일반논문 (Regular Paper)
방송공학회논문지 제30권 제3호, 2025년 5월 (JBE Vol.30, No.3, May 2025)
https://doi.org/10.5909/JBE.2025.30.3.437
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

1인칭 영상으로부터의 실시간 3차원 전신 자세 추정을 위한 단일 단계 방법

나 소 연[a)], 장 주 용[a)‡]

One-Stage Method for Real-time 3D WholeBody Pose Estimation from Egocentric Images

Soyeon Na[a)] and Ju Yong Chang[a)‡]

요 약

1인칭 영상으로부터의 3차원 휴먼 자세 추정 연구는 머리 착용형 카메라로부터 얻어지는 1인칭 영상에서 사용자의 3차원 관절 좌표를 추정하는 것을 목표로 한다. 해당 연구는 어안 카메라로 인한 1인칭 영상의 왜곡 보정, 실시간 추정 성능 향상이 도전 과제이다. 기존의 1인칭 영상으로부터의 3차원 전신 자세 추정을 위한 모델은 손과 몸을 따로 추정한 뒤 결합하는 두 단계 방법으로 손 검출 성능이 저하되면 손 자세 추론 성능도 저하된다. 이를 해결하기 위해 손 검출 성능에 영향을 받지 않고 관절 사이의 관계 정보를 활용할 수 있는 단일 단계 기반의 방법을 제안한다. 제안하는 방법을 평가하기 위해 SceneEgo Hand Occlusion 데이터셋을 구축하고, 제안 방법이 손 가리워짐에 강인한 손 자세 추정 성능을 보임을 확인한다. 또한, 자세 추정 속도 측정을 통해 실시간 활용 가능성을 보인다.

Abstract

Research on 3D human pose estimation from egocentric images aims to estimate the 3D joint coordinates of a user from egocentric images captured by a head-mounted camera. This research faces challenges such as correcting distortions in egocentric images caused by fisheye cameras and improving real-time estimation performance. Existing models for 3D wholebody pose estimation from egocentric images typically adopt a two-stage approach, where hand and body poses are estimated separately and then combined. However, in this method, a decline in hand detection performance leads to a degradation in hand pose estimation accuracy. To address this issue, we propose a one-stage approach that is not affected by hand detection performance and can leverage the relational information between joints. To evaluate the proposed method, we construct the SceneEgo Hand Occlusion dataset and demonstrate that our approach achieves robust hand pose estimation even under occlusion. Furthermore, we assess pose estimation speed to validate its potential for real-time applications.

Keyword : Egocentric wholebody pose estimation, Graph convolution network, Deep learning

Ⅰ. 서 론

1인칭 영상 기반의 3차원 휴먼 자세 추정(egocentric 3D human pose estimation)은 머리 착용형 카메라(head mounted camera)로부터 얻어지는 1인칭 영상을 기반으로 카메라를 착용하고 있는 사람의 3차원 관절 좌표를 추정하는 것을 목표로 한다. 최근 가상 현실(virtual reality; VR), 증강 현실(augmented reality; AR), 혼합 현실(mixed reality; MR) 산업의 발전과 함께 1인칭 영상으로부터의 휴먼 자세 추정 연구가 크게 주목받고 있다. 1인칭 영상은 주로 어안 카메라(fisheye camera)를 사용하여 획득되는데, 이로 인해 심한 왜곡과 하체가 상체에 가려지는 현상이 빈번히 나타난다. 이는 일반적인 3인칭 영상에 비해 1인칭 영상에서의 휴먼 자세 추정 성능 저하를 초래하는 주요 원인이 된다. 또한, VR/AR 기기에 해당 기술을 효과적으로 활용하기 위해서는 실시간 추정 속도가 요구된다. 따라서 해당 연구에서는 어안 카메라로 인한 왜곡 보정과 실시간 추정 성능 향상이 중요한 도전 과제이다.

[1]에서는 1인칭 영상을 입력으로 하여 전신에 대한 3차원 관절 좌표를 추정하는 모델인 EgoWholeBody가 제안되었다. 이 모델은 손과 몸의 관절 좌표를 개별적으로 추정하는 두 단계 방법(two-stage method)이라고 할 수 있다. 이 방법은 검출된 손 영역의 크기를 확장하여 손 자세 추정을 위한 입력으로 활용한다. 따라서 영상에서 손의 영역이 몸의 영역보다 작아서 손 관절 좌표를 추정하기 어려운 문제가 완화된다.

그러나 왜곡이 심한 1인칭 영상에서는 일반적으로 손 검출 성능이 저하된다. 이러한 손 검출 성능의 저하는 손 영역을 정확하게 잘라내지 못하게 한다. 이는 잘려진 손 영역을 입력으로 사용하여 손 관절 좌표를 추정하는 두 단계 방법의 손 관절 추정 성능을 저하시킨다. 또한, EgoWholeBody는 3차원 히트맵(3D heatmap) 기반의 방법을 사용한다. 3차원 히트맵의 추정은 일반적으로 많은 계산량을 필요로 하고, 이는 모델의 추정 속도를 느리게 만들어 실시간 처리가 어렵다는 한계점을 가져온다. 따라서 본 논문에서는 1인칭 영상에서 손 검출 성능 저하로 인한 문제와 모델의 추정 속도를 개선하기 위해 전신의 관절 좌표를 한 개의 네트워크로 추론하는 단일 단계 방법(one-stage method) 기반의 3차원 전신 자세 추정 방법을 제안한다. 제안하는 모델은 3차원 히트맵을 사용하지 않고, 관절 좌표를 직접 회귀(direct-regression)함으로써 모델의 추론 속도를 향상시키고, 이를 통해 실시간 처리가 가능하다. 또한, 그래프 합성곱 신경망(graph convolutional network; GCN)으로 하여금 관절 사이의 관계 정보를 효과적으로 학습하게 함으로써 개선된 관절 좌표를 출력하도록 한다.

본 연구에서는 제안하는 방법을 평가하기 위해 기존의 데이터셋뿐만 아니라 추가적으로 SceneEgo Hand Occlusion 데이터셋을 구축하고, 손 가리워짐(occlusion) 상황에서의 실험을 통해 제안 방법에 대한 정량적, 정성적 평가를 수행하였다. 실험 결과에 따르면 제안하는 방법은 기존 두 단계 방법보다 손 가리워짐 환경에 대해 강인한 손 자세 추정 성능을 보인다. 또한 추론 속도 실험을 통해 제안 방법의 실시간 활용 가능성을 입증하였다.

