Regular paper | Journal of the KIECS. pp. 1321-1330, vol. 18, no. 6, Dec. 31. 2023, t. 122, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2023.18.6.1321

증강현실 캐릭터 구현을 위한 AI기반 객체인식 연구

이석환* · 이정금** · 심현***

AI-Based Object Recognition Research for Augmented Reality Character Implementation

Seok-Hwan Lee* · Jung-Keum Lee** · Hyun Sim***

요 약

본 연구는 증강현실에서 적용할 캐릭터 생성에서 단일 이미지를 통해 여러 객체에 대한 3D 자세 추정 문제를 연구한다. 기존 top-down 방식에서는 이미지 내의 모든 객체를 먼저 감지하고, 그 후에 각각의 객체를 독립적으로 재구성한다. 문제는 이렇게 재구성된 객체들 사이의 중첩이나 깊이 순서가 불일치 하는 일관성 없는 결과가 발생할 수 있다. 본 연구의 목적은 이러한 문제점을 해결하고, 장면 내의 모든 객체에 대한 일관된 3D 재구성을 제공하는 단일 네트워크를 개발하는 것이다. SMPL 매개변수체를 기반으로 한 인체 모델을 top-down 프레임워크에 통합이 중요한 선택이 되었으며, 이를 통해 거리 필드 기반의 충돌 손실과 깊이 순서를 고려하는 손실 두 가지를 도입하였다. 첫 번째 손실은 재구성된 사람들 사이의 중첩을 방지하며, 두 번째 손실은 가림막 추론과 주석이 달린 인스턴스 분할을 일관되게 렌더링하기 위해 객체들의 깊이 순서를 조정한다. 이러한 방법은 네트워크에 이미지의 명시적인 3D 주석 없이도 깊이 정보를 제공하게 한다. 실험 결과, 기존의 Interpenetration loss 방법은 MuPoTS-3D가 114, PoseTrack이 654에 비해서 본 연구의 방법론인 Lp 손실로 네트워크를 훈련시킬 때 MuPoTS-3D가 34, PoseTrack이 202로 충돌수가 크게 감소하는 것으로 나타났다. 본 연구 방법은 표준 3D 자세 벤치마크에서 기존 방법보다 더 나은 성능을 보여주었고, 제안된 손실들은 자연 이미지에서 더욱 일관된 재구성을 실현하게 하였다.

ABSTRACT

This study attempts to address the problem of 3D pose estimation for multiple human objects through a single image generated during the character development process that can be used in augmented reality. In the existing top-down method, all objects in the image are first detected, and then each is reconstructed independently. The problem is that inconsistent results may occur due to overlap or depth order mismatch between the reconstructed objects. The goal of this study is to solve these problems and develop a single network that provides consistent 3D reconstruction of all humans in a scene. Integrating a human body model based on the SMPL parametric system into a top-down framework became an important choice. Through this, two types of collision loss based on distance field and loss that considers depth order were introduced. The first loss prevents overlap between reconstructed people, and the second loss adjusts the depth ordering of people to render occlusion inference and annotated instance segmentation consistently. This method allows depth information to be provided to the network without explicit 3D annotation of the image. Experimental results show that this study's methodology performs better than existing methods on standard 3D pose benchmarks, and the proposed losses enable more consistent reconstruction from natural images.

키워드

3D Pose, AR, DeepLearning, HumanPose, Markerless AR, R-CNN, SMPL
3차원 자세, 증강 현실, 딥러닝, 휴먼포즈, 마커리스 증강 현실

* 순천대학교 (lsh76@scnu.ac.kr, missljk1004@scnu.ac.kr)
** 교신저자 : 순천대학교 스마트농업전공
· 접 수 일 : 2023. 10. 27
· 수정완료일 : 2023. 11. 19
· 게재확정일 : 2023. 12. 27
· Received : Oct. 27, 2023, Revised : Nov. 19, 2023, Accepted : Dec. 27, 2023
· Corresponding Author : Hyun Sim
Dept. Smart Agriculture, Sunchon National University
Email : simhyun@scnu.ac.kr

