비디오 셧의 감정 관련 특징에 대한 통계적 모델링

학생회원 박 현 재*, 정회원 강 행 봉*

Statistical Model for Emotional Video Shot Characterization

Hyun-Jae Park* *student member*, **Hang-Bong Kang*** *Regular Member*

요 약

비디오 데이터에 존재하는 감정을 처리하는 것은 지능적인 인간과 컴퓨터와의 상호작용을 위해서 매우 중요한 일이다. 이러한 감정을 추출하기 위해서는 비디오로부터 감정에 관련된 특징들을 검출하기 위한 컴퓨팅 모델을 구축하는 것이 바람직하다. 본 논문에서는 비디오 셧에 존재하는 저급 특징들의 확률적인 분포를 이용하여 감정 이벤트 발생에 관련된 통계학적인 모델을 제안한다. 즉, 비디오 셧의 기본적인 특징을 추출하고 그 특징을 통계적으로 모델화 하여 감정을 유발하는 셧을 찾아낸다. 비디오 셧의 특징으로는 칼라, 카메라 모션 및 셧 길이의 변화를 이용한다. 이러한 특징들을 EM(Expectation Maximization) 알고리즘을 이용하여 GMM(Gaussian Mixture Model)으로 모델링하고, 감정과 시간과의 관계를 MLE(Maximum Likelihood Estimation)를 이용하여 시간에 따른 확률 분포 모델로 구성한다. 이런 두 개의 통계적인 모델들을 융합하여 베이시안 분류법을 적용하여 비디오 데이터로부터 감정에 관련된 셧을 찾아낸다.

Key Words : Video shot modeling; Low level feature; Emotion detection

ABSTRACT

Affective computing plays an important role in intelligent Human Computer Interactions(HCI). To detect emotional events, it is desirable to construct a computing model for extracting emotion related features from video. In this paper, we propose a statistical model based on the probabilistic distribution of low level features in video shots. The proposed method extracts low level features from video shots and then form a GMM(Gaussian Mixture Model) for them to detect emotional shots. As low level features, we use color, camera motion and sequence of shot lengths. The features can be modeled as a GMM by using EM(Expectation Maximization) algorithm and the relations between time and emotions are estimated by MLE(Maximum Likelihood Estimation). Finally, the two statistical models are combined together using Bayesian framework to detect emotional events in video.

Ⅰ. 서 론

효과적인 비디오 검색이나 비디오 요약본 추출은 대용량의 비디오 데이터를 처리하는데 있어 매우 필요한 일이다. 단순한 특징 레벨의 검색이나 의미에 관련된 검색에 대하여 많은 연구가 진행되고 있지만, 사람의 지각에 기반한 검색에 대한 연구는 최근 들어 시작되고 있다 [1]. 특히, 사람이 느끼는 감정을 처리할 수 있는 검색이 가능하다면, 교육 및 엔터테인먼트 분야 등 보다 많은 분야에서 활용이 가능해진다.

비디오로부터 사용자가 느끼는 감정을 분류하기 위해서는 감정에 관련된 저급 특징들로부터 고급 개념인 감정과의 관계를 추출하는 것이 필요하다[2-3]. 이런 관계를 추출할 수 있는 방법 중 하나는 비디오로부터 감정에 관련된 특징들을 찾아내고 이러한 특징들의 검출 및 특성을 기술 할 수 있는 저급 특징들에 대한 확률적인 분포를 이용하여 감정 이

* 가톨릭대학교 컴퓨터정보공학부 지능형 멀티미디어 시스템 연구실{hyunjapark, hbkang}@catholic.ac.kr

논문번호 : 030400-0915, 접수일자 : 2003년 9월 15일

※ 본 연구는 2003년도 가톨릭대학교 교비연구비 지원으로 수행되었음.

벤트 발생에 관련된 통계학적인 모델을 개발하는 것이다.

비디오 셧은 연속된 프레임의 집합으로 비디오 데이터를 구성하는 최소 단위이며, 비디오 관련 정보를 처리하는데 있어서 매우 중요한 단위이다. 비디오 셧의 여러 특징에 대한 확률 분포 모델은 셧 경계 검출이나 비디오 장르 구분 또 멀티 미디어 네트워크 상에서 비디오 트래픽 모델링 등 다양한 분야에 응용이 가능하다[4-6]. 하지만, 감정은 일반적으로 다양한 특징을 가지고 있어서, 모델링 하는 문제가 매우 까다롭다.

본 논문에서는 비디오 셧이 가지고 있는 저급 특징들인 칼라, 모션 및 셧의 길이의 확률 분포를 정확히 모델링 하여 감정 비디오 셧을 검출하는 방법을 제안한다. 제 2장에서는 본 논문과 관련 있는 연구들에 대하여 살펴본다. 그리고 제 3장에서는 비디오 데이터의 감정 관련 특징을 기술하고, 이 특징들을 감정 별로 모델링 하는 기법에 대하여 설명한다. 제 4장에서는 감정 검출을 위한 베이시안 분류법에 대하여 기술한다. 제 5장에서는 제안한 통계적 모델을 사용한 실험 결과를 분석한다.

II. 기존 연구

비디오 셧의 특징에 대한 통계적 모델링은 주로 셧의 경계 검출, 비디오 장르 구분 및 멀티미디어 네트워크상의 트래픽 모델에 적용되었다. 셧의 경계를 검출하기 위해서는 주로 비디오 프레임에 존재하는 칼라나 모션 등의 저급 특징들을 계산하여 연속된 프레임들 사이에서 이들 특징들의 차이가 많이 나는 곳을 셧의 경계로 정한다. 하지만, 이러한 고정된 임계값을 사용하는 방법은 비디오에 존재하는 다양한 특징들의 변화에 좋은 결과를 내기 어렵다. Vasconcelos et al.[4]은 셧 경계 검출을 위한 적응적인 임계값을 얻기 위해 셧의 길이나 움직임의 확률 분포를 모델링 하여 베이시안 프레임워크에서 사전 지식으로 사용하였다.

