# RGB 이미지와 Depth 이미지를 이용한 3D 휴먼 키포인트 탐지

*정근석 **이예지 ***윤경로
건국대학교대학원
*wjdrmstjrsla@naver.com, **leeyegi@gmail.com , ***yoonk@konkuk.ac.kr

# 3D Human Keypoint Detection With RGB and Depth Image

*Keunseok Jeong **Yegi Lee ***Kyoungro Yoon
Department of computer science and engineering, Konkuk University

## 요약

2019 발생한 COVID-19로 인하여 전 세계 사람들의 여가 활동이 제한되면서 건강관리를 위해 홈 트레이닝에 많은 관심을 기울이고 있다. 뿐만 아니라 최근 컴퓨팅 기술의 발전에 따라 사람의 행동을 눈으로 직접 판단했던 작업을 컴퓨터가 키포인트 탐지를 통해 인간의 행동을 이해하려는 많은 연구가 진행되고 있다.

이에 따라 본 논문은 Azure Kinect를 이용하여 촬영한 RGB 이미지와 Depth 이미지를 이용하여 3D 키포인트를 추정한다. RGB 이미지는 2D 키포인트 탐지기를 이용하여 2차원 공간에서의 좌표를 탐지한다. 앞서 탐지한 2D 좌표를 Depth 이미지에 투영하여 추출한 3D 키포인트의 깊이 값을 이용하여 3D 키포인트 탐지에 대한 연구 개발하였다.

## 1. 서론

현재 컴퓨팅 기술의 발전함에 따라서 인간의 행동을 눈으로 직접 판단하던 작업에서 키포인트 탐지 기술을 이용하여 사람의 행동을 이해하기 위한 연구를 활발히 진행되고 있는 중이다[1]. 또한, 2019년 중국 우한에서 시작한 COVID-19 바이러스로 인해 전 세계 국민의 취미 활동이 제안되면서 개개인의 건강관리를 위해 집에서 혼자서 운동할 수 있는 홈 트레이닝에 대한 관심이 집중되고 있고, 홈 트레이닝을 진행하기 위해 운동 자세를 코치해 주는 인공지능 코치를 개발하기 위해 키포인트 탐지가 필요하다.

2012년 알렉스넷 (AlexNet)이 등장한 이후에 이미지 분류 (Image Classification) 문제를 다룰 때, 심층신경망(Deep Neural Network)를 이용하여 기존 이미지 분류 문제를 다루기 위해 사용하던 방법보다 높은 정확도를 달성했다[2]. 이러한 성공으로 인하여 키포인트 탐지 분야에서도 마찬가지로 심층신경망을 적용하고자 하는 시도를 하였고, 그 결과 RGB(Red Green Blue) 이미지를 이용하여 높은 정확도와 속도로 2D 키포인트를 탐지할 수 있게 되었다. 하지만 RGB 이미지를 이용하여 탐지한 2D 키포인트는 인간의 행동을 이해하는 작업에서 3D 키포인트 보다 성능이 좋지 않다. 이러한 이유로, 인간의 행동을 잘 이해하기 위해서 3D 키포인트를 탐지하려는 연구가 활발하게 진행되고 있다[3]

본 논문은 Azure Kinect를 사용하여 RGB 이미지와 Depth 이미지를 촬영한다. 여기서 촬영된 RGB이미지와 Depth 이미지를 이용하여 3D 키포인트를 탐지하는 기법을 제시한다. 이 기법은 RGB 이미지를 이용하여 2D 좌표인 (X, Y)좌표를 추출하고, Depth 이미지를 이용하여 깊이 값(Z 좌표)를 추출한 후 이것들을 융합하여 3D 키포인트를 생성하여 각 관절에 해당하는 3D 키포인트를 탐지 한다.

본 논문의 구성은 다음과 같다. 제 2장에서 본 논문에서 제안하는 기법을 이용하여 3D 키포인트를 탐지하기 위한 배경지식 및 관련 연구에 대해 서술한다. 제 3장에서 본 논문에서 제안하는 3D 휴먼 키포인트 탐지 시스템(3D Human Keypoint Detection System)에 대한 구현방법에 대해 서술한다. 마지막으로 제 4장에서는 구현 결과에 대한 결론 및 향후 과제에 대해 서술한다.

