2D 렌더링 정보를 활용한 손-객체의 3D 복원

남현길, 박종일*
한양대학교
{ skagusrlf, jipark }@hanyang.com

Hand-Object 3D Reconstruction Based on 2D Rendering

Hyeongil Nam Jong-Il Park
Hanyang University

요 약

본 논문은 RGB 영상 데이터셋의 일부만을 지도학습하여(Sparsely-supervised learning) Annotation 되지 않은 영상에 대해 손-객체의 3D 포즈를 복원하기 위한 방법을 제안한다. 기존의 연구에서는 손-객체의 포즈에 해당하는 6DoF 만을 학습 데이터로 활용한다. 이와 달리, 본 논문에서는 정확도 향상을 위해 복원된 결과를 동일한 입력 영상 내에서 비교 가능하도록 3D 모델로 복원한 결과를 입력 영상의 마스크로 만들어 학습에 반영하였다. 구체적으로 추정된 포즈로 만들어낸 마스크를 입력 영상에 적용한 결과와 Ground-truth 포즈를 적용한 영상을 학습 시에 손실 함수에 반영하였다. 비교 실험을 통해 제안된 방법이 해당 방법을 적용하지 않은 경우 보다 3D 매쉬 오차가 적었음을 확인할 수 있었다.

1. 서론

최근 MR(Mixed Reality) 및 로봇 분야 등의 다양한 곳에서 RGB 영상으로 손-객체를 3D 로 복원하는 연구가 증가하고 있다[1, 2, 3]. 손-객체 복원을 통해 가상현실과 실제 현실을 연결해 주는 인터랙션 방법으로 사용된다. 이를 위해, 근본적으로 3D 손-객체의 포즈(6 DoF)를 복원할 수 있어야 한다. 본 연구는 데이터셋의 일부 데이터만을 학습하여, Annotation 이 되지 않은 영상에 대해서도 손-객체 포즈 추정 정확도를 보장하고자 하였다. 구체적으로 학습 과정에서 손-객체의 포즈를 추정한 정보로 3D 모델을 복원하여 입력 영상의 마스크를 생성한다. 그리고 이를 입력 영상에 해당 마스크를 적용한 결과와 Ground-truth 포즈 정보를 동일하게 만든 결과를 손실 함수에 반영하였다. 기존의 연구에서는 2D - 3D keypoint 들만의 6 DoF 정보만을 손실 함수에 사용하는 경우가 많다[7]. 또한 객체 만의 포즈를 추정하는 선행 연구에서는 occlusion 을 감안하기 위해서 3D 모델을 2D 렌더링한 결과와 영상 이미지 사이의 차이를 그대로 손실 함수에 적용한 연구들이 있었다[1, 2, 3, 4]. 그러나 손-객체 포즈를 모두 추정하는 연구에서 2D 렌더링한 결과를 사용한 효과를 직접 검증가 되지 않았다. 또한 3D 모델을 2D 렌더링한 것과 실제 이미지를 비교하는 경우, 렌더링 자체와 이미지 상태에 차이가 있기 때문에 직접적인 비교에 어려움이 있다. 합성 데이터를 사용하여 적용한 경우에는 데이터셋 자체부터 현실과 괴리감이 존재한다. 본 논문에서 역시 occlusion 을 감안하기 위해, 학습 과정에서 손-객체 포즈를 모두 이용하여 렌더링한 2D 정보를 사용하지만, 선행연구와는 달리 복원된 결과와 입력 영상을 직접적으로 비교하지 않는다. 대신, Ground truth 와 추정된 포즈로 렌더링하여 각각 실제 이미지 영역을 마스크로 만들고 이를 입력 영상에 반영함으로써 간접적으로 추정된 결과