Ⅱ. 관련 연구

1. 1인칭 영상 기반 3차원 휴먼 자세 추정

1인칭 영상 기반 3차원 휴먼 자세 추정 연구의 목표는 머리 착용형 카메라에서 촬영된 사용자의 영상을 입력으로 하여 사용자의 3차원 자세를 추정하는 것이다. 대부분의 기존 연구는 히트맵 기반의 방법이며, 이는 비교적 높은 정확도를 제공하지만 추정 속도의 실시간성 측면에서는 한계를 가진다. Mo2Cap2[2]는 1인칭 단안 영상에서 하체가 잘 보이지 않는 문제를 해결하고자 하였다. Mo2Cap2는 하체를

a) 광운대학교 전자통신공학과(Department of Electronics and Communications Engineering, Kwangwoon University)
‡ Corresponding Author : 장주용(Ju Yong Chang)
E-mail: juyong.chang@gmail.com
Tel: +82-2-940-5136
ORCID: https://orcid.org/0000-0003-3710-7314
※ This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2023-00219700, Development of FACS-compatible Facial Expression Style Transfer Technology for Digital Human, 90%) and the Excellent researcher support project of Kwangwoon University in 2024 (10%).
· Manuscript March 25, 2025; Revised April 21, 2025; Accepted April 21, 2025.

중심으로 확대한 1인칭 영상을 추가적인 입력으로 사용하여 2차원 히트맵을 추정하고 카메라로부터 각각의 관절 사이의 절대 거리를 계산하여 3차원 자세를 추정한다. [3]에서는 1인칭 영상에서 3차원 휴먼 자세 추정을 위한 인코더-디코더 구조 기반 모델인 xR-EgoPose가 제안되었다. xR-EgoPose는 1인칭 영상에서 상체와 하체 사이의 해상도 차이를 보정한다. 하지만 Mo2Cap2와 xR-EgoPose는 시야 밖 관절이나 심각한 가리워짐 상황에서 자세 추정 성능이 저하된다. [4]에서는 1인칭 스테레오 영상을 입력으로 하여 신체의 가리워짐 문제를 해결하고자 U-Net 구조의 네트워크를 통해 의사 팔다리 마스크(pseudo limb mask)와 2차원 히트맵을 활용하는 모델인 EgoGlass가 제안되었다. [5]에서는 어안 카메라 왜곡 문제를 해결하기 위해 카메라 파라미터와 3차원 자세를 동시에 추정하는 모델이 제안되었다. [5]는 자동 어안 카메라 왜곡 보정으로 자세 추정 성능을 향상시켰으나 실제 환경에서의 성능이 평가되지 않았다. UnrealEgo[6]는 스테레오 정보를 활용하기 위해 가중치 공유 인코더(weight-sharing encoder)를 사용하여 자세 추정 성능을 향상시키지만 상대적으로 복잡한 자세에서는 성능이 저하된다. SceneEgo[7]는 3차원 특징 복셀(feature voxels)과 배경 깊이(scene depth)를 활용하여 3차원 히트맵을 생성하고 이를 기반으로 휴먼 자세를 추정한다. 하지만 SceneEgo는 깊이 추정 네트워크의 성능에 많은 영향을 받는다는 한계점이 있다. Ego3DPose[8]는 팔다리 히트맵(limb heatmap)을 사용하여 UnrealEgo보다 향상된 성능을 보였지만 실제 환경에서의 평가 결과를 제시하지 않았다. [9]는 1인칭 영상에서 자기 가리워짐(self-occlusion)과 시야 밖 관절(out-of-view limbs) 문제를 해결하기 위해 처음으로 비전 트랜스포머(vision transformer) 기반의 3차원 휴먼 자세 복원 방법인 EgoTAP을 제안하였다. EgoCoord[10]는 카메라 캘리브레이션 정보(camera calibration data) 없이도 학습 가능한 왜곡 임베딩(learnable distortion embedding)과 3차원 히트맵을 활용하여 3차원 휴먼 자세를 추정한다. 하지만 히트맵 기반의 방법은 상대적으로 큰 모델 크기로 모델의 추정 속도가 느려 실시간 처리가 어렵다. 이를 해결하기 위해 EgoPoseFormer[11]는 스테레오 영상을 입력으로 하여 히트맵을 사용하지 않고 다층 퍼셉트론(multi-layer perceptron; MLP)을 통해 3차원 초기 자세를 추정하고 트랜스포머(transformer) 기반의 모듈을 통해 자세를 보정한다. 본 논문에서는 1인칭 단안 영상을 입력으로 실시간 자세 추정이 가능한 MLP 및 GCN 기반의 방법을 제안한다.

기존 연구들과 달리 [1]에서는 1인칭 영상을 입력으로 하여 몸과 손의 자세를 모두 추정하는 3차원 전신 자세 추정 모델인 EgoWholeBody가 제안되었다. EgoWholeBody는

그림 1. EgoWholeBody[1]의 손 검출 결과
Fig. 1. Hand detection results of EgoWholeBody[1]

몸과 손의 관절 좌표를 각각 추정한 뒤 결합하는 두 단계 방법에 기반한다. 하지만 1인칭 영상은 왜곡으로 손 검출이 어렵고 손 검출 성능의 저하는 손 자세 추정 성능을 저하시킨다. 그림 1은 기존 방법[1]의 잘못된 손 검출 결과를 보여준다. 본 논문에서는 두 단계 방법의 이러한 단점, 즉 1인칭 영상에서 손 검출 성능 저하로 인한 문제를 해결하기 위해 전신의 관절 좌표를 하나의 네트워크로 추정하는 단일 단계 기반의 3차원 전신 자세 추정 방법을 제안한다.

2. 그래프 합성곱 신경망 기반의 3차원 휴먼 자세 추정

그래프 합성곱 신경망 기반의 3차원 자세 추정 방법들은 일반적으로 2차원 자세를 입력으로 하여 그래프 구조로 표현된 신체 정보를 기반으로 3차원 자세를 출력한다. 이러한 방법들은 신체를 그래프 구조로 표현하기 위해 각 관절을 노드(node)로 활용하고 관절 사이의 연결을 엣지(edge)로 활용한다. [12]에서는 입력 영상에서 획득한 의미론적 특징(semantic feature)을 그래프 합성곱에 함께 사용하는 방법인 SemGCN이 제안되었다. Modulated GCN[13]은 weight modulation과 affinity modulation을 사용하여 작은 모델 크기를 유지하면서 기존 모델의 성능을 개선한다. 하지만 SemGCN과 Modulated GCN은 이웃 노드 정보만 고려한다는 한계점을 가진다. [14]에서는 이러한 한계점을 해결하고자 먼 거리의 관절 간 관계도 학습할 수 있는 방법인 FlexGCN이 제안되었는데, 이를 통해 가리워짐이나 깊이 모호함(depth ambiguity) 문제가 완화됨을 보였다. 본 논문에서는 초기 자세를 보정하는 그래프 합성곱 신경망 기반의 자세 보정 모듈을 제안한다. 기존 방법은 2차원 관절 좌표를 3차원 관절 좌표로 리프팅(lifting)하기 위해 GCN을 사용했지만 제안하는 방법은 초기 3차원 관절 좌표를 입력으로 하여 개선된 3차원 관절 좌표를 출력하기 위해 GCN을 사용한다.

