특집논문 (Special Paper)
방송공학회논문지 제27권 제2호, 2022년 3월 (JBE Vol.27, No.2, March 2022)
https://doi.org/10.5909/JBE.2022.27.2.174
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

다시점 준지도 학습 기반 3차원 휴먼 자세 추정

김 도 엽[a)], 장 주 용[a)‡]

Multi-view Semi-supervised Learning-based 3D Human Pose Estimation

Do Yeop Kim[a)] and Ju Yong Chang[a)‡]

요 약

3차원 휴먼 자세 추정 모델은 다시점 모델과 단시점 모델로 분류될 수 있다. 일반적으로 다시점 모델은 단시점 모델에 비하여 뛰어난 자세 추정 성능을 보인다. 단시점 모델의 경우 3차원 자세 추정 성능의 향상은 많은 양의 학습 데이터를 필요로 한다. 하지만 3차원 자세에 대한 참값을 획득하는 것은 쉬운 일이 아니다. 이러한 문제를 다루기 위해, 우리는 다시점 모델로부터 다시점 휴먼 자세 데이터에 대한 의사 참값을 생성하고, 이를 단시점 모델의 학습에 활용하는 방법을 제안한다. 또한, 우리는 각각의 다시점 영상으로부터 추정된 자세의 일관성을 고려하는 다시점 일관성 손실함수를 제안하여, 이것이 단시점 모델의 효과적인 학습에 도움을 준다는 것을 보인다. Human3.6M과 MPI-INF-3DHP 데이터셋을 사용한 실험은 제안하는 방법이 3차원 휴먼 자세 추정을 위한 단시점 모델의 학습에 효과적임을 보여준다.

Abstract

3D human pose estimation models can be classified into a multi-view model and a single-view model. In general, the multi-view model shows superior pose estimation performance compared to the single-view model. In the case of the single-view model, the improvement of the 3D pose estimation performance requires a large amount of training data. However, it is not easy to obtain annotations for training 3D pose estimation models. To address this problem, we propose a method to generate pseudo ground-truths of multi-view human pose data from a multi-view model and exploit the resultant pseudo ground-truths to train a single-view model. In addition, we propose a multi-view consistency loss function that considers the consistency of poses estimated from multi-view images, showing that the proposed loss helps the effective training of single-view models. Experiments using Human3.6M and MPI-INF-3DHP datasets show that the proposed method is effective for training single-view 3D human pose estimation models.

Keyword : 3D human pose estimation, Semi-supervised learning, Multi-view consistency, Deep learning

a) 광운대학교 전자통신공학과(Department of Electronics and Communications Engineering, Kwangwoon University)
‡ Corresponding Author : 장주용(Ju Yong Chang)
E-mail: jychang@kw.ac.kr
Tel: +82-2-940-5136
ORCID: https://orcid.org/0000-0003-3710-7314

※ 이 논문의 결과 중 일부는 한국방송·미디어공학회 “2021년 추계학술대회”에서 발표한 바 있음.
※ This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2021-0-00348, Development of A Cloud-based Video Surveillance System for Unmanned Store Environments using Integrated 2D/3D Video Analysis, 90%) and the Excellent researcher support project of Kwangwoon University in 2021 (10%).
· Manuscript received January 17, 2022; Revised February 15, 2022; Accepted February 15, 2022.

Ⅰ. 서 론

3차원 휴먼 자세 추정(3D human pose estimation)은 대상 휴먼 객체의 3차원 관절 좌표를 추정하는 것을 그 목표로 가지며, 컴퓨터 비전 분야에서 주요하게 연구되는 분야 중 하나이다. 딥러닝 기술의 발전과 함께 최근 3차원 휴먼 자세 추정 방법의 성능이 크게 향상되었고, 그 결과 AR, VR, 엔터테인먼트, 헬스케어 산업 등에서 활발하게 사용되고 있다. 그러나 영상으로부터 3차원 휴먼 자세를 추정하는 방법은 여전히 어려운 일이며, 특히 단일 시점의 영상 또는 비디오를 입력으로 하는 방법의 경우 깊이 모호성(depth ambiguity)은 해결해야 할 어려운 문제들 중 하나이다. 우리는 영상으로부터의 3차원 휴먼 자세 추정 방법을 입력에 따라서 크게 다시점 모델(multi-view model)과 단시점 모델(single-view model)로 구분한다.

한 자세에 대한 여러 카메라 시점의 영상을 입력으로 사용하는 다시점 모델은 단시점 모델보다 정확한 자세 추정이 가능하다. 그 이유는 다시점 모델이 3차원 휴먼 자세 추정 시 깊이 모호성 문제와 영상 시점에 따른 가리워짐(occlusion) 문제에 강인한 모델을 다시점 영상으로부터 학습할 수 있기 때문이다.

단시점 모델은 단일 시점의 영상 입력으로부터 3차원 휴먼 자세를 추정하는 방법이다. 단시점 모델은 한 장의 영상으로부터 3차원 자세를 추정할 수 있지만, 다시점 모델에 비하여 깊이 모호성 문제와 가리워짐 문제에 취약하다. 이러한 단시점 모델의 성능 개선은 다양한 시점과 자세를 포함하는 대량의 정제된 데이터를 필요로 한다. 그러나 3차원 자세에 대한 참값(GT; ground-truth)을 제공하는 데이터를 획득하는 일은 일반적으로 많은 시간과 비용을 필요로 한다.

본 논문에서 우리는 3차원 자세 GT가 제공되지 않는(unlabeled), 캘리브레이션된 다시점 데이터셋을 가정하고, 이러한 데이터셋을 활용하여 단시점 모델의 성능을 개선하는 방법을 제안한다. 기본적인 아이디어는 사전 학습된 다시점 모델[1]을 unlabeled 다시점 데이터셋에 적용하고, 그 추정 결과를 다시점 영상들에 대응하는 3차원 휴먼 자세에 대한 의사 참값(P-GT; pseudo-GT)으로서 단시점 모델의 학습에 활용하는 것이다. 또한 우리는 다시점 영상에 대한 단시점 모델의 자세 추정 결과들에 일관성을 부여하는 다시점 일관성 손실함수(multi-view consistency loss)를 제안한다. 이는 단시점 모델의 3차원 깊이 추정 성능과 가리워짐 발생 시 휴먼 자세 추정 성능을 개선한다.