*교신저자

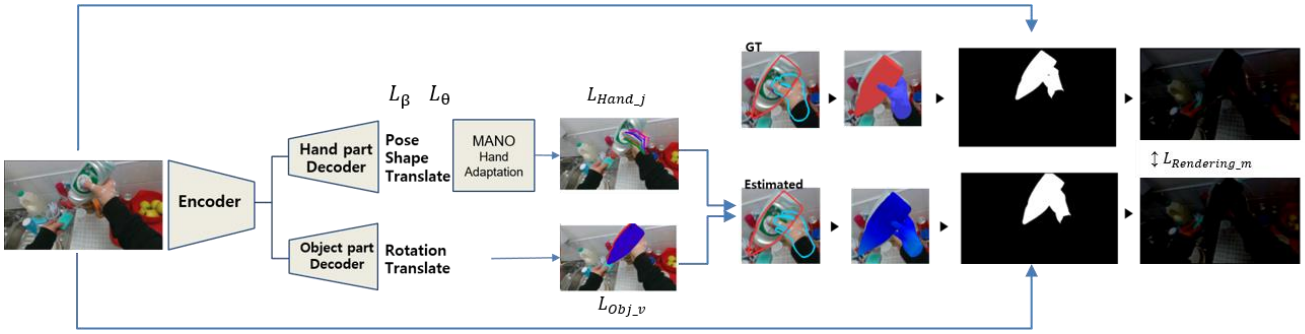

그림 1. RGB 영상에서 손-객체 3D 복원을 위한 네트워크 구조 및 렌더링 마스크 생성

오차를 학습에 반영하도록 하였다.

또한, 데이터셋의 일부만을 학습하여 Annotation 되지 않은 데이터에서 포즈를 추정하는 연구에서는 본 논문과 유사하게 입력 영상 정보를 비교하지만, optical flow 를 이용하여 영상 간의 포즈 변화를 추가로 학습하여 추정 정확도를 향상 시키는 연구가 있었다 [6]. 이와 달리 본 논문에서는 추가적인 학습을 거치지 않고 정확도를 향상시키고자 하였다. 이를 통해 비디오 혹은 일부 영상 데이터만을 이용해 손-객체 3D 포즈 추정이 가능함을 확인할 수 있었다. 본 논문은 다음과 같이 구성된다. 먼저 전체적으로 손-객체 포즈 추정을 위한 딥러닝 학습 과정 및 렌더링 마스크를 생성하는 방법에 대해 설명한다. 그리고 이를 손실함수에 어떻게 반영할 지 서술한다. 마지막으로 해당 방법을 First-Person Hand-Object 데이터셋에 적용한 실험 및 결과에 대해 확인하였다[5].

2. 전체 학습 과정 및 2D 렌더링 마스크 생성

본 논문에서는 RGB 입력 영상에서 손-객체 포즈를 추정하기 위한 딥러닝 기반의 학습 방법을 그림 1 과 같이 나타내었다(그림 1). 전체 학습 과정을 개괄적으로 살펴보면, 입력 영상으로부터 손과 객체의 6 DoF 에 각각 대해 학습을 진행하는 것을 알 수 있다. 그후에 손에 대해서는 3D hand model 에서 부자연스러운 모양 및 회전으로 학습되는 것을 방지하고자 하였으며, 복원된 결과를 2D 렌더링 마스크로 만들어 렌더링 결과를 비교하였다. 구체적인 학습 과정은 다음과 같다.

먼저 학습 네트워크는 선행연구와 같이 ResNet-18 을 사용하였으며, 이미지를 인코딩하고 마지막 레이어를 뽑아 2 개의 심층 레이어를 통해 뽑힌 특징점들로부터 손과 객체의 파라미터들을 계산해낸다[6]. 또한 손-객체의 전역 포즈를 추정하기 위해서 카메라 좌표계로 입력 영상의 이미지에 맞게 손-객체의 정확한 3D vertex 들을 투영하였다. 구체적으로 손-객체 이동을 추정하기 위해 카메라의 초점거리로 정규화한 깊이 오프셋을 선행연구와 같이 추정하여 이미지 상에서 손과 객체가 각각 어떻게 이동하였는지 추정하였다[6]. 이를 통해서 쉽게 손과 객체의 3 차원 이동을 도출해내고, 전역 회전에 대해서는 axis-angle 표현식을 이용하여 객체 중심 좌표계로 손과 객체의 회전을 예측하였다. 그리고 손의 포즈를 추정하기 위해서 선행연구들과 같이 추가적으로 MANO(hand Model with Articulated and Non-rigid defOrmations) 모델에 손의 이동 및 회전 정보를 반영한다[9]. 이를 통해 MANO 모델 연구에서 제공하는 저차원 손 포즈 공간에 대한 PCA(Principal Component Analysis)의 계수 값을 예측하여 MANO 모델의 포즈를 예측한다[9]. 그렇게 되면 사용되는 손 포즈 및 모양에 대해서도 부자연스러운 회전 등을 방지하도록 적응적으로 함께 학습하였다[6]. 그리고 알고 있는 손-객체의 3D 모델을 2D 로 렌더링시켜 마스크로 만든다. 만들어진 마스크를 입력 영상에 반영하여 입력 영상을 조작하여 Ground-truth 와 비교한다.

