광학식 동작 데이터에 대해 노이즈에 강인한 방위불변 팔 제스처 인식 기술

김성실○, 박고은, 이수아, 한다성
한동대학교
sskim5272@gmail.com, 21500250@handong.edu, 21700525@handong.edu
dshan@handong.edu

Noise-robust and orientation-invariant arm gesture recognition for optical motion data

Sungshil Kim ○, Goeun Park, Sua Lee, Daseong Han
Handong Global University

요 약

일반적으로 제스처 인식에 주로 쓰이는 2차원 영상 데이터는 제스처의 깊이 데이터를 수집할 수 없기 때문에 팔 관절의 움직임을 정확히 파악하기 어렵다는 한계점이 있다. 또한 3차원 공간 상에서 사용자의 방위에 대한 정보가 없으면 사용자의 방위에 따라 제스처 인식에 혼선이 생길 수도 있다. 이 문제를 다루기 위해 본 논문은 3차원 광학식 동작 데이터를 이용하는 Convolutional Neural Network (CNN) 모델에 기반한 팔 제스처 인식 방법을 제안한다. 또한 노이즈에 강인한 인식 결과를 얻기 위해 임의로 노이즈를 넣은 학습 데이터를 가지고 CNN을 효과적으로 학습시키는 방법을 제안한다.

1. 서 론

최근 IoT, VR/AR 등의 기술이 발전하면서 직관적인 인터렉션을 가능하게 하는 제스처 인식 기술에 관한 연구들이 활발히 진행되고 있다. 제스처 인식은 인체 각 부위의 움직임에 관한 일련의 데이터 속에서 특징을 추출해 제스처의 의미를 인식하는 과정을 거친다. 여기서 데이터 수집을 위해 영상 데이터가 많이 이용되는데, 2차원 영상 데이터를 사용하게 되면 깊이나 방위 정보를 얻기 어렵기 때문에 상당히 제약된 상황에서만 높은 정확도를 얻을 수 있다. 최근 몇 년간 깊이 감지 기술이 발전하면서, 깊이 센서를 장착한 마이크로소프트사의 키넥트(Kinect)를 이용해 팔 제스처를 인식하는 연구도 진행되었다[1]. 하지만 키넥트로부터 사용자의 방위에 대한 정보를 받아 오기 어렵기 때문에, 사용자가 바라보는 방향이 갑자기 달라지는 경우 제스처 인식의 정확도가 떨어질 수 있다.

본 논문은 이러한 문제점들을 해결하기 위해 광학식 모션 캡처 시스템을 이용한 제스처 인식 기술을 제안한다. 모션 캡쳐 시스템으로부터 실시간으로 측정되는 사용자의 위치와 방위를 기준으로 지역 좌표계를 설정하고 제스처 동작에 대한 캡쳐 데이터를 그 좌표계를 기준으로 표현하여 깊이 정보와 함께 사용자 방위에 불변한 캡쳐 데이터를 얻을 수 있다. 본 논문은 이러한 데이터에 기반하여 사용자 방위 변화에 강인한 팔 제스처 인식 기술을 제안한다. 또한 광학 센서에서 발생하기 쉬운 노이즈나 데이터 손실 문제를 다루기 위해, 학습 데이터에 임의로 노이즈를 넣어 가공한 데이터를 가지고 CNN 모델을 학습시키는 방법을 제안한다. 실험 결과로서 제안된 방법이 일반 데이터만 학습시켰을 때보다 정확도가 상당히 더 높고 노이즈가 있는 상황에서도 높은 정확도로 제스처를 인식한다는 것을 보여준다.

2. 관련 연구

다양한 데이터와 인식 방법이 제스처 인식에 사용되고 있는데, 손 제스처를 인식하는 연구들이 많은 부분을 차지하고 있다. 손 제스처 인식을 위해 수집되는 데이터에는 2D 이미지 데이터나 센서 데이터들이 있고, 인식을 위해 은닉 마르코프 모델(Hidden Markov Model, HMM), 조건적 랜덤 필드 모델(Conditional Random Field model, CRF), Support Vector Machines (SVM), CNN 모델 등이 사용된다. Keskin 와 그의 동료들의 연구[2]에서는 특정한 calibration object 를 이용해 색깔이 있는 장갑을 끼고 촬영한 2D 손 제스처 이미지를 3 차원으로 재구성하고, 이를 HMM 을 이용해 인식하였다. 또한 Kuraki 와 그의 동료의 연구[3]에서는 HMM 을 이용해 손 제스처에 대한 실시간 깊이 데이터를 인식하는 방법을 제안했다. 손 제스처 뿐 아니라 다른 제스처를 인식하는 연구로서, Ronao 와 Cho 의 연구[4]에서는 스마트폰에 내장된 가속도 센서와 자이로스코프 센서 데이터를 입력 데이터로

