손의 관절 위치 검출 및 모델 경량화를 통한 3D 시각화 시뮬레이션

이창화*, 이선경*, 김동욱*, 이승은*, 차준욱*, 이한솔*, 조윤성*, 백승렬*
울산과학기술원*

3D simulation via hand pose estimation and model simplification

Changhwa Lee*, Seongyeong Lee* , Donguk Kim*, Seungeun Lee*, Junuk Cha*, Hansol Lee*, Yunseong Cho*, Seungryul Baek*
Ulsan National Institute of Science and Technology*

Abstract - 이미지 혹은 영상의 프레임으로부터 사람의 손 위치를 검출하고 포즈를 추정해보며, 얻어진 정보를 활용해 모바일 기반의 AR 글래스 디바이스에서 시각적으로 시뮬레이션한다. 기존의 손 위치 및 관절 검출 모델을 경량화해 가상 현실 및 실시간 활용이 가능해지도록 모바일 애플리케이션으로의 개발 및 확장을 목표로 한다. 딥러닝 네트워크는 다양한 합성 및 실제 RGB 데이터를 활용해 손의 위치와 관절을 정확한 추론하는 지도 학습을 했으며, 실제 경제성이 높은 디바이스에 직접 실험해 봄으로써 결과를 시각화할 수 있었다. 이 논문에서 제안한 모델은 기존의 손 관절 추정 알고리즘와 대비해 속도와 성능의 열화성 측면에서 정량적으로 비교했다. 이러한 위치 정보를 활용한 실시간 경량화 딥러닝 모델의 설계는 효율성과 정확도를 높일 수 있고, 높은 이식성과 확장성으로 AR/VR 등의 다양한 분야에 활용이 가능하다.

1. 소 개

단일 RGB 이미지, 혹은 비디오의 프레임들로부터 2D 또는 3D 손 관절을 추론하는 연구는 과거부터 중요했다. 딥러닝 기반의 손 관절 추론을 위한 다양한 연구는 이미 많이 이루어진 만큼, 실시간으로 손 관절 및 제스처를 추론하는 연구는 AR/VR과 같은 사람-컴퓨터 상호작용(HCI) 애플리케이션에서 활용할 가능성이 크다. 딥러닝을 활용해 다양한 손 관절 정보를 추론해 이의 정확성을 높이는 연구들이 진행되는 만큼, 기존의 다양하게 연구된 모델로는 모바일 및 엣지 디바이스에 알맞게 경량화하기 어렵다. 따라서 실시간으로 딥러닝을 활용해 손의 위치를 빠르게 검출하고 관절의 추론이 요구된다. 이때, 손 RGB와 해당하는 2D 또는 3D 정답이 함께 있는 데이터셋을 활용한 손의 위치를 검출하고 관절을 추론하는 딥러닝 네트워크를 학습할 수 있다.

손에 관한 연구는 크게 손 위치 검출, 관절의 추론, 3D 메쉬 재건, 물체와의 상호작용 등으로 나눌 수 있다. 특히 모바일에 적합한 모델을 구성하기에는 이러한 연구적 문제를 모두 다루기 어려움이 존재한다. 이 논문에서는 손 연구의 가장 기본적인 손 위치 검출 및 관절을 추론하는 과정을 실제 모바일 기반 AR 장비를 이용해 시각적으로 시뮬레이션한다.

모바일용 딥러닝 네트워크에서 가장 중요한 부분은 모델의 경량성이다. 이때 정량적으로 평가하는 부분은 모델 파라미터의 개수 및 모델 추론에 소요되는 시간이다. 모델의 파라미터 개수가 적을수록 모델 추론에 소비되는 시간 또한 적어지며 평균 속도가 증가해, 실사용자 입장에서 즉각적인 응답이 가능해진다. 이 시스템에서는 기존의 합성곱 포즈 기계(Covolutional Pose Machine)[1] 방법을 MobileNet[2] 형태로 변형한 모델을 사용해 정확성과 신속성 모두 달성할 수 있었다. 구글에서 배포한 Mediapipe[3]의 손 관절 추론 네트워크와 비교했을 때에도 좋은 결과를 확인하였다.

2. 관련 연구

2.1 손 위치 검출 및 관절 추론

손 관절을 추론하기 전 손 위치를 검출하는 연구가 먼저 활발하게 이뤄졌다. 손 위치 검출은 기존의 물체 및 전경 위치 검출 알고리즘을 사용하는 경우가 일반적이다. 물체의 위치를 사각 박스 형태로 정답을 주고 이를 추론하는 연구는 마찬가지로 손 위치 검출에 충분히 활용될 수 있었고, 대부분 R-CNN 계열의 방법을 활용한다. 구글의 Mediapipe[3]에서는 SSD[4]과 Mobilenet[2]을 결합한 BlazeFace[4] 구조를 이용하여 손 위치를 검출하는 대신, 손바닥의 위치를 검출하였다.