3. 렌더링 마스크 손실

Ground-truth 와 추정된 손-객체 포즈 정보를 각각 입력 영상에 렌더링한 모양을 마스크로 사용하여 비교함으로써 손-객체가 모두 복원되었을 때의 입력 영상 내에서 어떻게 보여질 수 있는지를 학습 과정에 반영하였다. 학습 과정에 반영하기 위해서 수식(1)과 같이 rendering mask loss term($L_{\text{Rendering mask}}$)을 구성하였다.

$$L_{\text{Rendering mask}} = || I \cdot M(R(I, V_{est})) - I \cdot M(R(I, V_{gt})) ||_1 \cdots (1)$$

여기서 I 는 입력 영상이고, V_{est}, V_{gt} 는 각각 추정된 vertex 와 ground-truth 에 해당하는 vertex 에 해당한다. 이때 R 의 과정은 Neural Renderer 를 사용하여 입력 영상에 손-객체의 3D 모델을 렌더링하는 것을 의미한다[10]. 그리고 M 으로 렌더링된 영역을 기준으로 마스크를 만들어 낸다. 이때, Ground-truth 와 추정된 손-객체 포즈 마스크를 각각 입력 영상에 적용하여, 입력 영상에 대해 손-객체 포즈 이외의 영역을 L1 norm 으로 하여 비교하여 loss function 에 반영하였다. 손-객체 포즈 이외의 영역을 비교함으로써 비교 대상의 영역을 보다 넓은

범위에 대해 파악하고자 하였다. 그리고 해당 Rendering mask loss 를 반영한 Total loss 는 다음과 같다.

$$L_{Total} = \lambda_T (L_{Obj\_v} + \lambda_J L_{Hand\_j} + \lambda_\beta L_\beta + \lambda_\theta L_\theta) + \lambda_R L_{Rendering\_m} \cdots (2)$$

전체적으로 손-객체에 대한 회전과 이동에 관해 6 가지의 파마리터들을 예측한다(L_{Obj_v}, L_{Hand_j}). 또한 선행연구에서 적용한것과 같이 적응적으로 25 가지 MANO 파라미터들을 함께 학습하게 된다[6]. 예측된 손과 객체의 모델을 참고하여 추정할 수 있도록 손의 경우에는 MANO 손 모델에 적용하여 결과적으로 MANO 에 의해 손의 vertex, joint 위치를 학습하게 된다. 또한 손에 대해서는 손의 부자연스러운 joint 회전(L_θ)와 모양의 변형(L_β)에 대해 L2 norm 의 형태로 패널티를 주어 학습에 반영한다[6]. 여기에 앞서 설명한 Rendering mask loss 를 추가하여 손-객체가 함께 복원 되었을 때의 모습을 학습에 적용할 수 있다.

4. 실험 및 결과

본 논문에서 제안한 Rendering mask loss 방법을 적용하기 위한 데이터셋으로 FPHand 데이터셋을 사용하였다[5]. 해당 dataset 에서 RGB 영상 내에서 손의 마디의 2D-3D keypoint data 와 객체의 2D-3D vertex 를 이용해 전역 pose 를 추정하도록 하였다. 더불어 선행연구와 같이 드물게 지도 학습을 검증하기 위해서 데이터셋의 0.0625%를 학습하도록 하였다[6]. 제안된 방법을 비교 평가 하기 위해, 크게 3 가지의 학습 형태에 대해 비교 평가를 진행하였다. (1) Rendering mask loss 가 반영되지 않은 것, (2) Rendering 시에 mask 내부의 영역을 loss 에 반영된 것, (3) Rendering mask 외부의 영역을 loss 로 반영된 것이다. 각 방법의 적용 결과의 차이를 확인하기 선행연구의 방법과 동일하게 손-객체의 3D 메쉬 오차를 비교 평가하였다[6]. 손의 복원 결과를 확인하기 위해서 21 개의 joint 에 대해서 end-point 오차(mm)의 평균을 계산하였다. 또한 객체에 대해서는 카메라 좌표계에서 vertex 거리(mm)의 평균을 계산하였다. 또한, 추가적인 학습없이 데이터셋을 드물게 학습하기 위한 비율을 0.0625%에서 1% 이상으로 높였을 경우에 대해서도 부차적으로 실험을 진행하였다.

