특집 메타버스

영상과 비디오로부터의 3차원 휴먼 자세 및 형상 복원 기술

□ 조슈아 산토소 , *전성호, *장주용, 박인규 / 인하대학교, *광운대학교

요약

미래의 메타버스 환경에서 3차원 가상 휴먼 표현은 매우 중요한 기술이며 영상 또는 비디오로부터 3차원 가상 휴먼 모델링이 핵심 기술이다. 본 기고문은 이 분야에 대한 충분한 사전 지식의 제공을 목표로 한다. 휴먼 복원 문제를 다루는 연구가 늘어남에 따라, 본 기고문에서 우리는 단일 영상 혹은 비디오로부터의 3차원 휴먼 복원 연구들에 대해 조사하고 그 결과를 다음과 같이 체계적으로 제시한다. 첫째, 3차원 휴먼 복원에 대한 배경 개념을 정의한다. 둘째, 제안된 분류법, 기여도, 정량적 결과에 따라 기존의 방법들을 상세하게 분석한다. 셋째, 관련 데이터셋 및 정성적 결과를 요약하여 연구자들이 이를 쉽게 활용할 수 있도록 한다. 마지막으로, 우리는 각 연구들을 분석하여 해당 방법들의 장점과 약점을 제시한다.

Ⅰ. 서 론

3차원 휴먼 복원은 행동 감지(action detection), 증강/가상 현실(augmented/virtual reality), 모션 캡처(motion capture) 및 온라인 게임과 같은 다양한 응용 분야에서 광범위하게 사용됨에 따라 관련 연구자들로부터 큰 주목을 받고 있다. 최근 연구들[1, 2, 3, 6, 7, 9, 10, 12, 13, 14, 15, 18, 21, 22, 26]에서는 예전에 비해 상당히 개선된 휴먼 복원 성능을 보여주었다. 이를 간단히 분류하기 위해 기존 연구들은 최적화(optimization) 기반 접근법[6, 9]과 학습(learning) 기반 접근법[1, 2, 3, 7, 10, 12, 13, 14, 15, 18, 21, 22, 26]으로 크게 분류될 수 있다. 학습 기반 접근법들은 추가적으로 모델 기반(model-based) 접근법[1, 2, 3, 6, 7, 8, 12, 14, 18, 21, 26, 30]과 모델을 사용하지 않는(model-free) 접근법[10, 15, 22]으로 나눠질 수 있다.

기존에 많은 연구들이 제안되어 왔으므로 본 기고문을 통해 모든 방법을 이해하기는 어렵다. 따라서 본 기고문에서 우리는 분석 대상을 단일 영상 혹은 비디오를 입력으로 사용하는 휴먼 메쉬(mesh) 복원 방법들로 한정하며, 해당 연구에서 사용된 데이터셋(dataset), 손실

함수(loss function)와 함께 다양한 방법들에 대한 구체적인 분류를 제시한다. 또한 우리는 정량적 및 정성적 결과를 포함한, 각 방법들에 대한 분석 결과를 제시한다.

Ⅱ. 문제 정의

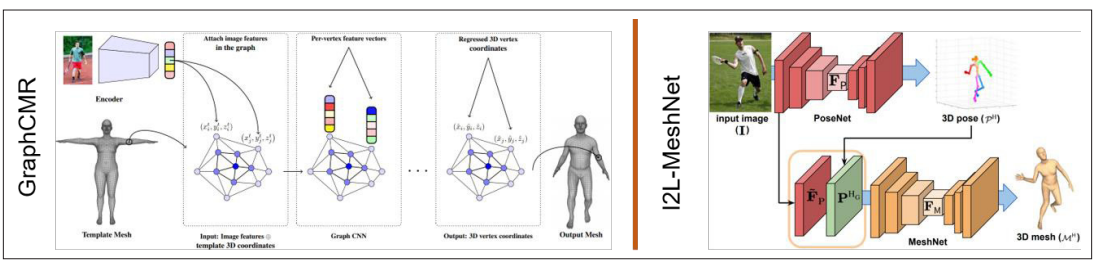
우리는 3차원 휴먼 복원 문제를 다음과 같이 정의한다. X_I를 단일 영상, X_{V_t}를 프레임 수가 n인 비디오 시퀀스 $X_V=(X_{V_1}, \cdots, X_{V_n})$ 에서 t-번째 프레임 영상이라고 가정하자. 이러한 X_I 혹은 X_{V_t}는 휴먼 복원 방법의 입력 $X \in \mathbb{R}^{w \times h \times c}$으로 사용된다. 여기서 w, h, c는 각각 입력 텐서의 너비, 높이, 그리고 채널 수를 의미한다. 휴먼 복원 방법의 목표는 휴먼 메쉬 $M=(V, F)$를 추정하는 것이다. 여기서 $V \in \mathbb{R}^{N \times 3}$와 F는 각각 개수가 N인 정점들(vertices)과 삼각형 면들(triangular faces)을 나타낸다. 일반적으로 휴먼 메쉬는 휴먼 모델을 사용하여 복원될 수 있다[9, 11, 13, 25]. 본 기고문에서 우리는 최근 휴먼 복원에서 널리 사용되는 skinned multi-person linear model(SMPL)[25]에 초점을 맞춰 그와 관련된 방법들에 대해서 자세히 살펴본다. SMPL은 휴먼 메쉬 M을 출력하는 미분가능한 함수 $M(\beta, \theta)$으로 표현된다. 형상 매개변수 $\beta \in \mathbb{R}^{10}$는 PCA공간에서 얻어지는 10개의 계수로 기술된다. 자세 매개변수 $\theta \in \mathbb{R}^{J \times D}$는 신체를 구성하는 관절들의 3차원 회전으로 기술된다. 여기서 J와 D는 각각 SMPL 인체 템플릿을 구성하는 주요 관절의 개수와 3차원 회전 표현의 차원을 나타낸다. SMPL 모델의 정점의 수 N은 6890이다.

