창의융합연구
Journal of Creativity and Convergence
2022, Vol. 2, No. 2, pp.1-23
http://dx.doi.org/10.23021/JCC.2022.2.2.1

비전 · 그래픽스 인공지능 기술을 활용하여 무엇을 연구할 수 있는가?[1)]

강지우 숙명여자대학교 ICT 융합공학부 교수[2)]

초록

본 연구에서는 비전 그래픽스 인공지능 기술을 활용한 다양한 연구 사례들을 소개하고자 한다. 공학뿐만 아니라 인문 사회 교육 예체능 등 다양한 분야에서 학문적 기반을 가진 연구자들에게 비전 그래픽스 인공지능 기술을 활용한 의미 있는 연구 사례들을 공유하여 인공지능을 통해 무엇을 해볼 수 있는지 어떤 융합을 진행할 수 있는지 고민해 볼 수 있는 기회를 제공해보고자 한다.

1, 서론

비전은 컴퓨터 비전이라는 용어에서 나온 단어로서, 현실을 인식하고 이해하는 기술을 의미한다. 예를 들어서 동영상에 있는 개체를 감지 또는 분류하는 기술, 원하는 개체를 추적하는 기술, 영상에서 일어난 일을 분류하는 기술 등을 포함한다. 인공지능 기술을 통해서 영상에서 정보를 인지할 때 3 차원 정보를 매개로 또는 결과로 필요로 하는 경우가 있다. 예를 들어서, 영상 안에 있는 사람의 행동을 분류할 때 영상 안의 사람의 자세를 2 차원으로 인지하려는 것보다는 3 차원으로 인지하려는 것이 훨씬 정확도를 크게 올려줄 수 있다. 이런 경우 3 차원 정보를 더 정확한 정보를 얻기 위한 매개로 사용했다고 볼 수 있다. 또는 얼굴 사진에서 3 차원 아바타 얼굴을 생성하거나 사람 사진에서 사람의 전신 아바타를 생성하는 기술은 3 차원 정보를 결과로서 얻기 위해서 사용했다고 볼 수 있다. 이렇게 컴퓨터 비전 기술 중에서 3 차원을 매개 또는 결과로서 필요로 하는 기술을 가리켜 3 차원 비전이라고 한다. 3 차원 비전 기술은 결과가 3 차원으로 나오기 때문에 결과를 효과적으로 가시화하기 위해서 또는 때로는 더 좋은 결과를 얻기 위해서는 컴퓨터 그래픽스 기술과 결합되어서 사용되고 있다. 예를 들어서, 한 사람의 얼굴

1) 본 원고는 2022 년 11 월 4 일 개최된 창의융합연구소 국내 학술대회의 특별 세션 발표 자료를 재구성한 원고임.
2) jwkang@sookmyung.ac.kr

사진을 입력으로 받아서 3 차원 얼굴 만들고, 그 3 차원 얼굴을 조작해서 다시 2 차원 영상을 만드는 딥페이크가 대표적인 3 차원 비전 그래픽스 기술이라고 볼 수 있다. 이 논문에서는 이러한 비전 그래픽스 인공지능 기술 분야의 최신 기술들에 대해서 공유하고 논의해보도록 한다.

2. Human Body

2.1. 2D Skeleton from Video

그림 1 OpenPose[3)]

2017 년에 공개된 'OpenPose'라는 기술은 이미지에 있는 사람들이 3 차원 관절의 위치를 실시간으로 구해줄 수 있는 인공지능 기술이다. 움직이는 영상에 대해서도 매우 정확한 검출 능력을 보이기 때문에 이 기술이 개발된 이후에 사람의 몸 자세 또는 관절 추정에 대한 인공지능 연구가 크게 발전하였다. 움직이는 동영상에서 각각의 프레임을 각각의 사진으로 보고 인공지능 네트워크를 적용해서 사람의 뼈대를 연속으로 검출하여 관절에 이어서 만들어 보여줄 수 있다. 실시간으로 나타나는 관절의 위치를 검출하기 위해서 이전 시간에 검출한 정보를 전혀 이용하지 않는데도 매우 부드럽고 정확하게 검출이 가능하다. 'OpenPose'는 꽤 정확하게 3 차원 관절의 위치를 검출해주지만 크게 두 가지 정도의 한계가 존재한다. 첫 번째로 정확한 검출을 위해서 고용량 GPU 를 필요로 한다는 점이다. 실시간으로 쓸 만한 데이터를 검출하려면 현재 최소 100 만 원 상당의 GPU 를 필요로 한다. 두 번째로 관절 정보는 완벽한 3 차원 자세 정보를 제공하지 않는다. 예를 들어, 팔을 앞으로 내밀고 손바닥을 앞뒤로 뒤집는 경우 손목 관절과 팔 관절이 회전하지만, 손목과 팔 관절의 위치는 변하지 않는다. 관절 정보는 이러한 관절 방향에

3) Cao et al. "Realtime Multi-person 2D Pose Estimation using Part Affinity Fields," IEEE Conference on Computer Vision and Pattern Recognition 2017.

대한 회전 정보를 포함하지 못하기 때문에 관절의 위치 정보만으로는 사람이 어떤 자세를 취하는지에 대한 3 차원 자세 정보를 정확하게 검출해 낼 수 없다.

2.2. 3D Shape and Pose from a Single Image

그림 2 A Skinned Multi-Person Linear Model[4)]

2016 년 컴퓨터과학 분야 권위 국제 학술대회 ECCV (European Conference on Computer Vision)에서 3 차원 사람 모델을 이용하여, 단일 사진에서 3 차원 자세를 취하고 있는 사람의 메시 모델을 만들어주는 기술이 발표되었다. 3 차원 사진과 앞에서 보인 오픈 포즈에서 뽑은 관절 정보를 입력으로 주어지면 결과로서 사진과 똑같은 자세를 취한 3 차원 아바타 모델을 출력한다. 이러한 입력과 출력을 가지는 기술은 계속 조금씩 정확도가 개선되어 여러 연구자에 의해 매년 발표되고 있다.

2.3. Multiple 3D Shapes and Poses from a Single Image

그림 3 Realtime Multi-person 2D Pose Estimation using Part Affinity Fields[5)]

4) Bogo et al. "Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image," European Conference on Computer Vision 2016.

5) Cao et al. "Realtime Multi-person 2D Pose Estimation using Part Affinity Fields," IEEE Conference on Computer Vision and Pattern Recognition 2017.

한 장의 이미지에서 3 차원 메쉬를 생성하는 문제가 여러 사람이 찍힌 사진에서 여러 3 차원 메쉬를 만드는 문제로 자연스럽게 확장되었다. 여러 사람의 사진에서 사람의 자세를 검출하는 건 가린 영역 가리는 영역이 존재하기 때문에 훨씬 어려운 문제이다. 보인 영역을 바탕으로 각각의 사람의 자세를 만들고 보이지 않는 부분은 통계적으로 사람이 자주 취하게 되는 자세를 임의로 선택하게 된다. 여러 사람을 찍은 사진은 한 사람에 대한 정보량이 그만큼 부족해지기 때문에 한 사람을 검출하는 기술과는 사용되는 애플리케이션이 다를 수 있다.

