Journal of KIIT. Vol. 21, No. 5, pp. 1-9, May 31, 2023. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2023.21.5.1

멀티뷰 카메라를 이용한 HMR 기반의 사람 자세 및 형태 추정 모델 성능 개선

권혁민*, 이준호**, 김화종***

Performance Improvement of HMR based Human Pose and Shape Estimation Model by using Multi-View Camera

Hyeok-Min Kwon*, Jun-Ho Lee**, and Hwa-Jong Kim***

요 약

인체생성모델 기반의 연구를 통해 센서가 필요없이 이미지만으로 모션데이터를 추정하는 연구가 활발히 진행되고 있다. 최근 대부분의 연구들이 단일 카메라 기반의 방식을 택하고 있으나 이는 가려진 영역에 민감하여 보이지 않는 영역이 있을 경우 모델의 성능이 떨어지는 문제가 있다. 이러한 부분을 해결하기 위해 본 연구에서는 다중카메라를 이용하는 멀티뷰 방식을 제안한다. 어떤 카메라에서 보이지 않는 영역이 다른 카메라에선 잘 보일 수 있다는 가정을 기반으로 본 연구에서는 두 개 카메라 이미지 입력에 대한 모델의 출력값을 반복 추정 과정에서 교차함으로써 두 카메라의 추정 정보를 활용할 수 있도록 하고 이를 통해 모델의 전체 성능을 개선하였다. 본 연구에서 제안하는 방식은 간단한 방식으로써 기존 모델구조를 확장하여 활용할 수 있는 장점이 있다. 제안한 방식의 효과에 관한 실험을 위해 기존 모델에 적용하여 10%이상의 성능향상을 보였다. 실험 코드는 https://github.com/kwonhyeokmin/MVHMR.git에서 확인할 수 있다.

Abstract

Through human body model based research, there is an active pursuit to estimate motion data using only image inputs without requiring sensors. However, most recent studies have opted for single-camera-based methods, which suffer from a decrease in performance when certain areas of the body are not visible in the image. To address this issue, this study proposes a multi-view-based method that using multiple cameras. Building on the assumption that a body area not visible in one camera may be visible in another, this study improves the overall performance of the model by utilizing the estimation information from two camera images in the iterative estimation process by cross-linking the output values of the model for the two camera inputs. The proposed method has the advantage of being a simple extension of existing model structures, allowing for broad applicability. The effectiveness of the proposed method was demonstrated by applying it to an existing model, resulting in more than a 10% performance improvement. The code is available in https://github.com/kwonhyeokmin/MVHMR.git.

Keywords

AI, computer vision, human pose estimation, human shape estimation, deep learning

* 강원대학교 데이터사이언스학과 석사과정
- ORCID: https://orcid.org/0009-0006-0886-960X
** 아주대학교 지능형소프트웨어학과 석사과정
- ORCID: https://orcid.org/0009-0001-9559-2699
*** 강원대학교 컴퓨터공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-3822-390X

· Received: Mar. 18, 2023, Revised: Apr. 25, 2023, Accepted: Apr. 28, 2023
· Corresponding Author: Hwa-Jong Kim
Dept. of Computer Engineering, Kangwon National University, Korea
Tel.: +82-33-250-6323, Email: hjkim3@gmail.com

I. 서 론

이미지로부터 사람의 3차원 관절과 형태를 추정하는 연구는 행동 예측이나 AR/VR 뿐만 아니라 로보틱스와 관련된 산업에도 다양하게 활용되고 있다[1]. 특히 인체생성모델을 활용한 연구들을 통해 단순히 관절의 위치를 추정하는 것을 넘어 사람의 관절의 각도를 추정하여 3차원 모션 데이터를 추출하는데 활용할 수 있다.

최근엔 주로 단일카메라 이미지를 통해 사람의 형태와 관절을 추정하는 연구들이 진행되고 있다. 그러나 단일카메라 이미지의 경우 사물이나 신체 등에 의해 추정하려고 하는 영역이 가려지는 상황이 쉽게 일어날 수 있다. 특히 인체생성모델 기반의 모델들은 가려진 이미지에서 정확도가 떨어지며 단순히 가려진 관절에 대한 예측 정확도가 낮아지는 것을 넘어 전체 관절에 대한 예측 정확도가 떨어지는 문제가 있다[2][3]. 이러한 부분에 강건한 모델을 만들기 위해 데이터를 다양하게 구성하거나 데이터를 증폭하여 학습시키는 방식으로 해결하고 있으나 본 연구에서는 근본적인 해결책은 아님을 지적한다.

다중카메라를 활용하게 되면 예측하고자 하는 관절이 특정 카메라에선 보이지 않지만 다른 카메라에서는 잘 보일 수 있어 가려진 부분을 보완할 수 있으며 이를 통해 모델의 정확도를 높일 수 있다. 이러한 방식을 멀티뷰 방식이라 하며 인체생성모델을 사용하지 않는 3차원 관절인식의 경우 멀티뷰 방식이 높은 성능을 내고 있다.