비디오 장르 구분 역시 셧에 관련된 정보를 이용하여 분류가 가능하다[4-5]. 또, 가변적인 비트율(Variable Bit Rate : VBR)로 압축된 비디오의 정확한 트래픽 모델은 기존의 큐잉 이론과는 잘 맞지 않는다. VBR의 트래픽을 모델링 하기 위해서는 셧의 길이 정보의 확률 분포로 모델링 하는 것이 바람직하다[6].

비디오로부터 감정에 관련된 이벤트를 검출하는 연구는 최근 들어 연구되고 있다. Moncrieff et al.[7]은 감정에 관련된 사운드 이벤트를 검출하기 위해 사운드의 에너지 역학을 이용하여 감정 이벤트를 검출하는 방법을 제안하였다. 제안된 방식을 이용하여 공포 감정을 갖는 장면을 비디오 데이터로부터 추출하였다. Hanjalic et al.[8]은 모션 및 사운드 정보를 이용하여 사용자의 감정 곡선을 생성하였다. 생성된 감정 곡선을 이용하여 사용자의 지각을 분석하였다.

본 논문에서는 비디오 셧의 저급 정보로부터 확률적인 분포를 모델링 하여 감정 이벤트를 검출하는 방법을 제안한다.

III. 감정에 따른 특징 모델링

본 논문에서 시도하는 감정 검출 기법은 비디오 데이터에서 추출할 수 있는 특징들을 확률적인 모델로 구성하고, 통계적인 분석을 통해 생성된 확률 모델들을 조합하여 분류하고자 하는 비디오 셧이 감정을 발생하는 셧인지 아닌지를 판단하게 된다. 이 때 사용하는 통계적인 모델을 두 가지로 나눠볼 수 있다. 하나는 시간에 따른 감정 발생 여부를 모델링 하는 것이고, 또 다른 하나는 셧 내에서의 특징값을 모델링 하는 것이다. 이러한 프레임워크는 Vasconcelos et al.이 제안하였으며[4], 본 논문에서는 이러한 모델링 기법을 변형하여 감정 셧 검출을 시도한다.

3.1. 감정 발생과 저급 특징과의 관련

비디오 데이터를 이용하여 감정 분류를 구현하기 위하여 비디오의 특징값을 계산한다. 비디오의 저급 특징으로는 칼라 정보와 모션 정보, 셧의 길이를 사용한다. 이를 정리하면 다음과 같다.

$$F=[\ f_C,\ f_M,\ f_L\ ]^T \qquad (1)$$

f_C 는 칼라 정보를 이용한 특징 벡터이며, f_M 은 모션 정보를 이용한 특징 벡터이다. f_L는 셧 길이의 시퀀스이다.

칼라 정보를 이용한 특징 벡터는 다음과 같다.

$$f_C = [f_{S-Cont}, f_{L-Cont}, f_{dominColor}]^T \quad (2)$$

f_{S-Cont}는 채도의 대비를 나타내는 값이며, f_{L-Cont}는 밝기의 대비를 나타낸다. 영상의 밝기와 채도는 영상의 분위기를 결정하는 영상의 성질이다. $f_{dominColor}$는 영상 내에 하나의 지배적인 색의 존재 여부를 나타내는 값이다. 영상 내에 여러 가지 색이 혼합되어 있는지 그렇지 않은지를 나타내는 값으로 영상의 일정 비율 이상을 차지하는 Connected-component가 존재하면 1, 그렇지 않으면 0을 갖는다.

f_M은 특정 셋에서의 카메라 모션을 나타낸다. 카메라 모션은 다음과 같이 정의된다.

$$f_M = [f_{pan}, f_{tilt}, f_{zoom}, f_{activity}]^T \quad (3)$$

f_{pan}, f_{tilt}, f_{zoom}, $f_{activity}$는 각각 pan, tilt, zoom, shot activity를 나타내며 [10]에서 제안한 템플릿 정합법을 이용하여 카메라 모션의 종류를 결정한 뒤 움직임의 크기를 계산하여 결정된다. 이러한 방식으로 카메라 모션의 종류를 결정하게 되면 결정된 카메라 모션을 제외한 나머지는 값을 가질 수 없으므로 0으로 결정되며, 결정된 카메라 모션의 크기는 영상의 전 영역에 대하여 옵티컬 플로우($\mathbf{v}(i,j)$)를 구하여 합을 계산하게 된다.

$$f_{pan\,|\,tilt\,|\,zoom} = \sum_i \sum_j \| \mathbf{v}(i,j) \| \quad (4)$$

pan, tilt, zoom, 세 가지의 카메라 모션 중 어느 것도 아니라면 모션이 없다고 가정하고 셋 내에서 프레임 차이를 구하여 $f_{activity}$를 구한다.

f_L은 셋의 길이로 영화의 성격을 나타내주는 간결하면서도 좋은 특징이 된다.

$$f_L = [f_L^t, f_L^{t-1}, \cdots, f_L^{t-9}] \quad (5)$$

Vasconcelos et al. 은 이러한 성질을 이용해 영화의 장르 구분을 시도하였다[4].

3.2. 시간에 따른 감정 변화 모델링

비디오 데이터에서 발생되는 감정은 대체로 시간에 따라 변화한다. 일정 시간 간격을 두고 발생하며, 한 번 발생한 감정은 어느 정도 지속하게 된다. 이러한 성질을 모델링하기 위하여 몇 가지 모델을 도입하여 통계적인 분석을 한다. 시간에 따라 발생할 확률이 변화하는 경우 Poisson, Erlang, Weibull, Exponential 과정 등이 사용된다[4]. 이러한 시간에 따른 확률 모델들을 통계적으로 분석하여 모델링하고, 이를 바탕으로 감정을 발생하는 부분들을 찾아내고자 한다.

