J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

RGB-D 정보를 이용한 2차원 키포인트 탐지 기반 3차원 인간 자세 추정 방법☆

A Method for 3D Human Pose Estimation based on 2D Keypoint Detection using RGB-D information

박 서 희[1] 지 명 근[2] 전 준 철[2*]
Seohee Park Myunggeun Ji Junchul Chun

요 약

최근 영상 감시 분야에서는 지능형 영상 감시 시스템에 딥 러닝 기반 학습 방법이 적용되어 범죄, 화재, 이상 현상과 같은 다양한 이벤트들을 강건하게 탐지 할 수 있게 되었다. 그러나 3차원 실세계를 2차원 영상으로 투영시키면서 발생하는 3차원 정보의 손실로 인하여 폐색 문제가 발생하기 때문에 올바르게 객체를 탐지하고, 자세를 추정하기 위해서는 폐색 문제를 고려하는 것이 필요하다. 따라서 본 연구에서는 기존 RGB 정보에 깊이 정보를 추가하여 객체 탐지 과정에서 나타나는 폐색 문제를 해결하여 움직이는 객체를 탐지하고, 탐지된 영역에서 컨볼루션 신경망을 이용하여 인간의 관절 부위인 14개의 키포인트의 위치를 예측한다. 그 다음 자세 추정 과정에서 발생하는 자가 폐색 문제를 해결하기 위하여 2차원 키포인트 예측 결과와 심층 신경망을 이용하여 자세 추정의 범위를 3차원 공간상으로 확장함으로써 3차원 인간 자세 추정 방법을 설명한다. 향후, 본 연구의 2차원 및 3차원 자세 추정 결과는 인간 행위 인식을 위한 용이한 데이터로 사용되어 산업 기술 발달에 기여 할 수 있다.

☞ 주제어 : 영상 감시, 객체 탐지, 키포인트 탐지, 인간 자세 추정, 딥 러닝

ABSTRACT

Recently, in the field of video surveillance, deep learning based learning method is applied to intelligent video surveillance system, and various events such as crime, fire, and abnormal phenomenon can be robustly detected. However, since occlusion occurs due to the loss of 3d information generated by projecting the 3d real-world in 2d image, it is need to consider the occlusion problem in order to accurately detect the object and to estimate the pose. Therefore, in this paper, we detect moving objects by solving the occlusion problem of object detection process by adding depth information to existing RGB information. Then, using the convolution neural network in the detected region, the positions of the 14 keypoints of the human joint region can be predicted. Finally, in order to solve the self-occlusion problem occurring in the pose estimation process, the method for 3d human pose estimation is described by extending the range of estimation to the 3d space using the predicted result of 2d keypoint and the deep neural network. In the future, the result of 2d and 3d pose estimation of this research can be used as easy data for future human behavior recognition and contribute to the development of industrial technology.

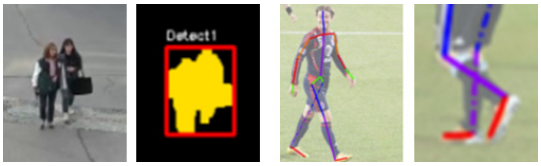
☞ keyword : Video Surveillance, Object Detection, Keypoint Detection, Human Pose Estimation, Deep Learning

1. 서 론

지능형 영상 분석은 CCTV(Closed Circuit Television)를 이용하여 영상 내 특성을 인식하고, 패턴을 추출함으로써 정보를 분석하는 분야이다. 최근 지능형 영상 분석 시스템은 효율적인 영상 감시(Video Surveillance)를 위해 딥 러닝(Deep Learning) 기반 학습 방법이 적용되어 다양하게 사전 정의된 이벤트를 강건하게 탐지함으로써 감시자에게 객체 탐지, 보행자 행위 예측과 같은 유용한 정보를 제공할 수 있게 되었다. 영상에서 인간의 행위를 인식하기 위해서는 움직이는 객체를 탐지하는 과정과 탐지된 인간의 자세를 추정하는 과정이 필요하다. 그러나 일반적으로 CCTV 영상은 3차원 실세계를 2차원 영상으로 투영

1 Human Care System Research Center, Korea Electronics Technology Institute(KETI), Seongnam, 13509, South Korea.
2 Department of Computer Science, Kyonggi University, Suwon, 16227, South Korea.
* Corresponding author (jcchun@kgu.ac.kr)
[Received 3 August 2018, Reviewed 21 August 2018(R2 4 October 2018), Accepted 10 October 2018]
☆ 본 논문은 2018년도 한국인터넷정보학회 춘계학술발표대회 우수 논문 추천에 따라 확장 및 수정된 논문임.
☆ 본 연구는 2018년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업 (No 2018R1D1A1B07042498)의 지원과, 2018학년도 경기대학교 대학원 연구원장학생 장학금 지원에 의하여 수행되었음.

시키면서 생기는 위상학적 정보의 손실 때문에 한 부분이 다른 부분에 의해 가려지는 폐색(Occlusion) 문제가 발생한다. 폐색 문제는 그림 1의 좌측과 같이 객체 탐지 과정에서 다른 보행자에 의해 가려짐으로써 발생하는 폐색[1]과 그림 1의 우측과 같이 자세 추정 과정에서 자신의 신체 부위에 의해 가려지는 자가 폐색(Self-occlusion)[2]으로 나눌 수 있다. 2차원 영상으로 객체를 탐지하고, 자세를 추정하기 위해서는 폐색 문제를 고려하는 것이 필요하다. 따라서 본 논문에서는 객체 탐지 과정에서 발생하는 폐색 문제와 자세 추정 과정에서 발생하는 자가 폐색 문제를 해결하기 위해 RGB-D 정보를 이용한 2차원 키포인트 탐지 기반 3차원 인간 자세 추정 방법을 설명한다.



