다중 인물 포함 단일 영상으로부터의 파라미터 기반 3차원 휴먼 모델 생성 기법 성능 비교 연구

*엄기문, **김정환, ***김원준, *이희경, *양승준, *서정일

*한국전자통신연구원 실감미디어연구실, **건국대학교 인공지능학과, ***건국대학교 전기전자공학부

*gmum@etri.re.kr

A comparative study on the performance of the parameter-based 3D human model generation techniques from a single image including multiple people

*Gi-Mun Um, **Jeong Hwan Kim, ***Wonjun Kim, *Hee Kyung Lee, *Seung-Jun Yang, *Jeongil Seo

*Electronics and Telecommunications Research Institute(ETRI), Communication & Media Research Laboratory, Media Research Division, Immersive Media Research Section

**Konkuk University, Department of Electrical and Electronics Engineering

***Konkuk University, Department of Artificial Intelligence

요약

본 논문에서는 다중 인물 포함 단일 영상으로부터 파라미터 기반 3차원 휴먼 모델 생성 기법 중 최근 발표된 SOTA 기법 4가지에 대해 대표적인 데이터 셋들에 대해 사전 학습 모델을 사용한 복원 성능 비교 실험을 수행하였다. 실험결과, CLIFF 기법과 PyMAF-x 기법이 PARE 기법이나 ROMP 기법에 비해 우수한 결과를 보였다.

1. 서론

AR/VR/XR(Extended Reality) 기술이 발전하고, 코로나19로 인해 비대면 서비스가 급격하게 성장하면서 현실 세계와 동기화된 메타버스 공간에서 제공하는 다양한 서비스가 등장하고 있고 관련 기술 개발 경쟁도 치열하게 일어나고 있다. 특히 메타버스 공간 내에서 사용자의 모습을 닮은 아바타나 가상인간 [1] 등 3D 휴먼 모델 복원 기술에 대한 관심이 높아지고 있다.

3D 휴먼 모델을 복원하기 위한 기술로는 크게 스튜디오 기반의 복원 기술과 일반 환경에서의 복원 기술로 나눌 수 있다 [2]. 이러한 스튜디오 기반의 복원 기술은 수십 대의 다중 카메라를 설치하고 획득된 영상으로부터 사람의 3차원 정보를 추출하고 이를 3D 휴먼 모델로 복원한 후 텍스처를 포함하여 사람의 3차원적 움직임을 렌더링하는 기술이며, 마이크로 소프트 Mixed Reality Capture Studio[3], 8i의 실감 스튜디오[4], 인텔 Studios [5] 등이 속한다.

이들 스튜디오는 최대 복원 대상 인원이 1 – 4인이 대부분이나 큰 규모를 가지는 스튜디오의 경우 50명이상도 가능하고, 카메라 대수는 최소 24대에서 100대 이상에 이른다. 따라서 이러한 스튜디오 기반의 복원기술은 고품질의 3D 휴먼 모델은 얻을 수는 있지만 스튜디오 구축을 위한 공간이나 비용, 처리 시간에 대한 제약이 많은 실정이다.

한편, 딥러닝 기반의 AI 기술이 발전하고 컴퓨터 비전 분야에도 딥러닝 기법이 많이 도입되면서 일반 환경하에서 3D 휴먼 모델 복원 기술도 많이 연구되고 있다. 여기에는 인텔의 RealSense[6], 마이크로소프트의 Kinect[7] 스테레오 랩스(Stereo Labs)의 ZED[8] 등의 범용 깊이 카메라를 1대 또는 여러 대 사용하여 복원하는 DynamicFusion 등 [9]과 같은 기법들과, 깊이 카메라 영상을 사용하지 않으면서 한 장 또는 소수의 RGB 컬러 영상만을 이용하여 3D 휴먼 모델을 복원하는 기법인 PIFu(Pixel-aligned implicit function) 기법과 같은 기법들도 최근 많이 제안되고 있다[10].

본 논문에서는 이렇게 다양한 3D 휴먼 모델 복원 기법 중에서 최근 ICCV, CVPR이나 ECCV 학회에 제출되고 발표된 논문 중 단일 RGB 컬러 영상을 입력으로 사용하는 대표적인 기법들을 선정하고, 입력 영상 내 다중 객체를 포함한 데이터 셋을 이용하여 파라미터 기반의 3차원 휴먼 모델 생성 성능을 정성적으로 비교 분석한 결과를 소개하기로 한다.