시간에 따라 감정의 발생 여부를 확률적으로 나타낼 수 있는 확률 모델은 poisson process, Erlang process, Weibull process 등이 있으며, MLE를 이용하여 주어진 샘플 집합에 적합한 모델 파라미터를 추정할 수 있다[4].

MLE는 널리 사용되는 파라미터 추정 방법으로서 주어진 샘플 분포의 확률을 최대로 만들어주는 모델 파라미터를 계산한다. MLE는 다른 파라미터 추정법보다 단순하고 계산이 용이하다는 장점을 가진다[11].

3.3. 셋의 특징값 모델링

주어진 특징 벡터 집합으로부터 추정해낼 수 있는 likelihood를 구하기 위해서는 주어진 벡터 집합을 발생시킬 확률이 가장 높은 모델 파라미터를 구해야 한다. 이 과정이 특징 모델링 과정이며 본 논문에서는 셋 내의 특징을 모델링 하기 위하여 GMM을 사용한다. 모델링의 목적은 주어진 샘플들의 분포를 잘 표현하는 확률 분포를 얻는 것이다. 그러나, 현실적으로 샘플들의 분포를 하나의 모델로 표현한다는 것은 불가능하므로 GMM을 사용한다. 현실에 존재하는 데이터들은 비슷한 성질에 따라 군집화 될 수 있으며 이 것은 가우시안 모델로 표현하기 적합하다. 특징 공간상에서 무질서하게 보여지는 샘플들의 분포도 지역적으로 작은 가우시안 모델 혹은 그와 유사한 변형된 모델을 형성한다. 본 논문에서는 특징 공간상에 분포하는 샘플들을 모델화 하기 위하여 가우시안 믹스춰 모델을 사용한다.

모델링 과정은 주어진 파라미터 공간상의 한 점에서 지역 최적점(local optimal region)으로 파라미터 벡터를 이동하는 과정으로 기울기 방법(Gradient Ascent method)과 EM(Expectation Maximization) 방법이 널리 사용된다. 본 논문에서는 EM 방식을 채택한다.

Ⅳ. 통계적 모델들을 이용한 베이시안 분류법

3장에서 소개한 특징 모델링을 이용하여 학습 샘플들의 분포를 바탕으로 구성된 모델은 새로운 샘플의 확률을 도출하고 그 값을 베이시안에 이용한다. 베이시안 분류법은 확률에 기초한 분류 방식으로 특징 모델링이 정확하다면 가장 이상적인 분류 방식이며 사전 확률비 테스트(Likelihood ratio test)를 이용한다. i번째 셧의 특징 벡터 F_i가 주어졌을 때 감정 발생 여부는 다음과 같다.

$$\begin{cases} H_1 & if \ \dfrac{P(F_i \mid S_i = 1)}{P(F_i \mid S_i = 0)} \geq 1 \\ H_0 & otherwise \end{cases} \quad (6)$$

위 식을 보면 사전 확률비 $\dfrac{P(F_i \mid S_i = 1)}{P(F_i \mid S_i = 0)}$는 전적으로 F_i의 분포에 의존한다. 그러나, 이러한 가정은 실제 비디오 데이터에서 발생하는 감정 발생을 모델링하기 어렵다. 이러한 가정은 감정 발생이 시간에 따라 변화한다는 성질을 고려하지 않는다. 따라서, 이러한 특성을 반영한 확률 모델이 구성되어야 한다. 본 논문에서는 감정 발생과 시간 간격과의 관계를 통계적으로 모델링 하여 시간에 따라 변화하는 감정 발생 여부를 예측하고자 한다.

비디오 데이터는 셧으로 최소한의 의미 있는 영상을 보여주게 되므로 셧 단위로의 특징 추출을 수행한다. 연속된 셧으로 구성된 비디오 데이터는 이산적인 과정이며, 셧의 감정 여부를 나타내는 확률 변수를 S라 하고 t번째 셧에서 감정 여부를 나타내는 확률 변수는 S_t로 나타낸다.

t번째 셧에서 감정을 유발하는 경우, $S_t = 1$로 나타내며, 그렇지 않을 경우 $S_t = 0$이다. i번째 셧에서 감정이 발생하고, e개 전의 셧에서 비감정에서 감정으로의 천이(transition)가 발생한 경우의 사후 확률을 다음과 같이 나타낸다.

$$P(S_i = 1 \mid S_{i-e-1} = 0, S_{i-e} = 1) \quad (7)$$

위와 유사하게 i번째 셧에서 감정이 발생하고, r개 전의 셧에서 감정으로부터 비감정으로의 천이가 발생한 경우, 사후 확률은 다음과 같다.

$$P(S_i = 1 \mid S_{i-r-1} = 1, S_{i-r} = 0) \quad (8)$$

식 (7), (8)을 종합하여, i번째 셧에서 감정이 발생하고, i-e에서 비감정에서 감정으로의 천이가 발생하고 i-r에서 감정에서 비감정으로의 천이가 발생할 확률은 다음과 같다.

$$P(S_i = 1 \mid S_{i-e-1} = 0, S_{i-e} = 1, S_{i-r-1} = 1, S_{i-r} = 0, F_i) \quad (9)$$

이 때 $\mathbf{F}_i$는 i번째 셧의 특징 벡터이다. 위의 과정과 동일한 조건하에 i번째 셧이 비감정일 확률을 구해보면 다음과 같다.

