특집 04

영상 학습 기반 손 포즈 추정 최신 연구 동향 분석

김대환 · 김용완 · 이기석 (한국전자통신연구원), 조동식 (울산대학교)

목 차

1. 서 론

영상 기반의 손 포즈 추정 기술이란 컬러(RGB) 또는 깊이(Depth) 영상 데이터에서 손과 손가락의 방향 또는 위치를 검출하고 추정하는 연구 분야를 의미한다. 이는 번거로운 장비 없이 시스템과의 자유로운 상호작용을 할 수 있다는 측면에서 인간-컴퓨터 상호작용(Human Computer Interface, HCI), 가상 및 증강현실 (VR/AR) 및 제스처 인식 시스템 (gesture recognition system) 분야에서 핵심적인 역할을 하고 있다. 특히 가상 및 증강현실 분야에서는 상당히 현실적인 가상 손 움직임을 가능하게 하여 사용자 경험을 대폭 향상할 수 있는 아주 중요한 연구이다.

몇 년 전부터 깊이 정보를 제공하는 센서들[1-2]의 대중화에 힘입어 상용화의 가능성을 높이기 시작한 손 포즈 추정 연구는 현재 깊은 신경망(Deep Neural Network, DNN) 알고리즘의 비약적인 발전으로 인하여 컬러 영상만으로도 3차원 손 포즈를 추정할 수 있는 수준까지 발전해 오고 있다[3].

하지만 높은 손가락 관절 자유도 (high-degree of freedom articulation), 심각한 가림 현상 (severe self/external occlusion), 빠른 손의 움직임(fast hand movement), 낮은 입력 영상 해상도 (low input image resolution) 및 불충분한 데이터셋(insufficient dataset)들과 같이 여전히 해결해야만 하는 많은 난제가 남아 있다.

※ This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIP) (2018-0-00999, Medical Digital Twin Generation and 3D Simulation Technology for Prediction and Computer Aided Diagnosis of Musculoskeletal Disease)

본 논문에서는 영상 기반 손 포즈 추정 연구에 대한 빠른 이해와 접근을 위해 연구 동향, 최신 연구 주제 및 데이터셋에 관해서 설명하고자 한다.

2. 연구 동향

영상 기반 손 포즈 추정에 관한 연구 동향은 크게 3가지의 관점에 따라 변화되어 왔다. 첫째, 입력 영상의 센서 형태, 둘째, 깊은 신경망 알고리즘의 발전, 마지막으로 양질의 데이터셋의 확보이다. 이장에는 3가지 관점에서의 연구 동향 변화를 알아보기로 한다.

2.1 센서 사용 형태

2000년대 후반까지의 손 포즈 추정에 관한 연구는 컬러 장갑을 활용하거나 아주 잘 고안된 형태의 손 영역 영상에서의 손 검출 (hand detection) 또는 손 포스처 인식(hand posture recognition) 정도가 대부분이었다[4]. 하지만 2010년에 깊이 센서의 출현과 대중화로 인하여 많은 연구자의 관심을 끌며 깊이 영상 기반의 3차원 손 포즈 추정 연구가 본격적인 궤도에 올라섰다. 깊이 영상은 3차원 공간 좌표 값을 제공하기에 쉽게 손을 분리할 수 있거나 손의 입체적인 형상들도 추정할 수 있는 장점이 존재한다[5,6]. 현재에도 깊이 영상의 간편하고 정확한 3차원 공간 좌표 값 제공으로 인하여 3차원 포즈 추정(3D hand pose estimation)이나 손 메시 복원 (3D mesh reconstruction) 연구가 활발히 진행 중이다. 하지만 최근에는 깊은 신경망 알고리즘의 혁신적인 발전으로 컬러 센서를 활용하는 연구가 다시 활기를 찾고 있다[3].

2.2 깊은 신경망 알고리즘 활용

최근 깊은 신경망 알고리즘(DNN)의 발전은 인공지능(artificial intelligence), 컴퓨터 비전(computer vision) 및 기계 학습(machine learning) 분야에서의 혁신적인 변화를 가져왔다. 특히 사람 포즈 추정 (human pose estimation), 영상 분류 (image classification), 물체 검출 (object detection), 음성 인식 (voice recognition) 및 행동 인식(action recognition) 기술들에 널리 사용되고 있다.

