2D Video 손동작 인식 및 얼굴 감정 인식을 통한 가상 인형 제어

문주은○ 고강연 윤희찬 한예원 이승규
경희대학교 컴퓨터공학과
cindy4741@khu.ac.kr, rhrkddus@khu.ac.kr, harryyoon7@khu.ac.kr,
hnyw00or@khu.ac.kr, seungkyu@khu.ac.kr

Virtual Puppet Control using 2D video Hand Tracking and Facial Emotion Recognition

Jueun Mun ○ Gangyeon Go Heechan Yoon Yewon Han Seungkyu Lee
School of Computing, KyungHee University

요 약

가상환경의 3D 데이터 컨트롤은 2D 기반의 기존 인터페이스보다는 공간상의 손의 움직임이나 표정 등 복합적인 Human-computer interaction을 활용하는것이 필수적이다. 이때, 다양한 환경에서 사용할 수 있도록 간단한 센서만을 이용하여 범용성을 높이는 것 또한 중요하다. 따라서 본 논문은 Virtual puppet이란 가상 환경에서 구현된 인형으로 2개의 RGB camera를 사용해서 가상의 인형을 조종한다. RGB camera을 이용해 손의 움직임을 추적하여 인형이 손가락의 움직임에 맞게 움직일 수 있도록 한다. 또한, 특정 손동작을 인식해서 정해진 행동을 취할 수 있도록 한다. 추가적으로, 다른 RGB camera에서는 사람의 얼굴을 촬영한다. 캡쳐한 사진으로 사람의 표정을 인식하고 이를 인형의 표정에 대입하여 다양한 표정을 구사할 수 있도록 한다.

1. 서 론

보통 가상환경에서 인형을 조종하기 위해서는 3D motion capture 또는 Remote manipulator와 같은 Motion control을 위한 특수한 센서를 사용했다. 하지만 이 장비들은 가격 등의 측면에서 진입장벽이 높아 일반 사용자들이 사용하기에는 많은 어려움이 있다. 본 논문에서는 이러한 문제를 해결하기 위해서 사용자들이 좀 더 쉽게 다가갈 수 있는 장비인 2대의 RGB camera를 사용하여 실감 나게 인형을 조종한다. RGB camera로 손의 움직임을 추적하여 인형이 손가락의 변화에 맞게 움직일 수 있도록 하고, 특정 손동작을 인식해서 정해진 행동을 취할 수 있도록 하여 가상으로 구현한 인형을 조종한다. 또한 다른 RGB camera에서는 사람의 얼굴을 촬영한다. 촬영한 얼굴을 CNN(Convolution Neural Network)에 넣어 사용자의 표정을 인식해서 기쁨, 슬픔 같은 감정들을 추출하고 이에 맞게 인형의 표정에 적용한다.

본 논문에서 구현한 Virtual puppet에 사용된 Hand Skeleton Tracking, Hand Gesture Estimation, Emotion Recognition은 RGB camera로만 구현되어 컴퓨터, 스마트폰, 스마트 tv 등 웹캠과 같은 카메라를 이용할 수 있는 다양한 전자기기에서 사용할 수 있도록 범용성을 높였다. 또한 Virtual puppet 뿐 아니라 어린이 fun용, 메타버스 아바타 컨트롤, 전문가용 3d 객체 컨트롤과 같은 다양한 분야에 적용할 수 있다.

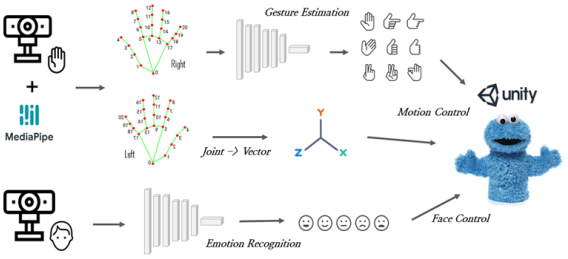

그림 1. 전체 네트워크 구조

특히, 현재는 3D 형태의 데이터를 마우스를 이용하여 2D 평면상에서만 제어할 수 있지만 본 논문에서 구현한 인식 및 트래킹 기능을 이용하면 건축 전문가들이 전체적인 설계도의 3D 단면을 직접 손을 이용해 돌려보며 확인하거나 의사들이 복잡한 종류의 수술을 위해서 3D MRI 혹은, 엑스레이 촬영 자료를 이용할 때 3차원으로 폭넓게 확인하는 등 다양한 분야에 큰 기여를 할 수 있다.[1]

2. 관련 연구

[1] "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학 사업의 연구결과로 수행되었음" (2017-0-00093)