## 2. 배경지식 및 관련 연구

Azure Kinect는 Microsoft사에서 음성 인식 작업이나 컴퓨터 비전같은 작업을 진행하기 위해서 고급 센서가 부착된 장치이다[4]. 원래 Kinect는 XBox의 게임용 인터페이스를 위해 활용되었다. 이후에 이를 발전시키면서 데스크탑과 연동 가능한 전용 드라이버를 제공하여, 음성 인식, 컴퓨터비전, 제스처 인식등 여러 분야에서 활용 가능한 장치로 발전하였다. Azure Kinect를 사용하기 위해서는 OS(Operating System)은 Windows10 64bit 이상, Ubuntu18.04 64bit 이상 버전에서 작동하기 때문에 Windows에서만 작동하던 기존 Kinect와 비교해 운영체제 선택이 자유로운 장점을 가지고 있다.

2D Human Keypoint Estimation은 이미지나 비디오 같은 영상에서 사람의 키포인트 위치를 추정하는 것으로 2D 휴먼 키포인트 탐지라고도 불린다. 최근 딥러닝 기술의 발전에 따라 키포인트 추정이 점점 정확해지는 추세에 따라 홈 트레이닝에서 운동 자세교정, CCTV에서 촬영된 영상을 이용하여 이상행동 감지등 다양한 분야에서 활발하게 연구가 진행되고 있다. 2D Human Keypoint Estimation을 수행하기 위한 대표적인 Framework는 Parts Based Framework와 Two Step Framework가 있다. Parts Based Framework는 인간의 신체 관절과

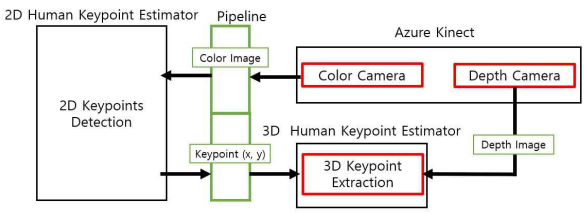
관절 사이의 연결 관계에 따라서 키포인트를 추정하는 방식으로 Bottom-Up Approach Method라고도 불린다[5]. Two-Step Framework는 2단계를 거쳐 인간의 신체 키포인트를 추정하는 Framework이다. 1단계에서는 이미지나 비디오 같은 영상에 사람이 있다면 사람이 위치하는 영역을 추출한다. 2단계에서는 1단계에서 사람이 위치하는 영역에 대해 키포인트를 탐지하는 심층신경망에 통과시켜 인간의 키포인트를 탐지한다[6].

3D Human Keypoint Estimation은 2D Human Keypoint Estimation과 비교했을 때 (X, Y) 좌표뿐만 아니라 Z 좌표까지 추정하여 3D 공간에서 각 키포인트의 (X, Y, Z) 좌표를 추정한다. 3D 키포인트를 추정하기 위해서는 2D 공간을 3D 공간으로 재구성하는 추가적인 작업이 필요하지만, 인간의 신체에 대해 2D 키포인트를 추정하는 것보다 3D 키포인트를 추정하는 것이 더 정확하게 자세를 추정할 수 있다는 장점이 있다. 또한, 인간의 행동인식, CCTV 감시 분야에서도 더 좋은 성능을 보일 수 있다. 3D 키포인트를 탐지하는 연구는 비디오를 이용하여 3D 키포인트를 추정하여 연속된 프레임마다 2D 키포인트를 추정한 후 연결될 수 있는 뼈의 길이를 이용하여 3D 키포인트를 추정하는 연구가 있다[7].

Point Cloud는 어떤 좌표계에 속해있는 Point들의 집합으로 정의된다[8]. 3D Coordinate System에서 Point는 일반적으로 (X, Y, Z)로 정의되며 Depth 이미지에 Point Cloud를 적용하면 2D Coordinate System을 3D Coordinate System으로 변환할 수 있다. 여기서 2D Coordinate System을 3D Coordinate System으로 변환하기 위해서는 카메라 렌즈에서 이미지 센서까지의 거리를 나타내는 Focal Length가 필요하다.

