단안 이미지 기반 3차원 사람 모델 복원 연구 동향

ㅁ **김원준** / 건국대학교

요약

본 기고에서는 한 장의 이미지로부터 3차원 사람 모델을 복원하는 최신 기술 동향을 살펴보고자 한다. 최근 심층학습의 성공에 힘입어 3차원 모델링 분야에서도 복잡한 최적화 기법을 이용하는 전통적인 모델링 방법 대신 심층학습을 기반으로 한 다양한 방법들이 활발히 소개되고 있다. 따라서, 본 기고에서는 심층학습을 기반으로 한 대표적인 3차원 사람 모델 복원 방법들을 중심으로 연구 흐름을 살펴보고, 실험 결과 분석을 통해 앞으로의 연구 방향을 제시하고자 한다.

I. 서론

최근 메타버스를 기반으로 한 다양한 플랫폼 및 서비스가 크게 증가하면서 정교한 사람 모델을 생성하는 연구에 대한 관심 또한 증가하고 있다. 한 장의 이미지만을 이용하여 3차원 사람 모델을 복원하는 방법은 대부분 최적화 과정을 반복하여 수행되어 왔으나, 복잡한 배경 및 가려짐이 존재하는 장면에서 복원 성능이 좋지 못한 단점이 있다. 한편, 다양한 컴퓨터 비전 분야에서 심층학습의 성공에 힘입어 최근 3차원 사람 모델 복원 연구에서도 이러한 심층학습을 기반으로 한 방법들이 활발히 소개되기 시작하고 있다. 심층학습을 기반으로 한 방법은 복잡한 최적화 모델을 이용하는 대신 대용량의 데이터셋을 기반으로 3차원 모델의 파라미터(Parameter)를 예측하는 학습 프레임워크를 이용하여 3차원 사람 모델 복원 성능을 성공적으로 향상시켰다.

심층학습을 기반으로 한 방법은 크게 모델 기반(Model-based) 방법과 모델을 이용하지 않는(Model-free) 방법

으로 나눌 수 있다. 먼저, 모델 기반의 방법은 가장 대표적인 3차원 사람 모델인 SMPL(Skinned Multi-person Linear) 모델[1]의 두 개의 파라미터, 즉, 자세(Pose)와 체형(Shape) 파라미터를 학습을 통해 예측하는 것을 목표로 한다. 모델 기반의 방법은 이러한 SMPL 파라미터를 이용하여 3차원 메쉬(Mesh) 구조를 정교하게 복원할 수 있는 장점이 있지만, 사람 이외의 다른 객체의 3차원 모델을 동일한 신경망을 통해 복원하지 못하는 단점이 있다. 이와 달리, 모델을 이용하지 않는 방법은 신경망을 통해 3차원 메쉬를 구성하고 있는 꼭짓점의 좌표를 직접 예측하기 때문에 SMPL 파라미터에 의한 제약이 없으며, 다양한 객체에 대한 3차원 모델 복원에 동일한 신경망 및 학습 전략을 적용할 수 있다. 그러나, 꼭짓점 간 공간적 제약이 없기 때문에 부드러운 메쉬 표면 생성에 어려움이 있다. 다음 장에서는 설명한 각 카테고리의 대표 방법들을 중심으로 연구 동향을 자세히 살펴보고자 한다.

