손 위치, 자세, 동작의 통합 심층 학습

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

Joint Deep Learning of Hand Locations, Poses and Gestures

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

요 약

본 논문에서는 사람의 손에 관한 개별적으로 분리되어 진행되고 있는 손 위치 추정, 손 자세 추정, 손 동작 인식 작업을 통합하는 Faster-RCNN기반의 프레임워크를 제안하였다. 제안된 프레임워크에서는 RGB 동영상을 입력으로 하여, 먼저 손 위치에 대한 박스를 생성하고, 생성된 박스 정보를 기반으로 손 자세와 동작을 인식하도록 한다. 손 위치, 손 자세, 손 동작에 대한 정답을 동시에 모두 가지는 데이터셋이 존재하지 않기 때문에 Egohands, FPHA 데이터를 동시에 효과적으로 사용하는 방안을 제안하였으며 제안된 프레임워크를 FPHA데이터에 평가하였다., 손 위치 추정 정확도는 mAP 90.3을 기록했고, 손 동작 인식은 FPHA의 정답을 사용한 정확도에 근접한 70.6%를 기록하였다.

1. 서론

사람의 손에 관련된 컴퓨터 비전 분야의 연구들(손 위치, 손 자세, 손 동작을 이해하고자 하는 연구들)은 상호 고려 없이 개별적으로 수행되고 있다. 그림 1에 손 위치, 손 자세, 손 동작을 이해하는 3가지 개별 작업에서 주로 다루는 데이터 형태와 정답 형태를 예시로 보여주고 있다. 예를 들어, 손 자세를 추정하는 연구에서는 손 위치는 정답으로 주어진 상황을 가정하고, 손 자세를 수행하는 모델의 정확도만을 올리고자 데이터 셋을 구축하고 방법론을 개발하고 있으며, 손 위치 추정 연구의 경우 위치 추정 이후의 작업에 대한 성능은 신경 쓰지 않고, 손 위치 추정의 정확도만을 높이는 연구를 수행하였다. 손 동작의 이해에서도 손 자세/손 위치를 추정하는 정확도가 최종단에 영향을 미침에도 불구하고 이 부분에 대해서는 최대한 정답을 사용하는 방식으로 앞

(a) 손 위치 추정 데이터셋 예시

(b) 손 자세 추정 데이터셋 예시

(c) 손 동작 인식 데이터 예시

그림 1. 손 관련 연구 (손 위치, 손 자세, 손 동작 이해) 데이터셋 형태 및 정답 예시

단에서 전달되는 에러 영향은 최소화한 채 동작을 잘 이해하는 부분에만 집중하였다. 각 작업에 대한 데이터 셋 구축과 방법론 개선 측면에서는 이러한 연구 방법이 효율적일 수 있으나, 실제 상황에서의 손 이해와는 괴리가 있는 연구 방법이라 보여진다.

본 연구에서는 이전 연구에서 고려하지 못한 손에 대한 통합적 이해를 통해 실제 환경과의 괴리를 최소화함과 동시에 성능향상 가능성을 시험하고자 한다. 손 위치, 자세, 동작을 개별적으로 이해하는 딥러닝 기반 프레임워크들을 통합하여 위치, 자세, 동작을 통합적으로 이해할 수 있는 단일 프레임워크를 구축하고 여러 개의 작업을 동시에 학습했을 때 상호 에러 전달을 고려한 최적의 학습 기법을 제안하고자 한다.

2. 관련 논문 분석

손 연구의 중요성. 손은 우리가 일상생활에서 외부 세계와 상호작용하는 가장 원초적인 도구로써, 머리에 달린 (Head-mounted) 카메라가 촬영하는 1인칭 시점의 연속된 Single RGB 동영상에서 손의 위치, 자세, 동작을 동시적으로 이해하는 것은 잠재적으로 일상생활에서의 편의를 증대시킬 수 있는 매우 중요한 연구주제이다. 기술의 발전은 손의 위치, 자세, 동작을 이해하는 연구가 상호 고려없이 서로 다른 작업들을 정답이 주어졌다고 가정한 채 실험실 환경 수준으로 실험이 진행되고 있으며, 기술의 발전은 초기 단계에 있는 반면, 응용적인 면에서 마이크로소프트사의 홀로렌즈 (Hololens), 페이스북이 인수한 오큘러스사의 리프트 (Rift), 구글 글래스, 삼성 기어 VR 등 다수의 스마트 글래스에서 채택되는 1인칭 시점에서 손의 위치, 자세, 동작의 이해는 중요한 문제로 인식되고 있다.