손 관절을 추론하는 연구는 다양하게 이뤄졌는데 크게 RGB 이미지와 뎁스 이미지(RGB-D)가 결합된 연구가 있다. 뎁스 이미지로 3차원의 손 관절을 추론하는 방법은 일반적으로 뎁스의 특징맵을 활용해 반복적인 최적화 기법을 활용했다. 이후 학습-기반의 방법으로 네트워크 단에서 관절과 손 모양까지 추론까지 하는 연구가 이뤄졌다.

RGB 이미지 또한 최적화 기법부터 학습-기반의 CNN을 활용한 구조가 많이 연구되었다. 특히 3차원 손 메쉬 추정의 경우 파라미터 기반의 MANO 모델을 대부분 활용하고 있다. 실제 학습은 3차원 관절 및 메쉬간의 차이를 이용한 손실 함수가 최소화가 되게끔 구성한다.

이러한 딥러닝 학습이 하기 위해서는 다양한 종류의 RGB 및 RGB-D 이미지와 정답이 있어야 한다. 존재하는 모든 손 관절에 해당하는 이미지를 얻기에는 현실적인 어려움이 있으므로, 실제 이미지 기반 데이터셋과 함께 컴퓨터 그래픽스를 활용한 생성한 합성 이미지 기반 데이터셋도 함께 사용된다. 대표적인 RGB-D 데이터셋은 FHAD[5], BigHand2.2M[6], NYU[7]가 있으며, RGB로는 InterHand2.6[8]가 있다.

2.2 딥러닝 경량화를 이용한 연구

이미지에서 특징을 뽑는 대표적인 네트워크로는 VGG[9], ResNet[10]등이 있다. 하지만 이러한 네트워크는 파라미터의 개수가 많고 추론 속도가 비교적 느리므로, 모바일에 적합한 네트워크로 경량화되고 구조를 개선시켰는데, 대표적으로 MobileNet이 있다. 이러한 MobileNet 및 다른 경량화된 네트워크를 활용해 실제 신체의 포즈를 추론하는 연구가 있으며, 특히 모바일 기기를 목표로 성능과 시간을 정량적으로 비교한 연구도 존재한다.

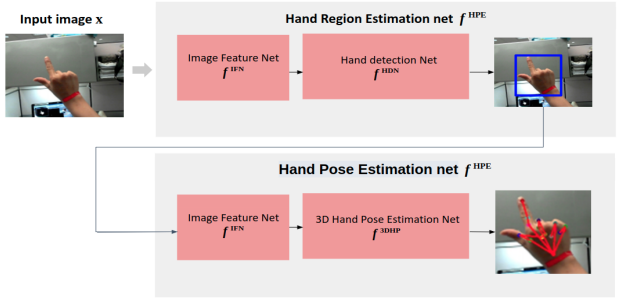
3. 방 법

3.1 손 위치 검출 네트워크

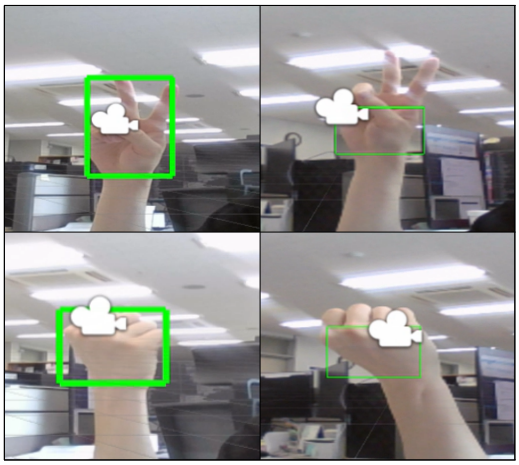
<그림 1>의 손 영역 추정 네트워크를 참고하면, 이미지의 특징을 추출한 후, 출력된 특징을 CNN 기반의 손 위치 검출 네트워크에 입력한 후 기존의 만들어진 손 위치 정답과 회귀적으로 비교해 지도 학습한다.

Mediapipe는 Anchor Box의 개수를 줄이기 위해 손 전체가 아닌 손바닥의 위치를 검출하고 일정 비율을 곱하여 손의 크기를 추론하였다. 우리는 중심점을 기준으로 손 위치를 검출하는 Centernet[11]구조를 사용하여 anchor box의 개수를 줄여 손바닥 대신 손 전체를 추론함에도 Anchor box의 개수를 유지하였다. 또한 합성곱 연산을 depthwise separable convolution 연산

으로 대체하여 연산 속도가 향상되었다.



