멀티채널 정보학습 기반 드라마 동영상 정서 인식

이범진[O], 김병희, 장병탁
서울대학교 컴퓨터공학부
{bjlee, bhkim, btzhang}@bi.snu.ac.kr

Sentiment Recognition of TV Drama Videos by Learning Multichannel Information

Beom-Jin Lee[O], Byoung-Hee Kim, Byoung-Tak Zhang
School of Computer Science and Engineering
Seoul National University

요 약

동영상에서의 정서 인식 기술은 양산된 동영상 데이터의 빠른 검색을 용이하게 하고, 증가하고 있는 유해 영상의 필터링 역할을 할 수 있다. 동영상이 유발하는 정서를 인식하고 정량화함으로써, 사용자의 감성에 기반한 검색, 추천 등의 서비스가 가능하게 된다. 본 논문에서는 동영상에서의 소리, 영상, 자막 등의 멀티채널 정보 기반의 정서 인식 기법을 제안한다. 동영상 시청자에 의해 특정 정서가 발생된 구간을 수집하고, 각 채널 별로 특성값을 추출하여 감독학습 기반의 정서 인식 모델 생성 및 무감독 학습 기반의 정서 프로파일을 생성한다. 실험 예로서 4가지 드라마에서 추출한 웃음, 울음, 화남의 정서 태깅 데이터를 구축하고, 소리와 영상 정보를 추출하여, 정서별 인식 모델 및 프로파일 생성 결과를 보인다. 개별 정서 인식 모델의 인식율은 8~90%대이며, 정서별 프로파일과의 상호 보완적인 조합을 통해 실제 동영상에서의 다양한 정서 인식이 가능하다.

1. 서론

비디오 데이터는 1970년 가정용 비디오 테이프의 보급 이래, 수십 년 동안 그 보존의 형태를 변화시키며 현재는 인터넷의 발달로 전 세계 어디서든 볼 수 있게 발전해 왔다. 방대한 양의 비디오 데이터에서 사용자가 원하는 컨텐츠를 얻기 위해서는 다양한 평가 기준이 필요하다. 영상 평가 요소로는 영상의 장르, 주요 등장인물, 영상 제작비용 등 많은 요소가 있겠지만 영상이 주는 감성은 해당 영상에 대해 시청자가 주되게 경험한다는 점에서 가장 영향력 있는 요소라고 할 수 있다[1].

본 논문에서는 동영상 데이터 중 주요 비중을 차지하는 드라마에 대해 주요 정서가 나타나는 구간을 소리와 영상의 특이점 변화 값들을 인지적 기계학습 기법을 이용하여 구분하고, 해당 감정을 가장 잘 표현하는 데이터를 확보하는 방법에 대해 서술한다. 그리고 더 나아가 정서를 기반으로 하는 비디오 데이터의 탐색 기법 개발 방법을 제안한다.

2. 관련 연구

시청각 자료에 대한 정서정보 추출 연구는 약 20년 전부터 꾸준히 이루어지고 있다[2]. Wang 등은 SVM (Support Vector Machine)으로 영화의 정서를 구분하였는데, 오디오 신호에 대해서 중요 특징들을 타입별로 분류를 하였고, 시각정보는 HSL (hue, saturation, lightness)을 이용하였다[3]. 또한 Xu 등은 감정과 이벤트의 강한 연결이 있는 소리에 대해 AEE (Audio Emotional Events)라는 방식을 소개하였다[4].

감정 중심의 연구로는 사용자의 감정이 나타나는 지점을 확인한 후, GMM (Gaussian Mixture Model)을 이용하여 추측하는 연구[5]와 HMM (Hidden Markov Model)을 이용, 사용자가 느끼는 3가지 정서(intended, expected. experienced emotion)에 대해 영화의 정서를 추적하는 시스템이 발표되었다[6]. 또한 Banda 등은 SAVEE (Surrey Audio Visual Expressed Emotion) 데이터[7]를 이용하여 오디오 데이터로 사용자의 정서를 구분하는 작업과 잡음이 많은 데이터에서 시각자료를 합침으로써 성능이 올라감을 보고하였다[8]. Zhang과 Wang은 오디오 데이터나 시각정보와 같은 하위 단계 특징값들을 이용하지 않은 인지적 단계에서의 정서 구분 및 검색방법의 구조에 대해서 소개하였다[9].