우리는 제안하는 3차원 휴먼 자세 추정 방법을 정량적, 정성적으로 평가한다. 그리고 평가 결과로부터 다시점 모델로부터 획득된 P-GT가 단시점 모델의 학습 및 성능 개선에 활용될 수 있음을 보인다.

Ⅱ. 관련 연구

1. 3차원 휴먼 자세 추정을 위한 다시점 모델

다시점 영상은 사람의 자세를 다양한 시점에서 제공하기에 자세에 대한 모호성을 크게 개선할 수 있는 정보를 포함한다. 특히 triangulation을 활용한 다시점 모델[1,2]들은 일반적인 단시점 모델에 비하여 매우 높은 정확도를 보인다. [2]는 다시점 영상에서 추정된 2차원 자세들에 대해 epipolar 기하학을 적용하여 3차원 자세를 복원하고 이를 학습에 사용한다. [1]은 3차원 공간에서 다시점 영상들로부터 얻어진 feature맵을 triangulation하는 volumetric triangulation 방법을 제안하였다. 또한 다시점에서 획득된 2차원 자세에 대한 고전적인 algebraic triangulation[3] 기반의 방법으로도 높은 정확도의 3차원 자세 추정이 달성 가능함을 보였다.

2. 3차원 휴먼 자세 추정을 위한 단시점 모델

단시점 모델에 대한 선행 연구들은 세부적으로 입력 단시점 영상으로부터 직접 영상 내 대상의 3차원 휴먼 자세를 추정하는 방법[4,5,6]과 영상으로부터 2차원 휴먼 자세 추정을 수행한 후 2차원 휴먼 자세로부터 3차원 휴먼 자세를 추정하는 리프팅(lifting) 방법으로 나뉜다[7,8,9]. 본 연구에서 제안하는 방법에서 사용하는 단시점 모델은 입력 단시점 영상에서 직접 자세 추정을 수행하는 방법에 속한다. [4]는 Stacked-hourglass[10] 구조에 기반하여, 3차원 자세를 복셀(voxel) 공간에서 coarse-to-fine 방식으로 추정하는 방법을

제안하였다. [5]는 관절의 확률 분포를 가지는 3차원 히트맵(heatmap)에서 argmax를 취하는 기존 방법과 달리 기댓값을 취하는 soft-argmax를 사용하여 간단하지만 효과적인 히트맵 회귀 기반 방법을 소개하였다. [4,5]에서 제안된 모델은 3D GT 기반의 지도학습을 통해 학습된다. [6]은 한 사람에 대한 다시점의 in-the-wild 영상으로부터 weak-supervision을 통해 모델을 학습한다. [6]은 단시점 모델의 학습 시 카메라 외부 파라미터 캘리브레이션과 3차원 자세 GT를 요구하지 않는 이점을 가진다. 본 연구는 카메라 내, 외부 파라미터 캘리브레이션과 3차원 자세 GT를 가정하며, 다시점 모델로부터 획득한 P-GT가 단시점 모델을 학습하기 위한 추가적인 데이터로써 성능 개선에 도움이 될 수 있는지를 밝히는 것이 목표이다. 또한 본 연구에서는 단시점 모델의 학습을 위해 준지도 학습 방법이 사용된다.

Ⅲ. 제안하는 방법

본 연구에서 제안하는 단시점 모델의 성능 개선을 위해 unlabeled 다시점 영상 데이터셋을 활용하는 방법의 대략적인 절차는 다음과 같다. 첫 번째, 사전 학습된 다시점 모델로부터 P-GT를 생성한다. 두 번째, GT를 포함하는 labeled 데이터셋으로 사전 학습된 모델로 단시점 모델을 초기화한다. 세 번째, P-GT를 포함하는 unlabeled 다시점 영상 데이터셋 및 GT를 포함하는 labeled 데이터셋을 함께 사용하여 단시점 모델을 추가 학습한다. 이 때 단시점 모델을 최적

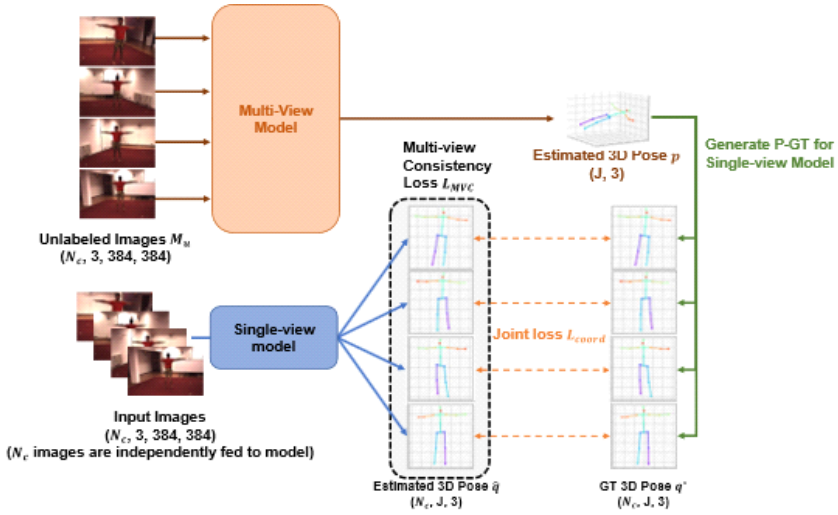

그림 1. 제안하는 다시점 준지도 학습 방법의 개요
Fig. 1. Overview of the proposed multi-view semi-supervised learning method

