Journal of Korean Institute of Intelligent Systems, Vol. 24, No. 3, June 2014, pp. 316-322
http://dx.doi.org/10.5391/JKIIS.2014.24.3.316

CogTV를 위한 생체신호기반 시청자 선호도 모델

A Viewer Preference Model Based on Physiological Feedback

박태서* · 김병희** · 장병탁*,**,†
Tae-Suh Park*, Byoung-Hee Kim**, and Byoung-Tak Zhang*,**,†

*서울대학교 인지과학협동과정, **서울대학교 컴퓨터공학부
*Cognitive Science Program, **† School of Computer Science & Engineering
Seoul National University

요 약

본 논문은 TV를 이용한 영화시청 환경에서 해당 컨텐트에 대한 시청자의 암묵적 반응과 컨텐트의 멀티모달 피쳐를 실시간으로 측정 및 동기화하여 이를 기반으로 동영상 선호모델을 지속적으로 개선하고 필요시 영화추천을 수행하는 시스템을 제안한다. 제안한 시스템에선 이미지, 소리, 자막 스트림으로부터 실시간 추출되는 저수준 피쳐들과 동기화되어 측정된 얼굴표정, 자세 및 생체신호로부터 해당 동영상이 유발한 시청자의 감정상태를 추정하여 선호모델 학습에 사용한다. 제안한 컨텐트-시청자 연계 추천모델의 일례로서 컨텐트의 오디오 및 자막 정보를 이용하여 시청자의 피부전기활성도로 측정된 arousal반응을 예측할 수 있음을 보인다.

키워드 : 추천 시스템, 동영상 선호도 모델, 생체신호기반 피드백, 피부전기활성도, 각성도 예측

Abstract

A movie recommendation system is proposed to learn a preference model of a viewer by using multimodal features of a video content and their evoked implicit responses of the viewer in synchronized manner. In this system, facial expression, body posture, and physiological signals are measured to estimate the affective states of the viewer, in accordance with the stimuli consisting of low-level and affective features from video, audio, and text streams. Experimental results show that it is possible to predict arousal response, which is measured by electrodermal activity, of a viewer from auditory and text features in a video stimuli, for estimating interestingness on the video.

Key Words : Recommender System, Video Preference Model, Physiological Feedback, EDA, Arousal Estimation.

1. 서 론

무한복제가 가능한 디지털 컨텐츠의 시대가 도래하면서 소비자의 선택권은 유례없는 규모로 커졌고, 검색 및 추천 기술의 발전에 힘입어 가장 대중적인 일부 컨텐츠 위주로 소비되던 과거와 달리, 소수취향의 다수 컨텐츠로의 접근 및 소비가 기술적으로 가능해졌다. 선택의 폭이 소비자가 종래의 탐색적 접근으로는 감당할 수 있는 수준을 넘어서면서, 영화나 도서 분야를 필두로 한 각종 온라인시장에서 컨텐츠 추천기능은 서비스의 핵심기능으로 자리잡았고[1], 이러한 상업적 활용을 가능케한 추천기술의 핵심은 협업 필터링(collaborative filtering)이다[2]. 협업 필터링은 "군중의 지혜"를 기반으로 추천을 수행하는 기법으로서, 다수의 사용자가 추천 대상 아이템에 부여한 평점 데이터를 기반으로 사용자와 아이템의 연관 관계를 분석하는 기법의 총칭이다[1]. 협업 필터링은 유사한 사용자 또는 유사한 아이템의 평점 정보를 활용하거나, 대규모 데이터에 숨겨진 사용자와 아이템의 선호 관계 요인을 파악하여 이를 추천에 적용하는 방식을 동작하며, 적용 분야에 크게 구애받지 않고 사용할 수 있는 장점이 있다[3].

그러나, 신규 사용자 혹은 신작영화의 경우에서와 같이 참조할 학습데이터가 존재하지 않거나, 사용자의 명시적인 선호도 정보를 획득하기 어려운 환경에서는 협업 필터링의 적용이 불가능하다[3]. 특히, TV시청환경에서는 "뒤로 기대기(lean-back)"로 대표되는 사용경험 특성상 시청자의 명시적인 만족도 피드백을 기대하기 어렵고, 프라이버시에 민감하여 시청자 개인정보나 시청만족도 자체를 수집하고 공유하기 어렵다는 제약이 존재하므로 협업 필터링을 적용하는

접수일자: 2014년 3월 9일
심사(수정)일자: 2014년 4월 1일
게재확정일자 : 2014년 5월 8일
† Corresponding author
본 논문은 미래창조과학부의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(NRF-2010-0017734), 산업통상자원부의 재원으로 한국산업기술평가관리원의 지원(KEIT-10035348)을 일부 받았음.

데에 있어 명백한 한계가 존재한다.