Ⅲ. 단일 영상으로부터의 3차원 휴먼 복원

휴먼 메쉬는 모델 기반 접근법 또는 모델을 사용하지 않는 접근법 두 가지 방식을 통해 복원될 수 있다. 모델을 사용하지 않는 접근법은 각 정점의 위치를 직접 추정하는 방법으로 우리는 이를 3.1에서 구체적으로 설명한다. 모델 기반 접근법은 각 정점의 위치를 추정하는 대신 입력 영상 혹은 비디오로부터 휴먼 모델의 매개변수를 추정하는 방법이며, 우리는 이를 3.2에서 구체적으로 설명한다.

1. 모델을 사용하지 않는 접근법

〈그림 1〉은 모델을 사용하지 않는 접근법의 전반적인 구조를 보여준다. 모델을 사용하지 않는 접근법은 Kolotouros 등이 제안한 GraphCMR[22]을 통해 처음으로 제안되었다. 입력 영상으로부터 추출된 특징은 그



<그림 1> GraphCMR[22](왼쪽)과 I2L-MeshNet[10](오른쪽)의 모델 구조

래프 네트워크에 임베딩(embedding)되어 휴먼 메쉬의 3차원 좌표를 추정한다. 여기서 그래프의 노드 수는 휴먼 메쉬 모델의 정점의 수와 같다. GraphCMR과 달리 Moon과 Lee는 image-to-lixel 구조에 기반한 I2L-MeshNet[10]을 제안하였다. 이 방법은 자세 네트워크(PoseNet)와 메쉬 네트워크(MeshNet)의 두 모듈로 구성된다. 먼저 단일 RGB 영상이 PoseNet에 입력되면, PoseNet은 3차원 관절 위치를 나타내는 3차원 가우시안 열지도(Gaussian heatmap)를 추정한다. MeshNet은 PoseNet의 앞부분에서 추출된 특징과 마지막으로 추정된 3차원 가우시안 열지도를 입력으로 받아 최종 휴먼 메쉬를 출력한다.

Lin 등은 METRO[15]라고 불리는 self-attention transformer 기반의 인코더 네트워크를 사용하여 휴먼 메쉬를 복원하는 방법을 제안했다. METRO는 먼저 convolutional neural network(CNN) 네트워크를 사용하여 주어진 단일 RGB 영상으로부터 특징을 추출한다. 추출된 영상 특징은 SMPL 인체 템플릿의 3차원 관절 및 정점 좌표와 합쳐지며, transformer에서 처리될 관절 쿼리와 정점 쿼리를 생성한다. 다음으로 transformer 인코더 네트워크는 관절 쿼리와 정점 쿼리를 입력받아 3차원 관절 좌표와 정점 좌표를 병렬적으로 출력한다.

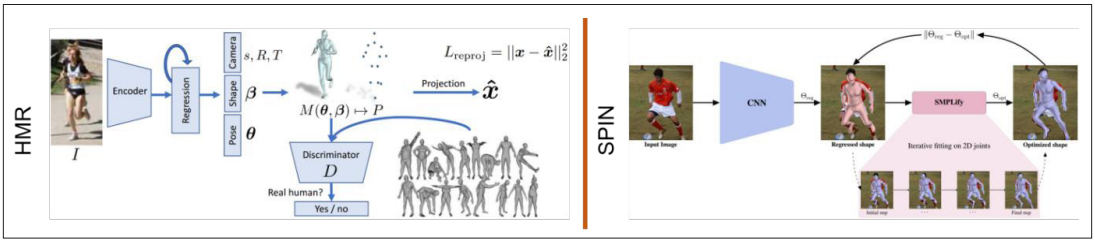
2. 모델 기반 접근법

모델 기반 접근법은 각 정점의 위치를 직접 추정하지 않고 휴먼 모델의 자세 및 형상 매개변수를 추정한다. 이러한 접근법은 모델을 사용하지 않는 접근법보다 더 활발히 사용되어 왔다. 최근 다수의 연구들[1, 2, 3, 6, 7, 8, 12, 14, 18, 21, 26, 30]에서 모델 기반 접근법이 사용되었으며, 주목할 만한 결과를 보여주었다. 모델 기반 접근법은 직접 추정 방식과 간접 추정 방식으로 다시 분류될 수 있다.

간접 추정 방법 : 이러한 접근법에서는 SMPL 매개변수를 추정하기 전에 입력 영상이 키포인트(keypoint), 휴먼 마스크(human mask) 또는 가우시안 열지도와 같은 다른 표현으로 변환된다. Bogo 등은 SMPLify[6]를 제안하였는데, 주요 아이디어는 다음과 같다. 입력 영상을 DeepCut[16]에 입력하여 2차원 자세를 추정한다. 추정된 2차원 자세는 피팅(fitting) 알고리즘의 입력으로 사용된다. 피팅 알고리즘의 목표는 SMPL 휴먼 모델에서 투영된 관절, 자세 및 형상의 오차를 최소화하여 SMPL의 자세 및 형상 매개변수를 DeepCut에서 추정된 2차원 자세에 맞추는 것이다. 피팅 알고리즘은 CNN 기반 알고리즘이 아니므로 반복적인 방법을 통해 오차 함수를 최소화한다.