<그림 1> 제안한 알고리즘의 네트워크 구성도



<그림 2> 손 객체 검출 Unity 시뮬레이션

3.2 손 관절 추론 네트워크

손의 관절을 추론하기 위해서는 입력되는 이미지의 손 영역이 비교적 커야한다. 따라서 앞서 3.1에서 다룬 손 영역 추정 네트워크에서 추정한 결과에 따라 이미지를 자른 후, 해당 이미지를 이미지 특징을 추출한 네트워크와 손 관절을 추정하는 네트워크에 입력된다(<그림 1> 참고). 최종적으로 출력되는 텐서의 차원은 (21,3)이며, 이는 손 관절의 개수와 일치한다. 이를 기존의 3차원 손 관절과 지도 학습하며 네트워크를 학습한다. 해당 네트워크는 Mediapipe[3]의 구조와 동일하다.

4. 실 험

4.1 AR 디바이스 시뮬레이션

실제 AR 글래스 애플리케이션의 시뮬레이션을 위해 경제성이 높은 제품은 Nreal 기기를 사용했다. 안드로이드 운영체제 기반의 모바일 디바이스 내의 연산처리장치를 활용해 성능을 높였으며, AR 애플리케이션을 개발할 수 있는 SDK는 Unity 응용프로그램으로 지원된다. Unity에서 손 객체 검출 모델을 이용하여 시뮬레이션한 결과는 **<그림 2>**와 같다. 첫 번째 열은 본 논문에서 제안한 모델로 손 위치를 추론한 결과이고, 두 번째 열은 구글 Mediapipe로 추론한 결과이다. 기존의 Mediapipe는 손바닥 위치를 추론하므로 정확한 손의 영역을 추론하기는 어렵다. 반면 우리가 제안한 방식은 같은 Anchor box 개수를 사용함에도 더 정확한 손 위치를 추론함을 확인할 수 있다.

4.2 정량적 성능 비교

기존에 구글에서 배포한 Mediapipe[3]와 성능을 비교했다. 정확도 비교는 학습한 데이터셋이 서로 상이하며 탐지하려는 객체(손, 손바닥)이 서로 다르므로 **4.1**과 같이 정성적인 측면으로 측정하였다. 추론 시간은 Mediapipe[3]가 약 20% 정도 더 빠르지만, Mediapipe[3]의 입력 영상의 크기가 2배 작다는 점을 고려하면 제안한 모델이 충분히 가벼운 모델임을 알 수 있다.

<표 1> 손 위치 검출 네트워크 성능 비교

	입력 영상 해상도	추론 시간
Mediapipe	256 x 256	5.4ms
제안한 모델	512 x 512	6.5ms

3. 결 론

AR/VR과 같은 사람-컴퓨터 상호작용(HCI) 애플리케이션에서 필수적인 손 위치 검출기를 제안하였다. 제안한 검출기는 Centernet기반의 손 위치 검출 네트워크를 구현하여 Anchor box의 개수를 줄이고, depthwise separable convolution을 이용해 연산 속도를 감소시켜 성능을 크게 개선하였다. 이러한 손 위치 검출기에 손 관절 추론 네트워크를 붙여 AR/VR 기기에서도 동작할 수 있도록 제작하였다. 향후 목표는 AR/VR 기기에서 관절 추론 네트워크를 이용한 3차원 손 자세 추론뿐만 아니라, 메쉬까지 복원하여 더욱 풍부한 손 정보를 추론하는 것이다.

감사의 글

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임. (No. 2020-0-00537 5G 기반 저지연 디바이스-엣지클라우드 인터렉션 기술 개발)

[참 고 문 헌]

[1] Shih-En Wei, Varun Ramakrishna, Takeo Kanade and Yaser Sheikh, “Convolutional Pose Machines”, CVPR, 2016

[2] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov and Liang-Chieh Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks”, CVPR, 2018

[3] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang and Matthias Grundmann, “MediaPipe Hands: On-device Real-time Hand Tracking”, CVPR workshop, 2020

[4] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran and Matthias Grundmann, “BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs”, CVPR workshop, 2019

[5] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek and Tae-Kyun Kim, “First-Person Hand Action Benchmark With RGB-D Videos and 3D Hand Pose Annotations”, CVPR, 2018

[6] Shanxin Yuan, Qi Ye, Bjorn Stenger, Siddhant Jain, and Tae-Kyun Kim, “BigHand2.2M Benchmark: Hand Pose Dataset and State of the Art Analysis”, CVPR, 2017

[7] Jonathan Tompson, Murphy Stein, Yann Lecun and Ken Perlin, “Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks”, ACM Transactions on Graphics, 2014

[8] Gyeongsik Moon, Shoou-i Yu, He Wen, Takaaki Shiratori, Kyoung Mu Lee, “InterHand2.6M: A Dataset and Baseline for 3D Interacting Hand Pose Estimation from a Single RGB Image“, ECCV, 2020

[18] Karen Simonyan and Andrew Zisserman, ”Very Deep Convolutional Networks for Large-Scale Image Recognition“, Arxiv, 2014

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, “Deep Residual Learning for Image Recognition”, CVPR, 2016

[20] Xingyi Zhou, Dequan Wang and Philipp Krähenbühl, “Objects as Points”, Arxiv, 2019