

# 딥러닝과 제스처 인식 기술

□ 양희덕 / 조선대학교

## I. 서론

컴퓨팅 장치, 센서 성능의 발전과 대용량 자료를 분석할 수 있는 머신 러닝 기술의 팔목할 만한 발전으로 인해 자동차, 휴대폰, TV와 같은 제품명 앞에 언제부터인가 ‘스마트’라는 단어가 자리잡게 되었다. 초기에 등장한 스마트 제품들은 기능이나 성능에 중점을 둔 기술의 발전에 초점이 맞추어 졌다면, 최근에 출시되고 있는 스마트 제품들은 사용자에게 제품의 활용 과정에서 최적의 경험을 제공하기 위한 사용자와 제품 간의 상호작용에 초점을 맞추고 있다. 그 중에서도 사용자가 제품 및 서비스를 얼마나 편리하게 이용할 수 있느냐에 관심을 갖고 있으며, 이는 자연스러운 사용자 인터페이스(Natural User Interface: NUI) 기술 발전에 전환점을 제공하고 있다.

NUI의 가장 대표적인 방법은 음성을 들 수 있으며, 현재 많은 제품에 적용되었다. 스마트 자동차에

적용된 음성인식 기술은 운전 중에 사용자가 내비게이션 및 음악 재생을 쉽게 할 수 있도록 도와주고 있고, 스마트 폰 등에 탑재된 음성 인식 기술은 다양한 어플리케이션(App)에 적용되어 활용되고 있다. 최근 애플과 구글은 각 사의 운영체제에 음성인식 기술을 통합하여 제공하고 있다. 애플은 ‘시리’라는 서비스를 제공하고 있으며, 외부 개발자도 활용할 수 있도록 공개하였다[1].

음성과 함께 많이 사용되고 있는 사용자 인터페이스는 제스처이다. 제스처는 사용자가 자신의 의도를 전달하기 위해 수행하는 의도된 동작뿐만 아니라 무의식중에 의미 없이 수행하는 동작을 포함하고 있다. 또한, 제스처라고 하면 보통 일반인이 수행하는 것만 생각하고 있지만, 청각 및 시각 장애인이 수행하는 동작도 모두 제스처이다. 청각 장애인이 의사전달을 목적으로 사용하는 제스처를 수화라고 한다.

구글은 프로젝트 솔리(Soli)라는 명칭으로 레이더

센서를 이용한 동작인식 기술을 개발하고 있다. 염지와 검지의 움직임을 통해 특정 사물을 제어할 수 있는 수준까지 개발이 진행됐으며, 최근에는 레이저 센서를 이용하여 객체를 식별하는 연구도 진행하고 있다[2]. 다양한 연구소, 회사, 학교 등에서 스마트 홈, 스마트 자동차, 스마트 TV, 게임, 공연 등에서 주변 사물, 공간 및 사람에게 정보를 전달하기 위한 직관적이고, 편리한 인터페이스 방법을 심도 있게 연구하고 있는 상황에서, 제스처 인식 기술에 대한 관심은 더욱 커져가고 있다[3,4,5].

최근 딥러닝 모델은 다양한 분야에서 눈부신 성능 향상을 보여주고 있다. 초창기의 딥러닝 기술은 주로 이미지 영상을 분석하는 분야에 적용이 되었는데, 높은 성능을 기반으로 다양한 분야에 적용이 되고 있다. 딥러닝 기술은 각종 인공신경망을 이용해 데이터로부터 유용한 특징을 직접 추출 및 학습하는 방식이다. 이 방식은 사람은 찾지 못하는 유용한 특징을 스스로 찾을 수 있다. 따라서 대용량의 데이터로부터 사람은 특징을 찾기 쉽지 않지만 반대로 딥러닝 기술은 유용한 특징을 효과적으로 찾을 수 있다. 딥러닝 기술의 발전으로 물체 인식 및 검출[6], 얼굴 인식[7], 제스처 인식[8,9] 등의 다양한 분야에서 기존 알고리즘 성능의 발전이 진행되고 있다.

