동영상 카툰 시스템을 위한 자동 프레임 추출 기법

Auto Frame Extraction Method for Video Cartooning System

김대진, 구떠올라
(주)다이렉트미디어

Dae-Jin Kim(djkim@directmedia.net), Ddeo-Ol-Ra Koo(rskoo@directmedia.net)

요약

멀티미디어 산업의 발달과 함께 디지털 콘텐츠 시장의 확산을 가져오고 있다. 그 중 인터넷 만화와 같은 디지털 만화 시장의 확장은 급속하게 커지고 있어서, 콘텐츠의 부족과 다양성 때문에 동영상 카툰에 대한 연구가 계속되고 있다. 지금까지는 동영상 카툰은 비사실적 렌더링과 말풍선에 초점이 맞추어졌으나, 이러한 것들을 적용하기 위해서는 카툰 서비스에 적합한 프레임 추출이 우선시 되어야만 한다. 기존의 방법으로는 동영상의 장면전환이 일어나는 샷(shot)안의 프레임을 추출하여, 사용자가 지정한 영역을 임의의 색상으로 렌더링(Rendering)하는 시스템이 있다. 하지만 이러한 방법은 사람의 손을 거치는 반자동적인 방법으로서 정확한 프레임 추출을 위해 사람의 손을 거쳐야하는 단점이 있다. 따라서 본 논문에서는 이러한 문제점을 해결하고, 보다 정확한 카툰에 적용할 프레임을 추출하기 위해 오디오 및 비디오 분리를 통한 방법을 제안한다. 먼저 동영상으로부터 오디오와 비디오를 분리한다. 오디오는 먼저 MFCC와 영교차율의 특징을 추출하고, 이 특징 정보를 미리 학습된 데이터와 GMM 분류기를 통하여 음악, 음성, 음악+음성으로 분류한 후 음성 영역을 설정한다. 비디오는 히스토그램을 이용한 방법과 같은 일반적인 장면 전환 프레임을 추출 후 얼굴 검색을 통해서 만화에서 의미가 있는 프레임을 추출한다. 그 후 음성 영역 내에 얼굴이 존재하는 장면전환 프레임이나 일정 시간동안 음성이 지속되는 영역 중 장면전환 프레임을 추출하여 동영상 카툰에 적합한 프레임을 자동으로 추출한다.

■ 중심어 : | 카툰 | 음성 영역 | 얼굴 검색 | 프레임 추출 |

Abstract

While the broadband multimedia technologies have been developing, the commercial market of digital contents has also been widely spreading. Most of all, digital cartoon market like internet cartoon has been rapidly large so video cartooning continuously has been researched because of lack and variety of cartoon. Until now, video cartooning system has been focused in non-photorealistic rendering and word balloon. But the meaningful frame extraction must take priority for cartooning system when applying in service. In this paper, we propose new automatic frame extraction method for video cartooning system. At first, we separate video and audio from movie and extract features parameter like MFCC and ZCR from audio data. Audio signal is classified to speech, music and speech+music comparing with already trained audio data using GMM distributor. So we can set speech area. In the video case, we extract frame using general scene change detection method like histogram method and extract meaningful frames in the cartoon using face detection among the already extracted frames. After that, first of all existent face within speech area image transition frame extract automatically. Suitable frame about movie cartooning automatically extract that extraction image transition frame at continuable period of time domain.

■ keyword : | Cartooning | Speech Area | Face Detection | Frame Extraction |

I. 서론

광대역 네트워크의 발달과 함께 멀티미디어 산업의 발전은 디지털 콘텐츠 시장의 확산을 가져왔다. 또한 스마트폰의 대중화는 영화, 음악, 만화와 같은 콘텐츠들을 실생활에서 생활패턴의 변화를 주고 있다. 특히 이중에서도 만화 콘텐츠는 짧은 시간에 조금씩 볼 수 있고, 독자들이 이해하기에 쉽고 직관적이기 때문에 정보 전달에 대한 매체로 자리 잡았다.