Ⅰ. 서 론

최근 메타버스 기술이 발달하면서 캐릭터의 실시간 객체 생성에 대한 연구가 활발히 이루어지고 있다. 특히 3D자세 분석 분야에서 다양한 연구가 이루어지고 있다. 현재의 연구 방향은 3D 키포인트 추정, 3D 형상 재구성, 전체 몸의 3D 자세와 형상 복구, 그리고 더 상세한 재구성 추정에서 뛰어난 성과를 나타내고 있다. 메타버스와 그 안에서 활동하는 캐릭터들에 대한 중첩 문제 해결이 필요하며, 이를 위해 단일 이미지에서 여러 사람의 3D 재구성이 중점적으로 연구되고 있다[1-2]. 다중 인물의 자세 추정에서는 bottom-up 접근법이 주목받고 있다. 이 방식은 처음에 장면 내의 모든 관절을 감지하고, 그 후에 이를 적절한 사람에 할당하는 방식이다. 그렇지만 이 bottom-up 방식은 관절을 제외하고 활용하는 것은 쉽지 않다. 반면에, top-down 접근법은 먼저 모든 객체를 감지하고, 그 후 각 사람의 자세를 추정하는 방식을 따른다. 이 접근법은 초기에는 난이도 있는 기술을 필요로 하지만, 현대의 최첨단 시스템 기술로 인해서 캐릭터 감지 및 자세 추정 기술에 기반하여 2D 자세에서 탁월한 성과를 보여주고 있다. 그러나 3D 상에서 여러 사람이 중첩된 경우의 자세를 추정할 때 예측하기 힘든 어려움이 있을 수 있다. 예를 들어 재구성된 사람들이 3D 공간에서 중첩되거나, 실제 깊이와 다르게 추정될 수 있다. 이런 문제점 때문에, 개별 사람의 3D자세를 정확히 예측하는 것뿐만 아니라, 장면의 모든 사람에 대한 일관된 재구성도 중요하다. 본 연구에서는 top-down 방식을 기반으로 하여, 깊은 학습 네트워크를 훈련시켜 장면의 모든 사람에 대한 일관된 3D 재구성을 추정하고자 한다. 기존의 R-CNN 프레임워크를 시작점으로, SMPL 모델을 사용하여 3D 재구성의 정밀도를 높이는 것을 목표로 하고 있다. 이 연구의 결과는 다양한 평가에서 우수한 성과를 보여주며, 전체적인 재구성의 품질을 향상 시키고 있다.

Ⅱ. 관련 연구

2.1 단일 인물의 3D 포즈와 형태

최근 연구에서는 3D 포즈의 뼈대 형태 추정[3-4]과 3D 형태의 비파라메트릭 방식 추정[6]에 주목하였다. 본 연구에서는 SMPL[5]와 같은 파라메트릭 모델의 사용을 중점적으로 고려하였다. Bogo 등의 SMPLify[7]는 주목할 만한 연구 중 하나로, 2D 관절 감지를 통해 SMPL을 반복적으로 맞추는 방식을 제안하였다. 이 접근법은 후속 연구에서 다양한 확장을 거치게 되었다[6]. 또한 딥 네트워크를 활용하여 이미지에서 직접 포즈와 형태 매개변수를 회귀하는 방식도 연구되었다. 이 중에서도 중간 표현 형태를 먼저 추정하는 연구가 주를 이루었다[8]. 특히 Kanazawa 등의 연구[9]는 교육 중에 불가능한 3D 형상을 패널티로 적용하여 효과적인 결과를 보였다.

2.2 여러 사람의 3D 포즈

R-CNN 기반 작업들 성공[10]을 기반으로, 여러 사람의 3D포즈 추정에 대한 연구도 활발히 이루어졌다. LCR-Net 접근법[11]은 여러 사람의 포즈 추정에서 주요한 방법론 중 하나로 자리매김하였다. 본 연구에서는 이러한 기존의 연구들과는 달리, 인스턴스 분할의 정보를 활용하여 깊이 순서를 추론하는 새로운 방식을 제안하였다. 다양한 사람들의 일관된 3D 재구성을 위한 제약조건 도입은 본 연구의 중요한 측면이다. 이에 대한 여러 연구가 이루어졌으며, 본 연구에서는 이러한 기존 연구를 바탕으로 새로운 제약 조건을 제안하였다. 특히, 침투 손실[12]과 깊이 순서 인식 손실을 도입하여 더욱 정확하고 일관된 3D 재구성을 달성하였다.

2.3 Object detection / tracking

다중 객체 추적을 위한 대표적인 알고리즘은 SORT(: Simple Online and Realtime Tracking) [13]이다. SORT는 이미지에 칼만 필터를 적용하여 경계 상자의 위치에 대한 예측을 추론하고, 알고리즘을 사용하여 검출기의 결과 경계 상자와 예측된 경계 상자를 연관시킨다. 그러나 이 방법은 물체 또는 인체가 가려지는 문제에 대해 취약하고, 복잡한 상황에서 문제를 해결하기 충분하지 않다. DeepSORT는 이 연관 문제를 보완하기 위해 알고리즘에 운동 및 외형 정보를 추가한다. 또한, OC-SORT, Strong SORT, Deep OC-SORT와 같은 다양한 SORT 기반 알고리즘이 연구되고 있다. MOTR[14]는 DETR의 객체 쿼리를 기반으로 한 트래킹 쿼리를 사용하여 비디오의 객체