그러나 인체생성모델을 이용한 연구의 경우 결국 관절의 위치가 아닌 각도를 추정해야하기 때문에 각 카메라에서 나온 출력값 중 대푯값 선택에 대한 모호한 부분이 생긴다. 예를 들어 A 카메라와 B 카메라가 있는 상황에서 두 카메라 중 어떤 카메라가 더 정확한 값을 출력할지 아닐지 판단하기도 난해하고 더 정확한 값을 출력하는 카메라를 찾았다하더라도 결국 두 카메라 중 하나의 카메라로부터 나온 이미지만 사용하기 때문에 결론적으론 한 개 카메라만 사용했을 때와 같은 결과가 나오게 된다.

위 문제를 해결하기 위해 본 연구에서는 간단한 방식을 통해 멀티뷰 방식으로 사람 관절 및 형태추정 모델을 구축하는 방법을 제안한다. 본 연구에서는 가장 대표적인 모델인 HMR(Human Mesh Recovery)의 반복 추정 과정을 응용하여 다른 카메라의 출력을 반영하였다. 본 연구는 현재 주로 연구되고 있는 단일카메라 모델에 쉽게 적용할 수 있다. 제안 모델은 두 개의 캘리브레이션이 완료된 카메라로부터 받은 이미지를 활용하여 기존 모델보다 10%이상의 성능향상을 보였다.

본 논문의 다음과 같이 구성하였다. 2장에서는 본 연구와 관련된 연구들을 소개하고 3장에는 실제 제안하는 모델의 네트워크 구조를 소개한다. 그리고 4장에는 공개 데이터셋에서 본 연구에서 기존 모델과 제안하는 방식을 사용한 모델의 성능을 비교하고 마지막 5장에서 결론 및 결과 이미지와 향후 연구에 대한 계획을 기술하였다.

II. 관련 연구

2.1 SMPL(Skinned Multi-Person Linear model)

SMPL[4]은 대표적인 인체생성모델 중 하나로 파라미터를 통해 인체의 3차원 메쉬를 생성한다. 이 모델은 3차원 형태와 움직임을 표현하기 위해 파라미터의 형태로 설계되었으며 인체 동작을 시뮬레이션 하거나 가상의 인체를 생성하는 등 여러 분야에서 사용되고 있다. SMPL의 파라미터는 관절파라미터(Pose parameter)와 형태파라미터(Shape parameter)로 구성되며 이를 입력으로 받아 6890개의 3차원 메쉬 V를 반환한다. 형태파라미터 $\beta \in R^{10}$는 사람의 체형에 관여하는 파라미터이고 관절파라미터는 중심관절 $\phi \in R^{1\times3\times3}$과 나머지 23개 관절에 해당하는 $\theta \in R^{23\times3\times3}$로 총 24개 관절의 회전행렬로 구성된 파라미터다. 최근엔 관절 파라미터를 6차원 행렬로 표현하여 성능을 높이는 연구[5]가 진행되었다.

2.2 3D 사람 관절 추정 연구

사람 3차원 관절 추정연구는 인체생성모델의 활용유무에 따라 모델기반 방식과 모델 프리 방식으로 구분하거나 입력 이미지의 활용방식에 따라 상향식 방식과 하향식 방식으로 구분한다.

모델 프리 방식은 이미지로부터 관절의 3차원 좌표를 바로 추정하지만 모델 기반 방식의 경우 3차원 좌표를 바로 출력하는 것이 아닌 인체생성모델의 파라미터를 넣어서 3차원 좌표를 출력하는 방식이다. 이에 따라 모델 프리 방식의 연구는 이미지로부터 관절의 3차원 좌표를 추정하기 위해 모델을 설계하나 모델 기반 방식의 경우 정확한 3차원 관절을 출력하기 위한 인체생성모델의 파라미터를 추정하는 모델을 연구한다.

상향식 방식은 각 사람의 경계박스 정보를 통해 사람의 영역만을 잘라내어 입력값으로 활용한다. 때문에 경계박스의 정확도에 영향을 크게 받으며 이미지 속 사람의 수를 n명이라 할 때 모델에 n개의 이미지를 반복하여 입력해야 하기 때문에 사람 수가 많을수록 비례하여 시간이 오래 걸리는 단점이 있다. 또한 경계박스 정보가 없을 경우 이를 추정하는 모델이 선행되는 단점 또한 존재한다[6]. 하향식 방식의 경우 이미지로부터 바로 여러 사람의 관절을 추정하기 때문에 사람끼리 겹친 이미지를 입력으로 받더라도 좋은 성능을 보인다. 이 방식은 사람 수에 상관없이 한 번만 모델에 입력하면 되지만 상향식 방식에 비해 관절의 예측 정확도가 떨어지고 입력 이미지의 해상도가 높아야 한다는 단점이 있다[3].

본 연구에서는 모델 기반의 상향식 방식을 채택하였다.

2.3 HMR

HMR[7] 모델은 모델기반의 상향식 방식 모델 중 하나로 경계박스 정보로부터 사람의 영역만을 잘라낸 이미지를 입력값으로 하여 SMPL의 파라미터를 출력한다. HMR은 바로 SMPL 파라미터를 추정하는 구조가 아닌 이전의 네트워크에서 추정했던 출력값을 같이 입력받아 업데이트하는 과정을 반복하여 추정하는 구조이다. 즉, 모델의 출력 SMPL 파라미터를 Θ라 하고 t번째 출력값을 Θ_t라고 했을 때 $\Theta_{t+1} = \Theta_t + \Delta\Theta_t$가 되며 HMR의 경우 $\Delta\Theta_t$를 추정하는 과정을 반복하여 최종적으로 정확한 Θ를 출력한다. 초기 입력값 Θ_0는 SMPL의 파라미터 평균 $\overline{\Theta}$로 설정하였고 반복 횟수는 3으로 설정하였다.

