손을 다루는 컴퓨터 비전 작업들을 위한 멀티 모달 합성 데이터 생성 방법

이창화*, 이선경*, 김동욱**, 정찬양***, 백승렬****
*UNIST 컴퓨터공학과
**영남대학교 컴퓨터 공학과
***아주대학교 사이버보안학과
****UNIST 인공지능대학원
changhwalee@unist.ac.kr, skwithu@unist.ac.kr, donguk.kim@yu.ac.kr,
cksdid4993@gmail.com, srbaek@unist.ac.kr

Generating A Synthetic Multimodal Dataset for Vision Tasks Involving Hands

Changhwa Lee*, Seongyeong Lee*, Donguk Kim**, Chanyang Jeong***,
Seungryul Baek****
*Dept. of Computer Science and Engineering, UNIST
**Dept. of Computer Science, Yeungnam University
***Dept. of Cyber Security, Ajou University
****Artificial Intelligence Graduate School, UNIST
changhwalee@unist.ac.kr, skwithu@unist.ac.kr, donguk.kim@yu.ac.kr,
cksdid4993@gmail.com, srbaek@unist.ac.kr

요 약

본 논문에서는 3D 메시 정보, RGB-D 손 자세 및 2D/3D 손/세그먼트 마스크를 포함하여 인간의 손과 관련된 다양한 컴퓨터 비전 작업에 사용할 수 있는 새로운 다중 모달 합성 벤치마크를 제안 하였다. 생성된 데이터셋은 기존의 대규모 데이터셋인 BigHand2.2M 데이터셋과 변형 가능한 3D 손 메시 (mesh) MANO 모델을 활용하여 다양한 손 포즈 변형을 다룬다. 첫째, 중복되는 손자세를 줄이기 위해 전략적으로 샘플링하는 방법을 이용하고 3D 메시 모델을 샘플링된 손에 피팅한다. 3D 메시의 모양 및 시점 파라미터를 탐색하여 인간 손 이미지의 자연스러운 가변성을 처리한다. 마지막으로, 다중 모달리티 데이터를 생성한다. 손 관절, 모양 및 관점의 데이터 공간을 기존 벤치마크의 데이터 공간과 비교한다. 이 과정을 통해 제안된 벤치마크가 이전 작업의 차이를 메우고 있음을 보여주고, 또한 네트워크 훈련 과정에서 제안된 데이터를 사용하여 RGB 기반 손 포즈 추정 실험을 하여 생성된 데이터가 양질의 질과 양을 가짐을 보여준다. 제안된 데이터가 RGB 기반 3D 손 포즈 추정 및 시맨틱 손 세그멘테이션과 같은 품질 좋은 큰 데이터셋이 부족하여 방해되었던 작업에 대한 발전을 가속화할 것으로 기대된다.

1. 서론

우리는 물건을 조작하는 것부터 다른 사람들과의 비언어적 의사 소통에 이르기까지 환경과의 상호 작용의 주요 수단으로 손을 사용한다. 이러한 손 제스처의 형태와 의미를 이해하는 것은 인간-컴퓨터 상호 작용, 컴퓨터 그래픽, 가상 및 증강 현실과 같은 다양한 애플리케이션에 필수적이다. 손을 이해하는 작업은 2D 손 세그멘테이션 [1], RGB 이미지로부터 3D 손 포즈 추정 [3] 및 깊이 이미지로부터 3D 손 포즈 추정 [2]와 같은 것들이 있으며, 3D 메시 재구성 [4] 및 손동작 인식 [5]에서도 손은 주요 물체로 인식되고 있다. 컴퓨터 비전 관점에서 손을 이해하는데 있어 가장 중요한 과제 중 하나는 정답이 있는 많은 양의 데이터에 액세스할 수 있는지 여부이다. 일반적으로 비선형적 혹은 고차원 매핑 학습이 필요하기 때문이다. 그러나 우리가 아는 한 대부분의 기존 데이터셋은 카메라 시점, 모양 및 손 자세 변형의 공간에서 제한이 된다. 대규모 데이터셋을 얻는 데 있어 많은 노동이 필요하기 때문에 이런 공간 제한이 생긴다. 본 논문에서는 손의 세 가지 주요 변형, 즉 모양, 관점 및 손 자세에 대해 보다 완전한 형태로 데이터를 모을 수 있는 데이터 수집 파이프라인을 제안한다. 본 논문의 기여는 아래와 같이 요약할 수 있다.

