이상치 억제를 통한 얼굴의 표정 조작

김성호, *송병철
인하대학교, *인하대학교
sbde500@gmail.com, *bcsong@inha.ac.kr

Facial Expression Manipulation with Outlier Suppression

Seong Ho Kim *Byung Cheol Song
Inha University *Inha University

요 약

얼굴 표정 데이터셋에는 특정 감정 부류로 분류하기 어려운 이상치들이 존재한다. 이러한 이상치들은 얼굴 표정 인식과 더불어 얼굴 표정 조작의 성능을 저하시키는 원인 중 하나이다. 따라서, 본 논문에서는 이상치 억제를 통한 개선된 얼굴 표정 조작 프레임워크를 제안한다. 우리는 이상치 억제를 위해 의미론적 속성 분류 측면에서 우수한 성능을 보여주는 CLIP 을 활용하였다. 우리는 정성적인 비교를 통해 기존의 얼굴 표정 조작 기법보다 개선된 성능을 제시한다.

1. 서론

얼굴 표정은 사람의 얼굴과 관련된 컴퓨터 비전 분야에서 큰 관심을 받고 있는 주제 중 하나이다. 최근 딥러닝의 발전과 함께 얼굴 표정 인식 기술을 넘어 자연스러운 얼굴 표정 조작을 위한 연구가 활발히 진행 중이다. 최신 얼굴 표정 조작 프레임워크인 NED [1]는 감정의 categorical label 을 기반으로 연구를 진행하였고 의미있는 결과를 보여주었다.

하지만 NED [1]는 아직까지 정성적 결과 측면에서 한계를 보이고 있다. 우리는 그 이유 중 한가지를 categorical label 기반 감정 데이터셋에 존재하는 이상치 때문이라 가정하였다. 우리가 정의한 이상치란 얼굴 표정 상에서 약하게 표현된 감정이나 모호하게 표현된 감정들로 인해 특정 부류로 분류하기 어려운 영상들을 의미한다.

본 논문에서는 이러한 문제점을 해결하기 위해 이상치 억제를 통한 개선된 얼굴 표정 조작 프레임워크를 제안한다. 우리는 데이터셋에 존재하는 이상치를 탐지 및 억제하기 위하여 의미론적 속성 분류 분야에서 강력한 성능을 보이는 CLIP [2]을 활용한다.

본 논문에서 제안하는 기법을 통해 우리는 MEAD 데이터셋을 이용한 실제 얼굴 표정 조작 결과에서 기존의 NED [1]보다 향상된 정성적 결과를 제시한다.

2. NED

NED [1]는 parametric 3D 얼굴 복원 기법 기반의 얼굴 표정 조작 프레임워크이다. 이러한 기법은 얼굴에서 표정과 신원을 분리하여 잠재 벡터를 얻을 수 있다는 점에서 이점이 있다. NED [1]는 분리된 잠재 벡터 중 표정에 대한 벡터만을 조작하여 최종 얼굴 표정이 변화된 영상을 생성하게 된다.

이러한 이점에도 불구하고 NED [1]는 정성적 결과 측면에서 한계를 보인다. 이에 대한 이유로 우리는 데이터셋에 포함된 이상치를 원인으로 가정한다. 실제로 얼굴 표정

