웨어러블 응용을 위한 CNN 기반 손 제스처 인식

문현철, 양안나, *천승문, 김재곤
한국항공대학교, 인시그널
{hcmoon, nayang}@kau.kr, smchun@insignal.co.kr, jgkim@kau.ac.kr

CNN-Based Hand Gesture Recognition for Wearable Applications

Hyeonchul Moon, Anna Yang, Sungmoon Chun, and Jae-Gon Kim
Korea Aerospace University, *Insignal

요 약

손 제스처는 스마트 글라스 등 웨어러블 기기의 NUI(Natural User Interface)를 구현하기 위한 수단으로 주목받고 있다. 최근 손 제스처 인식에서의 인식률 개선을 위하여 다양한 인식기법이 제안되고 있으며, 딥러닝 기반의 손 제스처 인식 기법 또한 활발히 연구되고 있다. 본 눈문에서는 웨어러블 기기에서의 미디어 소비 등 다양한 응용을 위하여 CNN(Convolutional Neural Network) 기반의 손 제스처 인식 기법을 제시한다. 제시된 기법은 스테레오 영상으로부터 깊이 정보와 색 정보를 이용하여 손 윤곽선을 검출하고, 검출된 손 윤곽선 영상을 데이터 셋으로 구성하여 CNN 에 학습을 시킨 후, 이를 바탕으로 손 윤곽선 영상으로부터 제스처를 인식하는 알고리즘을 제안한다.

1. 서론

최근 VR 등의 응용에서 웨어러블 기기가 확산되면서 스마트 글라스 등의 웨어러블 기기에서 손 제스처는 NUI(Natural User Interface)로 주목받고 있다. 특히, 스마트 글라스 착용시 양손을 자유로이 사용하여 사용자 명령을 생성할 수 있다는 점에서 손 제스처 인식 및 그에 따른 명령을 생성하는 기법에 대한 연구가 활발히 진행중이다. 이에 따라, 손 제스처의 정확하고 효율적인 검출 및 인식 기능이 요구된다[1], [2]. 최근 딥러닝(Deep Learning) 기술의 발전으로 인하여 다양한 인식 기술 분야에서 딥러닝 알고리즘을 적용하고 있으며, CNN(Convolutional Neural Network)은 그 대표적인 알고리즘이다[3], [4].

본 논문에서는 웨어러블 응용을 위한 CNN 기반의 손 제스처 인식 기법을 제시한다. 제시된 기법은 스테레오 영상으로부터 깊이 정보와 색 정보를 이용하여 손 윤곽선을 검출하고, 이 검출된 제스처 이진 영상을 데이터 셋으로 구성하고, CNN 에 학습하여, 이를 바탕으로 손 제스처를 인식한다.

본 논문에서는 제 2 장에서는 손 제스처 검출 기법을 제시하고 제 3 장에서는 검출된 손 제스처를 인식을 위한 CNN 기법를 제시한다. 제 4 장에서는 본 논문의 실험결과와 결론을 제시한다.

2. 손 제스처 검출

손 제스처 검출을 위해서 스마트 글라스에 장착된 스트레오 카메라를 이용하여 스트레오 비디오를 획득하고 이로부터 손의 윤곽선을 찾는다. 그림 1 과 같이 스테레오 카메라로 입력된 좌, 우 영상(RGB 영상)을 스트레오 매칭을 통하여 깊이(depth) 영상을 획득한다. 스트레오 기기 특성상 사용자의 손은 카메라로부터 일정한 거리 범위(30 ~ 50 cm)에 있다고 가정을 한다.

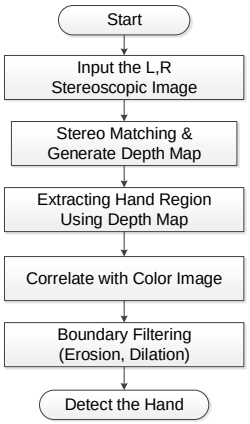

그림 1. 손 윤곽선 검출 순서도

그림 2. 검출된 손 윤곽선 예

깊이 영상과 색정보를 적용하여 손 제스처 영역을 검출한다. 보다 정확한 손 제스처의 영역을 얻어내고 잡음을 제거하기 위해 모폴로지(morphology) 연산을 한다. 그림 2 는 제시된 기법의 검출 결과 예이다[1], [2].

3. CNN 기반 제스처 인식

CNN 은 다층신경망의 한 종류로 인식 분야에서 매우 좋은 결과를 보이고 있다. CNN 은 기존의 신경망과 다르게 원본 데이터를 바로 연산하여 출력하지 않고, 특징(feature) 추출 단계와 분류화(classification) 단계를 거쳐 결과값을 내는 것이 특징이다. CNN 의 구조는 컨벌루션(convolution) 층, Pooling 층, Fully-Connected(FC) 층으로 구성되어 있다[4].

