모바일 Deep Residual Network 을 이용한 뎁스 영상 기반 1 인칭 시점 VR 손동작 인식

박혜민, 박나현, 오지헌, 이철우, 최형우, 김태성
경희대학교 전자정보대학 생체의공학과
{hmp9669, nhpark, dhwlgjs3, tskim}@khu.ac.kr
{lcwoo0604, chesai}@naver.com

Depth Image based Egocentric 3D Hand Pose Recognition for VR Using Mobile Deep Residual Network

Hye Min Park, Na Hyeon Park, Ji Heon Oh, Cheol Woo Lee, Hyoung Woo Choi, Tae-Seong Kim
Dept. of Biomedical Engineering, College of Electronics and Information,
Kyung Hee University, Republic of Korea

요 약

가상현실(Virtual Reality, VR), 증강현실(Augmented Reality, AR), 혼합현실(Mixed Reality, MR) 분야에 유용한 인간 컴퓨터 인터페이스 기술은 필수적이다. 특히 휴먼 손동작 인식 기술은 직관적인 상호작용을 가능하게 하여, 다양한 분야에서 편리한 컨트롤러로 사용할 수 있다. 본 연구에서는 뎁스 영상 기반의 1 인칭 시점 손동작 인식을 위하여 손동작 데이터베이스 생성 시스템을 구축하여, 손동작 인식기 학습에 필요한 1 인칭(Egocentric View Point) 데이터베이스를 촬영하여 제작한다. 그리고 모바일 Head Mounted Device(HMD) VR 을 위한 뎁스 영상 기반 1 인칭 시점 손동작 인식(Hand Pose Recognition, HPR) 딥러닝 Deep Residual Network 를 구현한다. 최종적으로, 안드로이드 모바일 디바이스에 학습된 Residual Network Regressor 를 이식하고 모바일 VR 에 실시간 손동작 인식 시스템을 구동하여, 모바일 VR 상 실시간 3D 손동작 인식을 가상 물체와의 상호작용을 통하여 확인 한다.

1. 서론

인간 컴퓨터 상호작용(Human Computer Interaction, HCI) 관련 기술들은 미래를 주도하는 연구의 주축으로 각광받고 있다. 이 분야에서 중요한 기술 분야 중 하나는 3D 손 동작 인식으로, 별도의 디바이스 컨트롤러를 사용하는 것 보다 가상의 공간에서 편하고 자연스러운 사용자 친화적 인터페이스를 제공할 수 있다[1].

최근 뎁스 센서와 딥 러닝[2] 기술의 발전으로 단일 뎁스 이미지를 이용한 3D 손 동작 인식 기술 개발에 큰 향상이 있었다. 특히 다량의 뎁스 이미지에서 자동으로 특징을 찾는 데 유용하고, 직접 학습할 수 있는 딥 러닝 Convolutional Neural Network(CNN)을 이용한 방법이 좋은 성능을 보였다[3]. 또한 고성능 GPU 를 사용한 행렬 연산을 통해 딥 러닝 모델의 학습 및 실행 속도는 실용 가능한 수준까지 발전했다[4,5].

현재 대부분의 손 동작 인식 시스템은 3 인칭 시점에 중점을 두었다. 1 인칭 손 동작 인식에 사용 가능한 오픈 데이터베이스는 대표적으로 First-Person Hand Action(FHAD) Dataset[6] 와 BigHand2.2M Dataset[7]을 합쳐 구성한 Hands2017 Dataset[8]가 있고, Synthetic 모델로 만든 SynthHands Dataset[9]가 있다. 그러나 실제 휴먼 손으로 촬영하지 않은 데이터베이스라는 한계, 정제된 손 동작 정보가 아닌 물체가 포함된 뎁스 영상, 그리고 다양한 제스처가 아닌 한정된 손동작만 존재하여 범용 사용이 제한적이다. 따라서 본 연구에서는 실제 휴먼 손을 모델로 하고 손동작을 직접 촬영하여 모바일 디바이스 VR 상에서 1 인칭 손동작 인식에 적합한 데이터베이스를 생성하였다.