2.4. 3D Shapes and Poses from a Dynamic Sequence

그림 4 Video Inference for Human Body Pose and Shape Estimation[6)]

VIBE (Video Inference for Human Body Pose and Shape Estimation)는 컴퓨터과학분야 국제학술대회 CVPR (IEEE Conference on Computer Vision and Pattern Recognition) 에서 2020 년에 발표된 자세 검출 기술을 매우 빠른 동영상으로 확장한 기술이다. 동영상에 대해서 자세를 정확하게 그리고 안정적으로 검출하는 기술은 지속적으로 발전 중인 기술이다. 제시한 영상의 사례에서는 2 년 전 가장 최근 기술 중 하나임에도 불구하고 사람이 모델이 작아졌다가 커졌다가 하며 크기가 일정하지 않은 것을 확인할 수 있다. 그럼에도 현재 기술은 매우 빠르게 움직이는 사람의 동영상에 대해서도 꽤 정확하게 검출한다고 볼 수 있다.

6) Kocabas et al., "Vibe: Video Inference for Human Body Pose and Shape Estimation," IEEE Conference on Computer Vision and Pattern Recognition 2020.

2.5. 3D Shapes and Poses from Video Clips

그림 5 Reconstructed 3D Humans and Environments in TV Shows[7)]

2022 년 컴퓨터과학분야 권위 국제학술대회 CVPR 에서 우리가 흔히 보는 TV 쇼 영상에서 나온 주인공들의 3 차원 자세를 검출하는 인공지능 기술이 발표되었다. 제시한 예제에서도 볼 수 있듯이 몸에 많은 부분이 가려져 있거나 형체를 알아보기 어려운 모자나 옷을 착용하고 있어도 꽤 정확하게 검출하는 것을 확인할 수 있다. 이러한 기술은 TV 쇼가 재생되는 동안 배경을 모두 분석해서 공간 정보에 대해서 분석하고 TV 쇼에서 나오는 각각의 장면에 대한 배경의 3 차원 정보를 얻어내는 기술을 바탕으로 한다. 이러한 공간 정보를 바탕으로 주인공에 대한 정보를 장면에서 분리해내고 인공지능 기술을 통해서 3 차원으로 만들어낼 수 있다.

2.5.1. 3D Human and Object Reconstruction from a Single RGB Image

그림 6 Contact, Human and Object Reconstruction[8)]

7) Pavlakos et al. "The One Where They Reconstructed 3D Humans and Environments in TV Shows," European Conference on Computer Vision 2022.

8) Xie et al. "CHORE: Contact, Human and Object REconstruction from a single RGB image," European Conference on Computer Vision 2022.

사진에서 인물뿐만 아니라 인물이 착용하고 있는 객체 또는 인물이 상호작용하고 있는 객체에 집중한 인공지능 연구도 진행되고 있다. 사람의 자세가 제대로 검출되려면 상호작용하고 있는 객체도 같이 검출해야 한다는 시각에서 나온 연구이다.

2.6. Optimal View Point Selection for 3D Human Pose

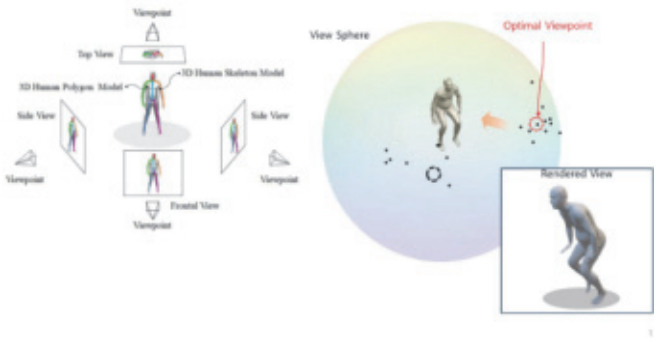

그림 7 Optimal Camera Point Selection Toward the Most Prefer able View of 3-D Human Pose[9)]

위 연구들과는 다른 관점의 기술로써 3 차원 사람 모델이 특정 자세를 취할 때 어떤 3 차원 각도에서 바라봐야 그 자세를 파악하기가 좋은지 또는 그 자세를 보기가 좋은지 연구하는 기술이 2022 년에 세계 권위 학술지 IEEE Transactions on Systems, Man, and Cybernetics: Systems 에 출판되었다.

지금까지 소개한 기술들은 3 차원 사람 체험 모델을 통해서 3 차원 메시를 예측한 기술로, 모델이 옷을 입은 상태가 아니라 모두 체형이 드러나 있다. 3 차원 채용 모델은 몇백 명 이상의 사람을 3 차원 스캔을 하고 이를 통해서 수많은 사람의 3 차원 통계적 모델을 만들어 놓은 다음 인공지능 모델을 통해서 입력 사진상의 사람이 통계적 모델의 어디쯤 위치하는지를 예측하게 하는 방식이다. 3 차원 모델이 어떠한 체험과 어떠한 자세를 가질 때 사진상에 있는 사람과 비슷해질지를 학습하는 방식으로 인공지능이 학습된다. 그래서 앞선 기술들은 사진상의 사람의 체형을 예측하거나 자세를 예측하는 기술이며 사진상의 사람과 똑같은 3 차원 메시를 만들어주는 기술은 아니다. 물론 사진상 사람의 3 차원 자세를 알게 되면 임의의 캐릭터에 똑같은 자세를 취할 수 있게 만들 수 있기 때문에 앞선 기술을 통해서 우리가 미리 만들어 놓은 캐릭터로 나와 똑같은 자세를 취하게 할 수 있다.

9) Kwon et al. “Optimal Camera Point Selection Toward the Most Prefer able View of 3-D Human Pose,” IEEE Transactions on Systems, Man, and Cybernetics: Systems 2022.

3. Full Avater (Clothed Body)

3.1. 3D Avatar from a Single Facial Image

그림 8 Normalized Avatar Synthesis Using StyleGAN and Perceptual Refinement[10]

2021 년 컴퓨터과학 분야 세계 권위 학술대회 CVPR 에서는 사진 한 장만으로 꽤 현실적인 '나'와 닮은 아바타를 만드는 기술이 발표되었다. '제페토'와 같은 VR 플랫폼은 '나'와 비슷한 캐릭터를 만들어주는 방식을 취하지만 이 플랫폼은 '나'와 비슷한 아바타를 만들어준다. '나'와 똑같은 또는 유사한 아바타를 만드는 것에 대한 필요와 수요가 형성됨에 따라 만들어진 기술이다. 이 기술은 현재 핀 스크린이라는 미국 회사에서 현재 상용화 중인 인공지능 기술로 수준이 매우 높은 기술이다. 예제의 영상에서 실제로 아바타가 움직이는 것은 '나'의 움직임을 따라하는 것이 아니라 미리 정의되어 있는 모션을 실행한 것이다.