사용하는 CNN 모델을 통해 사용자의 동작을 인식하도록 하였다.
많은 노이즈를 포함한 광학식 마커 데이터로부터 캐릭터의 관절체 동작 데이터를 강인하게 추출하기 위해, Holden 의 연구[5]는 임의로 노이즈를 넣은 마커 데이터를 가지고 인공 신경망을 학습시켜서 마커 데이터로부터 캐릭터 동작을 안정적으로 추출하는 학습방식을 제안하였다. 본 논문은 이 방식을 팔 제스처 인식에 적용하여 인식률을 높이는 방법을 제안한다.

3. 시스템 개요

본 논문에서 제안하는 시스템은 크게 동작 캡처, 전처리, 학습의 세 가지 구성요소들을 갖는다. 동작 캡처에서는 먼저 사용자의 등과 손목에 마커를 붙이고 이들에 동작을 강체 동작으로써 추적한다. 전처리에서는 등 쪽 강체의 위치와 방위를 기준으로 지역 좌표계를 생성하여 손목 강체의 위치 데이터를 그 좌표계에 대한 데이터로 변환하고 하나의 제스처 단위로 분절한다. 본 논문에서는 더 다양한 학습 데이터를 얻기 위해 기존 데이터에 임의적으로 노이즈를 넣는다. 저장된 데이터는 CNN 모델에 학습시키기에 적절한 형식으로 맞춰 주기 위해 정규화(normalization)과정을 거친다. 정규화 된 데이터들은 학습용 데이터와 시험용 데이터로 나눠지고, 학습용 데이터를 CNN 모델에 학습시킨다. 이런 과정들을 거쳐 데이터 학습이 끝나면 CNN 모델에 두 종류의 시험용 데이터를 입력해 제스처 분류에 있어서의 정확도를 확인하여 성능을 평가한다. [그림 1]은 시스템 개요의 도식화이다.

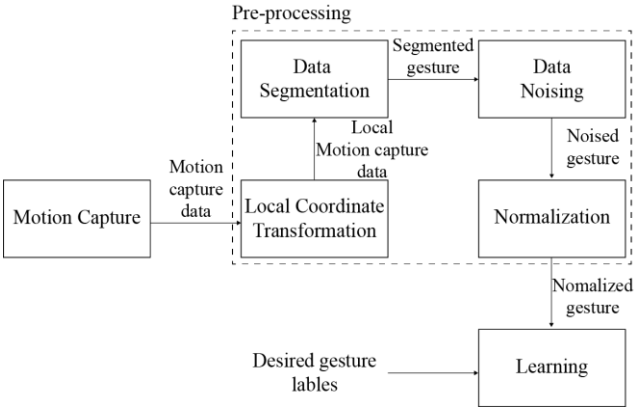

그림 1. 시스템 개요

4. 팔 제스처 모션 데이터 수집

본 논문에서는 차후 데이터 분리과정을 단순화하기 위해 동작의 시작점과 끝점이 정해진 순환적인 동작들을 인식할 제스처 대상으로 삼고, 데이터 수집 시 각 동작 사이에 충분한 시간차를 주어 제스처 간의 구분을 분명하게 해 주었다. 인식할 제스처 대상은 네 종류로, 원 제스처, 삼각형 제스처, 교차형 제스처와 별 제스처이다 [그림 2]. 또한 데이터에 다양성을 주기 위해 각 제스처의 이동 방향과 동작의 크기, 사용자가 바라보는 방향 등을 다르게 하여 모션 데이터를 수집했다. 수집과정에서 실험의 효율성을 위해 사용자는 약 6 분간 같은 종류의 제스처를 반복하였다.