깊은 신경망 알고리즘들에는 합성곱 신경망 (Convolutional Neural Network, CNN)[7], 순환 신경망 (Recurrent Neural Network, RNN)[8], 자동 인코더(Auto-Encoder, AE)[9], 생성적 적대 신경망 (Generative Adversarial Network, GAN) [10] 등이 있다. 이러한 알고리즘들은 고차원 (high dimensionality)으로 구성되는 데이터 구성 공간 (data configuration space)을 표현할 수 있다는 사실이 직간접적으로 입증되었다.

따라서 현재 최신 손 포즈 연구에서는 깊은 신경망 알고리즘 및 변형들을 이용하여 최소 20 자유도(DOF)의 움직임을 가지고 있는 3차원 손 포즈를 추정하거나 손 메시를 복원하기 위한 연구들이 활발하게 진행되고 있다.

2.3 데이터셋

데이터셋은 알고리즘들 간 사이를 비교 평가(benchmark)하거나 알고리즘을 학습하기 위해서 필수적이다. 일반적으로 데이터셋은 컴퓨터 그래픽 기술을 사용하여 합성 생성하거나 깊이 또는 컬러 센서를 이용하여 만들어진다. 데이터셋 제작의 가장 중요한 부분은 연구에 목적에 따라 실제 결과 데이터를 사전에 만들어 내는 것이다.

이 과정을 라벨링(labeling) 또는 주석(annotation)이라고 부른다. 현재까지는 사람이 직접 수동으로 표시(marking)를 해왔지만, 최근에는 이를 자동화하는 이슈들도 제기되고 있다 [11-13].

영상 기반 손 포즈 추정 데이터셋들은 연구 목적과 주제에 따라 크게 2가지 관점에서 만들어지고 있다. 데이터 형태와 손의 상황이다. 데이터 형태는 기존과 같이 합성 영상(synthetic image)과 실 획득 영상(real captured image)으로 나눌 수 있고, 손의 상황은 연구의 방향성에 따라 한 손, 양 손, 손-물체 및 손-손 상호작용으로 나눌수 있다. 추후 4장에서 공용으로 사용되고 있는 손 포즈 데이터셋에 대해 자세히 설명하도록 한다.

3. 최신 연구 주제

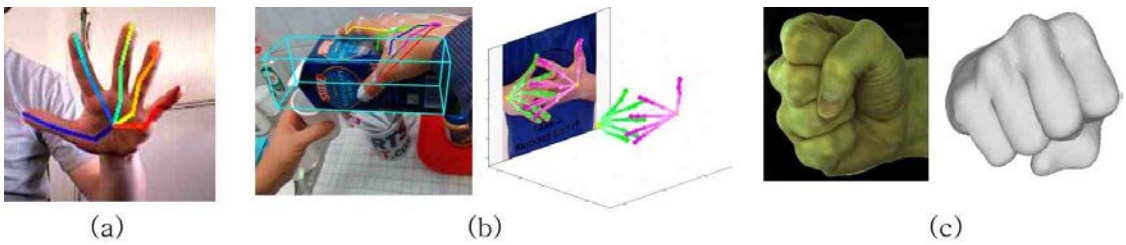
영상 학습 기반의 손 포즈 추정에 관한 최신 연구 주제들은 (1) 3차원 포즈 추정, (2) 손-물체 및 손-손 상호작용 시의 3차원 손 포즈 추정 및 (3) 3차원 손 메시 복원 등으로 분류할 수 있다. 이는 최근 4년간 (2017년 ~ 2020년)의 컴퓨터 비전, 기계 학습 및 가상/증강현실 분야에서 최상위 계층(Top-tier)으로 평가되는 학회들(CVPR, ICCV, WACV, SIGGRAPH, BMVC)에 출판된 논문들 위주로 정리 요약하였다. 그림 1은 최근 3가지 연구 주제의 상황별 결과 예제들을 보여주고 있다.

3.1 3차원 포즈 추정

3차원 손 포즈 추정은 깊이 또는 컬러 영상에서 손의 3차원 방향과 손가락 관절 (finger joints)들의 3차원 위치들을 추정하는 것이다. 이는 지극히 전통적인 연구 주제이지만, 최근 깊은 신경망 (DNN) 알고리즘 덕분에 컬러 영상에서 3차원 포즈 정보를 추정할 수 있기에 재조명을 받으며 이슈화되고 있다.