본 논문에서는 드라마 동영상에서의 주요 정서(관객 웃음, 웃음, 울음, 화남) 인식에 초점을 두며, 특히 음성과 영상 두 채널에 대한 감독학습과 무감독 학습 기반의 정서 인식 모델의 조합을 통해 다중 정서 인식도 가능하게 한다.

3. 정서 인식 모델

3.1 개요

제안하는 동영상 정서 인식 프레임워크는 그림1과 같이 요약되며 이러한 개요를 진행하기 위해서는 먼저 사용할 특성과 정서의 선정이 필요하다.

비디오는 동적 이미지, 텍스트, 음악, 소리 등 풍부한 특성들이 제공되는 미디어 매체이다. 이러한 특성들 가운데 정서인식에 사용될 특성은 동적 이미지(시각정보)와 소리(소리정보)로 선택하였다. 그 이유는 먼저 시각정보의 경우 화남, 놀람 등의 정서에서는 화자의 움직임의 변화가 상당히 큰 것을 볼 수 있었기 때문에 각 정서별

그림 1. 제안하는 드라마 동영상 정보 인식 기법 개요

로 움직임의 변화가 다르게 나타날 것이라 예상되었기 때문이다.. 또한 소리정보의 경우 기쁨, 슬픔 등 정서가 소리로 확연히 표현되어 특징적인 값이 도출될 수 있을 것이라 가정하였다.

두 번째 사용할 정서 선택 측면에서는 valence/arousal model[10]에 기반을 두어 선정하였다. 다양한 정서들이 존재하지만 valence/arousal model의 축이 되는 정서들을 정확히 인식할 수 있다면 다른 정서들은 파라미터들의 조정으로 인식이 가능할 것이라는 예측 하에 웃음, 울음, 화남, 무정서를 인식 목표 정서로 선택하였다.

이로써 첫 번째 무정서 대비 목적 정서에 대한 탐지 모델과 코믹드라마에서 빈번히 등장하는 관객들의 웃음과 화자들의 웃음을 구분하는 모델을 제안한다.

3.2 멀티채널 정보 추출

위 개요에서 설명하였듯이 정서에 따른 화자들의 특징적인 움직임이 있을 것이라는 가정을 기반으로 해당 이미지에서의 특이점들의 움직임 변화값을 수집한다.

초당 24프레임의 정지 이미지를 추출하고, 연속된 프레임 간의 특징점의 움직임 패턴[1]의 총합을 에너지로 정의하여, 깁스 분포값 변화에 대한 평균, 표준편차, 왜도, 첨도의 4개의 값을 추출한다.

소리 신호는 정서인식에 사용될 가장 중요한 특징값(feature)이다. 각 정서별로 표현되는 소리가 다양하고, 특징적이기 때문이다. 정서가 태깅된 동영상에서 128비트, 모노 타입, 44.1kHz 샘플링 레이트의 신호 처리 후, 25ms의 해밍 윈도우를 설정하여 13차원의 MFCC (Mel-frequency cepstral coefficients) 값과 1차원의 에너지 값, 총 14차원의 벡터를 추출한다. 정서가 태깅된 동영상에서 1초 단위의 음성에 대해 추출한 14차원*42구간의 값을, 각 차원별로 평균값을 구하여 최종적으로 14차원의 특성값을 추출한다.

3.3 정서 인식 및 요약 프로파일

각 정서별로 상기와 같은 영상 및 소리 특성값을 추출하여, 1초를 단위로 하는 학습 데이터를 구성하고 이를 감독학습 기반의 정서 인식 모델 및 무감독 학습 기반의 정서별 프로파일을 구축한다. 감독학습 기반의 정서 인식 모델은, 지정 정서가 나타나지 않은 '무정서 구간' 대비 지정 정서의 발생 지점을 인식하도록 구축한다. 정서별 프로파일은 각 채널 별로 특성값 벡터 집합의 평균 및 공분산(동형 가정, $\Sigma = \sigma I$, $\sigma = \max \sigma_i$)을 고려한 고차원 구체의 내부로 정의한다. 새로운 동영상에 대해 1초 길이의 슬라이딩 윈도우를 적용하여, 정서별 인식기를 적용한 인식 및 특성값 벡터의 프로파일 대비 위치 확인을 통한 인식이 모두 가능하다.