$$P(S_i = 0 \mid S_{i-e-1} = 0, S_{i-e} = 1, S_{i-r-1} = 1, S_{i-r} = 0, F_i) \quad (10)$$

두 사후 확률 (9), (10)을 이용하여 베이시안 분류법을 적용하면 식 (11)을 구성할 수 있다.

$$\frac{P(S_i = 1 \mid S_{i-e-1} = 0, S_{i-e} = 1, S_{i-r-1} = 1, S_{i-r} = 0, F_i)}{P(S_i = 0 \mid S_{i-e-1} = 0, S_{i-e} = 1, S_{i-r-1} = 1, S_{i-r} = 0, F_i)} \quad (11)$$

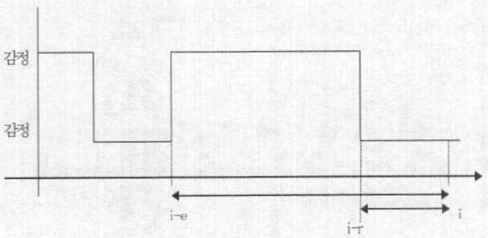


그림 1 : 시간 간격 i와 r

i번째 셧에서의 감정 여부의 판별은 (11)번 식으로부터 시작한다. i번째 셧이 감정인지 비감정 셧인지는 시간 간격 r과 e에 의존한다. 그림 1은 시간 간격 e와 r에 대하여 보여준다. i-e번째 셧은 최근의 비감정 영역에서 감정 영역으로 변화한 후 첫 번째 셧이며, i-r은 최근 감정 영역에서 비감정 영역으로 변화한 후 첫 번째로 나타난 셧이다. (11)은 다음과 같이 정리 할 수 있다.

$$\frac{P(S_{i-e-1} = 0, S_{i-e} = 1 \mid S_i = 1) P(S_{i-r-1} = 1, S_{i-r} = 0 \mid S_i = 1)}{P(S_{i-e-1} = 0, S_{i-e} = 1 \mid S_i = 0) P(S_{i-r-1} = 1, S_{i-r} = 0 \mid S_i = 0)} \cdot \frac{P(F_i \mid S_i = 1) P(S_i = 1)}{P(F_i \mid S_i = 0) P(S_i = 0)} \quad (12)$$

식 (12)에 연산의 용이성을 위하여 로그를 취하면 식 (13)를 구할 수 있다.

$$\log\left(\frac{P(S_{i-e-1}=0,S_{i-e}=1|S_i=1)}{P(S_{i-e-1}=0,S_{i-e}=1|S_i=0)}\right)$$
$$+\log\left(\frac{P(S_{i-r-1}=1,S_{i-r}=0|S_i=1)}{P(S_{i-r-1}=1,S_{i-r}=0|S_i=0)}\right)+\log\left(\frac{P(F_i|S_i=1)}{P(F_i|S_i=0)}\right)$$
$$+\log\left(\frac{P(S_i=1)}{P(S_i=0)}\right) \quad (13)$$

식(13)의 각 항을 아래와 같이 간단한 표현으로 나타낼 수 있다.

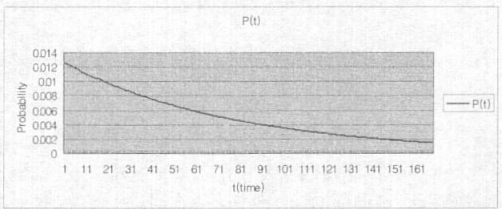
$$g(e)+h(r)+f(\boldsymbol{F}_i)+C \quad (14)$$

$$g(e)=\log\left(\frac{P(S_{i-e-1}=0,S_{i-e}=1|S_i=1)}{P(S_{i-e-1}=0,S_{i-e}=1|S_i=0)}\right)$$
$$h(r)=\log\left(\frac{P(S_{i-r-1}=1,S_{i-r}=0|S_i=1)}{P(S_{i-r-1}=1,S_{i-r}=0|S_i=0)}\right)$$
$$f(\boldsymbol{F}_i)=\log\left(\frac{P(F_i|S_i=1)}{P(F_i|S_i=0)}\right)$$
$$C=\log\left(\frac{P(S_i=1)}{P(S_i=0)}\right)$$

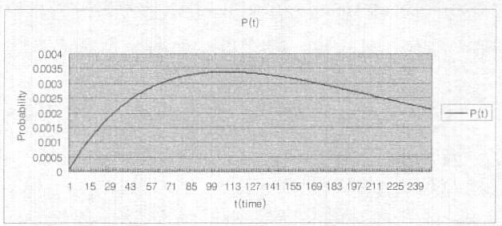
이를 이용하여 (6)과 같은 형식으로 표현하면 다음과 같다.

$$\begin{cases} S_i=1 & \text{if } f(F_i)\geq -(g(e)+h(r)+C)=T(r,e) \\ S_0=0 & otherwise \end{cases} \quad (15)$$

$g(e)$는 비감정 영역에서 감정 영역으로의 천이가 일어나고 e개의 셧 이후에 감정 영역이 일어날지에 대한 시간 분포 함수이다. 그림 2는 $g(e)$의 한 예를 보여준다.



(a)

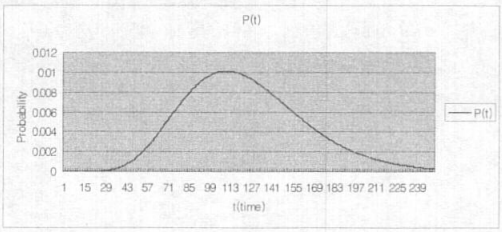


(b)

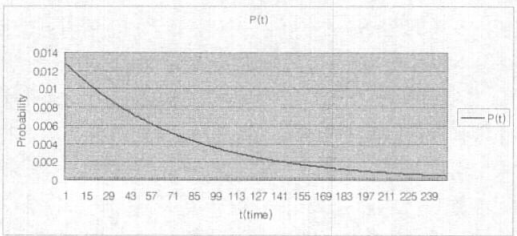
그림 2 : $g(e)$의 구성 요소

그림 2의 (a)는 $P(S_{i-e}=1,S_{i-e-1}=0|S_i=1)$를, (b)는 $P(S_{i-e}=1,S_{i-e-1}=0|S_i=0)$을 통계적인 모델로 구성한 예이다.