2. 비교 대상 단일 RGB 컬러 영상 기반 다중 객체 3차원 휴먼 모델 생성 기법

본 논문에서 비교 실험 대상으로 선택된 단일 RGB 컬러 영상 기반 다중 객체 3차원 휴먼 모델 생성 기법들은 ICCV2021, CVPR2022 학회와 ECCV2002 학회에 제출되었으며, Github 상에 소스가 공개된 기법들 중심으로 선정하였다. 실험 대상으로 선정된 첫 번째 기법은 ROMP

(Regress One-stage Multiple 3D People) 기법[11]이다. 이 기법은 전체 영상으로부터 다중 분리 가능한 세가지 맵인 신체 중심 히트맵, 카메라 맵, SMPL(Skinned Multi-Person Linear Model) 파라미터 맵을 추정하며, 신체 중심별로 휴먼 3D 자세와 형태를 복원하는 기법이다. 그림 1은 ROMP 기법의 구성도이다. 다음으로 선정된 기법은 PARE(Part Attention Regressor) 기법 [12]으로서 그림 2와 같이 신체 부위별로 주목되는 부위를 알아내고 주목되는 가중치를 학습하고, 보이는 영역의 특징을 취합하여 가려진 부위의 성능을 개선한 기법이다. 세 번째 비교 실험 대상으로 선정된 기법은 CLIFF(Carrying Location Information in Full Frames)[13] 기법이다. 이 기법은 다중 객체 검출시 cropping 된 영상과 사용된 바운딩 박스(bounding box) 위치를 이용하여 3차원 관절정보를 2차원 영상으로 투영했을 때의 2차원 투영오차를 계산하여 이를 손실함수를 사용함으로써 전역적 위치와 회전정보를 보다 정확하게 예측함으로서 추정된 자세의 정확도를 높이는 기법이다. 그림 3은 CLIFF 기법의 구성도를 나타내고 있다. 마지막 네 번째 비교 대상 기법으로 선택된 기법은 PyMAF(Pyramidal Mesh Alignment Feedback Loop)-x 기법 [14]인데, 이 기법은 특징 피라미드를 구성하여 예측된 파라미터의 정확도를 개선하는 regression 기반 PyMAF 기법을 확장하여 손목과 팔 부분의 미세 정확도를 개선한 기법이다. 그림 4는 PyMAF-x 기법의 전체 구성을 나타내고 있는데, 그림에서 보듯이 몸통과 손, 얼굴별로 PyMAF 신경망[15]를 개별적으로 적용한 후 통합하는 구조로 되어 있다.

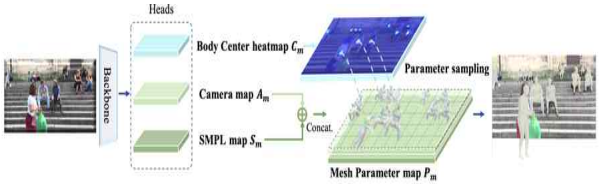


그림 1. ROMP 기법 구성도 [10]

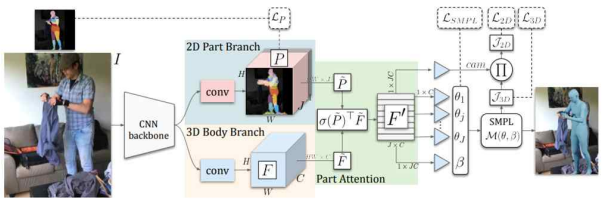


그림 2. PARE 기법 구성도 [11]

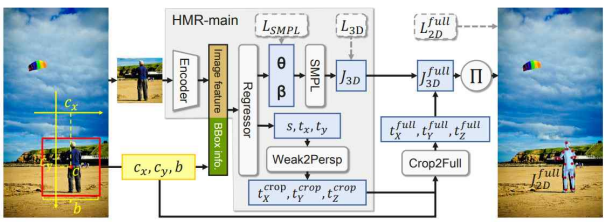


그림 3. CLIFF 기법 구성도 [12]