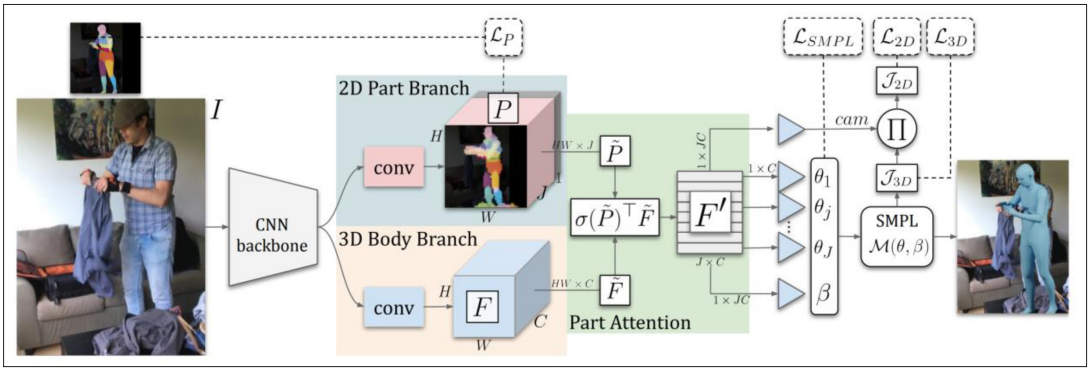
II. 심층학습 기반 3차원 사람 모델 복원 방법

본 장에서는 한 장의 이미지로부터 3차원 사람 모델을 복원하는 방법을 모델 기반의 방법과 모델을 이용하지 않는 방법으로 나누어, 각 방법의 연구 흐름과 대표 연구 방법들의 특징을 자세히 살펴보고자 한다.

1. 모델 기반(Model-based) 방법

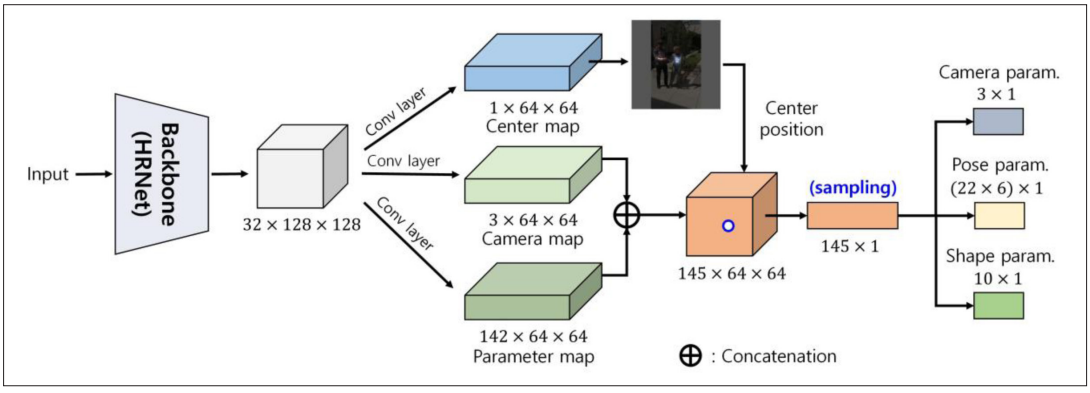
모델 기반의 방법은 먼저 심층학습을 통해 SMPL 모델의 파라미터를 예측할 수 있도록 압축기-복원기(Encoder-Decoder) 신경망 구조를 개발하였다. 이후 정밀한 예측을 위해 사람의 신체 관절 간의 관계를 고려한 운동학적 위상 기반 복원기를 연구하였으며, 가장 최근에는 가려짐에 강인한 3차원 모델 복원을 위한 다양한 학습 전략이 소개되고 있다.

자세히 살펴보면, Kanazawa[2] 등은 압축기를 통해 예측한 SMPL 파라미터를 이용하여 3차원 메쉬 구조를 생성하고, 이를 2차원 평면으로 투영하여 정답 관절 위치와의 비교를 통해 학습하는 방법을 제안하였다. 자연스러운 자세 생성을 위해 제작된 3차원 모델과의 분별력을 향상시킬 수 있는 적대 손실(Adversarial Loss) 함수를 추가로 사용하였다. Kolotouros[3] 등은 [2]와 마찬가지로 기본적인 압축기를 통해 예측된 SMPL 파라미터를 반복적 최적화 과정을 통해 정제하여 이를 학습에 이용하는 방법을 제안하였다. 2차원 평면으로 투영된 관절 위치 대신 파라미터 값의 차이를 직접 이용한 학습을 통해 보다 정밀한 3차원 사람 모델 복원을 수행할 수 있음을 보였다. 한편으로, 단안 비디오 영상에서 프레임 간 시간적 연관

