일반논문 (Regular Paper)
방송공학회논문지 제29권 제1호, 2024년 1월 (JBE Vol.29, No.1, January 2024)
https://doi.org/10.5909/JBE.2024.29.1.19
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

3D 피트니스 인체 포즈 추정을 위한 MoveNet의 확장

박 병 준[a], 윤 경 로[a]‡

Extending MoveNet for 3D Fitness Human Pose Estimation

Byeongjun Park[a], and Kyoungro Yoon[a]‡

요 약

별도의 센서를 사용하지 않고 RGB 영상으로 사람의 자세를 추정하는 것은 비용 절감과 사용성을 위한 중요한 포인트이며, 최근에는 HCI, 로봇공학, 비디오 분석, 메타버스 등 다양한 분야의 필요성으로 인해 점차 대중화되고 있다. 자세 추정에 대한 많은 연구 개발 시도가 있었지만, 한 시점의 사진으로 3차원 사람의 자세를 추정하는 것은 깊이 모호성, 객체 폐색, 배경 무질서, 훈련 데이터의 부족 등으로 연구에 어려움을 겪고 있다. 최근 논문들은 폐색에 초점을 맞추어 모델을 설계하고 있지만, 기존의 많은 3차원 자세 훈련 데이터에는 폐색 정보가 포함되어 있지 않다. 본 논문에서는 RGB 영상을 이용하여 폐색 없는 데이터에서 빠르게 추론할 수 있는 3차원 사람의 자세 추정기를 제안한다. 제안한 추정기는 2차원 인체 자세 추정기인 MoveNet을 기반으로 3차원 자세 추정이 가능하도록 변환하였다. 3차원 자세 추정을 위한 하이퍼파라미터를 최적화하기 위해 하이퍼파라미터를 다양화하는 실험을 진행하였으며, 단일 시점 RGB 이미지를 이용한 최첨단 논문과 성능을 비교하였다.

Abstract

Human pose estimation from RGB images without using other sensors is an important point for cost reduction and usability, and it has become increasingly popular in recent years due to the need in various fields such as HCI, robotics, video analytics, and metaverse. There have been many researches and development attempts on pose estimation, but 3D human pose estimation from a monocular image is experiencing difficulties in research due to depth ambiguity, object occlusion, background disorder, and lack of training data. Recent papers design models with a focus on occlusion, but many existing 3D pose training data do not include occlusion information. In this paper, we propose a 3D human pose estimator that can infer quickly from data without occlusion using RGB images. Based on MoveNet, which is a 2D human body pose estimator, it was transformed to be able to estimate 3D pose. In order to optimize the hyperparameters for 3D pose estimation, we conducted an experiment to diversify the hyperparameters, and compared the performance with the state-of-the-art papers using a monocular RGB image.

Keywords : Neural Network, 3D Pose Estimation, MoveNet, RGB Image, Fitness

a) 건국대학교 스마트ICT융합공학과(Dept. of Smart ICT Convergence Engineering, Konkuk University, Korea)
‡ Corresponding Author : 윤경로(Kyoungro Yoon)
E-mail: yoonk@konkuk.ac.kr
Tel: +82-2-450-4129
ORCID: https://orcid.org/0000-0002-1153-4038
※본 연구 논문은 2023년 건국대학교 교내연구비 지원으로 이루어진 연구결과입니다.
· Manuscript September 13, 2023; Revised January 25, 2024; Accepted January 25, 2024.

Ⅰ. 서 론

2020년 코로나바이러스 감염증 (COVID-19)의 확산으로 전 세계 인구의 이동과 활동이 제한되고 운동 부족으로 인한 건강 문제가 발생하기 시작했다. 이러한 문제점을 해결하기 위해 비대면 홈 트레이닝 플랫폼도 개발이 가속화되었다. 비대면 홈 트레이닝 플랫폼에서 운동 코칭을 받기 위해서는 트레이너를 대신하여 컴퓨터가 사람의 행동을 인식하고 3D 자세 추정을 통해 피드백을 주어야 한다. 인체 자세 추정은 단일 시점의 단일 카메라에서 획득한 이미지나 동영상에서 인체 부위를 추정하는 과정이다. 인간 자세 추정은 인간 중심 작업(예: 인간 감지, 인간 추적 및 인간 행동 인식)에서 감시, 행동 분석 및 자율 주행에 이르기까지 다양한 응용 분야에 대한 풍부한 인간 동작 정보를 제공한다[1]. 로보틱스, 이미지 분석, 메타버스 등의 분야에 대한 관심이 높아지면서 인간과 컴퓨터의 상호작용에 대한 관심이 잇달아 높아지고 있다.

인체의 자세 추정은 관절의 자유도가 높고, 이미지에 대한 변수가 많으며, 시점의 변화, 복잡한 배경 등으로 인해 어려운 문제이다. 이러한 어려움은 굴절 마커[2][3], 깊이 센서[4][5], 관성 측정 장치(IMU)[6][7]와 같은 센서를 사용하고 제한된 환경에서 촬영하는 MoCap 시스템에서도 나타난다. 이러한 시스템에는 추가 하드웨어가 필요하며 동시에 데이터 수집 프로세스에서 사람의 움직임을 제한하는 등의 제약이 따르는 문제점이 있다.