## II. 제스처 인식을 위한 다양한 센싱 기술

제스처 인식 기술을 위해서는 사용자가 수행한 제스처를 입력 받을 수 있는 센서가 필요하며 두 가지의 형태로 나뉜다. 센서나 장치를 사용자가 신체로 직접 접촉하여 데이터를 획득하는 접촉식 방식과 원거리 및 근거리 센서를 이용하여 데이터를 획득하는 비접촉식 방식이 있다. 최근에는 사용자가 착용할 수 있는 다양한 형태의 웨어러블 센서가 개발되고 있는데, 사용자가 직접 센서를 접촉하지 않기 때문에 비접촉식 방식으로 분류를 많이 하며, 사용자의 행동환경이나 장소에 제약을 주지 않기 때문에 연구가 많이 진행되고 있는 분야이다[4].

접촉식 방식은 사용자가 접촉하기 때문에 비교적 정확한 제스처 정보를 획득할 수 있는 반면, 사용자가 센서나 장비를 접촉해야 하는 번거로움이 있다. 반면, 비접촉식 방식은 사용자가 센서나 장비를 접촉하지 않기 때문에 사용하기 편리하고, 사용자의 동작이 자연스러운 반면, 센서에 따라서 사용자의 행동환경 및 거리의 제약이 발생하고, 획득된 데이터의 정확도가 센서 및 환경의 영향을 받는다. 그래서 이 두 가지의 장·단점을 갖고 있는 웨어러블 센서에 대한 연구와 발전이 많이 되고 있는 상황이다.



<그림 1> 닌텐도의 위(Wii) : 접촉식 센서[9]



〈그림 2〉 애플의 3D 터치: 접촉식 센서[10]

〈그림 1〉, 〈그림 2〉는 각각 접촉식 센서인 위(Wii)와 애플의 3D 터치스크린을 보여주고 있다. 애플은 아이폰에 3D터치라 불리는 인터페이스를 탑재했으며, 이는 사용자가 화면을 누르는 압력을 측정해 다양한 응용을 가능하게 하는 기술이다.

〈그림 3〉~〈그림 5〉는 각각 비접촉식 센서인 립모션(Leap motion), 리얼센스(RealSense), 토비의

아이엑스(EyeX)를 보여주고 있다. 립모션은 2개의 적외선 카메라를 이용하는 스테레오 비전 시스템이다. 기존의 대형 TV와 같은 원거리 제스처 인식 기술에 초점을 맞춘 것이 아니고, 노트북과 같은 근거리 환경에서의 제스처 인식을 위한 인터페이스에 초점을 준 제품이다.

인텔은 3대의 카메라를 이용하여 3차원 공간을 인식할 수 있는 리얼센스라는 기술의 개발과 활용에 많은 연구를 진행하고 있다. 리얼센스 기술에는 동작인식을 비롯해 안면인식, 증강현실, 3D 스캐닝 등이 포함되어 있다. 리얼 센서를 탑재한 태블릿, 노트북 등과 이를 이용한 어플리케이션 프로그램이 많이 개발 중에 있다.

토비의 아이엑스(EyeX)는 눈의 움직임을 감지해



〈그림 3〉 립모션: 비접촉식 센서[11]



〈그림 4〉 리얼센스: 비접촉식 센서[12]



〈그림 5〉 토비의 눈동자 추적 및 안경 장치[13]



〈그림 6〉 Myo: 웨어러블 센서[14]

주변 기기를 컨트롤할 수 있는 눈동자 추적 장치다. 눈동자 움직임 분석은 사용자의 명령에 반응하는 것뿐만 아니라 사용자가 무심코 한 반응까지도 데이터로 획득할 수 있기 때문에 다양한 산업분야에 적용이 가능하다. 눈동자의 깜빡임을 감지하여 운전 중 졸음을 감지하는 연구는 하나의 응용 예이다.