또한 기존의 전통적인 컴퓨터 비전 알고리즘 기반 손 동작 인식이 아닌 모바일 AP 에 적합하고, 정확도 개선을 위한 Residual Learning[10] 기반의 ResNet Regressor[3]를 구현하여 1 인칭 시점의 3D 손동작을 인식하였다.

마지막으로 실시간 모바일 VR 3D 손 포즈 인식 시스템을 구현 검증하였다. 학습된 손 동작 인식 딥러닝 모델을 안드로이드 모바일 어플리케이션 가상현실(VR)에 이식하고, 뎁스 손영상으로부터 3D 손 관절 위치를 추정하고, 추정된 관절 정보를 기반으

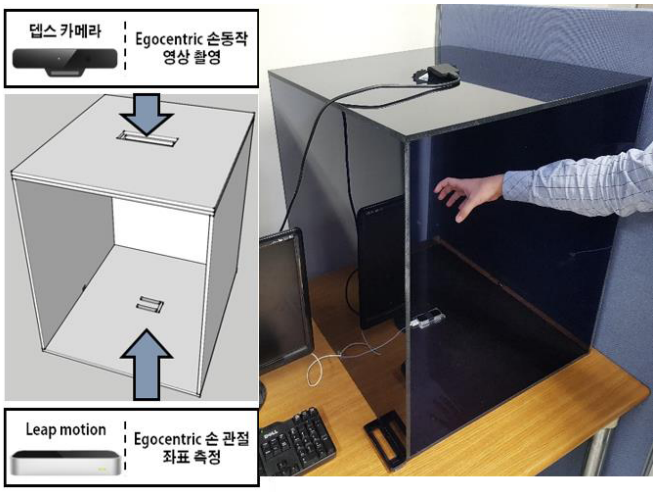
로 3D 손동작을 재구성하였다. 인식된 손동작으로 HMD VR 상에서 가상 현실 내의 물체와 상호작용을 통하여 기능을 검증하였다.

2. 방법

A. 1 인칭(Egocentric) 시점 손동작 Database 생성

딥 러닝 모델의 1 인칭 손동작 학습에 사용할 1 인칭 시점의 손동작 데이터베이스를 구축하기 위하여, 직접 Egocentric 손동작 촬영 시스템을 구축하고 데이터를 생성하였다.

그림 1 에 제작 촬영 시스템을 도시하였다. 제작한 박스의 하단에 Leap Motion[14]를, 상단에 Intel SR 300[15]를 설치하였다. 두 센서 모두 적외선을 감지하기 때문에 외부의 빛에 의한 오류가 발생할 수 있으므로 외부 환경을 제어하기 위해 암실로 박스를 제작하였다. 하부 Leap Motion 센서로 21 개의 관절 정보를 측정하여 저장하였고, 상부 Intel SR300 카메라로 1 차원 손 뎁스 영상을 얻어 딥 러닝 학습에 적합한 1 차원 시점 손동작 데이터베이스를 생성하였다. 두 기기의 초당 촬영하는 프레임 수가 다르기 때문에 PC 의 시스템 시간을 통해 뎁스 이미지에 대해 관절 데이터를 동기화하였다. 학습에 알맞은 16 개의 관절 정보와 640 x 480 16bit PNG 형태로 저장하여 데이터베이스를 생성하였다.