를 식별한다. 객체 쿼리 세트는 객체가 이미지에 나타나는 시점부터 사라지는 시점까지의 각 프레임에 대해 예측을 수행한다. 디코더의 입력으로 사용되며 현재 프레임에 대한 추적 예측을 생성하고, 업데이트 된 세트는 다음 프레임의 디코더 입력으로 사용된다. 또한 TAN(:Temporal Aggregation Network) 및 QIM(:Query Interaction Module)의 결합된 구조를 사용하여 과거 프레임에서 처리된 객체의 트랙 쿼리가 집계되고 처리되며, 현재 프레임의 트랙 쿼리는 다중 헤드 주의를 통해 각 과거 프레임의 트랙 쿼리와 상호 작용한다. 이 방법은 IDS(:IDS on Switch)를 줄여 부드러운 추적 결과를 보여준다.

2.4 Pose estimation algorithm

사람의 자세 추정은 이미지나 비디오에서 인간의 신체 관절 또는 부위의 위치를 추정하는 기술이다. 사람의 포즈 추정에는 두 가지 주요 접근법이 있다: 상단-하단(top-down), 이는 별도의 객체 탐지기를 사용하고, 이미지의 모든 사람의 키포인트와 사람들의 연결(모서리)를 사용한다. 또한 사람 수(단일, 다중), 키포인트의 좌표 차원(2D, 3D), 추정에 사용된 카메라의 유형 (RGB, RGB-D)등을 지정할 수 있다. 2차원 추정의 경우, 실시간 환경에서 다중 사용자 추정을 위한 OpenPose와 최근의 Transformer 기반 연구가 있다. Transformer 기반 연구는 더 나은 결과를 보여주었다. 3차원 추정의 경우, 2차원 추정 결과를 3차원으로 재구성하거나 SMPL, SMPL-X 모델을 사용하여 인간의 메시를 추정하고 얼굴과 손 영역으로 확장한다. 또한, 실시간 환경에서 부드럽게 작동할 수 있도록 추정 정확도와 효율성 간의 절충점에 대한 연구도 진행되고 있다.

III. 기술적 접근 방법

이 섹션에서는 본 연구의 기술적 접근법을 설명한다. 본 연구는 SMPL 모델에 대한 정보를 제공하며 시작 방법과 사용하는 기본 아키텍처에 대해서 설명한다. 그 다음 제안한 손실에 대해 상세하게 설명하며, 이는 중첩 없는 재구성과 일관된 depth를 촉진한다. 마지막으로, 구현 세부 사항을 상세하게 제공한다.

3.1 SMPL 매개변수 모델

사람의 몸의 표현 및 구현을 위해 SMPL 매개변수 모델[5]을 사용한다. 다른 표현과 비교해 SMPL이 작업에 매우 적합한 이유는 그것이 네트워크의 훈련에 새로운 손실을 포함하여 중첩과 가림막에 대한 추론을 가능하게 해주기 때문이다. SMPL 모델은 입력으로 자세 매개변수와 형태 매개변수를 받고 메시 M을 출력한다.

3.2 기본 아키텍처

본 연구의 아키텍처 측면에서는 R-CNN 프레임워크를 따르며, 그 구조는 Mask R-CNN 반복과 매우 유사하다. 네트워크는 백본 (ResNet50), 지역 제안 네트워크, 탐지 및 SMPL 매개변수 회귀를 위한 헤드로 구성된다. SMPL 분기의 경우, 그 구조는 Kanazawa 등이 제안한 반복적인 회귀 분석기와 유사하다. 카메라 매개변수는 바운딩 박스 당 예측되며, 후에 전체 이미지에서 바운딩 박스의 위치를 기반으로 업데이트 한다. 각 바운딩 박스는 인접한 사람들과 그들의 자세를 알기 때문에, 그들과 일관된 자세를 예측한다. 기본 네트워크의 경우, 다양한 구성 요소는 종단 간 방식으로 공동으로 훈련된다. 3D 지상 진리가 제한적으로 사용 가능한 경우, SMPL 매개변수와 3D 키포인트에 대한 손실, L3D,를 적용한다. 2D 관절만 사용 가능한 경우, 2D 재투영 손실, L2D,를 사용하여 지상 진리 2D 키포인트와 3D 관절의 투영 간의 거리를 최소화한다. 또한, 판별자를 사용하며, 훈련 중 불가능한 3D 형태에 대한 처벌로 적대적 선행 사항 Ladv를 적용한다. 위의 각 손실은 지상 진리 바운딩 박스에 할당된 후 각 제안에 독립적으로 적용된다.

