웨어러블 센서를 이용한 라이프로그 데이터 자동 감정 태깅

박경화[1], 김병희[2], 김은솔[2], 조휘열[2], 장병탁[1,2]
[1]서울대학교 뇌과학협동과정, [2]서울대학교 컴퓨터공학부
{kwpark, bhkim, eskim, hyjo, btzhang}@bi.snu.ac.kr

Automated Emotional Tagging on Lifelog Data with Wearable Sensors

Kyung-Wha Park[1], Byoung-Hee Kim[2], Eun-Sol Kim[2], Hwi-Yeol Jo[2], Byoung-Tak Zhang[1,2]
[1]Brain Science Program, Seoul National University
[2]Department of Computer Science and Engineering, Seoul National University

요 약

본 논문에서는 실생활을 영상으로 촬영한 웨어러블 센서 데이터에서 연속적으로 변하는 사용자의 감정 태그 정보를 자동으로 출력하는 시스템을 제안한다. 주제나 매체에 구애받지 않고 사용가능한 감정 태그는 텍스트 정보에서 추출하거나 전체 비디오에 대해 하나의 태그를 추출하는 것이 일반적이나, 이 시스템은 감성 컴퓨팅을 기반으로 하여, 웨어러블 센서 데이터에서 감정 태그를 연속적으로 추출한다. 감정 발화 벤치마크 데이터셋을 학습한 감독 학습 기반 감정 예측 모델을 이용해 웨어러블 장치에 녹음된 소리로부터 긍정적, 부정적 반응을 인식하고, 웨어러블 장치인 피부전도도 센서(EDA) 데이터를 모델링하여 흥분 정도를 인식해 감정 태그를 결정한다. 이 시스템에 적용된 감정 발화 인식 성능을 최근 뛰어난 성능을 보인 컨벌루션 신경망을 이용한 결과와 비교 분석한다. 이 시스템으로 웨어러블 장치로 수집한 영상의 감정 태그를 발굴할 때, 음성만으로 발굴하는 것보다 EDA를 포함하여 발굴하는 것이 더 좋은 결과를 보였다.

1. 서론

소셜 네트워크 서비스에 자기 이야기를 올리는 개인 라이프로깅(lifelogging)은 요즘 일상처럼 이뤄지는 일 중 하나이다. 또한, 근래 널리 보급된 시계형 웨어러블 장치 덕분에 라이프로깅은 더욱 풍족해졌다. 그러나 풍족해진 만큼 관리가 어려워졌다. 재검색(retrieval)할 때는 단순하면서 여러 매체에서 사용 가능한 공통점이 있는 태그를 사용하는 것이 중요하다. 이런 점에서 감정 태그는 이런 점을 충족시키면서 웨어러블 센서 데이터에도 사용할 수 있는 장점이 있다.

기존에는 표정이나 대화 목소리를 이용해서 사람의 감정을 예측했다. 대화에서 감정을 인식하는 감정 발화 인식(Emotional Speech Recognition, 이하 ESR) 문제는 최근 대두되는 딥러닝으로 문제를 해결하고자 하고 있고, 감정 교류를 앞세우는 가정용 로봇에 탑재되면서 실용화 단계에 있다. 그러나 시장에 출시된 감정 교류 로봇이 ESR을 잘 하는지는 명확하지 않다. 사람이 감정을 겉으로 숨기거나[1] 한정된 장소에서만 작동하는 좁은 문제공간, 카메라에 포착되지 않는 물리적 한계 등이 발생했을 때에도 대처를 해야한다. 다른 모달리티를 이용해 감정 예측을 해야 하는 경우에 최근 보급된 사용자의 숨김없는 무의식적인 반응을 연속적으로 측정하는 피부전도도(Electrodermal Activity, EDA) 이용을 고려할 수 있다.

웨어러블 장치에서 쓰이는 센서 중에 EDA 센서는 감정 변화와 직접적인 연관이 있는 것으로 알려져 있다[1]. EDA는 감정 중에서도 특히 흥분도(arousal)에 연관되어 있으며, 화가 났을 때나 긴장했을 때 그 변화가 두드러진다[2].