HMR는 가장 대표적인 모델 중 하나로써 [8]-[10] 연구 등 여러 연구에서 이 구조를 활용하여 좋은 성능을 내고 있다. 본 연구에서는 해당 네트워크 구조를 응용하여 단일 카메라 이미지가 아닌 다중카메라 이미지를 이용하는 멀티뷰 기반 모델을 통해 성능을 높이는 방식을 제안한다.

2.4 멀티뷰 기반 모델 연구동향

모델 기반 방식 연구의 대부분은 단일 카메라 이미지를 통해 SMPL의 파라미터를 추정하나 최근 멀티뷰 기반 방식의 모델 또한 활발히 연구되고 있다. [11]는 여러 카메라 이미지를 통해 메쉬 포인트를 추정하고 해당 포인트에 SMPL파라미터를 피팅하여 출력한다. [8]는 본 연구에 많은 영향을 준 연구로써 두 개의 이미지로부터 추정된 SMPL의 파라미터를 합쳐 다음 반복의 입력값으로 넣어 성능을 높였다. 그러나 [8][11]연구들은 새로운 모델 구조로써 기존 성능을 잘 보이는 단일 이미지 기반 모델을 활용하지 못한다. 그러나 본 연구에서 제안하는 모델은 기존 단일 이미지 기반 모델을 멀티뷰 기반 모델로 확장하는 방법을 제안함으로써 사용성을 높이고 새로 학습시킬 필요 없이 잘 학습된 단일 이미지 기반을 미세조정하여 정확한 SMPL 파라미터를 추정할 수 있도록 하였다.

III. 멀티뷰 기반의 사람 자세 및 형태 추정

본 연구는 멀티뷰 기반으로써 두 개의 이미지로부터 SMPL 모델 파라미터를 추정하는 방식을 제안한다. 유사한 방식을 적용한 연구[8]의 경우 반복 추정 과정에서 두 개의 이미지로부터 나온 출력값을 합쳐 다음 반복의 입력값으로 활용하여 좋은 성능을 보였다. 그러나 현재 대부분의 모델들은 단일 이미지 기반으로써 한 개의 이미지로부터 나온 출력값을 다음 반복의 입력값으로 활용한다. 따라서 [8]의 방식은 다음 반복에 들어가기 위한 출력값의 차원이 단일 이미지를 활용하는 방식과 다르기 때문에 기존의 단일이미지 방식의 네트워크를 미세조정하여 사용할 수 없다는 한계가 있다.

이를 해결하기 위해 본 논문에서는 이전 반복의 출력을 받는 과정에서 두 개의 카메라 이미지 입력으로부터 나온 관절파라미터와 형태파라미터 출력값을 서로 교차하여 입력값으로 활용하는 방식을 제안한다. 본 제안 방식을 MVHMR(Multi-View Human Mesh Recovery)라하고 이 방식을 통해 캘리브레이션된 두 개의 카메라로부터 얻은 이미지를 통해 SMPL모델 파라미터를 추정할 수 있다.

본 제안의 핵심은 반복 추정 과정에서 두 개의 카메라의 출력값을 교차함으로써 다른 쪽 카메라에서의 출력값을 반영하는 것이다. 이를 통해 최종적으로 타겟 카메라에서 보이지 않거나 판단하기 어려운 관절값을 더 정교하게 추정할 수 있다. 또한 본 제안방식은 기존 반복 추정 방식을 활용하는 좋은 성능을 보이는 모델 네트워크의 구조와 모델 파라미터를 수정할 필요 없이 미세조정을 활용할 수 있어 다양한 모델에 적용할 수 있다.

그림 1는 본 연구에서 제안하는 멀티뷰 기반의 네트워크 구조를 나타낸다. 네트워크는 특징맵을 추출하는 백본 네트워크와 SMPL의 입력값에 해당하는 ϕ, θ, β, γ값 추출을 위한 회귀모델로 구성되어 있다.

회귀모델은 그림 2과 같이 두 개의 전결합층과 두 개의 드롭아웃층을 지난 후 세 개의 헤더 네트워크로부터 사람의 관절 각도와 관련된 파라미터 ϕ, θ와 형태와 관련된 파라미터 β, 그리고 평행이동과 관련된 파라미터 γ를 출력한다. 백본 네트워크는 HRNet-48[12]과 Resnet-50[13]을 활용하여 특징맵을 추출하였다.

먼저 각 이미지는 백본 네트워크에 입력값으로 들어간다. 그 후 첫 번째 반복의 회귀모델에 초기값에 해당하는 ϕ, θ, β, γ와 백본 네트워크에서 출력된 특징맵을 입력으로 받는다.