$h(r)$은 감정 영역에서 비감정 영역으로의 천이가 일어나고 r개의 셧 이후에 감정 영역이 나타나는 것에 대한 시간 분포 함수이다. 그림 3은 $h(r)$의 한 예를 보여준다.



(a)



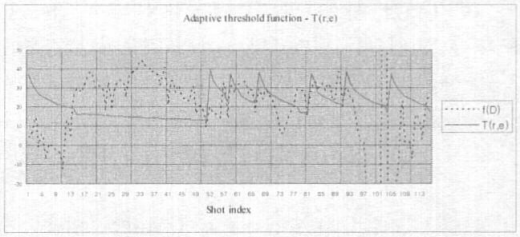
(b)

그림 3 : $h(r)$의 구성 요소

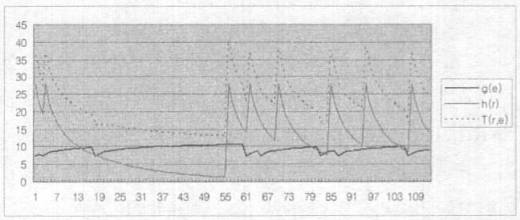
그림 3의 (a)는 $P(S_{i-r}=0,S_{i-r-1}=1|S_i=1)$를, (b)는 $P(S_{i-r}=0,S_{i-r-1}=1|S_i=0)$을 모델로 구성한 예이다.

$T(r,e)$ 는 시간 간격 r과 e의 함수이고, r과 e는 비감정에서 감정으로 변화하는 시간 t_e와 감정에서 비감정으로 변화하는 시간 t_r에 의해 결정된다. $T(r,e)$는 비감정에서 감정 영역으로 변화가 있은 후 높은 확률을 유지하다가 일정 시간이 지나면 값이 작아지게 된다. 또, 감정에서 비감정 영역으로 변화가 있은 후에는 낮은 확률을 유지하다가 일정 시간이 지나면 높은 확률이 나타나게 된다.

이는 비디오 데이터들이 주기적으로 일정 시간동안 감정을 발생시키는 셧들로 구성된 성질을 이용한 것이다. 이러한 시간 분포 모델을 도입함으로써 얻을 수 있는 이득은 일정 시간 동안 연속적으로 나타나는 감정 영역을 끊지 않고 연속적으로 검출해 낼 수 있다는 점이다. 그림 4를 보면 이러한 예를 볼 수 있다. 각 그래프는 식 (13)를 간략하게 표현한 식 (14)의 항들을 도시한 것이다.



(a)



(b)

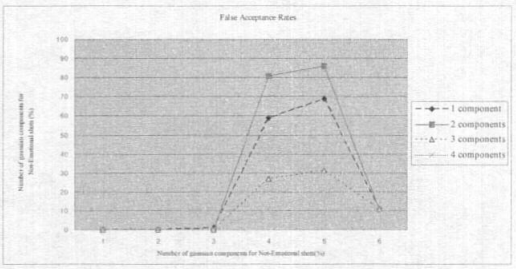
그림 4 : 임계 함수의 구성

그림 4의 (b)를 보면 $T(r,e)$를 구성하는 각 항들이 나타나 있다. 13번 셧에서 감정 영역이 발생하였음을 볼 수 있다. 감정 영역이 발생하면 $g(e)$의 값이 떨어지는 것을 볼 수 있다. 이렇게 $g(e)$는 감정이 발생하면 값이 급격하게 떨어졌다가 일정 시간이 지나면 다시 상승하여 감정 영역을 지속하는 효과를 낳는다. 이것은 감정 영역이 지속되는 것을 모델링 한 결과이다. 그리고 52번에서 감정 영역이 끝나는 것을 볼 수 있는데, 이 때 $h(r)$이 급격하게 상승하여 새로운 감정이 발생할 수 없게 하는 효과를 낳는다. $g(.)$, $h(.)$는 이렇게 시간에 따라 임계 함수를 조절한다. 그림 4의 (a)를 보면 실제 임계 함수 $T(r,e)$와 특징값의 함수 $f(F_i)$를 볼 수 있으며 $f(F_i)$가 $T(r,e)$보다 높은 영역 13~52번 셧이 감정 영역으로 분류된다.

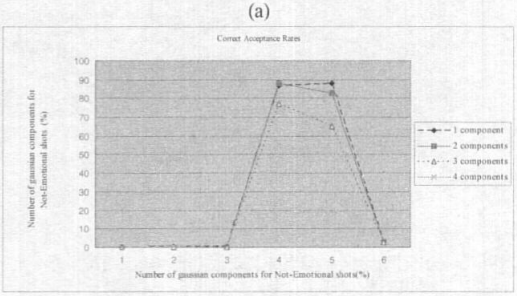
V. 실험 결과 및 분석

본 논문에서 제안된 방법을 이용하여 15개의 영화(I know what you did last summer part 1/ part 2, Scream, Dying Young, Autumn in New York, Home alone, Mask part 1/part 2, Jurasic Park part 1/ part 2, Ring part 1/part 2, Titanic, Saving Private Ryan, When Harry met Sally)를 실험하였다. 각 영화 데이터에 대하여 10명의 사용자들이 감정을 표시하여 7명 이상이 같은 감정을 가질 때 감정 셧으로 분류하였다. 실험은 leave-one-out 방식을 사용한다. 1개의 영화를 테스트하기 위하여 나머지 데이터들을 모두 학습 데이터로 이용하는 방식이다.