그림 2. 인식할 제스처의 종류

5. 데이터 전처리 과정

본 논문에서 수집된 데이터에 기반하여 CNN을 학습시키기 전에 일련의 전 처리 과정을 거친다. 먼저 전역 좌표계로 수집된 손목의 위치 데이터를 등 쪽 강체를 기준으로 생성한 지역 좌표계에 대한 좌표로 변환하는데, 그 과정은 다음과 같다. 등 쪽 강체의 위치값을 P_b, 회전행렬을 R_b, 손목 강체의 위치값을 P_w 이라 할 때, P_b로 생성되는 지역 좌표계의 P_w값은 식(1)로 나타낼 수 있다. 변환식을 통해 서로 다른 방위을 가지던 두 좌표계의 데이터가 같은 방위를 가지게 된다.

$$P'_w = R_b^T(P_b - P_w) \quad (1)$$

로컬 좌표계로 변환된 데이터는 여러 개의 제스처에 대한 하나의 데이터로 이루어져 있다. 이를 각 제스처에 대한 데이터로 나눠 주기 위해 다음 식(2)을 이용한다. Δ_i는 0.5초 단위의 마커 좌표계 x,y,z의 변화량이며 이 값이 실험적으로 얻은 임계값 θ =0.4를 넘을 때를 기점으로 데이터를 분할한다.

$$\Delta_i=\sum_i^{i+fps/2}\{(x_{i+1}-x_i)^2+(y_{i+1}-y_i)^2+(z_{i+1}-z_i)^2\} \quad (2)$$

분할된 데이터는 시험용 데이터로 사용하기 위해 임의적으로 노이즈를 넣어 별도로 저장된다. 본 논문에서는 모션 캡쳐 시 발생하는 노이즈로 폐쇄 마커(occluded markers)와 Jitter 노이즈의 두 가지를 가정한다. 폐쇄 마커는 마커가 어떤 물체에 의해 가려져 데이터가 존재하지 않는 경우를 의미하고 Jitter 노이즈는 실제 값과는 동 떨어진 값이 발생하는 경우를 의미한다
CNN 을 이용하여 제스처를 학습시키기 전 마지막으로 서로 크기가 다른 제스처 데이터를 정규화하는 과정을 거친다. 이 때 정규화 하는 데이터의 크기가 작아질수록 각 제스처의 특징이 사라지게 되기 때문에 정규화 할 데이터 크기(N)는 제스처의 특징을 살릴 수 있도록 실험적으로 정해진다. 본 논문에서는 fps(frame per seconds)=120 으로 데이터를 수집하였는데, 이를 N=200 (1.6sec)이 되도록 정규화 하였다.

6. CNN 모델

CNN은 사물 분류 분야를 비롯한 여러 응용 분야에서 높은 성능을 보여주고 있는 네트워크 중 하나이다[4]. 본 논문에서 사용한 CNN 모델의 구조는 다음과 같다. 먼저 정규화 된 제스처의 데이터 크기와, 3차원 위치를 나타내는 데이터가 200*3 로 input 데이터에 입력된다. 전개된 데이터는 크기가 60인 필터(filter)와 합성곱 연산을 수행하는 1차원 합성곱 계층(1-D Convolution layer)과 활성화 계층(Activation layer)을 거친다. 이때 활성화 계층은 비선형 함수인 ReLU(Rectified Linear Unit)함수가 사용되었다. 이후 크기가 20, 간격(stride)이 2인 1차원 최대 풀링층(1-D Max-pooling layer), 그리고 다른 1차원 합성곱 계층과 활성화 계층(ReLu)을 거친다. 다음으로 데이터는 완전연결 계층(Fully-connected layer)으로 넘어가고 tanh 함수를 활성화 함수로 가지는 활성화 계층을 거친다. 그리고 1000개의 가중치(weight)를 곱하고 편향(bais)를 더한 뒤 마지막으로 Softmax 활성화 함수를 거쳐 각 모션 데이터가 어떤 제스처인지 분류하게 된다.

7. 실험 결과

실험은 피실험자 3 명을 대상으로 진행되었고, 수집한 제스처의 총 개수는 1474 개로 한 종류당 평균 368 개의 제스처 데이터를 수집했다. 이 데이터에 Jitter 노이즈와 폐쇄 마커 노이즈를 주기 위해 먼저 각 제스처의 모든 값을 -0.2 에서 0.2 사이의 소수점 6 자리 값으로 변환하고, 데이터 값의 20 퍼센트를 무작위로 골라 0 값으로 바꾸었다[그림 3].