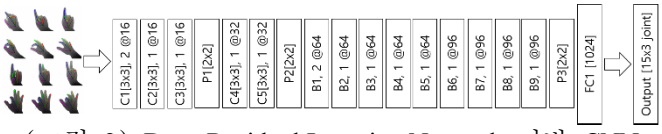


(그림 1) 제작된 박스의 구조도. (좌측) 측정 셋업 구조도 (우측) 손 동작 데이터 촬영 시연

B. 3D 손 동작 인식 Deep Residual Network

손 뎁스 영상으로부터 3D 손 동작을 인식하기 위해 3D 손 관절의 위치를 추정하는 Deep Residual Learning Network 기반 CNN Regressor 를 그림 2 과 같이 설계하였다. Regressor 는 총 5 개의 Convolution Layer, 3 개의 Pooling Layer, 1 개의 Fully-connected Layer 그리고 9 개의 Bottleneck Layer 로 구성된다. Bottleneck Layer 는 Residual Network 를 기반으로 하며 계산 복잡도와 연산 시간을 줄이기 위해 1x1 커널로 dimension 을 줄이고 3x3 Convolution Layer 수행 후 마지막 1x1 커널로 다시 차원을 복구한다. 깊은 신경망일수록 학습 데이터 속 존재하는 개념을 잘 추출할 수 있어 학습 결과가 좋아지지만, gradient 값이 너무 큰 값이나 작은 값으로 포화되어 학습의 효과가 없어지고 학습 속도가 느려지는 vanishing/exploding gradients 문제로 인해 단순히 더 많은 레이어를 쌓는 것 만으로는 성능이 보장되지 않는다. 따라서 깊은 신경망을 최적화하고, 정확도 개선을 위한 Residual Network 의 개념을 도입하여 Bottleneck Layer 를 구성하였다[11]. 이는 입력과 출력을 더해주는 Skip Connection 에 의해 신경망은 입력과의 차이만을 학습하면 된다. 따라서 입력의 작은 움직임을 쉽게 검출할 수 있으며 입력이 바로 출력에 연결되어 파라미터의 수에 영향이 없고 연산 복잡성도 해결된다는 장점이 있다. 본 논문에서 구성된 Regressor 는 계층적으로 입력에서 특징을 추출하고 최종 16 개의 손 관절 정보를 3 차원 좌표로 예측한다.

본 연구에서 제안하는 손 동작 추정 모델을 학습, 검증하기 위해 i7-6850 CPU, 16GB RAM, 4 개의 NVIDIA Geforce GTX 1080 그래픽카드가 장착된 데스크탑을 사용하였다. Python2 와 딥 러닝 프레임워크인 TensorFlow[12]를 사용하였다.



(그림 2) Deep Residual Learning Network 기반 CNN Regressor 의 구조. 뎁스 영상을 입력으로 받으며 출력으로 16x3 의 3D 손 관절 정보를 추정한다.

C. 모바일 디바이스 VR 시스템 구성

뎁스 센서로는 Time-of-Flight(ToF) 기반 PMD 사의 Pico Flex[16] 카메라를 사용하였다. 카메라는 Android 8.0 가 탑재된 LG G6 에 C++로 구현된 Royal SDK 를 Java Native Interface(JNI)를 통하여 연동한다.

입력된 손 뎁스 영상의 3 차원 중심 좌표의 정확도를 높이기 위해 실시간으로 들어오는 프레임 당 중심점(Center of Mass, CoM)의 값들의 5 개의 평균을 계산하여 최종 CoM 을 결정한다. 이 후 깊이 정보에 따른 유동적인 손 추출을 위해 얻은 3 차원 중심 좌표를 실제 좌표계에서 x, y 축으로 임계값을 설정하여 거리를 계산한 뒤 이를 2D 뎁스 영상 좌표계인 u, v 로 변환하여 관심 영역(Region of Interests, ROI)을 추출한다. 3 차원 좌표와 이미지 좌표를 서로 변환할 수 있는 수식 (1)을 이용하면 3 차원 실제 좌표에서 ROI 의 좌표를 찾을 수 있다. cu, cv 는 이미지의 크기를 나타내고 fx, fy 는 뎁스 카메라의 focal length 를 나타낸다.

