# 실시간 손 제스처 인식을 위한 덴스넷 기반 이중흐름 3차원 합성곱 신경망 구조

**노대철, 최현종, 김태영**[1)]
서경대학교 컴퓨터공학과
sheocjf1025@skuniv.ac.kr, ipip4652@skuniv.ac.kr, tykim@skuniv.ac.kr

# DenseNet based Two-Stream 3D Convolutional Neural Networks for Real-time Hand Gesture Recognition

**Dae-Cheol Noh, Hyeon-Jong Choi, Tae-Young Kim**
Department of Computer Engineering, Seokyeong University

**요 약**

가상현실이 의료, 교육, 군사 훈련 등 다방면에서 사용됨에 따라 가상환경에서 자유로운 상호작용을 제공하기 위한 손 제스처 인식에 대한 연구가 활발히 진행되고 있다. 그러나 대부분은 별도의 센서를 요구하거나, 낮은 적중률을 보이고 있다. 본 논문은 정적 손 제스처 인식과 동적 손 제스처 인식을 위해 일반적인 USB 카메라 이외의 별도의 센서나 장비 없이 딥러닝 기술을 사용한 손 제스처 인식 방법을 제안한다. 입력된 손 제스처를 고주파 영상들로 변환한 다음, 각 손 제스처 영상과 고주파 영상에 대해 각각 덴스넷 기반의 이중 흐름 합성곱 신경망을 수행한 다음 융합된 정보로 보다 정확한 손 제스처를 인식한다. 그리고 실시간 인터페이스 검증을 위해 가상현실 기반 3D 디펜스 게임을 개발하여 실험한 결과, 6개의 정적 손 제스처와 9개의 동적 손 제스처 인터페이스에 대해 기존의 단일 흐름의 덴스넷에 비해 4.58%의 성능이 향상된 평균 92.6%의 인식률을 보였다. 본 연구의 결과는 마우스나 키보드 없이 다양한 가상현실 응용 분야에서 입력 인터페이스로 활용될 수 있다.

## 1. 서 론

최근 HMD와 그래픽 프로세서 성능의 발달로 가상현실에 대한 관심이 급증함에 따라 가상현실 상에서 몰입감을 높이기 위해 사용자 친화적 인터페이스(NUI: Natural User Interface)에 대한 연구가 활발히 진행되고 있다[1]. 가상현실 속의 인터페이스는 키보드나 마우스와 같은 단순한 입력 장치를 넘어서서 사용자에게 몰입감을 줄 수 있어야 하며 자연스럽고 직관적으로 제공되어야 한다.

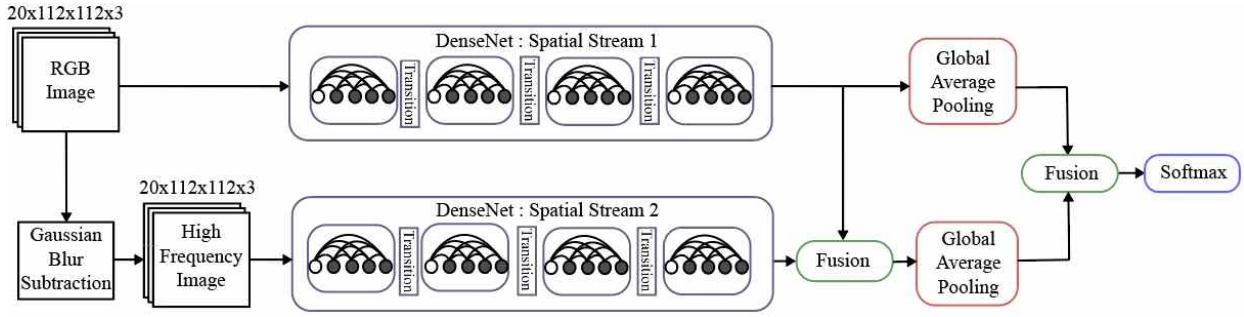
사용자에게 가장 자연스러운 인터페이스는 일상생활에서 손을 사용하여 취하는 제스처를 가상공간 속에서도 인식할 수 있도록 하는 것이다. 이와 같은 연구로 키넥트나 립모션을 이용한 손 제스처 인터페이스에 관한 연구[2]가 있었지만 특화된 별도의 센서가 필요하고 조명이나 거리에 제약을 받는 단점이 있었다. 이후 컴퓨팅 기술과 고성능 GPU의 발달로 딥러닝(Deep Learning) 기술을 손 제스처 인식에 적용하는 연구[3]가 진행되고 있다. 하지만 이 연구들 역시 특정 장비나 센서를 필요로 하거나, 낮은 적중률을 보이는 등의 문제점을 지니고 있다.