그림 3 은 본 논문의 제스처 검출을 위한 CNN 구조도이다. C1, C2 는 Convolution 층으로 입력계층에서 입력되는 영상 데이터를 5×5 크기의 필터를 사용하여 간격 1(stride=1) 로 컨벌루션을 수행하고, 데이터의 특징을 추출해 낸다. S1, S2 는 Pooling 층으로 여기서는 Max Pooling 기법을 사용하여 특징 맵의 크기를 가로 세로 각각 반으로 줄이는 역할을 수행한다. Max Pooling 층은 일정 크기 범위 내에서 가장 특징이 두드러진 부분을 그 범위의 대표값으로 지정하는 방식으로 그림 4 는 그 예를 보여준다.

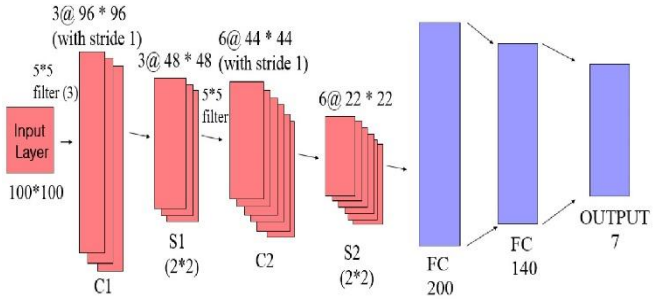

그림 3. 제안하는 CNN 구조

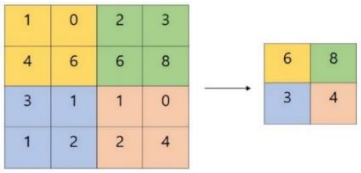

그림 4. Max Pooling 예

F1, F2 는 Fully-Connected 층이며, 이 계층은 일반적인 다층 신경망의 구조와 같이 추출된 특징을 바탕으로 신경망에서 연산하여 출력값을 결정한다.

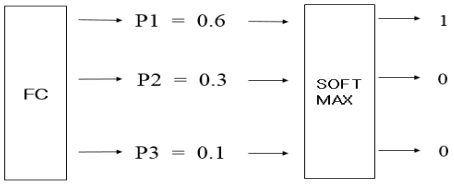

그림 5. Fully-Connected 층의 분류화 과정

그림 5 은 Fully-Connected 층의 분류화 과정의 간단한 예를 나타낸 것이며, 여기서 P1, P2, P3 는 영상에서 추출된 각 특징에 해당하는 확률 값을 나타낸다. 본 논문에서는 그림 6 의 제스처를 사용하여 P1~P7 로 변경된다. 확률 값이 계산되면 Softmax 함수를 사용하여 확률 값이 가장 큰 제스처의 값을 입력 영상의 제스처로 인식한다(결과 값 '1').

4. 실험결과 및 결론

본 논문의 실험에서는 학습을 위한 데이터 셋으로 5 명의 피실험자의 손 제스처 영상을 스테레오 카메라로 획득하였다. 그림 6 의 10 개의 제스처에 대한 5,600 개의 학습 데이터, 1,400 개의 테스트 데이터를 구성하고 인식의 정확도를 확인하였다.

표 1 에 제시된 실험결과와 같이 95.3%의 인식 정확도를 얻었으며, 각 제스처마다 91.68%~98.3%의 정확도 보임을 확인하였다. 앞으로 제스처 인식 정확도를 높이기 위한 CNN 구조도 연구를 진행할 예정이다.

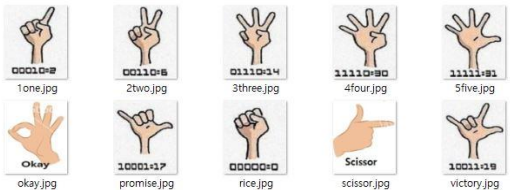

그림 6. 제스처 종류

표 1. 손 제스처 인식 결과(인식 정확도)

	1	2	3	4	5
Acuracy (%)	93.12	96.09	98.3	93.89	95.21
	6	7	8	9	10
	94.18	96.12	91.68	95.68	94.56
	Total Accuracy = 95.3 %				

ACKNOWLEDGMENT

이 논문은 2016 년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원(R0127-15-1015))을 받아 수행된 연구임.

참 고 문 헌

[1] A. Yang, S. Chun, H. Ko, J. G. Kim, "Hand Gesture Description for Wearable Applications in M-IoTW," ISO/IEC JTC1/SC29/WG11 M38526, Geneva, Swiss, May. 2016.

[2] A. Yang, S. Chun, and J.-G. Kim, "Detection and Recognition of Hand Gesture for Wearable Applications in IoMTW," In Proc. ICACT2017, Feb. 2017, pp. 598 - 601.

[3] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." Nature 521.7553 (2015): 436-444.

[4] J. Nagi, and et al. "Max-pooling convolutional neural networks for vision-based hand gesture recognition," In Proc. Signal and Image Processing Applications (ICSIPA), 2011.