TV 시청환경에서의 협업 필터링 적용이 불가한 상황의 대안으로서, 아이템과 사용자의 추가의 속성 정보를 활용하는 내용 기반(content-based) 추천 방식의 접근법에 주목할 수 있다. 내용 기반 추천 기법은 명시적인 신규 컨텐츠에 대해서도 내용 정보가 제공되면 추천에 바로 포함시킬 수 있는 장점이 있으며, 사용자의 속성을 추가로 고려하기 때문에 TV 시청과 같은 개별적 체험과 결합시키기 유리한 면이 있다. 그러나, 기존의 내용 기반 추천 방식은 텍스트로 표현되는 태그 및 메타 데이터에 주로 의존하기 때문에 컨텐츠 내의 다양한 요인과 변화를 반영하기 힘든 점과, 추천 결과가 다양하지 못한 한계가 있다[4].

본 논문에서는 TV시청환경에서 상술한 문제를 극복하기 위해 영상 컨텐츠의 저수준 특성값(feature)의 변화 및 시청자가 별도의 의지적 행동 없이 시청 중의 정서적 변화에 따라 무의식적으로 내비치는 암묵적 피드백(implicit feedback)을 활용한 선호도 모델의 학습 방법과 이를 이용한 컨텐츠 추천 시스템을 제안한다.

예를 들어 특정 시청자가 강한 흥미를 보이고 집중해서 시청한 영화의 구성요소를 분석하여 해당 시청자에게 특화된 선호모델(preference model)을 학습할 수 있고, 더 나아가 시청자의 집중여부를 지속적으로 모니터링하여 최적의 대안제시 시점을 결정할 수 있다.

암묵적 피드백은 시청 중에 발생하는 시청자의 동영상에 대한 비언어적이고 수동적인 반응으로서, 시선(eye gaze)과 자세(posture), 생리적 반응(physiological responses)을 포함한다. 통상 “별점”으로 대표되는 명시적 피드백과 달리 암묵적 피드백은 시청자의 의지적 표현을 거치지 않으므로 측정 자체에 불확실성이 따르지만, 상대적으로 설문방식에 따르는 응답자의 의식적 왜곡을 피할 수 있다는 장점이 있다. 기존 협업 필터링에서 평점정보를 수집하듯이 TV에서 매번 프로그램 별로 만족도 입력을 요구하는 것은 휴식 목적의 TV고유의 사용자경험을 훼손하는 것이므로 실질적으로 채택 불가능하며, 만족도 추정을 위한 암묵적 피드백의 선택은 TV시청환경 하에서는 불가피하다.

아이템 내의 저수준 피쳐는 이미지의 경우 color histogram [5], motion intensity [6], complexity [7][8] 등이 시청자 감성에 영향을 주는 요인으로 보고되고 있고[9-12], 오디오의 경우에는 MFCCs (Mel Cepstrum Coefficients) [13]나 pitch [12]등의 영향이 보고되고 있다[14-16].

또한, 최근의 디지털방송이나 DVD영화는 시각장애인을 위한 Closed Caption트랙(이하 CC) 혹은 외국인을 위한 자막트랙을 기본적으로 포함하고 있으므로, 이를 토대로 자연어 처리기법(natural language processing)을 적용하여 컨텐츠의 의미적 피쳐(semantic feature), 그리고 더 나아가 스토리 자체를 이해하려는 노력이 있어왔으며[17], 이러한 연구들을 수행하는 과정에서 단어별로 연관된 감정정보를 데이터베이스화한 사례도 보고되고 있다[18].

본 논문에서는 먼저 제2장에서 시청자의 암묵적 피드백을 측정하고 이로부터 관심도를 추정하는 모델을 소개한다. 제3장에서 전체시스템 구성 및 동작을 설명하고, 제4장에서 구성된 시스템의 일례로 오디오-텍스트 기반 생체신호측정 및 선호모델학습을 수행한 사례를 소개하며, 제5장에서 보여준 실험결과를 통해 도출된 결론을 제6장에서 논한다.

2. 암묵적 피드백 측정

암묵적 피드백은 앞서 정의한 바와 같이 동영상이 유발한 비언어적-수동적 반응이므로, 시선, 자세 및 생리적 반응이 대표적인 측정대상이다. 이러한 암묵적 피드백의 측정을 통해 추정하고자 하는 궁극적인 대상은 시청자의 감정(affect)으로서, 추천 목적에 한정할 경우 즐거움이나 몰입도와 같은 하위개념에 초점을 맞추게 된다.

감정(affect)을 체계적으로 기술하고자 하는 시도는 심리학에서 오랫동안 연구되어온 주제이고, Ekman이 6가지 기본 감정 - anger, disgust, fear, happiness, sadness, surprise - 이 문화나 인종에 상관없이 인간이라면 공통적으로 얼굴표정을 통해 인식가능하다는 걸 입증한 이후[19][20], 감정을 상기 6개 범주의 조합으로 묘사하고자 하는 categorical approach가 주로 연구되어 왔다.

한편, 동일한 문제를 다른 측면에서 본 연구자들은 상기 6가지 기본감정을 포함한 인간의 다양한 감정을 2개 혹은 3개의 기저감정의 조합으로 표현할 수 있다는 dimensional approach를 제안한 바 있고, 표현 및 적용의 용이함으로 인해 계산적모델링 분야에서 널리 적용되고 있다. 본 논문에서는 대표적인 dimensional approach로서 Arousal과 Valence를 인간 감정 표현의 주요 축으로 간주하는 모델을 채택하였다[21][22].