(그림 1) 폐색 문제[1, 2]
(Figure 1) The problem of occlusion

2. 관련 연구

2.1 컴퓨터 비전 기반 객체 탐지

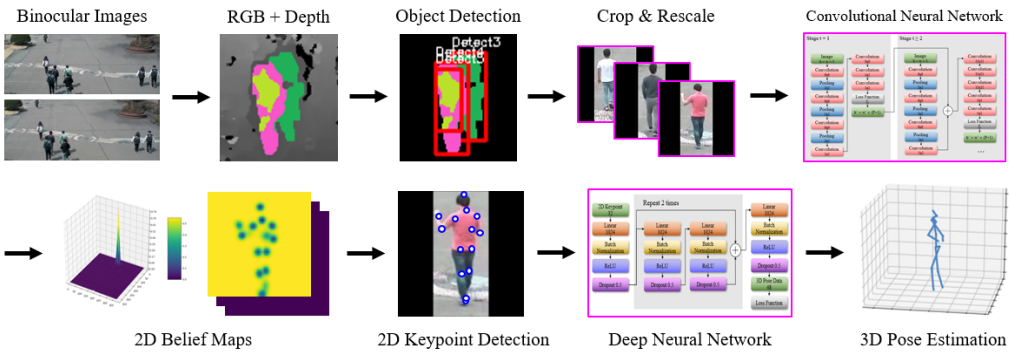
컴퓨터 비전 기반 객체 탐지는 영상 처리 기법을 이용하여 연속된 영상에서 관심 있는 객체를 추출하는 방법이다. 2차원 영상 기반의 객체 탐지 방법으로는 대표적으로 차 영상(Frame Differencing), 광류(Optical Flow), 배경 분리(Background Subtraction)를 이용한 방법이 있다[3]. 먼저 차 영상을 이용한 객체 탐지 방법은 두 개의 연속된 영상의 차이를 계산함으로써 움직이는 객체가 존재한다고 판단하는 방법이다. 광류를 이용한 객체 탐지 방법은 연속된 영상에서 2차원 모션 벡터(Motion Vector)를 이용하여 객체의 움직임을 탐지하는 방법이다. 마지막으로 배경 분리를 이용한 객체 탐지 방법은 가우시안 혼합 모델(Gaussian Mixture Model)을 이용하여 초기 배경을 모델링하고, 모델링 된 배경으로부터 움직이는 객체를 차 연산함으로써 객체를 탐지하는 방법이다[4].

이러한 2차원 영상 기반 객체 탐지 방법은 폐색 문제가 발생하여 객체를 정확하게 탐지 할 수 없는 한계를 가진다. 이와 같은 문제를 해결하기 위해 마이크로소프트사의 키넥트(Kinect) 장치 및 스테레오 비전(Stereo Vision)을 이용하여 깊이(Depth) 정보를 획득함으로써 객체를 탐지하는 연구들이 시도되어왔다[5]. 키넥트 장치는 깊이 센서를 이용하는 방식이며, 스테레오 비전은 에피폴라 기하학(Epipolar Geometry) 기반으로 같은 물체를 서로 다른 위치에서 촬영한 2개 이상의 영상 차이를 계산하는 방식이다[5]. 그러나 키넥트 장치는 인간 검출 거리가 제한되어 있기 때문에 비교적 원거리 영상을 처리하는 지능형 영상 감시 시스템에는 스테레오 비전으로 깊이 정보를 계산함으로써 객체를 탐지하는 것이 필요하다.

2.2 딥 러닝 기반 인간 자세 추정

인간의 자세를 추정하는 것은 신체의 구성을 추정하는 과정이기 때문에 인간의 관절인 키포인트의 위치를 예측하는 것이 필요하다. 인간의 자세는 스켈레톤 모델(Skeleton Model)로 표현되며, 추정 방식은 하향식(Top-down) 및 상향식(Bottom-up) 접근법으로 나눌 수 있다. 하향식 접근법은 경계 상자(Bounding Box)를 탐지하여 내부에서 자세를 추정하게 되고, 상향식 접근법은 영상으로부터 자세를 추정하기 위한 키포인트를 수집하고, 키포인트 간에 연관성을 계산하여 자세를 추정하는 방식이다. 아래와 같이 최근 Human3.6M[6], HumanEva[19] 등의 자세 데이터를 일련의 딥 러닝 신경망 구조를 통해 학습함으로써 2차원 및 3차원 키포인트 위치를 예측하는 자세 추정 연구가 활발히 진행되고 있다[6-14].

- LinKDE (2014) [6] : 기본적으로 Human3.6M[6]와 함께 제공되며, 단일 프레임 회귀(Regression)를 사용하여 KDE(Kernel Dependency Estimation)에 기반 한 예측을 나타내는 선형 푸리에 근사법이다.
- Tekin et al.(2016) [7] : 비디오의 연속된 프레임에서 컨볼루션 신경망(Convolutional Neural Network)과 모션 정보를 통해 3차원 자세를 복원하는 방법이다.
- Chen et al.(2017) [8] : 심층 신경망(Deep Neural Network)을 이용하여 단일 RGB 영상으로부터 2차원 자세 추정을 거치고 3차원 데이터와 정합하여 3차원 자세를 추정 하는 방법이다.
- Zhou et al.(2016) [9] : 단안(Monocular) 시퀀스로부터 EM(Expectation Maximization) 알고리즘과 컨볼루션 신경망을 통해 3차원 자세를 추정하는 방법이다.
- Du et al.(2016) [10] : 단안 RGB 카메라로 인간의 높이를 계산한 높이지도(Height-map)와 컨볼루션 신경망을 통해 3차원 자세를 추정하는 방법이다.



(그림 2) RGB-D 정보를 이용한 2차원 키포인트 탐지 기반 3차원 인간 자세 추정의 개요
(Figure 2) The overview of 3D human Pose Estimation based on 2D Keypoint Detection using RGB-D information

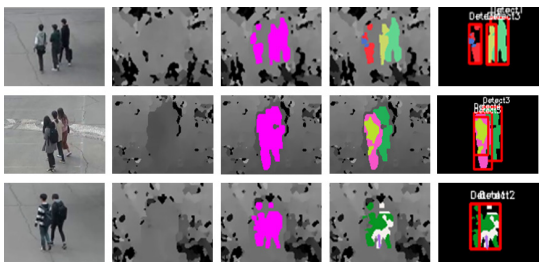
- Park et al.(2016) [11] : 컨볼루션 신경망을 통해 2차원 자세를 추정한 결과와 영상 특징을 이용하여 3차원 자세 추정을 수행하였으며, 루트 관절에서 멀리 떨어진 관절의 오류를 효과적으로 줄인 방법이다.
- Zhou et al.(2016) [12] : 운동학적 모델(Kinematic Model)과 컨볼루션 신경망을 이용하여 3차원 자세를 추정한 방법이다.
- Tome et al.(2017) [13] : 3차원 자세에 대한 확률론적 지식(Probabilistic Knowledge)을 다단계 컨볼루션 신경망에 융합하여 개선된 2차원 키포인트를 탐지하고 3차원 자세를 추정하는 방법이다.
- Martinez et al.(2017) [14] : 간단한 심층 신경망을 이용하여 2차원 키포인트와 3차원 실제 위치 값(Ground Truth)과의 정합을 학습시킴으로써 3차원 자세를 빠르게 추정한다.