Ⅲ. 제안하는 방법

본 연구에서 제안하는 방법은 자세 초기화(pose initialization) 모듈과 GCN 기반의 자세 보정(pose refinement) 모듈로 구성된다. 그림 2는 본 논문에서 제안하는 방법의 개요를 보여준다. 자세 초기화 모듈의 입력은 휴먼 객체를 포함하는 1인칭 영상이고 출력은 3차원 초기 전신 관절 좌표 $P \in \mathbb{R}^{55\times3}$이다. GCN 기반의 자세 보정 모듈의 입력은 3차원 초기 전신 관절 좌표 P와 영상 특징 벡터 F_G이고 최종 출력은 개선된 3차원 전신 관절 좌표 $P' \in \mathbb{R}^{55\times3}$이다.

1. 자세 초기화 모듈

1인칭 영상이 주어지면 자세 초기화 모듈은 3차원 초기

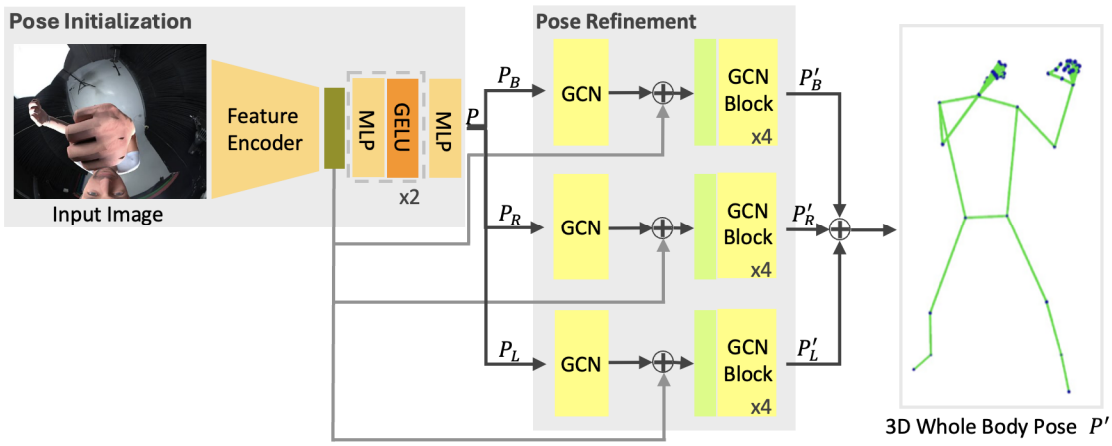

그림 2. 제안하는 방법의 개요
Fig. 2. Overview of the proposed method

전신 관절 좌표 P를 출력한다. 먼저, ResNet50[15] 인코더를 통해 입력 영상에서 영상 특징 맵(image feature map) $F \in \mathbb{R}^{2048\times8\times8}$을 출력한다. 영상 특징 맵 F는 전역 평균 풀링(global average pooling)을 통해 영상 특징 벡터 $F_G \in \mathbb{R}^{2048}$로 변환된다. 전역 평균 풀링으로 변형된 영상 특징 벡터 F_G는 3층의 MLP와 GELU 활성화 함수[16]로 구성된 네트워크를 통과하여 3차원 초기 전신 관절 좌표 P를 출력한다.

2. 자세 보정 모듈

자세 보정 모듈은 3차원 초기 전신 관절 좌표 P와 영상 특징 벡터 F_G를 입력으로 하여 개선된 3차원 전신 관절 좌표 P'를 출력한다. 자세 보정 모듈은 3개의 네트워크로 구성되며 각각의 네트워크는 modulated GCN[13]을 기반으로 한다. 표 1은 각각의 네트워크에서 사용되는 관절 좌표들에 대한 요약된 정보를 제시한다. 첫 번째 네트워크는 몸 관절을 보정한다. P의 몸 관절 좌표 $P_{Body} \in \mathbb{R}^{15\times3}$와 왼손 중수지(metacarpophalangeal; MCP) 관절 좌표 $P_{LH\_MCP} \in \mathbb{R}^{5\times3}$, 오른손 중수지 관절 좌표 $P_{RH_MCP} \in \mathbb{R}^{5\times3}$를 포함하는 $P_B \in \mathbb{R}^{25\times3}$을 입력으로 한다. 출력은 개선된 3차원 몸 관절 좌표 $P'_B \in \mathbb{R}^{25\times3}$이다. 두 번째 네트워크는 오른손을 보정하기 위한 네트워크로 P의 오른손 관절 좌표 $P_{RH} \in \mathbb{R}^{20\times3}$, 오른쪽 손목 관절 좌표 $P_{RW} \in \mathbb{R}^{1\times3}$, 그리고 오른쪽 팔꿈치 관절 좌표 $P_{RE} \in \mathbb{R}^{1\times3}$을 포함하는 $P_R \in \mathbb{R}^{22\times3}$을 입력으로 하여 개선된 3차원 관절 좌표 $P'_R \in \mathbb{R}^{22\times3}$를 출력한다. 세 번째 네트워크는 왼손을 보정하기 위한 네트워크로 P의 왼손 관절 좌표 $P_{LH} \in \mathbb{R}^{20\times3}$, 왼쪽 손목 관절 좌표 $P_{LW} \in \mathbb{R}^{1\times3}$, 그리고 왼쪽 팔꿈치 관절 좌표 $P_{LE} \in \mathbb{R}^{1\times3}$을 포함하는 $P_L \in \mathbb{R}^{22\times3}$를 입력으로 하여 개선된 3차원 관절 좌표 $P'_L \in \mathbb{R}^{22\times3}$를 출력한다. 최종 출력 $P' \in \mathbb{R}^{55\times3}$는 P'_B에서 손의 중수지 관절 좌표를 제외한 15개의 몸 관절 좌표와 P'_R, P'_L에서 각각 손목과 팔꿈치 좌표를 제외한 20개의 손 관절 좌표를 결합한 것이다. 제안 방법에서는 팔꿈치, 손목, 중수지 관절이 3개의 네트워크에서 공통적으로 사용된다. 그림 3은 이러한 공통 관절을 시각적으로 보여준다.

표 1. 자세 보정 모듈에서 사용되는 관절 좌표 표기
Table 1. Notations of joint coordinates used in the pose refinement module

