이미지 기반 완전 3D 인간 복원 기술 동향

*송대영 **이희경 **서정일 *조동현

*충남대학교 **한국전자통신연구원

*eadyoung@naver.com **lhk95@etri.re.kr **seoji@etri.re.kr *cdh12242@cnu.ac.kr

Trends of Full 3D Human Reconstruction Technology Based on Image

*Song, Dae-Young **Lee, HeeKyung **Seo, Jeongil *Cho, Donghyeon

*Chungnam National University **Electronics and Telecommunications Research Institute

요 약

이미지 기반 3D 형상 복원에 있어서, 이미지에 보이지 않는 폐색(Occlusion) 영역 부분에 대한 정보가 손실되므로 완전한 복원에 어려움이 있으며, 세밀한 텍스쳐(Texture) 표현이 이루어지지 않고 심한 평활화(Smoothing)나 고립된 노이즈 메쉬(Isolated Noise Mesh) 등 구조적 훼손이 발생한다. 주로 깊은 신경망을 이용하여, 음함수(Implicit Function) 기반 방법은 사전 훈련이 완료된 보조 신경망들을 전면부에 배치하거나, Hourglass 등 임베딩(Embedding) 아키텍처를 추가하거나, 또는 표면 법선(Surface Normal)과 같은 환시(Hallucination)를 생성하여 신경망에 입력하기도 한다. 이 논문에서는, 인물의 이미지를 입력받아 색상, 머리카락 및 의상을 포함하는 완전 3D 인간 복원 기술들을 조망해본다.

1. 서론

최근 메타버스(Metaverse) 관련 기술에 대한 수요 증가로, 딥러닝 기반의 인간 아바타(Avatar) 복원이 시도되고 있다. 크게 분류해보면, 이미지로 사람의 자세(Pose)만을 추정하는 연구[1,2,3,4], 파라미터(Parameter) 기반 복원 연구[5,6,7,8,9], 앞 연구와 다르게 색상, 의상, 머리카락 등을 모두 복원하는 음함수 기반 연구[10,11,13,14,19],, NeRF[15]와 유사하게 접근하는 연구[16,17,18] 등이 있다. 본 논문은 그 중 다층 퍼셉트론(Multi-Layer Perceptron, 이하 MLP)에 기반하여, 이미지를 임베딩하여 모든 3차원 형상을 복원하는 음함수 기반 복원 기술에 초점을 둔다.

물체를 3차원으로 표현하는 방법에 대한 연구는 다양한 방향으로 진행되어 왔다. 복셀(Voxel) 기반 표현, 포인트 클라우드(Point Cloud) 기반 표현, 점유(Occupancy) 기반 표현 등이 그렇다. 복셀 표현은 공간 복잡도가 크다는 단점이 있고, 포인트 클라우드 기반 방법은 클라우드를 양자화하는 등 후처리 단계를 추가 요구한다는 단점이 있다. 점유 맵(Map) 기반 방법은 3차원 그리드(Grid) 상에서 점유 여부를 [0, 1] 범위의 확률로 표현하게 되며, 메모리 문제에서 비교적 효율적이고, 점유 맵을 병렬 처리가 가능한 마칭 큐브(Marching Cubes)[20] 알고리즘으로 3차원 메쉬(Mesh) 시각화가 가능하다는 장점이 있어, PIFu[10]를 선두로 하는 음함수 방법의 최종 표현 형식으로 사용되고 있다.

입력 이미지로부터 완전한 3차원 이미지를 추론하는 것은 깊이 추정, 표면의 고변동(High Frequency) 표현, 폐색 영역에 대한 모호성 등 다양한 문제가 산재해있다. 깊이 추정의 경우, 일반적인 단일 이미지 환경에서는 3차원 복원에 대한 360° 임베딩을 충분히 다룰 수 없기 때문에, StereoPIFu[19]와 같은 연구를 제외하면 직접적으로 깊이를 이용한 시도는 많지 않다. 대신 PIFu와 같이 2D 정보로부터 직접 점유맵을 추정하거나, 깊이가 아닌 3D 임베딩을 도와주는 여러 사전 훈련된 모듈(Module)을 이식하는 시도[11,13,14]들이 존재한다.

인간 형상을 표현하는 음함수 방법들은 대개 단일 이미지만 입력받아도 3차원 복원이 가능하다는 특장점이 있지만, 깊이 추정에 대한 모호함, 폐색 영역에 대한 추론이 취약하다는 한계가 있으며, 본론에 소개될 연구들은 이를 극복하는 일련의 시도들이라고 할 수 있겠다.

2. 본론

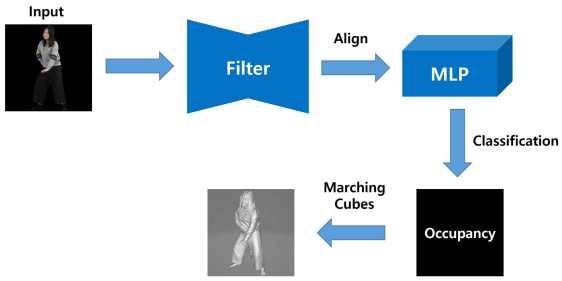

그림 1. PIFu[10]의 문제 정의.