최근 딥러닝에 관한 연구로 객체 인식을 위한 다양한 신경망 구조들이 등장했는데, 그 중 덴스넷(DenseNet)[4]과 이중흐름(Two-Stream)[5] 구조를 예로 들 수 있다. 덴스넷은 1x1x1 커널의 합성곱(Convolution)을 사용하여 특징 맵의 크기를 조정하기 위한 병목 계층(BottleNeck Layer)과 압축 계층(Compression Layer)을 삽입하여 적은 파라미터로 높은 속도와 높은 인식률을 제공한다. 이중흐름 구조는 입력 데이터의 공간적 흐름에 따른 정적인 정보 외에 시간적 흐름에 따른 동적인 정보를 보완하기 위하여 광학 흐름(Optical Flow) 처리를 한 데이터를 입력으로 각각 학습 시킨 다음 각 신경망으로부터 출력된 특징 맵(Feature Map)을 융합(Fusion)함으로써 인식의 정확도를 높였다.

본 연구에서는 이중흐름 구조를 사용하되 손 제스처의 특성상 시간 흐름 정보보다 각 타임구간의 손 포스처 영상이 인식에 많은 영향을 끼치는 점을 고려하여 손 영상의 고주파(High Frequency) 정보를 별도로 학습하여 기존 결과와 융합한다. 또한 합성곱 신경망 구조로 적은 파라미터와 높은 인식률을 제공하는 덴스넷을 사용한다. 본 논문에서 제안하는 손 제스처 인식 방법은 다음과 같다. 일반적인 USB 카메라로 최근 20 프레임의 손 제스처를 입력 받아, 이중흐름 구조의 첫 번째 흐름의 입력으로 가공하지 않은 RGB 영상을, 두 번째 흐름의 입력으로 원본 영상의 고주파 영상을 사용한다. 각 흐름의 마지막 합성곱 연산을 수행한 후 RGB 흐름과 고주파 흐름에서 나온 특징 맵에 대하여 1차 융합을 실시한다. 이후 RGB 흐름에서 출력된 특징 맵과 융합된 특징 맵에 대해 각각 전역 평균 풀링(Global Average Pooling) 연산을 수행한다. 그 후 최종적으로 출력된 결과에 대해 2차 융합을 실시하고, 소프트맥스(Softmax) 함수를 거쳐 인식 결과를 출력하게 된다. 본 방법의 검증을 위하여 조명, 배경, 위치와 거리 등 다양한 상황을 고려한 15가지 손 제스처 데이터 세트로 실험한 결과 시간에 따라 위치와 모양이 변하는 동적 제스처와 변하지 않는 정적 제스처에 대하여 이중흐름 구조를 사용하지 않는 일반적인 덴스넷과 비교했을 때 평균 5%의 향상률을 보였다.

본 논문의 구성은 다음과 같다. 2장에서 본 논문에서 제안한 손 제스처 인식을 위한 덴스넷 기반의 3차원 이중흐름 합성곱

1) 교신저자




(그림 1) 덴스넷 기반 이중흐름 3차원 합성곱 신경망

신경망에 대해 설명한다. 3장에서 학습을 위한 손 제스처 정의 및 실험 결과에 대해 기술한 후 4장에서 가상현실 게임에 적용한 결과를 기술한다.

## 2. 덴스넷 기반의 3차원 이중흐름 신경망 구조

본 논문은 실시간 손 제스처 인식을 위한 덴스넷 기반 이중 흐름 신경망 구조(그림 1)를 제안한다. 기존 연구에서 사용한 광학 흐름 처리는 긴 처리 시간을 필요로 하고, 각 흐름에서 사용한 신경망인 VGGNet은 계층이 깊어질수록 더 많은 연산을 필요로 한다는 단점이 있다. 이러한 단점을 개선하기 위하여 광학 흐름 처리 대신 손의 상세 정보를 부가적으로 학습하여 성능을 높이기 위해 상대적으로 처리 시간이 짧은 가우시안 블러를 활용한 고주파 영상을 두 번째 흐름 신경망 구조의 입력으로 사용하고, VGGNet 대신 상대적으로 얕은 커널을 사용하면서 적은 연산량과 높은 인식률을 가진 덴스넷을 사용하였다.