<그림 6>은 웨어러블 센서인 탈밀랩(Thalmic Labs)의 ‘마이요(Myo)’이다. 이 제품은 근육의 움직임을 감지해 동작을 인식하는 밴드 형태의 기기다. 손바닥을 펴거나 쥐는 동작, 손의 회전 등을 인식할 수 있다[14].

이 외에도 다양한 형태의 제스처 인식을 위한 센서가 개발 중에 있다. 터치스크린과 같은 접촉식 방식에서 웨어러블 혹은 카메라와 같은 비접촉식 방식으로 사용자의 제스처를 인식할 수 있는 센서의 개발이 진행 중이다[15].

### III. 제스처 인식 기술

본 장에서는 최근에 활발히 연구가 되고 있는 제스처 인식 기술을 손 제스처, 전신 제스처, 풀바디 모션 인식 기술, 이미지·동영상에서 제스처 인식 등으로 구별하여 소개한다.

#### 1. 손 제스처 인식

손 제스처를 인식하기 위해서 많은 연구에서는 Ⅱ장에서 소개한 다양한 센서를 이용하여 손의 위치, 모양, 궤적 정보를 이용한다. 손의 궤적 정보를 이용하는 연구에서는 시계열 데이터 분석을 위하여 HMM(Hidden Markov Model), DBN(Dynamic Bayesian Network), CRF(Conditional Random

Field)와 같은 모델이나 이의 변경 모델을 이용하고 있다[16]. 손의 모양 정보를 이용하는 경우에는 TOF(Time of Flight)나 스테레오 카메라로부터 측정된 3차원 정보를 분석하여 손의 구조적 특징을 이용하거나 손의 모양 정보를 부스팅(boosting)하는 방법을 이용하여 분석한다. <그림 7>은 BMW 자동차에서 주변 기기를 제어하기 위해서 사용하고 있는 손 제스처 인식의 예이고, <그림 8>은 조지아 공대에서 연구를 수행하고 있는 수화를 텍스트로 번역해 주는 예이다. 손 제스처는 사람이 자신의 의



<그림 7> 자동차에 적용된 손 제스처[17]



<그림 8> 수화 인식 : Copycat[18]

도를 전달할 수 있는 가장 효과적인 제스처 중 하나이기 때문에 응용 분야가 다양한 편이다.

## 2. 전신 제스처 인식

전신 제스처 인식을 위해서는 인체 구성 요소의 관계를 분석하는 것이 필요하다. 이를 위해서, 키넥트 센서나 다양한 형태의 센서를 이용하여 신체의 관절 정보를 그래프 구조로 모델링한다.

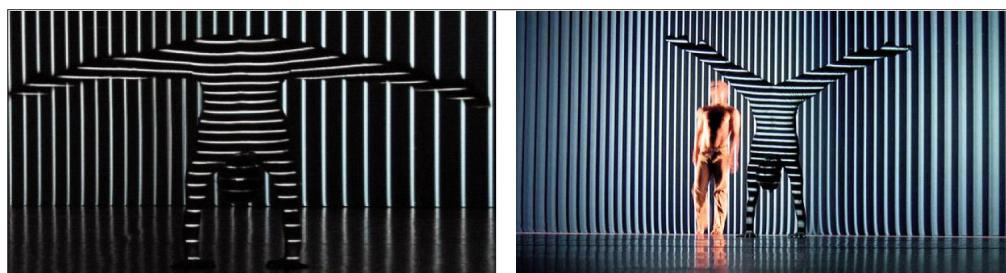
미디어 아티스트 클라우스 오베마이어는 키넥트 센서를 이용하여 동작인식 기술을 공연에 접목하였다. 무대 반대편에 키넥트 카메라를 설치하고 이를 활용하여 동작을 분석한 뒤 무대 배경의 변화를 수행하는 공연을 하였다. <그림 9>는 공연 장면의 일부를 보여주고 있다.