특히 긍정적 감정요인과 부정적 감정요인이 동시에 영향을 끼치는 시청만족도의 복합적인 특성을 감안하여 상기 dimensional approach의 양대 축 중 Arousal에 초점을 맞춰, Arousal을 해당 컨텐트가 유발한 몰입도 및 감정적 고양수준의 중요한 지표로 간주한다.

Arousal의 추정은 상기 언급된 암묵적 피드백(얼굴표정, 자세, 생체신호) 중 어느 것을 사용하더라도 일정수준 가능하나, 본 논문에서는 이를 시청자의 피부에서 접촉식 센서로 측정한 피부전기활성도(Electrodermal Activity, 이하 EDA)를 통하여 달성하고자 한다[23]. EDA는 주로 자율신경계(autonomic neuronal system)의 지배를 받는 생체신호 중 하나이므로 시청자의 의도적 왜곡을 피할 수 있고 반응시간이 충분히 빠르며(0.8-4초) arousal과 높은 상관관계를 보이는 것으로 알려져 있다[24].

그림 1은 본 연구에서 사용된 시스템을 통해 취득한 영화시청 중 EDA수준과 자막감성수준의 시계열 데이터의 일부이다. 이 샘플 내에서는 자막내 단어의 긍정적 속성(하늘색) 및 부정적 속성(보라색)의 전체적 추이와 실시간 취득

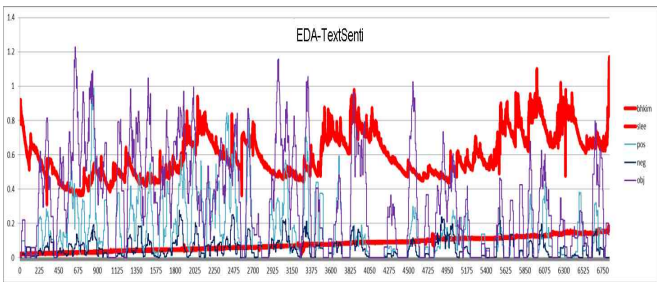


그림 1. 동일영상 시청 중인 피험자 두 명의 EDA반응 (굵은선) 및 영상 자막 기반 세 가지 감성 지수(가는선)

Fig. 1. An example of time series data consisting of EDAs of two subjects (bold lines) and emotional tuples of words in dialogues from a close caption track (narrow lines)

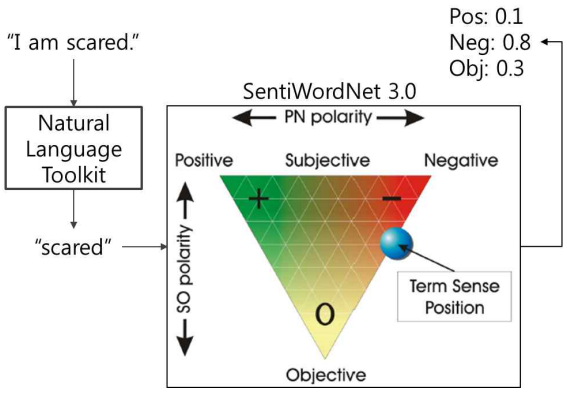


그림 2. SentiWordNet기반 자막내 단어 감정지수 추출
Fig. 2. Extraction of polarity scores of subscript words based on SentiWordNet

된 EDA반응(굵은 빨간선)의 추이 간에 약한 상관관계가 관찰되는데, 동일한 스케일 내에서 피험자 2명간 EDA범위에 큰 차이가 있음을 확인할 수 있다. 이러한 개인차는 생체신호에서 흔히 관찰되는 특성으로서, 개인 내에서의 일관성 이상은 기대하기 어렵다.

그림 2는 SentiWordNet [18]을 이용하여 영문자막내 단어에 내포된 감성적 요소를 3축의 실수공간으로 변환하는 과정을 도시한 것이다. 자막에서 사용된 문장에는 이중부정 등 단어의 의미를 왜곡시키는 다수의 문법적 장치가 반영되어 있으므로 단어를 추출하기 전에 자연어처리 라이브러리를 적용하여 이를 보정해주어야 한다.

만족도의 또다른 간접적인 지표로 시청자의 시선집중 수준을 사용할 수 있다. 구체적으로 특정 컨텐트의 시청중 임의 시구간 내에서의 시선(eye gaze)이 시청면을 향하는 시간의 점유율 및 깜박임 정도를 측정함으로써 해당 컨텐트에 대한 주의집중 수준을 정량적으로 측정할 수 있다.

3. 시스템 구성

제안하는 시스템은 TV를 이용한 동영상 시청 환경에서, 동영상 컨텐츠를 표현하는 다양한 피쳐 및 시청자의 암묵적 반응을 실시간에 측정 및 동기화하여 이를 기반으로 시청자의 동영상 선호모델을 지속적으로 개선하고 필요시 영화추천을 수행하는 것을 목표로 한다.