3.2. RGB Video to Avatar

그림 9 DeepCap: Monocular Human Performance Capture using Weak Supervision[11)]

10) Luo et al. "Normalized Avatar Synthesis Using StyleGAN and Perceptual Refinement." IEEE Conference on Computer Vision and Pattern Recognition 2021.

11) Habermann et al. "DeepCap: Monocular Human Performance Capture using Weak Supervision," IEEE Conference on Computer Vision and Pattern Recognition 2020.

2020 년 CVPR 에서 발표된 DeepCap 은 임의의 동영상에 있는 사람을 3 차원 Mesh 로 만들어주는 기술이다. 아주 미세한 동작까지 정확하지는 않지만, 앞의 기술과는 달리 옷의 주름 등이 모두 반영되어 3 차원으로 만들어지는 것을 확인할 수 있다. 또한 이렇게 만들어진 '나'의 아바타는 '나'의 동작뿐만 아니라 누가 미리 찍어둔 모션으로 움직이게 만들 수도 있다.

3.3. RGB Video (Image) to Avatar

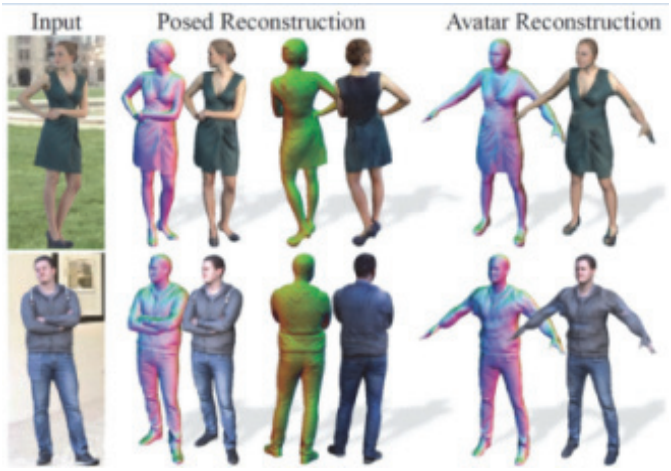

그림 10 Animation-ready Clothed Human Reconstruction Revisited[12]

2021 년 CVPR 에서 발표된 ARCH++기술은 위와 유사한 종류의 기술이다. 단일 사진 또는 여러 사진으로부터 3 차원 아바타를 생성해 낼 수 있다. 주름 등이 꽤 잘 잡히지만 당연하게도 정보량이 부족하기 때문에 이음새 또는 가린 부분이 어색할 수밖에 없다. 이렇게 만들어진 아바타는 미리 찍어둔 모션에 의해서 움직일 수 있다.

12) He et al. "ARCH++: Animation-ready Clothed Human Reconstruction Revisited." IEEE International Conference on Computer Vision 2021.

4. Facial Avatar

4.1. 3D Avatar Face from a Single Image

그림 11 Learning an Animatable Detailed 3D Face Model from in-the-wild Images[13)]

최근 얼굴에 대한 많은 연구가 이루어지고 있어 인공지능 모델을 통해서 상당히 좋은 퀄리티로 3 차원 생성이 가능하다. 이렇게 생성된 모델은 사진 속 표정뿐 아니라 자유롭게 다양한 표정을 취할 수 있다.

4.2. 3D Single-Image Face Modeling

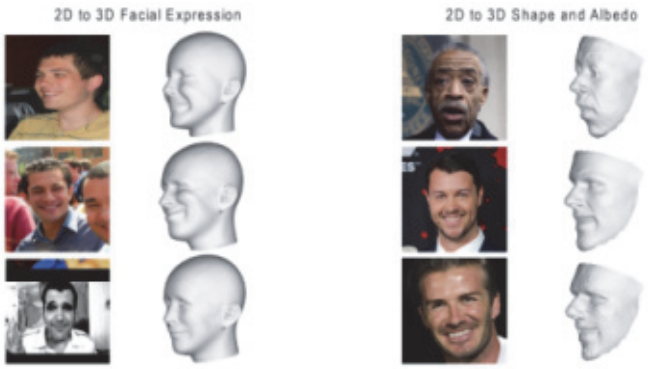

그림 12 2D to 3D Facial Expression[14)] / 2D to 3D Shape and Albedo[15)]

13) Feng et al. "Learning an Animatable Detailed 3D Face Model from in-the-wild Images." ACM Transactions on Graphics 2021.
14) Jiwoo Kang and Sanghoon Lee, "A Greedy Pursuit Approach for Fitting 3D Facial Expression Models," IEEE Access 2020
15) Suwoong Heo, Hyewon Song, Jiwoo Kang, and Sanghoon Lee, "Local Spherical Harmonics for Facial Shape and Albedo Estimation," IEEE Access 2020

사람이 표정을 짓는 것을 근육의 움직임이라고 생각해서 근육 움직임 간의 유사도를 최대화하고 중복을 최소화하는 방향으로 사진에 있는 사진에 있는 표정을 생성하는 기술이 2020 년에 세계 우수학술지 IEEE Access 에 공개되었다. 이 연구는 단일 이미지에 있는 사람의 표정을 3 차원으로 과하지 않으면서 가장 효과적으로 드러내는 연구이다. 또한 사람 얼굴에 있는 빛을 정밀하게 모델링해서 사람 얼굴에 있는 주름을 거의 수염 수준까지 나타낼 수 있도록 개발한 기술도 컴퓨터그래픽스 권위 학술대회인 Eurographics 에서 2020 년에 공개되었으며, 이 기술을 통해 이미지에 있는 사람의 얼굴을 매우 정확하게 3 차원으로 나타낼 수 있다.

4.3. Single-image 3D Face Modeling with Texture

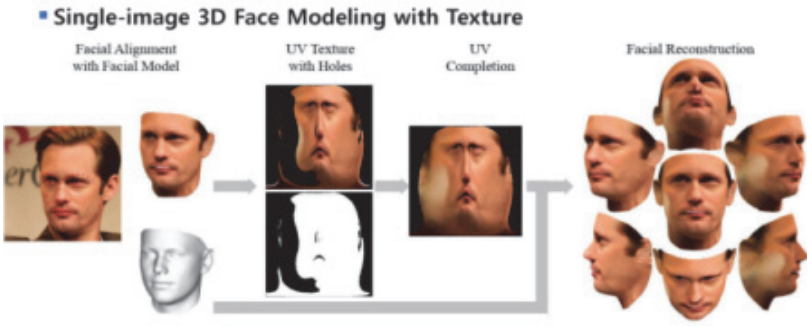

그림 13 UV Completion with Self-referenced Discrimination[16)]

보통 사람의 얼굴 사진은 얼굴이 회전해 있기 때문에 한 장의 사진에서 얼굴 영역이 반 정도밖에 보이지 않는다. 그래서 나머지 반 정도의 영역을 매우 정확하게 예측해서 3 차원의 얼굴을 정확하게 완성하는 기술이 연구되었다. 좌측에 있는 사진에 있는 사람의 얼굴을 우측에 있는 3 차 모델로 완벽하게 복원하는 것을 확인할 수 있다.