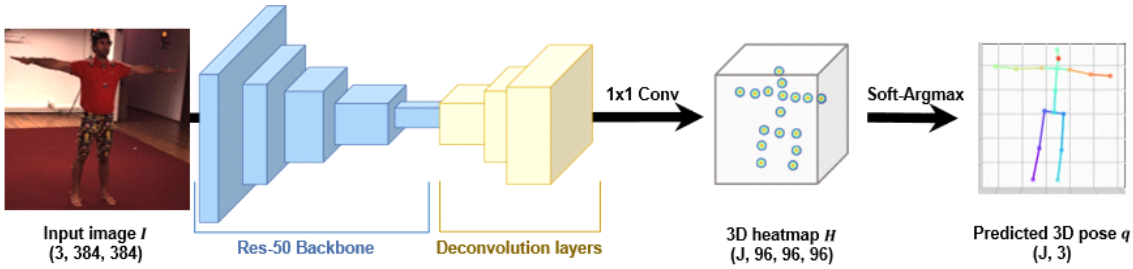

그림 2. 단시점 모델[5]의 구조
Fig. 2. The pipeline of the single-view model [5]

화하기 위해 제안하는 다시점 일관성 손실함수를 사용한다.

1. 단시점 모델의 사전학습

본 연구에서 3차원 휴먼 자세 추정을 수행하는 단시점 모델은 ImageNet[11]으로 사전 학습된 ResNet-50[12]을 백본(backbone)으로 하는 적분 회귀(integral regression) 모델[5]이다. [5]는 영상으로부터 3차원 휴먼 자세를 추정하는 방법들 중 높은 성능을 보였으며, 간단한 구현이 가능한 용이성 때문에 제안하는 방법을 위한 단시점 모델로써 선택되었다. 단시점 모델을 구성하기 위해 ResNet-50 모델에서 global average pooling 층과 fully connected 층을 3개의 연속된 deconvolution 층과 하나의 1x1 convolution 층으로 바꿔 fully convolutional 네트워크 F를 구성한다. 각 deconvolution 층의 필터 크기와 stride는 각각 4와 2로 설정한다. 단시점 모델은 F의 출력 텐서에 soft-argmax[5] 연산을 적용하여 3차원 휴먼 자세를 구성하는 관절 좌표를 획득한다. 본 연구에서 사용하는 단시점 모델의 구조는 그림 2에 나타나 있다.

네트워크 F는 입력 영상 $I \in \mathbb{R}^{3\times 384\times 384}$을 입력 받아 J개 관절에 대한 3차원 히트맵 $H_j \in \mathbb{R}^{96\times 96\times 96}(j=1,...,J)$를 출력한다. 그 후, H_j에 soft-argmax를 적용하여 각 관절의 3차원 좌표 $q_j=[q_{j,x},q_{j,y},q_{j,z}]^T \in \mathbb{R}^3$를 획득한다.

Soft-argmax 연산은 히트맵을 확률 분포로 만들고 기댓값(expectation)을 계산하여 공간 좌표를 획득하는 방법으로 다음 식은 j번째 관절에 대한 3차원 히트맵 $H_j(x,y,z)$를 확률 분포 $\tilde{H}_j(x,y,z)$로 만들기 위해 softmax 연산을 적용하는 과정을 나타낸다:

$$\tilde{H}_j(x,y,z)=\frac{e^{H_j(x,y,z)}}{\sum_{z'}\sum_{y'}\sum_{x'}e^{H_j(x',y',z')}}. \tag{1}$$

그 후 $\tilde{H}_j(x,y,z)$에 다음과 같은 기댓값 연산을 적용하여 3차원 관절 좌표 q_j를 획득한다:

$$\begin{cases} q_{j,x}=\sum_{x'}\sum_{y'}\sum_{z'}x'\tilde{H}_j(x',y',z') \\ q_{j,y}=\sum_{x'}\sum_{y'}\sum_{z'}y'\tilde{H}_j(x',y',z') \\ q_{j,z}=\sum_{x'}\sum_{y'}\sum_{z'}z'\tilde{H}_j(x',y',z') \end{cases}. \tag{2}$$

우리는 식 (2)를 모든 관절에 적용하여 3차원 휴먼 자세 $q=\{q_1,...,q_j\}$를 획득한다.

단시점 모델의 사전 학습을 위해 우리는 모델이 출력한 q에 labeled 데이터셋의 GT에 기반한 L1 손실 함수를 적용한다.