딥 러닝 기술의 발전과 함께[8][9], 최근 연구는 사용 환경의 편의성을 고려하여 RGB 이미지에서 사람의 자세를 추정하는 것에 초점을 두고 있다[10][11]. 그러나 2D 키포인트 주석을 쉽게 수집할 수 있지만, 3D 키포인트 주석을 정확하게 수집하기는 여전히 어려운 문제이다. 또한 입력 요소, 자세 정보의 표현, 추정 가능한 인원수에 따라 적용 분야 및 난이도가 달라질 수 있다. 데이터의 재구성 모호성으로 인해 단일 관점에서 수집한 데이터의 경우 3D 포즈 주석을 만들기 어렵게 되고 벤치마크 데이터 셋을 이용한 3D 포즈 추정은 더 어려워진다. AI-Hub의 피트니스 자세 데이터는 5가지 시점에서 동시에 촬영한 영상의 2D 포즈 주석과 이들을 병합한 하나의 3D 포즈 주석을 제공하므로 본 논문에서 사용하는 데이터 셋으로 채택했다.

본 논문은 기존의 실시간 2D 인체 자세 검출기인 MoveNet의 구조를 차용하여 3D 자세를 추론할 수 있도록 확장하였다. 이를 위하여 AI-Hub의 피트니스 데이터를 바탕으로 3D 월드 좌표를 2D 픽셀 좌표로 변환하는 방법을 제시하였으며 2D 자세 검출기를 3D로 확장하는 과정에서 히트맵 노이즈를 제거하기 위한 적절한 클리핑 계수를 실험을 통해 설정했다.

본 논문의 구성은 다음과 같다. 2장에서는 3차원 인체 자세 검출에 필요한 배경지식 및 관련 연구에 대해 설명한다. 3장에서는 데이터 셋의 처리 방법과 제안한 3D 인체 자세 추정 시스템의 구현에 대해 설명한다. 4장에서는 하이퍼파라미터와 최신 논문 비교 실험 결과와 분석을 기술한다. 마지막으로 5장에서는 결론 및 향후 과제를 제시한다.

Ⅱ. 배경지식 및 관련 연구

이 부문에서는 키포인트를 통해 인체 자세 추정을 정의한 관련 연구와 제안한 모델의 기반이 되는 2D 키포인트 검출기인 MoveNet에 대해 설명한다.

1. 2D Human Pose Estimation

2D 인체 자세 추정은 신체 관절의 키포인트를 정의하고 2D 좌표로 키포인트의 위치를 추정하는 것을 말한다. 이미지로부터 키포인트를 추출하기에 이미지 품질에서 문제가 발생할 수 있다. 예를 들어, 다른 사물이나 다른 신체 부위에 의해 가려지거나 빛의 밝기나 배경의 무질서 등으로 인해 키포인트가 보이지 않아 사람의 포즈를 추정하기 어려운 문제가 있다. 최근 몇 년간 딥 러닝이 발전함에 따라[9]-[11] RGB 이미지에서 2D 키포인트를 감지하는 성능이 크게 향상되었다[12][13]. 인간의 자세 추정 정확도가 높아짐에 따라 감시, 행동 분석, 자율주행 등 다양한 응용 분야에 대한 연구가 활발히 진행되고 있다.

2. 3D Human Pose Estimation

기존의 3D 인체 자세 추정 방법은 세 가지 범주로 분류

할 수 있다[14]. 첫째, 3D 포즈 추적은 프레임 간 추적을 기반으로 하며 3D 인체 자세 추정에 대한 초기 논문의 대부분이 이 범주에 속한다. 둘째, 2단계로 나뉜 2D-3D 포즈 리프팅으로 2D 자세를 감지한 후 3D 자세로 끌어올리는 것이다. 셋째, 이미지 자체의 픽셀에서 3D 포즈를 직접 유추하는 포즈 회귀 분석 작업이다.

SR-Net[15]은 희귀한 포즈 학습을 위해 다른 포즈에서 부분 관절 구성이 나타남을 이용한다. 신체를 몇 개의 하위 영역으로 나누고 이를 별도의 네트워크 분기로 취급한다. 글로벌 일관성은 각 분기의 컨텍스트를 신체의 나머지 부분에서 저차원 벡터로 재조합하여 유지한다. 제안된 세분화 및 재조합 접근법은 폐색(occlusion)이 발생한 포즈의 예측에서 상당한 정확도 향상을 가져온다.

EvoNet[16]은 훈련 데이터에서 볼 수 없는 포즈에서 추론이 실패할 수 있다는 점을 고려하여 2D-3D 네트워크를 훈련하기 위해 방대한 양의 훈련 데이터를 합성할 수 있는 새로운 데이터 세트로 증강하여 데이터 세트 편향을 효과적으로 줄인다. 계층적 인간 표현과 사전 지식에서 영감을 얻은 휴리스틱을 기반으로 보이지 않는 3D 인간 골격을 합성하기 위해 제한된 데이터 세트를 증강한다.

