# 딥러닝 기반의 자세추정 알고리즘 개발

서경덕[1], 황의재[2+], 유화익[2+], 강현영[1], 류지승[1], 이예린[1], 권오윤[2*], 양세정[1*]

연세대학교 의공학과[1], 연세대학교 물리치료학과[2]

# Development of Pose Estimation Algorithm Based on Deep Learning

Kyungdeok Seo[1], Uijae Hwang[2+], Hwaik Yoo[2+], Hyeonyoung Kang[1], Jiseung Ryu[1], Yerin Lee[1], Kwon Ohyoon[2*] Sejung Yang[1*]

Department of Biomedical Engineering, Yonsei University, Korea

Department of Physical Therapy, Yonsei University, Korea

rud395@yonsei.ac.kr, *syang@yonsei.ac.kr, *kwonoy@yonsei.ac.kr

**Abstract**

Since the recent COVID-19 incident, the problem of estimating a person's posture has attracted considerable attention in the computer vision field as the demand for home healthcare and VR games increases. Existing pose estimation algorithms have a problem in that spatio-temporal loss occurs in the process of combining context information with keypoint location information. The purpose of the study is to develop a high-performance posture estimation algorithm by supplementing the problem by modifying the method of coupling the residual block used in the existing algorithm. Pose estimation was performed on five subjects using a model pre-trained with MPII datasets, an open-source keypoint dataset. As a result of the experiment, it is predicted that faster and more convenient analysis will be possible in the actual field.

## 1. 연구 배경

뇌가 받아들이는 외부 정보의 약 90%를 시각정보가 차지하고 있으며, 정[1]보를 저장하고 기억하는 시간도 가장 길어, 약 65%의 사람이 시각정보만을 이용해 학습한다고 한다[1]. 따라서 시각정보는 인간의 지능을 구성하는 가장 중요한 요소라고 할 수 있다. 컴퓨터 비전이란 컴퓨터에게 인간과 같은 시야를 제공하여 물체 및 환경 등을 인식 및 이해하게 하는 인공지능 기술이다. 최근 COVID-19 사태 이후, 컴퓨터비전 분야에서는 재택 진료 및 VR 게임 등의 수요가 증가함에 따라 사람의 자세를 추정하는 문제가 상당한 관심을 받고 있다.

자세란, 몸가짐이나 일정한 태도를 취하고 있는 모습으로 정의되며, 대표적인 비언어적 의사표현 수단이자 행동의 기본 단위이기도 하다. 다시 말해 자세는 사람의 움직임을 이해하는 데 필요한 중요한 정보라고 할 수 있다. 이러한 자세 정보를 영상에서 추출하여 사람의 자세를 추정하는 문제를 Human pose estimation이라고 한다. 자세는 관절의 움직임에 의해 나타나므로 자세 추정 문제는 관절의 위치에 대한 예측 문제로 재정의될 수 있다[2].

근 10년 사이 이 분야의 연구 흐름은 수작업으로 생성한 feature나 graphical method를 사용해 관절을 개별적으로 추정하는 것에서, 딥러닝 알고리즘을 사용하여 신체 전반을 동시에 추정하는 것으로 옮겨갔다. 개별적인 추론은 기존의 전통적인 방법들에 비해 관절 하나하나를 비교적 정확히 검출하긴 하지만 맥락 정보가 없어 전체적인 자세를 추정하는 데 있어 매우 낮은 정확도를 보였다. 이러한 문제점은 딥러닝 알고리즘을 이용해 관절의 위치 정보를 개별적으로 획득한 후, 신체 전반의 방향, 크기, 형태 등을 함께 신경망에 입력하여 자세를 추정하는 방식을 사용함으로써 획기적으로 개선되었다[3].

그러나 맥락 정보를 관절의 위치 정보와 결합시키는 과정에서 추가되는 모듈에 의해 시공간적 손실이 발생한다. 이번 연구에서는 이러한 손실을 방지하기 위해 모듈의 결합 방식을 수정하여 전체적인 맥락 정보를 강화함으로써 성능을 향상시키는 것을 목표로 한다. 모델은 오픈소스 키포인트 데이터셋인 MPII 데이터셋[5]을 사용하여 사전학습되었으며, 모델의 성능은 직접 촬영한 Deep squat 영상 데이터를 사용해 평가했다.