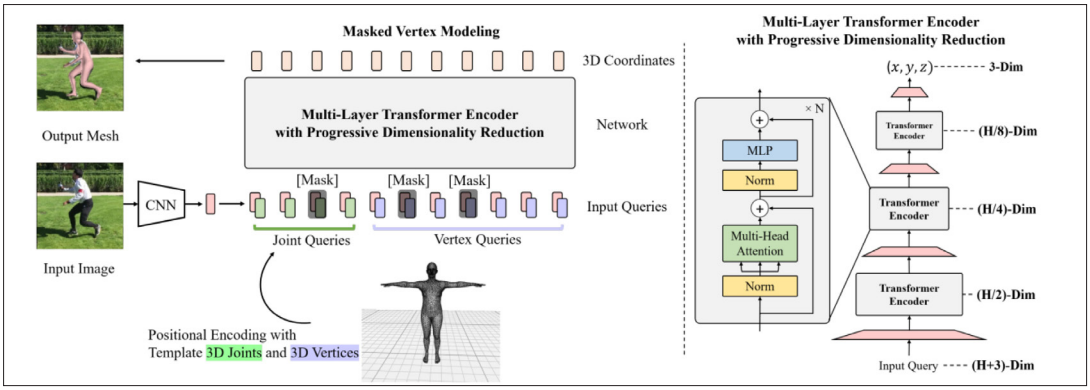


<그림 1> PARE[7] 모델 신경망 상세 구조 (각 신체 부위별 특징을 함께 고려)

성을 바탕으로 3차원 사람 모델을 복원하는 연구도 소개되었다. Kocabas[4] 등은 프레임 간 관계를 순환 신경망(Recurrent Neural Network)을 이용하여 학습하였으며, Wan[5] 등은 시공간 트랜스포머(Transformer)를 이용하여 해당 정보를 인코딩(Encoding)하여 3차원 사람 모델 복원에 사용하였다. Zhang[6] 등은 피라미드 구조를 통해 예측한 SMPL 파라미터로 초기 3차원 메쉬 구조를 복원하고 이를 압축된 특징 지도에 투영하여 해당 좌표의 특징들만을 이용하여 자세를 정제하는 방법을 제안하였다. 가장 최근에는 다양한 가려짐에 강인한 3차원 사람 모델 복원을 위해 Kocabas[7] 등은 각 신체 부위 분할 결과와 압축된 특징 간 자기 주의(Self-attention) 연산을 통해 각 신체 부위의 가시성(Visibility)을 학습에 이용하였다(〈그림 1〉 참조). 이를 통해 가려진 신체 부위를 나머지 신체 부위와의 관계를 통해 예측할 수 있도록 하였다. Sun[8] 등은 예측한 사람의 중심 좌표에서 SMPL 파라미터를 샘플링하는 간단한 신경망 구조를 제안하였다. 사람 영역을 분할하여 신경망 입력으로 사용하는 대부분의 기존 방법과 달리, 중심 좌표를 예측하는 상향식(Bottom-up) 방법을 제안하여 여러 명이 존재하는 이미지에서도 정확하게 각 사람의 3차원 모델을 성공적으로 복원할 수 있음을 보였다(〈그림 2〉 참조). 그밖에도 3차원 깊이 정보를 기반으로 스케일 모호성을 억제하여 3차원 모델 복원 성능을 높인 최신 방법도 소개되었다[9].



<그림 2> ROMP[8] 모델 신경망 상세 구조 (사람 중심 좌표에서 특징을 샘플링)



<그림 3> 트랜스포머 기반 3차원 사람 모델 복원의 예[11]

2. 모델을 이용하지 않는(Model-free) 방법

모델을 이용하지 않는 방법은 각 꼭짓점 좌표를 추정하기 위한 신경망 구조 설계 연구부터 시작되었으며, 부드러운 메쉬 표면 복원을 위해 트랜스포머 기반 꼭짓점 간 관계 학습 방법이 제안되었다. 이후, 각 신체 부위 간 관계 또한 지역적 특성을 기반으로 학습하기 위해 그래프 모델을 트랜스포머와 함께 사용하는 구조가 소개되었으며, 최근에는 트랜스포머를 다양하게 변형하여 정확도 및 수행 속도 향상을 위한 연구가 활발히 진행되고 있다.