Notation	Description	Dimension
P	Initial wholebody joint coordinates	$\mathbb{R}^{55\times3}$
P'	Refined wholebody joint coordinates	$\mathbb{R}^{55\times3}$
Body joint refinement network		
P_{Body}	Body joint coordinates of P	$\mathbb{R}^{15\times3}$
P_{LH_MCP}	Left-hand MCP joint coordinates of P	$\mathbb{R}^{5\times3}$
P_{RH_MCP}	Right-hand MCP joint coordinates of P	$\mathbb{R}^{5\times3}$
P_B	$P_{Body}+P_{LH\_MCP}+P_{RH\_MCP}$	$\mathbb{R}^{25\times3}$
P'_B	Refined joint coordinates of P_B	$\mathbb{R}^{25\times3}$
Right-hand joint refinement network		
P_{RH}	Right-hand joint coordinates of P	$\mathbb{R}^{20\times3}$
P_{RW}	Right wrist joint coordinates of P	$\mathbb{R}^{1\times3}$
P_{RE}	Right elbow joint coordinates of P	$\mathbb{R}^{1\times3}$
P_R	$P_{RH}+P_{RW}+P_{RE}$	$\mathbb{R}^{22\times3}$
P'_R	Refined joint coordinates of P_R	$\mathbb{R}^{22\times3}$
Left-hand joint refinement network		
P_{LH}	Left-hand joint coordinates of P	$\mathbb{R}^{20\times3}$
P_{LW}	Left wrist joint coordinates of P	$\mathbb{R}^{1\times3}$
P_{LE}	Left elbow joint coordinates of P	$\mathbb{R}^{1\times3}$
P_L	$P_{LH}+P_{LW}+P_{LE}$	$\mathbb{R}^{22\times3}$
P'_L	Refined joint coordinates of P_L	$\mathbb{R}^{22\times3}$

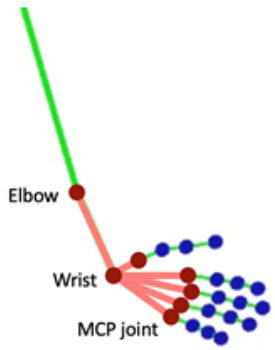

그림 3. 공통 관절
Fig. 3. Common joints

그림 4는 자세 보정 모듈의 네트워크 구조를 보여준다. 이는 1층의 입력 그래프 합성곱 신경망, 배치 정규화(batch normalization)[17], RELU[18] 활성화 함수, 4개의 그래프 합

성곱 신경망 블록(GCN Block), 그리고 1층의 출력 그래프 합성곱 신경망으로 구성된다. 그래프 합성곱 신경망 블록은 2층의 그래프 합성곱 신경망, 배치 정규화, 그리고 RELU 활성화 함수로 구성된다. 입력 그래프 합성곱 신경망은 3차원 초기 관절 좌표 P_B, P_R, P_L 입력으로 하여 그래프의 노드 특징(node feature) $H_B \in \mathbb{R}^{128\times25}$, $H_R \in \mathbb{R}^{128\times22}$, $H_L \in \mathbb{R}^{128\times22}$을 출력한다. 제안하는 방법에서는 입력 영상의 정보를 활용하기 위해 영상 특징 벡터 F_G를 노드 특징과 합쳐서 그래프 합성곱 신경망 블록의 입력으로 사용한다. 보다 구체적으로, 영상 특징 벡터 F_G는 2층의 1차원 합성곱(1D convolution)을 통과하여 512차원으로 축소된다. 이후 차원 축소된 특징 벡터를 25번 반복하여 $F_{G\_B} \in \mathbb{R}^{512\times25}$, 22번 반복하여 $F_{G\_R} \in \mathbb{R}^{512\times22}$, $F_{G_L} \in \mathbb{R}^{512\times22}$를 생성한다. 그 후 각각의 노드 특징과 연결(concatenate)하여 $H_{B\_T} \in \mathbb{R}^{640\times25}$, $H_{R\_T} \in \mathbb{R}^{512\times22}$, $H_{R\_T} \in \mathbb{R}^{640\times22}$, $H_{L_T} \in \mathbb{R}^{640\times22}$를 생성한다. 노드 특징 H_{B_T}, H_{R_T}, H_{L_T}는 관절 사이의 연결 정보를 나타내는 스켈레톤 그래프를 포함하고 있는 4개의 그래프 합성곱 신경망 블록에 의해, 개선된 노트 특징 H'_{B_T}, H'_{R_T}, H'_{L_T}으

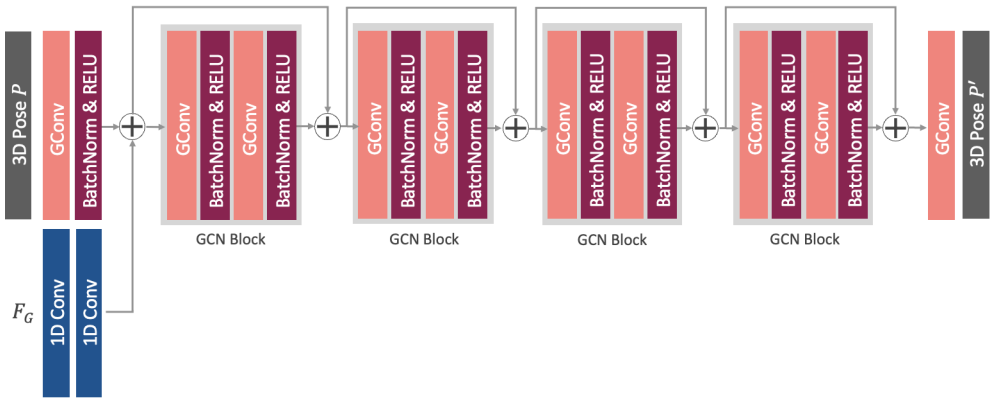

그림 4. 자세 보정 모듈의 네트워크 구조
Fig. 4. Network architecture of pose refinement module

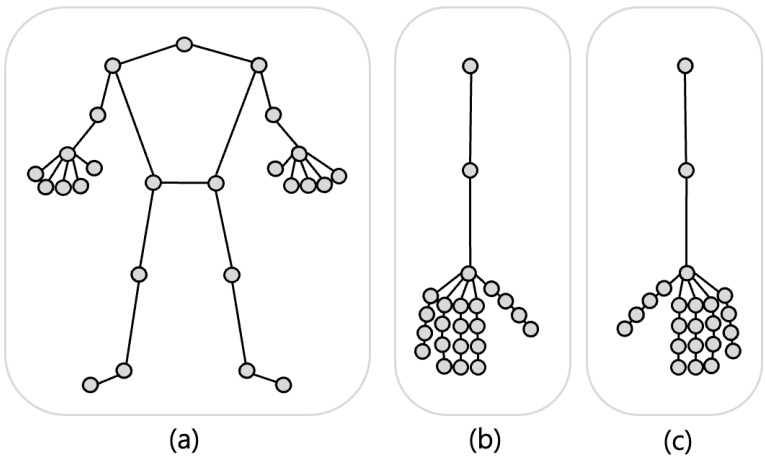

그림 5. (a) 몸 관절 보정 네트워크의 스켈레톤 그래프, (b) 오른손 관절 보정 네트워크의 스켈레톤 그래프, (c) 왼손 관절 보정 네트워크의 스켈레톤 그래프
Fig. 5. Skeleton graphs used in (a) the body joint refinement network, (b) the right-hand joint refinement network, and (c) the left-hand joint refinement network

로 변환된다. 그림 5는 몸 관절 보정 네트워크, 오른손 관절 보정 네트워크, 그리고 왼손 관절 보정 네트워크의 스켈레톤 그래프를 보여준다. 마지막으로 H'_{B_T}, H'_{R_T}, H'_{L_T}는 1층의 출력 그래프 합성곱 신경망을 통과하여 개선된 3차원 관절 좌표 P'_B, P'_R, P'_L을 출력한다.