손 제스처 인식을 위해 일반적인 USB 카메라에서 초당 30 프레임의 속도로 사용자의 제스처를 촬영한 후. 최근 20 프레임의 영상으로 입력 데이터를 제작한다. 입력 데이터는 가공하지 않은 RGB 손 제스처 영상 데이터와 고주파 영상 데이터로 나누어지며, 고주파 영상은 원본 영상을 가우시안 블러 처리하여 저주파 영상으로 변환한 후, 원본 영상에서 저주파 영상을 뺌으로써 제작한다(그림 2). 이후, 메모리 절약을 위해 정규화 및 112x112 크기로 재조정(Resize)을 거치고, 두 개의 공간 흐름 신경망인 덴스넷에 각각 입력되어 학습이 진행된다. 신경망에 입력된 손 제스처 영상은 3x7x7 커널을 사용하는 합성곱을 한 번 수행하고 이후 4개의 고밀도 연결 구역을 거친다. 고밀도 연결 구역에서 발생하는 합성곱 연산의 입력에 대해 배치 정규화(Batch Normalization)와 ReLU 활성 함수를 실시하고 1x1x1 커널 합성곱과 3x3x3 커널 합성곱을 4번 반복한다. 이후 이행계층(Transition Layer)에서 각 프레임의 정보를 보존하기 위해, 영상의 크기를 다운 샘플링 할 때에도 프레임의 크기는 유지시킨다. 4개의 고밀도 연결 구역을 거친 후, 1차 융합을 통해 두 개의 흐름 신경망에서 출력된 특징 맵을 하나로 융합한다. 융합 기법은 두 개의 특징 맵을 서로 합친 후 개수만큼 나누어주는 평균 융합(Average Fusion) 기법을 사용하였다. 융합이 끝난 후 첫 번째 전역 평균 풀링 계층에는 RGB 흐름에서 출력된 특징 맵이 그대로 입력되고, 두 번째 전역 평균 풀링 계층에는 융합된 특징 맵이 입력된다. 이 때 과적합(Overfitting)에 대한 대책으로 드롭아웃(Dropout)을 수행한다. 위와 같은 단계를 거쳐 출력된 각각의 특징 맵은 2차 평균 융합을 거친 후 소프트맥스 함수를 통해 인식 결과를 출력한다.

(그림 2) 원본 영상(좌)과 저주파 제거 영상(우)

## 3. 손 제스처 정의 및 실험 결과

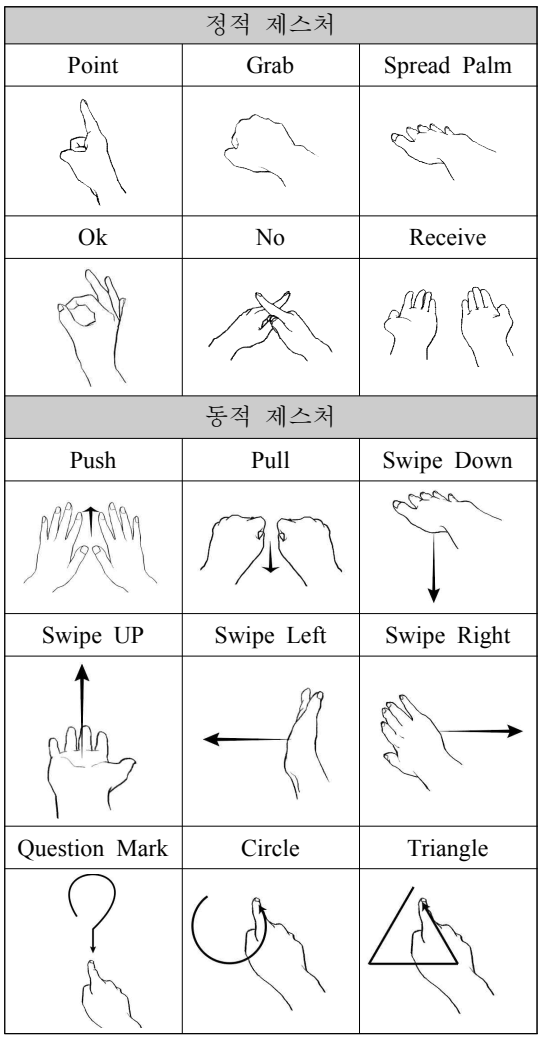
실험에 사용된 손 제스처는 손 제스처를 시작부터 끝까지 손의 모양과 위치가 변하지 않는 정적 제스처와 시간에 따라 위치가 변하는 동적 제스처로 분류하여 6가지 정적 제스처와 9가지 동적 제스처를 정의하였다(그림 3).