4.3.1. Facial Expression and Pose Transfer

그림 14 Competitive Learning of Facial Fitting and Synthesis Using UV Energy

16) Kang et al. "UV Completion with Self-referenced Discrimination," Eurographics (NRF Outstanding Conference) 2020

3DMM (3-Dimensional Morphable Model)은 앞선 기술을 확장하여 3 차원 인공지능 얼굴 모델링으로 확장한 기술이다. 인공지능 기술을 이용해서 한 사람을 3 차원으로 만들고 또 다른 한 사람의 얼굴을 3 차원으로 만들면 두 사람 중 한 사람의 표정을 다른 사람의 표정으로 바꿀 수 있다. 제시된 사진은 중앙에 있는 사람의 표정이 위에 있는 사람의 표정으로 바뀌는 예제이다.

4.4. Light Transfer (Re-lighting)

그림 15 Competitive Learning of Facial Fitting and Synthesis Using UV Energy[17)]

얼굴에 있는 빛의 방향을 바꾸는 예제이다. 자료에서는 얼굴의 빛을 좌측에서 우측으로 조금씩 바꿔주고 있다. 인공지능 기술을 통해서 얼굴을 모델링할 때 정밀하게 예측하기 위해서 얼굴의 색상뿐만 아니라 얼굴의 빛의 방향도 예측하고 있다. 이를 통해 빛의 방향이 잘 예측되었다면 출력하는 3 차원에 나타나는 얼굴의 빛의 방향도 바꿔줄 수 있다. 빛의 방향을 바꾸는 건 쉬워도 자연스럽게 바꾸는 것은 여전히 쉽지 않다. 예제에서는 상대적으로 자연스럽게 나온 사진을 뽑았지만, 현재의 기술로는 모든 사진이 자연스럽게 연출되는 건 아니다.

17) Kang et al. “Competitive Learning of Facial Fitting and Synthesis Using UV Energy,” IEEE Transactions on Systems, Man, and Cybernetics: Systems 2022.

4.5. Real-Time Tracking and Facial Animation

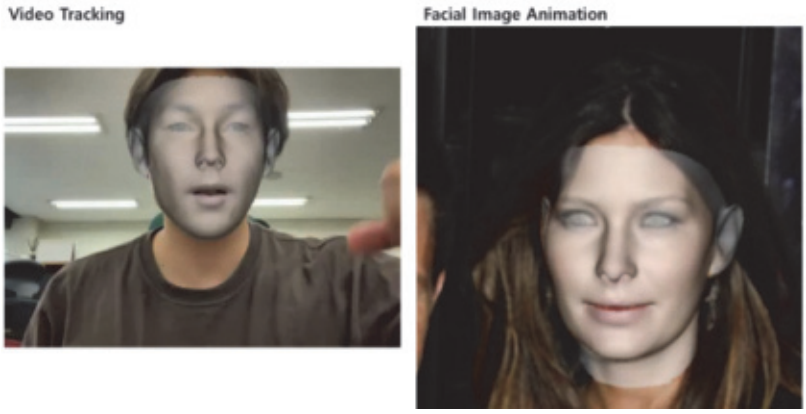

그림 16 Competitive Learning of Facial Fitting and Synthesis Using UV Energy[18)]

얼굴 모델링 기술 통해서 실시간으로 얼굴을 추적하거나 한 장의 사진에 있는 얼굴을 얼굴 모델과 색상 합성 기술을 통해서 움직임이 가능하다.

4.6. Real-time Character Animation

그림 17 Real-time Character Animation

얼굴의 IR 마커를 통해서 얼굴에 매우 미세한 움직임까지 추적하고 추적된 움직임을 통해서 3 차원 캐릭터를 움직일 수 있다. 마커만 붙인다고 추적이 되는 것이 아니라 마커를 정밀하게 3 차원으로 추적할 수 있는 기술이 필요하고, 또한 3 차원 점을 바탕으로 3 차원 캐릭터 표정을 만드는 기술이 사용된다. 작은 IR 마크 하나에 1 만 원 정도 하며, 위 자료의 얼굴에는 110 개 정도 붙어있다. IR 마커 하나가 너무 작아서 자주 분실이 되고 뗐다 붙였다는 게 너무 번거롭다는 단점으로 최근에는 정밀한 추적이 필요할

18) Kang et al. “Competitive Learning of Facial Fitting and Synthesis Using UV Energy,” IEEE Transactions on Systems, Man, and Cybernetics: Systems 2022.

때는 IR 마커가 아니라 펜으로 칠하는 펜 마커가 이용되고 있다. 펜 마커를 이용할 경우 정확도가 조금 떨어지지만 붙이는 IR 마커보다는 편리하게 추적 가능하다.

4.7. 2D Style GAN

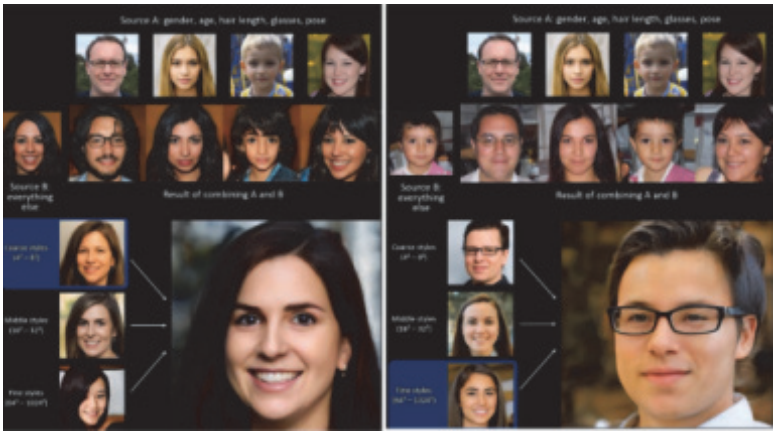

그림 18 A Style-Based Generator Architecture for Generative Adversarial Networks[19)]