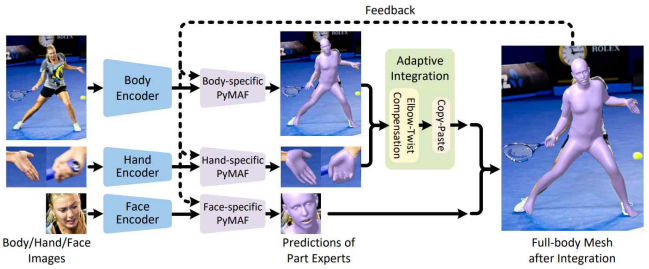


그림 4. PyMAF-x 기법 구성도 [13]

3. 비교 실험 결과

II장에서 소개한 4 가지 제안한 단일 RGB 컬러 영상 기반 다중 객체 3차원 휴먼 모델 생성 기법들 간의 성능을 정성적으로 분석하기 위해 Crowd Pose 데이터 셋 [16] 중 다중 인물을 포함한 정지영상 9장과 3DPW[17] 데이터 셋 1장을 사용하여 비교실험을 수행하였다. 비교실험에 사용된 신경망 학습모델은 사전 학습된(pre-trained) 모델을 사용하였다. 그림 5~그림 9는 각 기법의 비교 실험 결과의 예를 나타내고 있다.

먼저 그림 5와 6을 보면, ROMP 기법의 경우 사람의 몸이나 얼굴 시선의 방향을 제대로 복원하지 못하는 경우가 많으며, PARE 기법의 경우 ROMP 기법보다는 얼굴의 방향을 잘 복원하고 있으나, 일부 작은 사람의 경우 검출이 안되는 문제점이 보여주고 있다.

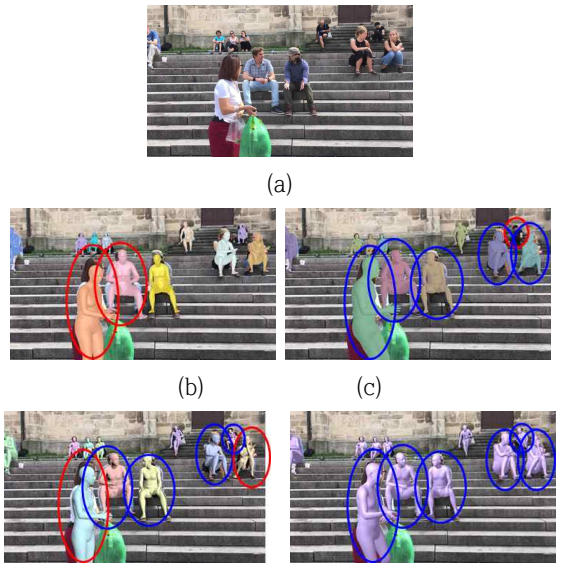
다음으로 CLIFF 기법의 경우 전반적으로 우수한 성능을 보이고 있으나 일부 손이나 팔 모양에서 오차를 보이고 있다. 한편, PyMAF-x 기법의 경우 CLIFF 기법과 거의 유사한 성능을 보이고 있으며, 팔이나 손 모양에서도 가장 좋은 성능을 보이고 있다. 한편, 그림 7과 8의 결과를 보면, ROMP 기법의 경우에는 객체는 빠짐없이 검출되었으나, 객체 간 순서 및 얼굴이나 몸 방향이 맞지 않는 결과를 보이고 있다. PARE 기법의 경우에도 객체 간 순서가 맞지 않고, 일부 객체가 검출되지 않은 경우도 발생하고 있다. CLIFF 기법의 경우에는 객체간 순서도 잘 유지되고 있으며, PyMAF-x 기법의 경우에는 일부 후방 객체가 검출되지 않는 오류가 있었으나 그 외에은 CLIFF 기법과 거의 유사한 복원 결과를 보여주고 있으며, 얼굴 표정까지 잘 복원하고 있다.

또한, 그림 9의 경우에는 객체 수가 2명임에도 객체 간 겹침 정도가 높음으로 인해 CLIFF 기법을 비롯한 4 가지 기법 모두가 좋은 복원결과을 보여주지 못하고 있다. ROMP 기법의 경우에는 남녀 두 선수가 검출은 되나 여자 선수와 남자 선수의 팔이 섞이고, 선수 간 앞뒤 순서와 몸과 시선방향이 맞지 않은 결과를 보였다. PARE 기법과 CLIFF 기법의 경우에는 중첩에 의해 여자 선수가 검출되지 않는 결과를 보였고, PyMAF-x 기법의 경우에는 두 선수가 모두 검출되고 몸과 시선방향 측면에서도 가장 좋은 결과를 보여주지만, 여전히 남녀 선수 순서는 뒤바뀌는 결과를 보여주고 있다.