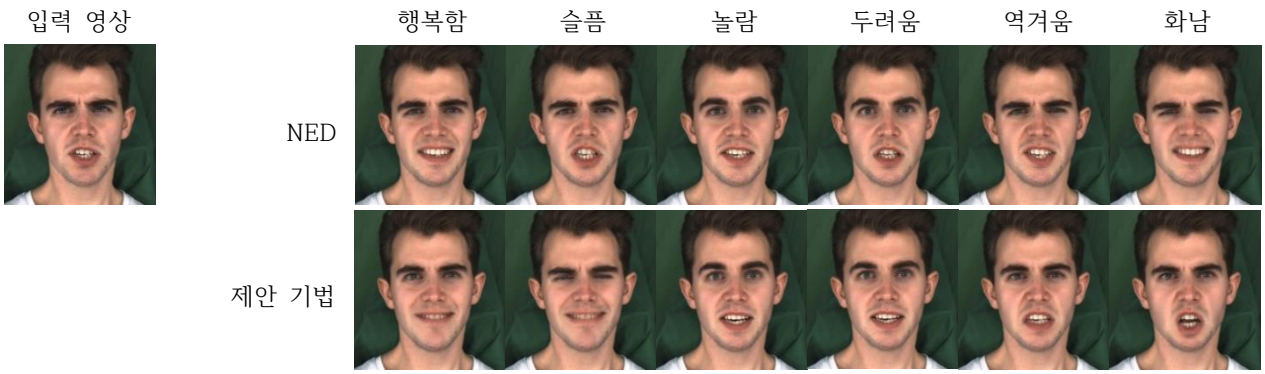

그림 1. NED (위)와 제안하는 기법 (아래)의 정성적 성능 비교

데이터셋에 존재하는 이상치는 얼굴 표정 기반 감정 분류 분야에서 성능하락의 원인 중 하나라고 밝혀진 바 있다 [3].

3. 제안 기법

우리는 NED [1]의 정성적 한계를 극복하기 위하여 CLIP [2]을 활용한 이상치 억제를 통한 개선된 얼굴 표정 조작 기법을 제안한다. CLIP [2]은 대용량의 영상-텍스트 쌍을 통해 학습되어 영상의 의미론적 특징을 분류하는데 특화된 모델이다. 이러한 이점은 얼굴 표정을 통한 감정 분류 문제에서도 탁월한 성능을 보여준다.

우리는 학습 전에 CLIP [2]을 활용하여 훈련 데이터셋의 label 별 감정 분류 예측 값을 확인한다. 이때, 예측 값이 threshold 보다 낮은 경우 해당 값을 기존 NED [1]학습을 위한 목적함수의 계수로 사용한다. 수식은 다음과 같으며,

$$\lambda = \begin{cases} 1 & \hat{y}_{CLIP} \geq 0.6 \\ \hat{y}_{CLIP} & \hat{y}_{CLIP} < 0.6 \end{cases} \quad (1)$$

λ는 목적함수의 계수이고 $\hat{y}_{CLIP}$은 CLIP [2]을 통한 label 별 감정 분류 예측 값이다. 이 방법은 얼굴 표정 발현이 클수록 감정 분류 예측 값이 커진다는 사전 지식에 기인한다 [4]. 최종 목적함수는 다음과 같다.

$$\mathcal{L} = \lambda \mathcal{L}_{NED} \quad (2)$$

4. 정성적 결과

그림 1 는 MEAD 데이터셋을 이용해 기존 NED [1]와 우리 기법의 감정 부류 별로 얼굴 표정 조작을 한 결과이다. 얼굴의 요소별로 비교를 해보면 이전의 결과 보다 해당 감정을 더 잘 표현하는 것을 알 수 있다. 특히, 행복함, 놀람, 역겨움, 화남의 감정에서 개선된 정성적 결과를 보임을 알 수 있다. 그림 1 의 정성적 결과를 통해 데이터셋에 포함된 이상치가 얼굴 표정 조작의 성능 하락 시키는 원인 중 하나라는 것을 알 수 있다.

5. 결론

우리는 얼굴 표정 데이터셋에 존재하는 이상치를 억제하는 것이 얼굴 표정 조작 프레임워크의 성능 향상에 도움이 될 수 있다는 것을 실험적으로 증명하였다. 해당 방향을 기준으로 의미있는 억제 방법을 사용한다면 더 나은 성능을 기대해 볼 수 있을 것이다.

6. Acknowledgement

The authors greatly thank the three anonymous reviewers for their kind comments which have really helped improve the quality of our paper. This work was supported by IITP grants funded by the Korea government (MSIT) (No. 2021-0-02068, AI Innovation Hub and RS-2022-00155915, Artificial Intelligence Convergence Research Center (Inha University)), and was supported by the NRF grant funded by the Korea government (MSIT) (No. 2022R1A2C2010095 and No. 2022R1A4A1033549).

7. 참고문헌

[1] Papantoniou, F. P., Filntisis, P. P., Maragos, P., & Roussos, A. (2022). Neural Emotion Director: Speech-Preserving Semantic Control of Facial Expressions in" In-the-Wild" Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18781-18790).

[2] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748-8763). PMLR.

[3] She, J., Hu, Y., Shi, H., Wang, J., Shen, Q., & Mei, T. (2021). Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6248-6257).

[4] d'Apolito, S., Paudel, D. P., Huang, Z., Romero, A., & Van Gool, L. (2021). Ganmut: Learning interpretable conditional space for gamut of emotions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 568-577).