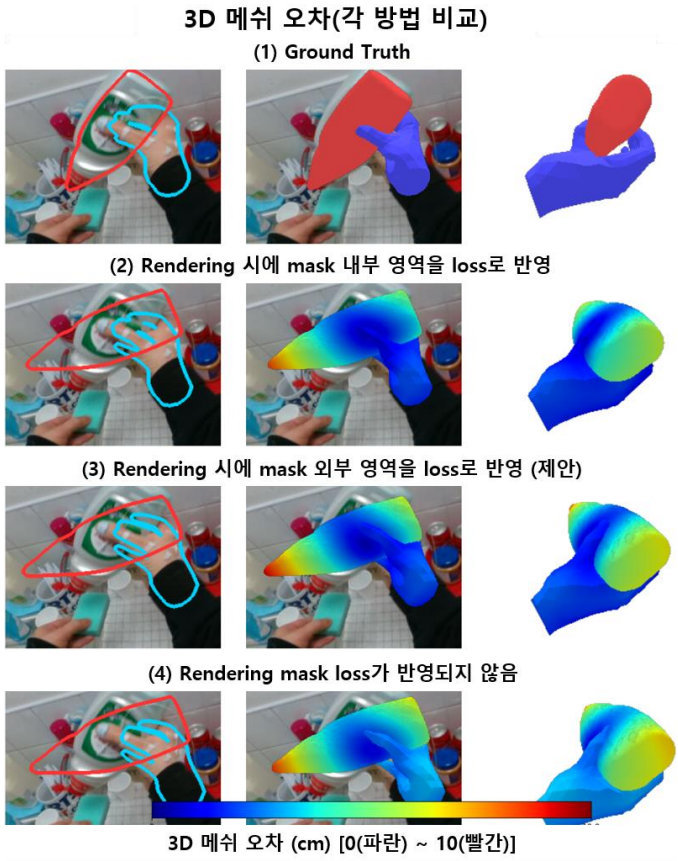

그림 2. 3D 메쉬 오차 결과(파란색일수록 오차가 적고, 빨간색일수록 오차가 높음) (Dataset 의 0.0625%학습): Ground-truth/ Rendering 시 mask 내부 영역 비교/ Rendering 시 mask 외부 영역 비교/ Rendering mask 미적용,

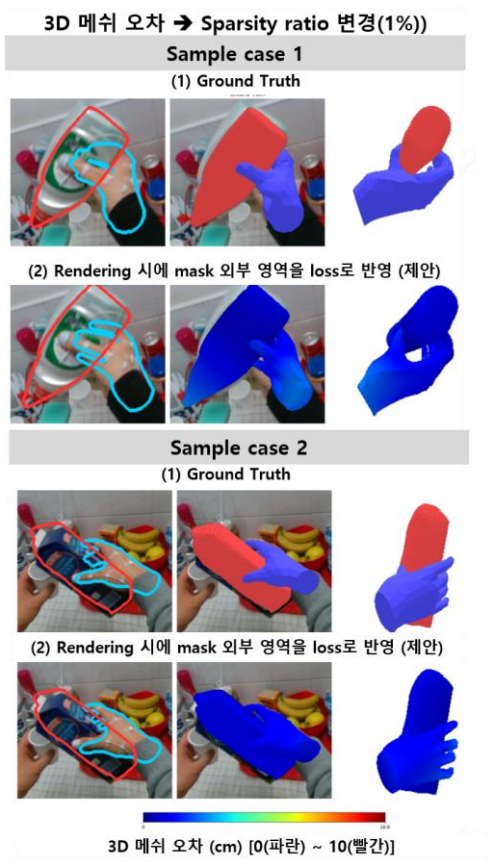

그림 3. 3D 메쉬 오차 결과 (파란색일수록 오차가 적고, 빨간색일수록 오차가 높음) (Dataset 의 1%학습) Sample case 2 가지: Ground-truth/ Rendering 시 mask 외부 영역 비교