그림 1 인체 좌표 구성 시퀀스
Fig. 1 Human body coordinate construction sequence

그림1은 다중 인물 이미지에서 1인의 포즈를 감지하고 해당 포즈의 각도를 계산하여 인체의 좌표를 구

성하는 시퀀스이다. 상기 시퀀스는 사람의 객체를 감지하고 포즈를 추정하는 알고리즘을 적용하여 인체의 랜드마크 및 키포인트를 추출한다. 추출된 랜드마크를 기반으로 다양한 부위의 각도 및 좌표를 계산하고 이를 바탕으로 인체의 좌표를 구성하는 절차로 진행된다. 각 프레임마다 좌표와 각도 정보를 동적으로 계산 및 추정한다.

3.3 Interpenetration loss

회귀 네트워크가 자주 겹치는 위치에 사람들을 예측하는 문제로, 여러 사람들의 일관된 재구성을 위한 주요 장애물이 발생한다. 중첩되지 않는 사람들의 예측을 촉진하기 위해, 재구성된 사람들 사이의 중첩에 패널티를 적용하였다. 아래의 수식은 [15]에서 모티프를 얻었다. 중요한 차이점은 본 연구의 장면에는 여러 사람이 포함되어 있고 훈련 중에 동적으로 생성된다는 점이다. 재구성된 장면에 대한 수정된 부호 거리 필드 (SDF)를 식(1)과 같이 정의한다.

$$\phi(x,y,z) = -\min(SDF(x,y,x),0) \quad \cdots (1)$$

수식(1)의 정의에 따라 사람 내부에서 x,y,x는 표면으로부터 거리에 비례하는 양의 값을 가지게 되며 사람의 외부에서는 그 값이 0이다. 일반적으로 x.y,x 는 np~~ 차원의 voxel 그리드에서 정의된다.

전체 장면에 대한 단일 복셀 표현을 생성하는 것이 가능하나 세밀한 복셀 그리드가 필요한 경우가 많으며, 장면의 확대 및 확장에 따라 메모리와 연산 측면에서 처리가 어려울 수 있다. 각 사람에 대해 별도의 i 함수를 계산하여, 사람 주변에 타이트한 상자를 계산하고 복셀화한다는 것을 발견한다.

사람 i와 j 사이의 충돌에 대한 사람 j의 충돌 벌칙은 식(2)과 같이 정의된다.

$$P_{ij} = \sum_{v \in M_j} \phi_i(v) \quad \cdots (2)$$

여기서 (v)는 3D 버텍스 v의 i 값을 3D 그리드에서 삼선 보간을 사용하여 차별 가능한 방식으로 샘플링한다. N명의 사람이 포함된 장면에 대한 최종 중첩 손실은 식(3)과 같이 정의된다:

$$-L_p = \sum_{j=1}^{N} p\Big(\sum_{i=1, i \neq j}^{N} P_{ij}\Big) \quad \cdots (3)$$

여기서는 Geman-McClure의 오류 함수를 사용한다. 동일한 사람에 해당하는 상자 간의 교차를 처벌하지 않기 위해 실제 상자에 할당된 가장 확실한 상자 제안만을 사용한다.

전체 장면을 위한 단일 복셀화 표현을 생성하는 것은 분명 가능하다. 그러나, 종종 아주 세밀한 복셀 그리드가 필요하며, 장면의 확장에 따라 메모리와 연산 측면에서 처리가 불가능해진다. 여기에서 중요한 관찰은 각 사람에 대해 별도의 i 함수를 계산함으로써, 사람 주변에 타이트한 상자를 계산하고 복셀화할 수 있다는 것이다.

3.4 Depth ordering-aware loss

중첩 외에도, 여러 사람의 3D 재구성에서 일반적인 문제는 사람들이 잘못된 depth의 순서로 추정된다는 것이다. 사람들이 2D 이미지 평면에서 겹치는 경우 이 문제가 더욱 두드러진다. 인간의 눈에는 어떤 사람이 더 가까운지 명확하지만, 네트워크의 예측은 여전히 일관성이 없을 수 있다. 픽셀 수준의 깊이 주석에 접근할 수 있다면 이 깊이 정렬 문제를 쉽게 해결할 수 있을 것으로 예상된다.