국내외 연구 동향. 2010년대 초반부터 마이크로소프트사의 Kinect 센서 등을 통해 얻은 single 뎁스 영상에서 3차원 모델 미팅 및 랜덤 포레스트 등의 머신러닝 알고리즘을 통해 손 자세 추정을 하는 알고리즘이 개발되었다 [1, 2]. 딥 러닝 이후, 3차원 모델 피팅 기법보다는 딥 러닝을 통한 학습 기반의 손 위치, 자세 및 동작 인식 알고리즘들이 제안되었으며 [3], 좀 더 실용적인 성능을 얻기 위해 큰 크기의 데이터베이스를 구축하는 연구도 진행이 되었다 [4, 5, 6]. 특히, 자세에 대한 데이터를 수집할 때 어려운 점은, 주요 자세에 대한 3차원 좌표에 대한 정답을 얻어야 한다는 점인데, 시간과 노력이 많이 필요하다. 손 자세의 주요 변화 축인 손 모양, 손가락 구성, 카메라 시점에 대해 magnetic 센서를 착용하고 센서 위치에서 손 관절 위치를 자동으로 추정할 수 있는 inverse kinematic 방법을 제안한 연구가 뎁스 영상과 정확한 정답을 자동으로 수집할 수 있어서 좋았으나, RGB영상에서는 magnetic센서가 보여서 RGB영상에 대해서는 일반화되지 못한 방법이며, Kinect센서와 magnetic센서 사이에 동기화가 이루어지지 않아 동작이 빠른 손에 대해서는 정답이 정확하지 않을 수 있다.

최근 연구 동향. 최근에는 좀 더 활용이 간단한 single RGB 카메라를 기반으로 하는 손 자세 인식 알고리즘이 개발 되었다 [7, 8]. Depth 입력에 비해 RGB 입력은 noise 등이 없다는 장점이 있으나, 깊이감 정보가 입력에 전혀 없는 상태에서 3차원 손 자세 정보를 추정해야 한다는 점에서 비선형적 사상 (non-linear mapping)이 필요하여 쉽지 않은 문제이다. 더욱이, 데이터 수집 관점에서는 깊이감에 대한 정답 수집에 어려움이 있다. 해결을 위해, calibrated 된 depth, RGB 카메라로 동시에 손을 촬영하여 3차원 좌표를 depth 카메라로부터 manual하게 읽어 더해주거나 오프라인에서의 3차원 모델 피팅을 여러 번 수행하여 가장 잘된 depth 값을 읽거나 하는 등의 방법론이 제안되었으나, 아직까지 더 체계적인 데이터 수집 방법론에 대한 연구는 진행 중인 상태에 있다. 그래픽스 엔진을 통해 정확한 정답과 합성된 영상을 동시에 생성해내는 방법론이 제안되었으나 [9], 합성된 데이터로 학습된 딥러닝 모델은 실제 testing영상에 대해 성능이 제한되는 한계가 있다. 결론적으로 데이터 수집에서의 어려움으로 인해 현재까지는 합성된 데이터 및 소량의 실제 데이터를 섞어서 딥러닝 하는 형태의 학습 방법을 택하고 있다. 1인칭 시점에서 사람의 손동작을 이해하려는 방법들이 최근 제안되었으나 [6, 10], 손 위치 및 자세 추정을 직접적으로 수행하지는 않고, 동영상에서 제공되는 RGBD 시퀀스 자체를 활용하거나 [10], 아니면 손 자세에 대한 정답을 이용한 연구가 있었다 [6]. 반면, 손 자세를 이용할 경우 RGBD 시퀀스 자체를 이용하는 것보다는 성능이 향상됨을 확인하였다 [6]. 또한 손 위치를 찾는 연구는 아직 부족한 편이며 [21], 딥러닝을 활용한 최근의 손 위치 찾는 연구에서는 세그멘테이션 기법이 활용되었다 [11].

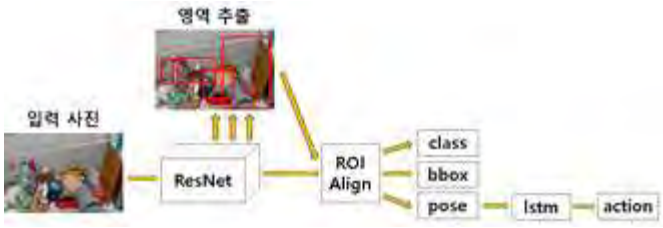


그림 2. 손 위치, 자세, 동작 통합 인식 파이프라인

3. 방법

3.1 통합 알고리즘.