최근 CMU의 Robotics Institute 연구팀은 단일 영상으로부터 여러 사람의 키포인트를 검출하는 실시간 시스템 OpenPose 라이브러리[15]를 발표하였다. 이는 컨볼루션 신경망 기반 순차적 예측 프레임워크인 컨볼루션 포즈머신(Convolution Pose Machine)[16]으로 모델을 학습하여 키포인트를 탐지한다. 따라서 본 연구에서는 객체 탐지 과정에서 나타나는 폐색 문제를 해결하기 위해 스테레오 비전을 통해 RGB-D 정보를 획득함으로써 객체를 강건하게 탐지하고, 하향식 방식을 채택하여 탐지된 영역에서 컨볼루션 포즈머신[16]을 통해 2차원 키포인트를 탐지하는 연구를 수행한다. 또한 2차원 키포인트 탐지 과정에서 나타나는 자가 폐색 문제를 해결하기 위하여 심층 신경망을 통해 2차원 키포인트를 3차원 공간상으로 확장시킴으로써 신체의 구성을 추정하는 방법을 설명한다.

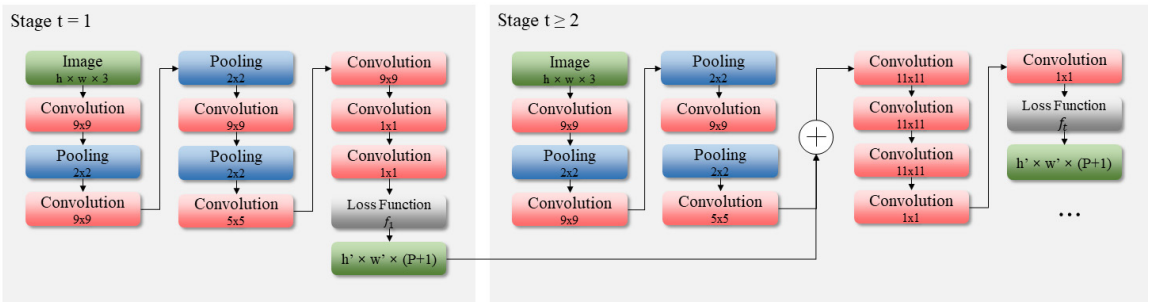
3. RGB-D 정보를 이용한 2차원 키포인트 탐지 기반 3차원 인간 자세 추정 방법

지능형 영상 감시 시스템에서 인간의 행위를 인식하기 위해서는 객체 탐지 및 자세 추정 과정에서 나타나는 폐색 문제를 해결하여 객체를 강건하게 검출할 필요가 있다. 본 논문에서는 이러한 문제를 해결하기 위하여 그림 2와 같이 양안 영상(Binocular Images)을 이용하여 깊이 정보를 계산하고 RGB 정보와 병합하여 객체를 탐지한다. 그 다음 탐지된 경계 상자 영역을 컨볼루션 포즈머신의 입력으로 설정하여 신뢰 지도(Belief Map)들을 반환함으로써 2차원 키포인트 탐지를 수행한다. 탐지된 키포인트 정보는 심층 신경망의 입력으로 설정되어 3차원 인간 자세 추정을 수행하게 된다. 본 논문에서 제안하는 방법을 RGB-D 정보 기반 객체 탐지, 컨볼루션 신경망 기반 2차원 키포인트 탐지, 심층 신경망 기반 3차원 인간 자세 추정으로 나누어 설명한다.

3.1 RGB-D 정보 기반 객체 탐지



(그림 3) RGB-D 정보 기반 객체 탐지
(Figure 3) Object Detection based on RGB-D information

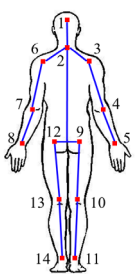


(그림 4) 2차원 키포인트 탐지를 위한 컨볼루션 신경망 구조[16]
(Figure 4) The structure of Convolutional Neural Network for 2D Keypoint Detection

본 연구에서는 두 대의 CCTV 영상을 이용하여 객체 탐지를 수행한다. 먼저 왼쪽 단안 영상을 기반으로 배경으로부터 움직이는 객체를 1차적으로 분할하는 과정을 거친다. 가우시안 혼합 모델을 이용하여 모델링 된 배경 영상으로부터 변화하는 영역을 차 연산 하여 움직이는 객체를 탐지한다[4]. 그 다음, RGB 기반으로 탐지된 결과에 깊이 정보를 추가하여 2차적으로 분할하기 위해 양안 영상을 이용하여 스테레오 비전 기반의 깊이지도(Depth Map)를 생성한다. 깊이지도는 두 영상 간 픽셀의 차이를 통해 영상의 유사도를 계산하는 블록 정합을 수행하며, 이는 1차원 명암 값을 가진다[5]. RGB 정보만으로 분할된 영역 내부에서 깊이 값을 차례대로 탐색하여 이전 픽셀 값과 현재 픽셀 값을 비교하고, 일정한 깊이 값 범위를 갖는 픽셀끼리 군집화(Clustering)를 수행하여 2차 분할을 수행한다. RGB-D 정보 기반 객체 탐지 결과는 그림 3에 나타나 있다. 그림 3의 좌측부터 원 영상, 깊이지도, RGB 정보 기반 분할, RGB-D 정보 기반 분할, 객체 탐지 결과이다[1].

3.2 컨볼루션 신경망 기반 2차원 키포인트 탐지

본 연구에서는 키포인트 위치를 예측하기 위해 객체 탐지 과정으로부터 계산된 영역 내부에서 컨볼루션 신경망을 기반으로 자세를 추정하는 하향식 접근을 수행한다. 또한 그림 5와 같이 14개의 2차원 키포인트를 가지는 스켈레톤 모델을 사용하였다. 먼저 각 다른 크기로 탐지된 경계 상자 영역을 재조정하여 컨볼루션 신경망의 입력으로 설정한다. 그 다음 2차원 키포인트 예측 결과를 나타내는 N개의 신뢰 지도들을 반환하는 컨볼루션 포즈머신[16]을 이용하여 키포인트를 추정한다. 컨볼루션 신경망 구조는 그림 4를 따른다.



