FACS 기반 GAN 기술을 이용한 가상 영상 아바타 합성 기술

김건형*, 박수현*, 이상호*
*주식회사 그루크리에이티브랩
dream201@gamecrewlab.com, suhyun.park@gamecrewlab.com,
gamecrew@gamecrewlab.com

Video Synthesis Method for Virtual Avatar Using FACS based GAN

Geonhyeong Kim*, Suhyun Park*, Sang Ho Lee*
*GREW Creative Lab. Inc.
dream201@gamecrewlab.com, suhyun.park@gamecrewlab.com,
gamecrew@gamecrewlab.com

요 약

흔히 DeepFake로 불리는 GAN 기술은 소스 영상과 타겟 이미지를 합성하여 타겟 이미지 내의 사람이 소스 영상에서 나타나도록 합성하는 기술이다. 이러한 GAN 기반 영상 합성 기술은 2018년을 기점으로 급격한 성장세를 보이며 다양한 산업에 접목되어지고 있으나 학습 모델을 얻는 데 걸리는 시간이 너무 오래 소요되고, 감정 표현을 인지하는 데 어려움이 있었다. 본 논문에서는 상기 두가지 문제를 해결하기 위해 Facial Action Coding System(FACS) 및 음성 합성 기술[4]을 적용한 가상 아바타 생성 방법에 대해 제안하고자 한다.

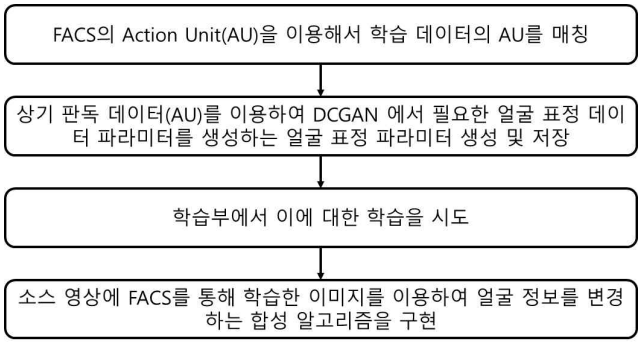
1. 서론

최근의 AI 기술은 기술 개발 자체에 대한 집중도가 높았지만 연구 개발이 어느정도 범용화될 수 있는 단계에 다다름에 따라 적용 분야에 특화한 서비스 개발이 활발해 지고 있다. 이 가운데 최근 딥러닝을 활용하여 제작한 콘텐츠인 GAN 기술 기반의 딥페이크(Deepfake)가 확산 중이며, 기존의 GAN 기법은 불안정하고 복잡한 영상의 조합에 어려움을 겪었으나 DCGAN[1]이 부상하면서 딥러닝을 통해 이러한 문제점을 해결할 수 있게 되었다. 이러한 GAN 기반의 딥페이크 기술이 차세대 ICT 기술로 부각되고 있으며, AI 기술을 융합하고자 하는 다양한 분야의 주요 기업들이 해당 기술을 응용하여 다양한 서비스 개발을 위해 집중 투자하고 있다. 하지만, 상기 단점을 해결하려는 GAN 기반의 많은 시도가 있었지만 기존의 edge 검출 기반의 특징점 선정 알고리즘은 영상 전체의 프레임에 대해 처리 시간이 오래 걸리고, 이는 곧 학습 품질을 보장하기 위해서는 컴퓨팅 파워와 시간이 많이 소비된다는 것을 의미하여 어려움이 있었다. 또한 edge 검출 기반 방식은 다양한 사람 얼굴의 감정 형태를 인식하는 데 정확도가 높지 않다는 단점이 있다[5]. 이에 본 논문에서는 FACS[2]를 이용하여 표준화된 사람의 감정 표현 방법을 DCGAN 학습을 위한 특징점으로 정의하여 정확도를 높이고 결과물의 품질을 끌어올리는 기법에 대해 [그림 1]과 같이 제안한다. 이를 위해 2장에서는 FACS 기반 GAN 기술[3, 4]에 대해 간단하게 설명하고, 3장에서는 실제 테스트를 예시로 한 성공/실패에 대한 영상 합성 예시 및 원인 분석을 실시한다. 마지막으로 4장에서 본 논문에 대해 요약하고 향후 연구 방향에 대해 서술한다.

[그림 1] 제안하는 기술 개요

2. FACS 기반 GAN

본 논문에서 제안하는 기술을 구현하기 위해서는 [그림 2]와 같이 4가지 단계로 구분된다.