그 후 Pavlakos 등은 단일 영상에서 SMPL 매개변수를 추정하는 CNN 기반 방법[8]을 제안했다. 제안된 방법은 초기화 모듈, 형상 모듈, 그리고 자세 모듈로 구성된다. 초기화 모듈은 멀티 태스크 학습(multi-task learning) 패러다임에 기반하여 입력 영상으로부터 2차원 열 지도와 실루엣(silhouette)을 동시에 추정한다. 다음으로 2차원 열지도는 자세 매개변수를 추정하기 위해 자세 모듈에 입력되고, 실루엣은 형상 모듈에서 형상 매개변수를 추정하기 위해 사용된다.

최근에 Li 등은 휴먼 메쉬를 복원하기 위해 역운동학(inverse kinematics) 접근법을 사용하는 하이브리드 솔루션[14]을 제안했다. 좀 더 구체적으로, CNN 네트워크는 두 개의 헤드로 구성되는데, 첫 번째 헤드는 3차원 관절 위치를 추정하고, 두 번째 헤드는 형상과 트위스트(twist) 각도 매개변수를 추정한다. 이후 추정된 3차원 관절 정보로부터 SMPL 모델을 활용하여 휴먼 메쉬를 복원한다. 상대 회전 문제를 다루기 위해 회전은 트위스트 회전과 스윙(swing) 회전, 두 가지로 나뉘어질 수 있는데, 트위스트 회전은 관절을 돌리는(turn) 것을 나타내고 스윙은 전체 관절을 움직이는(move) 것을 나타낸다.



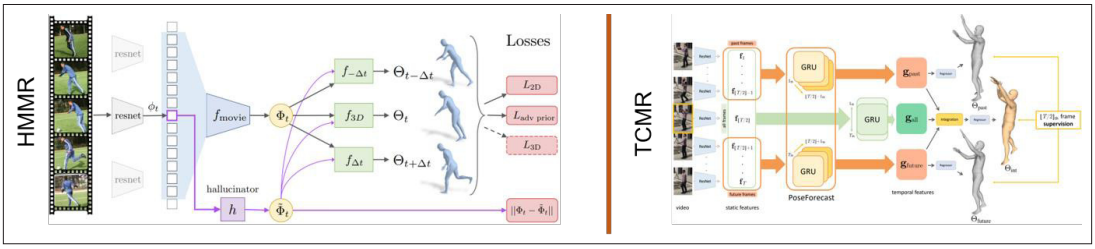
<그림 2> HMR[1](왼쪽)과 SPIN[21](오른쪽)의 모델 구조

직접 추정 방법 : 이 접근법에서는 입력 영상으로부터 SMPL 매개변수가 직접적으로 추정된다. Kanazawa 등은 human mesh recovery(HMR)[1]을 제안하였다. HMR은 단일 CNN 네트워크를 사용하는 대신 〈그림 2〉처럼 네트워크를 디자인하기 위해 생성적 적대 신경망(GAN) 개념을 사용했다. 생성 네트워크는 단일 영상으로부터 SMPL 매개변수를 추정한다. HMR은 신뢰할 수 없는 휴먼 메쉬가 생성되는 것을 막기 위해 판별기(discriminator)를 사용하여 추정된 휴먼 메쉬가 진짜인지 가짜인지를 판단한다.

Kolotouros 등은 CNN 기반, 그리고 최적화 기반의 방법을 혼합한 SPIN[21]을 제안했다. 일반적으로 CNN 기반 방법은 메쉬를 추정하는데 있어서 빠르고, 만족스러운 성능을 보이지만 잘 설계된 최적화 기반 방법에 비해서는 좋은 성능을 내지 못한다. 그에 반해 최적화 기반 방법은 좋은 피팅 성능을 보이지만 초기 예측값에 따라 성능이 많이 좌우되고 매우 느리다는 단점을 가진다. SPIN의 주된 아이디어는 초기 SMPL 매개변수를 추정하기 위해 영상을 CNN 네트워크에 입력하는 것이다. 초기 매개변수는 일반적으로 SPIN 모델의 최종 결과의 절반 정도의 성능을 보인다. 최적화 기법은 CNN에서 추정된 매개변수를 최적화의 시작점으로 사용해서 반복적인 피팅을 수행한다. 이를 통해 SPIN은 최적화 소요 시간을 크게 줄일 수 있다.

앞서 소개된 연구[1, 21]에서는 CNN 네트워크에서 추출된 특징으로부터 SMPL 매개변수를 직접 추정하였다. 최근, Georgakis 등은 복원 결과를 개선하고 가리워짐(occlusion)에 대처할 수 있는 최적화 접근방식으로 계층적 운동학 모듈(hierarchical kinematic module)을 도입했다[7]. SMPL 휴먼 모델은 23개의 관절로 구성된다. 모든 관절은 6개의 클러스터(골반, 머리, 오른팔, 왼팔, 오른다리, 왼다리)로 나누어질 수 있다. 예를 들어, 오른쪽 팔은 오른쪽 어깨, 오른쪽 팔꿈치, 오른쪽 손목 관절들로 구성될 수 있다. 순방향 패스(forward pass)에서는 오른쪽 어깨가 입력되는데, 이에 따라 오른쪽 팔꿈치의 추정값은 오른쪽 어깨에 의존하게 된다. 마찬가지로 오른쪽 손목의 추정값은 오른쪽 어깨와 오른쪽 팔꿈치 등에 달려있다. 마지막으로 추정 결과로 얻어진 모든 잔차(residuals)를 기반으로 현재 자세가 갱신된다. 역방향 패스(backward pass) 단계에서는 순방향 패스 단계와 유사한 절차가 수행된다. 그러나 몸의 말단에서부터 시작해 중심부로 진행된다는 점이 주된 차이점이다.