2.1 Hand tracking and Gesture recognition

핸드 트래킹(Tracking)은 이미지 상에서 손의 연속적인 위치를 추적하여 손 모양의 2D 또는 3D 골격(Skeleton)을 계산하는 것을 말한다. 현재 다양한 종류의 핸드 트래킹(Tracking) 기술[3, 4,5]들이 사용되고 있으며 AR/VR 게임, 터치 리스 스크린 등 활용 분야가 계속해서 확대되고 있다. 적외선 카메라를 이용하는 방법부터, 일반 RGB 카메라를 이용하는 방법 등 핸드 트래킹(Tracking)은 다양한 방식으로 시도되고 있다. 핸드 트래킹은 실시간성이 강하기 때문에, 잡음으로 인해 발생하는 혼선 해결과 순간적인 움직임의 포착은 주요한 연구과제중 하나이다. 또한, Gesture Recognition 기술은 2D, 혹은 3D 이미지나 각 손가락의 Skeleton을 이용하여 손의 동작을 파악하는 것을 말한다. 주로 Skeleton의 순서에서 얻을 수 있는 정보나 2D 또는 3D 컨볼루션 네트워크를 통하여 얻는 정보 등을 활용하여 제스처를 판단한다.

2.2 Emotion Recognition

감정 인식(Emotion Recognition)은 카메라에서 비치는 사용자의 얼굴 데이터를 통해서 감정을 인식하는 기술[2]이다. 표정만이 아니라 목소리 등의 매체를 통해서도 판단되기도 한다. 사람의 감정을 표정을 통해서 명확하게 판단하는 이론은 아직 확립되어 있지 않지만 보통은 다양한 표본에서 얻은 자료로 각각의 감정마다 일반적인 표정을 알아내는데 중점을 두고 있다. 이를 위해서 주로 기계 학습(Machine Learning)을 통해 얻어진 모델[1,5]이 쓰이고 있으며 주로 SVM, RF, KNN 등의 알고리즘이 현재 사용되고 있고, 서로 다른 액션 유닛(Action unit)을 기반으로 얼굴의 감정과 강도를 측정하는 방식 역시 사용되고 있다.

3. 시나리오 제안 및 구현

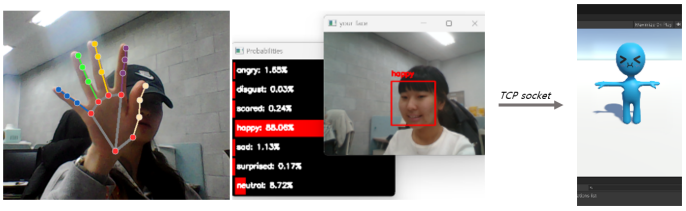

그림 2. 프로그램 실행 화면

3.1 캐릭터 모션 컨트롤

캐릭터를 컨트롤 하고 있는 사용자의 왼쪽 손을 RGB 카메라를 통해 촬영하고 이때 획득한 이미지를 Mediapipe hand tracking 모델에 입력하여 Hand skeleton 정보를 결과 값으로 얻는다. 획득한 정보를 바탕으로 다음 두가지 방식으로 캐릭터를 컨트롤한다. 첫째, 엄지, 검지, 중지 손가락의 Skeleton 정보를 중점적으로 사용하여 캐릭터의 신체 일부분을 조종한다. 먼저 캐릭터의 오른쪽 팔, 캐릭터의 상반신, 캐릭터의 왼쪽 팔 부분에 엄지, 검지 그리고 중지를 각각 할당한다. 그 다음 각 손가락 Joint 값을 이용하여 해당 손가락의 방향 벡터를 구하고, 이 벡터값을 이용하여 캐릭터를 조종한다.

왼손이 캐릭터 신체 부분을 컨트롤 한다면, 오른손은 정해진 제스처를 취해서 게임 캐릭터가 해당 제스처에 해당하는 모션을 행동하도록 한다. 우선 카메라를 통해 사용자의 오른손 이미지를 받아오고, 이를 Hand tracking 모델에 넘겨준다. 그러면 Hand tracking 모델이 오른손의 Skeleton을 추출하고 추출된 Skeleton의 Joint 부분을 Pre-trained된 인공신경망에 입력 데이터로 넣어주면 인공신경망은 각 마디의 각도를 이용하여 현재 어떤 제스처를 취하고 있는지 판단한다. 현재 모델에 학습시킨 제스처의 종류는 총 6가지로, 학습을 하는 데에 사용된 데이터의 개수는 표 1과 같다. 이때 제스처들 중에서 손을 모두 피는 제스처와 손을 오므린 제스처는 캐릭터 조종의 시작과 종료를 시스템에게 알려주는 역할을 한다. 종료를 하게 되면 이전에 저장되어 있는 데이터들은 모두 초기화하여 처음부터 다시 시작할 수 있도록 구현하였다. 또한 제스처를 이용하여 캐릭터를 앞뒤, 좌우로 이동할 수 있게 구현하였다. 이후 모델이 판단한 제스처를 소켓을 통해 유니티로 전송하여 사용자가 특정 제스처를 취하고 있다고 알리면 유니티는 해당 제스처에 해당하는 모션을 찾고, 캐릭터가 이 모션을 취할 수 있도록 한다.