본 연구의 주요 아이디어는 인스턴스 분할 주석, 예를 들면 대규모 COCO 데이터셋[16]과 같은 것을 활용할 수 있다는 것이다. 이미지 평면에 모든 재구성된 사람들의 메시를 렌더링하면 각 픽셀에 해당하는 사람을 지정하고 주석된 인스턴스 주석과의 일치에 기반하여 최적화한다. 이 아이디어는 직관적으로 들리지만, 실제로는 더 복잡하다. 명백한 구현은 Neural Mesh Renderer (NMR)[17]와 같은 차별 가능한 렌더러를 사용하고, 실제 인스턴스 분할과 이미지에 메시를 렌더링하여 생성된 것 사이의 불일치를 처벌한다. [17]의 실제 문제는 오직 보이는 메시 버텍스에만 오류를 역전파 한다는 것이다. 깊이 정렬 오류가 있으면, 보이지 않는 버텍스를 카메라에 더 가깝게 움직이도록 유도하지 않는다. 실제로, 대부분의 객체들이 더 멀리 움직이게 되어 학습이 붕괴되는 경향을 관찰한다. 또한 숫자적 불안정성에 직면하기도 한다. 장면의 의미론적 분할만 렌더링하는 대신, NMR[17]를 사용하여 각 사람에 대한 깊이 이미지 Di도 독립적으로 렌더링한다. 장면에 N명의 사람이 있다고 가정하면, 그들 각각에게 고유한 인덱스를 할당한다.

y(p)는 지상 진실 분할에서 픽셀 위치 p의 사람 인덱스이고, 는 3D 메시의 렌더링을 기반으로 한 예측된 사람 인덱스이다. 0을 사용하여 배경 픽셀을 나타낸다. 픽셀 p에 대해 두 가지 추정치가 사람을 나타내는 것으로 판단되고 동의하지 않으면, 즉, 이면, 이 픽셀의 두 사람, y(p) 및 의 깊이 값에 손실을 적용하여 올바른 depth의 순서를 촉진한다. 적용하는 손실은 [8]과 유사한 depth 손실이다. 보다 구체적으로, 완전한 손실 표현은 수식(4)와 같다.

$$Lp = \sum_{p \in S} \log(1 + \exp(D_{y(p)}(p) - D_{\tilde{y}(p)}(p))) \quad \cdots (4)$$

3.5 구현 세부사항

본 연구는 PyTorch와 공개적으로 사용 가능한 mmdetection 라이브러리를 사용하여 진행하였다. 모든 입력 이미지의 크기를 원래의 COCO 훈련과 동일한 종횡비로 512×832로 조정하였고 기본 모델의 경우 3.2.에서 지정된 손실만으로 훈련을 진행하며, 본 연구의 전체 모델에는 3.3 및 3.4.에서 제안된 손실을 훈련에 포함하였다. 훈련에는 2개의 3080Ti GPU를 사용하며, GPU당 이미지 배치 크기는 4이다. 그림2는 단일 인물의 관절 keypoint와 관절별 각도를 계산하여 출력하고 있는 테스트 수행 화면과 이를 구현한 알고리즘이다.

그림 2 관절 keypoint와 관절별 각도 계산 및 구현 알고리즘

Fig. 2 Calculate joint keypoints and angles for each joint & Implementation Algorithm

SDF 계산을 위해, 본 연구에서는 [18, 19]를 CUDA에서 다시 구현했다. 32×32×32 voxel 그리드에서 단일 메시를 voxel화하는 데 약 45ms가 1080Ti GPU에서 소요된다. 효율성을 위해, 3D 경계 상자 검사를 수행하여 중첩된 3D 경계 상자를 탐지하고 관련 메시만 voxel화한다. 또한, 큰 이미지를 더 효율적으로 렌더링하기 위해 NMR[27]의 일부를 다시 구현한다. 이로 인해 평균적으로 전달 패스 복잡도가 O(Fwh)에서 O(F+wh)로 줄어들었으므로, 속도가 한 차수 이상 빨라진다. 여기서 F는 면의 수, w와 h는 각각 이미지의 너비와 높이를 나타낸다.

아래 그림3은 다중 인체의 디텍션 및 자세 추정을 구현한 이미지이다.

그림 3 다중 인체의 디텍션 및 자세 추정 구현

Fig. 3 Implementation of detection and posture estimation of multiple human bodies

Ⅳ. 실험 결과

여기서는 본 연구의 접근 방식에 대한 경험적 평가를 제시하였다. 먼저, 훈련과 평가에 사용된 데이터셋을 설명한다. 그런 다음, 정량적 평가에 중점을 둔다. 그리고 좀 더 질적인 결과를 제시한다.

