비디오 스트림 구조를 활용한 동적 키프레임 기반 사용자 개성 예측

이미라[1,3], 우사이먼성일[2], 정혜동[3]
[1]성균관대학교 전자전기컴퓨터공학과 석사과정, 한국전자기술연구원
[2]성균관대학교 소프트웨어학과 교수
[3]한국전자기술연구원
olmizr@g.skku.edu, swoo@g.skku.edu, hudson@keti.re.kr

Predicting User Personality Based on Dynamic Keyframes Using Video Stream Structure

Mira Lee[1,3], Simon S.Woo[2], Hyedong Jung[3]
[1]Department of Electrical and Computer Engineering, Sungkyunkwan University
[2]Department of Computer Science & Engineering, Sungkyunkwan University
[3]Korea Electronics Technology Institute

요 약

기술이 발전함에 따라 복합적인 모달리티 정보를 포함하는 멀티미디어 데이터의 수집이 용이해지면서, 사람의 성격 특성을 이해하고 이를 개인화된 에이전트에 적용하고자 하는 연구가 활발히 진행되고 있다. 본 논문에서는 비디오 스트림 구조를 활용하여 사용자 특성을 예측하기 위한 동적 키프레임 추출 방법을 제안한다. 비디오 데이터를 효과적으로 활용하기 위해서는 무작위로 선택한 프레임에서 특징을 추출하던 기존의 방법을 개선하여 영상 내 시간에 따른 정보와 변화량을 기반으로 중요한 프레임을 선택하는 방법이 필요하다. 본 논문에서는 제 3자가 평가한 Big-five 지표 값이 레이블링된 대표적인 데이터셋인 First Impressions V2 데이터셋을 사용하여 외면에서 발현되는 특징들을 기반으로 영상에서 등장하는 인물들의 성격 특성을 예측했다. 결론에서는 선택된 키프레임에서 멀티 모달리티 정보를 조합하여 성격 특성을 예측한 결과와 베이스라인 모델과의 성능을 비교한다.

1. 서론

사람의 성격 특성은 인간의 행동을 측정 가능한 지표로 설명하기 위한 심리학적 요소이다[1]. 이러한 성격 특성 정보는 채용 심사, 추천 시스템, 포렌식 등에서 유용한 정보를 제공하는 판단 도구로 활용되고 있다[2], [3]. 학계에서는 소셜 미디어, 영상 플랫폼, 온라인 게임 등에서 사용자에 대한 정보를 추출하여 성격 특성을 예측하고 평가하는 성격 컴퓨팅(Personality Computing) 연구가 지속적으로 이루어지고 있다[4]. 성격 컴퓨팅 분야는 다양한 형태의 데이터를 활용하여 사용자의 특성을 인식하는 것을 넘어, 지능형 에이전트를 구현하기 위한 성격 특성의 합성에 대한 연구도 포함한다. 사람의 성격을 측정하고 생성하는 이러한 연구는 사람과 상호작용하며 사교적인 관계를 형성하고 때로는 주어진 역할을 수행하는 사회적 인지 로봇(Socially-aware Robot) 구현을 위한 핵심 기술이 되기도 한다. 사회적 인지 로봇을 인간에게 신뢰를 얻을 수 있을 만큼 지능적으로 만드는 것은 감정, 공감과 같은 추가적인 사회적 기술이 요구되기 때문이다[5]. 로봇과 인간이 사회 작용을 하기 위해서는 다양한 성격 유형과 그 복잡성을 고려하여 인간과 로봇의 성격 연관성을 파악하는 것이 필요하다. 따라서 사용자의 성격 특성을 예측하는 것은 인간과 에이전트가 상호 이해를 바탕으로 한 소통을 하기 위해 반드시 필요한 기술이다.

일반적으로 심리학 분야에서는 성격 특성을 측정하기 위해 리커트 척도에 따라 응답하도록 구성된 NEO-FFI[6], MBTI[7] 등 널리 알려진 설문지에서 질문의 의도를 반영한 일련의 계산 방법을 통해 결과를 산출한다. 반면 컴퓨터 과학 분야의 성격 컴퓨팅 연구에서는 행동, 표정, 발화 내용 및 속도 등 시각과 음성 정보를 통해 외면적인 특징을 분석하는 방법이 일반적이다. 본 논문에서는 영상으로부터 추출할 수 있는 준언어적, 비언어적 정보를 조합하여 성격 특성을 예측하는 방법을 제시한다.