4. 비디오 정서 인식 실험 결과

4.1 데이터

제안하는 기법의 적용 예로서, 4가지 미국 드라마- 빅뱅 이론(Big Bang Theory), 프렌즈 (Friends), E.R., 그레이즈 아나토미 (Grey's Anatomy)-를 기반으로 정서 인식 실험을 수행하였다. 각 드라마 별로 1개의 에피소드[2]에 대해 12명의 20대 시청자가 지정한 관객 웃음, 웃음, 울음, 화남의 정서 발생 구간에 대해, 1초 길이의 슬라이딩 윈도우를 적용하여 특성값을 추출한 결과 표 1과 같이 학습 데이터가 구축되었다. 무정서는 각 정서 발생 구간의 좌/우 3초 지점 이외의 곳에서 무작위로 추출하였다.

표 1 학습 데이터 구성: 정서별 인스턴스 수

정서	인스턴스 수
관객 웃음	400
웃음	149
울음	127
화남	422
무정서	199

4.2 실험 구성

실험은 2가지 정서에 대한 분류 성능 실험과 관객웃음과 화자웃음의 프로파일 비교 실험으로 구성하였다[3].

표 2 정서 대비 목표정서 인식(분류) 결과 표/차트

비교 정서	정서 인식률 (3-fold cross validation 평균)	
	Accuracy	F-measure
관객 웃음	85.48%	0.902
웃음	81.61%	0.766
화남	80.03%	0.864
울음	93.23%	0.915

4.3 실험 결과 및 분석

4.3.1 정서 인식 결과

첫 번째 실험 결과는 표 2와 같이 상당히 뛰어난 성능을 보여주었다. 총 분류 결과는 모두 80%를 상회했음을 볼 수 있다. 이로써 실험에 사용된 데이터들은 각 정서

1) OpenCV의 CalcOpticalFlowPyrLK() 함수 적용

2) BBT 시즌2 1화, FR 시즌8 1화, ER 시즌10 1화, GA 시즌6 1화

3) 실험에서는 초 단위로 함축된 시각정보 값과 초당 14*42개의 행렬로 구성된 소리정보를 사용하여 각 정서를 SVM (SMO, 2차 Poly-Kernel)에 학습 및 분류하였다.

를 잘 표현하고 있음이 증명되었고, 정서 인식 모델의 사용이 충분한 성능결과를 도출할 것을 예상할 수 있다.

두 번째 관객과 화자 웃음 비교 실험은 93.44%의 성능을 보여줌으로서 무작위적인 수집된 데이터들과 차이가 확실히 있으며, 이는 향후 각 정서 프로파일이 더욱 명확히 작성된다면 비디오 정서인식 시스템에 상당한 기여 가능성을 보여준다.

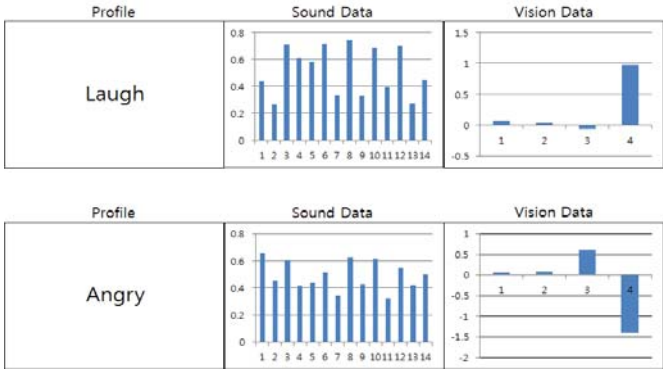

그림 2 제안하는 정서별 프로파일의 예.