깊이 영상 기반의 3차원 포즈 추정 연구는 2D 깊이 맵(depth map)[14-19], 포인트 클라우드(point cloud)[20-21] 또는 3D 복셀 표현(voxel representation)[22-23]과 같은 특징들을 기반으로 다양한 깊은 신경망을 활용하는 방법들이 있다. 기본적으로 깊이 영상에서 3차원 위치 정보를 추출할 수 있기에 3차원 정보를 회귀(regression)하는 과정 없이 직접적인 깊은 신경망 학습을 통해 3차원 포즈를 추정할 수 있는 장점이 있다. 특정한 형태의 새로운 개념보다는 합성곱 신경망(CNN)을 기반으로 한 다양한 변형 알고리즘들이 제안되었다.

컬러 영상 기반의 3차원 포즈 추정 연구에는 (1) 관절 점수 지도(score map)나 열 지도(heat



(그림 1) 손 포즈 추정 상황별 결과 예제 영상들. (a) 손 포즈 추정[24], (b) 손-물체[41] 및 손-손[59] 상호작용 시의 포즈 추정, (c) 손 메시 복원 [51].

map)를 직접 회귀(regression)하는 방법, (2) 컬러 영상과 3차원 포즈 사이의 잠재 공간(latent space)을 연동(mapping)하는 방법 그리고 (3) 잠재 공간을 구분(disentangling)하는 방법들이 있다.

첫째, 깊은 신경망 기반의 회귀 방법들은 관절의 2차원 점수 지도[24-25] 또는 열 지도[26-27]와 3차원 관절의 위치 사이의 관계를 포즈 회귀 신경망(pose regression network)을 통하여 학습하는 것이다. 일반적으로 2차원 관절 검출 알고리즘[28]을 사용하여 점수 지도나 열 지도를 획득한 뒤, 회귀 신경망을 학습하는 방식을 사용한다. 이에 더하여 3차원 포즈 추정의 정확성을 높이기 위해, 손의 기하학적(geometry) 또는 운동학적(kinematic) 구조를 반영하는 3차원 모델을 사용하여 포즈를 예측하는 연구들[29-31]이 있다.

둘째, 컬러 영상과 3차원 포즈 사이의 관계성을 내포할 수 있는 잠재 공간을 학습하고 생성하여 3차원 포즈를 추정하는 방법이다. 보통 변형 자동 인코더 (Variational Auto Encoder, VAE)[32-34]나 생성적 적대 신경망 (GAN)[35-36]을 사용하여 잠재 공간을 생성한다. 이는 이종의 모달리티(modality) 사이의 관계성을 연동시켜 쉽게 컬러 영상에서 3차원 포즈를 쉽게 추정할 수 있게 해 준다.

셋째, 컬러 영상을 다양한 요소들(포즈, 배경, 카메라 뷰 등)이 결합되어 존재하는 잠재 공간이라 가정하고, 이들을 서로 분리하여 3차원 포즈 요소를 획득하는 방법이다. 잠재 공간을 분리할 수 있는 표현 형태(disentangled representation)[37-38]를 학습하여 서로 다른 모달리티를 분리하도록 한다. 이는 복잡하게 얽힌 하나의 잠재 공간을 생성하기보다는 3차원 포즈 연동 공간만을 분리할 수 있게 해 준다.

3.2 손-물체 및 손-손 상호작용 시의 3차원 손 포즈 추정

손-물체 및 손-손 상호작용 시의 3차원 손 포즈 추정이란 손에 물체를 들고 있거나 두 손이 서로 겹쳐지는 상황에서 3차원 관절의 위치들을 찾아내는 것이다. 물체에 의해서 손이 심각하게 가려지거나 비슷한 텍스처 형태를 띄는 양 손이 서로 밀접하게 결합되어 분리가 힘든 상황에 놓인 경우를 말한다. 실제 상호작용 상황에서 빈번히 일어나는 상황이기에 매우 실용적인 연구 주제이다.