손 자세를 담은 임의의 크기의 영상을 인풋으로 해서, 먼저 손 검출을 수행한뒤, 검출된 손 박스에 대해 손 자세를 복원해내고, 복원된 손 자세를 활용하여 손 동작을 인식하는 프레임워크를 그림 2 과 같이 제안하였다. 손 검출 모듈은 Faster R-CNN [14] 기반의 네트워크로 구성되었으며 손 자세 복원 알고리즘은 Mask-RCNN [12]와 유사한 형태로 구성되었다. 마지막으로 손 동작 인식 알고리즘은 LSTM 알고리즘 기반으로 구성되었다. 이후 섹션들에서 각 알고리즘의 특성에 대해 세분해 설명하고자 한다.

손 검출 알고리즘. RGB 영상에서 다중 객체를 검출하기 위한 알고리즘인 Faster R-CNN [14] 구조를 이용하여 손을 검출한다. Faster R-CNN은 CNN 모델을 이용해 특징을 추출하며, 추출된 특징들을 이용해 물체가 존재한다고 추정되는 영역을 추출한다. 추출된 영역의 특징을 이용하여 영역의 클래스 정보를 찾고 Bounding Box의 좌표를 조정한다. 해당 Faster R-CNN을 이용하여 손에 대한 Bounding Box를 추정할 수 있는 알고리즘을 재 학습하였다.

손 자세 복원 알고리즘. Mask-RCNN [12] 에서 영역분할 손실함수를 추가하여 영역분할을 객체탐지와 동시에 수행한 것과 같이, 수식 1과 같은 손 자세 손실함수를 추가하여 검출된 손에 대해 자세 추정을 추가로 해주는 방식을 사용한다.

$$L_{pos} = \|f(x) - y\|_2^2$$

이 때, x는 인풋 영상, f는 손 추정 네트워크, y는 21x3 크기의 손 자세 정답 행렬을 의미한다. 영역 추출 네트워크가 추출한 영역의 특징을 CNN layer에 통과시키고 히트맵 방식을 이용해 21개의 손 관절에 대해 가장 확률이 높은 좌표를 찾는다.

손 동작 인식 알고리즘. 손동작 영상은 시계열 데이터이므로 이를 주로 처리하는 LSTM 구조를 사용하여 손 동작을 인식하였다. LSTM은 hidden state와 cell state를 이용해 순서가 있는 데이터를 순차적으로 처리하는 알고리즘이다. 각 프레임에서 구한 21개의 손 관절을 최소값을 0, 최대값을 1로 정규화하여 LSTM layer의 입력값으로 사용하고 수식 2와 같은 cross entropy 손실함수를 추가하여 손동작 클래스를 학습하였다.

$$L_{act} = -\sum_{i=1}^{N} p(x_i) \log q(x_i)$$

3.2 학습 방법.

손 위치, 자세, 동작을 동시에 학습할만한 형식의 정답을 가지는 데이터베이스가 부재하기 때문에, 본 논문에서는 손 위치에 대한 정답을 가지는 Egohands로부터 손에 대한 자세와 동작을 동시에 가지는 FPHA 데이터로 '왼손', '오른손' 위치 정답을 먼저 '전이' 하도록 하였으며, 그렇게 모든 자세에 대한 정답을 가지게 된 FPHA 데이터에 대해 학습을 시도하여 최종 결과를 얻었다.

FPHA 데이터 [6] 소개. FPHA(First-Person Hand Action) 데이터셋은 3가지 시나리오 주제에서 6명을 대상으로 45개 서로 다른 손동작 카테고리에서 수행되는 1,175가지 손동작 비디오를 포함한다. 총 105,459개의 RGB-D 프레임들에 정확한 손 자세와 손동작 카테고리 형태로 annotation이 구성된다.

EgoHands 데이터 [13] 소개. EgoHands 데이터는 총 4개의 손이 상호작용하는 데이터로 구성되어 있으며 총 4800개의 프레임에 대해 정답을 가지는 데이터이다. '왼손', '오른손'에 대한 클래스 정답과 각 손에 대한 위치를 박스형태로 제공해주고 있다.