• **보다 완벽한 멀티 모달 데이터셋을 체계적으로 생성하기 위한 파이프라인 제안**: 해당 파이프라인에서는 변형 가능한 3D 손 메시 (mesh) 모델을 완벽하게 관절 공간이 있는 기존의 실제 깊이 벤치마크에 피팅한다. 모델의 해당 파라미터를 변경하여 모양과 시점을 생성한다. 마지막으로, 2D 프로젝션 방법을 사용하여 3D 메시 모델에서 다중 모달리티 데이터가 생성한다. 파이프라인은 Fig. 1에 나와 있다.

• **대규모 합성 데이터셋 제안** : 3D 메시 재구성과 같은 컴퓨터 비전 작업에서 2D / 3D 포즈 추정-커뮤니티에서 종종 완전한 데이터셋의 부족을 지적했었다. 이 데이터셋을 공개적으로 액세스 할 수 있도록 함으로써 간극을 메우고 여러 모달 연구를 장려하고자 한다.

• **데이터셋 품질에 대한 분석** : 제안 된 데이터셋을 세 가지 주요 손 도메인 변형 (모양, 카메라 시점 및 손 자세) 측면에서 기존 데이터셋과 비교했다. 또한 데이터셋의 완성도를 실험적으로 보여 주고자 했다.

2. 본론

[2] 논문은 10명에 대한 총 496개의 가능한 최대 범위의 손 자세 간의 변화 과정을 동영상 형태로 수집했다.

Figure 1. 핸드 벤치마크 생성을 위해 제안 된 파이프라인의 개략도. 먼저 BigHand2.2M 데이터베이스에서 고유 한 관절을 선택한 다음 MANO 손 모델을 골격에 맞추고 마지막으로 RGB-D, 골격, 세그멘테이션 마스크가 생성된다.

이러한 체계적인 데이터 수집 방법으로 얻어진 데이터셋은 손 관절 공간에서 완전하다고 생각될 수 있으나 데이터셋이 연속적인 동영상으로 수집되었기 때문에 여러 중복된 자세가 캡처되었다. 모든 중복 관절을 활용하는 것은 심층 신경망 훈련에 비효율적이다. 따라서 우리는 총 957,032 개의 학습 데이터에서 관절 공간의 중복성을 줄인 다음 MANO 모델을 샘플링 된 손 관절에 맞추는 방법을 활용하도록 동기 부여되었다.

구별가능한 손 관절 선택. 21개 관절의 3D 좌표 값 x, y 및 z로 구성된 63 차원 원시 손 관절 벡터는 주요한 세 손의 변형에 영향을 받는다. 즉 관점, 모양 및 손 자세 만 추출하기 위해 25 차원 각도 특징, 각 손가락에 대해 5개 각도를 추출 할 것을 제안한다. K-평균 알고리즘은 이러한 각 특징 벡터 위에 적용되었고 K=100,000의 클러스터 크기는 실험적으로 설정되었으며, 32 개의 말단 관절 범위 사이에 496 연속 전환을 포함하는 각 공간은 이 숫자로 충분히 커버된다. 이 과정을 통해 K개의 관절 $z=\{z_j\}, j\in[1,K]$이 고유한 손 관절로 선택되었다.

MANO [6] 모델 피팅. MANO 모델을 이전 단계에서 얻은 고유한 손 관절들에 피팅하는 단계이다. 각 고유 손 관절 z_i에 대한 MANO 모델의 모양 $s=\{s_j\}, j\in[1,10]$, 카메라 $c=\{c_j\}, j\in[1,8]$ 및 관절 $a=\{a_j\}, j\in[1,45]$ 파라미터를 다음 방정식을 해결하여 구한다.

$$(s_i,c_i,a_i)=\mathrm{argmin}_{(s,c,a)} \| f(V(s,c,a))-z_i \|_2^2+R(V(s,c,a))$$

여기서 f는 MANO 모델에서 나온 메시의 꼭지점을 손 관절로 매핑하는 함수이며, R은 메시의 삼각 면을 부드럽게 해 주는 라플라시안 함수이다.

피팅된 손 모델의 다양한 모양 및 관점 파라미터. 더욱 완성된 손 포즈 공간을 확보하고자 우리는 MANO 메시의 모양과 관점 파라미터를 조작해야 함을 제안한다. 피팅된 손 메시는 x 및 y 축에 관한 각도로 구성된 회전 행렬을 사용해 회전할 수 있고, 10개의 모양 파라미터를 조작해 메시 모양을 변경할 수도 있다.