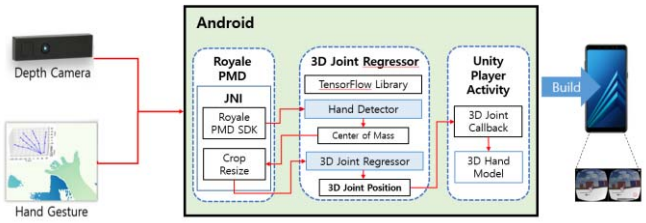
$$(x, y, z) = \left(\frac{(u - cu) \times z}{fx}, \frac{(v - cv) \times z}{fy}, z\right)$$

$$(u, v)_{start} = \left(\frac{(x - 125) \times fx}{z} + cu, \frac{(y - 125) \times fy}{z} + cv\right) \qquad (1)$$

$$(u,v)_{end} = \left(\frac{(x+125)\times fx}{z} + cu, \frac{(y+125)\times fy}{z} + cv\right)$$

TensorFlow Mobile 을 사용하여 구현한 딥 러닝 모델을 Freeze 및 Optimize Model 을 통해 최적화하여 안드로이드 플랫폼에 이식하였다. 해당 모델을 통해 입력 뎁스 손 영상으로부터 3D 관절 정보를 추정하였다.

모바일 디바이스에서 가상현실(VR) 환경을 구축하기 위하여 Unity3D[17]를 이용하였다. 실제 손 관절을 움직여서 실시간으로 가상현실의 물체를 잡아 옮기거나 화면을 터치하여 색을 변화시키는 상호작용이 가능하였다. 안드로이드 모바일 디바이스 상에서의 손동작 인식 VR 시스템 및 함수 구조를 그림 3 에 도시하였다.



(그림 3) 안드로이드 모바일 시스템 및 함수 구조 개요.

3. 결과 및 분석

A. Egocentric DB 생성 결과

Egocentric 데이터베이스의 샘플 데이터 관절 정보와 뎁스 영상 정보를 그림 4 에 도시하였다. 640x480 뎁스 영상 해상도에 손 관절 수는 16 개이며 각 관절의 3D 위치로 구성된다.

(a) (b)

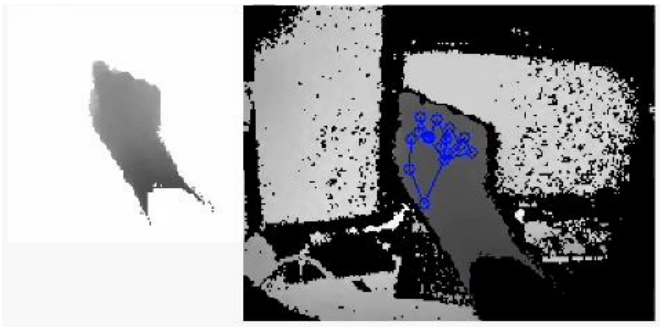
(그림 4) 3D Egocentric DB 예시. (a) & (b) 뎁스 손영상 및 16 조인트 위치

B. 3D Egocentric 손 동작 추정 학습 및 결과

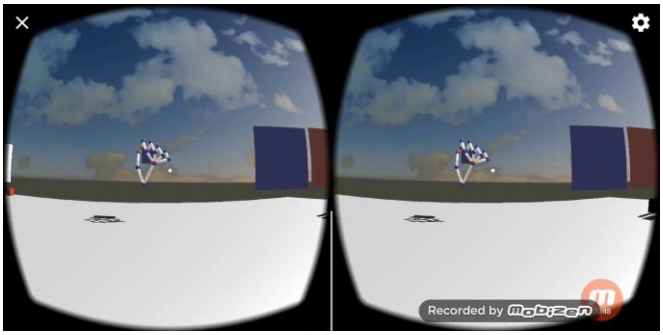
Egocentric 데이터베이스를 사용하여 손 동작 추정을 통해 얻은 시스템의 최종 출력인 예측된 관절 좌표와 Ground-truth 사이의 에러를 계산하였다. 에러는 3 차원 유클리드 거리로 계산하였다. 약 10,000 개의 학습용 데이터와 1500 개의 테스트용 데이터를 사용했고, 평균 거리 에러는 6.312, 표준편차는 2.204 의 결과를 갖는다. 기존 연구의 평균 거리 에러 10.31mm[13]에 대비하여 뛰어난 성능을 보여준다.