2019 년에 CVPR 에서 발표된 2D Style GAN 기술은 3 차원 기술은 아니지만 최근 얼굴 모델링의 한 획을 그은 기술이라고 할 수 있다. 매우 추상적인 수준부터 매우 상세한 수준까지 여러 단계 수준에 걸쳐 얼굴의 스타일을 변경해 가면서 얼굴 사진을 생성할 수 있도록 인공지능 네트워크를 학습시킨 기술이다. 상세 수준이라는 것은 인공지능 네트워크가 학습하면서 정하는 부분이고 연구자는 그것을 관찰하면서 파악하기 때문에 무 자르듯이 정해지지는 않는다. 추상적인 수준은 자세나 머리카락, 얼굴형 정도이고, 중간 수준은 눈, 코, 입 등 얼굴의 요소들, 상세 수준은 색상 같은 요소들이다. 이러한 수준별로 학습시킬 수 있다는 점을 이용하면 얼굴 모양의 추상적인 수준은 A 의 얼굴에서 가져오고 얼굴 모양의 상세 수준은 B 의 얼굴에서 가져와 합쳐서 A 플러스 B 의 얼굴을 만드는 등의 응용을 할 수 있다. 예제에서 볼 수 있듯 왼쪽에 있는 3 개의 스타일이 합쳐져서 하나의 얼굴이 만들어진다.

19) Karras et al. “A Style-Based Generator Architecture for Generative Adversarial Networks,” IEEE Conference on Computer Vision and Pattern Recognition 2019

4.8. View Synthesis from Photo Collection

그림 19 Nerfies: Deformable Neural Radiance Fields[20)]

이 기술은 최근에 'NeRF'라고 불리는 트렌디한 기술이라고 할 수 있다. 의료 분야에서 사용하던 볼륨랜더링 기술을 인공지능 분야로 가져와서 다수의 이미지를 입력해서 3 차원 색상 공간을 예측해서 임의의 카메라 뷰에 대해서도 사진을 생성할 수 있는 기술이다. 예제에서 왼쪽이 네트워크를 학습시킬 때 사용한 영상이고 오른쪽이 인공지능 네트워크에 의해서 만들어진 실제 학습할 때 사용하지 않은 찍지 않은 시선에서의 영상이다. NeRF 기술이 처음 나왔을 때는 많은 양의 사진이 필요하고 한 번 학습시킬 때 며칠씩 걸리는 등 시간이 오래 걸렸는데 기술이 많이 발전해서 NeRF 의 학습 시간이 매우 짧아졌다.

4.9. 3D Face from a Single Image to make Image in Different View

그림 20 Efficient Geometry-aware 3D Generative Adversarial Networks[21)]

최근 NeRF 기술이 발전해 가면서 NeRF 의 응용 사례로 단일 이미지에서 비슷한 형태로 3 차원을 예측을 해서 새로운 뷰에서 이미지를 생성하는 기술이 개발되고 있다. 예제에서 보이듯 실제로 한 장의

20) Park et al. "Nerfies: Deformable Neural Radiance Fields." IEEE Conference on Computer Vision and Pattern Recognition 2021.
21) Chan et al. "Efficient Geometry-aware 3D Generative Adversarial Networks," IEEE Conference on Computer Vision and Pattern Recognition 2022.

이미지를 입력으로 받아서 다른 회전된 어떤 각도에서의 얼굴이 출력될 수 있다. 출력되는 이미지의 결과는 이런 3 차원 정보의 예측을 바탕으로 한다.

4.9.1. Editable StyleGAN

그림 21 SemanticStyleGAN[22)]

2022 년에 발표된 SemanticStyleGAN 기술은 앞선 기술이 확장되어 얼굴을 단순하게 생성할 뿐만이 아니라 얼굴에 요소들이 어디 있는지까지도 예측하는 기술이다. 코나 눈, 눈썹 등의 위치와 모양을 예측하고 이를 임의로 바꾸게 되면 바꾸는 것에 맞춰서 실제로 생성하는 이미지가 변하게 된다. 이러한 인공지능 네트워크를 학습하게 되면 우리가 얼굴 눈 입 머리카락의 모양을 편하게 수정할 수 있다.

4.9.2. Editable 3D-Aware GAN

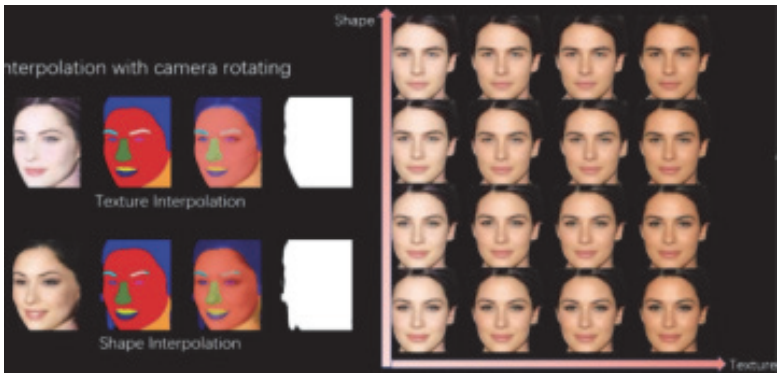

그림 22 Face Editing in Neural Radiance Fields[23)]

Editable 3D-Aware GAN 기술은 앞에서 소개했던 한 장의 사진으로부터 얼굴의 요소를 구하는 기술과 3 차원을 예측하는 기술 두 가지가 합쳐져서 하나의 기술로 개발된 것이다. 이렇게 얼굴의 요소가

22) She et al. "SemanticStyleGAN: Learning Compositional Generative Priors for Controllable Image Synthesis and Editing," IEEE Conference on Computer Vision and Pattern Recognition 2022

23) Sun et al. "FENeRF: Face Editing in Neural Radiance Fields," IEEE Conference on Computer Vision and Pattern Recognition 2022

같이 구해지기 때문에 눈, 코의 위치 등 임의로 설정한 부분을 편집해주면 인공지능 네트워크가 해당 부분에 맞게 얼굴을 생성해준다.

4.9.3. 3D Face from a Phone Scan

그림 23 Learning an Animatable Detailed 3D Face Model[24)]

NeRF 기술의 개념을 이용한 기술로, 핸드폰으로 얼굴을 둘러서 찍으면 3 차원으로 모델링되고, 임의에 의해 재구성할 수 있는 기술이 컴퓨터 그래픽스 권위 학술지 ACM Transactions on Graphics 에 2021 년에 출판되었다. 이 기술을 통해 재구성된 얼굴은 3 차원으로 모델링 되었기 때문에 자유롭게 표정이나 얼굴의 모양을 바꿀 수 있다.

5. Scene

5.1. Single RGB to 3D Scene

그림 24 Panoptic 3D Scene Reconstruction[25)]

24) Feng et al. "Learning an Animatable Detailed 3D Face Model from in-the-wild Images," ACM Transactions on Graphics 2021.
25) Dahnert, Manuel, et al. "Panoptic 3D Scene Reconstruction from a Single RGB Image." Advances in Neural Information Processing Systems 2021.