3. 손실 함수

자세 초기화 모듈과 자세 보정 모듈로 구성된 제안 모델의 학습을 위해 L2 손실 함수(L2 loss function)를 사용한다. 자세 초기화 모듈의 학습을 위한 손실 함수 $\mathcal{L}_{3D}$는 다음과 같이 정의된다:

$$\mathcal{L}_{Body} = \frac{1}{15}\sum_{j=1}^{15}\left\|P_{Body,j} - P^*_{Body,j}\right\|_2, \quad (1)$$

$$\mathcal{L}_{RH} = \frac{1}{20}\sum_{j=1}^{20}\left\|P_{RH,j} - P^*_{RH,j}\right\|_2, \quad (2)$$

$$\mathcal{L}_{LH} = \frac{1}{20}\sum_{j=1}^{20}\left\|P_{LH,j} - P^*_{LH,j}\right\|_2, \quad (3)$$

$$\mathcal{L}_{3D} = \mathcal{L}_{Body} + \mathcal{L}_{RH} + \mathcal{L}_{LH}. \quad (4)$$

여기서 $\mathcal{L}_{Body}$, $\mathcal{L}_{RH}$, $\mathcal{L}_{LH}$는 각각 몸 관절 좌표에 대한 손실 함수, 오른손 관절 좌표에 대한 손실 함수, 왼손 관절 좌표에 대한 손실 함수를 나타낸다. P_j는 j번째 관절 좌표 추정 값을 의미하고 P^*_j는 j번째 관절 좌표에 대한 참값(ground truth)을 의미한다. 다음으로 자세 보정 모듈의 학습을 위한 손실 함수 $\mathcal{L}_G$는 다음과 같이 정의된다:

$$\mathcal{L}_B = \frac{1}{25}\sum_{j=1}^{25}\left\|P'_{B,j} - P^*_{B,j}\right\|_2, \quad (5)$$

$$\mathcal{L}_R = \frac{1}{22}\sum_{j=1}^{22}\left\|P'_{R,j} - P^*_{R,j}\right\|_2, \quad (6)$$

$$\mathcal{L}_L = \frac{1}{22}\sum_{j=1}^{22}\left\|P'_{L,j} - P^*_{L,j}\right\|_2, \quad (7)$$

$$\mathcal{L}_G = \mathcal{L}_B + \mathcal{L}_R + \mathcal{L}_L. \quad (8)$$

Ⅳ. 실험 결과 및 분석

1. 데이터셋, 평가 척도, 구현 세부사항

본 연구에서는 제안 방법의 학습을 위해 EgoWholeBody 학습 데이터셋[1]을 사용하였다. EgoWholeBody 데이터셋은 1인칭 영상 및 전신 3차원 자세가 어노테이션(annotation)되어 있는 합성(synthetic) 데이터셋이다. 이는 2,367개의 Mixamo[19] 모션 시퀀스(motion sequence)에 의해 구동되는 14개의 RenderPeople[20] 모델이 렌더링된 총 700,000장의 프레임으로 구성된다. 평가에 사용한 데이터셋은 EgoWholeBody 평가 데이터셋과 SceneEgo Hand Occlusion 데이터셋이다. EgoWholeBody 평가 데이터셋은 4개의 RenderPeople 모델이 렌더링된 46,794장의 프레임을 포함한다. SceneEgo Hand Occlusion 데이터셋은 COCOA 데이터셋[21,22]의 다양한 물체를 실제 영상으로 구성된 SceneEgo 데이터셋[7]의 1인칭 영상 속 휴먼 객체의 손 위치에 합성하여 생성된 데이터셋이다. VR/AR 기기의 경우 헤드셋뿐만 아니라 사용자가 손에 쥐는 컨트롤러도 함께 포함하는 경우가 많은데, 이는 손의 가리워짐을 빈번히 유발한다. 이러한 손 가리워짐에 대해서 제안하는 방법의 강인함을 평가하기 위해서 SceneEgo Hand Occlusion 데이터셋에 대한 평가를 진행하였다. 그림 6은 SceneEgo Hand Occlusion 데이터셋의 예를 보여준다. SceneEgo Hand Occlusion 데이터셋은 2명의 배우가 걷기, 앉기, 신문 읽기 등 일상 생활에서의 행동들을 촬영한 13,242장의 프레임으로 구성된다.

평가 척도로는 Mean Per Joint Position Error(MPJPE), Procrustes Alignment MPJPE(PA-MPJPE)를 사용한다. MPJPE는 추정된 3차원 관절 좌표 집합과 참값 3차원 관절 좌표 집합의 기준 관절을 일치시킨 후 각 관절에 대한 오차의 평균을 측정한 값으로 단위는 mm이다. PA-MPJPE는 각 관절 집합에 Procrustes analysis[23]를 적용해 스케일(scale), 이동(translation), 회전(rotation)의 영향을 제거한 후 오차의 평균을 측정한 값으로 단위는 mm이다.

제안하는 방법은 단계적 학습 방법을 사용한다. 자세 초기화 모듈을 먼저 학습한 후 자세 초기화 모듈을 고정(freeze)시킨 후 자세 보정 모듈을 학습한다. 자세 초기화

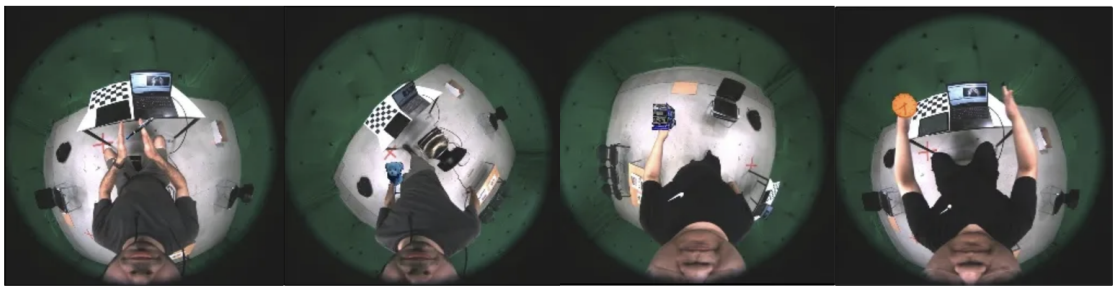

그림 6. SceneEgo Hand Occlusion 데이터셋의 예
Fig. 6. Examples of SceneEgo Hand Occlusion dataset

모듈을 학습하기 위한 최적화 방법으로는 Adam[24]을 사용하고 학습률(learning rate)은 10^{-4}이다. 배치 크기(batch size)는 128, 학습 에포크(epoch)는 20으로 설정하여 학습한다. 자세 보정 모듈을 학습하기 위한 최적화 방법은 Adam을 사용하고 학습률은 10^{-3}이고 10에포크마다 10^{-1}배씩 작아지게 한다. 배치 크기는 150, 학습 에포크는 20으로 설정한다.