C. 모바일 가상환경 어플리케이션 구현 결과

그림 5 에서 모바일 VR 의 1 차원 손동작 인식 예를 도시하였다. (a)는 손 중심을 통해 찾은 ROI 영역이고, (b)는 연동된 뎁스 카메라 화면이며 재구성된 손 관절을 표시했다. (c)는 해당 손 동작을 인식하여 가상 환경에서 물체를 잡은 화면이다. 실제 손 관절을 움직여 가상환경 속의 물체를 잡아 옮기거나 스크린을 터치하여 해당 부분의 색을 변경할 수 있으며 초당 18~23 프레임 정도의 성능을 보였다.



(a) (b)



(c)

(그림 5) (a) 뎁스 손 영상 (b) 3D 손 조인트 및 뼈대 인식 (c) 가상현실 3D 손동작 인식 결과.

4. 결론

본 연구에서는 PMD Pico Flex 뎁스 센서로부터 실시간으로 들어오는 영상을 안드로이드 모바일 시스템에서 Deep Residual Network 기반 Deep Residual Regressor 을 학습시키고, 손 동작 추정을 통해 가상환경 상에 실제 구현하여 실시간 작동을 검증하였다. 또한 안드로이드 모바일 상 Unity3D 로 가상현실 물체와 상호작용할 수 있는 시스템을 구현했다.

ACKNOWLEDGEMENT

이 논문은 2019 년도 정부(미래창조과학부)의 재원으로 한국연구재단 -현장맞춤형 이공계 인재양성 지원사업의 지원을 받아 수행된 연구임(No. 2017H1D8A1031522). 본 논문은 산업통상자원부 국제

공동기술개발사업으로 지원된 연구임 ((MOTIE.Korea)(N0002252))

참고문헌

[1] Guleryuz, Onur G., and Christine Kaeser-Chen. "Fast Lifting for 3D Hand Pose Estimation in AR/VR Applications." 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018.
[2] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436.
[3] Yuan, Shanxin, et al. "Depth-based 3d hand pose estimation: From current achievements to future goals." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[4] Oberweger, Markus, Paul Wohlhart, and Vincent Lepetit. "Hands deep in deep learning for hand pose estimation." arXiv preprint arXiv:1502.06807 (2015).
[5] Ye, Qi, Shanxin Yuan, and Tae-Kyun Kim. "Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation." European conference on computer vision. Springer, Cham, 2016.
[6] Yuan, Shanxin, et al. "Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
[7] Yuan, Shanxin, et al. "Depth-based 3d hand pose estimation: From current achievements to future goals." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[8] Yuan, Shanxin, et al. "The 2017 hands in the million challenge on 3d hand pose estimation." arXiv preprint arXiv:1707.02237(2017).
[9] Malik, Jameel, et al. "Deephps: End-to-end estimation of 3d hand pose and shape by learning from synthetic depth." 2018 International Conference on 3D Vision (3DV). IEEE, 2018.
[10] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[11] He, Kaiming, et al. "Identity mappings in deep residual networks." European conference on computer vision. Springer, Cham, 2016.
[12] Abadi, Martín, et al. "Tensorflow: A system for large-scale machine learning." 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16). 2016.
[13] Gi, Geon, et al. "Real Time 3D Pose Estimation of Both Human Hands via RGB-Depth Camera and Deep Convolutional Neural Networks." International Conference on the Development of Biomedical Engineering in Vietnam. Springer, Singapore, 2018.
[14] https://www.leapmotion.com/
[15] https://www.intel.com/
[16] https://pmdtec.com/picofamily/
[17] https://unity3d.com/