2. 관련 연구

2.1 성격 특성 분류 체계

많은 심리학자들은 사람의 성격 특성을 분류하고 정량화하기 위한 수많은 모델을 제안해왔다. 그 중

Big-five 성격 지표[8]는 사람의 성격을 다섯 가지로 분류하고 세분화된 하위 범주에 대해 정량화할 수 있어 컴퓨터 연산에 용이하다는 장점을 가지고 있다. 이는 시간과 상대에 따라 변하는 사람의 성격 특성을 표현하기에 적합하다는 근거가 되기도 한다[9]. Big-five 성격 지표에서는 personality trait를 개방성(Openness to experience), 성실성(Conscientiousness), 외향성(Extraversion), 우호성(Agreeableness), 신경성(Neuroticism)으로 분류하고 각 성격 특성에 해당하는 하위 항목들을 정의한다. 다섯 개의 성격 지표는 각각 0에서 1까지의 값을 가지며 분류 및 회귀를 통해 예측된다.

2.2 동적 키프레임 추출

키프레임(Keyframe)은 영상으로부터 객체의 동적인 정보를 요약하기 위해 추출하는 것으로, 주로 행동 또는 제스처 인식에 활용되어 왔다[10], [11]. 많은 정보를 포함하고 있는 영상에서 중요한 정보를 추출하기 위해서는 장면과 프레임에 대한 이해를 바탕으로 시간에 따른 객체의 상호적인 관계를 고려해야 한다[12]. 이에 대한 예시로 영상 전체가 아닌 현재 시점의 프레임 정보와 연속된 프레임 간의 관계를 학습하여 비디오 캡셔닝에 적용한 연구가 있었다[13]. 일반적으로 성격 특성 예측을 위한 연구에서는 입력 영상을 균등하게 분할하여 구간 내에서 무작위로 프레임을 선택하는 방법을 채택한다[14], [15], [16]. 하지만 이 방법은 영상 내 객체의 움직임이나 장면에 대한 변화량을 전혀 반영하지 못한다는 한계를 가진다. 따라서 본 논문에서는 비디오 스트림 구조를 활용하여 영상에서 중요한 정보를 포함하는 키프레임을 추출하고, 더 나아가 키프레임 간의 변화량이 큰 상위 프레임만 활용하는 방법을 제안한다.

2.3 영상 스트림의 구조와 영상 압축 알고리즘

영상은 프레임의 연속적인 시퀀스로 구성되며, 비디오 압축은 네트워크와 스토리지의 제한된 환경에서 효율적인 서비스를 제공하는데 중요한 역할을 한다[17]. 그 중 MPEG 비디오 압축 프로세스의 일부인 GOP(Group of pictures)로 비디오 압축 및 스트리밍 효율 향상을 위해 영상 내 장면 변화와 독립성에 따라 비디오 시퀀스를 조직화할 수 있다. 비디오는 조직화된 구조에 따라 데이터를 압축(Compression) 및 해제(Decompression)하는 코덱을 거쳐 다양한 멀티미디어의 응용 분야에 사용된다. GOP는 I-프레임(Intra-frame), P-프레임(Predictive-frame), B-프레임(Bi-directionally predictive-frame)으로 구분된다[18]. 이 중 I-프레임은 영상의 시작점이나 장면에 전환이 발생하는 등 중요한 시점에서 독립적인 정보를 가졌다고 판단되는 프레임으로 선택되며, 이는 입력 영상으로부터 사용자의 성격 특성 및 감정을 예측할 때 유용한 정보로 활용될 수 있다. 이러한 GOP의 특성을 활용하면 영상의 모든 프레임을 입력으로 활용하지 않고도 효율적으로 동적인 키프레임 추출을 할 수 있다.

3. 데이터셋

사람의 성격 특성은 내면적 요소와 외면적 요소로 분류할 수 있으며 Big-five 지표에 대해 본인이 직접 평가한 내면의 정보로 레이블링된 데이터셋에는 UDIVA[19] 데이터셋이 대표적이다. 반면에 타인에게 표출되는 정보를 바탕으로 제 3자가 성격 특성을 평가한 대표적인 데이터셋에는 First Impressions V2[20] 데이터셋이 있다. 본 논문에서는 영상으로부터 시각과 청각 등 멀티모달 정보를 인식하여 외면으로 발현되는 사용자의 성격 특성을 평가하는 것이 목표이므로 제 3자에 의해 평가된 성격 특성이 레이블링 되어있는 First Impressions V2 데이터셋을 활용하였다.