일반적으로 만화는 작가들이 쓴 각본을 기반으로 하여 내용을 한 밑그림에 넣어서 구성하게 된다. 이러한 전통적인 방법은 한권의 만화를 만들기 위해서 많은 시간을 소요할 뿐만 아니라, 사람의 손이 많이 거치는 작업이다. 따라서 이러한 단점을 보완하고자 카툰(Cartooning)기법을 이용하여 이미 만들어진 동영상상을 이용하여 만화처럼 만들기 위한 다양한 연구가 이루어지고 있다[1]. Wang 등은 동영상의 장면전환이 일어나는 샷(shot)안의 프레임을 추출하여, 사용자가 지정한 영역을 임의의 색상으로 렌더링(Rendering)하는 시스템을 제안하였다[2]. Ryu 등은 사람의 색상 인지와 유사한 특성을 가지는 HSV 색상 모델을 사용하여, 각 영역의 색상 유사도를 계산하고 일반적으로 활용되는 색상과 프레임의 이미지의 픽셀값을 고려한 이진화를 수행하여 흑백 만화를 구성하는 알고리즘을 연구하였다[3]. Hwang 등은 비디오 스트림으로부터 추출한 이미지에 배경효과(Background Effect), 속도선(Speed line), 회전궤도(Rotation Trajectory)등과 같은 만화적 특징요소를 첨가하여 보다 효과적인 이미지를 구성하였다[4]. 그리고 Preu 등은 동영상의 대본을 이용하여 장면(Scene)에 대한 시간 및 특징정보를 미리 구성한다. 그리고 이 정보를 통하여 말풍선의 형태정보를 구성한 후 장면의 시작점으로부터 얼굴이 검출된 프레임에 말풍선들을 반자동으로 렌더링 하는 시스템을 제안하였다[5].

이러한 연구들은 대부분 단순히 비디오 스트림으로부터 장면전환점을 찾아 그곳의 프레임을 추출하여 카툰을 하거나 대본을 이용하여 장면 정보를 미리 구성하는 등 반자동으로 카툰이 된다. 따라서 카툰의

주목적에 맞게 사람의 손을 거치지 않고 자동으로 카툰 되는 시스템이 요구되어진다. 그러기 위해서는 등장 인물의 대화와 말풍선의 동기(Synchronization)가 정확히 이루어 져야 되고, 의미 있는 프레임을 추출하는 것이 카툰 시스템의 성능을 좌우하게 된다.

이러한 카툰 시스템의 요구사항들을 만족시키기 위해서 일반적인 장면전환 검출 후 얼굴이 검출되는 프레임과 대화가 이루어지는 시점의 프레임을 추출하여 본 논문에서 제시하는 조건에 맞추어 카툰 시스템을 구성하였다. 따라서 기존에 사람이 직접 수동으로 장면 전환점으로부터 프레임을 추출하는 방법에서 벗어나 카툰에 의미 있는 프레임을 추출하여 사용함으로써 만화의 이야기 전개가 자연스럽게 이루어지도록 하였다.

본 논문의 구성은 2장에서는 기존 동영상 카툰의 일반적인 방법에 대하여 설명하고, 3장에서는 제안하는 자동 동영상 카툰 시스템을 설명한다. 4장에서는 앞의 기술을 이용하여 데이터베이스 구축 및 실험을 한 후 마지막으로 5장에서 결론을 맺었다.

II. 일반적인 카툰 기법

기존의 동영상 카툰 시스템은 처음에 동영상으로부터 장면전환점을 검출하여 프레임을 추출하고, 이를 만화적인 요소처럼 구성하기 위하여 비사실적 렌더링을 한다. 이때 Mean-Shift를 통하여 색의 구성을 변경하고 윤곽선을 표시함으로써 만화 화면을 구성한다. 여기에 말풍선을 넣음으로써 이야기 전개가 가능한 카툰 시스템이 완료된다. 이러한 과정을 자세히 살펴보면 아래와 같다.

1. 장면전환 검출

기존의 영상 데이터에서 장면 전환점을 찾기 위해서 키프레임을 추출하는 기법이 주로 사용된다. 이 프레임은 동영상을 가장 효율적으로 대표하는 장면들로 구성되어 있다. 따라서 이 프레임을 추출하기 위해서는 동영상의 의미 단위를 알아야 한다. [그림 1]은 동영상 시

퀵스 구조를 나타낸다.

그림 1. 동영상 시퀀스 구조

장면전환점은 A에서 B로 변화가 일어나는 부분을 말하며, 샷 내부에 높은 유사성을 갖는 프레임은 동일한 장면으로 간주하여 장면전환점으로 추출하지 않는다.

장면전환점을 찾기 위한 다양한 방법들이 연구되고 있다.

- ① 인접프레임의 변화량을 이용하여 픽셀 값의 차이를 비교하는 방법
- ② 히스토그램(Histogram) 차이를 비교하는 방법
- ③ 에지를 추출하여 비교하는 방법
- ④ 인접한 블록사이의 유사도를 비교하는 방법
- ⑤ DCT(Discrete Cosine Transform) 계수나 모션 벡터를 비교하는 방법

최근에 와서는 하나 이상의 특징들을 혼합하여 특징들을 비교함으로써 장면 전환점을 찾는 연구가 계속되고 있다.