본 논문에서는 1) 웨어러블 장치 영상 데이터의 녹음 데이터로 긍정부정도(valence)를 인식하고 피험자의 EDA 데이터로 흥분도(arousal)를 인식해 2차원적 감정 모델에 기반을 두는 연속적 감정 태깅 시스템[그림 1]을 제안한다. 2) 이 시스템 안의 감정 발화 인식 모델은 비교적 간단한 기계학습 모델을 사용하며, 그 성능을 최신의 것과 비교한다. 3) 웨어러블 센서 영상에서 소리만으로 감정 태깅을 한 결과와 EDA 정보를 함께 모델링하여 예측한 결과를 분석해 본 시스템의 정당성을 논한다.

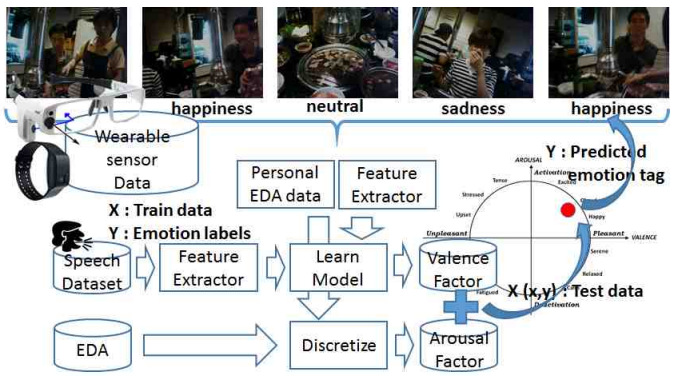

그림 1. 감정 자동 태깅 프레임워크

2. 관련 연구

2.1 라이프로그 태깅

태그를 자동으로 발굴하는 연구는 많은 연구가 있었으나, 웨어러블 센서의 보급은 비교적 근래 이뤄져 라이프 로깅 데이터를 재검색 하는 연구는 진행 중에 있다. 웨어러블 센서 데이터 중 소리 정보를 태그로 이용해 재검색을 하려고 한 연구가 있다[3]. 본 논문에서는 소리 정보와 웨어러블 센서 데이터를 함께 사용해 감정 태그를 발굴한다.

2.2 감성 컴퓨팅

감성 컴퓨팅(affective computing)은 인간의 감성을 분석하여 계산 모델을 설계하는 분야이다. 사람의 감정은 기본적으로 분노, 슬픔, 행복, 공포, 혐오, 놀람까지 6개로 알려져 있다[4]. 여기에 기준이 되는 중립 상태를 포함하는 경우도 있다. 감정을 명확히 구분 짓는 카테고리식 접근과 흥분도와 긍정부정도를 두 축으로 가지는 차원적 접근법이 있다[5]. 본 논문에서는 차원적 접근을 통해 감정을 모델링했다.

2.3 감성 대화 인식

감성 대화 인식(Emotional Speech Recognition)은 감정을 담아 발화한 것을 녹음하여, 발화에 담긴 감정을 분석하거나 분류한다. 이를 위해 공개된 데이터셋들이 많으며 성능 벤치마킹을 하게 된다. 이런 감정 발화 데이터셋을 분류 문제에 쓸 때 중요한 점은 어떤 특징을 발굴했느냐가 된다. 감성 컴퓨팅에 기반을 둔 특징을 다양하게 많이 실험에 이용하는 것이 보통이다. 가장 최근에는 딥러닝을 이용해 특징을 발굴하고, 이 데이터로 SVM을 학습시켜 분류 문제를 풀고자 하는 시도가 있었다[6].

3. 감정 태깅 시스템

본 시스템 [그림 1]은 웨어러블 장치로 촬영된 영상에 감정 태그를 발굴하는 시스템으로, 피험자가 직접 말하는 1인칭적 발화 뿐 아니라 타인의 말소리나 환경 소음과 같은 3인칭적 외부음도 고려한 입력 X로부터 감정 태그 Y를 추출해야한다. 이러한 문제 특성을 [표 1]에서 나타낸다. 입력으로 피험자의 발화 뿐 아니라 타인의 발화까지 포함되는 데이터 특성상, 발화를 다루는 ESR 벤치마크 데이터셋을 학습한 모델로 감정을 인식하며, 개인화에 초점을 맞추는 웨어러블 센서 EDA를 모델링하여 rule-based로 개인의 감정을 포괄하는 동시에 환경의 분위기까지 포함하는 감정 태그를 결정한다.