실험은 각 영화에 대하여 테스트하며, 특징 모델링을 위한 GMM의 가우시안 컴포넌트의 개수에 따라 변화하는 성능을 실험하였다. 특징을 모델링하는 부분은 두 부분으로 감정 셧을 모델링 하는 부분과 비감정 셧을 모델링 하는 부분이다. 각 부분에서 모델 컴포넌트의 개수를 달리하여 실험하였다.



(a)



(b)

그림 5 : 비감정 셧의 모델링을 위한 가우시안 컴포넌트 개수에 따른 성능 평가

두 번째 실험은 제안한 베이시안 분류식과 다른 방식의 분류식을 이용하였을 경우의 차이를 실험하였다. 본 논문에서 제안하는 분류 방식은 현재 셧의 감정 발생 여부가 최근 일어났던 감정의 시작부터 현재까지 흘러간 셧 개수에 의존한다는 가정과 최근 감정이 사라진 때부터 현재 셧까지의 시간 간격에 의존한다는 두 개의 가정에 기반한다. 이러한 의존 관계를 달리하여 성능을 평가한다. 마지막으로 특징 벡터의 구성 요소들을 변화시켜 전체적인 결과가 어떻게 변화하는지 살펴보았다.

그림 5는 비감정 비디오셧을 GMM을 이용하여 모델링 할 때 사용되는 가우시안 컴포넌트의 개수의 변화에 따른 결과이다. 실선과 점선으로 표시된 선들은 각각 감정셧들을 모델링 하기 위한 가우시안 컴포넌트의 개수 별로 도시되어 있다. 그림 5의 (a)를 보면 FAR(False Acceptance Rate)이 나타

난다. Gaussian Component의 개수가 3개까지는 감정 셧 검출에 실패하여 FAR이 0인 것을 볼 수 있으며 3 이후에 4~5에서 올바른 검출이 이루어졌음을 (a)와 (b)에서 확인할 수 있으며 4에서 5보다 낮은 FAR을 보인다. 그림 5의 (b)는 가우시안 컴포넌트의 개수에 따라 변화하는 Correct Acceptance Rate을 확인할 수 있으며 4에서 5보다 좋은 결과를 얻었음을 볼 수 있다. 이 결과를 그림 6에서 볼 수 있다.

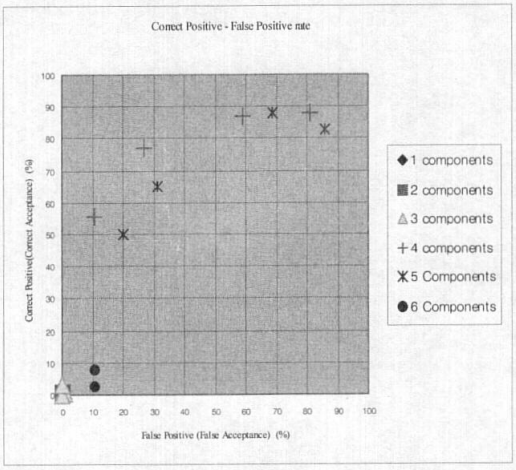
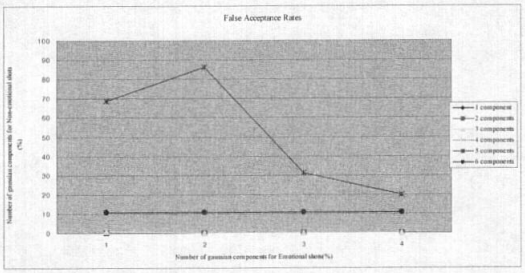
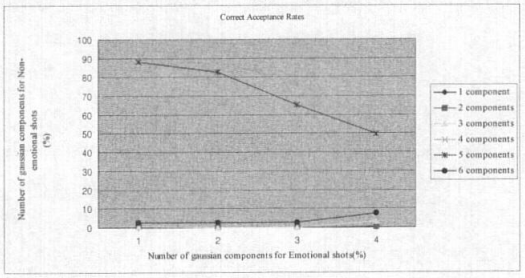


그림 6 : 비감정 비디오 셧의 모델링을 위한 컴포넌트 개수에 따른 성능 비교



(a)



(b)

그림 7 : 감정 비디오 셧의 모델링을 위한 가우시안 컴포넌트 개수 변화에 따른 성능 평가.

그림 6에서 시스템의 성능은 좌상단으로 접근할수록 좋은 결과를 나타낸다. 4개의 가우시안 컴포넌트를 사용하는 경우 5보다 더 좌상단으로 접근함을 볼 수 있으며, 따라서 4개의 가우시안 컴포넌트를 사용하는 경우 비감정셧의 특징을 적합하게 모델링한다.

그림 7은 감정 비디오 셧을 GMM을 이용하여 모델링 시 사용되는 가우시안 컴포넌트의 개수의 변화에 따른 결과이다. 그래프는 비감정 셧들을 모델링 하기 위한 가우시안 컴포넌트의 개수 별로 실선과 점선으로 도시되어 있다. 그림 7의 (a)를 보면 FAR(False Acceptance Rate)이 2 이후에 급격히 하강하는 것을 볼 수 있다. 반면 (b)를 보면 Correct Acceptance Rate은 하강하는 속도가 (a)보다 느린 것을 알 수 있다. 성능이 좋게 나오는 두 개의 선은 각각 컴포넌트의 개수가 4, 5인 경우이며 이 때 GMM을 이용한 모델링이 좋은 결과를 내는 것을 알 수 있다. 이 결과를 그림 8에서 볼 수 있다.

