얼굴표정 인식 기법의 최신 연구 동향

이민규, *송병철
인하대학교
*bcsong@inha.ac.kr

Recent Research Trends of Facial Expression Recognition

Min Kyu Lee, *Byung Cheol Song
Inha University

요 약

최근 딥러닝의 급격한 발전과 함께 얼굴표정 인식(facial expression recognition) 기술이 상당한 진보를 이루었다. 얼굴표정 인식은 컴퓨터 비전 분야에서 지속적으로 관심을 받고 있으며, 인포테인먼트 시스템(Infotainment system), 인간-로봇 상호작용(human-robot interaction) 등 다양한 분야에서 활용되고 있다. 그럼에도 불구하고 얼굴표정 인식 분야는 학습 데이터의 부족, 얼굴 각도의 변화 또는 occlusion 등과 같은 많은 문제들이 존재한다. 본 논문은 얼굴표정 인식 분야에서의 위와 같은 고유한 문제들을 다룬 기술들을 포함하여 고전적인 기법부터 최신 기법에 대한 연구 동향을 제시한다.

1. 서론

얼굴표정 인식은 컴퓨터 비전 분야에서 지속적으로 관심을 받고 있으며, 인포테인먼트 시스템(Infotainment system), 인간-로봇 상호작용(human-robot interaction) 등 다양한 분야에서 활용되고 있다. 최근 기계학습의 일종인 딥러닝(Deep learning)의 급격한 발전으로 얼굴표정 인식 기술 수준이 상당한 진보를 이루어 왔으며, 다양한 산업 분야에 적용되고 있다. 그럼에도 불구하고 얼굴표정 인식 분야는 학습 데이터의 부족, 조도와 얼굴 각도의 변화 등과 같은 많은 문제들이 존재한다. 초기 얼굴표정 인식 기술들은 단일 이미지 속 인물(들)의 얼굴표정을 분류하는 것이 주류였지만, 최근에는 동영상 속 인물(들)의 얼굴표정을 분석하는 연구도 활발하다.

본 논문은 얼굴표정 인식 분야에서의 위와 같은 고유한 문제들을 다룬 기술들을 포함하여 고전적인 기법부터 최신 기법에 대한 연구 동향을 제시한다. 본 논문의 구성은 다음과 같다. 2 절에서는 고전적인 얼굴표정 인식 기술들을 설명한 후 3 절에서 딥러닝 기반의 얼굴표정 기술을 확인한다. 마지막으로 4 절에서는 본 논문에 대한 결론을 맺는다.

2. 고전적인 얼굴표정 인식 기술

고전적인 얼굴표정 인식은 그림 1(a)와 같이 단일 이미지 또는 비디오 시퀀스를 입력 받아 얼굴표정의 고유한 특징들을 추출하고 그들의 변화를 포착한 후 Support vector machine(SVM), Random forest 등과 같은 분류기로 얼굴표정을 인식한다. 오랫동안 사용된 특징 추출 기술들은 hand-crafted feature이다. Hand-crafted feature는 얼굴의 질감, 기하학적 특성, scale 등과 같은 얼굴 고유의 특징을 효과적으로 추출할 수 있다. 대표적

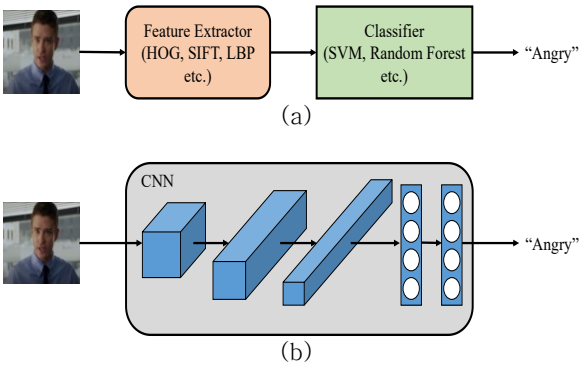

그림 1. (a) 고전적인 얼굴표정 인식 기술과 (b) CNN 기반의 얼굴표정 인식

인 hand-crafted feature로는, Local Binary Pattern (LBP)[5], Scale Invariant Feature Transform (SIFT), Histograms of Oriented Gradients (HOG), histograms of Local Phase Quantization (LPQ) 등이 있다. Hand-crafted feature를 시공간의 3차원으로 확장한 기법들도 등장하며 비디오 데이터에 대응하여 얼굴표정 인식을 수행할 수 있다. 초기 얼굴표정 인식 기법들은 주로 상기 hand-crafted feature에 기반하고 있다.