4.1 데이터셋

Human3.6M: 각 프레임에 하나의 사람만이 보이는 실내 데이터셋이다. 훈련과 평가를 위한 3D 지상 진리를 제공한다. 훈련을 위해 S1, S5, S6, S7 및 S8 주제를 사용하고, S9 및 S11 주제를 평가에 사용한다. Panoptic: Panoptic 스튜디오에서 촬영된 여러 사람의 데이터셋이다. MPI-INF-3DHP: 3D 인체 자세 추정 단일 인물 데이터셋이다. S1부터 S8까지의 주제를 훈련에 사용한다. PoseTrack: 야외에서의 2D 자세 주석이 있는 데이터셋이다. 각 시퀀스에 대해 여러 프레임이 포함된다. 이 데이터셋을 훈련과 평가에 사용한다. LSP, LSP Extended, MPII: 야외에서의 2D 관절 주석이 있는 데이터셋이다. COCO [16]: 2D 자세와 인스턴스 분할 주석이 있는 야외 데이터셋이다. 다른 야외 데이터셋과 마찬가지로 2D 관절을 훈련에 사용하며, 인스턴스 분할 마스크는 깊이 순서 인식 손실 계산에 사용된다.

4.2 최신 기술과의 비교

최신 기술과의 비교를 위해, 이 연구의 접근 방식의 성능을 표준 단일 인물 기준에 대해 평가한다. 이 연구의 목표는 항상 여러 사람의 3D 자세와 형태이지만, 이미지에 한 사람만 있을 때, 이러한 접근 방식도 경쟁력 있는 결과를 얻을 것으로 기대한다. 본 연구는 대중적인 Human3.6M 데이터셋에서 네트워크의 성능을 평가한다. 여기서 가장 관련 있는 방법은 Kanazawa 등이 제안한 HMR [9]이다. 이는 본 연구와 유사한 구조적 선택 (반복적 회귀 분석기, 회귀 대상), 훈련 관행 (적대적 선례) 및 교육 데이터를 공유하기 때문이다. 결과는 표 1과 같다. 본 연구의 방법은 HMR보다 우수한 성능을 보이며, 더 많은 데이터로 훈련된 Arnab 등의 방법보다도 뛰어나다.

표 1. Human3.6M 데이터셋의 결과
Table 1. Results from Human3.6M dataset

Method	HMR	Arnab et al.	This Resarch
Reconstruction Error	56.8	54.3	52.7

단일 인물 설정에서 본 연구의 방법이 경쟁력 있다는 것을 확인했으므로, 여러 인물을 기준으로 평가를 계속한다. 이 경우에, 본 연구는 여러 사람의 자세와 형태를 추정하는 방법도 고려한다. 가장 관련 있는 기준선은 Zanfir 등의 연구 [12]이다. 본 연구는 이러한 접근법과 Panoptic 데이터셋에서 비교하며, 그들의 평가 프로토콜을 사용한다 (Panoptic 스튜디오의 데이터는 훈련에 사용되지 않는다고 가정). 결과는 표 2에 정리하였다.

표 2. Panoptic 데이터셋의 결과
Table 2. Results from Panoptic dataset

Method	Haggling	Mafia	Ultim.	Pizza	Mean
Zanfir et al.	140.0	165.9	150.7	156.0	153.4
Zanfir et al.	141.4	152.3	145.0	162.5	150.3
This Resarch (basic)	141.2	140.3	160.7	156.8	149.8
This Resarch (full resource)	129.6	133.5	153.0	156.7	143.2

본 연구의 초기 네트워크는 제안된 손실 없이 훈련되며, Zanfir 등의 이전 연구에서 보고된 결과와 비슷한 성능을 보인다. 그러나 더 중요한 것은, 제안된 두 가지 손실 (전체)을 추가하면 모든 하위 시퀀스와 전체에서 성능이 향상되며, 이전의 기준선을 능가하기도 한다. 이러한 결과는 다인(多人) 설정에서 본 연구의 접근법의 강력한 성능, 그리고 이 연구에서 제안하는 손실로부터 얻는 이점을 보여준다.

다인(多人) 3D 자세 추정을 위한 또 다른 인기 있는 벤치마크는 MuPoTS-3D 데이터셋이다. 이 벤치마크에 대한 결과를 확인할 수 있는 여러인물의 3D 자세와 형태의 접근법이 없기 때문에, 본 연구는 단일 인물 3D 자세와 형태에 대한 최첨단 접근법을 기반으로 두 개의 강력한 상하방향 기준선을 구현한다. 구체적으로, 본 연구는 회귀 접근법인 HMR [9]와 최적화 접근법인 SMPLify-X를 선택하고, 이들을 OpenPose[5]에서 제공하는 감지에 적용하거나 또는 Mask-RCNN 에 적용한다. 결과는 표 3과 같다.