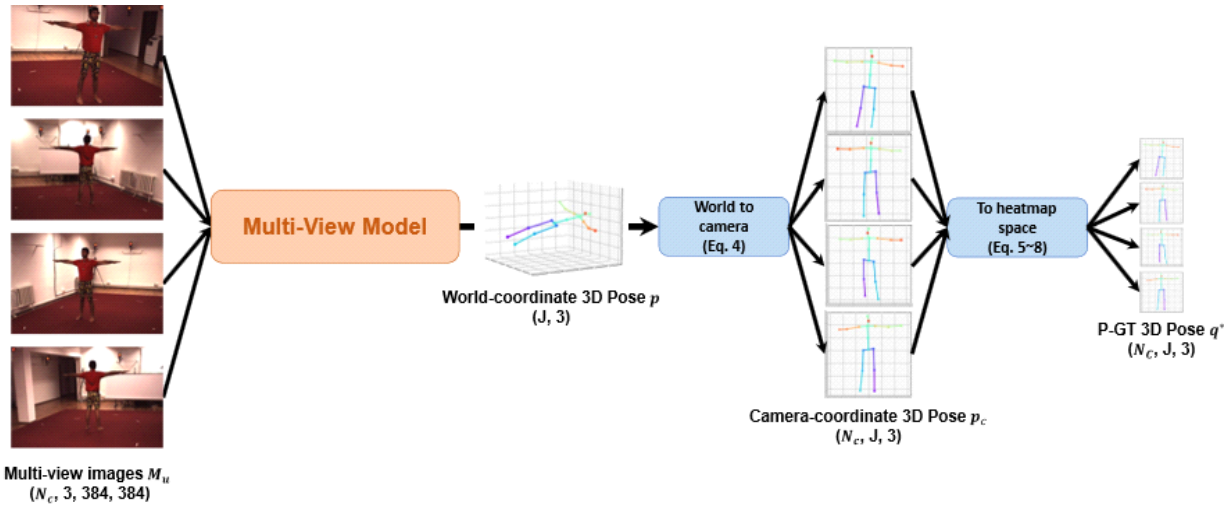

그림 3. P-GT의 생성 과정 개요
Fig. 3. Overview of P-GT generation process

2. P-GT 데이터셋 생성

우리는 unlabeled 다시점 영상 데이터셋을 활용하기 위하여 P-GT 데이터셋을 생성한다. P-GT 생성 과정은 그림 3에 나타나 있다. 먼저, 사전 학습된 다시점 모델 G로부터 J개 관절들의 3차원 좌표 $p=\{p_1,...,p_J\}$를 추정한다. 다시점 모델 G로 다시점 영상으로부터의 3차원 자세 추정 연구 중 높은 성능을 보이는 [1]의 방법을 사용한다. [1]의 algebraic triangulation 모델은 volumetric triangulation 모델보다 상대적으로 메모리 소모가 적고 정량적 성능에서 큰 차이가 나지 않기 때문에 본 연구에서는 algebraic triangulation 모델을 사전 학습하여 사용한다.

$$p=G(M_u), \tag{3}$$

여기서 $M_u=\{m_c\}_{c=1}^{N_c}\in\mathbb{R}^{N_C\times 3\times 384\times 384}$는 unlabeled 다시점 영상 데이터셋의 한 sample이다. m_c와 N_c는 각각 시점 c에서 촬영된 영상과 한 자세를 관찰하는 서로 다른 시점의 개수를 나타낸다. 여기서 3차원 휴먼 자세 p는 월드 좌표계(world coordinate system)에서 정의된다.

p를 단시점 모델의 학습에 활용하기 위해 우리는 p를 각 시점 c의 영상 m_c에 대응하는 히트맵 공간으로 정규화(normalization)한다. 이는 p의 m_c에 대한 원근 투영(perspective projection), 픽셀 좌표에 대한 정규화, 그리고 깊이 좌표에 대한 정규화 과정들로 이루어진다.

다음 식은 관절 좌표 p_j를 m_c에 투영하여 2차원 좌표를 획득하고, 픽셀 좌표에 대해 정규화 하는 과정을 나타낸다:

$$p_{j,c}=[p_{j,c,x},p_{j,c,y},p_{j,c,z}]^T=R_cp_j+t_c, \tag{4}$$

$$s\cdot\ [p_{j,c,u},p_{j,c,v},1]^T=K_cp_{j,c}, \tag{5}$$

$$[q_{j,c,x},q_{j,c,y}]^T=[\frac{p_{j,c,u}}{4},\frac{p_{j,c,v}}{4}]^T. \tag{6}$$

$R_c\in SO(3)$와 $t_c\in\mathbb{R}^3$는 카메라 c의 외부 파라미터를, 그리고 $K_c\in\mathbb{R}^{3\times 3}$는 내부 파라미터를 나타낸다. 식 (4)를 통해 월드 좌표계에서 정의된 p_j는 카메라 좌표계에서 정의된 관절 좌표 $p_{j,c}$로 변환된다. $p_{j,c}$는 식 (5)를 통해 픽셀 좌표 $[p_{j,c,u},p_{j,c,v}]^T$로 투영된다. 마지막으로 식 (6)을 통해 우리는 영상 크기와 히트맵 크기 사이의 비율인 4를 사용하여 정규화된 2차원 좌표 $[q_{j,c,x},q_{j,c,y}]^T\in\mathbb{R}^2$를 얻을 수 있다. 또한 우리는 $p_{j,c}$의 깊이 좌표 $p_{j,c,z}$를 다음과 같이 정규화한다:

$$q_{j,c,z}=\left(\frac{p_{j,c,z}-p_{pelvis,c,z}}{1000}+1\right)\times 0.5\times 96. \tag{7}$$

각 관절의 깊이 좌표를 정규화 하기 위하여 우리는 먼저 휴먼 객체의 크기가 2000 mm^2이하임을 가정한다. $p_{j,c,z}$에서 골반 관절의 깊이 값인 $p_{pelvis,c,z}$를 빼서 골반 관절을 기준으로 상대적으로 정의되는 깊이 값을 얻는다. 이러한 깊이 값은 $[-1000,1000]$의 범위에 존재하므로 우리는 추가적으로, 정규화된 깊이 값이 히트맵의 뎁스 축 범위인 $[0,96]$ 내에 존재하도록 만든다. 이러한 과정은 식 (7)에 나타나 있으며, 이를 통해 우리는 정규화된 깊이 값 $q_{j,c,z}$를 획득한다. 결국 시점 c에 대응하는 96x96x96 크기의 히트맵 공간에서 정의되는 P-GT 관절 좌표 $q_{j,c}^*$는 다음과 같다:

$$q_{j,c}^*=[q_{j,c,x},q_{j,c,y},q_{j,c,z}]. \tag{8}$$

식 (4)-(8)을 각 시점 c에 적용하여 우리는 unlabeled sample M_u에 대한 P-GT $\{q_c^*\}_{c=1}^{N_c}$를 획득할 수 있고, 이를 활용하여 우리는 P-GT 데이터셋 $M=\{m_c,q_c^*\}_{c=1}^{N_c}$을 구성한다.