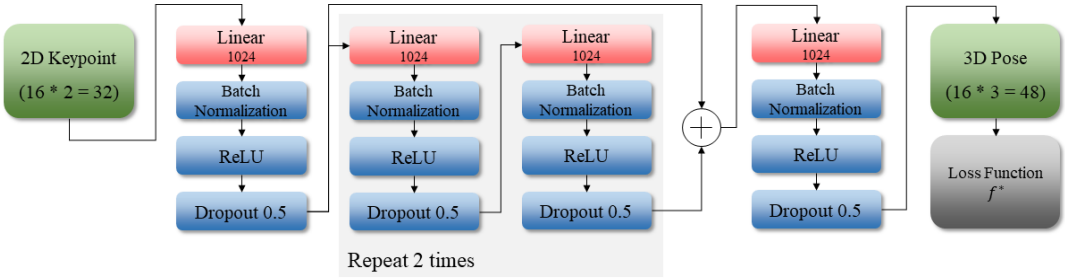
1 Head
2 Neck
3 Right Shoulder
4 Right Elbow
5 Right Wrist
6 Left Shoulder
7 Left Elbow
8 Left Wrist
9 Right Hip
10 Right Knee
11 Right Ankle
12 Left Hip
13 Left Knee
14 Left Ankle

(그림 5) 스켈레톤 모델
(Figure 5) Skeleton Model

본 연구에서는 키포인트에 대한 실제 값 주석이 달린 MPII 자세 데이터 세트를 이용하여 사전 훈련된 모델을 사용하였다. 컨볼루션 포즈머신[16]은 기존 포즈머신(Pose Machine)[17]의 자세 추정을 컨볼루션 신경망 구조로 구현한 것이며, 키포인트인 $g_t(\cdot)$의 위치를 예측하는 두 단계로 구성된다. 첫 번째 단계에서는 영상 특징을 기반으로 각 키포인트 위치의 신뢰도(Confidence) 추정치를 산출한다. 키포인트의 픽셀 위치는 $p-th$이고, Z는 모든 키포인트 P에 대한 위치 $Y=(Y_1,\ldots,Y_p)$의 순서쌍 (u,v)의 집합이다. 각 단계 $t\in\{1,\ldots,T\}$에서 위치 분류 g_t의 각 키포인트의 위치를 할당하기 위한 신뢰도는 픽셀 위치 z에 있는 영상으로부터 추출된 특징을 기반으로 예측된다. 여기서 X_z는 픽셀 위치 Z를 중심으로 하는 T단계에 대한 영상 패치의 특징 벡터이다. 첫 번째 단계인 $t=1$에서 신뢰도 값은 다음 식 1에 의해 산출된다.

$$g_1(X_z)=\left\{b_1^p(Y_P=z)\right\}_{p\in\{0,\ldots,P\}} \quad (1)$$

b_1^p는 위치 분류 g_1에 의해 예측된 신뢰 값이다. 키포인트 p의 신뢰도 값은 영상의 모든 위치 $z=(u,v)^T$에서 계산된다. 그 다음, 특징함수 ψ를 통해 이전 단계의 신뢰도에 대한 특징과 특징 벡터를 사용하여 신뢰도를



(그림 6) 3차원 인간 자세 추정을 위한 심층 신경망 구조[14]
(Figure 6) The structure of Deep Neural Network for 3D Human Pose Estimation

정제한다. 식 2와 같이 영상 위치 z의 특징 함수 ψ는 각 키포인트 위치에 대한 신뢰 지도를 입력으로 사용하고 신뢰 지도 위치에서 추출된 특징을 생성한다.

$$g_t(X_z^{'}, \psi(z, b_{t-1})) = \{b_t^p(Y_p = z)\}_{p\in\{0,\ldots,P+1\}} \quad (2)$$

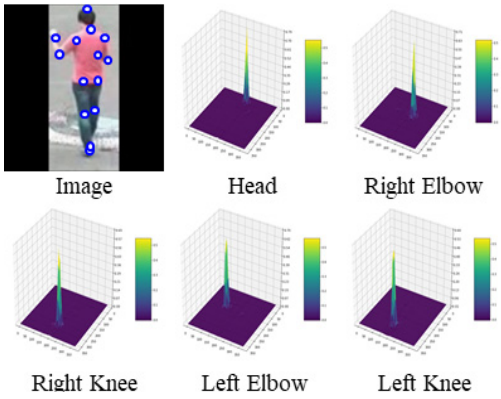
컨볼루션 포즈머신[16]의 신경망 구조는 입력 영상에서 특징을 추출하는 컨볼루션 계층과 추출된 특징을 서브 샘플링(Sub-sampling)하는 풀링(Pooling) 계층으로 구성된다. 의미 있는 출력 값을 추출하기 위한 활성화 함수는 ReLU(Rectified Linear Unit) 함수를 사용한다. 컨볼루션 신경망 구조는 각 단계마다 신뢰 지도를 생성하는 t 단계로 구성되어 있으며, 이전 단계의 예측 결과는 다음 단계의 입력으로 사용된다. 그러므로 단계 t에서는 모든 키포인트에 대한 위치를 예측하기 위해 반복적으로 신뢰 지도를 생성한다. 가장 이상적인 신뢰 지도 $b_*^p(Y_p = z)$는 각 키포인트 실제 위치에 가우시안 피크(Gaussian Peak)를 위치시킴으로써 생성된다. 또한 손실 함수(Loss Function)를 식 3과 같이 정의함으로써 단계를 반복할수록 예측된 키포인트 위치와 이상적인 키포인트 위치를 나타내는 신뢰 지도 간의 거리 l_2 값을 최소화하여 정확도가 향상 될 수 있도록 한다. 이러한 신뢰 지도 간의 거리 값을 각 키포인트와 배경에 대해 구하여 총 합인 f_t를 구한다.

$$f_t = \sum_{p=1}^{P+1}\sum_{z\in Z} \| b_t^p(z) - b_*^p(z) \|_2^2 \quad (3)$$

이러한 손실 함수를 이용하여 식 4와 같이 각 단계에서 손실을 더함으로써 키포인트 위치가 산출된다.

$$F = \sum_{t=1}^{T} f_t \quad (4)$$

2차원 키포인트는 컨볼루션 신경망에 의해 사전 학습된 모델을 사용하여 탐지되며, 컨볼루션 신경망에 입력되는 영상은 368×368 픽셀의 해상도로 정규화 되어 입력되고, 필터와 스트라이드(Stride)가 2×2 크기로 맥스 풀링(Max Pooling) 단계를 거쳐서 8배 만큼 다운 샘플링(Down-sampling)된다. 다운 샘플링을 통해 최종적으로 특징 지도(Feature Map)를 산출하게 된다. 컨볼루션 신경망을 통해 산출된 각 신뢰 지도에 표시된 지점은 생성된 신뢰도의 최대 높이 값을 뜻하며, 주어진 픽셀에서 키포인트가 얼마나 확실하게 예측되었는지를 나타낸다. 신뢰 지도를 나타내기 위한 신뢰 분포도에서 x, y축은 입력 영상의 해상도를 뜻하며 z축은 신뢰도 값을 나타낸다. 한 사람에 대한 신뢰 분포도 일부는 다음 그림 7과 같다.