어떤 방 사진을 입력으로 받았을 때 그 방에 있는 어떤 개체를 3 차원으로 모델링하는 기술이 2021 년에 세계 인공지능 권위 학술대회를 NeurIPS (Advanced in Neural Information Processing Systems)를 통해 공개되었다. 예제에서 보이듯 단순히 방 한 장의 이미지를 받았는데 그럴듯하게 3 차원으로 만들어주는 것을 확인할 수 있다. 이런 결과를 보면 아주 상세하지는 않지만, 인공지능 네트워크가 3 차원을 어느 정도 제대로 구성하는 것이 가능하다고 생각할 수 있지만 실제로는 이 기술을 구현하기 위해 여러 데이터베이스를 바탕으로 한다.

5.2. Indoor Scene Dataset

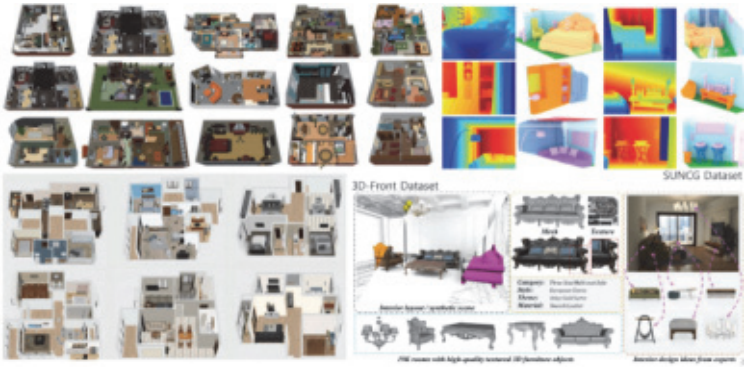

그림 25 Indoor Scene Dataset

위 데이터베이스는 CAD (Computer-Aided Design)를 이용해서 사람이 직접 손으로 실내 공간에 있는 가구들을 정밀하게 만든 데이터셋이다. 이 데이터셋을 인공지능에 학습시켜서 앞서 소개한 방의 3 차원 모델링 기술을 예측하는 데 사용하고 평가하는 데 사용했기 때문에 실제 이 데이터셋에 없는 가구나 많이 벗어난 가구의 경우에는 예측할 수가 없다. 그래서 실제로 아예 관측하지 못한 어떤 개체의 경우는 인공지능 네트워크가 여전히 예측하는 데 한계가 있다고 할 수가 있다.

5.3. RGB Sequence to 3D Scene

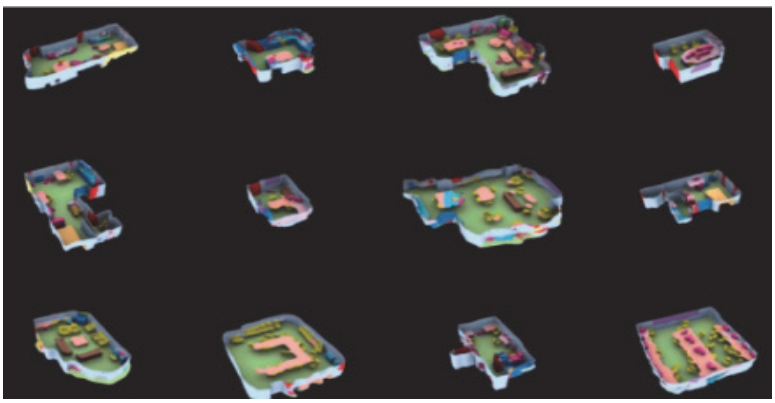

그림 26 Atlas: End-to-end 3D Scene Reconstruction from Posed Images[26)]

26) Murez et al. "Atlas: End-to-end 3D Scene Reconstruction from Posed Images." European Conference on Computer Vision 2020.

2020 년에 ECCV 에서 공개된 Atlas 기술은 이미지를 찍으면 일종의 지도를 만들어주는 기술로 공간에 대한 어떤 3 차원 공간 지도를 만들어준다고 볼 수 있다. 어떤 공간의 동영상을 찍어 나가면 그 공간에 대해서 3 차원으로 만들어주고 그 공간에 있는 객체가 무엇인지 분류해주는 기술이다. 같은 공간을 여러 번 찍어 나갈수록 공간이 3 차원으로 재구성되면서 그 공간에 있는 개체가 의자인지 벽인지 등을 분류해줄 수 있다.

5.4. View Synthesis from Photo Collection

그림 27 NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections[27)]

이어 2021 년에 CVPR 에서는 앞에서 소개한 NeRF 기술이 대량 사진에서 쓰여서 어떤 큰 공간을 3 차원으로 만드는 데 사용된 기술이 공개되었다. 이 NeRF 기술이 대량의 사진에 사용되었을 경우에는 거대한 풍경 또는 공간에 대해서 임의의 뷰의 사진을 생성하는 데 이용될 수 있다. 예제에서 보이는 뷰는 인공지능 네트워크에 의해서 생성된 뷰이고 실제로 학습에 사용된 사진들이 아니며, 실제로는 다양한 각도의 임의의 사진들이 제공되어 학습되었다.

27) Martin-Brualla et al. "NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections." IEEE Conference on Computer Vision and Pattern Recognition 2021.

6. Virtual Garment (Cloth)

6.1. Virtual Garment Modeling

그림 28 Self-Supervised Neural Dynamic Garments[28)]

2022 년도에 CVPR 에서 공개된 SNUG (Self-Supervised Neural Dynamic Garments) 기술은 3 차원 옷 가상 데이터를 다양하게 생성하고 그 데이터를 바탕으로 모델링해서 인공지능의 학습을 시킨 기술이다. 그래서 임의의 옷을 인공지능 네트워크가 3 차원으로 표현 가능하게 만들었는데, 수학적 용어로는 파라미터화 시켰다고 표현한다. 예제의 사진에서는 아래 막대를 조정하여 8 개의 숫자로 옷이 표현될 수 있도록 인공지능 네트워크가 학습되었다. 그래서 이 숫자를 바꾸면 옷의 소매가 길어지기도 하고 짧아지기도 하는 등 다양하게 변하게 된다. 현재로서는 몇 가지 옷 분류에 대해서 가능성을 보여주는 단계 정도로 연구되었고, 이러한 파라미터화가 현실에 있음 직한 옷에 대해서 일어날 수 있는 대부분의 경우에 대한 변화를 나타낼 수 있게 된다면 내가 입은 옷과 똑같은 3 차원 옷을 만드는 문제는 내가 입은 옷과 가장 비슷한 3 차원 옷 모델을 만드는 8 가지 숫자를 찾는 문제로 바뀔 수 있다. 이런 파라미터화의 가장 큰 장점은 내가 입은 옷과 가장 비슷한 옷 모델을 만드는 8 가지 숫자를 찾았을 때 숫자를 조금 조절해서 옷소매를 짧게 만든다거나 총 기장을 늘린다거나 하는 변화가 쉽게 가능해진다는 점이다. 이 파라미터와 인공지능 기술이 2 차원에서 적용된 대표적인 예로 나와 똑같은 얼굴 모델을 찾아서 나이가 더 들어 보이게 한다거나 나이가 더 어려 보이는 사진으로 바꿔주는 애플리케이션이 있다.

