동영상 기반 감정인식을 위한 DNN 구조

이민규, 최준호, *송병철
인하대학교
*bcsong@inha.ac.kr

Deep Neural Network Architecture for Video - based Facial Expression Recognition

Min Kyu Lee, Jun Ho Choi, *Byung Cheol Song
Inha University

요 약

최근 딥 러닝의 급격한 발전과 함께 얼굴표정인식 기술이 상당한 진보를 이루었다. 그러나 기존 얼굴표정인식 기법들은 제한된 환경에서 취득한 인위적인 동영상에 대해 주로 개발되었기 때문에 실제 wild 한 환경에서 취득한 동영상에 대해 강인하게 동작하지 않을 수 있다. 이런 문제를 해결하기 위해 3D CNN, 2D CNN 그리고 RNN 의 새로운 결합으로 이루어진 Deep neural network 구조를 제안한다. 제안 네트워크는 주어진 동영상으로부터 두 가지 서로 다른 CNN 을 통해서 영상 내 공간적 정보뿐만 아니라 시간적 정보를 담고 있는 특징 벡터를 추출할 수 있다. 그 다음, RNN 이 시간 도메인 학습을 수행할 뿐만 아니라 상기 네트워크들에서 추출된 특징 벡터들을 융합한다. 상기 기술들이 유기적으로 연동하는 제안된 네트워크는 대표적인 wild 한 공인 데이터세트인 AFEW 로 실험한 결과 49.6%의 정확도로 종래 기법 대비 향상된 성능을 보인다.

1. 서론

얼굴표정인식은 컴퓨터 비전과 human-computer interaction 분야에서 지속적으로 관심을 받고 있는 기술 분야이다. 얼굴표정인식 기술은 차량 내 운전자 모니터링, 로봇과의 상호작용 등 다양한 분야에서 활용될 수 있다. 최근 딥 러닝 같은 기계학습의 발달과 함께 얼굴표정인식 기술 수준이 상당한 진척이 이루기는 했지만 여전히 강인하게 동작하는데 있어서는 어려움이 있다. 왜냐하면 실제 wild 환경은 occlusion, 저조도 등 여러 가지 성능 저하 요인을 포함하고 있기 때문이다. 초기 얼굴표정인식 기술은 한 장의 정지 영상 속 인물(들)의 감정을 분류하는 것이 주류였지만, 최근에는 동영상 속 인물(들)의 감정을 분석하는 연구도 활발하다. 종래 연구에서는 비교적 제한된 환경에서 수집된 CK+ [1]와 같은 공인 데이터세트를 이용하여 알고리즘의 성능을 검증해왔다. CK+ 영상들은 피험자들이 정면을 보고 부자연스럽게 인위적인 감정을 표현하는 영상들이 대부분이다. 또한, 일정한 조도 환경에서 정면의 얼굴 영상들이 취득되었기 때문에 얼굴 각도와 조도 등이 알고리즘 개발에서 고려되지 않는다. 그러나 실제 wild 환경에서 촬영된 동영상들은 다양한 얼굴 각도 및 조도를 갖기 때문에 이에 강인한 얼굴표정인식 기법이 요구된다.

본 논문은 실제 wild 한 비디오 데이터에 대해 강인하게 얼굴표정인식을 수행하기 위해 3D CNN, 2D CNN 그리고 RNN 으로 결합된 새로운 DNN 구조를 제안한다. 먼저 보조 네트워크를 가지는 3D CNN 는 주어진 동영상으로부터 전반적인 시공간 정보를 담고 있는 특징 벡터를 추출하는 역할을 한다. 그리고 2D CNN 으로서 미세조정된 DenseNet [2]은 동영상 내 각 프레임이 갖는 세부적인 공간 정보를 담고 있는 특징 벡터를 추출한다. 마지막으로, RNN 은 상기 두 특징 벡터들을 적절히 융합하고, 최종적인 얼굴표정 분류를 수행한다. 실험을 통해 우리는 제안 기법이 대표적인 wild 공인 데이터세트인 AFEW 에서 49.6%의 우수한 분류 정확도를 보임을 알 수 있었다.