가우스 분포를 기반으로 평균 제곱 오차를 최소화하는 기존의 딥 러닝 접근 방식과 달리 MDN[17]은 다중 모드 혼합 밀도 네트워크를 기반으로 3D 포즈의 여러 실현 가능한 가설을 생성한다. MDN 접근법에 의해 2D 인간 포즈 추정기에서 2D 관절을 입력함으로써 추정된 3D 포즈가 2D 재투영에서 일관성이 있음을 보여주고 2D 관절에서 3D 포즈의 여러 가능한 가설을 생성하는 새로운 접근 방식을 제안한다.

3. MoveNet

MoveNet[18]은 17개의 2D 키포인트를 감지하는 빠르고 정확한 모델이다. 이 모델은 TensorFlow Hub[19]에서 표준 모델과 경량 모델의 두 가지 형태로 제공된다. 신체 자세 추정은 지난 5년 동안 상당한 발전을 했지만 아직까지 많은 응용 프로그램에 적용되지 않았기에 MoveNet은 빠르게 만들고 어디에서나 실행할 수 있도록 하는 데 중점을 둔다. MoveNet의 목표는 최신 논문에 적용된 시스템 구성의 장점을 활용하여 모델을 설계하고 추론 시간을 최대한 짧게 최적화하는 것이다. 그 결과 자세, 환경, 하드웨어 설정 등의 변수가 있음에도 정확한 키포인트를 추정한다.

MoveNet은 히트 맵을 사용하여 사람의 키포인트를 정확하게 지역화하는 상향식 추정 모델이다. 그림 1은 크게 특징추출기와 예측 헤더 두 부분으로 구성된 MoveNet의 구조를 나타낸 그림이다. 예측을 위한 콘셉트는 속도와 정확도가 눈에 띄게 변경된 CenterNet을 기반으로 한다. MoveNet의 특징 추출기는 MobileNetV2[20]를 사용하며 추출된 특징 맵은 4개의 예측 헤더에 입력된다. MobileNetV2는 FPN[21] 구조를 가진 피처 추출기로서 stride를 4로 채택하여 의미적으로 풍부한 고해상도 피처 맵 출력을 얻는다.

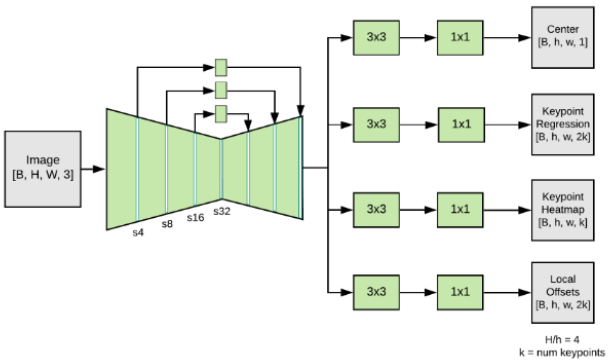

그림 1. MoveNet 구조
Fig. 1. MoveNet structure

Ⅲ. 3D 휴먼 키포인트 추정기

1. 피트니스 자세 이미지 데이터

피트니스 자세 이미지 데이터는 15초 내외의 동영상 20만 클립을 촬영한 후 이미지를 추출하여 생성한 데이터 셋으로 올바른 운동 자세에 대한 피드백을 주고자 하는 목적으로 제작되었다. 총 41종의 운동이 있고 각 운동의 운동 상태는 816개의 시나리오로 구성되어 있으며 5대의 카메라로 360도 촬영되었다. 촬영된 영상에서 총 300만 개의 이미지를 추출하여 주석을 달아 구축된 데이터 셋이다. 촬영 카메라는 FHD (1920×1080) 해상도로 2 FPS 이상으로 촬영했다.

대부분의 자세 추정 데이터 셋에서 공통적으로 사용되는 키포인트인 코, 왼쪽 눈, 오른쪽 눈, 왼쪽 귀, 오른쪽 귀, 왼쪽 어깨, 오른쪽 어깨, 왼쪽 팔꿈치, 오른쪽 팔꿈치, 왼쪽 손목, 오른쪽 손목, 왼쪽 엉덩이, 오른쪽 엉덩이, 왼쪽 무릎, 오른쪽 무릎, 왼쪽 발목, 오른쪽 발목의 17개 키포인트를 사용한다. 데이터 전처리는 회전 및 크기 변환, 이상 값 제거, 이미지 자르기, 키포인트 좌표 정규화 등의 전처리 과정을 거친다.