2. 비사실적 렌더링

(Non-PhotoRealistic Rendering)

2.1 Mean Shift

각각의 포인트(비디오나 이미지의 픽셀)에서 비슷한 포인트 간의 지역적인 조밀도를 바탕으로 결과를 추측한다. 즉 Mean-Shift 알고리즘은 비슷한 데이터들 간의 지역적인 조밀도의 변화도(Gradient)를 추측한다. 여기서 지역적인 조밀도는 입력된 이미지나 비디오에서 구성된 확률 밀도함수로 구성된다. 계산된 변화도 벡터

(Gradient Vector)는 초기 위치에서 Mean-Shift 포인트를 도출시키며, 연산된 포인트에서 또 다시 변화도 벡터를 구하는 반복적인 실행을 통하여 지역적 조밀도 분포에서 가장 높은 정점(Peak)을 찾게 된다. 같은 정점에서 수렴되는 모든 포인트들은 같은 세그먼트에 속한 멤버로 취급된다. 모든 포인트가 각각 지역적인 정점을 찾음으로 인해 데이터는 분할 될 수 있다.

2.2 윤곽선 추출

이미지를 펜으로 그린 것과 같은 효과를 두기위해서 영상의 강인한 스트로크(Stroke)를 이용하여 만화적인 표현을 완성할 수 있다. 이 때 캐니에지(Canny Edge) 검출 기법과 같은 것을 사용한다.

2.3 비사실적 화면 구성

2.1의 Mean-Shift를 통해서 색의 단순화를 구성하고 2.2의 윤곽선을 통해서 펜으로 그림을 그린 듯 한 효과를 나타낼 수 있다. [그림 2]에서는 이 두 가지를 이용하여 만화적인 비사실적 화면을 구성하였다.

그림 2. 만화적인 비사실적 화면 구성[5]

[그림 2]는 1의 동영상 프레임으로부터 2의 캐니에지 검출을 통한 윤곽선 화면과 3의 Mean-Shift를 통한 색의 단순화를 통하여 4의 만화적인 비사실적 화면 구성을 보여준다[5].

3. 말풍선

카투닝에서 말풍선 삽입은 대부분 프레임을 추출하고, 직접 위치를 선정하여 수작업으로 삽입하였다. 이때 대부분의 독자들은 만화의 장면에서 오른쪽보다는 왼쪽에, 아래쪽 보다는 위쪽에 말풍선의 대사를 먼저 읽게 되는 것을 고려하여 말풍선을 넣는다. 즉 읽는 순서(Reading Order)에 맞게 배치한다[6].

Chun등은 확장된 특징점 기반의 레이블 배치(EPFLP: Extended Point Feature Label Replacement) 방법과 Qlayout 평가 함수를 사용하여, 말풍선을 읽는 순서에 맞게 자동으로 배치하였다[7]. 그러나 이야기 전개상 적절한 장면전환 프레임 추출을 전제로 접근하였다.

따라서 이야기 전개를 부드럽게 하기위해서 카투닝에 적절한 장면전환 프레임 추출이 요구된다.

III. 제안하는 카투닝 알고리즘

장면전환 검출을 하기 위해서 기존의 방법은 영상이 변화하는 시점을 이용하여 추출하였다. 또한 카투닝을 하기 위해서는 장면전환 프레임 중 사용자가 직접 선택하여 카투닝에 사용하였다. 그러나 시스템 자동화라는 측면에서 적합하지 않고, 이야기 전개가 이루어지는 시점을 찾아서 사용자가 프레임을 직접 추출하는 것은 시스템에 적합하지 않다. 즉 추출된 프레임들의 순차적 집합에 말풍선을 추가하였을 때 만화로서의 이야기 흐름이 이어져야 하는 자동화 시스템이 필요하다.

따라서 본 논문에서는 기존의 방법대로 장면전환 검출 후 얼굴 검출을 이용한 비디오 필터링과 음성 검출을 이용한 오디오 필터링 기법을 제시하는 조건에 맞추어 효과적인 자동 카투닝 시스템을 구축하였다.

1. 비디오 필터링

만화는 특성상 인물들의 대화를 통해서 이야기가 전개 된다. 대화가 이루어지는 시점에는 대부분 사람이 존재하며, 만화는 말풍선을 통해서 인물의 생각이나 의사를 전달한다. 특히 인물의 감정을 표시하기 위해서

이 시점에는 얼굴이 대부분 존재한다. [그림 3]에서는 일반적인 만화의 이야기 전개모습을 보여준다. 얼굴을 통해서 인물의 감정을 표현하고, 말풍선을 통해서 이야기 전개를 한다. 본 이미지는 “스포츠 투데이”에서 서비스 했던 “샤인”이라는 제목의 만화이다.

그림 3. 일반적인 만화 이야기 전개

따라서 일반적인 장면전환 검출을 통해서 뽑아낸 프레임을 얼굴 검출을 통해서 다시 한 번 필터링 함으로써 카투닝에 보다 의미 있는 프레임을 추출 할 수 있다.