(그림 7) 신뢰 분포도
(Figure 7) Distribution plot of belief

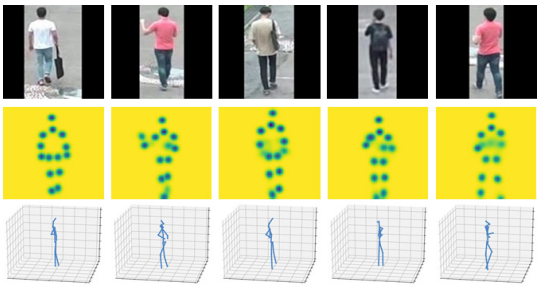
3.3 심층 신경망 기반 3차원 인간 자세 추정

2차원 키포인트 좌표에 대응하는 3차원 좌표는 무한하기 때문에 3차원 인간 자세 추정은 어려운 문제 중 하나이다. 본 연구에서는 비교적 정보량이 적기 때문에 작업

하기 유용한 2차원 키포인트 좌표를 입력으로 설정하여 간단한 심층 신경망을 통해 3차원 공간상의 자세 추정을 수행한다[14]. 예측된 신뢰도 값의 최댓값을 식 5를 통해 계산하여 키포인트 좌표를 산출한다.

$$Y_p = \arg max\, b_p[u, v] \tag{5}$$

본 연구에서 사용된 심층 신경망의 입력 데이터는 2차원 키포인트 좌표 $x \in R^{2n}$들로 구성되며, 심층 신경망에 의해 산출되는 데이터는 3차원 자세 좌표 $y \in R^{3n}$들로 구성된다. 또한 임의의 좌표 공간에서 본래의 3차원 키포인트 위치와 유사하게 추정함으로써 누락된 3차원 데이터를 복구하기 위해 카메라 프레임으로부터 전역 좌표계를 고정하여 훈련된 모델을 사용한다. 심층 신경망은 Human3.6M 데이터 세트[6]의 16개의 키포인트를 가지는 스켈레톤 모델을 이용하여 사전 훈련되었으며, $2n$ 크기인 32개의 2차원 입력 데이터는 선형 계층을 거쳐 1024개로 증가되고, 그 다음 선형 계층에서 최종 예측 값 산출 전에 적용되어 $3n$ 크기인 48개의 3차원 자세 데이터를 산출한다[14]. 또한 활성화 함수인 ReLU 함수를 사용하여 활성화 값을 조절하며, 활성화 함수의 출력 값들을 정규화 하기 위한 배치 정규화(Batch Normalization) 기법과 가중치 중 일부만 사용하는 드롭아웃(Dropout) 기법을 사용한다. 선형 계층은 두 번 반복되며, 두 블록은 스킵 연결(Skip Connection)을 수행함으로써 두 개의 가중치 계층 다음에 입력 값을 그대로 더해주는 잔여 연결(Residual Connection) 방법으로 연결됨으로써 총 6개의 선형 계층으로 구성된다. 3차원 인간 자세 추정을 위한 심층 신경망 구조는 그림 6에 나타나 있다.



(그림 8) 3차원 인간 자세 추정
(Figure 8) 3D Human Pose Estimation

이러한 심층 신경망을 통해 3차원 자세 데이터는 식 6의 손실 함수 $f^* : R^{2n} \rightarrow R^{3n}$를 이용하여 실제 값과 예측된 값 사이의 예측 오차를 최소화함으로써 계산된다. 식 6의 x_i값은 주어진 카메라 매개변수 하에 실제 2차원 키포인트 위치 좌표를 사용하여 계산된다.

$$f^* = \min \frac{1}{N} \sum_{i=1}^{N} L(f(x_i) - y_i) \tag{6}$$

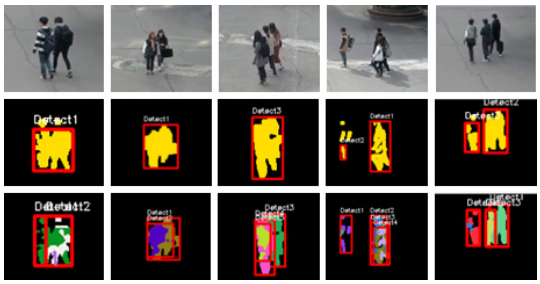
본 연구에서는 Human3.6M 데이터 세트[6]를 이용하여 훈련된 모델을 사용하였고, 훈련 시 전체 데이터에 대한 한 번의 학습을 의미하는 에폭(Epoch)은 200, 학습 속도(Learning Rate)는 0.001, 64 크기의 미니 배치 값을 이용하였다[14]. RGB-D 정보를 이용한 객체 탐지 기반 3차원 인간 자세 추정 결과는 그림 8과 같으며, 상단부터 차례대로 재조정 된 입력 영상, 2차원 키포인트 탐지, 3차원 인간 자세 추정의 결과를 나타낸다.

4. 실험결과

(표 1) 실험 환경
(Table 1) Experimental Environments

구분	세부 환경
CPU	AMD Ryzen 7 1700 3GHz
GPU	NVIDIA GeForce GTX 1080 Ti
RAM	32.00 GB
OS	Windows 10
Camera	Hanwha Techwin SNO-6084R
Resolution	800×450
Language	C, C++, Python 3.0
Develop Tool	Visual Studio 2015, Jupyter Notebook
Library	OpenCV, TensorFlow

4.1 객체 탐지 방법 비교



(그림 9) 객체 탐지 결과 비교
(Figure 9) Comparison of Results of Object Detection

(표 2) Human3.6M 데이터 세트[6]를 이용한 3차원 인간 자세 추정 결과 비교 (관절 위치 오류 당 평균)
(Table 2) Comparison of results of 3D Human Pose Estimation using Human3.6M (MPJPE)

Model	Direction	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
LinKDE [6]	132.7	183.6	132.4	164.4	162.1	205.9	150.6	171.3
Tekin et al. [7]	102.4	147.2	88.8	125.3	118.0	182.7	112.4	129.2
Chen et al. [8]	89.9	97.6	90.0	107.9	107.3	139.2	93.6	136.1
Zhou et al. [9]	87.4	109.3	87.1	103.2	116.2	143.3	106.9	99.8
Du et al. [10]	85.1	112.7	104.9	122.1	139.1	135.9	105.9	166.2
Park et al. [11]	100.3	116.2	90.0	116.5	115.3	149.5	117.6	106.9
Zhou et al. [12]	91.8	102.4	96.7	98.8	113.4	125.2	90.0	93.8
Tome et al. [13]	65.0	73.5	76.8	86.4	86.3	110.7	68.9	74.8
Martinez et al. [14]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1
Ours	**59.5**	**86.9**	**73.5**	**88.6**	**93.6**	**116.3**	**62.9**	**109.9**