2. 제안 기법

그림 1 과 같이, 제안 기법은 비디오 신호를 2 가지 측면으로 활용한다. 한 가지는 프레임 단위 공간적 정보이다. 각 프레임은 공간적 신호만을 가지고 있기 때문에 해당 시점에서의 인물의 표정과 같은 기하학적 구조에 대한 정보를 제공한다. 이 신호는 우리가 파악하고 싶은 세부적인 공간적 정보에 해당한다. 다른 하나는 시공간 정보를 가지는 비디오 시퀀스 자체이다. 시퀀스는 기본적으로 시간 정보를 포함하기 때문에 표정의 변화 및 분위기에 대한 정보를 제공한다. 따라서 전체 시퀀스는 우리가 파악하고 싶은 전반적인 맥락에 대한 정보에 해당한다. 우리는 입력 동영상으로부터 이 두가지 정보를 정확히 추출하고, 그들을 효과적으로 융합할 수 있는 구조를 제안한다.

2.1 미세조정된 DenseNet

우리는 전이 학습을 통해 얼굴표정인식에 대해 학습된 모델을 세부적인 공간적 정보를 추출하는데 효과적으로 활용되

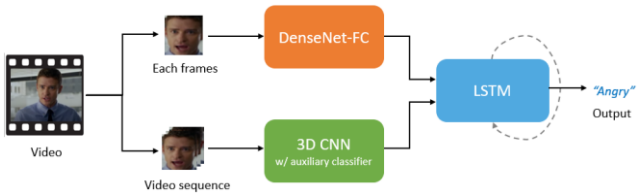

그림 1. 제안하는 DNN 구조에 대한 블록도

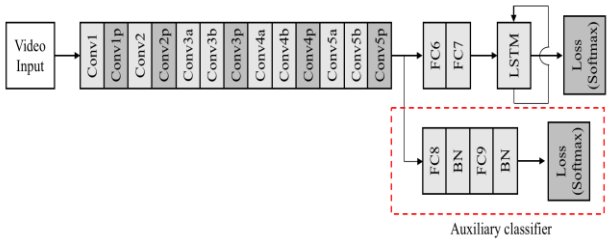

그림 2. 보조 분류기를 가지는 C3D 구조

도록 미세조정된 DenseNet 을 채택하였다. 우리는 기존 DenseNet 의 마지막 layer 인 global average pooling (GAP) 를 두 개의 fully-connected layer (FC) 로 대체하였다. 우리는 이 수정된 네트워크를 DenseNet-FC 라고 명명한다. DenseNet 의 미세조정 과정은 두 단계로 구성된다. 먼저, ImageNet [3] 으로 학습한 이후 얼굴표정인식 관련 공인 데이터세트인 FER2013 [4] 로 미세조정 학습을 수행한다.

2.2 보조 분류기를 가지는 3D CNN

3D CNN 은 입력 영상 내 전반적인 맥락을 포착하는 역할을 한다. 3D CNN 은 C3D [5]를 채택한다. DenseNet-FC 와 마찬가지로 전이 학습을 위해 사전 학습을 수행한다. 또한 우리는 C3D 의 마지막 layer 에 보조 분류기 (auxiliary classifier)를 추가한다. 보조 분류기는 vanishing gradient 를 완화할 뿐만 아니라 정규화 효과가 있기 때문에 학습을 더욱 안정적으로 할 수 있게 해준다. 결과적으로 보조 분류기를 가지는 C3D 의 구조는 그림 2 와 같이, FC, batch normalization, dropout 순서로 구성되면서 마지막은 softmax 함수가 온다

2.3 특징 간 융합 RNN 구조

우리는 DenseNet-FC 와 보조 분류기를 가지는 C3D 로부터 추출한 특징 벡터들을 효과적으로 융합하기 위해 특징 간 융합 RNN 구조를 제안하며. 각 신호의 고유 정보들을 보존하면서 융합할 수 있다. 이 때 RNN 의 모델로 Long Short-term Memory (LSTM) [6]을 채택한다.

기존 LSTM 에서는 시계열 데이터를 입력으로 받지만, 제안 기법은 시계열 데이터에 해당하는 DenseNet-FC 로 추출한 각 프레임의 특징 벡터뿐만 아니라 C3D 로부터 추출된 특징 벡터를 입력한다. 수식적으로 다음과 같다.