다음으로, 정성적 결과와의 일관성 확인을 위해 SOTA(state-of-the-art) 알고리즘 벤치마크 사이트 [18] 에 제시된 4 가지 기법의 정량적 척도인 PA-MPJPE(Procrustes Analysis)-MPJPE(Mean Per Joint Position Error)값과 MPJPE값은 표 1과 같다 [17]. 각 척도의 정의는 참고문헌 [19]에 정의되어 있다.

표 1. 3DPW 데이터 셋에 대한 3D 휴먼모델 복원 기법 정량적 평가 결과

비교 기법	MPJPE	PA-MPJPE
ROMP	76.7	47.3
PARE	74.5	46.5
CLIFF(HR-W48)	69	43
PyMAF-x	74.2	45.3

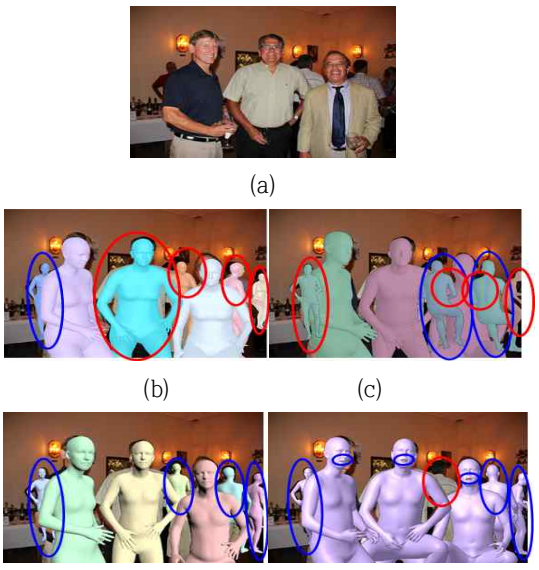


(a)

(b) (c)

(d) (e)

그림 5. 3DPW 데이터 영상에 대한 3D 휴먼 모델 생성 결과

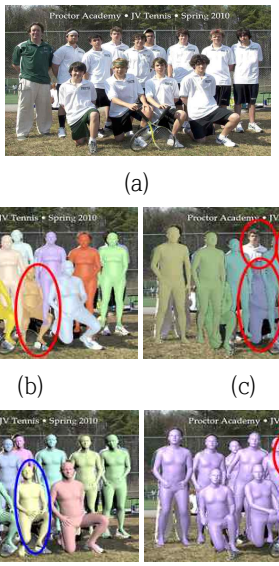


(a)

(b) (c)

(d) (e)

그림 6. Crowd Pose 데이터 영상 1에 대한 3D 휴먼 모델 생성 결과 예 (a) 입력 영상(Crowd Pose 1) (b) ROMP (c) PARE (d) CLIFF (e) PyMAF-x

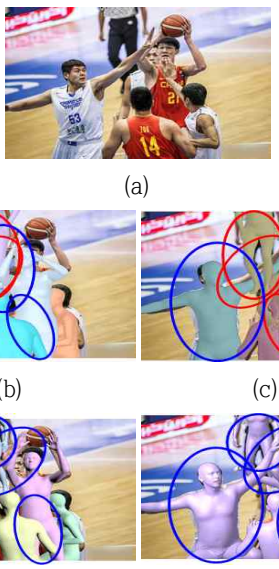


(a)

(b) (c)

(d) (e)

그림 7. Crowd Pose 데이터 영상 2에 대한 3D 휴먼 모델 생성 결과 예 (a) 입력 영상(Crowd Pose 2) (b) ROMP (c) PARE (d) CLIFF (e) PyMAF-x



(a)

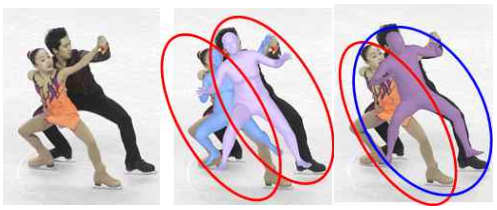
(b) (c)

(d) (e)