종류	Hand Open	Hand Close	Point	Thumbs Up	Check	Peace
개수	1522	1572	1992	1288	2911	1841

표 1. 제스처 학습 데이터 개수

3.2 캐릭터 표정 컨트롤

사용자의 표정과 캐릭터의 표정을 매칭시켜 캐릭터의 표정을 컨트롤한다. 웹캠을 이용해 실시간으로 찍고 있는 얼굴 영상을 캡쳐하여 얼굴 부분의 Image를 검출한다. 이를 Pre-train된 CNN 모델[6]에 넣어 사용자의 현재 표정 정보에 대한 결과값을 리턴받는다. 이 결과값을 유니티에 전송해 미리 구현한 Mesh texture 데이터를 이용해 해당 표정과 일치하는 캐릭터의 표정으로 바꿔준다. 영상 싱크에 맞게 실시간으로 결과를 출력해야 하므로 연산량을 줄이기 위해 매개변수를 줄여 경량화한 CNN 모델을 사용한다. 이 모델은 Facial Expression Recognition 2013 Dataset으로 Pre-train되어 있는데 해당 데이터셋의 표정 모델은 대부분 서양인이라 한국인과 인종이 달라 실제로 사용했을 때 민감도가 떨어진다는 단점이 있다. 따라서 Korean Drama Multi-Label Facial Emotion Recognition Dataset[7] 및 사용 환경에 맞는 데이터로 Fine-tuning을 진행했다.

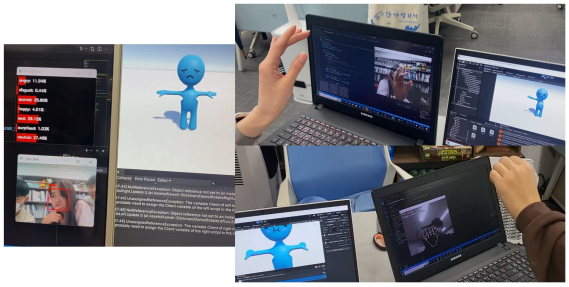

그림 3. 시연 화면, 왼쪽부터 시계 방향으로
Emotion Recognition, Hand Skeleton Tracking,
Hand Gesture Estimation

3.3 안정화를 위한 데이터 전처리

RGB 카메라로 손을 트래킹하는 과정에서 트래킹에 실패하거나 추출된 손의 Skeleton 정보가 비정상적으로 나올 가능성이 있다. 이러한 경우 사용자와 캐릭터 간의 인터랙션을 방해할 수 있으므로, 유니티 상으로 값이 넘어가기 전에 이러한 값들을 전처리해주는 것이 중요하다. 이를 위하여 각 프레임에서 얻은 값을 전후 프레임과 비교하여 현재 값이 적절한 값인지를 판단하고, 만약 적절하지 않다면 이를 조절할 수 있도록 다음과 같이 구현하였다.

3.3.1 트래킹 실패

Frame상에서 손의 Skeleton 값이 구해지지 않는 경우는 두가지 존재하는데, 첫번째로는 실제로 손이 없는 경우, 두번째로는 일시적으로 트래킹에 실패하는 경우이다. 전자의 경우에는 프레임 상에서 손의 Skeleton 정보를 얻지 못하는 것이 정상적이지만, 후자의 경우에는 손의 Skeleton 정보가 구해져야만 한다. 따라서 먼저 손이 카메라 상에서 보이지 않는 경우인지 아니면 실제로 트래킹에 실패한 경우인지 판단해야 한다. 이를 판단하는 방법으로 우선 각 프레임에서 얻은 손 Skeleton 정보를 Queue에 넣고, Queue의 크기를 일정하게 유지하기 위해서 순차적으로 들어온 Skeleton 정보를 차례로 Queue에서 내보낸다. 매번 Queue에서 뽑은 데이터들은 유니티로 값을 전송한 후에 이를 이전 Frame의 데이터 값을 저장해 놓는 곳에 업데이트 하며 값을 저장한다. 이때 만약 Queue에서 뽑아낸 Skeleton 정보에 아무런 값이 없다면, Queue에 있는 모든 Skeleton 데이터 정보들을 확인한다. 만일 Queue에 있는 모든 Skeleton 값 역시 존재하지 않는다면 이는 손이 카메라에서 사라졌다는 것을 의미하지만, Queue에 있는 skeleton 정보들이 값을 가지고 있다면 이는 일시적인 트래킹 실패를 의미하게 된다. 따라서 이러한 경우에는 미리 저장해 두었던 이전 Frame의 데이터 값을 불러와 이 데이터와 동일하게 Skeleton 값을 저장하고, 이를 유니티 상으로 전송해준다. 마지막으로 현재 프레임의 값을 이전 Frame 데이터 값으로 업데이트 한다.