각 방법을 비교 평가하였을 때, (그림 2)와 같이 나타난다. 또한, 총 1375 의 샘플 배치에 대해서 손-객체의 3D joint(손)와 vertex(객체) 오차를 비교하였을 때, 배치 영상 마다 오차의 격차가 미미하여 평균 값을 그대로 평가에 사용하기에 어려움이 있다. 때문에 각 배치 영상 마다 적용된 방법 중 가장 오차가 적거나 높은 경우를 확인하였다. 그 결과, 각 방법들 중에서 Rendering mask loss(mask 바깥쪽 영역 비교)를 적용한 경우가 가장 오차가 적음을 확인할 수 있었다 (표 1). 또한 데이터셋의 0.0625%에서 1%로 학습할 데이터 양을 늘려서 학습하였을 때에, Rendering mask 를 적용한 결과를 육안으로 확인하였을 때 오차가 거의 없음을 확인할 수 있었다(그림 3).

표 1. 3D vertex/joint 오차의 최소, 최대 빈도 비교

개수 / 각 방법	최소	최대
Mask-out(제안)	559(가장 우수)	320(가장 우수)
Mask-in	330	611
Without Mask	486	444

5. 결론

본 논문에서는 RGB 데이터셋의 일부만을 학습하여 Annotation 되지 않은 데이터로부터도 손-객체의 포즈를 정확하게 추정하여 3D 로 복원하기 위해, 본 논문에서 제안하는 방법인 Rendering mask loss 를 이용하여 정확도를 향상시키고자 하였다. 구체적으로 손-객체의 포즈가 추정된 결과를 알고 있는 3D 모델로 2D 마스크를 만들어 입력 영상에서 관련 영역을 Groud-truth 의 것과 비교하여 손실 함수에 반영하였다. 해당 방법을 실험에 적용하여 정확도를 향상을 확인 할 수 있었다. 데이터셋을 0.0625% 학습한 결과로 해당 방법의 효과를 검증하였고, 학습한 데이터셋의 양을 1%로 늘렸을 때를 추가적으로 실험하여 해당 방법을 적용하였을 때, 정확도가 높았음을 더불어 확인할 수 있었다. 이후의 연구에서는 야외 및 다양한 조명 환경에서의 강인한 복원이 가능할 수 있도록 하여 혼합현실 및 로봇비전 분야에서 확대 적용 될 수 있을 것이다.

감사의 글

이 논문은 2021 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2019R1A4A1029800).

참고문헌

[1] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. “Deepim: Deep iterative matching for 6d pose estimation.” In The European Conference on Computer Vision (ECCV), 2018.

[2] Kiru Park, Timothy Patten, and Markus Vincze. “Pix2Pose: Pixel-wise coordinate regression of objects for 6D pose estimation.” In The IEEE International Conference on Computer Vision (ICCV), 2019.

[3] Martin Sundermeyer, Marton Zoltan-Csaba, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. “Implicit 3D orientation learning for 6D object detection from RGB images.” In The European Conference on Computer Vision (ECCV), 2018.

[4] Javier Romero, Hedvig Kjellstrom, and Danica Kragic. “Hands in action: real-time 3D reconstruction of hands in interaction with objects.” 2010.

[5] Garcia-Hernando, Guillermo & Yuan, Shanxin & Baek, Seungryul & Kim, Tae-Kyun. “First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations.” (2017).

[6] Y. Hasson, B. Tekin, F. Bogo, I. Laptev, M. Pollefeys and C. Schmid, "Leveraging Photometric Consistency Over Time for Sparsely Supervised Hand-Object Reconstruction," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 568-577, doi: 10.1109/CVPR42600.2020.00065.

[7] Bardia Doosti, Shujon Naha, Majid Mirbagheri, David J. Crandall: “HOPE-Net: A Graph-Based Model for Hand-Object Pose Estimation”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6608-6617

[8] Javier Romero, Dimitrios Tzionas, and Michael J. Black. “Embodied hands: modeling and capturing hands and bodies together.” ACM Trans. Graph. 36, 6, Article 245 (November 2017), 17

[9] Boukhayma, Adnane, Rodrigo de Bem and P. Torr. “3D Hand Shape and Pose From Images in the Wild.” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019): 10835-10844.

[10] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural “3D mesh renderer.” In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018