표 3. MuPoTS-3D의 결과
Table 3. Results fromc MuPoTS-3D

Method	All	Matched
OpenPose + SMPLify-X	62.84	68.04
OpenPose + HMR	66.09	70.90
Mask-RCNN + HMR	65.57	68.57
This Resarch(basic)	66.95	68.96
This Resarch(full resource)	69.12	72.22

본 연구의 기준 모델은 다른 접근법과 유사한 성능을 보이며, 제안된 손실로 훈련된 본 연구의 전체 모델은 기준선을 크게 능가한다. 이전의 결과와 마찬가지로, 이 실험은 본 연구의 일관성 손실 사용을 더욱 정당화한다. 뿐만 아니라, 본 연구는 단일 인물을 염두에 두고 훈련된 기준선이 3D 자세의 다인(多人) 설정에는 최적이 아니라는 것을 보여준다. 이것은 2D 경우와 다르다. 여기서 단일 인물 네트워크는 다인 상하 방향 파이프라인에서도 매우 잘 수행될 수 있다.

4.3 모델별 비교

이 연구에서 여러 사람의 3D 자세 추정에 대한 관심은 일반적인 3D 자세 지표를 통해 자세를 추정하

는 것을 넘어서서 이미지 장면의 일관된 재현을 구현하는 것이다. 예를 들어, 객체들의 깊이 순서가 잘못될 수 있거나, 재구성된 메시가 겹치게 배치될 수 있다. 이러한 일관성 지표에서 훈련 중에만 적용되더라도 본 연구의 제안된 손실이 어떻게 네트워크 예측을 개선하는지 보여주기 위해, 본 연구는 더 상세한 평가를 위한 두 가지 성분별 연구를 수행한다.

먼저, 본 연구는 중첩 손실을 통해 예측에서 겹치는 사람들을 자연스럽게 제거할 것으로 예측된다. 본 연구는 이를 평가하며, 중첩 손실의 유무에 따른 충돌 수를 보여주는데 결과는 표 4와 같다.

표 4. Interpenetration loss 결과
Table 4. Interpenetration loss results

Method	MuPoTS-3D	PoseTrack
This Resarch(basic)	114	654
This Resarch(basic)+L_p	34	202

예상대로 LP 손실로 네트워크를 훈련시킬 때 충돌 수가 크게 감소하는 것을 볼 수 있다.

더불어 깊이 순서를 인식하는 손실은 장면에 있는 사람들의 자세 변환 추정 방법을 개선해야 한다. 단안 방법에 대한 척도별 변환 추정을 평가하는 것은 의미가 없으므로, 본 연구는 반환된 깊이 순서만 평가할 것을 제안한다. 본 연구는 장면의 모든 사람 쌍을 고려하고, 본 연구의 방법이 이 쌍의 순서 깊이 관계를 올바르게 예측했는지 평가한다. 예상대로, 깊이 순서를 인식하는 손실은 본 연구의 기준선을 개선한다.

마지막으로, 본 연구는 테스트 시간에 이러한 일관성 손실을 적용하지 않는다. 대신, 훈련 중에, 본 연구의 손실들은 재구성에 제약조건으로 작용하며, 명시적인 3D 주석이 사용 가능하지 않은 이미지에 대해 네트워크에 더 나은 지도를 제공한다. 개선된 지도는 테스트 시간에도 더 일관된 결과로 이어진다.

Ⅴ. 결론

본 연구에서는 단일 이미지에서 다중 인물 3D 자세 및 형태 추정을 위한 종단 간 접근 방식을 제시한다. R-CNN 프레임워크를 사용하여 이미지 내에서 감지된 각 사람에 대한 SMPL 모델 매개변수를 회귀하는 상향식 접근 방식을 설계한다. 본 연구의 주요 공헌은 문제를 보다 전체적인 관점에서 평가하고 각 사람에 대한 독립적인 자세 추정에만 중점을 둔 대신 일관된 장면 재구성을 추정하는 것에 목표를 둔다. 이를 위해, a) 중첩된 인간을 생성하는 것을 피하고 b) 사람들을 일관된 깊이 순서로 위치시키도록 격려하는 네트워크를 훈련하는 두 가지 새로운 손실을 프레임워크에 통합한다. 본 연구의 접근 방식을 다양한 벤치마크에서 평가하며, 전통적인 3D 자세 메트릭에서 매우 경쟁력 있는 성능을 보이며, 재구성된 장면의 일관성 측면에서도 질적 및 양적으로 훨씬 더 우수한 성능을 보인다. 향후 연구에서는 사람들 간의 상호 작용을 보다 명시적으로 모델링하여 더욱 정확하고 상세한 장면 재구성을 더욱 세밀한 수준에서도 달성하려고 한다. 비슷한 맥락에서, 장면의 전체적인 재구성을 향해 추가 정보를 통합할 수 있다. 이에는 지면, 배경 또는 사람이 상호 작용하는 객체에 대한 제약 조건이 포함될 수 있다.