1.1 Adaboost를 이용한 얼굴 검출

기존의 Preu 등은 장면의 시작점으로부터 얼굴을 검출하였다. 이때 사람의 피부색과 유사한 색을 검출하여 얼굴을 검출하였다[5]. 그러나 이것은 인종이나 사람 피부색의 차이가 존재하기 때문에 한계가 있다. 본 논문에서는 Viola[10] 및 Lienhart[8] 등이 제안한 하르분류기(Harr Classifier)를 통한 Adaboost 알고리즘을 사용

하여 얼굴이 검출된 프레임을 추출 하였다.

Adaboost 알고리즘을 이용한 얼굴 검출을 위해선 먼저 분류기의 학습이 필요하다. 이때 훈련에 사용되는 특징으로 [그림 4]와 같이 사각형 특징(Rectangle Features)이 사용된다. 사각형 특징 들은 2개 혹은 3개의 검은색 과 흰색의 사각형으로 구성되었다[12].

그림 4. 사각형 특징

이런 특징 들은 수평과 수직 그리고 대각 방향으로 나타나며 이는 얼굴의 구성요소를 간략히 나타낸 것이라 할 수 있다.

영상 내 얼굴 검출을 위해 [그림 4]와 같은 특징들을 사용하는데 있어 연산 속도를 증가시키기 위해 Viola 등은 적분영상(Integral Image)을 사용하였다. 적분영상은 영상의 어떤 한 위치 x, y 에 대해 픽셀들의 밝기 값의 합을 나타낸다. 수식 (1)은 적분 영상을 구하는 식이다.

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y') \quad (1)$$

식 1에서 $ii(x, y)$ 는 적분영상을 나타내며, $i(x', y')$ 은 원 영상을 나타낸다. 식 (1)은 다음의 식 (2) 및 식 (3)과 같이 순환 식을 이용하여 빠르게 구할 수 있다[13].

$$s(x, y) = s(x, y - 1) + i(x, y) \quad (2)$$

$$ii(x, y) = ii(x - 1, y) + s(x, y) \quad (3)$$

이때, $s(x, y)$ 는 열 방향의 픽셀 값들의 합을 나타내며, $s(x, -1) = 0, ii(-1, y) = 0$ 이다.

적분영상을 이용하면 어떠한 사각형 영역내의 픽셀 값들의 합도 네 점을 이용하여 빠르게 구할 수 있다 [14]. [그림 5]와 같은 사각형 영역에서 A의 영역의 픽셀 값을 구하기 위해 식 (4)와 같이 사각형 A의 네 꼭지점을 이용하여 구할 수 있다.

$$\sum(A) = ii(p_1) + ii(p_4) - ii(p_2) - ii(p_3) \quad (4)$$

그림 5. 두 개의 사각 특징 점 계산 예

Adaboost 분류기의 훈련은 얼굴 영상과 비 얼굴 영상을 잘 판단 할 수 있도록 [그림 4]의 특징을 이용해 수행된다. Adaboost 분류기를 구축하는데 식 (5)와 같은 Viola 분류함수를 이용한다.

$$F = \text{sign}(w_1 f_1 + w_2 f_2 + \dots + w_n f_n) \quad (5)$$

이때 $\text{sign}()$ 함수는 음수에 대해서는 -1, 양수에 대해서는 1, 0에 대해선 0을 반환한다[11]. 수식 (5)에서 f_n 은 약한 분류기로서 학습되고 누적되며, 여러 값이 작을수록 가중치 w_n 은 점점 커진다. 식 (5)의 분류함수를 통해서 높은 검출율과 약한 거절 비율을 갖는 노드를 만든다. 이때 노드를 만들기 위해 통계적 부스팅을 사용한다. 이렇게 분류기를 통해 영상내의 얼굴을 검출하며 분류기의 단계를 통한 얼굴의 검출과정은 [그림 6]과 같다.

그림 6. 분류기에 의한 얼굴 검출 과정

[그림 7]에서는 하르분류기를 이용한 얼굴 검출을 한 프레임 모음을 나타낸다. 1시간 22분의 영화중 초반 30초의 동영상으로부터 추출하였다. 테스트에 사용한 영화는 “식객 2: 김치전쟁”이다.

그림 7. Adaboost를 이용한 얼굴 검출 결과

그러나 얼굴 검출을 이용한 필터링은 아래와 같은 경우에 문제점을 가진다.

- ① 주인공의 대사가 없는데 주변인물의 얼굴 때문에 장면 프레임이 검출되는 경우
- ② 주인공이 화면에 존재하며 뒷모습을 보이고 대사하는 경우

따라서 이러한 문제점을 보완하기 위해 오디오 필터링을 같이 적용한다.

2. 오디오 필터링

만화에서는 말풍선을 통해서 이야기 전개가 이루어지기 때문에 대사가 일어나는 부분의 프레임을 카투닝에 이용할 수 있다. 영화나 드라마 같은 영상데이터는

크게 3가지 오디오 형태를 가진다.