시청자 암묵적 피드백 Y는 만족/선호모델 상태(R)의 각 생체신호별 특성(g)에 따른 발현으로서 콘텐트 구성요소 X와의 관계는 다음과 같이 기술된다:

$$\begin{aligned} \vec{R} &= f(\vec{X};\theta) \\ \vec{Y} &= g(\vec{R}+e) = g(f(\vec{X};\theta)+e) \end{aligned} \quad (1)$$

여기서 g는 개인차가 상당한 발현수준 및 소요시간의 함수이나, 개인별 추정으로 한정하고 샘플링 시구간을 충분히 크게 잡으면 만족/선호모델 파라미터 Θ의 학습과정은 식 (2)와 같이 근사된다.

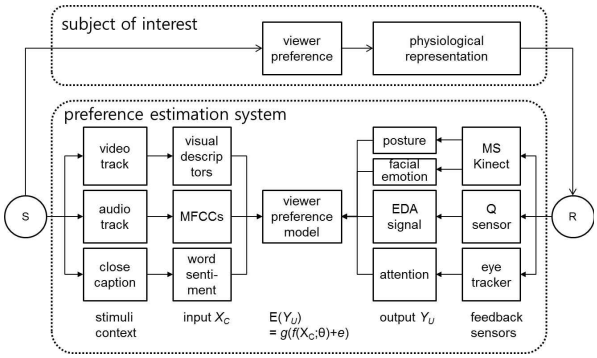


그림 3. CogTV 시스템 구성도
Fig. 3. System architecture of CogTV

$$\hat{\theta} = \arg\min \sum_i |r_i - \tilde{r}_i| \approx \arg\min \sum_i |y_i - \tilde{y}_i| \quad (2)$$
$$where\ r \in \vec{R}, y \in \vec{Y}$$

따라서, 모델 파라미터가 주어진 상태에서의 추천은 백그라운드에서 취득된 대안 컨텐츠들의 R 혹은 Y의 추정값을 토대로 상위 N개를 후보로 제안하게 된다.

상술한 컨텐츠 구성요소 X에 따른 만족/선호모델 온라인 학습을 지원하는 추천시스템의 구조는 그림 3과 같다.

그림 3에서 도시되었듯이, 동영상 컨텐트(S)에 노출된 시청자의 암묵적 피드백(R)은 시청자에 내재된 만족/ 선호(viewer preference)상태가 반영된 생체신호표현(physiological representation)에서 유래한 것으로서, 본 시스템에서는 이를 다양한 접촉식 및 비접촉식 센서를 이용하여 다각도로 측정하되, 본 논문에서는 이 중에서 Q Sensor를 통해 측정된 EDA로부터 추정된 Arousal에 한정한다.

따라서, 상기 시스템 내의 만족/선호모델(viewer preference model)의 학습을 통해 상기 식(2)와 같이 입력된 피쳐 X_c와 이에 의해 영향받은 것으로 여겨지는 Arousal (Y_u)간의 관계를 가장 잘 설명하는 모델 파라미터 Θ가 결정되면, 이 모델에 다른 컨텐트로부터 추출된 피쳐를 입력하여 해당 컨텐트가 야기할 시청자의 Arousal $E(Y_u)$를 추정할 수 있게 된다.

그림 4는 그림 3의 시스템 구성도를 기반으로 현재 구현된 실험환경을 보여준다. 해당 환경에서 암묵적 반응을 측

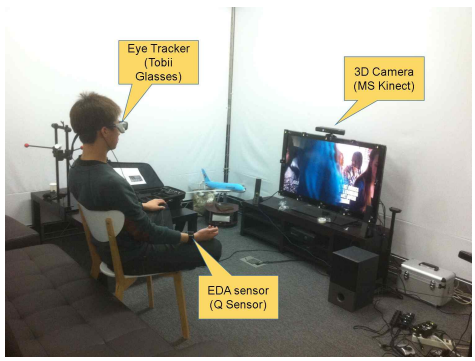


그림 4. 시청자의 암묵적 반응 측정을 위한 시스템 구축사례
Fig. 4. An example system for measuring implicit feedback of a watcher

정하기 위해 Microsoft사의 Kinect (3D 카메라)와 Tobii사의 아이트랙커(시선검출) 및 Q Sensor(접촉식 EDA센서)를 기본으로 구성하였고, 용도에 따라 확장 가능하다. 컨텐트 피쳐로는 OpenCV와 HTK (Hidden Markov model Toolkit) [25], SentiWordNet [18], Python Natural Language Toolkit [26]을 활용하여 영상/음성/자막에서 다양한 수준의 피쳐를 추출하여 시청자 반응과 동기화된 형태로 학습 데이터를 생성하도록 구성하였다.

4. 실험설계

본 논문에서는 오디오-텍스트 피쳐를 기반으로 시청자의 arousal 수준을 추정한 실험사례를 일례로서 소개한다.

20~30대의 여성 2명과 남성6명에게 각각 1편의 임의 선정된 2시간 분량의 영화를 보여주고, 시청 중 매 25ms마다 취득한 EDA 데이터와 동영상에서 추출한 피쳐를 동기화시킨 학습-평가 데이터셋을 구축하였다.

또한, Valence효과를 구분하기 위해 긍정적인 성격이 강한 영화 2편과 부정적인 성격이 강한 영화 3편으로 구성된 stimuli set을 생성하여 피험자에게 할당하였다.