4.3.2 정서별 프로파일

정서 프로파일은 그림 2와 같은 형태로 작성될 수 있다. 이렇게 정의된 프로파일은 비디오 데이터에서 추출되는 각 채널들의 공분산의 Euclidean Distance를 비교함으로써 어느 정서에 해당되는지 판별한다. 이 방법을 적용한 예(그림 3)의 결과는 표 3과 같다. 그림3의 (a)는 화자가 웃고있는 장면이다. 그에따른 표 3의 분류결과는 (a)의 인식모델, 프로파일 모두 해당 정서를 감지(+)한 상태라 나타난다. 그리고 (b)와 (c)의 경우는 어느 하나의 정서 감지 모델에서 정서를 인식하지 못하여도 나머지 모델에서 상호보완을 하여 해당 정서 상태를 인식함을 보여준다. 마지막 (d)는 추후 다중 정서인식에 해당 모델들의 기여도에 따라 다중 정서 상태임을 알 수 있는 결과를 나타내고 있다.

표 3 정서 대비 목표정서 인식(분류) 결과 표/차트

	정서	인식모델	프로파일	
			소리정보	시각정보
(a)	웃음	+	+	+
(b)	화남	-	+	+
(c)	웃음	+	-	+
(d)	웃음/화남	+/+	-/+	+/-

5. 결론 및 향후 연구

본 논문에서는 비디오 드라마에 등장하는 화자들이 나타내는 4가지 정서를 분류하는 방법에 대해 논의하였다. 기존 연구들과 달리 여러 드라마를 사용하여 등장하는 화자들이 수시로 변하였고, 장면과 소리 또한 많은 변화가 나타나는 데이터를 사용하였다. 그리고 시청자들의 정서가 아닌 화자들의 감정에 대한 프로파일 데이터를 사용하여 오직 해당 영상의 특성들에만 의존하여 정서인식을 수행하는 난이도 있는 문제에 대해 접근을 시도하였다. 이러한 환경 잡음을 고려하지 않고 실험을 진행했음에도 멀티채널정보 기반의 목표 정서 인식률은 80%이상을 보여주었다. 향후 더욱 다양한 정서 프로파일링과 추가적인 시각 특성추출방법을 사용하여 진화하는 실시간 정서 프로파일 변화방법 및 정서인식 특화 모델을 제안하고 다중 정서 인식 시스템의 구축이 수행될 것이다.

그림 3 정서 인식 결과 예.

감사의 글

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(No. 2012-0005643, Videome), 정부(지식경제부)의 재원으로 한국산업기술평가관리원의 지원(10035348, mLife) 및 교육과학기술부의 BK21-IT 프로그램에서 일부 지원되었음.

참고문헌

[1] 김광수, 영화선택 및 평가에 관한 연구, 광고연구, 제 48호, pp. 139-164, 1999.

[2] A. Hanjalic, Extracting Moods from Pictures and Sounds: Towards truly personalized TV, Signal processing magazine, IEEE, 23(2):90-100, 1995.

[3] HL. Wang, L. Cheong, Affective Understanding in Film, Circuits Systems for Video Technology, IEEE, 16(6):689-704, 2006

[4] M. Xu, LT. Chia, J. Jin, Affective Content Analysis in Comedy and Horror Videos by Audio Emotional Event Detection, IEEE international conference on Multimedia and Expo, 2005.

[5] 박현재, 강행봉, 비디오 셧의 감정 관련 특징에 대한 통계적 모델링, 한국통신학회논문지, 28(12C):1200-1208, 2003.

[6] N. Malandrakis, A. Potamianos, G. Evangelopoulos, A. Zlatintsi, A Supervised Approach to Movie Emotion Tracking. In Proc. ICASSP. pp. 2376-2379, 2011.

[7] S. Haq, P. Jackson, J. Edge, Audio-visual feature selection and reduction for emotion classication, In Proc. AVSP, pp. 185-190, 2008.

[8] B. Ntombikayise, R. Peter, Handling Noise Analysis in Audio-Visual Emotion Recognition, MMCogEmS: Inferring Cognitive and Emotional States from Multimodal Measures, ICMI 2011 Workshop, 2011.

[9] L. Zhang, J. Wang, Design for Emotion Classification and Retrieval of Video Based on User Experience, ICECC, pp. 802-805, 2011.

[10] J. A. Russell, A circumspect model of affect, Journal of Psychology and Social Psychology, 39(6):1161, 1980.