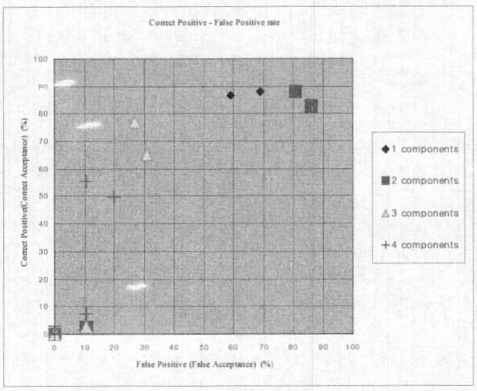


그림 8 : 감정 비디오 셧의 모델링을 위한 컴포넌트의 개수에 따른 성능 비교

그림 9는 임계 함수를 구성하는 항들의 변화에 따른 결과를 보여준다. 그림 9에는 다음 5가지 경우의 결과들이 도시되어 있다.

$$T = 0 \tag{16}$$

$$T = C \tag{17}$$

$$T = g(e) + C \tag{18}$$

$$T = h(r) + C \tag{19}$$

$$T = g(e) + h(r) + C \tag{20}$$

(16)부터 (20)까지의 임계 함수는 다음과 같은 과정을 통해 얻어진 결과이다. (21)을 Si=1 이 될 조건으로 하고 이 때 prior probability를 모른다고 가정하면 (16)번의 임계 함수를 얻을 수 있다. (22)와 같이 식을 구성하고 prior probability를 안다고 가정하면 (17)과 같이 임계 함수를 얻을 수 있다. (23)과 같이 Si=1 일 확률이 Si-e , Si-e-1 에 의존한다고 가정하면 (18)과 같이 임계 함수를 얻을 수 있다. 마찬가지로 현재 Si의 상태가 최근 일어났던 감정에서 비감정으로의 천이 이후 흐른 시간에 의존한다 가정할 경우 (24)와 같이 조건식을 구성할 수 있고, 정리하면 (19)와 같이 임계 함수를 설정할 수 있다. 마지막으로 본 논문에서 제안한 방식인 (25)와 같이 조건식을 구성하면 임계 함수를 (20)과 같이 얻을 수 있다.

$$\frac{P(S_i=1|\,\boldsymbol{F}_i)}{P(S_i=0|\,\boldsymbol{F}_i)} \geq 1 \qquad (21)$$

$$\frac{P(S_i=1|\,\boldsymbol{F}_i)}{P(S_i=0|\,\boldsymbol{F}_i)} = \frac{P(\boldsymbol{F}_i|S+i=1)P(S_i=1)}{P(\boldsymbol{F}_i|S+i=0)P(S_i=0)} \geq 1 \qquad (22)$$

$$\frac{P(S_i=1|S_{i-e-1}=0,\, S_{i-e}=1,\, \boldsymbol{F}_i)}{P(S_i=0|S_{i-e-1}=0,\, S_{i-e}=1,\, \boldsymbol{F}_i)} \geq 1 \qquad (23)$$

$$\frac{P(S_i=1|S_{i-r-1}=1,\, S_{i-r}=0,\, \boldsymbol{F}_i)}{P(S_i=0|S_{i-r-1}=1,\, S_{i-r}=0,\, \boldsymbol{F}_i)} \geq 1 \qquad (24)$$

$$\frac{P(S_i=1|S_{i-e-1}=0,\, S_{i-e}=1,\, S_{i-r-1}=1,\, S_{i-r}=0,\, \boldsymbol{F}_i)}{P(S_i=0|S_{i-e-1}=0,\, S_{i-e}=1,\, S_{i-r-1}=1,\, S_{i-r}=0,\, \boldsymbol{F}_i)} \geq 1 \qquad (25)$$

얻어진 각각의 임계 함수(16)~(20)을 이용하여 감정 셧 검출을 한 결과는 그림 9와 같다. 그림 9를 보면 (16), (17)을 사용한 경우보다 (18)~(20)을 사용하는 경우 더 좋은 결과를 얻을 수 있음을 알 수 있다.

그림 10은 특징의 변화에 따라 결과가 어떻게 나타나는지를 보여준다. 11개의 특징을 사용한 경우와 14개, 17개로 실험하였다. 11개의 특징은 현재부터 이전 10개의 셧 길이와 현재 셧의 Activity를 사용하였다. 14개의 특징은 앞의 11개의 특징에 칼라 정보 3개를 추가하였다. 17개의 특징은 앞의 11개의 특징에 3개의 칼라 정보와 3개의 카메라 모션을 추가하였다. 결과를 보면 특징이 늘어남에 따라 좋은 결과를 나타내는 것을 볼 수 있다.

감정 분류 시스템의 전체적인 성능을 보면 편차가 심하다는 것을 알 수 있다. 시스템의 변수에 따라 많은 차이를 보인다. 그러나 변수 설정에 따라 FAR(False Acceptance Rate)이 30%대인 경우 80%에 근접한 Correct Acceptance Rate을 얻을 수 있으며, 40%대에서 90%에 근접한Correct Acceptance Rate을 얻을 있다(그림 9 참조)