28) Santesteban et al. "SNUG: Self-Supervised Neural Dynamic Garments," IEEE Conference on Computer Vision and Pattern Recognition 2022.

6.2. Virtual Garment Modeling

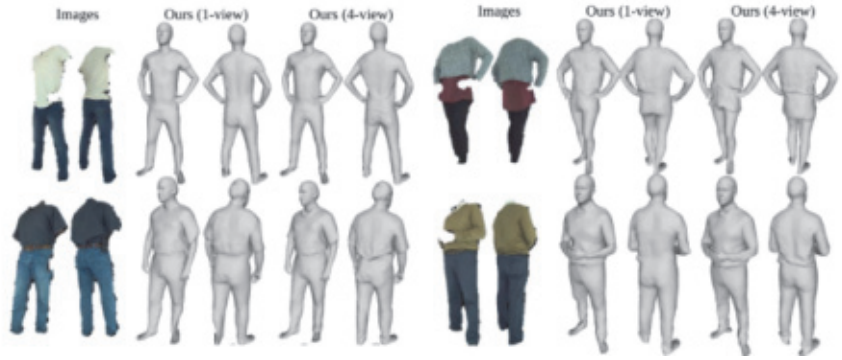

그림 29 Double Reverse Diffusion for Realistic Garment Reconstruction from Images[29)]

영상처리 분야 세계 권위 학술지 IEEE Transactions on Circuits and Systems for Video Technology 에 심사중인 Double Reverse Diffusion for Realistic Garment Reconstruction from Images 의 이름의 연구는 단일 이미지에서 옷을 예측하는 연구이다. 원래는 시간에 따른 물체의 변화를 예측하는 데 사용하는 인공지능 기술을 옷이 점점 주름이 져가는 것을 예측하는 데 반복적으로 적용하였다. 현재 추가 기술을 개발해서 실시간 비디오 영상의 옷을 3 차원으로 예측하는 방향으로 기술을 확장하고 있다.

7. Hand

7.1. Video to 3D Hand

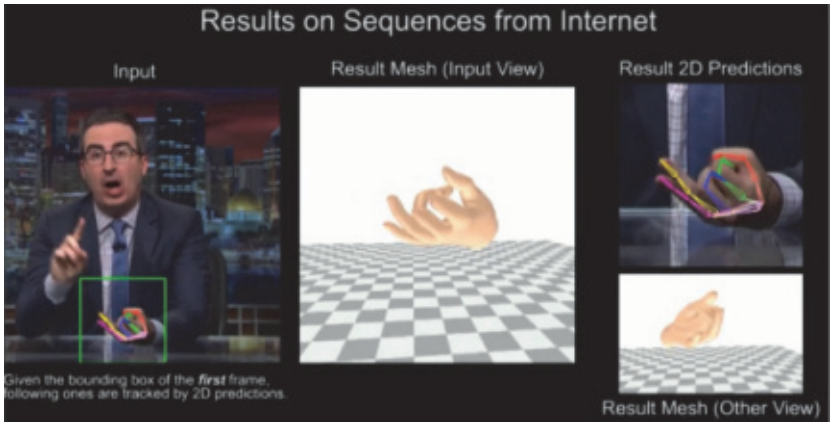

그림 30 Monocular Real-time Hand Shape and Motion Capture using Multi-modal Data[30)]

29) Nguyen et al. "Double Reverse Diffusion for Realistic Garment Reconstruction from Images," IEEE Transactions on Circuits and Systems for Video Technology, under review.

30) Zhou et al. "Monocular Real-time Hand Shape and Motion Capture using Multi-modal Data," IEEE Conference on Computer Vision and Pattern Recognition 2020.

단일 비디오 영상으로부터 손의 움직임을 정확하게 예측하는 기술이 2020 년 국제학술대회 CVPR 에서 발표되었다. 그럼에도 현재 기술로서는 정면 영상이 아닌 이상 매우 정확하게 예측하기는 힘들다.

7.2. Video to Two Hand

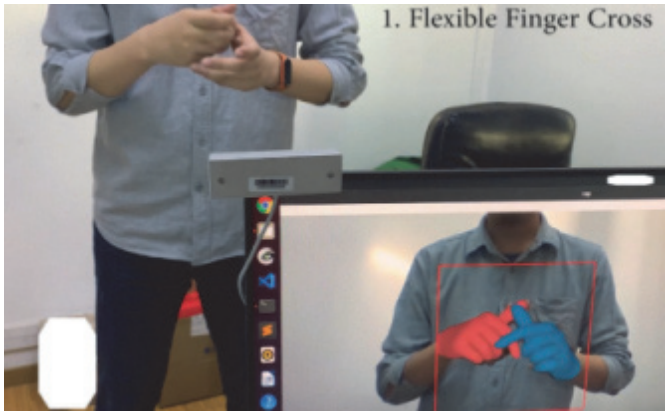

그림 31 Monocular Real-time Hand Shape and Motion Capture using Multi-modal Data[31)]

위 기술의 연장선으로 한 개의 손이 아니라 두 개의 손이 상호작용을 했을 때를 예측하는 연구도 진행되었다.

7.3. Hand and Object from an Image

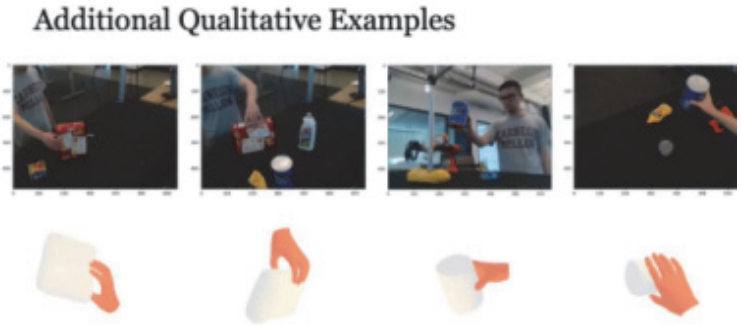

그림 32 Collaborative Learning for Hand and Object Reconstruction with Graph Convolution[32)]

2022 년 CVPR 에서는 손이 물체와 상호작용을 했을 때 물체와 손을 같이 3 차원으로 만들어주는 GCN (Graph Convolution Network)을 이용한 Collaborative Learning for Hand and Object Reconstruction 기술이 발표되었다.

---

31) Zhou et al. “Monocular Real-time Hand Shape and Motion Capture using Multi-modal Data,” IEEE Conference on Computer Vision and Pattern Recognition 2020.