선명한 영상을 입력으로 사용하는 기존의 방법들[1, 7, 21]과 달리 Xu 등은 저해상도 영상으로부터의 휴먼 메쉬 복원을 보여주었다[30]. 제안된 방법은 해상도 인지 네트워크(resolution-aware network), 자기 지도 학습(self-supervised learning), 그리고 대조 학습(contrastive learning)으로 구성된다. 해상도 인지 네트워크는 각기 다른 해상도의 영상을 입력 받아 각 해상도 레벨의 특징을 융합하는 방법을 학습한다. 자기 지도



<그림 3> HMMR[2](왼쪽)과 TCMR[12](오른쪽)의 모델 구조

학습 기반의 손실 함수는 각기 다른 해상도의 영상에서 추정된 인체 메쉬의 일관성을 유지하는 데 사용된다. 마지막으로, 가변 해상도의 동일 영상과 최대한으로 유사한 특징 표현을 생성하기 위해 대조적 학습이 활용된다.

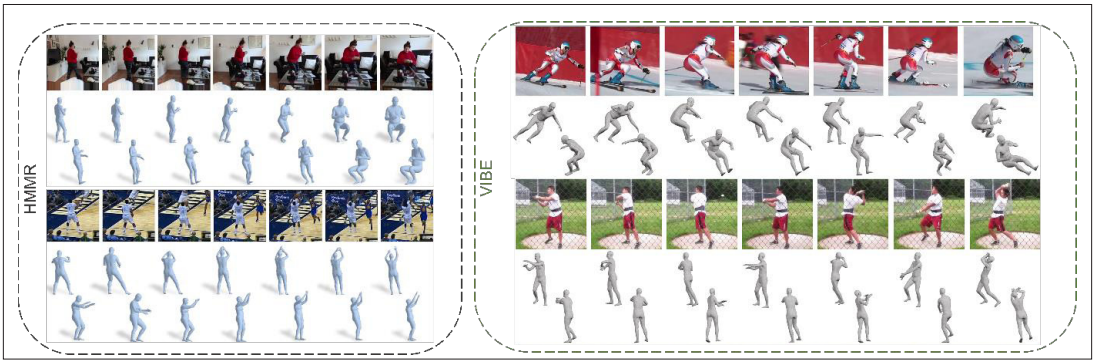
한편 Zhang 등[26]과 Biggs 등[3]은 가리워짐이 있는 영상에서 타당한 휴먼 메쉬를 복원하는 방법을 제안하였다. Zhang 등은 UV 지도 인페인팅(inpainting) 네트워크와 saliency 지도를 추정하는 네트워크를 제안했다. UV 지도 인페인팅 네트워크는 가리워짐이 있는 UV 지도를 입력으로 받아 가리워짐이 없는 UV 지도를 예측한다. 가리워짐이 있는 UV 지도는 물체로 가리워진 사람의 텍스처(texture)를 나타낸다. Saliency 지도 네트워크는 입력 영상으로부터 휴먼 마스크(human mask)를 추정한다. 휴먼 마스크는 입력 영상에서 배경 및 가리워짐같이 불필요한 정보를 줄이는 데 사용된다. 이후, 마스킹된 영상은 고차원의 UV 지도를 제공하는 UV 지도 네트워크의 추가적인 특징과 함께 회귀 네트워크(regression network)에 입력된다.

Biggs 등은 이와 다른 전략을 제안하였다. 모델을 훈련하는 동안, 자세 집합이 네트워크에서 출력된다. 이때, best-of-M 손실 함수는 출력된 자세들 중에 참값(ground truth)과 가장 유사한 자세를 선택하도록 만든다. 평가 과정에서 pretrained 네트워크는 입력 영상으로부터 다양한 자세 가설들(multiple pose hypotheses)을 추정한다. 다음으로 n개로 양자화된 best-of-M을 사용하여 자세 집합을 샘플링하고 flow 모델을 이용해 정규화하여 일부 비정상적인 자세를 제거한다.

Ⅳ. 비디오로부터의 3차원 휴먼 복원

Kanazawa 등은 비디오에서 3차원 휴먼 메쉬를 복원하는 방법을 제안했다[2]. 이를 위해 제안된 방법은 먼저 비디오 시퀀스로부터 시간적 특징을 추출한다. 이후에는 현재 프레임, 가까운 과거 프레임, 가까운 미래의 프레임에 대한 휴먼 메쉬를 추정한다. 입력으로 비디오를 사용할 경우의 이점은 시간적으로 인접한 프레임에서 휴먼 메쉬의 외관이 서로 유사하다는 사실을 이용할 수 있다는 점이다. 이런 이유로, 해당 논문은 〈그림 3〉에서 볼 수 있듯이 시간적 맥락 표현을 hallucinate하도록, 중간 프레임에 의해 학습되는 hallucinator 네트워크를 제안했다.

Kocabas 등은 GAN 기반 모델인 VIBE[18]를 제안하였다. 제안된 모델의 특징 추출기는 [2]와 달리 각 프레임의 특징을 사전 처리(preprocessing)로 추출한다. Gated recurrent unit(GRU) 레이어는 잠재 특징(latent feature)을 산출하기 위해 사용된다. 또한, 신뢰할 수 없는 자세 집합을 피하기 위해 모션 판별기가 사용되며, AMASS[31] 데이터셋을 사용하여 학습한다. Choi 등은 VIBE의 결과를 개선하기 위해 PoseForecast 모듈을 도입했다[12].



<그림 4> HMMR[2]과 VIBE[18]의 정성적 결과. 모든 결과는 해당 논문에서 가져왔다.