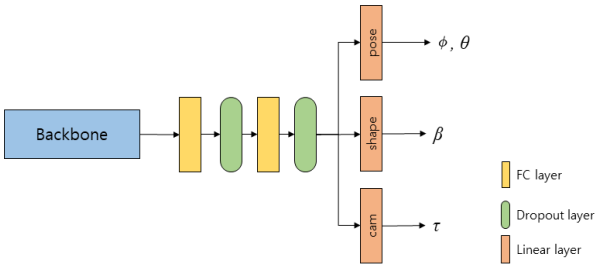


그림 2. 회귀모델 네트워크 구조
Fig. 2. Network architecture of regressor

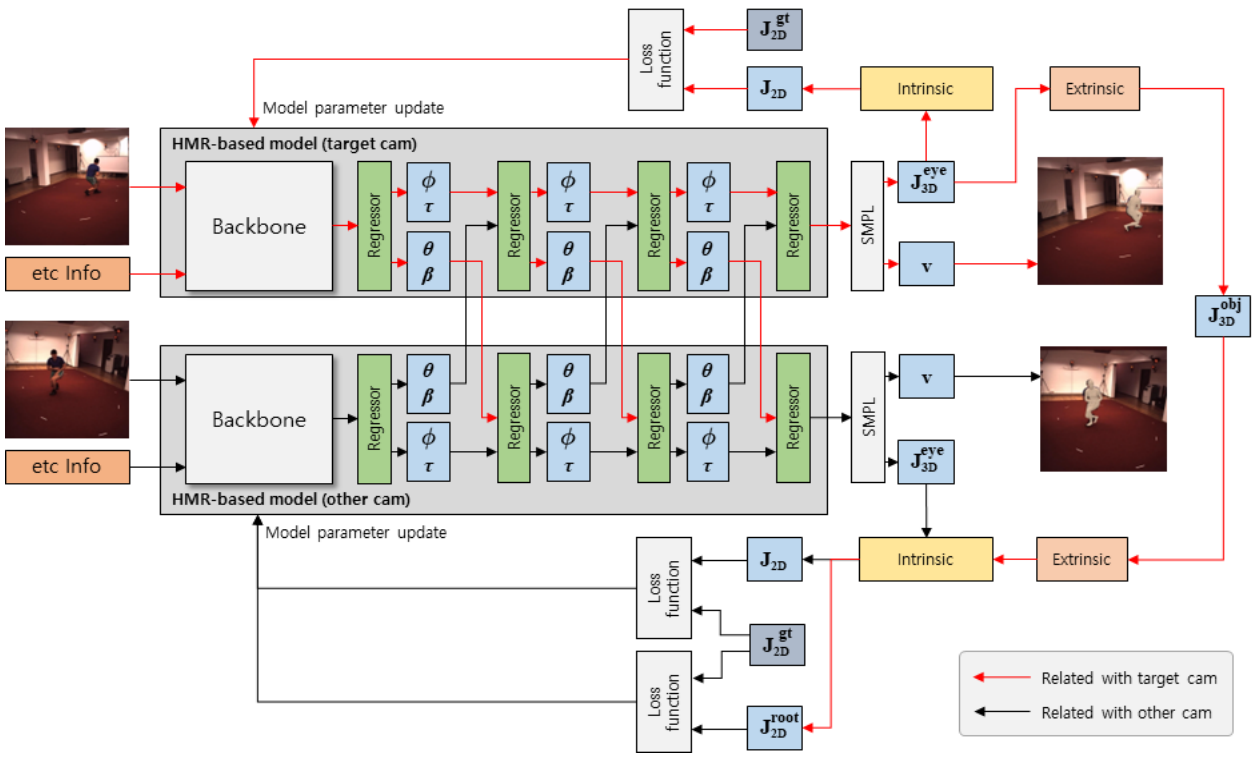


그림 1. 제안모델(MVHMR) 네트워크 구조
Fig. 1. Network architecture of proposed model(MVHMR)

이때, 초기값은 기존 HMR 모델과 마찬가지로 SMPL의 평균값을 활용하였다.

그 후, 첫 번째 회귀모델의 각 카메라에서 나온 출력값 ϕ, θ, β, γ 중 관절파라미터 θ와 형태파라미터 β를 서로 교차하여 다음 반복의 입력값으로 활용한다. 즉, 타겟 카메라 이미지에 대한 출력 ϕ^1, θ^1, β^1, γ^1의 θ^1, β^1와 다른 카메라 이미지에 대한 출력 ϕ^2, θ^2, β^2, γ^2의 θ^2, β^2를 서로 교차시켜 타겟 카메라 이미지를 입력으로 받는 회귀모델에는 ϕ^1, θ^2, β^2, γ^1가 입력값으로 들어가고 다른 카메라의 이미지를 입력으로 받는 회귀모델에는 ϕ^2, θ^1, β^1, γ^2가 입력값으로 들어가게 된다. 이 과정을 총 3번 반복하여 추정과정을 진행하도록 네트워크를 구성하였다. ϕ는 중심각도에 해당하며 이는 카메라 위치 및 각도에 종속되므로 교차하는 과정에서 ϕ값은 제외하고 θ값을 교차하였다.

IV. 제안 모델 성능평가 및 비교분석

4.1 평가 방법

성능평가를 위해 각 데이터에 매칭되는 관절 위치를 추출하고 아래 평가지표를 통해 성능을 측정하였다.

MPJPE(Mean Per Joint Position Error) 관절의 예측 정확도를 평가하는 지표로써 정답값과 예측값의 유클리디안 거리를 계산한다. 먼저 정답값의 중심좌표 기준으로 예측값을 정렬한 후 정답값과 예측값의 각 관절거리의 평균을 계산한다.