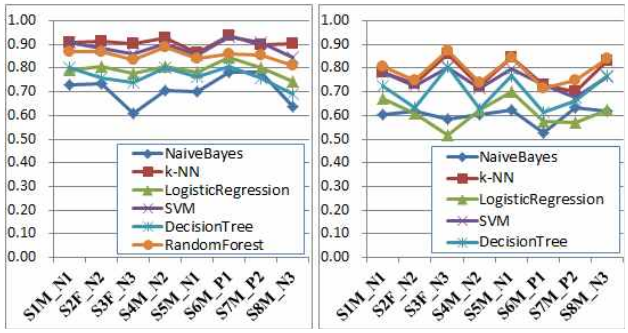
동영상의 피쳐 중 오디오 정보인 MFCCs (Mel-Frequency Cepstral Coefficients)는 25ms 프레임 별로 20단계 필터 계수 및 각 필터에 대한 차분값(delta)을 추출하였으며, 롱텀 피쳐로서 5초 길이, 50% 중복 속성의 이동 윈도우를 적용한 후 윈도우별로 필터별 피쳐군의 1~4차 모멘트(평균, 표준편차, 왜도, 첨도)를 추출하였다.

텍스트 정보는 영화 자막 혹은 TV방송신호의 close caption트랙에서 추출된 문자열을 Python Natural Language Toolkit [26]기반 negation등의 전처리를 수행한 후 이를 SentiWordNet [18] 조회를 통하여 각 관심 시구간 내 감성요소(sentiment)를 세 가지 준위(pos, neg, obj)로 변환하였고, Arousal에의 기여특성을 감안하여 (pos - neg) 및 sqrt(pos^2+neg^2) 피쳐 2종을 추가하였다.

피험자 반응은 착용형 EDA센서(Q Sensor)를 통해 8Hz로 샘플링된 시계열값을 상기 각 모달리티별 윈도우 내의 평균값으로 변환하여 동기화시켰다. 본 논문에서는 Arousal 예측 문제를 EDA값 기반 이진 분류문제로 설정하여 분석을 진행하였으므로, EDA 값의 이진화를 위해 측정된 EDA 값으로부터 각 피험자별 EDA 값의 1사분위수(Q1)보다 작은 값을 Low, 3사분위수(Q3)보다 큰 값을 High로 설정하였으며, 불명확한 신호의 배제를 위해 1사분위수와 3사분위수 사이의 값은 사용하지 않았다. 각 사용자별로 High/Low로 레이블된 관측 데이터의 수가 균형에 가까운(balanced) 데이터셋을 이와 같이 구성하였다.

표 1. 오디오/텍스트 피쳐 기반 EDA를 매개로 한 Arousal 추정을 위해 적용한 분류 알고리즘 및 모델 설정
Table 1. List of classification algorithms and their settings for audio/text-based arousal estimation (using EDA as the proxy signal for arousal)

Algorithm	Parameters
Naive Bayes	normal distribution assumption for numeric attributes
***k*-NN**	k = 1, 3, 5 normalized Euclidean distance
Logistic Regression	multinomial, ridge estimator
SVM	SMO, RBF Kernel
Decision Tree	C4.5, pruning on
Random Forest	various number of trees, unlimited tree depth



(a) (b)

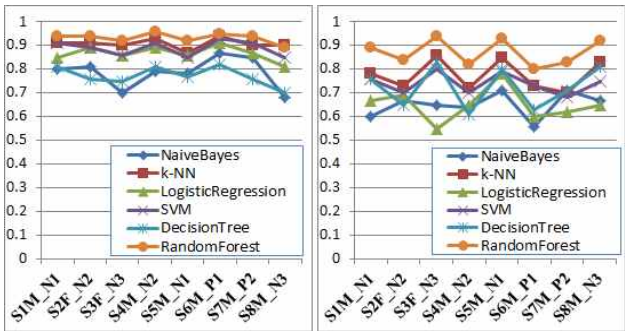
그림 5. 다양한 분류 모델을 이용한 시청자(S#[성별]) 및 영화 컨텐츠(N#)별 EDA 반응을 (a) 오디오 기반, (b) 텍스트 기반으로 예측한 결과의 정확도(accuracy)

Fig. 5. Accuracy plot of various classification models for EDA responses based on (a) audio and (b) text (x-axis: viewerID_contentsID)

오디오-텍스트 기반 EDA 예측을 위한 문제공간의 특성을 다각도로 접근하기 위해 다양한 종류의 분류 알고리즘(classifier)을 적용하였다. 본 실험에서는 나이브베이즈 분류기[27], 최근접 이웃 기반 분류기(k-NN)[28], 로지스틱 회귀모델[29], Support Vector Machine (SVM) [30], 결정트리(decision tree)[31] 및 Random Forest [32]를 적용하였고(표 1), 모달별 효과를 측정하기 위해 오디오 피쳐와 텍스트 피쳐 각각에 대해 학습을 별도로 수행하였다. 평가시 3-fold cross validation을 5회 적용하여 학습 결과 모델 간의 통계적 성능 비교를 수행하였다.