5개의 2D 픽셀 좌표와 1개의 3D 월드 좌표가 제공되고 각 이미지에 맞는 3D 좌표를 얻기 위해서는 3D 월드 좌표를 픽셀 좌표와 맞추는 작업이 필요하다. 그림 2는 기준(축, 크기)이 다른 3D 월드 좌표를 2D 픽셀 좌표의 기준으로 변경하기 위해 회전 및 크기 변환하는 과정이다. 3D 키포인트의 X축을 고정 후 180도 회전하여 2D 픽셀 좌표와 X, Y 축을 맞추고 Y 축을 고정 후 Θ도 회전, 크기 변환, 평행 이동을 거쳐 식 (1)과 같이 2D 키포인트와 3D 키포인트의 x, y 값의 2D JPE(Joint Position Error)를 계산한다. 설치된 카메라 별 3D 좌표와 각도가 다르기에 Θ도를 회전 및 변경하며 2D JPE가 가장 낮은 각도의 변환된 3D 좌표를 데이터로 선택한다.

$$2D\ JPE = \sum_{j}^{joints} \sqrt{(x_j - \hat{x}_j)^2 + (y_j - \hat{y}_j)^2} \qquad (1)$$

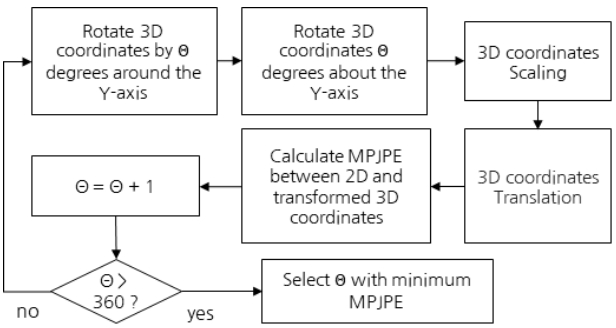

그림 2. 3D 데이터 회전 및 크기 변환
Fig. 2. 3D data rotation and scaling

추출된 사람의 키포인트의 바운딩 박스의 중심의 x좌표가 이미지의 좌측 끝 또는 우측 끝에서 전체 길이의 20% 미만으로 떨어져 있는 경우 다른 객체를 탐지하였거나 정합의 정확도가 떨어짐을 고려하여 훈련 데이터에서 제거하였으며, 3D 월드 좌표가 픽셀 좌표로 변환한 후 2D JPE가 계산되며 그 값이 10 이상이면 훈련 데이터에서 제거했다.

선택된 데이터의 좌표의 2D JPE가 17 초과인 데이터는 제외하였다. 키포인트 바운딩 박스 중심을 (x, y), 너비와 높이를 (w, h)라 할 때, 이미지를 중심이(x, y) 너비와 높이가 (w + h, w + h)인 정사각형 모양으로 잘라낸다. 전처리할 이미지의 크기가 원본 이미지의 크기를 초과할 시 검정 바탕으로 패딩하였다.

이상치를 제거한 단일 시점의 201,486개의 AI-Hub의 피트니스 자세 이미지를 8 : 1 : 1 비율로 나누어 학습 데이터는 160,873장, 검증 데이터는 20,503장, 테스트 데이터는 20,110장으로 분할하여 사용했다.

2. 제안모델

2D 인체 자세 추정기인 MoveNet 모델을 기반으로 3D 인체 자세 추정 시스템을 구축하였다. 그림 3 (a)와 같이 모델의 입력값은 192×192×3 컬러 이미지를 입력받아 피처 추출기 모델인 MobileNetV2를 통해 24×48×48 크기의 피처맵을 출력한다. 피처맵은 각 헤더에 대해 그림 3 (b)와 같이 3×3 및 1×1 합성곱 신경망 블럭을 거치며 센터 히트맵 헤더는 48×48×48×1, 키포인트 회귀 헤더는 48×48×48×51, 키포인트 히트맵 헤더는 48×48×48×17, 로컬 오프셋 헤더는 48×48×48×51의 차원을 갖는다. 헤더의 출력된 결과는 clipping 함수를 거치고 clipping 함수에 대한 실험은 4.2에 서술한다.

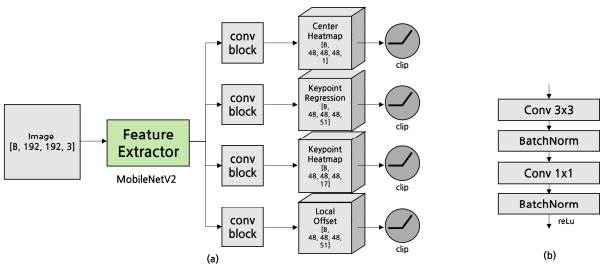

그림 3. 제안된 모델 구조
Fig. 3. proposed model structure

각 출력 히트맵과 회귀 필드에 따라 차이를 두고 손실

함수를 설정했다. 센터 히트맵과 키포인트 히트맵의 경우 출력의 손실 함수로 MSE(Mean Square Error)를 사용했다. 식(2)에 표시된 것처럼 J는 키포인트 관절의 수, Y는 출력 히트맵 벡터, α, β는 가중치 계수로 각 8, 1로 설정했다. 센터 히트맵의 경우 훈련 데이터에서 사람의 중앙을 기준으로 정사각형 형태로 잘라내는 전처리 과정을 거쳤기 때문에 중앙 히트맵이 중앙에 가까울수록 높은 가중치를 부여했다.