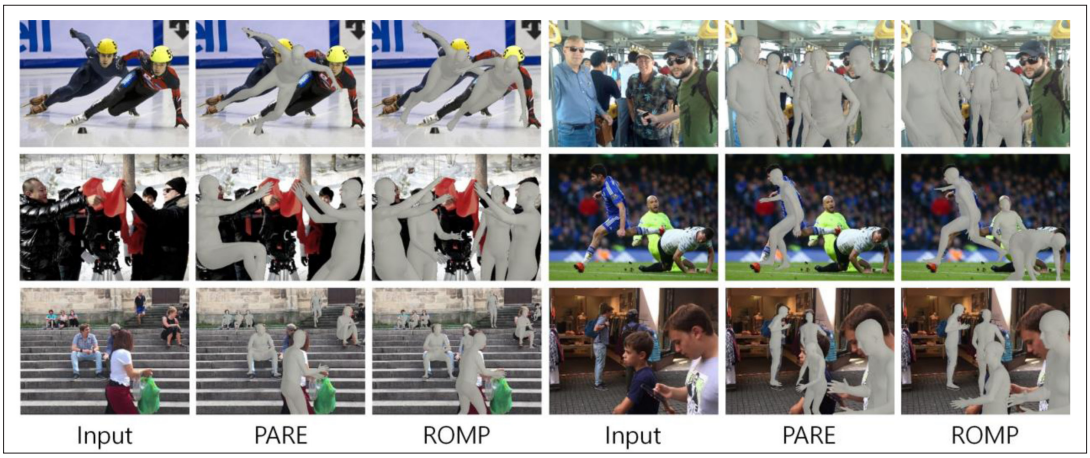
자세히 살펴보면, Kolotouros[10] 등은 메쉬를 구성하는 각 꼭짓점을 그래프 노드(Node)로 표현하고, 신경망을 통해 압축된 특징을 각 노드에 할당하여 그래프 합성곱(Graph Convolution) 연산을 통해 모든 꼭짓점의 좌표를 예측하는 방법을 제안하였다. Lin[11] 등은 신체 부위 관절과 메쉬의 각 꼭짓점과의 관계를 학습하기 위해 트랜스포머 기반 신경망 구조를 제안하였으며(〈그림 3〉 참조), 가려진 신체 부위도 이러한 전역적 관계 학습을 통해 효과적으로 복원할 수 있음을 보였다. 저자는 더 나아가 관절 및 꼭짓점의 기하학적 특징뿐만 아니라 이미지 특징을 함께 고려할 수 있도록 트랜스포머의 토큰(Token)을 설계하였으며, 각 신체 부위 간 지역적 관계 또한 학습할 수 있도록 그래프 합성곱 연산을 트랜스포머 구조에 적용하였다[12]. Cho[13] 등은 트랜스포머 구조의 경량화를 위해 이미지 특징과 관절 특징을 분리하여 학습할 수 있는 트랜스포머 압축기-복원기 구조를 제안하였다. Kim[14] 등은 메쉬 구조의 각 꼭짓점을 2차원 이미지 공간으로 투영한 후 해당 위치에서 특징을 샘플링(Sampling)하여 트랜스포머 구조를 통해 3차원 메쉬 좌표를 예측하는 방법을 제안하였다. 그밖에도 트랜스포머를 기반으로 하는 다양한 신경망 구조를 통해 보다 자연스러운 메쉬 표면을 생성하는 방법들이 최근까지도 활발하게 연구되고 있다.