이 모듈은 과거와 미래 프레임으로부터 추가적인 시간적 특징(temporal feature)을 예측하여 정적 특징(static feature)과 결합한다. 마지막으로, 과거, 미래 및 모든 프레임의 시간적 특징을 통합하여 최종 3차원 메쉬를 예측한다. 정성적 결과는 〈그림 4〉에서 확인할 수 있다.

V. 손실 함수

휴먼 메쉬 복원의 성공 여부는 학습에 사용된 손실 함수에 크게 의존한다. 일반적으로 널리 사용되는 손실함수로는 L1, L2 손실함수를 들 수 있다. 휴먼 메쉬 복원을 위해 기존에 제안된 많은 손실 함수들이 있는데, 본 기고문에서 우리는 이 손실 함수들을 키포인트, 모델 매개변수, 그리고 메쉬에 대한 세 가지로 분류한다. 키포인트 손실 함수는 2차원 및 3차원 공간에서 추정된 키포인트와 그에 대응하는 참값과의 차이를 계산한다. 2차원 키포인트는 3차원 키포인트와 예측된 카메라 파라미터를 이용해 생성된다. 그러나 키포인트 손실 함수만으로는 성공적인 휴먼 메쉬 복원이 어렵다. 이는 네트워크가 사람의 형상을 고려하지 않고 자세에만 의존하게 되기 때문이다. 따라서 예측된 메쉬는 일반적으로 신뢰하기 어려운 휴먼 형상이 된다. 이를 해결하기 위해 휴먼 모델 매개변수 손실 함수 또는 메쉬 손실 함수라는 두 가지 방법이 제안되었다. 메쉬 손실 함수(L_M)는 예측된 정점의 위치와 참값 간의 차이를 계산한다. 반면에, 모델 매개변수(L_{HM}) 손실 함수는 예측된 자세, 형상 매개변수와 그 참값 간의 차이를 계산한다.

[8]은 휴먼 메쉬를 영상 공간에서의 휴먼 마스크로 나타낸 후 이를 활용하는 실루엣 손실 함수(L_{Sil})를 제안했고, 이 손실 함수는 휴먼 복원 결과를 추가적으로 향상시킬 수 있는 것으로 알려져 있다. 그러나 2차원과 3차원 공간 사이의 모호함으로 인해 실루엣 손실 함수(L_{Sil})만으로는 불충분하다. 따라서 이를 보완하기 위해, [10]은 벡터 정규화 손실 함수(L_{Normal}) 와 엣지(edge) 길이 손실 함수(L_{Edge})를 통해 복원 결과를 개선할 수 있음을 보여주었다.

VI. 데이터셋

대부분의 휴먼 메쉬 복원 모델은 지도 학습(supervised learning)에 기반하며, 학습을 위해 참값과 함께 그에 대응하는 영상 또는 비디오가 필요하다. 이 장에서는 본 기고문의 목적과 가장 관련성이 높은 데이터셋들에 대해 설명한다. 편의를 위해, 우리는 데이터셋들

을 세 가지로 분류했고, 〈표 1〉은 분류된 데이터셋들의 요약을 보여준다.

3차원 주석(3D annotation) : Human3.6M[4]은 휴먼 메쉬 복원 연구에서 가장 일반적으로 사용되는 데이터셋이다. 이 데이터셋에는 11명의 배우가 15가지 유형의 행동을 수행한다. 또한 마커를 기반으로 정확하게 획득된 3차원 자세, 경계 상자(bounding box) 및 실루엣 정보를 포함하는 주석을 얻을 수 있다. MPI-INF-3DHP[5]는 8명의 배우에 의해 수행되는 8개의 서로 다른 행동 집합들로 구성되어 있으며, 3차원 자세 주석은 정확도가 다소 떨어지는 마커리스(marker-less) 모션 캡처로부터 얻는다. 두 데이터셋은 모두 제어된 실내 환경에서 만들어졌다. 이와는 대조적으로, 3DPW[27]는 in-the-wild 환경에서 60개의 비디오를 캡처해서 만들어졌고, IMU센서를 사용하여 SMPL 자세 및 형상 매개변수를 수집하였다. 앞서 언급된 모든 데이터셋은 비디오 기반 데이터셋에 속한다.

2차원 주석(2D annotation) : 최근 연구에서는 일반적으로 LSP[23], LSP-Ext[24], MPI[19], MSCOCO[28] 데이터셋이 학습에 사용된다. 이 데이터셋들은 2차원 주석이 있는 단일 영상을 제공하고 각자 다른 순서로 인덱싱되어 있다. 또한 Penn-Action[29] 및 PoseTrack[20] 데이터셋도 비디오로부터의 휴먼 메쉬 복원 문제에 일반적으로 사용된다. Jin 등은 인체 메쉬 복원 뿐만 아니라 손가락과 얼굴을 동시에 복원하는 모델의 학습에도 사용할 수 있는 손가락 및 얼굴 랜드마크에 대한 주석을 추가하여 MSCOCO 데이터셋을 확장했다[25].

대부분의 3차원 주석 데이터셋은 in-the-wild 시나리오를 충분히 반영하지 못하는 실내 환경에서 만들어진다. 정확한 주석을 확보하기 위해 3차원 주석에 다시점(multi-view) 기반의 피팅이 필요하기 때문이다. 그에 반해 2차원 데이터셋은 in-the-wild 환경에서 만들어지지만 앞서 설명한 이유로 3차원 주석을 포함하지 않는다.