$$HeatmapLoss = \frac{1}{J}\sum_{j}^{J}((Y_j - \hat{Y}_j)^2 * (\alpha Y + \beta)) \quad (2)$$

각 키포인트 간의 관계는 다양하게 나올 수 있지만 얼굴과 어깨의 위치와의 관계는 다양하게 나올 수 없으므로 키포인트 히트맵의 해당 키포인트 간의 변화가 적어지도록 얼굴 손실 함수를 식(3)과 같이 설정했다. Y는 출력 히트맵, N은 관절 쌍의 수이며, E는 코, 왼쪽 눈, 오른쪽 눈, 왼쪽 귀, 오른쪽 귀, 왼쪽 어깨, 오른쪽 어깨 키포인트들 간의 쌍으로 이루어진 집합으로 키포인트 간의 벡터를 뼈로 생성하여 MSE를 계산한다.

$$BoneLoss = \frac{1}{N}\sum_{(i,j)\in E}\sqrt{((Y_i - Y_j) - (\hat{Y}_i - \hat{Y}_j))^2} \quad (3)$$

키포인트 회귀와 지역 오프셋 회귀의 손실 함수는 식(4)와 같이 예측한 키포인트와 실제 키포인트 사이의 차이 값에 절댓값을 취하는 L1 손실 함수를 사용했다. 이 손실 함수들의 합이 낮아지도록 학습하였다.

$$RegLoss = \frac{1}{J}\sum_{j}^{J}|Y_j - \hat{Y}_j| \quad (4)$$

모델의 출력은 48×48×48 히트맵 형태로 표현되며 식(5)를 이용하여 히트맵의 활성화가 적게 된 부분을 0으로 설정하여 학습에 혼란을 주지 않도록 하였다. 기존 2D MoveNet의 경우 Threshold를 0.1로 설정하였으나 제안 모델에서는 2D에서 3D로 차원을 확장하여 출력 히트맵의 크기가 커져 활성화 정도가 전체에 미치는 영향이 감소하였다고 판단하여 0.001로 낮추어 학습하였다.

$$clip(x) = \begin{cases} x\ (x > threshold) \\ 0\ (x < threshold) \end{cases} \quad (5)$$

Ⅳ. 3D 휴먼 키포인트 추정기

1. 실험환경

실험을 진행하기 위한 3D 키포인트 탐지시스템의 실험 환경에 대하여 서술한다. 단일 시점의 RGB 이미지를 입력으로 표 1과 같은 실험 환경에서 3D 키포인트 탐지 시스템을 실행하였다.

표 1. 실험 환경
Table 1. Experimental environment

OS	Windows 10
CPU	3.8GHz Ryzen 7 5800X 8-Core
GPU	RTX-3090
CUDA	v11.6
RAM	96GB
Program Language	Python 3.9

2. 실험 결과 및 분석

실험은 하이퍼파라미터에 따른 피트니스 자세 데이터에 대해 본 논문에서 제안한 3D 자세 추정기의 성능을 MPJPE (Mean Per Joint Position Error) 척도에서 동일한 작업을 수행한 최신 논문들과 비교한다[3]. MPJPE는 식(6)과 같이 추정된 키포인트와 Ground-Truth 키포인트 위치 사이의 유클리드 거리를 산술평균하여 계산한 지수이다. MPJPE가 낮을수록 Ground-Truth와 근사하게 추정했음을 의미한다. 표 2의 하이퍼파라미터로 학습하고 검증 데이터 세트에 대해 평가했다. 그림 4에서는 모델의 정확도를 평가하기 위해 좌측 그래프는 손실 함수의 값을, 우측 그래프는 MPJPE

$$MPJPE = \frac{1}{frames}\frac{1}{joints}\sum_{f}^{frames}\sum_{j}^{joints}\sqrt{(x_{(f,j)} - \hat{x}_{(f,j)})^2 + (y_{(f,j)} - \hat{y}_{(f,j)})^2 + (z_{(f,j)} - \hat{y}_{(f,j)})^2} \quad (6)$$

표 2. 모델 하이퍼파라미터
Table 2. Model hyperparameter

num workers	8
img_size	192 × 192
optimizer	Adam
learning_rate	0.0015(1 ~44 epoch), 0.001(45 ~ epoch)
weight_decay	5.e-4
heatmap_threshold	0.001

값이 50 미만이 나온 비율을 각 epoch 별로 보여주고 있다. 학습의 종료 지점은 early stopping의 patience 값을 5로 설정하여 87 epoch에서 검증 데이터 세트의 정확도는 82.36%에 도달했고 손실 함숫값은 60으로 감소했다.

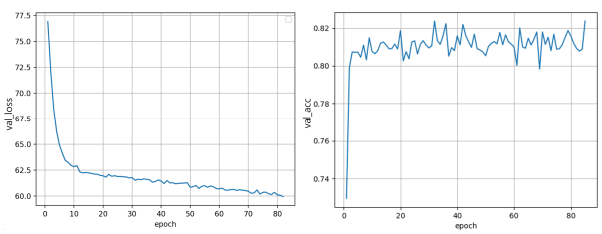

그림 4. 검증 손실 값 및 정확도
Fig. 4. Training loss and accuracy

그림 5는 실험을 위한 12가지 동작에 대해 본 논문에서 제안한 3D 휴먼 키포인트 검출 시스템으로 검출한 사진이다. 주어진 환경에서, 제안 모델의 처리 속도는 한 프레임당 약 0.78~0.79초가 걸려 13FPS(Frame Per Second)를 갖는다.