정적 제스처와 동적 제스처의 통합된 인식 모델 실험은 프로세서 Intel Core i5 8400, 그래픽 카드 GeForce GTX 1080Ti, RAM 16GB 등으로 구성된 장비와 Python 3.5 기반의 Tensorflow 1.5 GPU 버전을 개발도구로 사용한 환경에서 진행되었으며, 배치 크기 8, Epoch 20, 드롭아웃 비율 0.5의 조건에서 진행되었다. 2D 합성곱을 이용한 정적 제스처 인식과 3D 합성곱을 이용한 동적 제스처 인식은 서로 다른 신경망에서 충분한 성능을 보였지만 통합되지 않은 두 모델을 번갈아 사용하는 것은 GPU 메모리와 인식 속도 측면에서 제약을 가진다. 본 실험에서는 실시간 인식을 위해 정적 제스처와 동적 제스처를 모두 인식하는 통합 모델을 구현한 다음 그 성능을 실험하였다.

실험 결과, <표 1>에서 보는 바와 같이 동적 제스처보다 정적 제스처의 성능이 낮았으며, 정적 제스처 만으로 학습했을 때보다 크게 저하됨을 확인할 수 있었다. 그 중에서도 상대적으로 유사한 Point, Grab, Spread Palm, OK의 네 가지 정적 제스처의 성능이 떨어졌으며 이를 통해 서로 다른 특징으로 분류되는 두 제스처를 일괄적으로 학습할 시 인식률이 저하됨을 확인할 수 있었다. 이를 해결하기 위해 본 논문에서 제시한 고주파 정보를 입력한 덴스넷 기반 이중흐름 신경망은 전반적으로 4.58%의 성능이 향상된 92.5%의 인식률을 보였다.

## 4. 응용 사례

본 논문의 덴스넷 기반 이중 흐름 신경망 학습 모델의 성능을 확인하기 위하여 가상현실 기반의 응용 프로그램을 제작하였다. 응용 프로그램은 다리 건너편에서 다가오는 몬스터들을 막는 디펜스 게임 프로그램으로, 주인공 캐릭터는 손 제스처 15가지를 인터페이스로 사용하여 마법 기술을 사용할 수 있다.




(그림 3) 정의 된 손 제스처 15가지

몬스터의 종류와 상황에 따른 다양한 마법 기술을 사용하기 위해 각 기술의 컨셉에 맞는 다양한 인터페이스를 구현하였다.

USB 카메라를 통해 사용자의 손 제스처를 초당 30 프레임 촬영하여 신경망 모델에 적용하여 실험한 결과 손 제스처를 평균 34 ms로 실시간 인식하여 손 제스처 기반 인터페이스로 가상현실 게임 실행이 가능함을 알 수 있었다(그림 4).


(그림 4) 손 제스처 인식을 활용한 가상현실 게임

<표 1> 기존 덴스넷과 이중흐름 덴스넷의 제스처별 인식률

| 제스처 | 덴스넷-BC | 이중흐름 덴스넷-BC |
|---|---|---|
| Point | 68.75% | 70.83% |
| Grab | 70.83% | 79.17% |
| Spread Palm | 66.67% | 72.92% |
| Ok | 72.92% | 83.33% |
| No | 89.58% | 91.67% |
| Receive | 93.75% | 95.83% |
| Push | 89.58% | 97.92% |
| Pull | 95.83% | 100.00% |
| Swipe Down | 97.92% | 100.00% |
| Swipe Up | 95.83% | 97.92% |
| Swipe Left | 97.92% | 100.00% |
| Swipe Right | 95.83% | 100.00% |
| Question Mark | 100.00% | 97.92% |
| Circle | 83.33% | 100.00% |
| Triangle | 100.00% | 100.00% |
| Total | 87.92% | 92.50% |

**Acknowledgement**

이 논문은 2017년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임. (No. NRF-2017R1D1A1B03029834)

## 참 고 문 헌

[1] C. Perrenot, M. Perez, N. Tran, and JP. Jehl. “The virtual reality simulator dV-Trainer is a valid assessment tool for robotic surgical skills,“ Surg Endosc, 26:2587 – 2593, 2012.

[2] G. Marin, F. Dominio, and P. Zanuttigh. “Hand gesture recognition with leap motion and kinect devices,” IEEE International Conference on Image Processing, 2014.

[3] P. Molchanov, S. Gupta, K. Kim, and J. Kautz. “Hand gesture recognition with 3D convolutional neural networks,” IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015.

[4] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. “Densely connected convolutional networks,” IEEE Conference on Computer Vision and Pattern Recognition, pp 3-11, 2017.

[5] K. Simonyan, A. Zisserman. “Two-stream convolutional networks for action recognition in videos,” NIPS, 2014.