의사 주석(pseudo annotation) : 앞서 언급된 데이터셋의 단점은 구축을 위해 많은 시간과 비용을 필요로 한다는 것이다. 이를 보완하기 위해 Kanazawa 등은

<표 1> 휴먼 메쉬 복원에 일반적으로 사용되는 데이터셋 요약

주석 타입	데이터셋 이름	출처	크기	해상도	장면 타입		주석 내용		
					실내	실외	2D 자세	3D 자세	SMPL
3D 주석	Human3.6M [4]	TPAMI 2014	3,600,000	1000×1000	✓	x	✓	✓	x
	MPI-INF-3DHP [5]	3DV 2017	1,300,000	~2048×2048	✓	x	✓	✓	x
	3DPW [27]	ECCV 2018	51,000	~1080×1920	x	✓	✓	✓	✓
2D 주석	LSP [23]	BMVC 2010	2,000	다양함	x	✓	✓	x	x
	LSP-Ext [24]	CVPR 2011	10,000	다양함	x	✓	✓	x	x
	MPI [19]	CVPR 2014	25,000	~1920×1080	✓	✓	✓	x	x
	MSCOCO [28]	ECCV 2014	~200,000	다양함	✓	✓	✓	x	x
	COCO-WholeBody [25]	ECCV 2020	~200,000	다양함	✓	✓	✓	x	x
	Penn Action [29]	ICCV 2013	163,841	480×270	x	✓	✓	x	x
	PoseTrack [20]	CVPR 2018	46,000	~640×380	✓	✓	✓	x	x
의사 (Pseudo) 주석	InstaVariety [2]	CVPR 2019	2,100,000	다양함	✓	✓	✓	x	x

<표 2> 3DPW[27], MPI-INF-3DHP[5], 그리고 Human3.6M[4]에서의 state-of-the-art 방법들 평가표. 모든 수치는 해당 논문에 보고된 값이다.

입력	알고리즘	3DPW				MPI-INF-3DHP			Human3.6M		
		PA-MPJPE	MPJPE	MPVPE	Accel	PA-MPJPE	MPJPE	Accel	PA-MPJPE	MPJPE	Accel
영상	HMR [1]	76.7	130.0	-	37.4	89.8	124.2	-	56.8	88.0	-
	GraphCMR [22]	70.2	-	-	-	-	-	-	50.1	-	-
	SPIN [21]	59.2	96.9	116.4	29.8	67.5	105.2	-	41.1	-	-
	Zhang et al [26]	72.2	-	-	-	-	-	-	41.7	-	-
	I2L-MeshNet [10]	57.7	93.2	110.1	30.9	-	-	-	41.1	55.7	13.4
	HKMR [7]	-	-	-	-	-	-	-	-	59.6	-
	Biggs et al [3]	55.6	**75.8**	-	-	-	-	-	42.2	58.2	-
	METRO [15]	**47.9**	77.1	**88.2**	-	-	-	-	36.7	**54.0**	-
	HybrIK [14]	48.8	80.0	94.5	-	-	**91.0**	-	**34.5**	54.4	-
비디오	HMMR [2]	72.6	116.5	139.3	15.2	-	-	-	56.9	-	-
	VIBE [18]	56.5	93.5	113.4	27.1	63.4	97.7	29.0	41.5	65.9	18.3
	TCMR [12]	55.8	95.0	111.3	**6.7**	**62.8**	96.5	**9.5**	41.1	62.3	**5.3**

InstaVariety[2]라는 새로운 비디오 데이터셋을 제안했다. 해당 비디오 데이터는 인스타그램[34]에서 84개의 인간 활동 해시태그를 사용해 수집해서 만들어졌다. 2차원 자세는 모든 프레임에 대해 OpenPose[31]를 실행함으로써 획득되었고, 추가적으로 SMPL 매개변수를 생성하기 위해 SMPLify가 사용될 수 있다. 생성된 2차원 자세 정보가 의사 주석이기는 하지만, [2]에서 입증되었듯이 3차원 휴먼 복원을 위해 사용될 수 있고 이는 모델의 성능을 향상시킬 수 있다고 알려져 있다.

VII. 성능 평가

휴먼 메쉬 복원 성능을 평가하려면 영상 또는 비디오로부터 예측된 자세 및 메쉬를 다양한 측면에서 평가해야 한다. 첫째, 자세 유사성은 다음과 같은 MPJPE(Mean Per Joint Position Error)를 사용하여 평가될 수 있다.

$$MPJPE = \frac{1}{J}\sum_{j=1}^{J}\left\|\boldsymbol{P}_j - \boldsymbol{P}_j^*\right\|_2, \qquad (1)$$

여기서 P_j와 P_j^*는 각각 j번째 관절에 대한 추정치와 참값을 나타낸다. 다만, 각 데이터셋마다 관절 갯수와 형식이 다르기 때문에 일반적으로 LSP에서 정의된 표준 자세 포맷에 해당하는 14개의 관절만 비교한다. 이는 예측된 메쉬에 선형 회귀 행렬을 곱해 얻어진 관절 집합과 그 참값을 비교함으로써 수행 가능하다. 둘째, 추정된 3차원 자세와 참값 자세에 Procrustes analysis를 적용한 후 MPJPE를 계산하는 PA-MPJPE를 들 수 있다. 이는 회전과 스케일을 제외하고 순수한 자세의 차이를 나타낸다. 셋째, 메쉬 유사성은 일반적으로 MPVPE(Mean Per Vertex Position Error)를 사용하여 평가된다.

$$MPVPE = \frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{V}_i - \boldsymbol{V}_i^*\right\|_2, \qquad (2)$$

여기서 V_i와 V_i^*는 각각 i번째 정점에 대한 추정치와 참값을 나타낸다. 〈표 2〉에서 모든 평가 지표는 밀리미터(mm) 단위의 평균 유클리드 거리를 사용하여 계산되었다.