- ① 음성 : 대화전달
- ② 음악 : 배경음악
- ③ 음성+음악 : 배경음악을 덧붙인 대화전달을 통하여 극적효과를 높임

위 3가지를 구분하여 적용하면 보다 효과적인 카투닝 시스템을 구축할 수 있다. 음성만 존재하는 경우는 추출한 프레임에 일반적인 말풍선을 적용하여 이야기를 전개한다. 음악만 존재하는 경우에는 이야기 전개에 불필요한 장면이므로 추출한 프레임을 제외시킨다. 음성+음악이 존재하는 경우에는 말풍선을 적용한다. 단 극적효과를 극대화하기 위하여 일반적인 말풍선이 아닌 생각/ 강조/느낌을 표현하는 말풍선을 적용한다.

본 논문에서 오디오 신호의 음성/음악 분류를 위하여 MFCC(Mel-Frequency Cepstral Coefficient)과 영교차율(ZCR: Zero-Crossing Rate)의 조합을 이용하여 분류하였다. 두 특징들 간의 상호조합이 가장 양호한 결과를 생성한다[9]. 또한 오디오 신호의 분류 결정은 훈련된 데이터베이스에 대해 학습시킨 분류항목별 가우시안 혼합 모델(GMM: Gaussian Mixture Model)을 적용하여 그 중 최대 확률을 가지는 것을 선택하였다.

2.1 MFCC

MFCC는 오디오 신호 처리에 기본적으로 적용되는 특징 중 하나이다. 이것은 인간 청각 인지시스템을 모델링하였다. [그림 8]에서는 MFCC 처리과정을 보여준다.

그림 8. MFCC 처리과정

입력된 오디오 신호는 고주파 성분의 포먼트(Formant)가 저주파성분의 포먼트보다 상대적으로 크기가 작아서 Pre-Emphasis의 과정을 거쳐, 전 대역에 고른 에너지 분포를 가지도록 한다. 그 후 Hamming 또

는 Hanning 윈도우를 통하여 프레임 추출한다. 추출된 프레임은 DFT(Discrete Fourier Transform)을 통하여 주파수 대역으로 변환하고, 변환된 계수값들을 [그림 9]와 같이 Mel Filter Banks를 통과한다.

그림 9. Mel Filter Banks

Mel Filter Bank는 인간청각구조에 근접한 여러 개의 Band-Pass Filter들로 구성되었으며 1K 미만에서는 균일하게 구성되나, 그 이상에서는 로그 크기 단위로 구성된다. Mel Filter Bank를 통과한 신호는 Logarithm 처리를 한다. 이는 크기성분을 살리고 위상성분을 버려서 화자의 입과 입력디바이스(마이크 등) 간의 거리 변화에 덜 민감하게 한다. 그 후 IDFT(Inverse Discrete Fourier Transform)를 통하여 MFCC를 구한다. k번째 Mel Filter Bank를 통과한 신호를 $S[k]$ 라고 할 경우, 식 (6)과 같이 표현된다. M은 필터뱅크의 개수이고, L은 MFCC 차수이다.

$$C[n] = \sum_{k=1}^M \log(S[k]) \cos \left[(k-0.5) \frac{n\pi}{M} \right] \quad (6)$$

$$n = 1, \dots, L$$

본 논문에서는 26개의 필터로 구성된 필터뱅크로 필터링을 하였으며 분류항목의 특징을 가장 잘 나타내는 2개의 계수값을 이용하였다. 선별된 계수값의 변화를 살펴보면 음성, 음성+음악, 음악 순으로 변화를 보였다 [9].

[그림 10]에서는 분류항목별 MFCC의 변화를 보여준다.

그림 10. 각 분류항목별 MFCC 변화(1, 2차)
A: 음성 B: 음악 C: 음성+음악

2.2 영교차율

영교차율은 오디오 프레임 내에서 신호의 부호가 변하는 횟수를 의미한다. 식(7)에서는 영교차율을 나타내는 수식을 표시한다.

[그림 11]에서는 각 분류항목별 영교차율을 비교하였다. [그림 11]과 같이 음성신호는 특성상 무성음과 유성음이 번갈아 나타내기 때문에 스펙트럼(Spectrum)의 변화가 음악보다 더 크다.

$$ZCR = \frac{1}{2} \sum_{m=0}^{N-1} |sgn[x(m)] - sgn[x(m-1)]| \quad (7)$$

$$sgn[x(m)] = \begin{cases} x(m) = 1, & x(m) > 0 \\ x(m) = -1, & x(m) < 0 \end{cases}$$

그림 11. 각 분류항목별 영교차율 변화
A: 음성 B: 음악 C:음성+음악

2.3 가우시안 혼합모델

분류기에 사용하는 알고리즘으로 가우시안 혼합 모델을 사용하였다. 가우시안 혼합 모델은 특징 벡터의 분포를 몇 개의 가우시안의 가중치 합으로 모델링 하는 것이다. 가우시안 혼합 모델로 모델링 된 확률 분포는 다음과 같은 식으로 표현할 수 있다.