표 1. 본 실험에서 고려하는 입력 X와 목적 레이블 Y의 개인적, 환경적 구성 요소. 고려하지 않는 요소는 ×로 나타내며, 환경적 시각 요소인 표정은 본 논문에서 고려하지 않는다.

	X			Y
	Visual	Audio	EDA	감정 태그
개인	×	발화	○	개인
환경	표정(×)	외부음	×	분위기

이 시스템은 발화 데이터셋을 학습시킨 모델에 웨어러블 센서 영상 데이터의 소리 데이터를 입력하여 얻은 결과를 긍정부정도(valence) 척도로 사용한다. 그리고 [그림 1]에는 피험자의 EDA 데이터를 오랫동안 수집하여 개인화를 구현하는 것을 나타내고 있지만, 본 논문에서는 피험자의 EDA 데이터를 high, low, neutral로 단순 이산화 하여 흥분도(arousal) 척도로 사용하였다.

이 실험에서 소리 데이터는 긍정부정도(valence)를 보여주는 척도로 사용할 것이므로 원래의 7개 레이블이 아닌 행복, 분노, 슬픔, 중립 4개의 레이블을 가진 인스턴스만을 이용하여 모델을 학습시키고 감정 태그 분류 결과 또한 얻는다. 4개의 예측된 감정 상태는 [그림 2]의 우측에 있는 흥분도-긍정부정도 2차원 평면의 x축에 해당하는 긍정부정도를 나타내는 것으로 가정했다. 즉, 행복은 x축에서 (+)방향, 슬픔과 분노는 (−)방향, 중립은 (0)원점 근처로 가정한다. 앞서 high, low, neutral로 이산화한 EDA 값을 함께 고려하여 흥분도-긍정부정도 평면상에 예측해야 될 4가지 감정을 맵핑해보면 [그림 2]와 같다. 본 논문에서 제안하는 이 규칙 기반 모델은, 음성으로 예측한 감정이 (+)인데 EDA로 관측된 감정이 high라면 영상에서 예상되는 감정은 분노가 되는 방식이며, EDA가 중립적인 상황이라면 중립상태로 가정한다. 이후 소단원에서 이런 모델을 포함하여 설계한 전체 시스템의 출력 결과와 사람이 기록한 감정 기록과의 정확도를 비교한다.

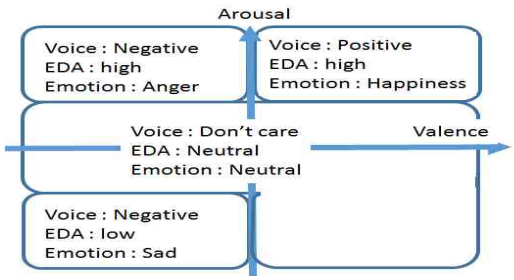

그림 2. 음성과 EDA 데이터 분석 결과의 조합을 흥분도-긍정부정도 평면에 맵핑하여 얻는 감정 태깅 모델의 개념도

4. 실험

4.1 학습 데이터 : 벤치마크 데이터셋

학습에 사용한 벤치마크 데이터셋은, 독일의 Burkhardt가 공개한 베를린 감정 데이터셋[7]을 이용한다. 이 데이터셋은 남자 5명, 여자 5명으로 총 10명의 피험자에게 대사를 1줄 읽게 했다. 그 대사는 7가지 감정: 분노, 지루함, 혐오, 공포, 행복, 슬픔, 그리고 보통 상태로 연기를 하도록 지도받았다. 10명이서 10종류의 대사를 7가지 감정 상태로 읽어서 약 500 여개의 음성 파일과 거기에 담긴 감정의 레이블로 구성되어 있다.