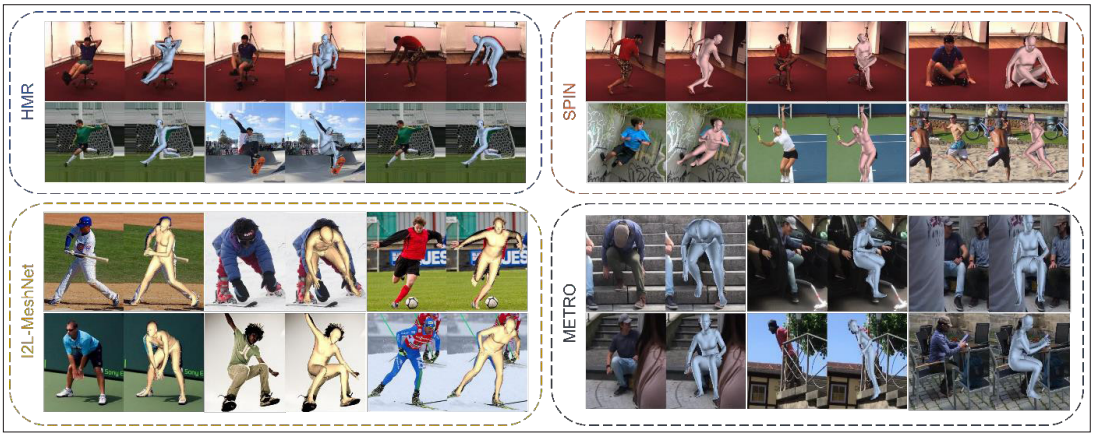
그 외에도 영상 기반 접근법에 적용될 수 있는 다른 추가적인 평가 기준들이 존재한다. 첫째는 LSP 테스트 데이터셋을 사용해 전경과 배경의 정확도와 F1 점수를 평가하는 방법이다. 전경은 예측된 메쉬와 그 참값 정보를 영상 공간에 투영하여 얻을 수 있다. 두 번째는 MPI-INF-3DHP 데이터셋을 사용한 PCK(Percentage of Correct Keypoint) 및 AUC(Area Under Curve)이다.

비디오 기반 휴먼 메쉬 복원 문제에서는 가속 오차(acceleration error)라는 평가 기준이 사용된다. 이 평가 기준은 각 키포인트의 3차원 가속도를 계산하여, 추정된 모션 시퀀스가 얼마나 매끄러운지를 평가한다.

〈그림 5〉는 모델 기반 방법과 모델을 사용하지 않는 방법들의 정성적 결과를 보여준다. 여기서 첫 번째 행은 전자의 방법에 대응하고, 두 번째 행은 후자의 방법에 대응한다. 모델 기반 방법은 자세 및 형상 매개변수를 추정함으로써 휴먼 메쉬를 추정할 수 있다는 단순성 때문에 널리 사용된다. 그러나 SMPL과 같은 휴먼 모델은 3차원 회전 표현에 의존하는데, 이는 periodicity, non-minimal representation, 혹은 불연속성(discontinuity)과 같은 몇 가지 문제를 야기할 수 있다. 그 결과 모델 기반 방식은 모델을 사용하지 않는 방법들에 비해 자세 추정 측면에서 상대적으로 낮은 성능을 보임을 〈표 2〉에서 확인할 수 있다. 반면에, 모델을 사용하지 않는 접근법은 상대적으로 더 나은 자세 추정 성능을 달성한다. 그러나 이 방법은 학습을 위해서 참값 메쉬를 필요로 하지만 대부분의 데이터셋에서는 메쉬 정보가 제공되지 않는다는 단점을 가진다. 또한, 정점의 수가 과도하게 많기 때문에 이를 고려해서 네트워크를 효율적으로 설계할 필요가 있다.

Ⅷ. 결 론

본 기고문에서 우리는 단일 RGB 영상 또는 비디오로부터 휴먼 메쉬를 복원하는 방법들에 대해서 살펴보았



<그림 5> HMR[1], SPIN[21], I2L-MeshNet[10], and METRO[15]에서의 정성적 결과. 모든 결과는 해당 논문에서 가져왔다.

다. 또한 손실 함수에 대해서도 조사하였고, 일반적으로 사용되는 최신 데이터셋 및 평가 기준들에 대해서도 정리하였다. 마지막으로 본 기고문에서는 각 접근법의 장단점을 분석하였고, 그들의 정량적 및 정성적 결과를 정리하였다. 필자들은 본 기고문이 국내의 3차원 휴먼 복원 연구에 도움이 될 수 있기를 희망한다.

참 고 문 헌

[1] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik, End-to-End Recovery of Human Shape and Pose, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018), pp. 7122-7131.

[2] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik, Learning 3D Human Dynamics From Video, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019), pp. 5614-5623.

[3] Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Benjamin Graham, and Andrea Vedaldi, 3D Multi-bodies: Fitting Sets of Plausible 3D Human Models to Ambiguous Image Data, Proc. Neural Information Processing Systems (2020).

[4] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu, Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments, IEEE Trans. on Pattern Analysis and Machine Intelligence (2014), 36(7):1325-1339.

[5] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt, Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision, Proc. International Conference on 3D Vision (2017), pp. 506-516.

[6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black, Keep It SMPL: Automatic Estimation of {3D} Human Pose and Shape from a Single Image, Proc. European Conference on Computer Vision (2016), pp. 561-578.

[7] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Kosecka, and Ziyan Wu, Hierarchical Kinematic Human Mesh Recovery, Proc. European Conference on Computer Vision (2020), pp. 768-784.