4. 실험

4.1 실험 방법

본 논문에서는 영상에서 추출한 멀티모달 정보를 활용한다. 먼저 영상에서 스트림 구조를 활용한 키프레임을 선택하여 영상 내 등장인물뿐만 아니라 배경에 대한 정보를 모두 포함하는 전역 프레임(Global frame)을 추출했다. 이 프레임에서는 히스토그램 평활화(Histogram equalization)를 적용하여 이미지의 밝기 분포를 균일하게 만들어 비교적 어두운 영역과 밝은 영역의 특징이 더 잘 드러나도록 하였다. 따라서 전 범위의 픽셀값이 고른 분포를 갖게 되어 인물과 배경에 존재하는 세부사항이 강조된다. 더 나아가 데이터셋에서 제공하는 영상의 길이가 모두 다르므로 영상마다 동일한 개수의 프레임을 추출하기 위해 모션 벡터(Motion vector)를 기준으로 키프레임 간 변화가 큰 상위 128개 프레임을 선택하여 입력으로 활용했다. 이러한 전역 프레임의 전처리 과정을 통해 영상에 등장하는 인물의 행동과 배경에 대한 정보를 극대화한다. 두 번째로는 등장인물의 얼굴을 크롭한 지역 프레임(Local frame)을 활용한

다. 사용자 성격 특성 및 감정 예측 등의 분야에서 사람의 얼굴에서 드러나는 특징은 가장 주요한 정보를 제공한다고 알려져 있으므로[21], [22] 얼굴만을 크롭한 이미지도 입력으로 활용하였다. 이렇게 각 영상에서 크롭된 이미지들은 동일한 크기로 전처리 된다. 마지막으로 준언어적인 정보를 제공하는 음성 특징을 입력으로 활용한다. 준언어적인 특징은 성격 특성을 예측하는 데 중요한 역할을 하는 음성의 높낮이, 속도, 운율 등의 정보를 포함한다. 길이가 서로 다른 영상들에서 준언어적인 특징들을 모델의 입력으로 활용하기 위해 가장 긴 영상의 길이를 기준으로 나머지 길이는 각각의 음성을 반복하여 패딩하는 전처리를 수행했다.

이렇게 각 모달리티에 대한 전처리를 마친 후 3D 컨볼루션 신경망을 통해 다섯 개의 Big-five 성격 특성 지표를 0에서 1까지의 값으로 예측할 수 있도록 학습하였다. 특히 추출된 전역 프레임들은 시간에 따른 시각적 특징의 변화도 포함하고 있으므로 시간에 대한 정보도 학습할 수 있게 하였다. 마지막으로 각각의 브랜치에서 학습된 특징들을 조합하여 Big-five 성격 특성 지표를 최종적으로 예측했다.

4.2 결과

모델의 성능은 다섯 개의 클래스에 대한 회귀 성능을 평가하기 위해 손실함수 L1 loss를 사용하였다. 이 때 본 논문에서 제안하는 방법을 '영상을 동일한 길이의 여섯 구간으로 나누어 각 구간에서 무작위로 프레임을 추출하는 방법'과 '전역 프레임과 음성 정보 각각의 모달리티만 활용했을 때'의 성능과 비교하였다. 표 1, 2에서는 1에서 평균절대오차를 뺀 평균 정확도를 계산하여 성능을 나타내었다.

<표 1> 영상으로부터 프레임을 무작위로 추출했을 때(1)와 영상 스트림 구조 및 변화량을 반영하여 추출했을 때(2)의 성능

Method	Average	Extra.	Agree.	Consc.	Neuro.	Open.
(1)	91.81	92.31	92.07	92.13	**91.44**	91.10
(2)	**91.87**	**93.49**	**92.30**	**92.17**	90.02	**91.36**

<표 2> 모달리티별 각각의 성능과 멀티모달(전역+지역+음성) 정보를 활용했을 때의 성능

Method	Average	Extra.	Agree.	Consc.	Neuro.	Open.
전역 프레임	91.65	92.09	91.89	91.43	**91.25**	**91.55**
음성	66.18	67.81	65.90	65.22	67.84	64.10
멀티모달	**91.87**	**93.49**	**92.30**	**92.17**	90.02	91.36