4.2 검증 데이터 : 웨어러블 센서 데이터셋

그리고 검증에 사용한 웨어러블 센서 데이터셋은, 아이트래커(eyetracker)를 실생활에서 착용하고 음식점에서 주문과 식사를 하면서 대화를 나누는 데이터셋[8]이다. 아이트래커에는 마이크도 장착되어 있어 음성도 녹음되며, 추가적으로 EDA 센서도 함께 착용하여 수집하였다. 2개의 데이터를 실험에 사용했으며, 각 영상은 평균 33분이다. 영상에서 음성만을 추출하여, 10초를 음성 클립 하나로 설정하고 5초씩 이동하면서 클립을 추출했다. 영상 하나 당 약 396개의 클립으로 나눠진다. EDA 값은 1초당 4회 측정하므로 음성 데이터와 싱크를 맞추기 위해 10초 단위로 묶어서 통계적 계산을 한 뒤 5초씩 넘어가며 반복했다. 통계적 계산에는 10초 단위로 묶은 EDA 값이 상향하는지 하향하는지 유지되는지를 계산하며. 계산된 각 영상의 값을 모두 모아 상향하는 경향을 보이는 상위 25%를 high, 하향하는 경향을 보이는 하위 25%를 low, 나머지 50%를 neutral로 이산화한다.

감성 컴퓨팅에서는 감정이 짧은 시간 내에 변하는 것으로 알려져 있으며 공개 데이터셋인 LIRIS-ACCEDE의 사례를 참고하여[9], 본 논문에서는 10초 단위로 데이터를 처리하고 5초씩 넘어가며, 실험에 참가하지 않았던 2명의 피험자가 영상을 보면서 행복, 슬픔, 분노, 중립 상태를 기록했다.

4.3 데이터 공통 처리 방법

학습 데이터와 검증 데이터에서 음성 파일 하나와 음성 클립 하나 당 [표 2]에 나타낸 6가지 특징을 특징별 채널을 다르게 추출하여 총 95개의 특징을 가지는 인스턴스 하나

로 만들었다. 데이터셋 간의 분포차이를 확인하여, 정규화 과정을 거쳤다. 각 데이터셋에서 추출한 인스턴스를 모아 각각의 처리된 학습 데이터와 검증 데이터로 만든다.

표 2. 음성 정보에서 감정 예측을 위해 추출하는 95차원의 feature 구성. 9차원 feature의 경우 지정 길이의 음성에 대해 평균과 분산 및 1, 2차 변화량, 중앙값, 최댓값, 최솟값의 9가지 통계량을 사용. Essentia[7]를 이용하여 추출.

Feature의 특성	Feature Name	구성(차원)
Spectral descriptors	MFCC	13 멜 계수
	spectral contrast coeffs	6채널*9
Time-domain descriptors	zerocrossing rate	9
	average loudness	9
Tonal descriptors	HPCP entropy	9
Rhythm descriptors	beats loudness	1

4.4 딥러닝 ESR 성능 비교 실험

앞서 베를린 감정 데이터셋에서 추출한 학습 데이터를 여러 기계학습 모델로 학습시켜 10등분 교차 확인한 결과를 얻는다. 딥러닝 ESR 성능 비교 검증 결과 SMO가 4개와 7개 레이블의 ESR 분류 모두 85.84%, 82.3%의 우수한 성능을 보였다.

표 3 베를린 감정 데이터셋을 이용한 다른 논문의 결과

논문	레이블 수	Details	정확도
본 논문	4	95 features, SMO	85.84%
	7		82.30%
Pan [10]	3	{MFCC+MEDC+Energy+Pitch }, SVM	95.04%
Bitouk [11]	7	{UL, CL}+spectral, LibSVM	78.2%
Huang[12]	7	spectrogram unsupervised FE, Semi-CNN (SVM)	88.3% 85.2%

[그림 3]에서 볼 수 있듯이, 음성만을 이용해 영상에 감정 태그를 분류하는 것 보다 EDA 정보를 추가한 모델링을 통해 감정 태그를 분류하는 것이 성능이 더 뛰어난 것을 볼 수 있다.