2. 정량적 평가 결과

우리는 제안하는 방법이 성능 개선에 정량적으로 도움을 준다는 것을 보인다. 이를 위해 EgoWholeBody 평가 데이터셋과 SceneEgo Hand Occlusion 데이터셋에서 EgoWholeBody 모델[1]과 기준 모델(baseline)과의 정량적 비교를 수행한다. 기준 모델은 선행 연구[25]에서 제안되었으며, 본 논문에서 제안하는 방법과 동일한 구조를 가진다. 한 가지 차이점은 자세 보정 모듈에 대한 입력 영상 특징 벡터의 차원이 128에서 512로 변경된 것인데, 이를 뒷받침하기 위한 차원 최적화 실험 결과가 제시된다.

표 2와 3은 각각 EgoWholeBody 평가 데이터셋에 대하여 몸 자세 추정 성능과 손 자세 추정 성능의 비교 결과를 보여준다. 제안 방법의 손 자세 추정 MPJPE는 EgoWholeBody 모델보다 3.69mm, baseline보다 0.12mm 감소했고 PA-MPJPE는 EgoWholeBody 모델보다 10.47mm, baseline보다 0.73mm 감소했다. 이를 통해 두 단계 방법인 EgoWholeBody 모델보다 제안하는 단일 단계 방법이 더 높은 손 자세 추정 성능을 달성함을 알 수 있다. 1인칭 영상의 어안 카메라 왜곡으로 인한 손 검출 오차가 손 영역을 정확

표 2. EgoWholeBody 평가 데이터셋에 대한 몸 성능 비교
Table 2. Comparison of body estimation performance on the EgoWholeBody test dataset

Model	MPJPE	PA-MPJPE
EgoWholeBody	66.28	43.14
Baseline	65.82	48.51
Ours	63.66	45.73

표 3. EgoWholeBody 평가 데이터셋에 대한 손 성능 비교
Table 3. Comparison of hand estimation performance on the EgoWholeBody test dataset

Model	MPJPE	PA-MPJPE
EgoWholeBody	33.10	19.68
Baseline	29.53	9.84
Ours	29.41	9.21

표 4. EgoWholeBody 평가 데이터셋에 대한 전신 성능 비교
Table 4. Comparison of wholebody estimation performance on the EgoWholeBody test dataset

Model	MPJPE	PA-MPJPE
EgoWholeBody	71.24	43.14
Baseline	55.04	35.13
Ours	54.73	34.78

하게 자르지 못하게끔 하기 때문에 잘린 손 영상을 사용하는 두 단계 방법의 성능이 저하된다. 하지만 baseline과 제안하는 방법은 단일 단계 방법으로 손 자세 추정 성능이 손 검출 성능에 영향을 받지 않는다. 따라서 손 검출이 잘 되지 않는 1인칭 영상에서 상대적으로 높은 손 자세 추정 성능이 관찰된다. 표 4는 EgoWholeBody 평가 데이터셋에

대하여 전신 자세 추정 성능을 비교한 표이다. 제안 방법의 전신 자세 추정 MPJPE는 EgoWholeBody 모델보다 16.51mm, baseline보다 0.31mm 감소했고, PA-MPJPE는 EgoWholeBody 모델보다 8.36mm, baseline보다 0.35mm 감소했다. 이를 통해 손 자세 추정 성능이 향상됨에 따라 전신 자세 추정 성능도 개선된 것을 확인할 수 있다.

표 5와 6은 각각 SceneEgo Hand Occlusion 데이터셋에 대한 몸 자세 추정 성능과 손 자세 추정 성능의 비교를 보여준다. 제안하는 방법의 손 자세 추정 MPJPE는 EgoWholeBody 모델보다 64.01mm 감소하였고, baseline 보다 0.75mm 감소하였다. SceneEgo Hand Occlusion 데이터셋은 손이 가리워지는 상황을 포함하고 있어 손 중심으로 잘려진 영상만을 입력으로 사용하는 두 단계 방법의 경우 손 자세 추정을 위한 충분한 정보를 얻기 어렵다. 이에 따라 두 단계 방법의 손 자세 추정 성능이 전반적으로 낮아지는 경향이 나타난다. 반면 단일 단계 방식을 사용하는 제안 방법과 baseline은 손 검출 과정에 의존하지 않으므로 손 가리워짐이 있는 상황에서도 비교적 높은 성능을 유지할 수 있다. 표 7은 SceneEgo Hand Occlusion 데이터셋에 대한 전신 추정 성능을 비교한 표이다. 제안하는 방법의 전신 자세 추정 MPJPE가 EgoWholeBody보다 37.51mm, baseline보다 1.59mm 감소했다. 이는 손 자세 추정 성능 개선이 전신 자세 추정에도 긍정적인 영향을 미쳤음을 의미한다.

표 8은 EgoWholeBody 모델, baseline, 그리고 제안하는 방법의 추정 시간 및 fps 비교 결과를 보여준다. 제안하는 방법은 baseline보다 고차원의 영상 특징 벡터를 자세 보정 모듈의 입력으로 사용하지만 추정 시간에는 큰 차이가 없다. 하지만 3차원 히트맵 기반 방법인 EgoWholeBody 모델보다 제안 방법의 추정 시간은 88.89% 작다. 제안 방법의 자세 추정 속도는 약 76.9fps로 EgoWholeBody 모델과 달리 실시간 추정이 가능하다.

표 9는 자세 보정 모듈에서 입력 영상 특징 벡터의 차원에 따른 성능 비교이다. baseline의 입력 영상 특징 벡터를 128차원에서 512차원으로 바꿨을 때 몸 자세 추정 MPJPE와 PA-MPJPE가 각각 2.16mm, 3.71mm로 감소한다. 1024차원과 512차원의 경우 성능 차이는 미미하였으며 차원이 증가할수록 모델의 추정 속도가 저하된다. 따라서 최적의 입력 영상 특징 벡터 차원으로 512차원을 선택하여 제안 방법에 사용한다.

표 5. SceneEgo Hand Occlusion 데이터셋에 대한 몸 성능 비교
Table 5. Comparison of body estimation performance on the SceneEgo Hand Occlusion dataset

Model	MPJPE	PA-MPJPE
EgoWholeBody	142.36	106.22
Baseline	161.16	115.73
Ours	151.29	108.43

표 6. SceneEgo Hand Occlusion 데이터셋에 대한 손 성능 비교
Table 6. Comparison of hand estimation performance on the SceneEgo Hand Occlusion dataset

Model	MPJPE	PA-MPJPE
EgoWholeBody	112.32	29.14
Baseline	49.06	16.63
Ours	48.31	16.50

표 7. SceneEgo Hand Occlusion 데이터셋에 대한 전신 성능 비교
Table 7. Comparison of wholebody estimation performance on the SceneEgo Hand Occlusion dataset

Model	MPJPE	PA-MPJPE
EgoWholeBody	157.55	95.37
Baseline	120.92	81.42
Ours	119.33	78.22

표 8. 추정 시간과 fps 비교
Table 8. Comparison ofinference time and fps

Model	Time(s)	fps
EgoWholeBody	0.117	8.55
Baseline	0.011	90.91
Ours	0.013	76.92

표 9. 자세 보정 모듈의 입력 영상 특징 벡터 차원에 따른 성능 비교
Table 9. Performance comparison according to the dimensionality of the input image feature vector for the pose refinement module

Feature dimension	MPJPE	PA-MPJPE	Inference time
128 (Baseline)	65.82	48.51	0.011
512 (Ours)	63.66	45.73	0.013
1024	64.10	44.80	0.015

3. 정성적 평가 결과

그림 7은 EgoWholeBody 평가 데이터셋에 대한 3차원 전신 자세 추정 결과를 정성적으로 보여준다. 두 단계 방법인 EgoWholeBody 모델보다 단일 단계 방법인 baseline과 제안하는 방법이 더 작은 손 자세 추정 오차를 보여준다.