III. 성능 평가

단안 이미지로부터 3차원 사람 모델을 복원하는 방법들의 성능 평가는 기본적으로 생성한 결과 내 관절 위치의 정확도를 기반으로 수행된다. 대표적으로 다음의 세 가지 평가 메트릭(Metric)을 사용한다. 자세히 살펴보면,

<표 1> 3차원 사람 모델 복원 성능 비교 (MB: Model-based, MF: Model-free)

Method	Type	Human3.6M		3DPW		
		MPJPE	PA-MPJPE	MPJPE	PA-MPJPE	MPVPE
SPIN [3]	MB	-	41.1	96.9	59.2	116.4
VIBE [4]	MB	65.6	41.4	82.9	51.9	99.1
MAED [5]	MB	56.4	38.7	79.1	45.7	92.6
PyMAF [6]	MB	57.7	40.5	92.8	58.9	110.1
PARE [7]	MB	-	-	74.5	46.5	88.6
ROMP [8]	MB	-	-	76.7	47.3	93.4
BEV [9]	MB	-	-	78.5	46.9	92.3
GCMR [10]	MF	-	50.1	-	70.2	-
METRO [11]	MF	54.0	36.7	77.1	47.9	88.2
Graphorm. [12]	MF	51.2	34.5	74.7	45.6	87.7
FastMETRO [13]	MF	52.2	33.7	73.5	44.6	84.1
PointHMR [14]	MF	48.3	32.9	73.9	44.9	85.5



<그림 4> 3차원 사람 모델 복원의 정성적 성능 비교 (CrowPose 데이터셋 기반)

<그림 5> 3차원 사람 모델 복원의 정성적 성능 비교 (Human3.6M과 3DPW 데이터셋 기반)

MPJPE(Mean Per Joint Position Error)는 예측된 관절의 3차원 위치와 정답 위치의 차이를 측정하는 메트릭이며, PA-MPJPE(Procrustes-aligned Mean Per Joint Position Error)는 사람 모델의 고관절 방향을 정렬한 후 관절 예측 정확도를 계산한다. 생성된 3차원 사람 모델의 메쉬 형태의 자연스러움을 측정하기 위한 메트릭으로는 MPVPE(Mean Per Vertex Position Error)가 사용된다. 이와 같은 메트릭을 통해 대표적인 방법들에 대한 정량적 성능 평가를 수행한 결과를 〈표 1〉에 나타내었다. 모든 메트릭의 값은 작을수록 성능이 우수함을 의미한다. 여기서 성능 평가를 위해 3차원 사람 모델 복원 분야의 가장 대표적인 데이터셋인 Human3.6M[15]과 3DPW[16]를 사용

하였다. 표의 결과를 살펴보면, 가장 최근에는 모델을 이용하지 않는 방법들의 성능이 크게 향상되고 있음을 확인할 수 있다. 이러한 성능 향상은 다양한 트랜스포머 기반 신경망 구조 및 학습 전략을 통해 메쉬 구조의 각 꼭짓점 간 상호 관계 등을 효과적으로 반영한 결과라고 볼 수 있다.

정성적인 성능 비교 결과를 〈그림 4〉와 〈그림 5〉에 나타내었다. 〈그림 4〉는 모델 기반 방법 중 대표적인 PARE[7]와 ROMP[8]를 이용한 복원 결과를 보여주고 있으며, 복잡한 가려짐에도 신뢰도 있는 성능을 확인할 수 있다. 특히, ROMP 방법의 경우 다른 사람에 의한 가려짐이나 다른 신체 부위에 의한 가려짐에도 좋은 복원 성능을 보여주고 있다. 〈그림 5〉는 모델을 사용하지 않는 방법 중 대표적인 방법들을 기반으로 한 복원 결과를 보여주고 있다. PointHMR[14]의 경우, 대상 사람의 스케일을 정확하게 예측하여 모델을 생성하고 있으며 장애물에 의한 다양한 가려짐에도 다른 방법 대비 정확하게 3차원 사람 모델을 복원하고 있음을 확인할 수 있다.

IV. 결 론

본 기고문에서는 한 장의 이미지로부터 3차원 사람 모델을 복원하는 최신 기술 동향을 살펴보았다. 모델 기반의 방법과 모델을 이용하지 않는 방법의 장단점을 심도 있게 분석하였으며, 각 카테고리의 대표적인 방법들의 특성 또한 함께 살펴보았다. 대표적인 데이터셋을 기반으로 한 정량적, 정성적 성능 평가 결과 및 분석 내용을 제시함으로써 3차원 사람 모델 복원 분야의 연구자들에게 연구 가이드를 효과적으로 제시하였다.