2D 이미지만을 활용하여 인간 복원을 위한 점유 맵을 추정하는 방법은 PIFu[10]에서 처음 시도되었다. 다음과 같이,

$$f_v^*(X) = \begin{cases} 1, X\text{가 표면 내부인 경우} \\ 0, \text{그 밖의 경우} \end{cases} \quad (1)$$

이진 분류(Binary Classification) 문제로 정의한다. 이 때, f_v^*를 음함수라고 칭하며, 그림 1과 같이 2D 입력 임베딩을 위한 Filter에는 Hourglass[2] 구조, 분류에는 MLP 구조를 활용한다.

Hourglass 구조는 자세 추정을 위해 디자인된 아키텍처로, 인코더

(Encoder)-디코더(Decoder) 간 대칭적인 구조를 띈다. 최종적인 자세 추정을 위해 합성곱의 수용 영역(Receptive Field)을 벗어난 전체 형상에 대한 이해가 필요하기 때문에, 인코더로부터 크기 별로 특징 맵의 잔차 연결(Skip-connection)을 만들고, 이를 그대로 디코더에 입력하는 것이 아니라, 합성곱 계층을 통과시켜 연결하는 특징이 있다.

1) 표면 법선 추정 방법

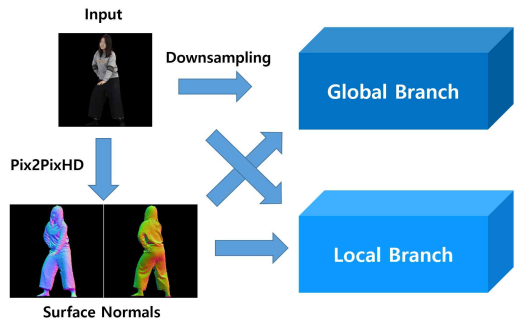

그림 2. PIFuHD[11]의 법선 입력 추가 모식도.

단일 이미지 입력만으로는 이미지에 보이는 전면부 정보만 알 수 있기 때문에, 보이지 않는 부분에 대한 추론 성능을 향상하는 데에 초점을 맞추어 PIFu의 후속 연구들이 진행되었다. 그 중에는 보조 신경망들을 전면부에 도입하여 모델의 표면 법선을 추정하고, 이를 추가 입력 자원으로 사용하는 시도가 있었다. 그림 2와 같이 PIFuHD[11]는 두 개의 PIFu 아키텍처를 사용하며 모델의 전체적인 특징과 지엽적인 특징을 임베딩하기에 앞서, Pix2pixHD[12] 모듈을 전면부에 배치하여 입력 이미지의 전면부, 후면부 표면 법선을 추정하게끔 훈련한 뒤 추정된 출력값을 추가 입력으로 사용한다. 이를 통해, 깊이 및 고주파 영역에 대한 세밀한 표현에 도움을 줄 뿐만 아니라, 폐색 영역에 대한 추론 시 데이터셋의 분포에 기반한 추가적인 조건 형성(Conditioning)이 가능하다.

2) 통계적 파라미터 모델 추정 방법

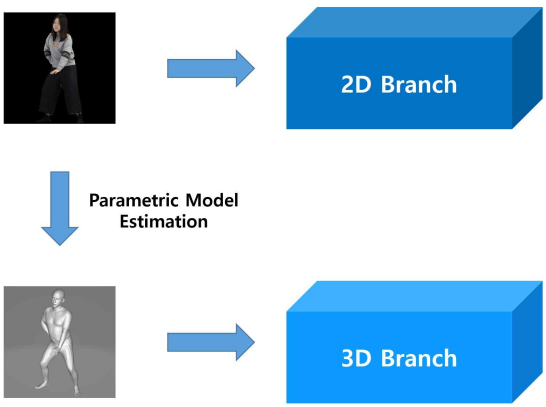

그림 3. PaMIR[13]의 3D 정보 입력 추가 모식도.

모델의 전·후면부에 대한 법선 정보를 추가 입력한다고 해도, 3D 복원에 반드시 필요한 깊이 정보가 부족할 수 있다. 그림 3에서, PaMIR[13]는 복원에 앞서 입력 이미지로부터 사전 훈련된 GraphCMR[21] 신경망을 활용하여 통계 모델인 SMPL[5]의 파라미터를 추정한 뒤, 이로부터 만든 메쉬를 복셀화(Voxelization)하여 3D 합성곱 신경망을 위한 모양으로 변형한다. 이 추가 정보는 3D 복원에 도움이 되는 깊이 정보를 MLP에 제공하게 된다.