이기종 작업을 위한 데이터 생성. 관절, 모양 및 관점 공간을 비교적 완전에서 메시를 생성 한 뒤 해당 메시를 사용하여 3D 손 관절 회귀 분석기 및 랜더링 엔진의 도움으로 다양한 형태의 데이터를 생성한다. 6 가지 형태(깊이맵, RGB맵, 2D 및 3D 손관절, 세그멘테이션 마스크, 3차원 메시)의 결과 예가 Fig. 2에 표시되었다.

3. 실험

이 섹션에서는 제안된 데이터셋의 품질을 분석한다. 먼저 PCA 프로젝션을 사용하여 관절 공간 측면에서 데이터셋를 공개적으로 사용가능한 다른 데이터(STB [7], RHD [3] 및 SH [8])과 비교한다. 둘째, 텍스처 모델을 언급 된 데이터셋의 모델과 비교한다. 마지막으로 제안된 데이터셋의 이점을 정량적으로 평

가하기 위해 RGB 이미지를 데이터셋을 사용 또는 사용하지 않고 손 자세 추정기를 학습한다.

3.1. 관련 데이터셋 비교

제안된 데이터셋의 모양과 시점 공간은 기존 RGB 기반 손 포즈 데이터베이스 (STB [7], RHD [3] 및 SH [8])에 비해 더 완벽하다. 제안된 데이터셋의 텍스쳐와 관절 공간이 Fig. 3(왼쪽 아래)에서 다른 데이터셋과 비교된다.

텍스쳐 비교. Fig. 3(왼쪽 위)의 상단에서 생성 된 텍스쳐를 다른 데이터셋의 텍스쳐와 질적으로 비교한다. 텍스쳐가 다른 합성 데이터셋(RHD[3] 및 SH[8])에 비해 실제 데이터셋(STB[7])에 약간 더 가깝다는 것을 알 수 있다.

손 관절 공간 비교. Fig. 3(왼쪽 위)의 하단에는 4 개의 비교 된 데이터셋의 관절 공간이 표시된다. 예상대로 데이터의 관절 공간은 관련 데이터셋에 비해 밀도가 더 높다. 이것은 STB에 간단한 손짓 계산이 포함되어있는 반면 RHD와 SH에는 합성 동작 포즈가 거의 포함되어 있지 않기 때문이라 할 수 있다.

Figure 2. 이기종 작업을 위한 데이터 생성 예시.

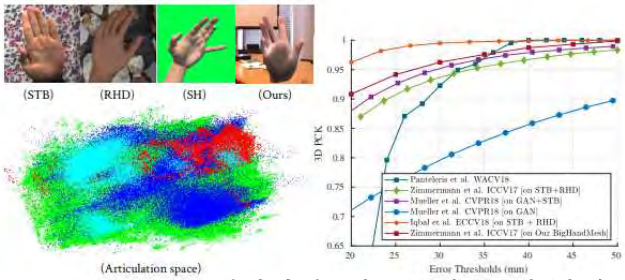

Figure 3. 관련 데이터셋 비교 (왼쪽 위/아래), RGB-to-3D실험 (오른쪽)

3.2. RGB-to-3D 실험

이 섹션에서는 제안 된 데이터셋의 이점을 정량적으로 평가한다. 이를 위해 제안된 데이터셋를 사용 또는 사용하지 않고 RGB 기반 손 포즈 추정기 [3]를 훈련하고 얻은 결과를 STB[8] 데이터셋의 최신 결과와 비교한다.

구현 디테일. RGB 기반 손 포즈 추정기의 경우 코드가 온라인에서 공개적으로 제공되는 [3]에 제시된 아키텍처를 사용했다. 전체 네트워크는 [3] 논문의 웹 사이트에서 제공하는 사전 훈련 된 가중치를 사용하여 아키텍처를 초기화했다. 각 세대에서 MANO 모델의 임의의 관점 및 모양 파라미터와 결합된 K = 100,000 관절을 사용하여 아키텍처를 미세 조정한다. 학습률이 10^{-3}이고 기본 β 파라미터로 Adam 최적화 알고리즘을 사용했다. 기존 Big Hands 2.2M에서 조밀하게 수집하고 [2]에서 데이터베이스의 1/20 배로 샘플링된 일부만 사용했으며 일부 데이터만을 사용했어도 약간의 정확도 저하가 관찰되었다. 또한 5 명의 사람만 공개적으로 사용 가능하므로 실제로 총 1,000,0000개 이미지로 공간을 늘렸다. 하위 집합을 사용해도 다른 연구보다 월등한 성능을 얻었다.

