | 4.0% | 4.0% | 8.0% | 80.0% |
| 정확도 | 56.0% | 64.0 % | 80.0% | 84.0 % | 80.0% |
| 전체정확도 | 72.8% | | | | |

표 2. 비주얼 및 샷 개수를 사용한 장르 분류

| 예측 / 실제 | 카툰 | 드라마 | 뮤직 비디오 | 뉴스 | 스포츠 |
|---|---|---|---|---|---|
| 카툰 | 76.0% | 16.0% | 4.0% | 0.0% | 4.0% |
| 드라마 | 20.0% | 56.0% | 4.0% | 20.0% | 0.0% |
| 뮤직비디오 | 4.0% | 0.0% | 92.0% | 0.0% | 4.0% |
| 뉴스 | 4.0% | 24.0% | 0.0% | 60.0% | 12.0% |
| 스포츠 | 12.0% | 8.0% | 4.0% | 0.0% | 76.0% |
| 정확도 | 76.0% | 56.0 % | 92.0% | 60.0% | 76.0% |
| 전체정확도 | 72.0% | | | | |

표 3. 오디오/비주얼 특징을 사용한 장르 분류

| 예측 / 실제 | 카툰 | 드라마 | 뮤직 비디오 | 뉴스 | 스포츠 |
|---|---|---|---|---|---|
| 카툰 | 92.0% | 0.0% | 8.0% | 0.0% | 0.0% |
| 드라마 | 12.0% | 76.0% | 0.0% | 8.0% | 4.0% |
| 뮤직비디오 | 0.0% | 0.0% | 96.0% | 0.0% | 4.0% |
| 뉴스 | 8.0% | 4.0% | 0.0% | 88.0% | 0.0% |
| 스포츠 | 4.0% | 0.0% | 4.0% | 0.0% | 92.0% |
| 정확도 | 92.0% | 76.0% | 96.0% | 88.0% | 92.0% |
| 전체정확도 | 88.8% | | | | |

표 4는 멀티모달 특징셋, 즉 $F_set_{multimodal}$를 적용했을 경우 각각의 특징들에 대한 처리시간을 보여준다. 1개의 60초 비디오 클립에 대한 전체 평균 처리시간은 약 102초이다. 비디오 클립의 수가 증가하면 처리시간은 선형적으로 증가한다.

표 4. 멀티모달 특징셋에 대한 처리시간

| 특징 / 시간 | $F_1$ | $F_2$, $F_3$, $F_4$, $F_5$ | $F_7$ | $F_{13}$, $F_{14}$, $F_{15}$ |
|---|---|---|---|---|
| 처리 시간 | 68.3 sec | 12.6 sec | 6.4 sec | 14.6 sec |

## 7. 논의

표 3에서 보여주는 분류 결과와 선행된 연구들의 결과와 비교할 수 있다. Ba Tu Truong, et al.[2]은 C4.5 결정트리 기반 인식기와 샷 길이, 카메라 모션, 칼라정보와 같은 비주얼 정보만을 이용하여 평균 83.1%의 분류 결과를 보여준다. 그들의 데이터베이스는 60초 길이의 카툰, 광고, 음악, 뉴스 그리고 스포츠 비디오로 구성되어 있다. Zhu Liu, et al.[10]은 28개 심볼을 가지는 5-states HMM과 오디오 정보만을 사용하고 있다. 광고, 농구, 풋볼, 뉴스, 기상뉴스로 구성된 10분 길이의 오디오 클립을 통해서 획득한 그들의 결과는 약 84.7%이다. Huang, et al.[11]의한 분류 결과는 약 91.40%이다. 그들의 실험은 오디오/비주얼 특징들과 Product HMM, 그리고, 뉴스, 기상뉴스, 광고, 농구, 풋볼로 구성된 10분 길이의 비디오 데이터베이스를 사용하고 있다. 장르의 종류와 양이 다르므로, 기존의 연구들과 제안하는 방법을 직접적으로 비교하기는 어렵다. 특히, J. Huang, et al.과 비교했을 때 제안된 방법에서 사용한 비디오의 길이가 훨씬 짧다. [2][10]에서 알 수 있듯이, 멀티모달 특징을 사용하는 본 실험의 결과는 유니모달 특징들을 사용하는 방법들에 비해 높은 것을 알수 있다.