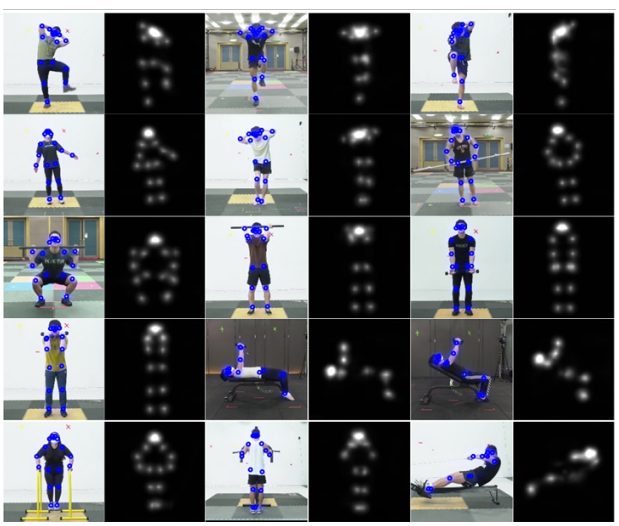

그림 5. 3D 자세 추정 결과 이미지와 히트맵
Fig. 5. 3D pose estimation results image and heatmap

표 3. 운동별 MPJPE
Table 3. MPJPE per excercise

Push-up	32.375
Y-Exercise	48.138
Standing side crunch	60.782
Crunch	45.605
Hip thrust	34.360
Dumbbell incline	29.279
Dumbbell chest fly	29.456
Knee push-up	38.337
Side lunge	50.001
Good morning	60.718
Cross lunge	26.481
Step backward dynamic lunge	58.945
Bicycle crunch	31.128
Lying leg raise	50.177
Standing knee up	41.476
Chest fly	36.078
Barbell lunge	29.137

본 연구에서 평가한 운동 동작 별 MPJPE 결과는 표 3과 같다. 특정 운동 동작이 다른 동작보다 더 복잡하거나 관절의 가림 현상이 발생한 경우, 인체 키포인트 검출의 정확도가 낮아질 수 있다. 예를 들어, 크로스 런지 (26.481)와 덤벨 인클라인 (29.279) 같은 운동에서는 상대적으로 높은 정확도를 보이고 굿모닝 (60.718)과 스탠딩 사이드 (60.782)는 상대적으로 복잡한 동작으로, 이로 인해 검출기가 더 낮은 성능을 보일 수 있다. 또한 낮은 성능을 내는 이유로 자세 변화의 빈도로 빠르게 변하는 동작이나 많은 자세 변화를 포함하는 동작은 인체 키포인트 검출에 어려움을 줄 수 있다. 예를 들어, 스텝 백워드 다이나믹 런지 (58.945)는 다른 동작에 비해 자세 변화가 많아 정확도가 낮을 수 있다.

그림 6 좌측에서는 출력된 히트맵을 키포인트로 변환하는 과정에서 클리핑 임곗값에 따른 모델 학습 성능을 나타

낸다. 임곗값 0.001을 기준으로 임곗값을 변경하며 검증 정확도를 비교했다. 기존 MoveNet 임곗값 0.1(녹색)로 학습을 진행하면 초기 검증 정확도가 현저히 떨어지고 70% 검증 정확도가 멈추는 것을 볼 수 있다. 예측된 히트맵 공간이 (48 × 48)에서 (48 × 48 × 48)로 넓어짐에 따라 출력값이 희소해지고 높은 임곗값이 학습을 방해한다고 추측할 수 있다.

학습 정확도가 80% 초반에 수렴함에 따라 60 epoch에서 학습된 가중치를 기준으로 히트맵 임곗값을 변경하여 매개변수를 학습했다. 그림 6 우측에서 히트맵 임곗값을 기준치보다 낮추어 테스트를 했을 때 정확도가 떨어지는 모습을 보였다. 위 실험에서 클리핑 임곗값이 낮을수록 학습이 잘 된다는 가정은 학습이 덜 된 초기에만 해당되고 학습이 진행될수록 임곗값을 높여야 된다는 결론을 얻었다.