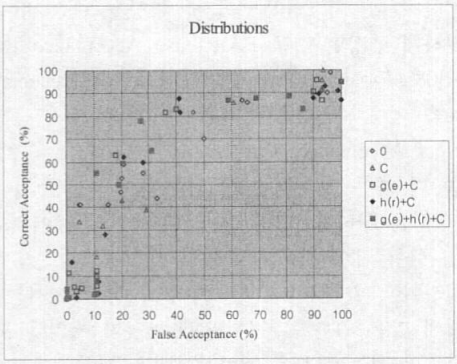


그림 9: 임계 함수를 구성하는 항들의 변화에 따른 결과

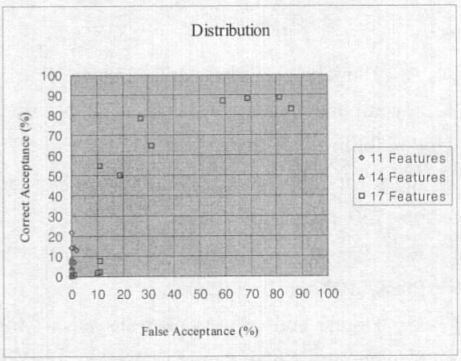


그림 10 : 특징의 변화에 따른 결과

VI. 결론

비디오 셧 단위로 감정 부분을 검출하는 방법은 기본적으로 비디오 셧의 특징들을 기반으로 이루어진다. 비디오 셧으로부터 추출하는 특징으로는 칼라 정보와 카메라 모션 정보, 셧의 길이 정보 등이 있다. 이러한 특징에 따라 감정 발생 셧을 추출한다. 그리고 시간에 따라 감정 셧의 지속 여부와 발생 빈도를 예측하는 방식으로 감정 발생 셧을 추출한다.

감정 발생 여부를 판단하는 방식은 확률에 근거한 베이시안 분류법이다. 특징값과 감정 발생과의 관계를 확률 모델로서 추정한 후 베이시안 분류에 적용하여 특정 특징값이 입력되었을 때 확률을 계산하여 감정 발생 여부를 결정하게 된다. 특징값 모델링을 위하여 사용되는 모델은 GMM이며 EM 알고리즘을 통하여 모델의 파라미터를 추정한다. 그리고 감정과 시간의 관계를 확률적으로 모델링 할 때의 파라미터 추정법은 MLE이다.

실험 결과 GMM 가우시안 컴포넌트의 개수와 특

징값 등 변수에 따라 편차가 있다는 단점이 있다. 그러나 분류 시스템의 성능에 영향을 주는 변수를 적절하게 설정할 경우 FAR(False Acceptance Rate)이 30%대에서 80%에 근접한 Correct Acceptance Rate을 얻을 수 있으며, 40%대에서 90%에 근접한 Correct Acceptance Rate을 얻을 있다.

좀 더 좋은 결과를 얻기 위해서는 적절한 파라미터 설정을 구하는 방법이 연구되어야 할 것이며, 더 많은 실험 데이터를 확보하여 편차가 적은 모델을 생성하여야 할 것이다. 또 비디오 데이터에서 감정 발생에 많은 영향을 주는 음향 효과의 특징을 도입함으로써 성능 향상을 꾀할 수 있을 것이다.

참 고 문 헌

[1] A. Hanjalic, "Video and Image Retrieval beyond the Cognitive Level: The Needs and Possibilities," *Proc. SPIE Storage and Retrieval for Media Databases*, San Jose, pp. 130-140, 200

[2] R. Picard, *Affective Computing*, MIT Press, 1997

[3] C. Dorai and S. Venkatesh eds., *Media Computing: Computational Media Aesthetics*, Kluwer Academic Publishers, 2002.

[4] Nuno Vasconcelos and Andrew Lippman, "Statistical Models of Video for Content Analysis and Characterization", *IEEE Trans. Image Processing*, vol. 9, pp. 3-19, Jan. 2000

[5] C. Taskiran, C. Bouman, and E. Delp, "Discovering video structure suing the pseudo-semantic trace", *Proc. SPIE Storage and Retrieval for Media Databases 2001,* Jan. 2001, pp.571-578

[6] U. Sarkar, S. Ramakrishnan and D. Sarkar, "Segmenting full-length VBR video into shots for modeling with markov-modulated gamma-based framework", *SPIE ITCom 2001,* 19-24, Aug. 2001

[7] S. Moncrief, C. Dorai and S. Venkatesh, "Affect Computing in Film through Sound Energy Dynamics", *Proc. ACM MM'01*, pp525-527, 2001

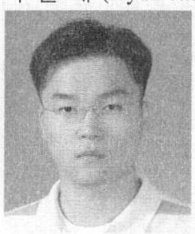
[8] A. Hanjalic and L. Xu, "User-oriented Affective Video Content Analysis", *Proc. IEEE Workshop on CBAIBL '01,* Kauai, HI, pp50-57, Dec 2001.

[9] E. Goldstein, *Sensation and perception*, Brooks/Cole, 1999.

[10] Sangkeun Lee, Hayes M.H., III, "Real-time camera motion classification for content based indexing and retrieval using templates", *Proc. ICASSP '02* , Volume: 4 , 2002 , pp.3664 -3667

[11] Richard O. Duda, Peter E. Hart, David G. Stork, *Pattern classification 2nd Ed.*, Wiley-interscience,

박 현 재 (Hyun-Jae Park) 학생회원



2002년 2월 : 가톨릭 대학교 컴퓨터공학과 졸업

2002년 3월~현재:가톨릭대학교 컴퓨터공학과 석사과정

<주관심분야> 패턴 인식, 기계 학습, 모델 기반 영상 처리

강 행 봉 (Hang-Bong Kang) 정회원

1980년 : 한양대학교 전자공학과 졸업

1986년 : 한양대학교 전자공학과 졸업(석사)

1989년 : 미국 Ohio State University 컴퓨터 공학 석사

1993년 : 미국 Rensselaer Polytechnic Institute 컴퓨터 공학 박사

1994년 ~ 1997년 : 삼성 종합기술원 수석 연구원

1997년 3월~현재 : 가톨릭 대학교 컴퓨터정보공학부 부교수

<주관심분야> 컴퓨터 비전, 멀티 미디어 시스템, 인공 지능, 생체 인식 및 Bioinformatics