## 3. 구현

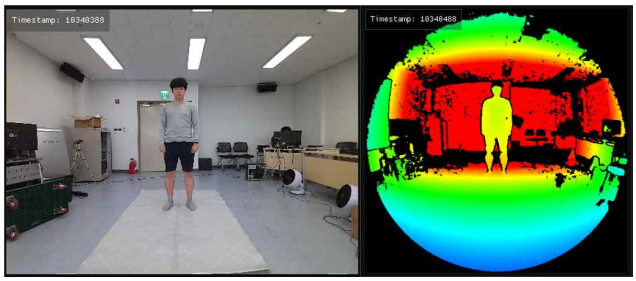
3D Human Keypoint Detection System의 구조도는 <그림 3-1>과 같다.




<그림 3-1> 3D Human Keypoint Detection System 구조도

위 구조도의 구성요소는 Azure Kinect, Pipeline, 2D Human Keypoint Estimator, 3D Human Keypoint Estimator이다. 각 구성요소의 역할은 다음과 같다. Azure Kinect는 RGB 이미지와 Depth 이미지를 촬영하는 역할을 한다. Pipeline은 C++와 Python 사이에 커뮤니케이션을 할 수 있도록 돕는 역할을 한다. 2D Human keypoint Estimator는 사전학습된 2D Human Keypoint Estimator를 이용하여 2D 키포인트를 탐지하는 역할을 한다. 여기서 2D Human Keypoint Estimator로 Alphapose를 사용한다. 3D Human Keypoint Estimator는 2D Human Keypoint Estimator에서 탐지한 (X, Y) 좌표를 Depth 이미지에 투영하여 Z 좌표 (깊이 값)을 추출한다. 추출한 깊이 값과 앞서 탐지한 2D 좌표를 융합하여 (X, Y, Z)좌표를 추출한다.

전체적인 흐름을 살펴보면 다음과 같다. Azure Kinect에서 RGB 이미지와 Depth 이미지를 촬영한다. 여기서 촬영된 Depth 이미지는 3D Human Keypoint Estimator로 전달되고 RGB 이미지는 파이프라인을 거쳐 2D Human Keypoint Estimator로 전달된다. 여기서 RGB 이미지 데이터는 파이프라인을 통해 2D Human Keypoint Estimator로 전달되기 위해서 데이터 직렬화 작업을 거친후 파이프라인을 통과한다. Azure Kinect를 이용하여 RGB이미지와 Depth 이미지를 촬영하면 <그림 3-2>와 같다.



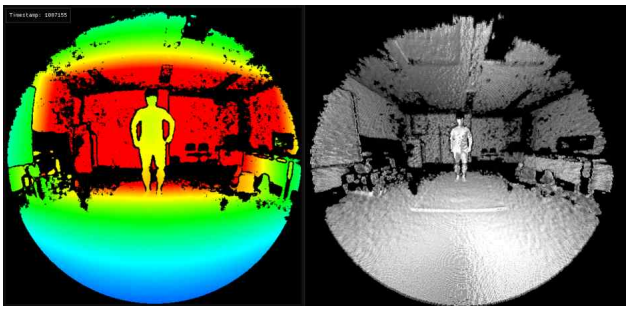

<그림 3-2> RGB 이미지와 Depth 이미지

2D Human Keypoint Estimator는 파이프라인을 통해 전달받은 RGB 이미지 데이터를 높이x너비x채널 형태를 가지는 이미지 데이터로 재구성한 후 사전학습된 2D Human Keypoint Estimator인 Alphapose에 통과시켜 2D 키포인트를 탐지한다. 탐지된 각 키포인트에 대한 2D 좌표는 직렬화 과정을 거친 후 파이프라인에 전달된다. 탐지된 2D 키포인트 좌표는 <그림 3-3>과 같다.