PA-MPJPE(Procrustes Alignment Mean Per Joint Position Error) 크기, 위치, 회전 등이 제외된 자세의 유사성을 측정하기 위한 지표로써 프로크러스테스 변환(Procrustes alignment)을 적용한 후 관절의 유클리디안 거리의 평균을 계산한다.

4.2 데이터셋

본 연구에서 제안한 멀티뷰 방식의 모델의 성능을 평가하기 위해 다중카메라를 이용하여 만들어진 데이터들을 활용하였다.

Human3.6M[14] 3차원 관절좌표 추론을 위한 데이터셋으로 실내에서 이루어진 여러개의 동작으로 데이터셋이 구성되어 있다. 동작은 식사하기, 앉기, 걷기 등으로 이루어져 있으며 4개의 카메라로부터 받은 이미지와 카메라 파라미터 등을 포함하고 있다. 본 연구에서는 다중카메라가 필요하기 때문에 평가데이터로써 본 데이터를 활용하였으며 이 중 9번 시나리오와 11번 시나리오를 선정하였다. 또한 멀티뷰 카메라 테스트를 위해 4개의 카메라 중 타겟 카메라와 나머지 3개 카메라 중 랜덤으로 하나를 선택하여 다른 카메라로 활용하였다. 평가 관절은 SMPL의 관절위치와 Human3.6M과 일치하는 관절 14개를 추출하여 평가에 반영하였다.

NIA 다중 객체 3차원 표현 데이터(NIA3D) 2022년 한국 지능정보사회진흥원(NIA) 주관사업으로 구축된 사람과 상호작용하는 다양한 3D 객체 인공지능 학습용 데이터셋으로써 일상 환경을 가정하여 실내/실외 환경에서의 경계박스, 세그멘테이션, 2D/3D 관절값, 카메라 데이터 등을 포함하는 데이터셋이다. 다중카메라 환경에서 구축되었기 때문에 본 연구의 테스트 데이터로 적합하다 판단하여 평가데이터로 선정하였다. 2D/3D 관절값은 29개 관절로 구성되어 있으며 이에 맞게 SMPL의 메쉬 포인트로부터 해당관절과 일치하는 29개 점을 추출하여 평가에 반영하였다. 카메라는 위 Human3.6M 데이터셋과 마찬가지로 타겟카메라와 나머지 2개 카메라 중 랜덤으로 하나를 선택하여 다른 카메라로 활용하였다. 또한 SMPL은 성인 남녀를 대상으로 구축된 모델이기 때문에 150cm 이하의 사람의 경우 소인으로 판단하여 평가데이터셋에서 제외하였다. 평가의 효율을 위해 실내/실외 데이터로 구분하여 최종적으로 각 데이터셋의 10%를 평가데이터로 선정하였다.

4.3 MVHMR model

현재 높은 성능을 보이는 모델인 CLIFF[9]에 제안방식을 적용하여 성능평가를 진행하였으며 이를 위해 해당 모델의 깃헙 저장소에서 제공하는 사전학습 모델을 다운받아 활용하였다.

실험은 백본 네트워크 종류로 나눠서 진행하였다. 2차원 관절 정답값을 적용하는 방식은 아래 표 1의 단계와 같이 진행하였으며 [15]연구에서 적용한 방식과 유사한 방식으로 적용하였다.

2차원 관절좌표를 이용한 미세조정을 위한 손실함수 아래 식 (3)에 해당한다. 2차원 관절좌표의 손실값은 식 (1)과 같이 타겟 카메라의 이미지에서 추정된 2차원 관절좌표와 정답값 사이의 L1 loss와 다른 카메라 이미지에서 추정된 2차원 관절좌표와 정답값과의 L1 loss의 합으로 구성하였다.

$$J_{2D} = \Pi J_{3D}, \qquad L_{2D} = \| J_{2D} - \hat{J}_{2D} \| \tag{1}$$

$$J_{2D}^{target \rightarrow other} = \prod_{other} (E_{other}(E_{target}^{-1}(J_{3D}^{target}))) \tag{2}$$

$$L_{root}^{target} = \| S(J_{2D}^{target}, idx) - S(\hat{J}_{2D}^{target}, idx) \|$$

$$L_{root}^{converted} = \| S(J_{2D}^{other}, idx) - S(\hat{J}_{2D}^{target \rightarrow other}, idx) \|$$

$$L^{MVHMR} = \lambda_{2D}^{target} L_{2D}^{target} + \lambda_{2D}^{other} L_{2D}^{other} + \lambda_{root}^{target} L_{root}^{converted} + \lambda_{root}^{converted} L_{root}^{converted} \tag{3}$$

식 (2)는 깊이정보에 해당하는 중심관절좌표의 손실값을 반영하기 위한 손실함수이다. $\prod$은 원근투영 함수이고 $S(J,idx)$는 관절 중 신체의 중심에 해당하는 좌표를 선택하는 함수로 본 연구에서는 골반에 해당하는 좌표를 중심관절좌표로 선택하였다. $J_{2D}^{target \rightarrow other}$는 J_{3D}^{target}에 타겟 카메라의 외부 매트릭스 E_{target}의 역행렬을 곱해 월드좌표계에서의 관절위치를 구한 후 다시 다른 카메라의 외부 매트릭스 E_{other}를 곱한 후 이미지 평면에 투영하여 다른 카메라에서의 2차원 관절좌표를 구한 값이다. 이 과정은 타겟 카메라 이미지로부터 추정된 3차원 중심관절좌표를 다른 카메라에 투영하여 2차원 중심관절좌표를 구하는 과정이다. 이를 다른 카메라에서의 중심관절좌표 정답값과 오차를 구하여 손실값에 추가하였다. 해당 과정은 표 1의 4번과 5번 과정에 해당한다. 이를 통해 깊이에 해당하는 값을 좀 더 정교하게 추정할 수 있도록 하였다.