손-물체 상호작용 상황에서의 손 포즈 추정 방법은 손과 물체를 서로 인식하여 분리한 후 손 포즈를 추정하는 방식[39-42]을 사용한다. 일반적으로 2개의 서로 다른 형태의 깊은 신경망을 사용하여 손-물체를 분리하고 손 포즈를 추정하도록 한다. 최근에는 가려진 손 영역의 부분을 생성적 적대 신경망(GAN)을 이용하여 생성한 뒤 포즈를 추정하는 방법[43]이 제안되었다. 이는 물체가 가려진 상황에서도 안정적인 포즈 추정 성능을 보여주었다.

서로 겹쳐진 양 손의 포즈를 동시에 추정하는 연구는 현재 아주 초보적인 단계에 머물러 있다. 최근 실시간으로 양 손을 동시에 추적하기 위해 손 내부의 상대적인 깊이 거리와 손 사이의 거리 지도의 정보를 활용한 알고리즘[44]이 제안되었다. 3차원 손 모델[45]의 매칭(matching) 과정을 통하여 포즈 매개변수를 추정하였다. 또한 물리적인 변형이 가능한 손 모델[46]을 이용하여 심각하여 가려진 상황에서 정밀한 추적이 가능한 연구도 제안되었다. 하지만 여전히 빠른 움직임이 있을 때나 복잡하게 가려지면 추적에 실패하기에 앞으로도 많은 발전이 필요하다.

3.3 손 메시 복원

최근 손 포즈 추정 성능 발전에 따라 영상으로부터 직접 손의 3차원 메시를 복원(reconstruction)하거나 모델링(modeling) 하고자 하는 연구들이 진행 중이다. 열지도 특징[47]이나 컬러 또는 깊이 영상을 입력[48-50]으로 다양한 깊은 신경망 구조를 이용하여 3차원 손 모델의 파라미터를 추정하여 메시를 복원하도록 한다. 최근 기존 3차원 손 모델인 MANO[45]의 낮은 해상도와 손가락 사이의 물리적 제한사항을 반영하지 못하는 단점들을 보안한 새로운 3차원 손 모델[51]이 제안되었다. 이는 깊은 신경망 프레임워크를 적용이 쉽게 가능하여 물리적 제한사항을 반영한 채 해상도 높은 메시 복원 결과를 보여 준다.

4. 데이터셋

데이터셋은 영상 학습 기반의 손 포즈 추정 연구를 진행할 때 아주 중요한 역할을 한다. 일반적으로 알고리즘 학습 시 데이터셋의 양(quantity)과 질(quality)에 따라 손 포즈 추정 결과 성능이 결정된다. 현재의 손 포즈 추정 학습 데이터셋(dateset)들은 크게 실데이터(real-world data)와 합성데이터(synthetic data)의 2가지 형태로 나눌 수 있다. 실데이터는 실제 손을 컬러 또는 깊이 센서들을 이용하여 촬영한 영상들이다. 합성데이터는 3차원 그래픽 모델링 기술을 활용하여 실데이터와 비슷한 형태의 손 모델을 생성하여 얻어진 영상들이다. 실데이터는 실질적인 학습이 가능하지만 부정확한 주석(annotation)

〈표 1〉 대표적인 손 포즈 추정 데이터셋

입력형태	데이터셋	형태	시점	해상도(WxH)	관절 수
깊이 (Depth)	ICVL[52]	Real	3rd	320 x 240	16
	NYU[53]	Real	3rd	640 x 480	36
	MSRA[54]	Real	3rd	320 x 240	21
	HandNet[55]	Real	3rd	320 x 240	6
	BigHand2.2M[56]	Real	3rd/Ego	640 x 480	21
	SynHand5M[57]	Synthetic	3rd	320 x 240	22
컬러 (RGB)	FreiHAND[58]	Real	3rd	224 x 224	21
	InterHand2.2M[59]	Real	3rd	512 x 334	21
컬러+깊이 (RGB+Depth)	Dexter1[60]	Real	3rd	320 x 240	6
	Dexter+Object[61]	Real	3rd	640 x 320	5
	RHD[24]	Synthetic	3rd	320 x 320	21
	STB[62]	Real	3rd	640 x 480	21
	EgoDexter[63]	Real	Ego	640 x 480	5
	SynthHands[63]	Synthetic	Ego	640 x 480	21
	FPHA[12]	Real	Ego	1920x1080(c) 640x480(d)	21
	HO3D[13]	Real	3rd	640x480(c,d)	15
	ContactPose[64]	Real	3rd	1920x1080(c) 512x424(d)	21

문제가 발생하는 반면, 합성데이터는 정확한 주석(annotation)을 쉽게 얻을 수 있지만, 실제 손 영상과는 다소 차이가 있을 수 있다. 표 1은 현재 손 포즈 추정 알고리즘 학습 및 테스트를 위해 가장 많이 사용하고 있는 대표적인 데이터셋들이다.