5. 실험결과

그림 5와 그림 6은 8명의 사용자가 세 가지 영화 컨텐츠



(a) (b)

그림 6. 다양한 분류 모델을 이용한 시청자(S#[성별]) 및 영화 컨텐츠(N#)별 EDA 반응을 (a) 오디오 기반, (b) 텍스트 기반으로 예측한 결과 모델의 성능(AUC)

Fig. 6. AUC (Area Under ROC) plot various classification models for EDA responses based on (a) audio and (b) text (x-axis: viewerID_contentsID)

표 2. Random Forest기반 학습결과 EDA 반응 예측성능
Table 2. EDA response classification results based on random forest classifiers

Subject# Gender_Movie#	Accuracy		AUC	
	Text	Audio	Text	Audio
S1M_N1	0.81	0.87	0.89	0.94
S2F_N2	0.75	0.87	0.84	0.94
S3F_N3	0.88	0.83	0.94	0.92
S4M_N2	0.74	0.89	0.82	0.96
S5M_N1	0.84	0.84	0.93	0.92
S6M_P1	0.72	0.86	0.80	0.95
S7M_P2	0.74	0.86	0.83	0.94
S8M_N3	0.84	0.81	0.92	0.89

를 시청하는 과정에서 기록한 EDA 반응을 각각 영화의 오디오 및 텍스트 피쳐를 기반으로 6가지의 분류 알고리즘을 적용하여 예측한 실험 결과이다. 주요 파라미터에 대한 선형 또는 로그 스케일 탐색을 통해 최적의 성능을 내는 모델을 선택하였다. 기본적인 성능 지표로서 정확도(accuracy)를 선택하였으며, High EDA를 기준으로 precision과 recall 간의 trade-off 관계를 표현하는 ROC (receiver operating characteristic) 곡선[33]의 대표 지표인 AUC (Area Under ROC Curve)를 추가로 측정하였다. 대부분의 설정에서 오디오 기반 모델의 정확도와 AUC값이 텍스트 기반 모델보다 높은 것을 확인할 수 있다. 또한 Random Forest (tree의 수=50) 및 k-NN (k=1) 모델의 성능이 거의 모든 경우에 가장 높게 측정되었다. 텍스트 기반의 EDA 반응 예측 성능의 경우 Random Forest 모델이 가장 높았으며, Random Forest 모델을 기준으로 오디오 및 텍스트 피쳐 기반의 예측 성능에 대해 보다 심화된 분석을 수행하였다.

표 2는 오디오 및 텍스트 피쳐 기반의 Random Forest 학습결과를 Accuracy(좌)와 AUC(우) 2개 척도로 측정한 결과를 요약하였다. 각기 다른 피험자들로부터 추출된 8벌의 데이터별 학습결과에 대한 3-fold cross validation결과, 평균 Accuracy는 텍스트 사용시 0.79, 오디오 사용시 0.85였고, 평균 AUC는 텍스트 사용시 0.87, 오디오 사용시 0.89였다.

Shapiro-Wilk 정규성 검정[34]결과 상기 표 2의 4가지 경우(Accuracy-Text, Accuracy-Audio, AUC-Text, AUC-Audio) 에 대하여 각각 p-value가 0.2448, 0.7956, 0.2643, 0.3858로서 모든 경우에 대하여 95% 신뢰수준에서 정규분포 가설을 기각할 수 없었다.

95%신뢰수준에서 오디오-텍스트간 등분산 검정결과, Accuracy의 경우 F = 5.4496, p-value = 0.0397로서 등분산 가설이 기각되었으며, AUC의 경우 역시 F = 6.3545, p-value = 0.0262로서 등분산 가설이 기각되었다.

상기 두 결과에 근거하여, stimuli가 다르고 개인차가 심한 EDA반응특성을 감안하여 paired t-test를 적용하여 텍스트와 오디오간의 EDA반응 예측성능간 차이가 존재하는지 여부를 단측검정으로 평가해보았다 (df = 7).

H1: 오디오가 텍스트보다 평균 예측성능이 높다.

로 대립가설을 설정한 경우, accuracy데이터에 대한 검정결과 t = 2.2393, p-value = 0.0301이었고 AUC데이터에 대한 검정결과 t = 2.3486, p-value = 0.0256이었다. 따라서,

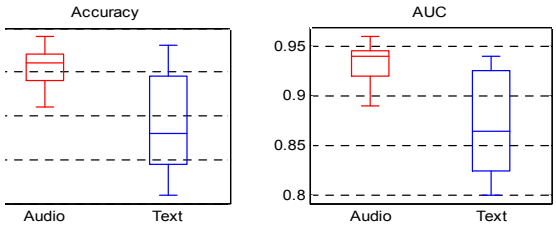


그림 7. 오디오-텍스트 기반 EDA 관측값 강도에 대한 이진 분류성능 비교결과. 오디오 기반 예측 성능이 텍스트에 비해 통계적으로 유의미하게 높다($p < 0.05$).
Fig. 7. Binary classification results of EDA signal level based on audio/text. Audio-based results are significantly better than text-based ones ($p<0.05$).