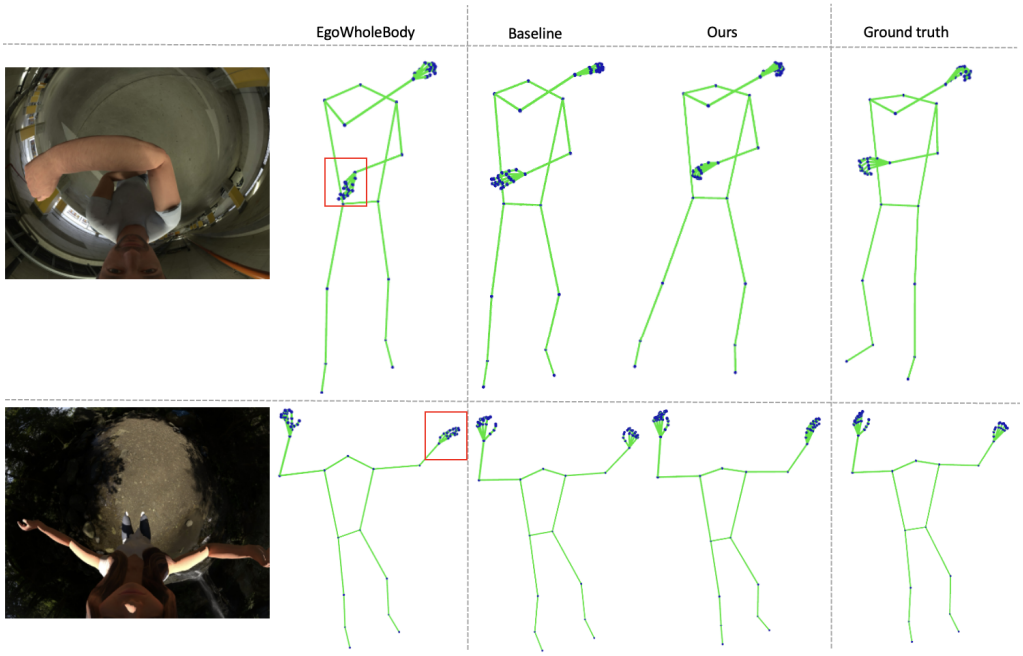

그림 7. EgoWholeBody 평가 데이터셋에 대한 정성적 비교
Fig. 7. Qualitative comparison on EgoWholeBody test dataset

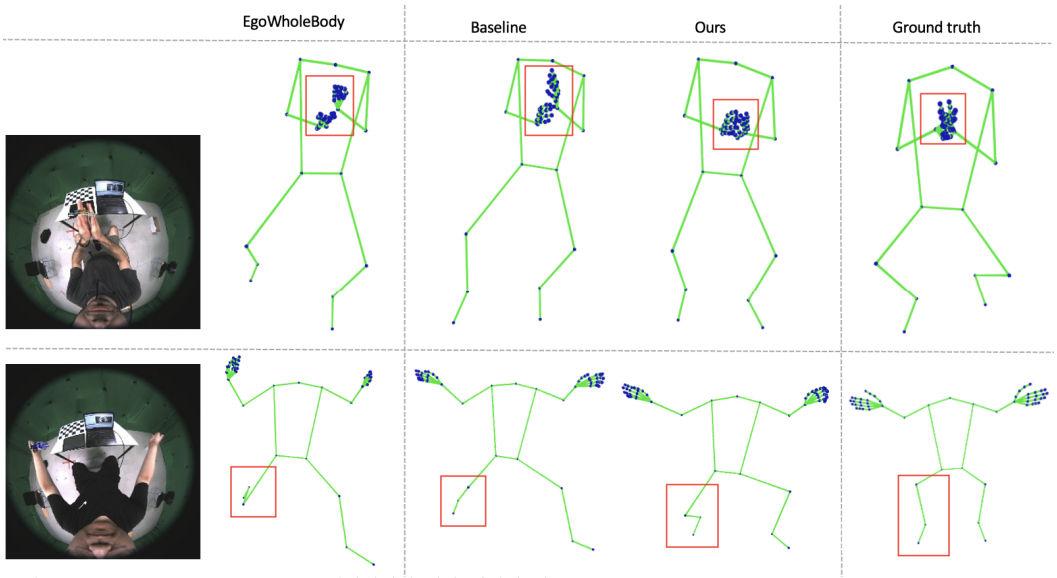

그림 8. SceneEgo Hand Occlusion 데이터셋에 대한 정성적 비교
Fig. 8. Qualitative comparison on SceneEgo Hand Occlusion dataset

그림 8은 SceneEgo Hand Occlusion 데이터셋에 대한 3차원 전신 자세 추정 결과를 정성적으로 보여준다. 이를 통해 손 가리워짐 상황에서도 제안하는 방법이 EgoWholeBody 및 baseline 모델보다 강인한 손 추정 성능을 보이는 것을 확인할 수 있다. 예를 들어, 그림 8의 첫 번째 행의 박수를 치는 동작에서 EgoWholeBody 모델과 baseline 모델은 두 손이 붙어 있지 않으나 제안하는 방법은 두 손이 붙어 있다. 또한 그림 8의 두 번째 행에서 제안하는 방법은 baseline보다 개선된 하체 추정 성능을 보여준다.

V. 결 론

본 논문에서는 1인칭 영상을 입력으로 하여 3차원 전신 자세를 출력할 때 손 검출 성능이 저하됨에 따라 손 자세 추정 성능도 저하되는 문제를 해결하고자 하였다. 이에 따라 손 검출 성능에 영향을 받지 않고 관절 사이의 관계 정보를 활용할 수 있는 단일 단계 방법을 제안하였다. 제안하는 방법을 통해 기존의 두 단계 방법보다 1인칭 영상에서 3차원 손 자세 추정 성능뿐만 아니라 전신 자세 추정 성능이 개선됨을 정량적, 정성적으로 증명하였다. 또한 자세 보정 모듈의 입력 영상 특징 벡터 차원에 대한 실험을 통해 최적화를 수행하였다. 제안하는 직접 회귀 기반의 방법은 히트맵 기반의 방법보다 자세 추정 속도가 빠르며 실시간 처리가 가능하다.