몇 년 동안의 손 포즈 추정 알고리즘의 발전에 따라 데이터셋의 종류도 변화되어 제작되었다. 예전에는 깊이 영상 데이터 위주[52-57]로 구성되었다면, 최근에는 컬러 영상을 활용할 수 있는 딥러닝 기술의 발전에 따라 컬러 영상이 포함된 데이터셋[12-13,24,60-64]이 제작되었다. 또한 가상/증강현실 영역에서 활용하기 위한 1인칭 시점에서 촬영된 데이터셋들 또한 제작되고 있다.

최근 손-물체 및 손-손 상호작용 상황에서의 손 포즈 연구를 위한 데이터셋들이 제작되고 있다. 물체를 들고 있거나[13,61,64] 양 손을 자유롭게 교차[59]하는 등의 현실적인 상황들을 반영하는 형태의 데이터셋들이 제작되어 활용되고 있다. 컨택트 포즈 (ContactPose)[64]는 물체를 쥐고 있을 상황에서의 손 포즈, 물체 포즈 및 컬러-깊이 (RGBD) 영상을 제공한다. 또한 인터핸드2.6M (InterHand2.6M)[59]은 손 사이의 심각한 상호작용 상황에서의 컬러 영상을 제공한다.

5. 결 론

본 논문은 영상 학습 기반의 손 포즈 추정 연구를 위한 최근 동향을 분석하였다. 손 포즈 추정 연구는 현재 매우 인기 있는 연구 분야이며, 최근 깊은 신경망 알고리즘의 등장으로 인해 추정 성능의 상당한 발전을 이루고 있다. 특히 손-물체 및 손-손 상호작용 시의 3차원 손 포즈 추정과 손 메시 복원 연구들은 인간-컴퓨터 상호작용 및 가상 및 증강현실 분야의 대중화를 이끌 수 있는 핵심 기술이 될 것이다.

하지만 여전히 높은 관절 자유도, 심각한 가림 현상 및 복잡한 상호작용 상황 등의 해결되지 않은 문제들이 존재한다. 따라서 이들을 해결하기 위한 알고리즘의 발전이 필요하다. 또한 이를 뒷받침하기 위한 주석 처리가 잘 반영된 데이터셋 제작이 필요한 실정이다. 앞으로의 미래 활용 가치가 높은 연구인 만큼 많은 발전이 이루어질 것으로 예상한다.

참 고 문 헌

[1] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake, “Efficient human pose estimation from single depth images,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pp. 2821-2840, 2012.

[2] https://www.intelrealsense.com

[3] T. Chatzis, A. Stergioulas, D. Konstantinidis, K. Dimitropoulos, and P. Daras, “A comprehensive study on deep learning-based 3D hand pose estimation methods,” Applied Sciences, vol. 10, no 19:6850, 2020.

[4] R.Wang and J. Popovic, “Real-time hand-tracking with a color glove,” ACM Transactions on Graphics, vol. 28, no. 3, pp. 1-8. 2009.

[5] D. Tang, T. Yu, and T. Kim, “Real-time articulated hand pose estimation using semi-supervised transductive regression forests,” In Proceedings of the IEEE International Conference on Computer