데이터 전이를 통한 통합 구조 학습. Ego hands 데이터의 왼손, 오른손 박스 형태 정답 정보를 Faster-RCNN으로 학습하여, 손 위치 추정기를 학습하고 FPHA데이터에 추정함으로써 FPHA데이터의 손 위치 정답을 만들어 주었고, 그렇게 얻은 FPHA의 모든 정답으로 통합구조를 학습하였다.

4. 실험

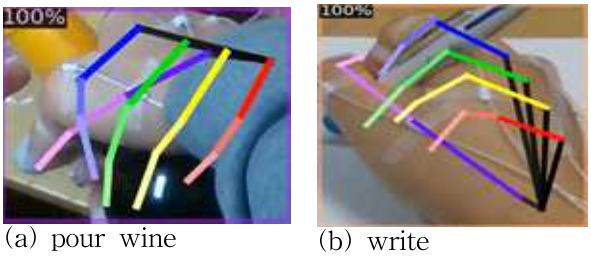
4.1 손 검출 알고리즘 성능 측정.

FPHA의 Bounding Box 정답의 부재로, 우리는 EgoHand 데이터에 대해 손 검출 알고리즘을 먼저 학습하였다. 표 1와 같이 IOU 0.5 기준 mAP 90 이상의 성능으로 손을 검출한다.

방법	mAP(IOU:0.5)	왼손	오른손
Faster-RCNN	90.3	90.1	90.5

표 1. 손 검출 알고리즘 성능표

그림 3. 손 검출 알고리즘의 결과 예시



(a) pour wine (b) write

그림 4. 손 동작 인식모델의 결과 예시

4.2 손 동작 인식 알고리즘 성능 측정. FPHA 데이터에 대한 손동작 인식 알고리즘의 성능은 표2 와 같다. FPHA에서는 손 관절의 위치가 주어졌다고 전제하였지만, 우리는 테스트 시 정답을 사용하지 않고 손 관절 위치를 추정하였으며 70%의 성능을 유지하였다.

방법	정답사용	정확도 (%)
3D 손 자세 정답 + LSTM	O	78.73
2D 손 자세 정답 + LSTM	O	77.32
제안한 모델	X	70.69

표 2. 손 동작 인식 알고리즘 성능표

그림 4 에서는 우리가 제안한 모델의 정성적인 성능을 보여준다. 손의 가려진 부분이 있음에도 관절의 위치를 비교적 잘 추정하며 이를 통해 손 동작 인식을 수행한다.

5. 결론

손에 대한 컴퓨터 비전 연구는 AR/VR향 스마트 글래스의 보급화로 인해 실용성이 증대될 것이라 예측된다. 손은 외부 세계와 상호작용하는 가장 원초적인 도구로써, 손의 위치, 자세, 동작을 동시적으로 동시에 이해하는 것은 AR/VR향 스마트 글래스, 사람-컴퓨터 상호작용 (HCI), 텔레오퍼레이션, 원격 수술 등의 미래 응용 기술을 선점하는 효과를 가져올 것이다. 앞으로 손 동작 인식 알고리즘에 물체 검출을 추가하여 물체와 어떤 상호작용하는지에 대해 연구할 계획이다.

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] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei, D. Freedman, P. Kohli, E. Krupka, A. Fitzgibbon, S. Izadi, Accurate, Robust, and Flexible Real-time Hand Tracking, CHI 2015.

[2] X. Sun, Y. Wei, S. Liang, X. Tang, J. Sun, Cascaded Hand Pose Regression, CVPR 2015.

[3] Q. Ye, S. Yuan, T-K. Kim, Spatial attention deep net with partial PSO for hierarchical hybrid hand pose estimation, ECCV 2016.

[4] S. Yuan, Q. Ye, T-K. Kim, BigHand2.2M benchmark: Hand pose dataset and state-of-the-art analysis, CVPR 2017.

[5] S. Baek, K. I. Kim, T-K. Kim, Augmented skeleton space transfer for depth-based hand pose estimation, CVPR 2018.

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

[7] 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.

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

[9] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, C. Schmid, Learning joint reconstruction of hands and manipulated objects, CVPR 2019.

[10] M. Ma, H. Fan, K. M. Kitani, Going deeper into first person activity recognition, CVPR 2016.

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

[12] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, ICCV 2017.

[13] S. Bambach, S. Lee, D. J. Crandall, C. Yu, Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions, ICCV 2015.

[14] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, NIPS 2015.