3. 다시점 일관성 손실 함수

제안하는 방법은 다시점 영상 데이터를 사용하여 단시점 모델을 학습한다. 단시점 모델이 한 자세에 대응하는 다시점 영상을 입력 받는 경우, 모델에 의해 추정된 각 시점에서의 휴먼 자세들은 일관된 자세를 취해야 한다. 따라서 우리는 학습된 모델로 하여금 이러한 조건을 만족시키게끔 하기 위해 다시점 일관성 손실함수를 제안한다. 다시점 일관성 손실함수는 각 시점에 대해 추정된 관절 좌표들을 월드

좌표계 기준으로 변환하고, 그 결과로 얻어지는 자세들 사이의 L1 손실 함수들의 합으로 정의된다. 히트맵 공간으로 정규화된 관절 좌표를 월드 좌표계로 변환하는 과정은 식 (4)-(8)의 역 연산으로 수행된다. 이제 다시점 일관성 손실 함수 L_{MVC}는 다음과 같다:

$$L_{MVC}=\frac{1}{J}\Sigma_{j=1}^{J}\Sigma_{c\neq c'}\parallel \Pi_c(\hat{q}_{j,c})-\Pi_{c'}(\hat{q}_{j,c'})\parallel_1, \quad (9)$$

여기서 $\hat{q}_{j,c}$와 $\hat{q}_{j,c'}$는 각각 시점 c와 c'에서 단시점 모델에 의해 추정된 정규화된 관절 좌표를 나타내며, Π는 정규화된 관절 좌표를 월드 좌표계 기준으로 변환하는 함수이다.

우리는 사전 학습된 단시점 모델을 미세 조정(fine-tuning) 하기 위한 손실 함수 L을 다음과 같이 정의한다:

$$L=\alpha L_{coord}+\beta L_{MVC}, \quad (10)$$

여기서 L_{coord}는 단시점 모델의 출력 $\hat{q}$에 적용하는 P-GT 및 GT에 기반한 L1 손실 함수들의 합으로 정의된다. α와 β는 각 손실 함수의 영향력을 결정하는 가중치이다. 단시점 모델의 미세 조정을 위해 우리는 식 (10)을 최소화한다.

Ⅳ. 실험 결과 및 분석

1. 데이터셋, 평가 방법, 구현 세부사항

본 연구는 제안하는 방법을 학습 및 평가하기 위하여 대규모의 3차원 휴먼 자세를 포함하는 Human3.6M[13] 및 MPI-INF-3DHP[14] 데이터셋을 사용한다.

Human3.6M 데이터셋에서 각 휴먼 객체(subject, 이하 S)는 15가지의 동작을 수행하며 각 휴먼 객체가 동작을 수행하는 비디오를 4개의 서로 다른 시점의 카메라(camera, 이하 Cam)로 촬영한다. 우리는 기존 연구들[7,8]의 학습 및 평가 방법에 따라 11명 중 5명(S1, S5, S6, S7, S8)의 인물에 대한 데이터를 학습 데이터셋으로 사용한다. 이 중 3명(S1, S5, S6)의 데이터는 labeled 데이터셋으로, 2명(S7, S8)의 데이터는 unlabeled 데이터셋으로 가정한다. 나머지 2명(S9, S11)의 데이터는 평가 데이터셋으로 사용한다. 본 연구에서는 데이터셋이 제공하는 4개 시점의 비디오(Cam1, Cam2, Cam3, Cam4)를 데이터로 사용한다. 평가 시, 평가 데이터셋을 64프레임 마다 서브 샘플링(sub-sampling)하여 사용한다.

MPI-INF-3DHP 데이터셋은 학습 데이터셋과 평가 데이터셋으로 구성된다[14]. 본 연구에서는 학습 데이터셋만을 사용하여 모델을 학습하고 평가하며, 이후 본문에서는 이를 MPI-INF-3DHP 데이터셋이라고 통칭한다. MPI-INF-3DHP 데이터셋은 8명의 휴먼 객체(S1, S2, S3, S4, S5, S6, S7, S8)가 동작을 수행하는 비디오를 14개의 서로 다른 시점의 카메라로 촬영한다. 이 중 8개 시점에서 촬영된 비디오(Cam0, Cam1, Cam2, Cam4, Cam5, Cam6, Cam7, Cam8)가 제공된다. 먼저, 학습 및 평가 데이터셋의 구성을 위하여 8명의 휴먼 객체 중 3명(S1, S2, S3)의 데이터를 labeled 데이터셋으로, 4명(S4, S5, S6, S7)의 데이터는 unlabeled 데이터셋으로 가정한다. 나머지 1명(S8)의 데이터는 평가 데이터셋으로 사용한다. 또한 본 연구에서는 각 휴먼 객체에 대하여 제공되는 8개 시점의 비디오 중 4개의 시점(Cam0, Cam2, Cam7, Cam8)에서 촬영된 비디오를 데이터로써 사용하였다. 본 연구에서는 MPI-INF-3DHP 데이터셋에서의 평가 시 평가 데이터셋에 대한 서브 샘플링을 수행하지 않는다.

우리는 단시점 모델의 성능을 정량적으로 평가하기 위해 MPJPE와 PA-MPJPE를 측정하여 보고한다. MPJPE는 평가 데이터셋에서 단시점 모델에 의해 추정된 관절과 그 GT 사이의 유클리드 거리를 나타낸다. PA-MPJPE는 추정된 관절과 GT 사이에 Procrustes alignment[15]를 수행한 후 MPJPE를 구한 값이다. MPJPE와 PA-MPJPE의 단위는 mm이다.

다시점 모델과 단시점 모델의 사전 학습은 labeled 데이터셋으로 수행된다. P-GT 데이터셋의 생성시 학습의 효율성을 위하여 사전 학습된 다시점 모델을 적용한 자세 추정 결과를 오프라인으로 저장한다. 그 후 단시점 모델의 미세 조정 시 사용한다.

다시점 모델의 사전 학습에서 에포크(epoch) 수, 배치 크기(batch size), 학습률(learning rate)은 Human3.6M에서 학

습 시 6, 8, 10^{-5}로 MPI-INF-3DHP에서 학습 시 10, 8, 10^{-5}로 각각 설정된다. 단시점 모델의 사전 학습의 경우 에포크 수, 배치 크기, 학습률은 Human3.6M에서 학습 시 20, 32, 10^{-4}로 MPI-INF-3DHP에서 학습 시 7, 32, 10^{-4}로 설정된다. 두 모델의 사전 학습에 사용된 optimizer는 Adam[16] 이다.