Vision, pp. 3224-3231, 2013.
[6] L. Ge, H. Liang, J. Yuan, and D. Thalmann, “Robust 3d hand pose estimation in single depth images: from single-view CNN to multi-view CNNs,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3393-3601, 2016.
[7] L. Yann, and Y. Bengio, “Convolutional Networks for images, speech, and time series,” In The Handbook of Brain Theory and Neural Networks, MIT Press: Cambridge, MA, USA, vol. 3361, no. 10, 1995.
[8] s. Hochreiter, and J. Schmidhuber, “Long short-term memory,” Neural Computing, vol 9. no. 8, pp. 1735-1780, 1997.
[9] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” Journal of Maching Learning, vol. 11, no. 10, pp. 3371-3408, 2010.
[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, “Generative adversarial nets,” In Proceedings of the Neural Information Processing Systems (NIPS), pp. 2672-2680, 2014.
[11] M. Oberweger, G. Riegler, P Wohlhart, and V. Lepetit. “Efficiently creating 3D training data for fine hand pose estimation,” In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4957-4965, 2016.
[12] G. Garcia-Hernando, S. Yuan, S. Baek, and T. Kim. “First-person hand action benchmark with RGB-D videos and 3D hand pose annotations, In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 409-419, 2018.
[13] S. Hampali, M. Rad, M. Oberweger, amd V. Lepetit, “Honnotate: A method for 3D annotation of hand and object poses,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3193-3203, 2020.
[14] Y. Zhou, J. Lu, K. Du, X. Lin, Y. Sun, and X. Ma, “Hbe: Hand branch ensemble network for real-time 3d hand pose estimation,” In Proceedings of the European Conference on Computer Vision (ECCV), pp. 501-516, 2018.
[15] K. Du, X. Lin, Y. Sun, and X. Ma, “Crossinfonet: Multi-task information sharing based hand pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9896-9905, 2019.
[16] P. Ren, H. Sun, Q. Qi, J. Wang, and W. Huang, “SRN: Stacked regression network for real-time 3D hand pose estimation,” In Proceedings of the British Machine Vision Conference (BMVC), p. 112, 2019.
[17] C. Wan, T. Probst, L. Gool, and A. Yao, “Self-supervised 3d hand pose estimation through training by fitting,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10853-10862, 2019.
[18] F. Xiong, B. Zhang, Y. Xiao, Z. Cao, Y. Yu, J. Zhou, and J. Yuan, J, “A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image, In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 793-802. 2019.

[19] L. Fang, X. Liu, L. Liu, H. Xu, and W. Kang, “JGR-P2O: joint graph reasoning based pixel-to-offset prediction network for 3D hand pose estimation from a single depth image,” In Proceedings of the European Conference on Computer Vision (ECCV), pp.120-137, 2020.

[20] S. Li and D. Lee, “Point-to-pose voting based hand pose estimation using residual permutation equivariant layer,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11927-11936, 2019.

[21] Y. Chen, Z. Tu, L. Ge, D. Zhang, R. Chen, and J. Yuan, “So-handnet: Self-organizing network for 3d hand pose estimation with semi-supervised learning,” In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 6961-6970, 2019.

[22] L. Ge, H. Liang, J. Yuan, and D. Thalmann, “3D convolutional neural networks for efficient and robust hand pose estimation from single depth images,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1991-2000, 2017.

[23] J. Malik, E. Abdelziz, A. Elhayek, S. Shimada, S. Ali, V. Golyanik, C. Theobalt, and D. Stricker, “HandVoxNet: deep voxel-based network for 3D hand shape and pose estimation from a single depth map,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7113-7122, 2020.

[24] C. Zimmermann, amd T. Brox, “Learning to estimate 3d hand pose from single rgb images,” In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 4903-4911, 2017.

[25] A. Boukhayma, R. Bem, and P. Torr, “3d hand shape and pose from images in the wild,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10843-10852, 2019.

[26] U. Iqbal, P. Molchanov, J. T. Gall, and J. Kautz, “Hand pose estimation via latent 2.5d heatmap regression,” In Proceedings of the European Conference on Computer Vision (ECCV), pp. 118-134, 2018.

[27] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, D. Sridhar, D. Casas, and C. Theobalt, “Ganerated hands for real-time 3d hand tracking from monocular rgb,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 49-59, 2018.

[28] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and T. Sheikh, “OpenPose: realtime multi-person 2D pose estimation using part affinity fields,” IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 43, no. 1, pp. 172-186, 2019.

[29] X. Zhang, Q. Li, H. Mo, W. Zhang, and W. Zheng, “End-to-end hand mesh recovery from a monocular rgb image,” In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2354-2364, 2019.

[30] S. Baek, K. Kim, and T. Kim, “Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1067-1076, 2019.

[31] A. Spurr, U. Iqbal, P. Molchanov, O. Hilliges, and J. Kautx, “Weakly supervised 3D hand pose estimation via bio-

mechanical constraints," In Proceedings of the European Conference on Computer Vision (ECCV), 2020.