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + W_{iv}v + b_i) \quad (1)$$
$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + W_{fv}v + b_f) \quad (2)$$
$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + W_{ov}v + b_o) \quad (3)$$
$$g_t = \sigma(W_{gx}x_t + W_{gh}h_{t-1} + W_{gv}v + b_g) \quad (4)$$

Method	Accuracy (%)
C3D [5]	41.25
ResNet 3D [7]	39.16
ResNeXt 3D [7]	41.01
LSTM [6]	46.74
GRU [8]	46.99
Proposed	**49.61**

표 1. AFEW 데이터세트에 대한 인식 성능 비교

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (5)$$
$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

여기서, x_t 는 DenseNet-FC 로부터 추출된 특징 벡터, v 는 C3D 로부터 추출된 특징 벡터를 나타낸다. 식 (1) 부터 (4) 는 LSTM 의 네 가지 게이트들의 식을 나타내며, i_t 는 입력 게이트, f_t 는 forget 게이트, o_t 는 출력 게이트, g_t 는 candidate 게이트이다. 식 (5), (6) 은 cell state c_t 와 hidden state h_t 를 나타낸다. 또한, σ 는 sigmoid 함수, $\tanh$ 는 hyperbolic tangent 함수, W 는 weight matrix, b 는 bias, 그리고 $\odot$ 는 hadamard product 를 나타낸다.

3. 실험 결과

제안 기법을 평가하기 위해 얼굴표정인식 공인 데이터세트인 AFEW 를 이용하여 실험하였다. Optimizer 는 SGD 를 사용하였고 mini-batch size 는 32 로 설정하였으며, GPU 는 GTX 1080Ti, 딥러닝 프레임워크는 PyTorch 를 사용하였다.

비교 기법으로 동영상 입력을 처리할 수 있는 모델을 채택하였으며, 3D CNN 에는 C3D 를 포함하여 ResNet 3D, ResNeXt 3D [7] 를 이용하였고, RNN 에서는 LSTM 과 GRU [8]를 이용했다.

먼저 표 1 은 종래 단일 모델들과 제안 기법의 성능을 나타낸다. 제안 기법의 성능은 49.61% 로 단일 모델 중 최고 성능 대비 약 3.3% 이상 향상된 성능을 보인다. 이는 단일 모델을 단독으로 사용되는 것보다 제안 기법과 같이 특징 간 융합하는 것이 긍정적인 시너지를 낸다는 것을 의미한다.

4. 결론

본 논문은 동영상 기반 얼굴표정인식을 할 수 있는 새로운 DNN 구조를 제안한다. 제안 네트워크는 2D CNN 과 3D CNN 에서 추출한 특징 벡터를 특징 벡터 간 융합하는 방식으로 시너지 효과를 유발한다. 실험을 통해 제안 기법은 wild 한 상황에서의 영상에서도 종래 기법과 비교하여 강인하게 동작함을 확인했다.

6. 감사의 글

본 논문은 산업통상자원부의 산업기술혁신사업으로 지원된 연구결과이며 [10073154, 인간 내면상태의 인식 및 이를 이용한 인간친화형 인간-로봇 상호작용 기술 개발], 2019 년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구이며 (N0002428 , 2019 년 산업전문인력역량강화사업), 한국연구재단의 지원을 받아 수행된 연구임(2016R1A2B4007353).

7. 참고문헌

[1] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010, June). The extended cohn - kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 94 - 101). IEEE.

[2] Huang, G., Liu, Z., Weinberger, K. Q., and van der Maaten, L. (2017, July). Densely connected convolutional networks. In Proceedi ngs of IEEE Conference on Computer Vision and Pattern Recognition (Vol. 1, No. 2, p. 3).

[3] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei - Fei, L. (2014). Large - scale video classification with convolutional neural networks. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (pp. 1725 - 1732).

[4] Goodfellow, I. J., et al. (2013, November). Challenges in representation learning: A report on three machine learning contests. In International Conference on Ne ural Information Processing (pp. 117 - 124). Springer, Berlin, Heidelberg.

[5] Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In Proceedings of IEEE International Conf erence on Computer Vision (pp. 4489 - 4497).

[6] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735 - 1780.

[7] Hara, K., Kataoka, H., and Satoh, Y. (2018, June). Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA (pp. 18 - 22).

[8] Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.