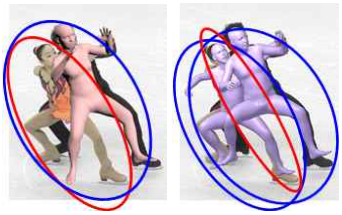
그림 8. Crowd Pose 데이터 영상 3에 대한 3D 휴먼 모델 생성 결과 예 (a) 입력 영상(Crowd Pose 3) (b) ROMP (c) PARE (d) CLIFF (e) PyMAF-x

4. 결론 및 추후 과제

본 논문에서는 단일 RGB 컬러 영상을 입력으로 사용하는 대표적인 기법들 중에서 최근에 발표된 단일 영상 내 다중 객체의 3D 휴먼 모델을 복원 가능한 4 종류 SOTA 기법들을 선정하고 비교 실험을 통해 성능을 비교하였다.



(a) (b) (c)



(d) (e)

그림 9. Crowd Pose 데이터 영상 4에 대한 3D 휴먼 모델 생성 결과 예 (a) 입력 영상(Crowd Pose 4) (b) ROMP (c) PARE (d) CLIFF (e) PyMAF-x

실험결과에서 보듯이 대부분의 영상에서 가장 좋은 성능을 보인 기법은 CLIFF 기법이었으며, PyMAF-x, PARE, ROMP 기법 순으로 정성적 및 정량적 복원 성능을 보여주었다. 향후 과제로는 각 기법의 장단점을 고려하여 얼굴 표정, 손 모양까지 고려하고, 텍스처와 옷 입은(clothed) 3D 휴먼 모델까지 확장하기 위한 기법에 대한 연구가 필요하다.

감사의 글

본 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2018-0-00207, 이머시브 미디어 전문연구실, No. 2021-0-02084, 비대면 실감 경험 공유를 위한 XR, Volumetric 실감미디어 생성 및 전송 기술 개발 및 한-유럽 국제공동연구)

참고문헌

[1] 한상열, “메타버스 신인류, 디지털 휴먼,” Issue Report, IS-135, 22.1.14.
[2] 박민규, 강주미 윤주홍, “메타버스 서비스를 위한 휴먼 모델링 기술 동향,” 3차원 객체 및 공간의 모델링 및 시각화 기술 특집, 방송과 미디어 제 26권 4호, pp. 61-71, 2021년 10월.
[3] https://www.microsoft.com/en-us/mixed-reality/capture-studios
[4] https://8i.com/
[5] https://newsroom.intel.com/press-kits/intel-studios/#gs.fh1he
[6] https://www.intelrealsense.com/
[7] https://azure.microsoft.com/ko-kr/services/kinect-dk/
[8] https://www.stereolabs.com/zed-2/
[9] R. A. Newcombe, D. Fox, and S. M. Seitz, “DynamicFusion: Reconstruction and Tracking of Non-rigid Scenes in Real-Time,” Proceedings of CVPR 2015.
[10] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li,“PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization,” Proceedings of the International Conference on Computer Vision(ICCV) 2019, 2019.
[11] Y. Sun, Q. Bao, W. Liu, Y. Fu, M. J. Black, T. Mei, “Monocular, One-stage, Regression of Multiple 3D People,” Proceedings of ICCV2021.
[12] M. Kocabas, C. P. Huang, O. H. M. J. Black, “PARE: Part Attention Regressor for 3D Human Body Estimation,” Proceedings of ICCV2021.
[13] Z. Li, J. Liu, Z. Zhang, S. Xu, Y. Yan, “CLIFF: Carrying Location Information in Full Frames into Human Pose and Shape Estimation,” Proceedings of ECCV2022.
[14] H.Zhang, Y. Tian, Y. Zhang, M. Li, L. An, Z. Sun, Y. Liu,“ PyMAF-X: Towards Well-aligned Full-body Model Regression from Monocular Images,” https://arxiv.org/abs/2207.06400v2.
[15] H. Zhang, Y. Tian, X. Zhou, W. Ouyang, Y. Liu, L. Wang, Z. Sun,“ PyMAF: 3D Human Pose and Shape Regression with Pyramidal Mesh Alignment Feedback Loop,” Proceedings of ICCV2021.
[16] https://github.com/Jeff-sjtu/CrowdPose
[17] https://virtualhumans.mpi-inf.mpg.de/3DPW/
[18] https://paperswithcode.com/sota/3d-human-pose-estimation-on-3dpw
[19] https://github.com/cbsudux/Human-Pose-Estimation-101