[32] T. Theodoridis, T. Chatzis, V. Solachidis, and K. Dimitropoulos, "Cross-modal variational alignment of latent spaces," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW), pp. 960-969, 2020.

[33] A. Spurr, J. Song, S. Park, and O. Hilliges, "Cross-modal deep variational hand pose estimation," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 89-98, 2018.

[34] L. Yang, S. Li, D. Lee, and A. Yao. "Aligning latent spaces for 3d hand pose estimation, In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2335-2343, 2019.

[35] C. Wan, T. Probst, L. Gool, and A. Yao, "Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 680-689, 2017.

[36] B. Zhu, C. Ngo, J. Chen, and Y. Hao, "R2gan: Cross-modal recipe retrieval with generative adversarial network," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11477-11486, 2019.

[37] L. Yang, and A. Yao, "Disentangling latent hands for image synthesis and pose estimation, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9877-9886, 2019.

[38] J. Gu, Z. Wang, W. Ouyang, J. Li, and L. Zhuo, "3d hand pose estimation with disentangled cross-modal latent space," In Proceedings of the IEEE Winter Conference on Applications on Computer Vision (WACV), pp. 391-400, 2020.

[39] H. Zhang, Z. Bo, J. Yong, and F. Xu, "Interaction fusion: Real-time reconstruction of hand poses and deformable objects in hand-object interactions, ACM Transactions on Graphics, vol. 38, no. 4, 2019.

[40] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. Black, I. Laptev, and C. Schmid, "Learning joint reconstruction of hands and manipulated objects," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11807-11816, 2019.

[41] B. Doosti, S. Naha, M. Mirbagheri, and D. Crandall, "HOPE-Net: A graph-based model for hand-object pose estimation,", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6608-6617, 2020.

[42] B. Tekin, F. Bogo, and M. Pollefeys, "H+O: Unified egocentric recognition of 3D hand-object poses and interactions," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4511-4520, 2019.

[43] S. Baek, K. Kim, and T. Kim, "Weakly-supervised domain adaptation via gan and mesh model for estimating 3D hand poses interacting objects," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6120-6131, 2020.

[44] J. Wang, F. Mueller, F. Bernard, S. Sorli, O. Sotnychenko, N. Qian, M. Otaduy, D. Casas, and C. Theobalt, "RGB2Hands:

real-time tacking of 3D hand interactions from monocular RGB video," ACM Transactions on Graphics (TOG), vol. 39, no. 6, 2020.

[45] J. Romero, D. Tzionas, and M. Black, "Embodied hands: modeling and capturing hands and bodies together," ACM Transactions on Graphics (TOG), vol 36, no. 6, 2017.

[46] B. Smith, C. Wu, P. Peluse, Y. Sheikh, J. Hodgins, and T. Shiratori, "Constraining dense hand surface tracking with elasticity," ACM Transactions on Graphics (TOG), vol 39. no. 6. 2020.

[47] X. Zhang, Q. Li, H. Mo, W. Zhang, and W. Zheng, "End-to-end hand mesh recovery from a monocular RGB image," In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2354-2364, 2019.

[48] L. Yang, J. Li, W. Xu, Y. Diao, and C. Lu, "BiHand: recovering hand mesh with multi-stage bisected hourglass networks," In Proceedings of the British Machine Vision Conference (BMVC), 2020.

[49] D. Kulon, R. Guler, I. Kokkinos, M. Bronstein, and S. Zafeiriou, "Weakly-supervised mesh-convolutional hand reconstruction in the wild," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4990-5000, 2020.

[50] C. Wan, T. Probst, L. Gool, and A. Yao, "Dual grid net: hand mesh vertex regression from single depth maps," In Proceedings of the European Conference on Computer Vision (ECCV), pp.442-459, 2020.

[51] G. Moon, T. Shiratori, and K. Lee, "DeepHandMesh: a weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling," In Proceedings of the European Conference on Computer Vision (ECCV), pp.440-455, 2020.

[52] D. Tang, H. Jin, A. Tejani, T. Kim, "Latent regression forest: Structured estimation of 3d articulated hand posture," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3786–3793, 2014.

[53] J. Tompson, M. Stein, Y. Lecun, and K. Perlin, "Real-time continuous pose recovery of human hands using convolutional networks,", ACM Transactions on Graphics (ToG), vol. 33, pp. 1-10, 2014.