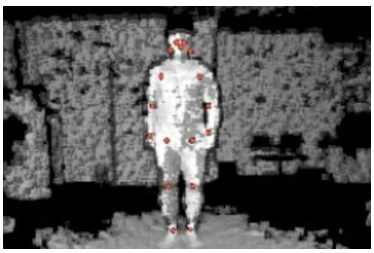

<그림 3-3> 2D 키포인트 추출

3D Human Keypoint Estimator는 전달받은 2D 좌표를 Depth 이미지에 투영한다. 여기서 RGB 이미지와 Depth 이미지의 Resolution이 서로 다르다. 이러한 이유로 Azure Kinect SDK에서 제공해주는 변환 함수를 사용하여 RGB 이미지에서의 (X, Y) 좌표를 Depth 이미지에서의 (X, Y)좌표로 변환한다. 여기서 추출한 깊이 값 (Z 좌표)를 위에서 추출한 (X, Y)와 융합하여 3D 키포인트를 탐지한다. 탐지한 좌표를 시각화하기 위해서 Depth 이미지에 Point Cloud를 적용하여 3D 공간을 <그

림 3-4>와 같이 생성한다. 그런 다음 각 키포인트에 해당하는 Point를 표현했다. 위 기법을 이용하여 탐지한 3D 키포인트는 <그림 3-5>과 같다.



<그림 3-4> Depth 이미지에 Point Cloud 적용한 3D 공간



<그림 3-5> 3D 키포인트 탐지

## 4. 결론 및 향후 과제

기존 3D Human Keypoint Estimation에 대한 연구는 모션 캡처 장비를 착용하여 데이터 세트를 제작한 후 심층신경망에 학습시켜 3D 키포인트를 탐지한다. 하지만 이 과정에서 많은 시간, 노력, 장비 구입비 등 많은 자원을 필요로 하기 때문에 본 논문에서는 RGB 이미지와 Depth 이미지를 이용한 3D Human Keypoint Estimation 기법을 제안한다.

2D Human Keypoint Estimator인 알파포즈를 이용하여 2D Human Keypoint에 대한 좌표를 추출한 후 이 좌표를 Depth 이미지에 투영하여 3D 좌표를 얻을 수 있었다. 이렇게 추정된 3D 키포인트 좌표의 위치를 확인하기 위해 Depth 이미지에 포인트 클라우드를 적용하여 3D 공간을 생성한 후 키포인트 위치를 확인하였을 때, 키포인트 위치는 사람의 표면에 위치하는 문제가 발생했다. 이러한 문제를 해결하기 위해서는 깊이 좌표를 조정할 수 있는 추가적인 실험이 필요하다.

위에서 발생한 문제를 해결하기 위해 깊이 좌표에 대한 조정하는 기법을 향후 과제로 제안한다. 예를 들어 Depth 이미지에 대해 포인트 클라우드를 적용한 3D 공간에서 인간의 키포인트에 대한 정답 좌표를 생성한 후 심층 신경망을 학습시키는 방법이다. 이 방법을 이용하면 인간의 신체 표면에 깊이 값이 위치하는 문제를 해결할 수 있을 것으로 보인다.

## 사사

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.20200021720012002, 인공지능을 이용한 맞춤형 홈트레이닝 플랫폼)

## 참고문헌

[1] Xiaoyang Wang and Qiang Ji「A Hierarchical Context Model for Event Recognition in Surveillance Video」, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014

[2] Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton「ImageNet Classification with Deep Convolutional Neural Networks」, Part of Advances in Neural Information Processing Systems 25 (NIPS), 2012

[3] Hanbyul Joo, Natalia Neverova and Andrea Vedaldi「Exemplar Fine-Tuning for 3D Human Model Fitting Towards In-the-Wild 3D Human Pose Estimation」, In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020

[4]Microsoft,「Azure Kinect Sensor SDK」, https://microsoft.github.io/Azure-Kinect-Sensor-SDK/master/index.html(2020.12.21)

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei and Yaser Sheikh「OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields」, IEEE, 2019

[6] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai and Cewu Lu「RMPE: Regional Multi-Person Pose Estimation」, In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016

[7] Dario Pavllo, Christoph Feichenholfer, David Grangier and Michael Auli「3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training」, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019

[8]「점구름」,https://ko.wikipedia.org/wiki/%EC%A0%90%EA%B5%AC%EB%A6%84, (2021.05.12.)