3. 딥러닝 기반의 얼굴표정 인식 기술

최근 딥러닝 중심으로 기계학습이 발달함에 따라 hand-crated feature보다 그림 1(b)와 같은 얼굴의 고유한 특징을 더욱 효과적으로 추출할 수 있다고 알려진 Convolutional Neural Network(CNN)을 이용한 얼굴표정 인식 기술들이 연구되고 있다. 그러나 단일 이미지는

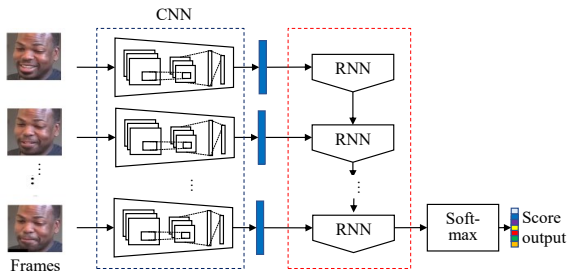

그림 2. CNN-RNN 구조

전후 상황에 대한 정보를 알 수 없기 때문에 극심한 얼굴 각도의 변화, occlusion에 취약할 수 밖에 없다. 이와 같은 한계에 대응하기 위해 더욱 최근에는 비디오 기반의 얼굴표정 인식 기술들이 활발히 연구되고 있다. 비디오는 이미지가 가지는 공간적인 정보뿐만 아니라 시간적 정보를 제공하기 때문에 전후 맥락을 파악할 수 있다. 이는 특정 비디오 프레임에서 occlusion 또는 얼굴 각도의 변화가 발생하더라도 이 전후 맥락 정보를 통해 얼굴표정 인식 성능을 보완할 수 있다. 하지만 비디오의 활용을 통해 주요 프레임의 선정 또는 효과적인 전후 맥락 정보의 추출이라는 크게 두 가지 추가적인 문제가 수반되며, 이에 대한 연구가 활발히 진행되고 있다.

비디오 기반의 대표적인 딥러닝 네트워크는 CNN-RNN 구조와 3D CNN 구조가 있다. 먼저, CNN-RNN 구조는 그림 2와 같이 CNN을 feature extractor로 사용하여 매 비디오 프레임의 feature를 순차적으로 추출한다. 추출된 feature들은 시간 순서에 맞게 순차적으로 Recurrent Neural Network(RNN)에 들어가며, 마지막 timestemp에서 최종적인 얼굴표정 결과를 출력한다. 반면, 3D CNN 구조는 기존의 convolution layer를 3차원으로 확장한 것이다. 즉, convolution 커널의 차원을 3차원으로 확대하여 시간축까지 필터링을 수행한다. 같은 맥락으로 pooling layer 또한 3차원으로 확대하였으며, 마지막은 기존 CNN과 동일하게 Fully-connected layer(FC)을 통해 최종적인 분류 결과를 출력한다. 위 두 가지 네트워크를 기반으로 다양한 네트워크가 연구되고 있다. Vielzeuf 등은 기존 CNN-RNN 구조에서 CNN을 3D CNN으로 대체하였으며, time-window를 통해 다수의 프레임들의 묶음이 효과적으로 고려될 수 있도록 설계하였다[1]. Lee 등은 전후 맥락 정보를 더욱 효과적으로 추출하기 위해 CNN-RNN 구조와 3D CNN 구조를 융합하였으며, 각 CNN으로부터 추출된 feature들을 세 가지 방식의 RNN 구조로 융합하였다[2].