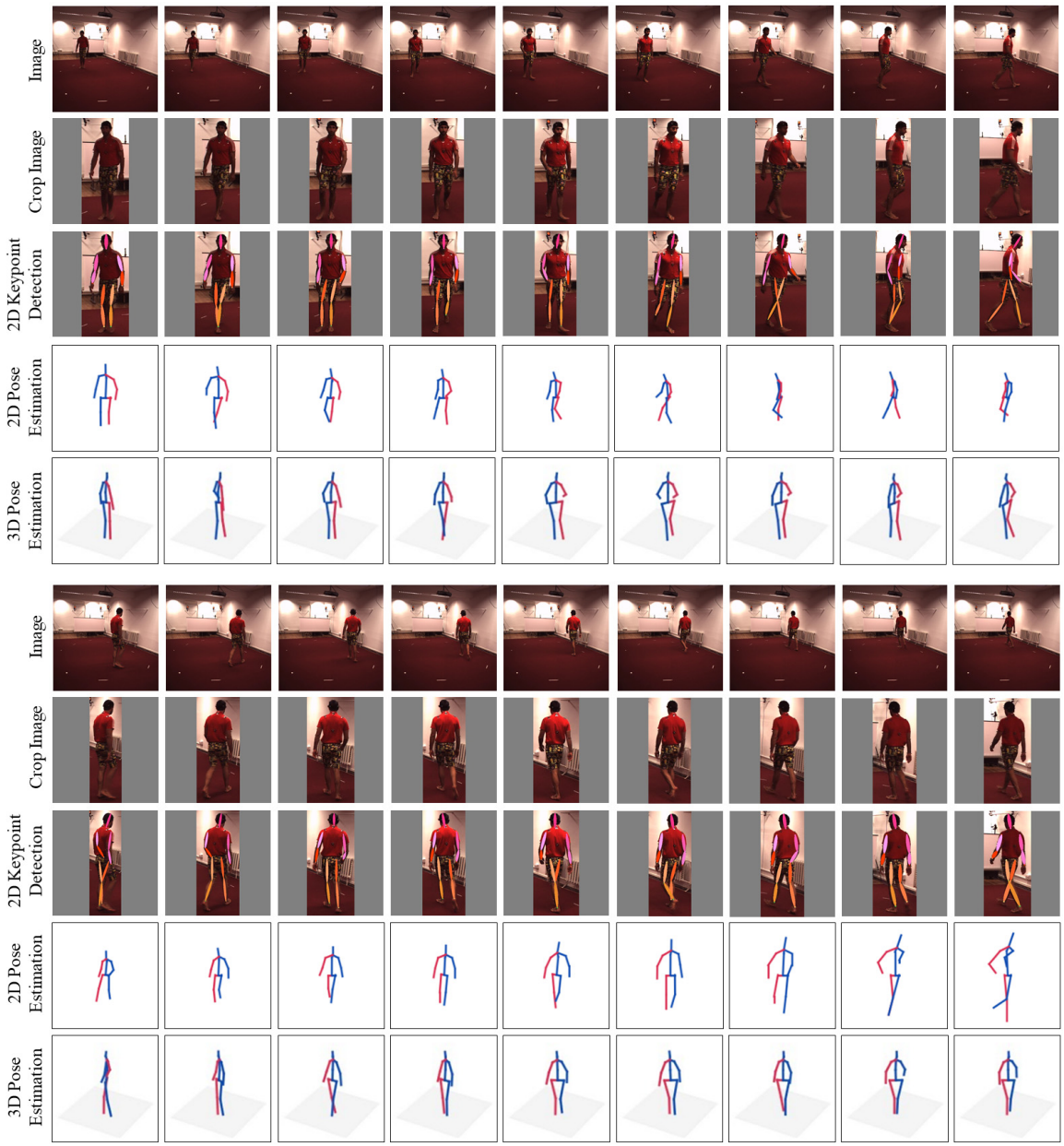
Model	Sitting	Sitting Down	Smoking	Waiting	Walk Dog	Walking	Walk Together	Average
LinKDE [6]	151.6	243.0	162.1	170.7	177.1	96.6	127.9	162.1
Tekin et al. [7]	138.9	224.9	118.4	138.8	126.3	55.1	65.8	125.0
Chen et al. [8]	133.1	240.1	106.7	106.2	114.1	87.0	90.6	114.2
Zhou et al. [9]	124.5	199.2	107.4	118.1	114.2	79.4	97.7	113.0
Du et al. [10]	117.5	226.9	120.0	117.7	137.4	99.3	106.5	126.5
Park et al. [11]	137.2	190.8	105.8	125.1	131.9	62.6	96.2	117.3
Zhou et al. [12]	132.2	159.0	107.0	94.4	126.0	79.0	99.0	107.3
Tome et al. [13]	110.2	173.2	85.0	85.8	86.3	71.4	73.1	88.4
Martinez et al. [14]	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Ours	**140.0**	**247.0**	**86.4**	**87.2**	**104.6**	**56.1**	**58.5**	**98.1**

본 연구의 객체 탐지 방법 비교는 기존 RGB 정보만을 이용한 경우와 RGB-D 정보를 이용한 경우로 나누어 실험하였다. 그림 9는 상단부터 차례대로 원 영상, RGB 정보만을 이용하여 객체를 탐지한 결과, 본 연구에서 제안하는 RGB-D 정보를 이용한 객체 탐지 결과이다. 기존 RGB 정보만을 이용하여 객체를 탐지하게 될 경우, 3차원 정보 부족으로 인해 객체 사이에 폐색 현상이 발생하여 객체가 하나로 탐지되는 문제가 발생한다. 본 논문에서는 이러한 폐색 현상을 고려하여 RGB-D 정보 기반의 객체 탐지를 수행한 결과, 폐색 문제로 인하여 두 개 이상의 객체가 하나의 객체로 탐지되었던 기존의 문제점을 해결하여 객체를 강건하게 분할하였다. 또한 움직이는 객체를 올바르게 계수 할 수 있음을 확인했으므로 지능형 영상 감시 시스템의 객체 계수 기능으로 구현이 가능 하다.