95% 신뢰구간에서 오디오가 텍스트보다 EDA반응 예측의 정확도(accuracy)와 AUC가 모두 유의미하게 높다고 말할 수 있다.

그림 7은 표 2의 데이터를 이용하여 8명의 피험자로부터 취득된 결과의 산포를 Box Plot으로 시각화한 것이다. 관찰된 Box Plot 패턴은 Arousal을 유발하는 감성요인이라는 관점에서 오디오 정보가 텍스트 정보에 대비하여 보다 예측성능이 우수하고, EDA반응의 개인차가 상대적으로 적은 경향성을 보인다. 오디오정보를 이용할 경우, 피험자에 따라 약 81~89%의 정확도(accuracy)로 개별 EDA를 예측할 수 있었다.

실험결과, 자막기반 감성분석과 저수준 오디오분석 모두 시청자의 EDA예측에 활용 가능하고, 상대적으로 저수준 오디오의 정보량이 더 많음을 확인할 수 있었다. 다만, 개인차가 큰 EDA 데이터의 특성상 예측값은 특정 개인에 대해서만 유효하며, 개인에 무관한 예측을 위해서는 적절한 정규화 과정이 필요하다.

6. 결론 및 향후 연구

본 논문에서는 종래의 협업 필터링을 적용할 수 없는 동영상 시청 상황에서의 추천 시스템을 구현하기 위하여 실시간으로 시청자의 감정 상태에 의해 영향받는 생체신호 및 동영상 내 멀티모달 피쳐를 추출하여 이들 간의 상관관계를 기계학습 기법으로 탐색하는 프레임웍을 제안하였다.

제안한 프레임웍의 타당성 검토의 일환으로, 주요 오디오 피쳐와 텍스트 피쳐에 대한 복수 피험자의 반응을 EDA로 측정하여 이들 간의 상관관계를 분석함으로써, 대사의 의미적 해석 뿐만 아니라 저수준의 오디오피쳐 만으로도 시청자의 Arousal수준을 상당한 정확도로 예측할 수 있다는 점을 보였다. 또한, 피쳐 선정관점에서 오디오와 텍스트간의 비교실험을 통하여 텍스트 대비 오디오 채널을 통해 전달되는 정보가 시청자의 Arousal에 미치는 영향이 보다 분명하고 개인차가 적다는 사실을 발견하였다.

상기 학습된 상관관계는 임의의 동영상 컨텐트에 대하여 별도의 메타데이터나 타인의 평점정보 없이 내용 그 자체만으로 해당 컨텐츠에 대한 시청자의 감정반응을 예측하는데 활용될 수 있다.

향후 연구에서는 비디오 피쳐를 추가하고, 각 모달별 피쳐의 추상화수준을 높이고, 보다 엄밀한 모델링을 위해 각

모달별 동적반응특성을 고려하여 확장하는 방향으로 연구를 진행할 예정이다.

References

[1] F. Ricci, L. Rokach, and B. Shapira, *Introduction to Recommender Systems Handbook*. Springer US, 2011.

[2] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, vol. 35. pp. 61 - 70, 1992.

[3] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30 - 37, 2009.

[4] P. Lops, M. Gemmis, M. De, & G. Semeraro, "Content-based Recommender Systems: State of the Art and Trends," in *Recommender Systems Handbook*, F. Ricci and L. Rokach, Eds. Springer US, pp. 73 - 105, 2011.

[5] S. Sural, G. Q. G. Qian, and S. Pramanik, "Segmentation and histogram generation using the HSV color space for image retrieval," *Proceedings of International Conference on Image Processing*, vol. 2, pp. 589-592, 2002.

[6] M. J. Black, "Combining intensity and motion for incremental segmentation and tracking over long image sequences," *Proceedings of European Conference on Computer Vision*, pp. 485 - 493, 1992.

[7] Y. K. Y. Ke, X. T. X. Tang, and F. J. F. Jing, "The design of high-level features for photo quality assessment," *Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 419-426, 2006.

[8] O. Le Meur, T. Baccino, and A. Roumy, "Prediction of the interobserver visual congruency (IOVC) and application to image ranking," *Proceedings of the 19th ACM International Conference on Multimedia*, pp. 373-382, 2011.

[9] P. Valdez and A. Mehrabian, "Effects of color on emotions," *Journal of Experimental Psychology General*, vol. 123, no. 4, pp. 394 - 409, Dec. 1994.

[10] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39 - 58, Jan. 2009.

[11] N. Liu, E. Dellandréa, B. Tellez, and L. Chen, "Associating textual features with visual ones to improve affective image classification," *Proceedings of the fourth International Conference on Affective Computing and Intelligent Interaction*, pp. 195 - 204, 2011.

[12] Y. Baveye, J.-N. Bettinelli, E. Dellandrea, L. Chen, and C. Chamaret, "A Large Video Data Base for Computational Models of Induced Emotion," *Proceedings of the sixth International Conference on Affective Computing and Intelligent Interaction*, pp. 13 - 18, 2013.

[13] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern Recognit. Artif. Intell.*, vol. 116, pp. 374 - 388, 1976.