참 고 문 헌 (References)

[1] J. Wang, Z. Cao, D. Luvizon, L. Liu, K. Sarkar, D. Tang, T. Beeler, and C. Theobalt, “Egocentric whole-body motion capture with FisheyeViT and diffusion-based motion refinement,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 777-787, 2024.
doi: https://doi.org/10.1109/cvpr52733.2024.00080

[2] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, P. Fua, H.-P. Seidel, and C. Theobalt, “Mo2Cap2: Real-time mobile 3D motion capture with a cap-mounted fisheye camera,” IEEE Transactions on Visualization and Computer Graphics, vol. 25, no. 5, pp. 2093-2101, 2019.
doi: https://doi.org/10.1109/tvcg.2019.2898650

[3] D. Tome, P. Peluse, L. Agapito, and H. Badino, “xR-EgoPose: Egocentric 3D human pose from an HMD camera,” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7728-7738, 2019.
doi: https://doi.org/10.1109/iccv.2019.00782

[4] D. Zhao, Z. Wei, J. Mahmud, and J.-M. Frahm, “EgoGlass: Egocentric-view human pose estimation from an eyeglass frame,” in Proceedings of the IEEE/CVF International Conference on 3D Vision (3DV), pp. 32-41, 2021.
doi: https://doi.org/10.1109/3dv53792.2021.00014

[5] Y. Zhang, S. You, and T. Gevers, “Automatic calibration of the fisheye camera for egocentric 3D human pose estimation from a single image,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1772-1781, 2021.
doi: https://doi.org/10.1109/wacv48630.2021.00181

[6] H. Akada, J. Wang, S. Shimada, M. Takahashi, C. Theobalt, and V. Golyanik, “UnrealEgo: A new dataset for robust egocentric 3D human motion capture,” in Proceedings of the European Conference on Computer Vision (ECCV), pp. 1-17, 2022.
doi: https://doi.org/10.1007/978-3-031-20068-7_1

[7] J. Wang, D. Luvizon, W. Xu, L. Liu, K. Sarkar, and C. Theobalt, “Scene-aware egocentric 3D human pose estimation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13031-13040, 2023.
doi: https://doi.org/10.1109/cvpr52729.2023.01252

[8] T. Kang, K. Lee, J. Zhang, and Y. Lee, “Ego3DPose: Capturing 3D cues from binocular egocentric views,” in Proceedings of the ACM SIGGRAPH Asia 2023 Conference, pp. 1-12, 2023.
doi: https://doi.org/10.1145/3610548.3618147

[9] T. Kang and Y. Lee, “EgoTAP: Attention-propagation network for egocentric heatmap to 3D pose lifting,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1234-1243, 2024.
doi: https://doi.org/10.1109/cvpr52733.2024.00086

[10] J.-B. Lee, H. Lee, B.-R. Lee, B.-G. Lee, and W.-H. Son, “EgoCoord: Self-calibrated egocentric 3D body pose estimation using pixel-wise coordinate encoding,” in Proceedings of the Asian Conference on Computer Vision (ACCV), pp. 1233-1249, 2024.
doi: https://doi.org/10.1007/978-981-96-0901-7_16

[11] C. Yang, A. Tkach, S. Hampali, L. Zhang, E. J. Crowley, and C. Keskin, “EgoPoseFormer: A simple baseline for stereo egocentric 3D human pose estimation,” in Proceedings of the European Conference on Computer Vision (ECCV), pp. 1-17, 2024.
doi: https://doi.org/10.1007/978-3-031-73001-6_23

[12] G. Zhao, X. Peng, and S. Lucey, “Semantic graph convolutional networks for 3D human pose regression,” arXiv preprint arXiv:1904.03345, 2019.
doi: https://doi.org/10.1109/cvpr.2019.00354

[13] Z. Zou and W. Tang, “Modulated graph convolutional network for 3D human pose estimation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11477-11487, 2021.
doi: https://doi.org/10.1109/iccv48922.2021.01128

[14] A. T. M. Shahjahan and A. Ben Hamza, “Flexible graph convolutional network for 3D human pose estimation,” arXiv preprint

arXiv:2407.19077, 2024.
[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016. doi:https://doi.org/10.1109/cvpr.2016.90
[16] D. Hendrycks and G. Gimpel, “Gaussian error linear units (GELUs),” arXiv preprint arXiv:1606.08415, 2016.
[17] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proceedings of the International Conference on Machine Learning (ICML), pp. 448-456, 2015.
[18] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems (NeurIPS), pp. 1097-1105, 2012. doi: https://doi.org/10.1145/3065386
[19] Mixamo. http://www.mixamo.com.
[20] Renderpeople. http://www.renderpeople.com.
[21] Y. Zhu, Y. Tian, D. Metaxas, and P. Dollár, “Semantic amodal segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1464 – 1472, 2017. doi: https://doi.org/10.1109/cvpr.2017.320
[22] P. Follmann, R. König, P. Härtinger, M. Klostermann, and T. B ttger, “Learning to see the invisible: End-to-end trainable amodal instance segmentation,” in Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1328 – 1336, 2019. doi:https://doi.org/10.1109/wacv.2019.00146
[23] J.C. Gower, “Generalized procrustes analysis,” Psychometrika, vol. 40, no. 1, pp. 33 – 51, 1975. doi: https://doi.org/10.4135/9781412952644.n186
[24] D.P. Kingma and J.Ba, “Adam: A method for stochastic optimization,” in Proceedings of the International Conference for Learning Representations (ICLR), 2015.
[25] S. Na and J. Y. Chang, “One-stage method for 3D wholebody pose estimation from egocentric images,” 37th Workshop on Image Processing and Image Understanding, 2025.

저 자 소 개

나 소 연

- 2024년 2월 : 광운대학교 전자통신공학과 학사
- 2024년 3월 ~ 현재 : 광운대학교 전자통신공학과 석사과정
- ORCID : https://orcid.org/0009-0004-4395-8750
- 주관심분야 : 컴퓨터비전 및 머신러닝

장 주 용

- 2001년 2월 : 서울대학교 전기공학부 학사
- 2008년 2월 : 서울대학교 전기컴퓨터공학부 박사
- 2008년 2월 ~ 2009년 1월 : Postdoctoral Researcher, Mitsubishi Electric Research Laboratories (MERL), US
- 2009년 4월 ~ 2011년 1월 : 삼성전자 DMC 연구소 책임연구원
- 2011년 4월 ~ 2012년 2월 : 서울대학교 BK 조교수
- 2012년 3월 ~ 2017년 2월 : 한국전자통신연구원 선임연구원
- 2024년 3월 ~ 2025년 2월 : Visiting Scholar, University of Birmingham, UK
- 2017년 3월 ~ 현재 : 광운대학교 전자통신공학과 교수
- ORCID : https://orcid.org/0000-0003-3710-7314
- 주관심분야 : 컴퓨터비전 및 머신러닝