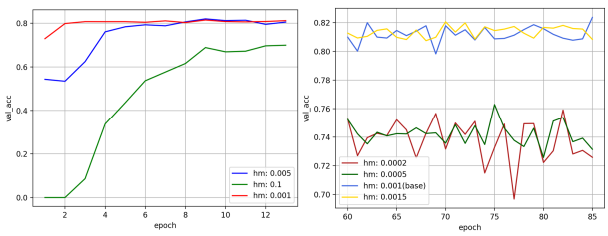

그림 6. 히트맵 임곗값에 따른 검증 정확도
Fig. 6. Validation accuracy as to heatmap threshold

표 4에서는 같은 task의 최신 논문 모델의 학습된 가중치를 대상으로 테스트 데이터 셋 또한 평가하였다. 표 2는 최신 제안한 3D 휴먼 포즈 추정기와 성능을 비교한 결과이다. 제안 모델의 MPJPE가 최신 논문의 최신 논문 SRNet[15]에 비해 약 9만큼 낮은 값을 얻었기에 더 정확한 추론을 했다고 해석할 수 있으며 그래프를 기반으로 한 Sem-GCN[22]이 56.4로 가장 높은 MPJPE를 보여준다. 실험 결과 제안하는 방법이 가장 낮은 MPJPE 값을 보이므로 타 모델과 비교하였을 때 키포인트를 정확하게 추정함을 알 수 있다.

표 4. 비교 실험 결과
Table 4. Comparison test result

Method	MPJPE
A Zeng et al. 2020[15]	50.2
S Li et al. 2020[16]	53.4
C Li et al. 2019[17]	51.9
Kanazawa et al. 2019[22]	56.4
Ours	41.9

V. 결 론

본 논문에서는 단일 RGB 영상을 이용한 3차원 휴먼 포즈 추정 시스템을 제안했다. 이 모델은 단일 RGB 영상만을 사용하여 3차원 키포인트를 추정하는 독특한 방법을 제시하였고 AI-Hub에서 제작한 피트니스 자세 이미지 데이터의 3D 월드 좌표를 픽셀 좌표로 변환하여 학습에 활용함으로써 데이터의 품질을 고려한 학습을 보였다.

제안된 모델은 2D 히트맵 기반 모델인 MoveNet을 확장하여 3D 키포인트를 출력하는 추출기를 설계하였다. 이 과정에서 192×192 컬러 이미지를 입력으로 사용하고, MobileNetV2를 통해 획득한 24×48×48 피처 맵을 활용하여 각 헤더에 넣어 48×48×48 히트맵으로 출력하였다. 2D 키포인트 추출에서 3D 키포인트 추출로 작업을 변경하면서 출력하는 공간이 넓어지기에 출력 신호를 clipping 하는 값을 기존 값 0.1보다 낮은 0.005로 설정할수록 학습이 잘 되고 어느 정도 학습이 진행된 후에는 clipping 계수를 높여야 학습이 더 잘 됨을 보였다.

제안된 모델은 초당 약 13프레임의 빠른 추론 속도를 보이며 주어진 데이터 셋에서 82%의 높은 정확도를 보인다. 피트니스 데이터 환경에서는 키포인트 오프셋과 같은 문제가 자주 발생하지 않기 때문에 중복으로 인한 키포인트 손실을 고려하지 않고 설계하였으므로 3D 키포인트 검출이 가능한 심층 신경망을 학습하여 3D 키포인트를 추정하였다. 테스트 데이터 셋에 대한 실험 결과 MPJPE는 다른 논문에 비해 17~25% 감소하여 향상된 예측 성능을 보였다.

최근의 인체 자세 추정 논문은 실생활에 적용했을 때 발생하는 변수를 고려한 3차원 인체 특징점 추정 논문이 주를 이룬다. 따라서 향후 과제로는 키포인트 폐색 정보를 고려하고 시간 정보를 활용하여 단일 영상이 아닌 영상 정보를 실생활에 적용할 수 있도록 활용하는 방법을

연구할 계획이다. 이를 위해 사람의 자세가 편향되지 않도록 전처리 및 증강 기법을 적용하고, 키포인트 폐색 정보를 인식하는 레이어를 추가하여 다양한 용도로 사용할 수 있는 견고한 3차원 키포인트를 추정할 수 있을 것으로 기대한다.

참 고 문 헌 (References)

[1] X Wang, Q Ji "A Hierarchical Context Model for Event Recognition in Surveillance Video", The IEEE Conference on Computer Vision and Pattern Recognition, pp. 2561-2568, 2014.
doi: https://doi.org/10.1109/cvpr.2014.328

[2] L Sigal, AO Balan, MJ Black. "HumanEva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion." International Journal of Computer Vision, 2010
doi: https://doi.org/10.1007/s11263-009-0273-6

[3] C Ionescu, D Papava, V Olaru, C Sminchisescu. "Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments." IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014
doi: https://doi.org/10.1109/tpami.2013.248

[4] A Haque, B Peng, Z Luo, A Alahi, S Yeung, L Fei-Fei. "Towards viewpoint invariant 3D human pose estimation" ECCV Cham: Springer International Publishing, 2016
doi: https://doi.org/10.1007/978-3-319-46448-0_10

[5] T Yu, Z Zheng, K Guo, J Zhao, Q Dai, H Li, G Pons-Moll, Y Liu. "DoubleFusion: real-time capture of human performances with inner body shapes from a single depth sensor." IEEE/
doi: https://doi.org/10.1109/cvpr.2018.00761