[14] K. R. Scherer, "Vocal affect expression: a review and a model for future research.," *Psychol. Bull.*, vol. 99, pp. 143 - 165, 1986.

[15] D. Neiberg and K. Laskowski, "Emotion Recognition in Spontaneous Speech Using GMMs," *Proceedings of INTERSPEECH - ICSLP Ninth International Conference on Spoken Language Processing*, pp. 809 - 812, 2006.

[16] R. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18 - 37, 2010.

[17] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "DBpedia - A crystallization point for the Web of Data," *J. Web Semant.*, vol. 7, pp. 154 - 165, 2009.

[18] S. Baccianella, A. Esuli, and F. Sebastiani, "SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," *Proceedings of the Seventh International Conference on Language Resources and Evaluation,* pp. 2200 - 2204, 2008.

[19] P. Ekman and W. V Friesen, "Constants across cultures in the face and emotion," *J. Pers. Soc. Psychol.*, vol. 17, pp. 124 - 129, 1971.

[20] P. Ekman, "Basic Emotions", in Dalgleish, T; Power, M, *Handbook of Cognition and Emotion*, Sussex, UK: John Wiley & Sons, 1999.

[21] L. A. Feldman, "Valence-focus and arousal-focus: Individual differences in the structure of affective experience," *Journal of Personality and Social Psychology*, vol. 69, pp. 153-166, 1995.

[22] L. F. Barrett, "Discrete emotions or dimensions? the role of valence focus and arousal focus," *Cogn. Emot.*, vol. 12, pp. 579 - 599, 1998.

[23] D. McDuff, A. Karlson, A. Kapoor, and M. Czerwinski, "AffectAura: an intelligent system for emotional memory," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 849-858, 2012.

[24] M. E. Dawson and A. M. Schell, "The Electrodermal System," in J. T. Cacioppo and L. G. Tassinary (Eds.), *Principles of Psychophysiology: Physical,social, and inferential elements*, The Cambridge Press, Cambridge, 1990.

[25] The Hidden Markov Model Toolkit (HTK), Available: http://htk.eng.cam.ac.uk/, 2009, [Accessed: March 3, 2014]

[26] S. Bird, E. Loper and E. Klein, *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.
[27] G. H. John, P. Langley, "Estimating continuous distributions in Bayesian classifiers," *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 338-345, 1995.
[28] D. Aha, D. Kibler, "Instance-based learning algorithms," *Machine Learning*. vol. 6, pp. 37-66, 1991.
[29] S. le Cessie, J. C. van Houwelingen, "Ridge estimators in logistic regression," *Applied Statistics*, vol. 41, no. 1, pp. 191-201, 1992.
[30] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, K. R. K. Murthy, "Improvements to Platt's SMO algorithm for SVM classifier design," *Neural Computation*. vol. 13, no. 3, pp. 637-649, 2001.
[31] R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, 1993.
[32] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
[33] K. A. Spackman, "Signal detection theory: Valuable tools for evaluating inductive learning," *Proceedings of the Sixth International Workshop on Machine Learning*. pp. 160 - 163, 1989.
[34] S. S. Shapiro, M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika* vol. 52, no. 3-4, pp. 591 - 611, 1965.

저 자 소 개

박태서(Tae-Suh Park)
1999년 : 인하대학교 전기공학과 공학사
2001년 : 인하대학교 전기공학과 공학석사
2002년~2008년 : 삼성종합기술원 전문연구원
2008년~2013년 : 삼성전자 책임연구원
2011년~2013년: 서울대학교 인지과학협동과정 박사과정 수료
2014년~현재 : SK텔레콤 성장기술원
관심분야 : Computer Vision, HCI, Cognitive Process Modeling
Phone : +82-2-880-1847
E-mail : taesuh@ieee.org

김병희(Byoung-Hee Kim)
2003년 : 서울대학교 컴퓨터공학부 공학사
2006년 : 서울대학교 컴퓨터공학부 박사과정 수료
2006년 : 독일 베를린공대 방문연구원
2006년~현재 : 서울대학교 컴퓨터공학부 연구원
관심분야 : Machine Learning, Artificial Intelligence, Probabilistic Graphical Models
Phone : +82-2-880-1847
E-mail : bhkim@bi.snu.ac.kr

장병탁(Byoung-Tak Zhang)
1986년 : 서울대학교 컴퓨터공학과 공학사
1988년 : 서울대학교 컴퓨터공학과 공학석사
1992년 : 독일 Bonn 대학교 컴퓨터과학 박사
1992년~1995년 : 독일국립정보기술연구소(GMD, 현 Fraunhofer Institutes) 연구원
1997년~현재 : 서울대학교 컴퓨터공학부 교수 및 인지과학, 뇌과학, 생물정보학 협동과정 겸임교수
2003년~2004년 : MIT 인공지능연구소(CSAIL) 및 뇌인지과학과(BCS) 객원교수
2012년~현재 : 서울대학교 인지과학연구소 소장
관심분야 : Biointelligence, Cognitive Machine Learning, Molecular Evolutionary Computation-based Neurocognitive Information Modeling
Phone : +82-2-880-1833
E-mail : btzhang@bi.snu.ac.kr