$$p(x|\theta) = \sum_{i=1}^M p(x|\omega_i, \theta_i) P(\omega_i) \quad (8)$$

여기서 $p(x|\omega_i, \theta_i)$ 는 혼합 모델의 기본 성분을 이루는 간단한 확률 밀도 함수이다. θ_i 는 i 번째 성분이 되는 확률밀도함수를 정의하는 파라미터(Parameter)로 평균과 공분산 행렬이 된다. ω_i 는 i 번째 성분임을 나타내는 확률변수이고, $P(\omega_i)$ 는 i 번째 성분이 전체 혼합

확률밀도 함수에서 차지하는 상대적 중요도를 의미한다. $P(\omega_i)$ 는 파라미터 α_i 로 표시하기도 하며 식(9)과 같은 성질을 만족해야 한다.

$$\sum_{i=1}^M \alpha_i = 1, \quad 0 \leq \alpha_i \leq 1 \quad (9)$$

이 혼합 확률 모델을 사용하여 데이터의 분포를 나타내기 위해 추정해야만 하는 파라미터 θ 는 식(10)과 같다.

$$\theta = (\mu_1, \mu_2, \dots, \mu_M, \sigma_1^2, \sigma_2^2, \dots, \sigma_M^2, \alpha_1, \alpha_2, \dots, \alpha_M) \quad (10)$$

이때 각 I 번째 성분의 가우시안 함수의 공분산 행렬은 $\sigma^2 I$ 이다.

가우시안 혼합모델의 훈련과정에서는 MLE(Maximum Likelihood Estimation) 알고리즘을 이용하여 최대 가우시안 혼합 분포 값을 가지는 파라미터를 추정하였다. 음성과 음악의 인식 과정에서는 추정된 파라미터를 이용하여, 입력된 신호 데이터의 특징 벡터에 대해 각각의 가우시안 혼합 분포를 구하고, 그 중 가장 큰 확률 값을 가지는 파라미터를 인식 결과로 표현할 수 있다.

2.4 음성/음악 분리

특정들의 단일조합을 통한 음성/음악 분류는 오분류 확률이 크다. 여기서 음성은 배경음악을 포함하고 있는 (음성+음악)을 포함한다. MFCC의 경우 음악구분에 있어서 영교차율 특징에 비해 좋지 않고, 영교차율의 경우 음성+음악 분류능력은 타 특징에 비해 가장 떨어진 다. 특히 음성 분류에 있어서 타 특징과 달리 음악으로 오분류 현상이 나타난다. 따라서 MFCC와 영교차율의 조합을 통해서 음성, 음악 분류능력을 향상 시켰다[9].

3. 동영상 카투닝을 위한 프레임 추출

3.1 카투닝 프레임 조건

본 논문에서 제안하는 동영상 카투닝을 위해서는 비

디오 및 오디오 추출의 결과를 통한 조건적 영상 프레임의 추출이 필요하다. 프레임 추출을 위한 조건은 다음과 같다.

- ① 얼굴과 음성이 동시에 존재하는 경우.
- ② 음성은 있지만 얼굴 검출이 되지 않은 경우.

① 경우는 오디오 필터링을 통한 결과와 비디오 필터링을 통한 결과에 맞춰 수행하면 된다. 하지만 ② 경우는 비디오 필터링의 얼굴 검출이 되지 못했지만, 오디오 필터링에서의 지속적인 음성데이터의 존재로 인해 화자의 얼굴이 화면상에 나타나지 않은 상태에서 대사를 했을 경우도 생각해야 하기 때문에 중요하다. 이러한 두 가지 경우에 대한 판단을 통해 카투닝에 사용할 정확한 프레임의 추출이 되어야 한다.

3.2 카투닝 프레임 추출

제안하는 시스템에서는 동영상 카투닝을 위한 장면 전환 알고리즘을 적용하기 위해서 먼저 동영상으로부터 오디오와 비디오로 각각 추출한다. 먼저 오디오로부터 특징을 추출 후 구성된 데이터베이스를 이용하여 테스트 하고자 하는 콘텐츠의 음성 영역을 추출한다.

비디오의 경우는 I 프레임을 추출한 후 히스토그램을 이용하여 먼저 장면전환 프레임을 추출하고, 그 후 Adaboost를 이용한 얼굴 영역이 존재하는 부분을 추출한다. 그 후 위의 두 가지 조건에 따라 프레임을 추출한다.