4.2 3차원 인간 자세 추정 방법 비교

본 연구에서 Human3.6M 데이터 세트[6]를 이용하여 벤치마크 함으로써 RGB-D 정보를 이용한 객체 탐지 기반의 3차원 인간 자세 추정 결과를 다른 3차원 인간 자세 추정 결과들과 비교하였다. Human3.6M 데이터 세트[6]는 360만개의 자세에 대한 정보와 함께 비디오 형식의 데이터로 제공된다. 또한 6명의 남자와 5명의 여자가 15개의 시나리오에 대한 가변성 있는 동작들을 수행하며 스켈레톤 모델과 실제 좌표 값을 제공한다. 이러한 시나리오는 Directions, Discussing, Eating, Greeting, Phone Call, Posing, Purchases, Sitting, Sitting Down, Smoking, Taking Photo, Waiting, Walking, Walking Dog, Walking Together 으로 구성된다. 본 연구에서는 이러한 시나리오들이 포함된 학습 및 테스트를 위해 나누어진 11가지 대상(Subject)들 중 S9, S11을 이용하여 자세 추정 결과를 평가하였다. S9, S11 데이터 세트에는 15가지 시나리오 및 초당 50 프레임을 가지는 256개의 동영상으로 구성된다.

본 연구에서는 비디오를 프레임 별로 분할하여 입력 영상을 추출한 후, Human3.6M 데이터 세트[6]에서 제공하는 깊이 정보인 TOF(Time of Flight) 데이터를 이용하여 RGB-D 정보 기반으로 객체를 탐지하였다. 탐지된 영상을 재조정하여 2차원 키포인트를 추정하고, 산출된 좌표들을 심층 신경망에 입력하여 3차원 인간 자세 추정을 수



(그림 10) Human3.6M 데이터 세트를 이용한 3차원 인간 자세 추정 결과
(Figure 10) The result of 3D Human Pose Estimation using Human3.6M dataset

행한 결과는 그림 10에 나타나있다. 그림 10은 Human3.6M 데이터 세트[6] 중 Walking 시나리오에 대해 2차원 및 3차원 자세 추정을 수행한 시퀀스 결과이며, 신체 왼편은 적색, 오른편은 청색으로 표시하였다.

3차원 인간 자세 추정의 비교 방식은 인간의 관절 위치의 실제 값과 추정된 예측 값 사이의 관절 위치 오류당 평균(Mean Per Joint Position Error, MPJPE)을 밀리미터(Millimeter, mm) 단위로 측정하여 평가한다. Human3.6M

데이터 세트[6]를 이용하여 본 연구의 3차원 자세 추정 결과를 다른 3차원 인간 자세 추정 연구들과 비교한 결과는 표 2에 나타나있다.

본 연구에서 3차원 인간 자세 추정을 수행한 결과 관절 평균 오류의 평균은 98.06mm 라는 결과를 얻었다. 또한 기존 Human3.6M 데이터 세트[6]의 벤치마크인 LinKDE[6]와 Tekin et al.의 연구[7], Chen et al.의 연구[8], Zhou et al.의 연구[9], Du et al.의 연구[10], Park et al.의 연구[11], Zhou et al.의 연구[12] 보다 관절 평균 오류가 낮게 측정된 것을 확인 할 수 있다. 비교된 연구들 중, 컨볼루션 포즈머신[16]을 사용하여 2차원 키포인트를 산출하고, 3차원 자세 라이브러리를 사용하여 3차원 인간 자세 추정을 수행한 Chen et al.의 연구[8] 보다 16.1mm 차이만큼 우수한 성능을 보였다. 또한 RGB와 깊이 영상을 컨볼루션 신경망의 입력으로 사용하여 2차원 키포인트를 산출하고, 3차원 모션을 추정한 Du et al.의 연구[10]보다 28.4mm 차이만큼 우수한 성능을 보였다. 그리고 본 연구와 유사하게 2차원 관절 위치 정보를 이용하여 3차원 자세를 추정한 Park et al.의 연구[11]보다 19.2mm 차이만큼 더 나은 성능을 보였다.

그러나 Tome et al.의 연구[13]와 Martinez et al.의 연구[14] 보다는 오류 측면에서 낮은 성능을 보인다는 것을 확인하였다. Tome et al.의 연구[13]는 컨볼루션 포즈머신[16] 기반으로 산출된 2차원 신뢰 지도에 3차원 키포인트 위치에 대한 확률론적 지식을 추가하여 기존 예측된 신뢰 값을 수정하는 과정을 거치기 때문에 추정된 3차원 자세 결과가 개선됨을 보였다. 또한 Martinez et al.의 연구[14]는 컨볼루션 포즈머신[16]과 비슷한 결과를 산출하는 누적 모래시계 신경망[18]을 기반으로 2차원 키포인트를 예측하였다. 또한 심층 신경망을 이용하여 3차원 자세를 추정하였으며, 추가적으로 신경망의 매개변수를 미세하게 조정하는 파인튜닝(Fine-tuning)을 수행하여 오류를 감소시켰다.

하지만 본 연구에서는 Human3.6M 데이터 세트를 이용하여 배치 사이즈(Batch Size) 1개 기준으로 평균 실행되는 시간을 계산 했을 때, 3차원 인간 자세 추정 모듈은 0.039초 정도 소요되었다. Tome et al., Martinez et al.의 연구에서 오류 측면에서는 낮은 성능을 보였지만, 지능형 영상 감시 시스템을 위한 속도 측면에서는 실시간 시스템에 적용 시킬 수 있는 가능성을 보였다. 따라서 본 연구에서는 향후 파인 튜닝을 수행하여 관절 위치 오류 당 평균을 감소시킴으로써 3차원 인간 자세 추정 결과를 개선할 필요가 있다. 향후 이러한 추정 방법을 이용하여 지능형 영상 감시 시스템에 실시간으로 적용 시킬 수 있다.

5. 결 론

본 연구에서는 3차원 실세계가 2차원 영상 정보로 투영되면서 발생하는 3차원 정보의 손실로 인한 폐색 문제를 객체 탐지 과정의 폐색과 자세 추정 과정의 자가 폐색으로 구분하여 두 가지 폐색 문제를 해결하기 위한 연구를 진행하였다. 객체 탐지 과정에서 발생하는 폐색을 해결하기 위해 RGB-D 정보 기반의 객체 탐지를 수행함으로써 객체를 강건하게 분할하여 탐지하고, 올바르게 계수할 수 있음을 확인하였다. 자세 추정과정에서 발생하는 자가 폐색 문제를 해결하기 위해 2차원 키포인트를 심층 신경망을 통해 3차원 공간상으로 확장함으로써 본래 3차원 자세와 유사하게 추정하고, 올바른 키포인트 데이터 복구를 시도 하였다.