표 1. 2차원 관절 정답값를 활용한 미세조정 알고리즘
Table 1. Fine-tuning algorithm with 2D keypoint gound true

1. Initialize the model with pre-trained weights.
2. Estimate SMPL parameters using the model.
3. Apply the estimated SMPL parameters to obtain 3D joint coordinates.
4. Transform the estimated 3D joint coordinates from camera coordinates (eye coordinate) to world coordinates (object coordinate) using the camera extrinsic matrix.
5. Further transform the transformed world coordinates into each camera's camera coordinates using the camera extrinsic parameters, and then project them onto 2D joint coordinates using the camera intrinsic matrix.
6. Calculate the error between the obtained 2D joint coordinates and the ground truth 2D joint coordinates for each camera.
7. Fine-tune the model using the optimization method that minimizes the error calculated in step 6. This process is repeated for all images in a batch.

λ_{2D}^{target}과 λ_{2D}^{other}는 5, λ_{2D}^{target}과 $\lambda^{converted}{}_{2D}$는 0.005, 배치 크기는 64, 옵티마이저는 아담 옵티마이저를 사용하였다. 학습률의 초기값은 5e-5으로 하고 학습횟수는 총 60번을 주고 45번째 학습부턴 학습률을 5e-6으로 하여 최적화 하였다. 미세조정과정에서 모델의 가중치를 업데이트할 때 배치정규화층과 드롭아웃층은 제외하였다.

아래 표 2와 표 3은 각각 Human3.6M 데이터셋과 NIA3D 데이터셋에서의 모델의 성능평가 결과이다. 성능지표는 MPJPE와 PA-MPJPE를 활용하였으며 평가 검증을 위해 현재 좋은 성능을 보이는 모델들을 함께 비교하였다.

표 2와 같이 제안된 모델의 평가결과 Human3.6M 데이터셋에서 MPJPE는 ResNet50 기준으론 3.2, HRNet-W48 기준으론 0.28의 성능향상이 있었고 PA-MPJPE는 ResNet50 기준으론 5.65, 3.27로 큰 폭으로 성능이 향상되었다. 표 3과 같이 NIA3D데이터셋에서 또한 실내 데이터의 MPJPE는 ResNet50 기준으론 1.6, 실외 데이터는 1.1의 성능향상이 있었으며 PA-MPJPE는 각각 1.9, 2.2의 성능이 향상되었다.

표 2. Human3.6M 데이터셋에서의 MVHMR 모델과 기존 모델 비교 성능평가
Table 2. Performance comparison between MVHMR and previous methods on Human3.6M dataset

	Human3.6M	
Method	MPJPE	PA-MPJPE
HMR	-	56.8
SPIN	-	41.1
HMR-EFT	63.2	43.8
CLIFF(Res-50)	50.5	35.1
CLIFF(HR-W48)	47.1	32.7
(ours)MVHMR(Res-50)	**46.85**	**29.45**
(ours)MVHMR(HR-W48)	**46.82**	**29.43**

표 3. NIA3D 데이터셋에서의 MVHMR 모델과 기존 모델 비교 성능평가
Table 3. Performance comparison between MVHMR and previous methods on NIA3D dataset

	NIA3D Interior	
Method	MPJPE	PA-MPJPE
CLIFF(Res-50)	73.3	44.0
CLIFF(HR-W48)	73.3	43.6
(ours)MVHMR(Res-50)	**71.7**	**42.1**
(ours)MVHMR(HR-W48)	**71.4**	**41.4**
	NIA3D Exterior	
Method	MPJPE	PA-MPJPE
CLIFF(Res-50)	78.3	46.0
CLIFF(HR-W48)	82.9	48.2
(ours)MVHMR(Res-50)	**77.2**	**44.5**
(ours)MVHMR(HR-W48)	**81.9**	**46.7**

같은 데이터에 대해 HRNet-W48 기준으로 실내 데이터에선 MPJPE는 1.9, PA-MPJPE는 2.2의 성능향상이 있었고 실외 데이터는 각각 1과 1.5의 성능향상이 있었다. PA-MPJPE는 회전과 스케일을 제외한 순수한 자세의 예측 정확도를 측정하기 위한 지표로 해당 지표의 높은 정확도 향상은 특히 본 구조가 정확한 관절파라미터를 추정하는데 강점을 보임을 알 수 있다. 즉, 아래 그림 3과 같이 한 쪽 이미지에서 3차원으로 판단하기 힘든 이미지가 입력으로 들어왔을 때 본 연구에서 제안하는 방식으로 더 정밀한 예측이 가능하다. 입력 이미지에서는 오른쪽 팔꿈치가 가려졌기 때문에 손의 위치(붉은색 영역)를 정확히 예측하기 힘드나 다른 이미지를 입력으로 같이 활용하면 이를 더 정확하게 추정할 수 있다.