단시점 모델의 미세 조정을 위해 우리는 labeled 데이터셋과 P-GT 데이터셋으로 추가 학습한다. 학습 시 배치 크기와 학습률은 각각 6, 10^{-4}로 설정하며, 에포크 수는 Human3.6M 데이터셋과 MPI-INF-3DHP 데이터셋에서 각각 9, 10 에포크로 설정한다. Optimizer로는 Adam을 사용한다. 손실 함수의 가중치는 $\alpha=1$과 $\beta=0.1$로 설정한다. 제안하는 모델은 Pytorch[17] 프레임워크를 사용하여 구현되었다.

2. 정량적 평가 결과

우리는 제안하는 방법이 단시점 모델의 성능 개선에 도움을 주는 것을 보이기 위하여 2가지 baseline 모델들과 제안하는 방법(Ours)을 정량적으로 비교한다. Baseline 모델로는 다음의 2가지 방법을 사용한다. 첫 번째는 GT 데이터셋으로 학습된 단시점 모델(Base)이고, 두 번째는 다시점 일관성 손실 함수를 적용하지 않고 P-GT와 각 시점의 추정 결과에 L1 손실 함수만을 적용한 모델(L1-only)이다.

표 1, 2는 각각 Human3.6M과 MPI-INF-3DHP 평가 데이터셋에 대해 baseline 방법들과 제안하는 방법의 정량적 성능을 보여준다. 우리는 다시점 데이터에 대한 성능 개선을 보이기 위하여 각 평가 데이터셋의 각 카메라 시점에 대하여 성능을 평가하고 결과를 제시하였다.

P-GT를 추가하여 미세조정을 하는 것이 모델의 성능 개선에 도움을 준다는 것을 보이기 위하여, Base와 L1-only 모델의 성능을 비교한다. Human3.6M 평가 데이터셋의 모든 시점에서 L1-only baseline이 Base 보다 높은 성능을 보였다. MPI-INF-3DHP 평가 데이터셋에서는 Cam2를 제외한 모든 시점에서 L1-only baseline이 Base보다 낮은 MPJPE를 달성하였으며, 모든 시점에서 PA-MPJPE가 Base 보다 낮다. 이 결과는 기존의 다시점 모델이 생성하는 P-GT가 단시점 모델의 학습에 효과적이라는 것을 보여준다.

다시점 일관성 손실 함수가 모델의 성능을 개선함을 보이기 위하여 L1-only와 제안하는 방법(Ours)의 결과를 비

표 1. Baseline 모델과 제안하는 방법(Ours)의 Human3.6M 평가 데이터셋에 대한 성능 비교
Table 1. Performance comparison of baseline models and proposed method (Ours) on Human3.6M evaluation dataset

	Base		L1-only		Ours	
	MPJPE (mm)	PA-MPJPE (mm)	MPJPE (mm)	PA-MPJPE (mm)	MPJPE (mm)	PA-MPJPE (mm)
Cam1	107.78	84.72	78.34	62.9	76.92	59.36
Cam2	99.7	79.11	106.12	86.41	98.19	77.66
Cam3	116.7	93.13	79.27	62.22	76.26	57.86
Cam4	96.57	76.76	93.76	74.79	84.71	67.15

표 2. Baseline 모델과 제안하는 방법(Ours)의 MPI-INF-3DHP 평가 데이터셋에 대한 성능 비교
Table 2. Performance comparison of baseline models and proposed method (Ours) on MPI-INF-3DHP evaluation dataset

	Base		L1-only		Ours	
	MPJPE (mm)	PA-MPJPE (mm)	MPJPE (mm)	PA-MPJPE (mm)	MPJPE (mm)	PA-MPJPE (mm)
Cam0	162.61	139.48	142.47	115.56	132.93	104.17
Cam2	183.31	156.48	188.87	149.03	166.40	134.20
Cam7	195.5	156.44	182.94	139.73	175.75	131.98
Cam8	161.26	140.60	149.02	126.76	132.39	108.00

교한다. 표 1, 2로부터, 각 평가 데이터셋에서의 모든 카메라 시점에 대해 제안하는 방법은 L1-only baseline보다 높은 성능을 보인다. 우리는 이 결과로부터 P-GT 데이터셋과 다시점 일관성 손실함수가 단시점 모델의 3차원 휴먼 자세 추정 성능을 실질적으로 개선할 수 있음을 확인하였다. 본 연구에서는 적분 회귀에 기반한 단시점 모델이 사용되었지만, 제안하는 방법은 단일 영상에 대응하는 3차원 자세를 추정하는 모든 단시점 모델의 학습에 적용되어 그 성능을 향상시킬 수 있다.

3. 정성적 평가 결과

그림 4, 5는 각각 Human3.6M 평가 데이터셋과 MPI-INF-3DHP 평가 데이터셋에 대한 Base 모델과 제안하는 방법의 자세 추정 결과와 그 GT를 시각적으로 보여준다. 그림 4의 각 입력 영상들은 복잡한 자세와 가리워짐 등을 포함하여 상대적으로 어려운 자세로 구성되어 있다. 그림 4, 5에서 각 영상 위에 투영된 자세 추정 결과들(좌)과 GT에 덧그려진 추정 결과(우)를 보였으며, 이를 통해 우리는