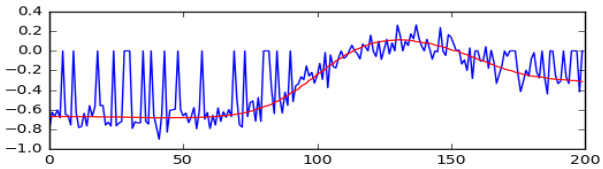

그림 3. 노이즈를 넣었을 때(파랑)와 넣지 않았을 때(빨강)의 데이터 그래프

그 결과 노이즈를 넣은 데이터의 수가 기존 데이터 수의 약 20 배가 되었다. 실험을 통해 나온 CNN 모델의 성능은 다음과 같다[표 1].

[표 1]. raw data, noise data 에 대한 accuracy

Training \ Test	Raw	Noise
Raw	0.9388	0.6582
Raw +noise	0.9691	0.9895

표와 같이 raw data 와 noise data 를 가지고 학습시킨 모델은 두 종류의 시험 데이터에 대해 0.9852, 0.9895 의 정확도를 보였고, 그렇지 않은 모델은 0.9388, 0.6582 의 정확도가 나와 노이즈를 넣어서 학습시켰을 때 정확도가 상당히 개선되는 현상을 보였다. 특히 noise data 로 성능을 시험했을 때 raw data 만 학습시킨 모델보다 정확도가 현저히 높아 모델이 노이즈에 강하다는 것을 확인할 수 있었다.

8. 결론 및 향후 연구

본 논문에서는 깊이 측정을 위해 광학식 마커 데이터를 이용하고, 사용자의 등 강체를 기준으로 로컬 좌표계를 형성해 사용자의 방위에 상관없이 높은 정확도로 인식하는 팔 제스처 인식 시스템을 제안하였다. 또한 노이즈가 발생하기 쉬운 모션 캡처 환경의 단점을 보완하기 위해 임의로 실제 상황과 비슷한 노이즈를 넣어 더 다양하게 가공한 데이터를 CNN 에 학습시켰다. 그 결과 데이터를 바로 학습시킨 모델보다 전반적으로 정확도가 개선되었고, 노이즈가 있는 데이터에도 문제없이 팔 제스처를 인식함으로써 시스템이 노이즈에 강하다는 것을 확인할 수 있었다.

그러나 본 논문은 3 명의 피실험자를 대상으로 학습 데이터를 수집했기 때문에 다양한 신체 특성을 고려하기 위해 실험 데이터를 개선할 필요가 있다. 또 향후 제스처 인식의 성능을 비교를 통해 CNN 외에 더 적합한 모델이 있는지 찾거나, 마커 개수가 더 많고 노이즈에 더 많이 노출되는 데이터를 다루어 본 논문에서 언급되지 않은 노이즈 종류를 해결하는 연구가 필요할 수 있다. 또한 더 난이도 있는 비순환형 제스처들을 추가하여 제스처 종류를 확장시키는 것도 흥미로운 연구가 될 것으로 보인다. 추가로 스스로 제스처의 시작과 끝을 판단하는 신경망 모델을 개발한다면 향후 실시간 제스처 인식 분야에서 효과적으로 사용될 것이다.

8. 참고문헌

[1] 조선영, et al. 키넥트 센서 기반 슈팅 게임을 위한 팔 제스처 인식. *정보과학회논문지: 소프트웨어 및 응용*, *39*(10), 796-805, 2012.

[2] Keskin, C., Erkan, A., & Akarun, L. Real time hand tracking and 3d gesture recognition for interactive interfaces using hmm. *ICANN/ICONIPP*, 26-29, *2003*.

[3] Kurakin, A., Zhang, Z., & Liu, Z. A real time system for dynamic hand gesture recognition with a depth sensor. In *EUSIPCO* , 2, 5, p. 6, 2012.

[4] Ronao, C. A., & Cho, S. B. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications*, *59*, 235-244, 2016.

[5] HOLDEN, D. Robust Solving of Optical Motion Capture Data by Denoising, 2018.