감사의 글

『이 논문은 2021년 순천대학교 학술연구비(과제번호: 2021-0320) 공모과제로 연구되었음.』

References

[1] H. Sim, “Development of Augmented Reality Character System based on Markerless Tracking”, *J. of the Korea Institute of Electronic Communication Sciences,* vol. 17, no. 6, 2022, pp. 1275-1282.

[2] J. Jung, G. Lee and B. Kim, “A Study on Stable Service of Marker based Augmented Reality Using 3D Location Measurement of Beacons”, *J. of the Korea Institute of Electronic Communication Sciences,* vol. 12, no. 5, 2017, pp.883-890.

[3] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H. Seidel, W. Xu, D. Casas, and C. Theobalt, “VNect: Real-time 3D human pose

estimation with a single RGB camera," *ACM Transactions on Graphics (TOG)*, vol. 36 no. 4, May 2017, pp. 21-44.

[4] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," *2017 IEEE International Conference on Image Processing(ICIP),* Beijing, China May. 2017.

[5] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM transactions on graphics (TOG)*, vol. 34, no. 248, Oct. 2015, pp. 1-16.

[6] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid, "BodyNet: Volumetric inference of 3D human body shapes," *European Conference on Computer Vision(In ECCV)*, Munich, Germany, Aug. 2018.

[7] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," *European Conference on Computer Vision(In ECCV),* Amsterdam, The Netherlands, July. 2016.

[8] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele, "Neural body fitting: Unifying deep learning and model based human pose and shape estimation," *2018 International Conference on 3D Vision,* Verona, Italy, Aug. 2018.

[9] A. Kanazawa, M. J. Black, D. W. Jacobs, and Jitendra Malik, "End-to-end recovery of human shape and pose," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(In CVPR)*, Salt Lake City, UT, USA, Dec. 2018.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *J. IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 39, 2017, pp. 1137-1149.

[11] G. Rogez, P. Weinzaepfel, and C. Schmid, " LCR-Net: Localization-Classification-Regression for Human pose," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition(In CVPR)*, Honolulu, HI, USA, July 2017.

[12] A. Zanfir, E. Marinoiu, and C. Sminchisescu, "Monocular 3D pose and shape estimation of multiple people in natural scenes－the importance of multiple scene constraints," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(In CVPR)*, Salt Lake City, UT, USA, June 2018.

[13] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," *2016 IEEE International Conference on Image Processing(ICIP),* Phoenix, AZ, USA, Sept. 2016.

[14] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, "MOTR: End-to-end multiple-object tracking with transformer," *2022 European Conference on Computer Vision(ECCV)*, Tel Aviv, Israel, July 2022.

[15] M. Hassan, V. Choutas, D. Tzionas, and M. J. Black, "Resolving 3D human pose ambiguities with 3D scene constraints," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, Oct. 2019.

[16] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, " Microsoft COCO: Common objects in context," *European Conference on Computer Vision*, Zurich, Switzerland, Sept. 2014, pp.740-755.

[17] H. Kato, Y. Ushiku, and T. Harada, "Neural 3D mesh renderer," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(In CVPR)*, Salt Lake City, UT, USA,, June 2018, pp. 3907-3916.

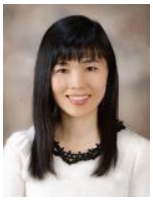
[18] D. Stutz, "Learning shape completion from bounding boxes with CAD shape priors," PhD thesis, Masters thesis, *RWTH Aachen University,* Sept. 2017.

[19] D. Stutz and A. Geiger, "Learning 3D shape completion from laser scan data with weak supervision," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(In CVPR)*, Salt Lake City, UT, USA,, June 2018.

저자 소개

이석환(Seok-Hwan Lee)

2002년 홍익대학교 판화과 졸업 (미술학사)
2018년 Bournemouth University, 3D Computer Animation (MA)
2020년~현재 순천대학교 만화애니메이션학과 조교수
※ 관심분야 : 3D 그래픽, 디지털트윈, 인공지능, 교육콘텐츠



이정금(Jung-Keum Lee)

2005년 순천대학교 교육대학원 국어교육과 졸업(교육학석사)
2014년 순천대학교 대학원 교육학과 졸업(교육학박사)
2008년~현재 순천대학교 교직과 강사
※ 관심분야 : 디지털트윈, 인공지능, 교육콘텐츠

심현(Hyun Sim)

2002년 순천대학교 컴퓨터과학과 졸업(이학석사)
2009년 순천대학교 대학원 컴퓨터과학과 졸업(이학박사)
2020년~현재 순천대학교 스마트농업전공
2023년~현재 순천대학교 정보전산원장
2021년~현재 순천대학교 공동훈련센터 센터장
2021년~현재 디지털트윈스마트시티연구소 소장
※ 관심분야 : 디지털트윈, 인공지능, 교육콘텐츠