먼저 동영상으로부터 음성 영역과 얼굴 검출 영역이 동시에 발생하는 영역이 동영상 카투닝을 위한 프레임 영역이다. 두 번째로 얼굴검출이 안되고 음성 영역이 계속 된다면, 즉 얼굴 검출 없이 대화만 계속 진행되는 경우가 5초 이상 발생된다면 대사로 상황 설명 등이 진행 되는 경우이므로 이 구간의 중간 부분의 프레임을 카투닝에 이용한다. [그림 12]에서는 이와 같은 과정을 나타낸 동영상 카투닝을 위한 장면전환 프레임 추출 순서도이다.

그림 12. 장면전환 프레임 추출 순서도

IV. 데이터베이스 구축 및 구현 결과

1. 오디오 데이터베이스 구축

오디오로부터 각각 음성영역을 추출하기 위해서는 먼저 오디오 데이터에 대한 데이터베이스를 구축해야 한다. 동영상 학습을 할 때, 배경음악을 포함한 대사 처리 부분도 카투닝을 할 때 대사처리가 들어가는 부분 이므로 카투닝 프레임 추출 영역에 사용하였다.

오디오 데이터는 각각 10분 분량의 영화 콘텐츠 30편, 공중파 드라마 15편, 음악 방송 15편, 뉴스 10편 및 3분 분량의 기타 음악파일로 구성하였다. 각 콘텐츠 중 처음과 끝부분의 묵음 부분이 3초 이상 존재하는 경우 그 부분은 제외 하였다. 또한 샘플 주파수를 통일하기 위하여 16kHz 이상 되는 콘텐츠는 다운샘플링을 통하여 주파수를 일치 시켰다. 한 프레임의 크기는 25ms이며 12.5ms씩 이동시켜가며 특징 파라미터를 추출하였다. 특징 파라미터 θ 로는 영교차율과 MFCC를 사용하였으며 MFCC 특징 파라미터는 각 분류항목 특징을 가장 잘 나타낼 수 있는 2개의 계수값(1, 2차)을 선별하여 사용하였다.

2. 구현 결과

[그림 13]에서는 알고리즘 적용 후 동영상 카투닝을 위하여 장면전환 프레임을 추출한 결과를 보여준다. ①의 경우는 “주인공의 대사가 없는데 주변인물의 얼굴 때문에 장면 프레임이 검출되는 경우”이고, ②의 경우는 “주인공이 화면에 존재하며 뒷모습을 보이고 대사하는 경우”에 해당한다. 동영상 카투닝에서 ①은 프레임이 추출되면 안 되고 ②는 추출되어야만 하는 프레임이다. 기존의 얼굴검출을 기반으로 한 동영상 카투닝과 달리 본 알고리즘을 적용 시에는 ②~⑨ 까지 보다 적절한 프레임을 추출하였다.

그림 13. 알고리즘 적용 후 동영상 카투닝을 위한 장면전환 프레임

본 논문에서는 말풍선이 필요한 장면인데 프레임 추출을 하지 않는 경우를 오분류로 분류하였다. 영상마다 오분류가 되는 확률은 영상의 화면 구도 및 영상 효과, 영화감독의 촬영 기법 등에 의해서 다르게 나온다. 따라서 영화의 장르에 따라 구분지어 오분류 확률을 조사하였다. 단 테스트에 사용한 영화들은 사람을 주인공으로 하는 영화들을 10분씩 실험에 사용하였다.

[표 1]에서 영화장르별 카투닝 프레임 오분류율을 나타낸다. 액션/SF 영화와 같이 움직임이 빠르고 배경음악을 많이 사용하는 영화에서는 오분류율이 높아진다. 그러나 기존의 얼굴검출만으로 프레임을 추출하는 방법에 비해 제안한 방법의 오분류율이 적어짐을 확인할 수 있다.

표 1. 영화장르별 카투닝 프레임 오분류율

	얼굴검출 오분류율	(얼굴검출+음성검색) 오분류율
드라마/멜로	26.2%	10.6%
액션	52.5%	25.7%
SF	55.7%	29.3%
휴먼	22.8%	10.6%
코믹	23.6%	12.1%
스릴러	32.4%	20.7%
공포	31.3%	19.8%

V. 결론

지금까지의 동영상 카투닝은 비사실적 렌더링과 말풍선에 초점이 맞춰졌고, 카투닝을 위한 프레임 추출인 기본적인 장면전환 프레임 추출에 대해서는 기존연구가 진행되지 않았다. 즉, 기존 연구는 영화를 화면에 만화화 하는 것이 주목적 이었다면, 본 논문에서는 비실사 렌더링을 하기 전 프레임 추출을 위한 기초연구가 목적이다. 기존의 카투닝을 위한 장면전환 프레임 추출은 사용자에 의한 반자동 적인 방법을 사용하였다. 그러나 카투닝을 위해서는 우선 적절한 프레임 추출이 우선이다. 따라서 본 논문에서는 동영상에서 비디오의 얼굴 검출과 오디오의 음성영역 추출을 이용하여 동영상 카투닝에 적합한 자동적 프레임 추출 방법을 제안했다는 점에서 새로운 연구 방향이라고 할 수 있다.