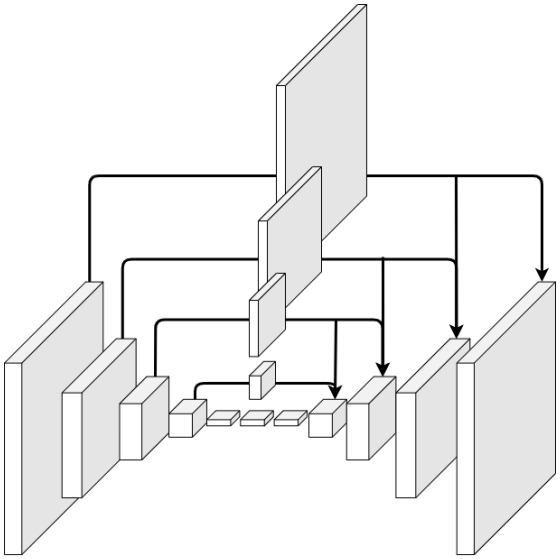

그림 1. Human pose estimation 모델 모식도

## 2.연구 방법

실험에는 20대 중반 남녀 5명이 참여했으며, Deep squat 영상은 스마트폰 후면 카메라로 촬영하였다. 영상은 하얀색 배경에서, 개별적으로 촬영되었다. 해당 영상의 정답 영상 또한 직접 제작하였으며, 이렇게 얻어진 Deep squat 영상과 정답 영상을 모델에 입력하여 영상 내 관절을 예측하도록 했다.

+: Contributed equally.

*: Corresponding author.

연구에는 인코더에서 획득한 각 관절의 위치 정보와 Residual block을 통해 얻은 전반적인 맥락 정보를 함께 디코더에 입력하여 심각한 폐색 및 왜곡 등 여러 과제에서 SOTA를 달성한 Stacked hourglass network(SHN)가 사용되었다. SHN 모델의 인코더와 디코더를 직접적으로 연결하는 Residual block은 디코더의 매 Convolutional block에 사용되는데, 기존 SHN은 가중치에 원본 영상의 특징을 합하는 방식으로 학습을 진행했으나 이번 연구에서는 잔차만 존재하는 차원을 추가하는 방식을 도입하여 원본 영상 특징의 손실이 전혀 없이 학습할 수 있도록 모델을 변경했다[4].

## 3. 연구 결과

실험에 사용된 MPII 데이터셋은 약 4만 명 이상의 인물이 포함된 2만 5천여 장의 이미지로 구성된 오픈소스 데이터셋으로, 유튜브 비디오에서 추출되었다. 각 이미지에는 사람의 관절 위치 정보 뿐 아니라 가려짐 여부, 머리, 상체의 방향, 활동 라벨 등의 정보까지 포함된다[5]. 관절 위치 정보와 가려짐 여부를 정답으로하여 이미지를 모델에 입력했으며, 이때 2만 장을 학습에, 나머지 5000장을 검증에 사용했으며, 모델 이외에 모든 조건을 똑같이 설정하여 100번 반복학습했다.

정확도를 판단하기 위한 지표로는 PCKh@0.5를 사용했다. PCK는 관절의 예측 좌표와 정답 좌표의 차이가 임계값보다 작으면 정확히 예측했다고 판단하는 지표다. PCKh@0.5는 이미지에 존재하는 사람의 머리 길이의 0.5배를 임계값으로 설정한다. PCK는 각 관절마다 확인할 수 있으며, 이번 연구에서는 Pelvis로 평가 대상을 제한했다.

그림 2에서 자세 추정 결과 이미지를 확인할 수 있다. 학습은 100번 수행했으나, 70번 반복한 후에는 검증 정확도가 크게 변하지 않는 것을 볼 수 있었다. 따라서 이 지점을 기준으로 최적의 결과를 평가했을 때, 검증 데이터셋의 정확도는 0.8418로 나타났으며, PCKh@0.5는 0.8263으로 나타났다.

그림 2. Deep squat 영상 입력시 자세 추정 결과

그림 2의 inference 시간은 약 9.5초로, 실제 현장에서 인간이 과제를 평가할 때와 유사하거나, 빠른 속도로 추정한다는 것을 알 수 있으며, 육안으로 봤을 때, 실제 사람의 관절 위치와 가깝게 추정된 것을 확인했다. 근골격계 통증 및 질환 등의 원인으로, 인체의 잘못된 움직임과 부적절한 자세가 신체조직의 미세손상(microtrauma)으로 축적되고, 이러한 축적이 거대손상(microtrauma)으로 발현되는데, 신체의 움직임을 영상으로 정량화 하는 기술을 통해서 운동역학(kinematics)을 기반으로 운동재활 또는 스포츠의학 분야에서 널리 활용될 수 있다.

## 4. Acknowledgements

이 연구는 한국연구재단의 4단계 두뇌한국21 사업(4단계 BK21 사업) 과제의 지원을 받아 수행하였음. (II21SS7606007)

## 5.참고 문헌

[1] In-Yeong Lee,Hyun-Sung Leem, “The Effect of the Visual Stimulus on the Auditory Acuity.” The Korean Journal of Vision Science (*KSVS)*, 2016, Vol 18, No 1, p.49-56

[2] A. Toshev, et al, ″DeepPose: Human Pose Estimation via Deep Neural Networks″, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014

[3] Newell, A., Yang, K., Deng, J. “Stacked Hourglass Networks for Human Pose Estimation.” Computer Vision – ECCV 2016, 2016, pp. 483-499.

[4] Deep Residual Learning for Image Recognition, Kaiming He, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778

[5] Mykhaylo Andriluka, et al, “2D Human Pose Estimation: New Benchmark and State of the Art Analysis.” , Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014