참 고 문 헌

[1] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “SMPL: A skinned multi-person linear model,” ACM Trans. Graph., vol. 34, no. 6, pp. 248:1-248:16, 2015.

[2] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, “End-to-end recovery of human shape and pose,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp. 7122-7131, Jun. 2018.

[3] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, “Learning to reconstruct 3D human pose and shape via model-fitting in the loop,” in Proc. IEEE Int. Conf. Comput. Vis., pp. 2252-2261, Oct. 2019.

[4] M. Kocabas, N. Athanasiou, and M. J. Black, “VIBE: Video inference for human body pose and shape estimation,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp. 5253-5263, Jun. 2020.

[5] Z. Wan, Z. Li, M. Tian, J. Liu, S. Yi, and H. Li, “Encoder-decoder with multilevel attention for 3D human shape and pose estimation,” in Proc. IEEE Int. Conf. Comput. Vis., pp. 13033-13042, Oct. 2021.

[6] H. Zhang, Y. Tian, X. Zhou, W. Ouyang, Y. Liu, L. Wang, and Z. Sun, “PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop,” in Proc. IEEE Int. Conf. Comput. Vis., pp. 11446-11456, Oct. 2021.

[7] M. Kocabas, C.-H. P. Huang, O. Hilliges, and M. J. Black, “PARE: Part attention regressor for 3D human body estimation,” in Proc. IEEE Int. Conf. Comput. Vis., pp. 11127-11137, Oct. 2021.

[8] Y. Sun, Q. Bao,W. Liu, Y. Fu, M. J. Black, and T. Mei, “Monocular, one-stage, regression of multiple 3D people,” in Proc. IEEE Int. Conf. Comput. Vis., pp. 11179-11188, Oct. 2021.

[9] Y. Sun, W. Liu, Q. Bao, Y. Fu, T. Mei, and M. J. Black, “Putting people in their place: monocular regression of 3D people in depth,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp. 13243-13252, Jun. 2022.

[10] N. Kolotouros, G. Pavlakos, and K. Daniilidis, “Convolutional mesh regression for single-image human shape reconstruction,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp. 4501-4510, Jun. 2019.
[11] K. Lin, L. Wang, and Z. Liu, “End-to-end human pose and mesh reconstruction with transformers,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp. 1954-1963, Jun. 2021.
[12] K. Lin, L. Wang, and Z. Liu, “Mesh graphormer,” in Proc. IEEE Int. Conf. Comput. Vis., pp. 12939-12948, Oct. 2021.
[13] J. Cho, Y. Kim, and T.-H. Oh, “Cross-attention of disentangled modalities for 3D human mesh recovery with transformers,” in Proc. Eur. Conf. Comput. Vis., pp. 342-359, Oct. 2022.
[14] J. Kim*, M.-G. Gwon*, H. Park, H. Kwon, G.-M. Um, and W. Kim, “Sampling is matter: point-guided 3D human mesh reconstruction,” in Proc. IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12880-12889, Jun. 2023 (* equally contributed).
[15] 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 Trans. Pattern Anal. Mach. Intell., vol. 36, no. 7, pp. 1325-1339, Jul. 2013.
[16] T. V. Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, “Recovering accurate 3D human pose in the wild using IMUs and a moving camera,” in Proc. Eur. Conf. Comput. Vis., pp. 601-617, 2018.

저 자 소 개

김 원 준

- 2012년 8월 : 한국과학기술원(KAIST) 박사
- 2012년 9월 ~ 2016년 2월 : 삼성종합기술원 전문연구원
- 2016년 3월 ~ 2020년 2월 : 건국대학교 전기전자공학부 조교수
- 2020년 3월 ~ 현재 : 건국대학교 전기전자공학부 부교수
- 주관심분야 : 컴퓨터 비전, 영상처리, 기계학습