5. 결론

본 연구에서는 비디오 스트림의 구조와 프레임의 변화량 정보를 활용하여 성격 특성을 예측하기 위한 키프레임 추출 방법을 제안한다. 이렇게 추출한 키프레임은 등장인물뿐만 아니라 배경 및 행동 변화량에 대한 정보를 제공하며, 얼굴을 크롭한 지역 프레임 및 음성 정보와 함께 특징 추출을 위한 정보로 활용된다. 실험 결과를 통해 모든 실험에서 외향성(Extraversion)이 상대적으로 높은 성능을 나타내었으며, 영상으로부터 프레임을 무작위로 추출했을 때보다 영상 스트림의 구조 및 객체의 변화량을 반영하여 프레임을 선택했을 때 더 높은 성능을 달성함을 확인하였다. 또한 단일 모달리티만을 사용하는 것보다 배경, 음성 등 다양한 정보를 조합하여 성격 특성 값을 도출했을 때 5개 중 3개 항목에서의 예측 성능이 더 우수하다는 사실을 확인하였다.

* 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2022-0-00043,개성 형성이 가능한 에이전트 플랫폼 기술 개발)

참고문헌

[1] Vinciarelli, Alessandro, and Gelareh Mohammadi. "A survey of personality computing." IEEE Transactions on Affective Computing, 5.3, 273-291, 2014.

[2] Suen, Hung-Yue, Kuo-En Hung, and Chien-Liang Lin. "Intelligent video interview agent used to predict communication skill and perceived personality traits." Human-centric Computing and Information Sciences, 10, 1-12, 2020.

[3] Suman, Chanchal, et al. "A multi-modal personality prediction system." Knowledge-Based Systems, 236, 107715, 2022.

[4] Phan, Le Vy, and John F. Rauthmann. "Personality computing: New frontiers in personality assessment." Social and personality psychology compass, 15.7, e12624, 2021.
[5] Mileounis, Alexandros, Raymond H. Cuijpers, and Emilia I. Barakova. "Creating robots with personality: The effect of personality on social intelligence." Artificial Computation in Biology and Medicine: International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2015, Elche, Spain, 2015, Part I, 6.
[6] Costa, Paul T., and Robert R. McCrae. Neo personality inventory-revised (NEO PI-R). Odessa, FL: Psychological Assessment Resources, 1992.
[7] Myers, Isabel Briggs, and Mary H. McCaulley, Myers-Briggs type indicator: MBTI, Palo Alto, Consulting Psychologists Press, 1988.
[8] Norman, W. T. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. The journal of abnormal and social psychology, 66(6), 574, 1963.
[9] Yang, Liang, et al. "Computational personality: a survey." Soft Computing, 26.18, 9587-9605, 2022.
[10] Beyan, Cigdem, et al. "Personality traits classification using deep visual activity-based nonverbal features of key-dynamic images." IEEE Transactions on Affective Computing 12.4, 1084-1099, 2019.
[11] Tang, Hao, et al. "Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion." Neurocomputing, 331, 424-433, 2019.
[12] Sadiq, Bashir Olaniyi, et al. "Keyframe extraction techniques: A review." ELEKTRIKA-Journal of Electrical Engineering 19.3, 54-60, 2020.
[13] Chen, Yangyu, et al. "Less is more: Picking informative frames for video captioning." Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 2018, 358-373.
[14] Yang, Karen, S. Mall, and N. Glaser. "Prediction of personality first impressions with deep bimodal LSTM." Technical report, arXiv, 1-10, 2017.
[15] Li, Yunan, et al. "Cr-net: A deep classification-regression network for multimodal apparent personality analysis." International Journal of Computer Vision, 128, 2763-2780, 2020.
[16] Giritlioğlu, Dersu, et al. "Multimodal analysis of personality traits on videos of self-presentation and induced behavior." Journal on Multimodal User Interfaces, 15.4, 337-358, 2021.
[17] Ma, Siwei, et al. "Image and video compression with neural networks: A review." IEEE Transactions on Circuits and Systems for Video Technology, 30.6, 1683-1698, 2019.
[18] Liu, Guozhu, and Junming Zhao. "Key frame extraction from MPEG video stream." 2010 Third International Symposium on Information Processing, Qingdao, Shandong China, 2010, 423-427.
[19] Palmero, Cristina, et al. "Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2021, 1-12.
[20] Ponce-López, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, I., ... & Escalera, S. Chalearn lap 2016: First round challenge on first impressions-dataset and results. Computer Vision - ECCV 2016 Workshops: Amsterdam, The Netherlands, 2016, 400-418.
[21] Mehta, Yash, et al. "Recent trends in deep learning based personality detection.", Artificial Intelligence Review, 53, 2313-2339, 2020.
[22] Chaudhari, Aayushi, et al. "ViTFER: facial emotion recognition with vision transformers." Applied System Innovation, 5.4, 80, 2022.