한편, 얼굴표정 인식 관련 데이터 세트는 다른 분야의 데이터 세트에 비해서 데이터의 양이 부족하다. 이를 보완하기 위해서 많은 종래 연구들이 transfer learning의 일종인 미세조정(fine-tuning) 학습 방식을 활용한다. 미세조정은 사전에 잘 학습된 네트워크의 파라미터를 이용하여 초기화 후에 학습하고자 하는 target 데이터 세트를 이용하여 학습을 수행하는 것을 말한다. 이를 통해 얻을 수 있는 이점은 네트워크를 랜덤으로 초기화하는 것보다 학습의 수렴속도가 빠르며, 더 높은 성능을 얻을 수 있다. 또한, target 데이터 세트와 비슷한 데이터를 사전에 학습함으로써 데이터의 부족 문제를 어느 정도 완화할 수 있다.

네트워크	유형	정확도(%)
LBP-TOP & SVM[5]	Hand-crafted feature	39.00
C3D [4]	3D CNN	39.69
SSE-HoloNet[3]	2D CNN	46.48
VGG-LSTM[1]	CNN-RNN	48.60
C3D-LSTM[1]	3D CNN-RNN	43.20
C3D-GRU[2]	CNN-RNN & 3D CNN	**49.87**

표 1. AFEW 데이터 세트에 대한 성능 비교

표 1은 공인 얼굴표정 데이터 세트인 AFEW에 대한 각 기술들의 성능을 나타낸다. AFEW에 경우 학습 데이터로 네트워크를 학습 후에 검증 데이터로 네트워크의 성능을 평가한다. Hand-crafted feature 기반의 방법이 가장 낮은 성능을 보이며, CNN-RNN과 3D CNN의 복합 구조인 [2]의 성능이 가장 우수하다. 흥미로운 점은 3D CNN 기반의 C3D를 단독으로 사용하였을 때 성능이 좋지 않았지만 CNN-RNN과 융합하였을 때 시너지가 남을 확인할 수 있다.

4. 결론

본 논문은 얼굴표정 인식 분야에서의 고유한 문제들을 다룬 기술들을 포함하여 고전적인 기법부터 최신 기법에 대한 연구 동향을 제시한다. 점차 커져가는 인포테인먼트 시스템, HRI 시장과 동시에 그에 상응하는 기술적 요구로서 정확한 얼굴표정 인식 기술에 대한 요구도 높아져만 가고 있다. 이에 따라 시선추정 기술의 주요한 요소 기술 중 하나인 얼굴표정 인식 기술 발전에 대해 그 귀추가 주목된다.

5. 감사의 글

본 연구는 산자부의 산업용지능융합부품기술개발사업의 연구비 지원으로 수행되었음 [과제명: 4K30p 급 Deep Learning 기반 Edge Computing IP 카메라용 시스템반도체 개발].

참고문헌

[1] V. Vielzeuf, S.Pateux, and F. Jurie, “Temporal multimodal fusion for video emotion classification in the wild,” in *Proceedings of the 2017 ACM International Conference on Multimodal Interaction*, pp. 569-576, ACM, 2017.

[2] M. K. Lee, D. Y. Choi, D. H. Kim, and B. C. Song, “Visual scene-aware hybrid neural network architecture for video-based facial expression recognition,” in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 1-8, 2018.

[3] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, “Learning supervised scoring ensemble for emotion recognition in the wild,” in *Proceedings of ACM International Conference on Multimodal Interaction,* pp. 553-560, ACM, 2017.

[4] Y. Fan, X. Lu, D. Li, and Y. Liu, “Video-based emotion recognition using CNN-RNN and C3D hybrid networks,” in *Proceedings of the 2016 ACM International Conference on*

Multimodal Interaction, pp. 445–450, ACM, 2016.

[5] S. Wang, W. Wang, J. Zhao, S. Chen, Q. Jin, S. Zhang, and Y. Qin, "Emotion recognition with multimodal features and temporal models," in *Proceedings of ACM International Conference on Multimodal Interaction,* pp. 598–602, 2017.