결과 분석. Fig. 3(오른쪽)에서 훈련된 손 관절 추정기 [3]를 기존 4 개의 최신 RGB 기반 3D 골격 추정 알고리즘과 비교하였다 [3, 7, 8, 9]. 각각의 방식들은 약간 다른 데이터셋을 활용하여 훈련되며 Fig. 3 (오른쪽)의 범례에서 이를 설명한다. 예를 들어, 'Zimmermann et al. ICCV17 [on STB + RHD]'는 Zimmermann와 Brox. [3]의 방법을 사용하여 STB 및 RHD 데이터셋에 대해 학습되었다. [7]의 경우 모델 피팅 방법에 관한 논문이므로 훈련 데이터셋이 필요하지 않다. Fig. 3(오른쪽)에서 우리는 테스트 데이터가 실제 이미지일 때도 RGB 기반 손 포즈 추정기를 훈련하기 위해 합성 데이터셋를 사용하는 이점을 관찰하고자 하였다. 제안된 데이터셋으로 [3]의 네트워크를 훈련하면 STB 및 RHD를 사용한 훈련에 비해 성능이 크게 향상된다. Iqbal et al. [8] 의 접근 방식은 평가 된 모든 접근 방식 중에서 가장 성능이 좋은 접근 방식이다. 그러나 네트워크 아키텍처 내에서 정교한 RGB-D 재구성 모듈을 사용하므로 데이터셋 사용과 관련하여 결론을 내리기가 어렵다. Mueller et al. [9]도 CycleGAN을 사용하여 SH 데이터셋를 강화하여 합성적으로 생성된 데이터를 사용하는 것을 제안했다. Mueller et al. [9]의 작업에 비해 우리의 훈련된 손 포즈 추정기는 더 높은 정확도를 달성하였다. 이 접근 방식은 심층적인 ResNet 아키텍처를 사용한다는 점을 고려할 때 더 강력한 아키텍처의 데이터셋를 사용하면 성능이 향상 될 여지가 있다. 또한 Mueller et al. [9] 의 작업은 실제 데이터셋 STB가 학습 단계에 포함되지 않은 경우 상당한 정확도 저하

를 보여준다. 우리가 제안한 데이터셋은 이러한 도움 없이도 더 높은 성능을 제공했다.

4. 결론

3D메시, RGB-D 영상과 2D / 3D 손 정답, 세그멘테이션 마스크를 포함한 새로운 멀티모달 합성 데이터 벤치마크가 제안되었다. 생성된 데이터셋은 기존의 대규모 깊이 손 포즈 데이터셋 BigHand2.2M과 변형 가능한 3D 손 메시 모델 MANO를 활용하여 다양한 손 자세의 변형을 커버할 수 있다. 우리는 RGB 기반 3D 손 자세 추정 및 시맨틱 손 세그멘테이션과 같이 이전에 품질이 큰 벤치마크가 부족하여 방해를 받았던 손을 활용한 다양한 컴퓨터 비전 작업에 대한 연구의 발전적인 진행을 위해 제안된 데이터셋이 유용하게 사용될 것이라 예상한다.

사사. 이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임. (No. 2020-0-01336 인공지능대학원 지원(울산과학기술원), No. 2020-0-00537 5G 기반 저지연 디바이스-엣지클라우드 인터랙션 기술 개발)

참고문헌

[1] A. U. Khan and A. Borji. Analysis of hand segmentation in the wild. In CVPR, 2018.

[2] S. Yuan, Q. Ye, B. Stenger, S. Jain, and T.-K. Kim. Bighand 2.2M benchmark: hand pose dataset and state of the art analysis. In CVPR, 2017.

[3] C. Zimmermann and T. Brox. Learning to estimate 3D hand pose from single RGB images. In ICCV, 2017.

[4] S. Baek, K. I. Kim, and T-K. Kim. Pushing the envelope for RGB-based dense 3D hand pose estimation via neural rendering. In CVPR, 2019.

[5] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim. First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In CVPR, 2018.

[6] J. Romero, D. Tzionas, and M. J. Black. Embodied hands:Modeling and capturing hands and bodies together. InSIG-GRAPH Asia, 2017.

[7] P. Panteleris, I. Oikonomidis, A. Argyros. Using a Single RGB Frame for Real Time 3D Hand Pose Estimation in the Wild, WACV 2018.

[8] U. Iqbal, P. Molchanov, T. Breuel, J. Gall, J. Kautz, Hand pose estimation via latent 2.5D heatmap regression, ECCV 2018.

[9] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, C. Theobalt, GANerated Hands for Real-Time 3D Hand Tracking from Monocular RGB, CVPR 2018.