표 3에서 보여지는 분류 결과는 약 89%이다. 이 결과를 바탕으로 분류 결과를 더욱 향상시키기 위해, 한 비디오에서 선택되는 비디오 클립수를 증가시켜 올바르게 분류될 수 있는 확률을 계산하고 사용할 수 있다. 계산된 결과는 표 5에서 보여준다. 즉, 분류를 위해 사용되는 비디오 클립수가 증가할수록 분류 성능은 높아진다. 특정한 비디오로부터 여러 개의 비디오 클립을 임의적으로 선택하고 그 중 절반 이상을 올바른 장르로 인식하면 분류가 성공이라고 규정한다. 예를 들면, 한 비디오에서 3개의 클립을 임의적으로 선택했을 때 분류 성능은 약 95.48%이다. 물론, 처리시간은 약 3배가 증가한다. 이 결과는 Huang, et al의 결과보다 훨씬 좋은 것을 알 수 있다.

$$Genre\_Classification\_Rate = \sum_{i=r}^{M} \binom{M}{i} P_g^i (1-P_g)^{(M-i)} \quad (21)$$

$$r = Round(M/2)+1, M = 1,3,5,\ldots\ (a\ odd\ number) \quad (22)$$

여기서 $M$은 한 비디오에서 선택되는 비디오 클립의 개수이고 $P_g$는 한 비디오 클립이 올바르게 분류되는 확률이다.

표 5. 한 비디오로부터 선택된 비디오 클립들이 정확히 분류될 확률

| 클립수 / 장르 | 표 3에서 1 클립 경우 | 3 클립중 2 클립이상 | 5 클립중 3 클립이상 |
|---|---|---|---|
| 카툰 | 92.0% | 98.2% | 99.5% |
| 드라마 | 76.0% | 85.5% | 90.7% |
| 뮤직비디오 | 96.0% | 99.5% | 99.9% |
| 뉴스 | 88.0% | 96.0% | 98.6% |
| 스포츠 | 92.0% | 98.2% | 99.5% |
| 정확도 | 88.8% | 95.48% | 97.64% |

실험결과와 이론적 분석을 통해, MPEG-7에 기반한 오디오/비주얼 특징을 사용하는 제안된 방법이 효율적이고 정확한 장르 분류를 하고 있음을 확인할 수 있었다.

## 8. 결론

본 논문은 멀티모달 특징을 사용하여 자동적인 비디오 요약 또는 검색 및 분류를 위해 방송 비디오 장르 분류 방법을 제안하고 있다. 제안하는 방법은 MPEG-7에서 오디오 및 비주얼 서술자들을 채택한다. 제안하는 방법의 유용성을 증명하기 위해 각 장르의 특성을 오디오 및 비주얼 특징들의 확률적인 분포를 분석하고, 멀티모달 특징들을 CART를 이용한 결정트리 기반의 인식기에 적용하여 분류한다. 여러 장르로 구성된 비디오 데이터베이스를 통한 실험은 제안하는 방법의 분류 성능을 평가한다. 실험을 통해 멀티모달 특징들의 적용한 제안된 방법이 방송 컨텐츠들의 정확하고 효율적인 분류하고 있음을 보여준다. 앞으로 수행될 과제로는 장르 범위의 확장 및 처리시간과 성능을 동시에 높일 수 있는 방법을 연구하는 것이다.

## 9. 참고문헌


[1] Ying Li et al., "Video Content Analysis using Multimodal Information: For Movie Content Extraction, Indexing and Representation," *Kluwer Academic Publishers*, 2003.

[2] Ba Tu Truong, et al., "Automatic genre identification for content-based video categorization," *IEEE Pattern Recognition 2000*, Vol. 4, pp. 230-233, Sep. 2000.

[3] Jasinschi, et al., "Automatic TV program genre classification based on audio patterns," *IEEE Euromicro Conference 2001*, pp. 370-375, Sep. 2001.

[4] Z. Liu, et al., "Classification of TV programs Based on Audio Information using Hidden Marcov Model," *IEEE WS. on Multimedia Signal Processing*, pp. 27-32, Dec. 1998.

[5] Nam, J., et al., Audio-visual content-based violent scene characterization," *IEEE ICIP 98*, Vol. 1, pp. 353-357, Oct. 1998.

[6] Rasheed, Z., at el., "Movie genre classification by exploiting audio-visual features of previews," *IEEE Pattern Recognition 2002*, Vol 2, pp. 1086-1089, Aug. 2002.

[7] Leo Breiman, et al., "Classification and Regression Trees," *Chapman & Hall/CRC*, 1984.

[8] Manjunath, et al., "Introduction to MPEG-7 Multimedia Content Description Interface. B.S.," *John wiley & Sons*, 2002.

[9] ISO/IEC 15938-4, "Information Technology Multimedia Conte -nt Description Interface Part 4: Audio," July 2001.

[10] Z. Liu, J. Huang, et al., "Classification of TV programs based on audio information using hidden Markov model," *IEEE MMSP-98*, pp. 27-32, Dec. 1998.

[11] Huang, J., et al., Integration of multimodal features for video scene classification based on HMM," *IEEE MMSP-99*, pp. 53-58, Sept. 1999.


## Acknowledgements


본 연구는 한국전자통신연구원(ETRI)의 지원받은 수행중인 SmarTV과제의 연구 결과 중 일부분임.