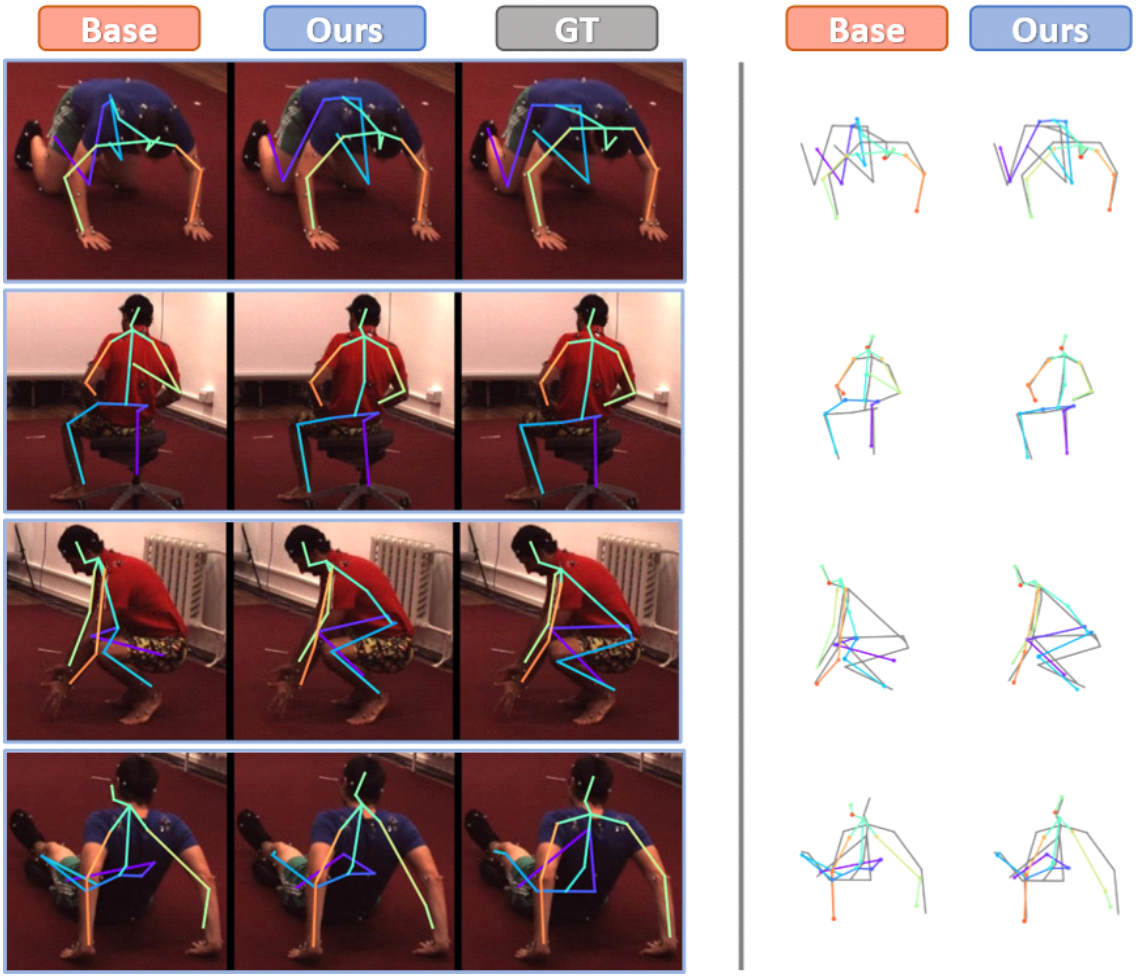

그림 4. Human3.6M의 평가 데이터셋에 대한 제안하는 방법(Ours)과 baseline(Base)의 정성적 결과. 입력 영상 위에 투영된 3차원 자세 추정 결과(좌), GT 자세(gray skeleton)와 각 방법으로 추정한 자세(colored skeleton)의 비교 결과(우)

Fig. 4. Qualitative comparison of the proposed method (Ours) and the baseline (Base) model on Human3.6M (left) evaluation datasets. 3D pose estimation results projected on the input images (left), and comparison results between the GT pose (gray skeleton) and the estimated pose (colored skeleton) by each method (right)