3.3.2 Outlier 데이터 조정

Hand tracking 모델이 항상 신뢰가능한 데이터를 도출해내지 않기 때문에, Outlier한 데이터를 판별하고 이를 적절한 데이터 값으로 변경해 주어야 한다. 우선 일차적으로 모델을 통해 얻은 데이터값의 신뢰도가 50% 미만이면 해당 값은 사용하지 않고, 이전 Frame의 데이터값을 넘겨주게 된다. 신뢰도가 50%를 넘겼지만, 여전히 값이 Outlier인 경우에는, 이를 해결하기 위하여 위와 동일하게 Queue와 이전 Frame의 Skeleton값을 이용한다. 만약 서로 인접하는 프레임(이전 Frame과 이후 Frame)과 현재 Frame의 Skeleton 손가락 벡터 차이가 모두 역치를 벗어나게 된다면, 이는 적절하지 않은 데이터라고 판단하고, 해당 데이터 값을 이전 Frame과 이전 Frame의 Skeleton 데이터 값의 평균으로 설정한다.

3.3.3 Smoothing

실험을 진행하는 과정에서, 손이 움직이지 않는데에도 불구하고 데이터의 값이 일정하게 나오지 않는 것을 확인할 수 있었다. 이에 따라 데이터의 값이 조금만 바뀌는 경우라도 캐릭터의 팔이 과도하게 떨리거나 움직이는 상황이 발생한다. 따라서 이러한 현상을 줄이기 위하여 Kalman filter를 적용하여 이전데이터들을 통해 예측 값을 구하고, 실제 값과 이를 비교하여 Smoothing한 Joint값을 얻고자 하였다. 이때 예측 값은 이전 값에서 지금까지 한 프레임 당 각 Joint의 변화 정도를 계산하여 구하였고, 분산의 경우 모델을 통해 얻은 데이터값의 신뢰도를 이용하여 분산을 정하였다. 이를 통하여 Smoothing하기 전보다 더욱 안정된 캐릭터의 움직임을 확인하였다.

4. 결론 및 향후 계획

본 논문에서는 사용자의 손의 움직임을 통해 자연스럽게 움직이고 사용자의 표정 변화를 따라하는 가상 인형을 만들었다.[8] 기존의 트래킹 모델과 인식 모델이 가지고 있는 잡음으로 인한 부정확함을 줄이고 순간적인 변화에 보다 더 잘 대처하는 모델을 만들었다. 또한, 특수한 장비를 없이 카메라만을 통한 컨트롤 방식이 더 발전하게 된다면, 다양한 분야의 3D data를 2D 상이 아닌 3D 상에서의 컨트롤을 적용할 수 있다.

5. 참고 문헌

[1] Ko, B.C., 2018. A brief review of facial emotion recognition based on visual information, sensors, 18, pp. 401, 2018

[2] Liyu Meng, Yuchen Liu, Xiaolong Liu, Zhaopei Huang, Yuan Cheng, Meng Wang, Chuanhe Liu, Qin Jin : Multi-modal Emotion Estimation for in-the-wild Videos arXiv preprint arXiv:2203.13032, 2022.

[3] Erol, Ali, George Bebis, Mircea Nicolescu, Richard D. Boyle, and Xander Twombly. Vision-based hand pose estimation: A review. Computer Vision and Image Understanding 108, no. 1-2 , pp. 52-73, 2007.

[4] Moon, Gyeongsik, Shoou-I. Yu, He Wen, Takaaki

Shiratori, and Kyoung Mu Lee, Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image, In European Conference on Computer Vision, pp. 548-564, 2020.
[5] Zhang F, Bazarevsky V, Vakunov A, Tkachenka A, Sung G, Chang CL, Grundmann M, Mediapipe hands: On-device real-time hand tracking, arXiv preprint arXiv:2006.10214. 2020.
[6] Octavio Arriaga, Matias Valdenegro-Toro, and Paul Plöger. Real-time convolutional neural networks for emotion and gender classification. arXiv preprint arXiv:1710.07557, 2017.
[7] H. Cho, W. K. Kang, Y. -S. Park, S. G. Chae and S. -j. Kim Multi-Label Facial Emotion Recognition Using Korean Drama Video Clips, 2022 IEEE International Conference on Big Data and Smart Computing (BigComp), 2022.
[8]https://drive.google.com/file/d/1CTpG_56LoOdsIRMHh4YNsuSvyFgYLkqv/view?usp=sharing