[그림 2] FACS 기반 DCGAN 학습기 프로세스

첫째, FACS의 Action Unit(AU)을 이용해서 학습 데이터의 AU를 매칭한다. 이를 이해서는 FACS의 각각의 표정(감정)에 대한 AU가 레이블링 된 표정 데이터 셋이 존재해야 하며, 제안하는 논문에서는 약 5만장 이상의 최소 4개 이상의 표정에 대한 사진 데이터를 취득하여 학습을 위한 데이터 셋 데이터베이스를 구축하고 이를 학습에 이용하였다. 데이터 셋 구축이 마무리 되면 두번째로 상기 판독 데이터(AU)를 이용하여 DCGAN 에서 필요한 얼굴 표정 데이터 파라미터를 생성하는 얼굴 표정 파라미터 생성 및 저장 기능을 구현한다. 얼굴 표정에 대한 파라미터 생성/저장이 완료되면, 세 번째로 학습부에서 이에 대한 학습을 시도한다. 마지막으로, 소스 영상에 FACS를 통해 학습한 이미지를 이용하여 얼굴 정보를 변경하는 합성 알고리즘을 구현하게 되며, DCGAN과 연계하여 Action Unit에 따른 표정 합성 및 재현 방법을 학습된 FACS 결합 모델을 이용하여 Inference 한다.

3. 학습 결과 및 분석

제안하는 기술을 실제로 학습하기 위한 PC 사양은 아래와 같다.

- CPU : Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz 2.80GHz
- GPU : Geforce GTX 2080 Ti
- HDD : HGST HTS721010A9E630 1TB
- OS : Windows 10 Pro (x64)

또한, 본 과제에서 제안하는 기술을 이용하여 GAN 네트워크를 학습 시킨 후 재현에 성공한 영상은 [그림 3]과 같다.

[그림 3] 소스 영상(좌측 상단), 변경할 얼굴이 포함된 타겟 영상(우측 상단), 및 두 영상을 합성해서 나온 결과 이미지(하단)

반면에, 합성에 실패한 영상은 아래 [그림 4]와 같으며, 소스 영상과 타겟 영상이 너무 이질감이 있는 경우에는 제대로 된 합성이 되지 않는다. [그림 4]의 영상에서는 성별이 다른 영상을 학습했기 때문에 이질감이 있어 학습이 제대로 되지 않았다.

[그림 4] 소스 영상(좌측 상단), 변경할 얼굴이 포함된 타겟 영상(우측 상단), 및 두 영상 합성에 실패한 결과 이미지(하단)

4. 결론

본 논문는 실감 영상에 이질감 없이 가상 아바타를 생성/적용할 수 있도록 하는 DCGAN 및 FACS 기반 특징점 추출 및 실감 아바타 영상 제작 방법을 제안하였다. 제안하는 기술들은 아직 완성도 측면에서 산업에 적용할 수 있을 만큼 자연스럽지는 않으나, 차차 개선해 나갈 예정이다. 이러한 기술을 적용하여 제작된 실감 영상은 산업적 활용가치가 높아 영화, 음반 등 다양한 분야에서 활용 중이며, 제작비용을 기존의 약 10%로 낮출 수 있을 것[6]으로 기대하고 있다.

사사(Acknowledge)

본 논문은 (재)인천테크노파크가 지원하는 'SW융합클러스터 2.0' 사업의 '2020년 SW융합 제품/서비스 상용화 지원사업(2차)'로 지원을 받아 수행된 연구 결과입니다. [과제명: DCGAN을 이용한 실감 아바타 영상 제작 플랫폼 제작 및 상용화 / 과제고유번호: S1225-21-1002]

참고문헌

[1] Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." arXiv preprint arXiv:1511.06434 (2015).

[2] Ekman, Rosenberg. What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA, 1997.

[3] Wang, Ting-Chun, et al. "Few-shot video-to-video synthesis." arXiv preprint arXiv:1910.12713 (2019).

[4] Prajwal, K. R., et al. "A lip sync expert is all you need for speech to lip generation in the wild." Proceedings of the 28th ACM International Conference on Multimedia. 2020.

[5] Chen, Xiaoming, and Wushan Cheng. "Facial expression recognition based on edge detection." International Journal of Computer Science and Engineering Survey 6.2 (2015): 1.

[6] Skiendziel, Tanja, Andreas G. Rosch, and Oliver C. Schultheiss. "Assessing the convergent validity between the automated emotion recognition software Noldus FaceReader 7 and Facial Action Coding System Scoring." PloS one 14.10 (2019): e0223905.