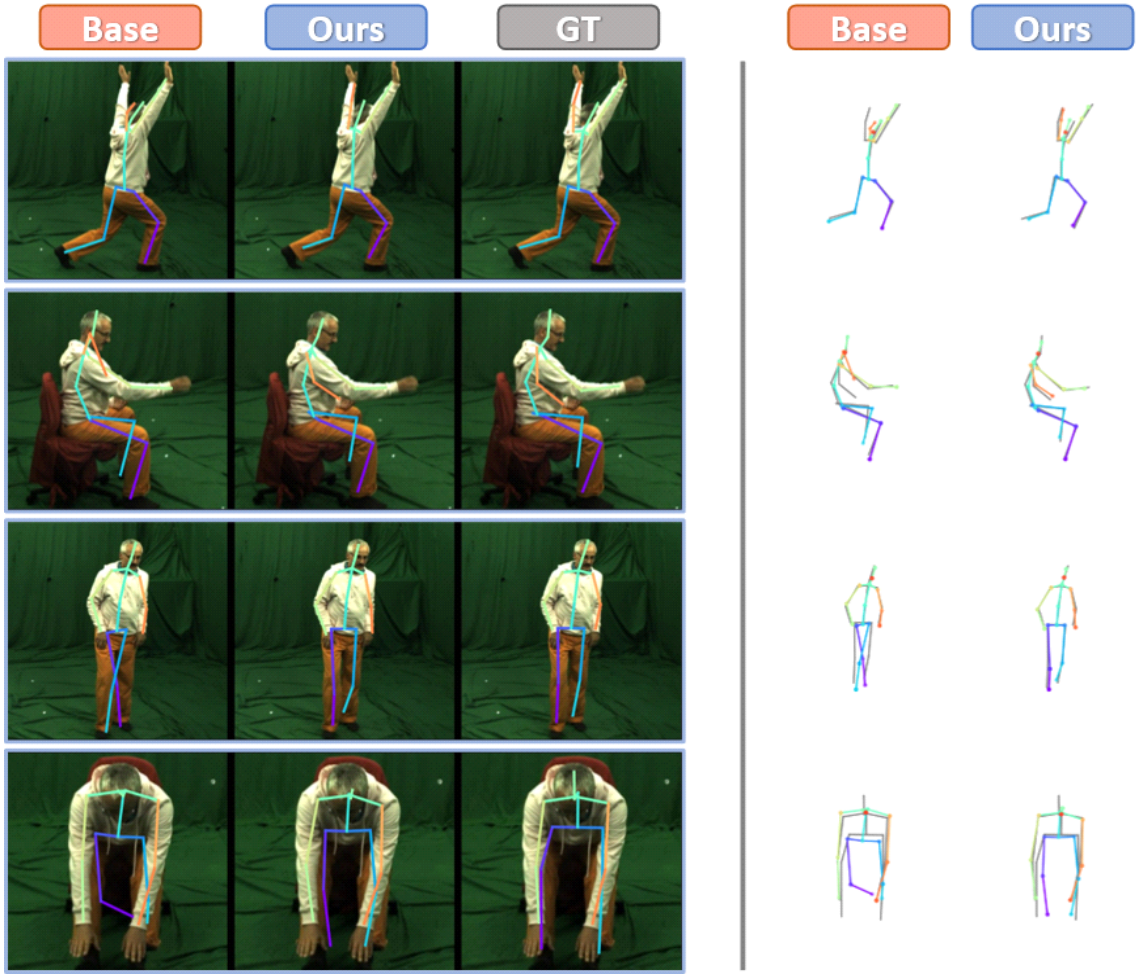

그림 5. MPI-INF-3DHP(우)의 평가 데이터셋에 대한 제안하는 방법(Ours)과 baseline(Base)의 정성적 비교. 입력 영상 위에 투영된 3차원 자세 추정 결과(좌), GT 자세(gray skeleton)와 각 방법으로 추정한 자세(colored skeleton)의 비교 결과(우)
Fig. 5. Qualitative comparison of the proposed method (Ours) and the baseline (Base) model on MPI-INF-3DHP (right) evaluation datasets. 3D pose estimation results projected on the input images (left), and comparison results between the GT pose (gray skeleton) and the estimated pose (colored skeleton) by each method (right)

Base 모델에 비하여 제안하는 방법이 GT에 보다 가까운 자세를 추정함을 알 수 있다.

V. 결 론

본 연구는 휴먼 객체의 3차원 자세 추정을 위한 단시점 모델의 성능을 개선하기 위해 캘리브레이션 된 unlabeled 다시점 데이터셋을 활용하는 준지도 학습 방법을 제안하였다. 제안하는 방법은 unlabeled 다시점 데이터에 사전 학습된 다시점 모델을 적용하여 P-GT를 생성하고, 이를 단시점 모델의 미세 조정에 활용한다. 또한 우리는 다시점 입력 영상에 대한 3차원 휴먼 자세 추정의 일관성을 고려하는 다시점 일관성 손실 함수를 제안하였다. 실험을 통해 우리는 기존의 사전 학습된 다시점 모델에 의해 생성된 P-GT와 다시점 일관성 손실 함수가 단시점 모델의 성능을 정량적, 정성적으로 향상시킴을 Human3.6M 데이터셋과 MPI-INF-3DHP 데이터셋에서 확인하였다.

참 고 문 헌 (References)

[1] K. Iskakov, E. Burkov, V. Lempisky, and Y. Malkov, “Learnable triangulation of human pose,” *IEEE International Conference on Computer Vision*, 2019.
doi: https://doi.org/10.1109/ICCV.2019.00781

[2] M. Kocabas, S. Karagoz, and E. Akbas, “Self-supervised learning of 3d human pose using multi-view geometry,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
doi: https://doi.org/10. 1109/CVPR.2019.00117

[3] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, Cambridge university press, 2003.
doi: https://doi.org/10.1017/ CBO9780511811685

[4] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3D human pose," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
doi: https://doi.org/10.1109/CVPR.2017.139

[5] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, “Integral human pose regression,” *European Conference on Computer Vision*, 2018.
doi: https://doi.org/10.1007/978-3-030-01231-1_33

[6] U. Iqbal, P. Molchanov, and J. Kautz, “Weakly-supervised 3d human pose learning via multi-view images in the wild,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
doi: https://doi.org/10. 1109/CVPR42600.2020.00529

[7] J. Martinez, R. Hossain, J. Romero, and J. J. Little, “A simple yet effective baseline for 3d human pose estimation,” *IEEE International Conference on Computer Vision*, 2017.
doi: https://doi.org/10.1109/ ICCV.2017.288

[8] H. Ci, C. Wang, X. Ma, and Y. Wang, “Optimizing network structure for 3D human pose estimation,” *IEEE International Conference on Computer Vision*, 2019.
doi: https://doi.org/10.1109/ICCV.2019.00235

[9] A. Zeng, X. Sun, F. Huang, M. Liu, Q. Xu, and S. Lin, “SRNet: Improving Generalization in 3D Human Pose Estimation with a Split-and-Recombine Approach,” *European Conference on Computer Vision*, 2020.
doi: https://doi.org/10.1007/978-3-030-58568-6_30

[10] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," *European Conference on Computer Vision*, 2016.
doi: https://doi.org/10.1007/978-3-319-46484-8_29

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in Neural Information Processing Systems*, pp. 1097-1105, 2012.
doi: https://doi.org/ 10.1145/3065386

[12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.
doi: https://doi.org/10.1109/CVPR.2016. 90

[13] 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, 2013.
doi: https://doi.org/10.1109/TPAMI.2013.248

[14] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, “Monocular 3D human pose estimation in the wild using improved CNN supervision,” *IEEE International Conference on 3D Vision*, pp. 506-516, 2017.
doi: https://doi.org/10.1109/3DV.2017.00064

[15] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, no. 2, 1975.
doi: https://doi.org/10.1007/BF02291478

[16] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *International Conference on Learning Representations*, 2015.
doi: https://doi.org/10.48550/arXiv.1412.6980

[17] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” *Advances in Neural Information Processing Systems Workshops*, 2017.
https://openreview.net/pdf?id=BJJsrmfCZ

저 자 소 개

김 도 엽

- 2019년 2월 : 광운대학교 전자통신공학과 학사
- 2019년 3월 ~ 현재 : 광운대학교 전자통신공학과 석박사통합과정
- ORCID : https://orcid.org/0000-0002-0624-5469
- 주관심분야 : 컴퓨터비전 및 머신러닝, Face Landmark Detection, 3D Shape Reconstruction

저 자 소 개

장 주 용

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