V. 결론 및 향후 과제

본 연구에서는 모델의 반복 추정 과정에서 출력값을 교체하는 방식을 통해 다중카메라에서 동작하는 사람관절 및 형태 추정 모델을 구축하는 방식을 제안하였다. 사람과 관절과 형태에 관여하는 출력값을 교차하여 다음 반복의 입력으로 넣음으로써 타겟 카메라에선 가려져 있으나 다른 카메라에서 보이는 관절이 있을 경우 다른 카메라 이미지를 입력값에 반영하여 관절을 좀 더 정확하게 추정할 수 있도록 하였다. 본 제안의 타당성 검증을 위해 HMR기반의 방식 중 좋은 성능을 보이는 CLIFF 모델에 본 제안방식을 적용하여 성능평가를 진행하였다. 또한 멀티뷰 방식 적용을 위해 다중카메라 기반의 데이터를 선정하여 성능평가를 진행하였다. 제안한 방식을 통해 기존 모델의 성능을 Human3.6M 데이터에서 Resnet50 기준 MPJPE는 7%, PA-MPJPE는 16% 정도의 성능이 개선되었다.

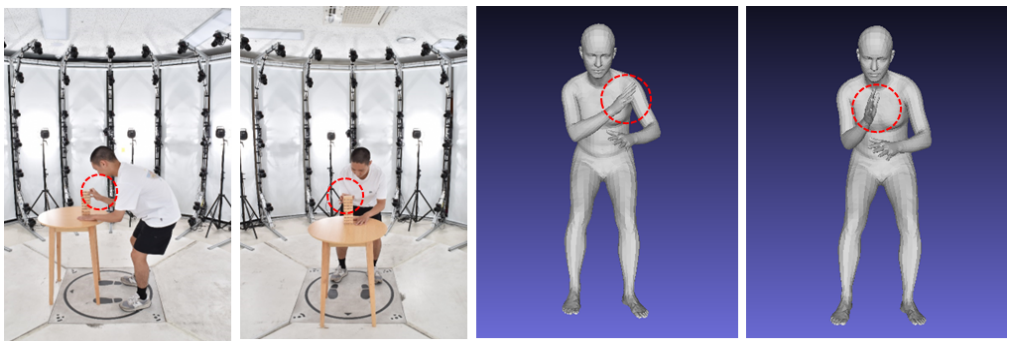


그림 3. 모호한 이미지에서의 결과값 예시 (왼쪽부터 입력이미지, 다른카메라에서 이미지, 기존모델 출력, 제안모델 출력)
Fig. 3. Example of results for ambiguous input image (From left to right: input images, other camera image, previous model result, proposed model result)

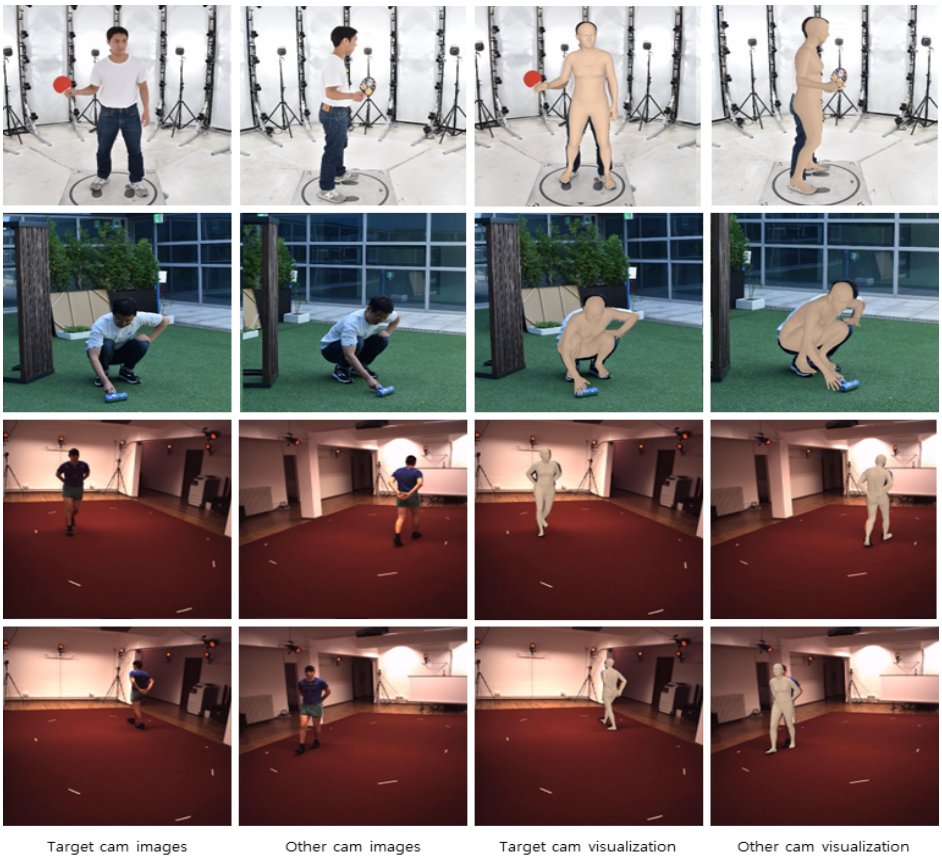


그림 4. 결과 예시 (왼쪽부터 타겟이미지, 다른카메라 이미지, 타겟카메라 시각화, 다른카메라 시각화)
Fig. 4. Example of results (From left to right: target camera images, other camera images, target camera visualization, other camera visualization)

본 연구에서 제안한 방식을 통해 캘리브레이션이 된 환경에서 구축된 데이터에 2쌍의 2차원 관절좌표로도 정확한 3차원 관절 및 메쉬를 추출할 수 있다. 또한 제안 방식은 HMR과 같이 반복 추정 구조로 되어 있는 모델들에 모두 적용할 수 있는 구조로 다양한 모델에 응용할 수 있다.