향후 연구로는 객체 탐지 과정에서 탐지된 객체의 정보는 객체 추적 연구로의 확장이 가능하며, 지능형 영상 분석 시스템에 보행자 침입 및 출입 탐지, 이동 방향 탐지, 객체 카운팅 등의 기능으로 구현이 가능하다. 또한 2차원 키포인트 탐지를 통한 3차원 인간 자세 추정 과정에서 컨볼루션 신경망과 심층 신경망의 파인 튜닝을 통해 관절 평균 오류를 최소화함으로써 3차원 키포인트 좌표를 실제 인간 자세와 유사하게 추정하여 정확도를 개선하는 연구가 추가적으로 필요하다. 산출된 2차원 키포인트는 3차원 자세 추정을 위한 용이한 단서로 제공 될 수 있으며, 최종 산출된 자세 데이터는 인간 행위 인식을 위한 데이터로 사용될 수 있으므로 확장 연구를 통해 지능형 영상 분석 시스템, 의료 분야의 행위 예측 기술을 이용한 환자의 행위 분석, 자율 주행 자동차에서의 보행자 탐지 분야 등에 적용되어 산업 기술 발달에 기여 할 수 있다. 또한, 추정된 자세 데이터를 기반으로 3차원 그래픽스 모델 생성으로의 확장 연구를 통해 자세를 유연하게 표현함으로써 게임, AR(Augmented Reality) 및 VR(Virtual Reality)과 같은 분야에서 응용 할 수 있다.

참고문헌(Reference)

[1] Seohee Park, Myunggeun Ji, and Junchul Chun, "2D Human Pose Estimation based on Object Detection using RGB-D information", KSII Transactions on Internet & Information Systems, Vol. 12, No. 2, pp. 800-816, 2018. https://doi.org/10.3837/tiis.2018.02.015

[2] Ramakrishna, Varun, Takeo Kanade, and Yaser Sheikh, "Reconstructing 3d human pose from 2d image landmarks", European conference on computer vision. Springer, Berlin, Heidelberg, pp. 573-586, 2012. https://doi.org/10.1007/978-3-642-33765-9_41

[3] Parekh, Himani S., Darshak G. Thakore, and Udesang K. Jaliya, "A survey on object detection and tracking methods", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, No. 2, pp. 2970-2978, 2014. http://www.ijircce.com/upload/2014/february/7J_A%20Survey.pdf

[4] Zivkovic, Zoran, "Improved adaptive Gaussian mixture model for background subtraction", Pattern Recognition, 2004. https://doi.org/10.1109/icpr.2004.1333992

[5] Hirschmuller, Heiko, "Stereo processing by semiglobal matching and mutual information", IEEE Transactions on pattern analysis and machine intelligence, Vol. 30, No. 2, pp. 328-341, 2008. https://doi.org/10.1109/tpami.2007.1166

[6] Ionescu, Catalin, et al, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments", IEEE transactions on pattern analysis and machine intelligence, Vol. 36, No. 7, pp. 1325-1339, 2014. https://doi.org/10.1109/tpami.2013.248

[7] Tekin, Bugra, et al, "Direct prediction of 3d body poses from motion compensated sequences", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. https://doi.org/10.1109/cvpr.2016.113

[8] Chen, Ching-Hang, and Deva Ramanan, "3d human pose estimation = 2d pose estimation + matching", CVPR, Vol. 2, No. 5, 2017. https://doi.org/10.1109/cvpr.2017.610

[9] Zhou, Xiaowei, et al, "Sparseness meets deepness: 3D human pose estimation from monocular video", Proceedings of the IEEE conference on computer vision and pattern recognition, 2016. https://doi.org/10.1109/cvpr.2016.537

[10] Du, Yu, et al, "Marker-less 3d human motion capture with monocular image sequence and height-maps", European Conference on Computer Vision. Springer, Cham, 2016. https://doi.org/10.1007/978-3-319-46493-0_2

[11] Park, Sungheon, Jihye Hwang, and Nojun Kwak, "3D human pose estimation using convolutional neural networks with 2D pose information", European Conference on Computer Vision. Springer, Cham, 2016. https://arxiv.org/abs/1608.03075

[12] Zhou, et al, "Deep kinematic pose regression", European Conference on Computer Vision. Springer, Cham, 2016. https://arxiv.org/abs/1609.05317

[13] Tome, Denis, Christopher Russell, and Lourdes Agapito, "Lifting from the deep: Convolutional 3d pose estimation from a single image", CVPR 2017 Proceedings, pp. 2500-2509, 2017. https://doi.org/10.1109/cvpr.2017.603

[14] Martinez, et al, "A simple yet effective baseline for 3d human pose estimation", International Conference on Computer Vision, Vol. 1, No. 2. 2017. https://doi.org/10.1109/iccv.2017.288

[15] OpenPose: A Real-Time Multi-Person Keypoint Detection and Multi-Threading C++ Library, 2017.

[16] Wei, Shih-En, et al, "Convolutional pose machines", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. https://doi.org/10.1109/cvpr.2016.511

[17] Ramakrishna, Varun, et al, "Pose machines: Articulated pose estimation via inference machines", European Conference on Computer Vision. Springer, Cham, 2014. https://doi.org/10.1007/978-3-319-10605-2_3

[18] Newell, Alejandro, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation", European Conference on Computer Vision. Springer, Cham, 2016. https://doi.org/10.1007/978-3-319-46484-8_29

[19] Sigal, Leonid, et al, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion", International journal of computer vision, 2010. https://doi.org/10.1007/s11263-009-0273-6

◐ 저 자 소 개 ◑

박 서 희(Seohee Park)

2017 B.S. in Computer Science, Kyonggi University, Suwon, South Korea
2018 M.S. in Computer Science, Kyonggi University, Suwon, South Korea.
2018~Present : Researcher of the Human Care System Research Center in Korea Electronics Technology Institute(KETI)
Research Interests : Computer Vision, Deep Learning, Human Pose Estimation
E-mail : eehoeskrap@keti.re.kr

지 명 근(Myunggeun Ji)

2017 B.S. in Computer Science, Kyonggi University, Suwon, South Korea
2017~Present : M.S. Student in Computer Science, Kyonggi University, Suwon, South Korea
Research Interests : Computer Vision, Augmented Reality
E-mail : jmg2968@kgu.ac.kr

전 준 철(Junchul Chun)

1984 B.S. in Computer Science, Chung-Ang University, Seoul, South Korea
1986 M.S. in Computer Science(Software Engineering), Chung-Ang University, Seoul, South Korea
1992 M.S. in Computer Science and Engineering(Computer Graphics), The Univ. of Connecticut, USA
1995 Ph.D. in Computer Science and Engineering(Computer Graphics), The Univ. of Connecticut, USA
2001.02~2002.02 Visiting Scholar, Michigan State Univ. Pattern Recognition and Image Processing Lab.
2009.02~2010.02 Visiting Scholar, Univ. of Colorado, Wellness Innovation and Interaction Lab.
Research Interests : Augmented Reality, Computer Vision, Human Computer Interaction
E-mail : jcchun@kgu.ac.kr