5. 결론 및 논의

본 논문에서는, 개인 요소와 환경 요소를 모두 고려하여 감정 태그를 생성하는 시스템을 제안했고, 그 시스템으로 실제 웨어러블 센서 데이터의 감정 태그를 수행한 결과 사람이 남긴 태그와 비교했을 때 좀 더 견고한 결과를 얻었다. 또한, 본 시스템의 앞부분에 포함된 ESR 모델이 같은 벤치마크 데이터셋을 학습시킨 딥러닝 ESR 성능 결과와 비교한 결과[표 3] 상대적으로 간단한 모델임에도 불구하고 고무적인 결과를 얻었다. 감정 태깅 결과에서도 음성만 사용한 것 보다 웨어러블 센서인 EDA를 함께 모델링하여 사용했을 때 정확도가 높아진 다는 것을 확인할 수 있었다. 그러나 독일어를 읽은 학습데이터와 한국어를 이야기한 검증데이터의 분포가 차이가 나는 covariant shift 문제는 향후 해결해야 될 문제이다. 이를 해결하기 위해 한국형 감정 발화 데이터셋이나 transfer learning이 필요할 것으로 보인다. 또한, 하루 생활 중에 감정이 격변하는 경우는 거의 없기에 레이블의 편향에 대처하는 것 또한 앞으로 추가적인 연구가 필요하다.

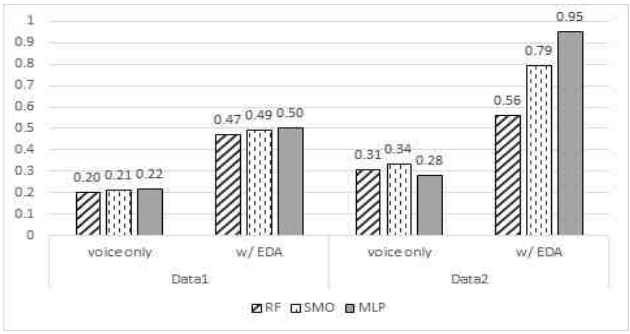

그림 3. 웨어러블 센서 검증 데이터를 음성만으로 감정 태깅을 했을 때와 EDA와 함께 감정 태깅을 했을 때의 성능 비교

감사의 글

이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원(R0126-15-1072-SW스타랩, 10035348-mLife, 10044009-HRI.MESSI)과 한국연구재단의 지원(NRF-2010-0017734-Videome)을 받아 수행된 연구임.

참고문헌

[1] P. Ekman and W. V. Friesen, “Detecting deception from the body or face.,” J. Pers. Soc. Psychol., vol. 29, no. 3, pp. 288-298, 1974.

[2] I. B. Mauss and M. D. Robinson, “Measures of emotion: A review,” Cogn. Emot., vol. 23, no. 2, pp. 209–237, 2009.

[3] M. Shah, B. Mears, C. Chakrabarti, and A. Spanias, “Lifelogging: Archival and retrieval of continuously recorded audio using wearable devices,” IEEE International Conference on Emerging Signal Processing Applications, pp. 99–102, 2012.

[4] J. Tao and T. Tan, “Affective Computing : A Review,” ACII 2005, 2005, pp. 981–995.

[5] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, “Affective Characterization of Movie Scenes Based on Content Analysis and Physiological Changes,” Int. J. Semant. Comput., vol. 03, no. 02, pp. 235–254, 2009.

[6] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, “ESSENTIA: an Audio Analysis Library for Music Information Retrieval,” ISMIR 2013, pp. 493–498, 2013.

[7] F. Burkhardt, a Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A Database of German Emotional Speech,” Eur. Conf. Speech Commun. Technol., vol. 2005, pp. 3–6, 2005.

[8] E.-S. Kim, et al., “Behavioral pattern modeling of human-human interaction for teaching restaurant service robots,” AAAI 2015 Fall Symposium on AI for Human-Robot Interaction, 2015.

[9] C. Chamaret, L. Chen, S. Member, Y. Baveye, and E. Dellandr, “LIRIS-ACCEDE : A Video Database for Affective Content Analysis,” IEEE Trans. Affect. Comput., vol. 6, no. 1, pp. 43–55, 2015.

[10] Pan, et al., "Speech emotion recognition using support vector machine," International Journal of Smart Home, 6(2):101-107, 2012.

[11] D. Bitouk, R. Verma, and A. Nenkova, “Class-level spectral features for emotion recognition,” Speech Commun., vol. 52, no. 7–8, pp. 613–625, 2010.

[12] Huang, et al., "Speech emotion recognition using CNN," ACM International Conference on Multimedia, pp. 801-804, 2014.