[54] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun, "Cascaded hand pose regression," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 824-832, 2015.

[55] A. Wetzler, R. Slossberg, and R. Kimmel, "Rule of thumb: Deep derotation for improved fingertip detection," arXiv:1507.05726, 2015.

[56] S. Yuan, Q. Ye, B. Stenger, S. Jain, amd T. Kim, "Bighand2.2m benchmark: Hand pose dataset and state of the art analysis," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4866-4874. 2017.

[57] J. Malik, A. Elhayek, F. Nunnari, K. Varanasi, K. Tamaddon, A. Heloir, and D. Stricker, "Deephps: End-to-end estimation of 3d hand pose and shape by learning from synthetic depth," In Proceedings of the International Conference on 3D Vision (3DV), pp. 110-119, 2018.

[58] C. Ziimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox, “Freihand:A dataset formarerless capture of hand pose and shape from single rgb images,” In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 813-822, 2019.

[59] G. Moon, S. Yu, H. Wen, T. Shiratori, and K. Lee, “InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a Single RGB Image,” In Proceedings of the European Conference on Computer Vision (ECCV), pp. 548-564, 2020.

[60] S. Sridhar, A. Oulasvirta, and C. Theobalt, “Interactive markerless articulated hand motion tracking using RGB and depth data,” In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2456-2463, 2013.

[61] S. Sridhar, F. Mueller, and M. Zollhöfer, D. Casas, A. Oulasvirta, and C. Theobalt, “Real-time joint tracking of a hand manipulating an object from rgb-d input,” In Proceedings of the European Conference on Computer Vision (ECCV), pp. 294-310, 2016.

[62] J. Zhang, J. Jiao, M. Chen, L. Qu, X. Xu, and Q. Yang, “A hand pose tracking benchmark from stereo matching,” In Proceedings of the IEEE International Conference on Image Processing (ICIP), pp. 982-986, 2017.

[63] F. Meller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt, “Real-time hand tracking under occlusion from an egocentric rgb-d sensor,” In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 1284-1293, 2017.

[64] S. Brahmbhatt, C. Tang, C. Twigg, C. Kemp, and J. Hays, “ContactPose: a dataset of grasps with object contact and hand pose,” In Proceedings of the European Conference on Computer Vision (ECCV), pp. 361-378, 2020.

저 자 약 력

김 대 환

이메일 : daehwank@etri.re.kr

- 2002년 동국대학교 컴퓨터멀티미디어공학과(학사)
- 2004년 POSTECH 컴퓨터공학과(석사)
- 2011년 POSTECH 컴퓨터공학과(박사)
- 2012년~현재 한국전자통신연구원 선임연구원 재직중
- 관심분야 : 컴퓨터비전, 머신러닝, 인공지능, HCI, 가상현실, 증강현실

김 용 완

이메일 : ywkim@etri.re.kr

- 1996년 인하대학교 전자공학과 (학사)
- 1998년 광주과학기술원 정보통신공학과 (석사)
- 2014년 한국과학기술원 전산학과 (박사)
- 1998년~현재 한국전자통신연구원 책임연구원 재직 중
- 관심분야 : 가상현실, 증강현실, 햅틱 인터랙션, 3D 인터페이스, 오감 인터랙션

이 기 석

이메일 : mvr_lks@etri.re.kr

• 1999년 성균관대학교 제어계측공학과 (학사)
• 2001년 성균관대학교 전기전자 및 컴퓨터공학과 (석사)
• 2001년-2019년 한국전자통신연구원 연구원
• 2020년~현재 한국전자통신연구원 VR/AR콘텐츠연구실장
• 관심분야 : 컴퓨터 그래픽스, VR/AR/XR

조 동 식

이메일 : dongsikjo@ulsan.ac.kr

• 2017년 고려대학교 컴퓨터학 (박사)
• 2004년~2018년 전자통신연구원(ETRI) 선임연구원
• 2018년~2020년 원광대학교 디지털콘텐츠공학과 교수
• 2018 가상현실 증강현실의 미래 저자
• 2020 MDPI Electronics Guest Editors (LifeXR)
• 2021년~현재 울산대학교 IT융합전공 교수
• 관심분야 : 홀로그램, VR/AR/MR, 컴퓨터그래픽스, HCI