[8] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis, Learning to Estimate {3D} Human Pose and Shape From a Single Color Image, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018), pp. 459-468.

[9] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black, Expressive Body Capture: 3D Hands, Face, and Body From a Single Image, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019), pp. 10975-10985.

[10] Gyeongsik Moon and Kyoung Mu Lee, I2L-MeshNet: Image-to-Lixel Prediction Network for Accurate 3D Human Pose and Mesh Estimation from a Single RGB Image, Proc. European Conference on Computer Vision (2020), pp. 752-768.

[11] Hanbyul Joo, Tomas Simon, and Yaser Sheikh, Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018), pp. 8320-8329.

[12] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee, Beyond Static Features for Temporally Consistent {3D} Human Pose and Shape from a Video, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021).

[13] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu, GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020), pp. 6183-6192.

[14] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu, HybrIK: A Hybrid Analytical-Neural Inverse Kinematics Solution for Human Pose and Shape Estimation, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021).

[15] Kevin Lin, Lijuan Wang, and Zicheng Liu, End-to-End Human Pose and Mesh Reconstruction with Transformers, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021).

[16] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V. Gehler, and Bernt Schiele, DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation, Proc. IEEE Conference on Computer Vision and Pattern Recognition (2016), pp. 4929-4937.
[17] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black, SMPL: A skinned multi-person linear model, ACM Trans. on Graphics (2015), 34(6):248:1-248:16.
[18] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black, VIBE: Video Inference for Human Body Pose and Shape Estimation, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020), pp. 5252-5262.
[19] Mykhaylo Andriluka, Leonid Pishchulin, Peter V. Gehler, and Bernt Schiele, 2D Human Pose Estimation: New Benchmark and State of the Art Analysis, Proc. IEEE Conference on Computer Vision and Pattern Recognition (2014), pp. 3686-3693.
[20] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele, PoseTrack: A Benchmark for Human Pose Estimation and Tracking, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5167-5176.
[21] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis, Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop, Proc. IEEE/CVF International Conference on Computer Vision (2019), pp. 2252-2261.
[22] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis, Convolutional Mesh Regression for Single-Image Human Shape Reconstruction, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019), pp. 4501-4510.
[23] Sam Johnson and Mark Everingham, Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation, Proc. British Machine Vision Conference (2010), pp. 1-11.
[24] Sam Johnson and Mark Everingham, Learning effective human pose estimation from inaccurate annotation, Proc. IEEE Conference on Computer Vision and Pattern Recognition (2011), pp. 1465-1472.
[25] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo, Whole-Body Human Pose Estimation in the Wild, Proc. European Conference on Computer Vision (2020), pp. 196-214.
[26] Tianshu Zhang, Buzhen Huang, and Yangang Wang, Object-Occluded Human Shape and Pose Estimation from a Single-Color Image, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020), pp. 7374-7383.
[27] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll, Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera, Proc. European Conference on Computer Vision (2018), pp. 614-631.
[28] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick, Microsoft COCO: Common Objects in Context, Proc. European Conference on Computer Vision (2014), pp. 740-755.
[29] Weiyu Zhang, Menglong Zhu, and Konstantinos G. Derpanis, From Actemes to Action: A Strongly-Supervised Representation for Detailed Action Understanding, Proc. IEEE International Conference on Computer Vision (2013), pp. 2248-2255.
[30] Xiangyu Xu, Hao Chen, Francesc Moreno-Noguer, Laszlo A. Jeni, and Fernando De la Torre, 3D Human Shape and Pose from a Single Low-Resolution Image with Self-Supervised Learning, Proc. European Conference on Computer Vision (2020), pp. 284-300.
[31] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh, OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields, IEEE Trans. on Pattern Analysis and Machine Intelligence (2019), 43(1):172-186.
[32] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black, AMASS: Archive of Motion Capture As Surface Shapes, Proc. IEEE/CVF International Conference on Computer Vision (2019), pp. 5441-5450.
[33] Ian Goodfellow, Jean Pouget-Abadi, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengi, Generative Adversarial Nets, Proc. Neural Information Processing Systems (2014), pp. 2672-2680.
[34] https://www.instagram.com

필자소개

조슈아 산토소

- 2019년 8월 : 인도네시아 비누스대학교 컴퓨터과학과 학사
- 2021년 8월 : 인하대학교 전기컴퓨터공학과 석사
- 주관심분야 : 3차원 휴먼 복원, 딥러닝

전성호

- 2021년 2월 : 광운대학교 로봇학부 학사
- 2021년 3월 ~ 현재 : 광운대학교 전자통신공학과 석사과정
- 주관심분야 : 컴퓨터비전 및 그래픽스(영상기반 3차원 휴먼 복원), 딥러닝

장주용

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

박인규

- 1995년 2월 : 서울대학교 제어계측공학과 학사
- 1997년 2월 : 서울대학교 제어계측공학과 석사
- 2001년 8월 : 서울대학교 전기컴퓨터공학부 박사
- 2001년 9월 ~ 2004년 2월 : 삼성종합기술원 전문연구원
- 2007년 1월 ~ 2008년 2월 : Mitsubishi Electric Research Laboratories 방문연구원
- 2014년 9월 ~ 2015년 8월 : MIT Media Lab 방문부교수
- 2018년 7월 ~ 2019년 6월 : University of California, San Diego(UCSD) 방문학자
- 2004년 3월 ~ 현재 : 인하대학교 정보통신공학과 교수
- ORCID : https://orcid.org/0000-0003-4774-7841
- 주관심분야 : 컴퓨터비전 및 그래픽스, deep learning, GPGPU