[6] TM Trumble, A Gilbert, C Malleson, A Hilton, J Collomosse. "Total capture: 3D human pose estimation fusing video and inertial sensors." In: Procedings of the British Machine Vision Conference 2017
doi: https://doi.org/10.5244/c.31.14

[7] T Von Marcard, R Henschel, MJ Black, B Rosenhahn, G Pons-Moll. "Recovering accurate 3D human pose in the wild using IMUs and a moving camera" Computer Vision – ECCV 2018
doi: https://doi.org/10.1007/978-3-030-01249-6_37

[8] RA Güler, N Neverova, I Kokkinos. "DensePose: dense human pose estimation in the wild." In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, IEEE, 2018
doi: https://doi.org/10.1109/cvpr.2018.00762

[9] A Toshev, C Szegedy. "DeepPose: human pose estimation via deep neural networks." IEEE Conference on Computer Vision and Pattern Recognition, 2014
doi: https://doi.org/10.1109/cvpr.2014.214

[10] A Newell, K Yang, J Deng. "Stacked hourglass networks for human pose estimation." Computer Vision – ECCV, 2016
doi: https://doi.org/10.1109/cac.2018.8623582

[11] Y Chen, Z Wang, Y Peng, Z Zhang, G Yu, J Sun. "Cascaded pyramid network for multi-person pose estimation." IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018
doi: https://doi.org/10.1109/cvpr.2018.00742

[12] M Andriluka, L Pishchulin, P Gehler, B Schiele. "2D human pose estimation: new benchmark and state of the art analysis." IEEE Conference on Computer Vision and Pattern Recognition, 2014
doi: https://doi.org/10.1109/cvpr.2014.471

[13] W Yang, S Li, W Ouyang, H Li, X Wang. "Learning feature Pyramids for human pose estimation." IEEE International Conference on Computer Vision (ICCV), 2017
doi: https://doi.org/10.1109/iccv.2017.144

[14] X Ji, Q Fang, J Dong, Q Shuai, W Jiang, X Zhou. "A survey on monocular 3D human pose estimation." Elsevier Virtual Reality & Intelligent Hardware, 2020
doi: https://doi.org/10.1016/j.vrih.2020.04.005

[15] A Zeng, X Sun, F Huang, M Liu, Q Xu, S Lin. "SRNet: Improving Generalization in 3D Human Pose Estimation with a Split-and-Recombine Approach" Computer Vision – ECCV, 2020
doi: https://doi.org/10.1007/978-3-030-58568-6_30

[16] S Li, L Ke, K Pratama, YW Tai, CK Tang, KT Cheng. "Cascaded Deep Monocular 3D Human Pose Estimation with Evolutionary Training Data" IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020
doi: https://doi.org/10.1007/978-3-030-58568-6_30

[17] C Li, GH Lee. "Generating multiple hypotheses for 3d human pose estimation with mixture density network" IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019
doi: https://doi.org/https://doi.org/10.1109/cvpr.2019.01012

[18] MoveNet https://blog.tensorflow.org/2021/05/next-generation-pose-detection-with-movenet-and-tensorflowjs.html (accessed Jan. 27. 2024)

[19] TensorFlow Hub https://www.tensorflow.org/hub (accessed Jan. 27. 2024)

[20] M Sandler, A Howard, M Zhu, A Zhmoginov, L Chen "Mobilenetv2: Inverted residuals and linear bottlenecks."Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018
doi: https://doi.org/10.1109/cvpr.2018.00474

[21] TY Lin, P Dollár, R Girshick, K He, B Hariharan, S Belongie. "Feature pyramid networks for object detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017
doi: https://doi.org/10.1109/cvpr.2017.106

[22] L Zhao, X Peng, Y Tian, M Kapadia, D Metaxas "Semantic graph convolutional networks for 3d human pose regression" Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019
doi: https://doi.org/10.1109/cvpr.2019.00354

저 자 소 개

박 병 준

- Feb. 2017: B.S. in Computer Science., Dankook University.
- Feb. 2023: M.S.E. in Smart ICT Convergence Eng., Konkuk University.
- ORCID : https://orcid.org/0000-0003-2273-6230
- Research interests : multi-media, computer vision and video encoding.

윤 경 로

- Feb. 1987 : B.S. in Electronics and Computer Eng., Yonsei University.
- Dec. 1989 : M.S.E. in Electrical Engineering/Systems, University of Michigan, Ann Arbor.
- May 1999 : Ph.D., in Computer and Information Science, Syracuse University.
- June 1999~Aug., 2003 : Group Leader, LG Electronics Institute of Technology.
- Sept. 2003~Present : Professor, Dept. of Smart ICT Convergence Eng., Konkuk University.
- Oct. 2017~Present : Chair, ISO/IEC JTC1 SC29 Korea Mirror Committee.
- July 2019~Present : Chair, Digital Virtualization Forum.
- Sept. 2019~Present : Chair, IEEE 2888 Working Group.
- ORCID : https://orcid.org/0000-0002-1153-4038
- Research Interests : Smart media system, Multimedia retrieval, Image processing, Multimedia information and metadata processing, Metaverse, Digital Twin.