여기에서 추정되는 SMPL은 의상이나 머리카락에 대한 정보가 없는 인체에 대한 통계적 템플릿(Template)으로, 안정적인 기하 정보를 모델에 제공하기는 하지만, 피부로부터 먼 거리에 존재하는 정보는 소실된다. 이러한 점에 착안하여, ICON[14]은 SMPL을 일단 사전 훈련된 PyMAF[9] 신경망으로 추정한 뒤, 이를 전·후면 표면 법선 이미지로 렌더링(Rendering)하고, 별도의 보조 신경망으로 잃어버린 기하 정보를 회복하게끔 훈련시킨다. ICON은 PaMIR에 비해 통계 모델에서 손실된 부분에 대한 추정에 비교적 강인하고, 복원된 모델을 워핑(Warping)시켜 동작을 부여했을 때 기하 구조가 잘 훼손되지 않지만, 치마와 같이 피부로부터 거리가 너무 멀고 세밀한 고변동 표면 표현이 필요할 때 성능이 저하되는 단점이 있다.

3. 결론

본 논문에서 서술한 인간 형상 복원 방법들은 다각도 입력으로 확장이 가능하지만, 주로 단일 이미지를 입력받을 때 폐색 영역에 대한 추론 모호성 및 깊이 추정에 집중하여 360° 임베딩을 시도한 것으로 요약할 수 있다. 이러한 방법으로는 표면 법선 추정, 파라미터 모델 사용, 보조 임베딩 모듈 추가 등이 있다.

현재까지도 폐색 영역을 포함하는 복원 기법을 다루는 음함수 기반 방법의 개선은 매우 어려운 문제인데, 얼굴이 보이지 않는 이미지를 입력하거나 몸의 일부분이 가려진 이미지를 입력으로 사용한다면 해당 부분에 심한 평활화나 끊어짐 등 구조적 훼손이 발생하기 때문이다. 또한 머리카락과 같은 흔들림이 심한 정보의 경우, 비디오에서 안정성을 보장하기 어렵다. 뿐만 아니라, 이진 분류를 수행하는 MLP는 합성곱 신경망에서 가정할 수 있는 귀납적 편향(Inductive Bias)을 활용하기 어렵다는 문제도 있다. 이러한 문제점들은 주로 필터 및 MLP와 함께 다양한 기능을 수행하는 보조 신경망 모듈을 두어 극복하는 방향으로 시도되고 있으며, 앞으로도 활발한 연구가 필요하다.

감사의 글

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2018-0-00207, 이머시브 미디어 전문연구실)

참고문헌

[1] Tompson, Jonathan J., et al. "Joint training of a convolutional network and a graphical model for human pose estimation." *Advances in Neural Information Processing Systems* 27 (2014).

[2] Newell, Alejandro, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation." *European Conference on Computer Vision*. Springer, Cham, 2016.

[3] Cao, Zhe, et al. "Realtime multi-person 2d pose estimation using part affinity fields." *Proceedings of the IEEE/CVF Conference on Computer Vision and*

Pattern Recognition. 2017.
[4] Cao, Zhe, et al. “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields” *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2019.
[5] Loper, Matthew, et al. "SMPL: A skinned multi-person linear model." *ACM transactions on graphics* 34.6. 1-16. 2015.
[6] Kanazawa, Angjoo, et al. "End-to-end recovery of human shape and pose." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2018.
[7] Choi, Hongsuk, et al. "Beyond static features for temporally consistent 3d human pose and shape from a video." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2021.
[8] Sun, Yu, et al. "Monocular, one-stage, regression of multiple 3d people." *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 2021.
[9] Zhang, Hongwen, et al. "Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop." *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 2021.
[10] Saito, Shunsuke, et al. "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization." *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 2019.
[11] Saito, Shunsuke, et al. "Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2020.
[12] Wang, Ting-Chun, et al. "High-resolution image synthesis and semantic manipulation with conditional gans." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2018.
[13] Zheng, Zerong, et al. "Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction." *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2021.
[14] Xiu, Yuliang, et al. "ICON: Implicit Clothed humans Obtained from Normals." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2022.
[15] Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." *European Conference on Computer Vision.* Springer, Cham, 2020.
[16] Peng, Sida, et al. "Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2021.
[17] Xu, Hongyi, Thiemo Alldieck, and Cristian Sminchisescu. "H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion." *Advances in Neural Information Processing Systems* 34 2021.
[18] Shao, Ruizhi, et al. "Doublefield: Bridging the neural surface and radiance fields for high-fidelity human rendering." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2022.
[19] Hong, Yang, et al. "Stereopifu: Depth aware clothed human digitization via stereo vision." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[20] Lorensen, William E., and Harvey E. Cline. "Marching cubes: A high resolution 3D surface construction algorithm." *ACM siggraph computer graphics* 21.4. 163-169. 1987.
[21] Kolotouros, Nikos, Georgios Pavlakos, and Kostas Daniilidis. "Convolutional mesh regression for single-image human shape reconstruction." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2019.