현재 본 제안방식은 캘리브레이션 된 2개 카메라에서만 적용이 가능하다. 이를 개선하여 카메라 대수에 비례하여 성능을 높일 수 있는 방식이나 캘리브레이션 되지 않은 환경에서 성능을 높일 수 있는 방법을 향후 연구로 진행하고자 한다.

References

[1] S. I. Mun, D. H. Kim, H. J. Chang, and J. H. Hong, "Human Shape and Pose Estimation from an RGB-D Image", The Institute of Electronics and Information Engineers conference, pp. 292-296, 2022.

[2] M. Kocabas, C. H. P. Huang, O. Hilliges, and M. J. Black, "Pare: Part Attention Regressor for 3d Human Body Estimation", Proc. of the IEEE/CVF International Conference on Computer Vision, pp. 11127-11137, 2021.

[3] R. Khirodkar, S. Tripathi, and K. Kitani, "Occluded Human Mesh Recovery", 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), pp. 1715-1725, 2022.

[4] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: a skinned multi-person linear model", ACM Transactions on Graphics, Vol. 34, No. 6, pp. 1-16, Nov. 2015. https://doi.org/10.1145/2816795.2818013.

[5] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the Continuity of Rotation Representations in Neural Networks", Proc. of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition(CVPR), pp. 5745-5753, 2019.

[6] J. J. Kim and C. B. Kim, "Implementation of Robust License Plate Recognition System using YOLO and CNN", Journal of KIIT, Vol. 19, No. 4, pp. 1-9, 2021. https://doi.org/10.14801/jkiit.2021.19.4.1.

[7] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-End Recovery of Human Shape and Pose", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 7122-7131, 2018.

[8] N. Saini, E. Bonetto, E. Price, A. Ahmad, and M. J. Black, "AirPose: Multi-View Fusion Network for Aerial 3D Human Pose and Shape Estimation", IEEE Robotics and Automation Letters, Vol. 7, No. 2, pp. 4805-4812, Apr. 2022. https://doi.org/10.1109/LRA.2022.3145494.

[9] Z. Li, J. Liu, Z. Zhang, S. Xu, and Y. Yan, "CLIFF: Carrying Location Information in Full Frames into Human Pose and Shape Estimation", Computer Vision-ECCV 2022, pp. 590-606, 2022. https://doi.org/10.1007/978-3-031-20065-6_34.

[10] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop", Proc. of the IEEE/CVF International Conference on Computer Vision, pp. 2252-2261, 2019.

[11] S. Chun, S. Park, and J. Y. Chang, "Learnable Human Mesh Triangulation for 3D Human Pose and Shape Estimation", Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2850-2859, 2023.

[12] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation", CVPR, pp. 5693-5703, 2019.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Identity Mappings in Deep Residual Networks", Computer Vision-ECCV 2016, pp. 630-645, Sep. 2016. https://doi.org/10.1007/978-3-319-46493-0_38.

[14] C. Ionescu, D. Papava, V. Olaru, and 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, Vol. 36, No. 7, pp. 1325-1339, Jul. 2014. https://doi.org/10.1109/TPAMI.2013.248.

[15] H. Joo, N. Neverova, and A. Vedaldi, "Exemplar Fine-Tuning for 3D Human Model Fitting Towards In-the-Wild 3D Human Pose Estimation", 2021 International Conference on 3D Vision(3DV), London, United Kingdom, Dec. 2021. https://doi.org/10.1109/3DV53792.2021.00015.

저자소개

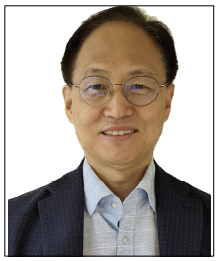
권 혁 민 (Hyeok-Min Kwon)

2019년 2월 : 강원대학교 컴퓨터정보통신공학(공학사)
2019년 3월 ~ 현재 : 강원대학교 데이터사이언스학과 석사과정
관심분야 : 컴퓨터비전, 인공지능, 사람관절추정, 머신러닝/딥러닝

이 준 호 (Jun-Ho Lee)

2003년 2월 : 아주대학교 전자공학과(공학사)
2022년 3월 ~ 현재 : 아주대학교 지능형소프트웨어학과 석사과정
관심분야 : 컴퓨터비전, 인공지능, 사람관절추정, 머신러닝/딥러닝

김 화 종 (Hwa-Jong Kim)



1982년 2월 : 서울대학교 전자공학과(공학사)
1984년 2월 : KAIST 전기 및 전자과(공학석사)
1988년 8월 : KAIST 전기 및 전자과(공학박사)
1988년 3월 ~ 현재 : 강원대학교 컴퓨터공학과 교수
관심분야 : 인공지능, 머신러닝/딥러닝, 연합학습