향후 연구방향으로는 본 논문에서 제시한 프레임 추출 방법을 위한 조건을 좀 더 확장하고, 영화 혹은 드라마의 사건이 일어날 장소를 미리 보여주는 인물도 음성도 없지만 이야기 전개상 꼭 필요한 장면에 대한 자동적 추출 방법에 대해 더욱 연구해야 한다. 또한 음성영역 추출의 결과를 보다 정밀하게 하여 동영상 카투닝에 보다 적절한 프레임 추출을 할 수 있도록 연구할 예정에 있다.

참고문헌

- [1] 이인권, "디지털 콘텐츠의 만화적인 스타일화", 한국콘텐츠학회지, 제6권, 제3호, pp.33-38, 2008.
- [2] J. Wang, Y. Q. Xu, H. Y. Shum, and M. F. Cohen, "Video Tooning," ACM Transactions on Graphics Vol.23, pp.574-583, 2004.
- [3] 류동성, 조환규, "HSV 색상 모델과 영역 확장 기법을 이용한 동영상 프레임 이미지의 흑백 만화 카툰닝 알고리즘", 정보과학회논문지 : 시스템 및 이론, 제35권, 제12호, pp.560-567, 2008.
- [4] W. I. Hwang, P. J. Lee, B. K. Chun, D. S. Ryu, and H. G. Cho, "Cinema Comics : Cartoon Generation From Video Stream," In Proc. of GRAPP, pp.299-304, 2006.
- [5] J. Preu and J. Loviscach, "From movie to comic, informed by the screenplay," In SIGGRAPH 2007:ACM SIGGRAPH posters, p.99, 2007.
- [6] D. Kurlander, T. Skelly, and D. Salesin, "Comic chat," In Proceedings of SIGGRAPH, ACM Press, pp.225-236, 1996.
- [7] B. K. Chun, D. S. Ryu, W. I. Hwang, and H. G. Cho, "An automated procedure for word balloon placement in cinema comics," The 2nd International Symposium on Visual Computing, pagesII, pp.576-585, 2006.
- [8] R. Lienhart and J. Maydt, "An Extended Set of Harr-like Features for Rapid Object Detection," IEEE ICIP, pp.900-903, 2002.
- [9] 이경록, 류시우, 광재영, "실험에 의한 음성·음악 분류 특징의 비교", 한국콘텐츠학회 추계종합학술대회 논문집, 제2권, 제2호, pp.308-313, 2004.
- [10] P. Viola and M. J. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," IEEE CVPR, Vol.1, pp.511-518, 2001.
- [11] G. Bradski and A. Kaehler, "Learning OpenCV: Computer Vision with the OpenCV Library," Cambridge, MA: O'Reilly, 2008.
- [12] S. U. Jung, "Efficient Rectangle Feature Extraction for Real-time Facial Expression Recognition based on AdaBoost," Master's Thesis, Korea Advanced Institute of Science and Technology, 2005.
- [13] K. G. Derpanis, "Integral image-based representations," Technical report, Department of Computer Science and Engineering, York University, 2007.
- [14] 이원오, 이의철, 박강령, 이희경, 박민식, 이한규, 홍진우, "얼굴 및 눈 위치 추적을 통한 IPTV 화면 인터페이스 제어에 관한 연구", 한국통신학회논문지, 제35권, 제6호, pp.859-993, 2010.

저자 소개

김 대 진(Dae-Jin Kim)

정회원

- 1998년 2월 : 대전대학교 전자공학과 졸업(공학사)
- 2000년 2월 : 동국대학교 전자공학과 졸업(공학석사)
- 2010년 2월 : 대전대학교 전자공학과 졸업(공학박사)

- 2000년 ~ 2003년 : 한빛소프트 주임연구원
 - 2003년 ~ 2007년 : 모토로라 코리아 전임연구원
 - 2007년 ~ 2008년 : 아이비인터넷 부장
 - 2008년 ~ 2010년 : 피어컴(미디어웹) 선임연구원
 - 2010년 ~ 현재 : 다이렉트미디어 기술연구소 연구소장
- <관심분야> : 저작권 보호, 멀티미디어 시스템, 디지털 콘텐츠, 멀티미디어 검색, IPTV, 오디오 핑거프린트, 비디오 카툰닝 등

구 떠올라(Ddeo-Ol-Ra Koo)

정회원

- 2008년 2월 : 남서울대학교 멀티 미디어학과 졸업(공학사)
- 2010년 2월 : 중앙대학교 첨단영상대학원 영상학과 (공학석사)
- 2010년 ~ 현재 : 다이렉트미디어 기술연구소 연구원

<관심분야> : 증강현실, HCI, 3D컴퓨터비전, 디지털 콘텐츠, 오디오 핑거프린트, 비디오 카툰링 등