32) Tse et al. “Collaborative Learning for Hand and Object Reconstruction with Graph Convolution,” IEEE Conference on Computer Vision and Pattern Recognition 2022.

7.4. Upper Body Poses (6 Pose Rots) to Hand Pose

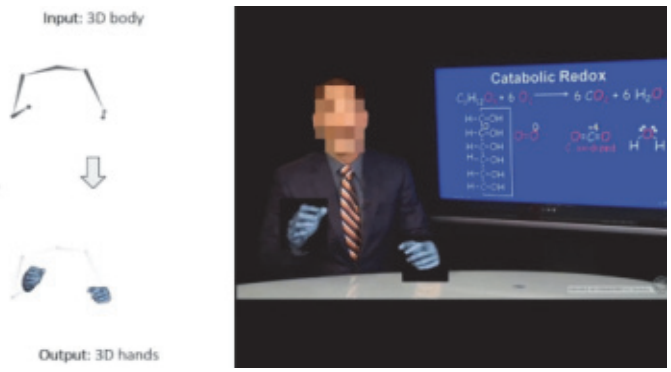

그림 33 Body2Hands: Learning to Infer 3D Hands from Conversational Gesture Body Dynamics[33)]

2022 년에 CVPR 에서 공개된 Body2Hands 기술은 사람의 손 모양을 예측할 때 손 영상 정보를 사용하는 게 아니라 사람의 어깨에서 손까지의 관절 정보를 이용해서 손이 어떤 모양을 취하는지를 예측한 기술이다. 예제의 왼쪽 사진에서 위쪽이 입력이고 아래쪽에 있는 손이 결과이다. 사람의 어깨 관절 정보만 이용해서 사람의 손 모양을 예측하면 자연스럽게 연출되는 것을 확인할 수 있다. 이 기술은 실제로 사람이 취한 손 모양과 얼마나 같은가를 확인하는 게 아니고 얼마나 손이 자연스럽게 연출이 되는가를 확인하는 것이다.

8. Modeling from Sound (Voice/Speech/Music)

8.1. Speech to 3D Gesture

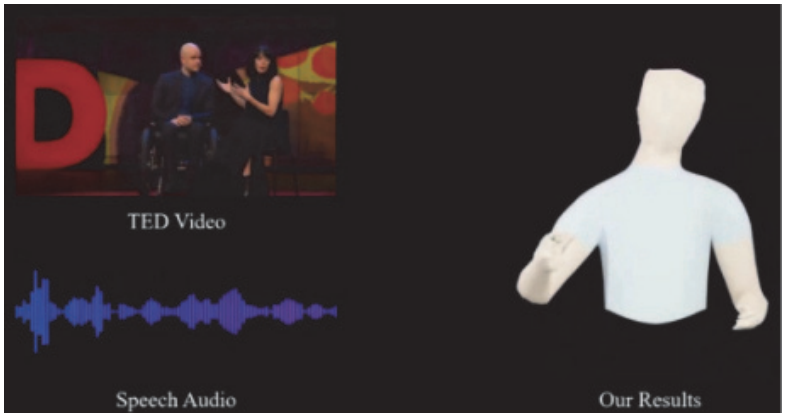

그림 34 Learning Hierarchical Cross-Modal Association for Co-Speech Gesture Generation[34)]

33) Ng et al. "Body2Hands: Learning to Infer 3D Hands from Conversational Gesture Body Dynamics." IEEE Conference on Computer Vision and Pattern Recognition 2022.

34) Liu et al. "Learning Hierarchical Cross-Modal Association for Co-Speech Gesture Generation," IEEE Conference on Computer Vision and Pattern Recognition 2022.

2022 년 컴퓨터과학 권위 학술대회 CVPR 에서 어떤 소리가 출력이 되면 그 소리에 맞는 자연스러운 제스처를 만들어주는 연구가 공개되었다. 위 예제는 영상에서 출력한 대화 소리에 따라서 3D 캐릭터로 어떤 움직임을 만들어주는 영상이다. 예제에서는 제스처를 하는데 계속 어깨만 들썩거리는 결과가 나오는데, 인공지능에 TED 영상을 기본으로 해서 트레이닝을 시켰기 때문에 데이터 의존적인 결과가 나오는 것을 확인할 수 있다.

8.2. Music to Dancing 3D Avatar

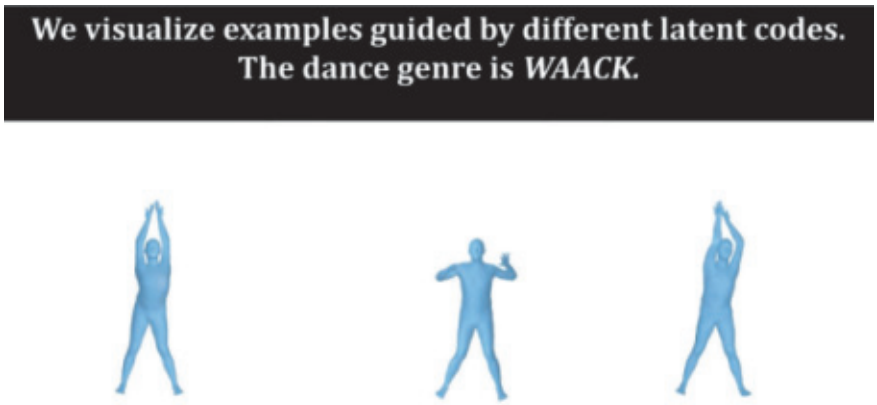

그림 35 Music-Conditioned Pluralistic Dancing Controlled by Multiple Dance Genres[35)]

위 연구와 비슷한 예로 음악에 따라서 춤을 생성해주는 인공지능 네트워크이다.

9. 결론

인공지능과 컴퓨터 비전, 그래픽스 기술을 통해서 할 수 있는 연구에 대해서 알아보았다. 특히 (1) 사람의 몸 모델링, (2) 옷을 포함한 전신 모델링, (3) 사람의 얼굴 모델링, (4) 개체 및 배경 모델링, (5) 의류 모델링, (6) 손 모델링, 그리고 (7) 음성과 융합한 3 차원 생성 및 모델링 영역으로 분류하여 나누어 보았다. 이러한 인공지능, 비전 그래픽스 기술은 다양한 예술, 사회 영역의 응용에서 사용될 수 있으며, 융합 및 확대될 수 있다. 이 논문이 최신 인공지능 기술의 동향을 파악하고, 현재 인공지능 및 비전 그래픽스 기술로 할 수 있는 다양한 응용으로 나아가는데 도움이 되기를 기원한다.

35) Kim. et al. “A Brand New Dance Partner: Music-Conditioned Pluralistic Dancing Controlled by Multiple Dance Genres,” IEEE Conference on Computer Vision and Pattern Recognition 2022.