국내의 업체에서도 공연 관객들과 상호작용을 수

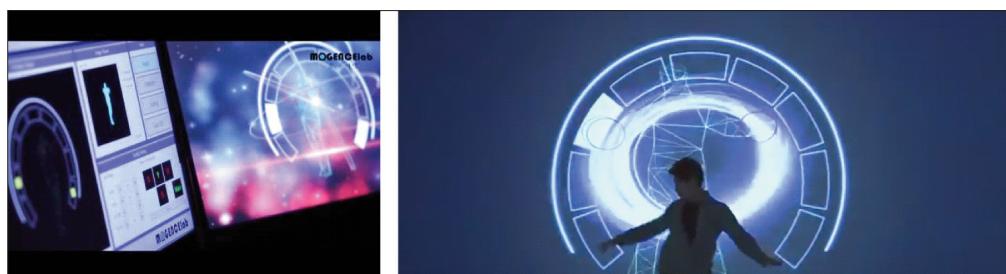
행할 수 있는 솔루션을 제공하고 있다. Playnut이라는 제품은 관객들의 반응에 대응하는 다양한 무대요소를 연출할 수 있고 공연자를 실시간으로 추적할 수도 있다. 또한, 공연자의 제스처를 인식하여 인식된 정보를 배경 및 다양한 형태로 변형하여 사용자에게 전달할 수 있다. 이를 통해 사용자의 흥미를 유발하고 공연의 재미를 증폭시킬 수 있다.

## 3. 풀바디 모션 인식 기술

3D 영화, 게임의 캐릭터 등의 모션을 생성하기 위해서 사용되는 방법으로, 몸에 여러 개의 마커를 부착하고 이를 센서로 사람의 움직임을 측정한다. 최근 가상현실에 대한 관심이 높아지고 있어, 사람이 수행한 행동을 인식하고 이를 가상의 공간에 적용하는 연구가 활발히 진행되고 있다. <그림 11>,



<그림 9> 클라우스의 공연 장면[19]



<그림 10> 전신 제스처를 활용한 공연 효과 증대[20]

〈그림 12〉는 각각 Xsens MVN 모션 캡처, Cyverith 사의 웨어러블 센서의 모션 캡처 장비와 예를 보여주고 있다.



〈그림 11〉 Xsens MVN 모션 캡처 장비[21]



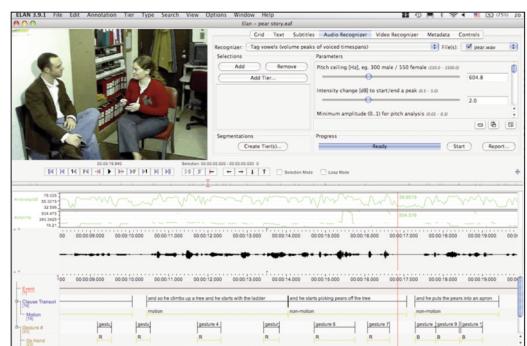
〈그림 12〉 Cyverith 사의 모션 캡처 장비[22]

#### 4. 이미지·동영상에서 제스처 인식

최근에는 CCTV 등의 증가로 인해 2D 영상에서 사람의 이상 행동 및 특별한 행동 패턴을 검출하기 위한 연구가 많이 진행되고 있으며, 이를 해결하기 위한 방법으로 딥러닝 기술이 적용되고 있다. 이미지 분석에 사용되는 컨볼류션 신경망 모델(Convolution Neural Network)을 변형한 모델이 사용

되고 있다. 또한, 인터넷 등에 존재하는 방대한 양의 이미지에서 원하는 이미지 검색을 위한 이미지 어노테이션(Image Annotation)분야에서도 활용되고 있다[23].

〈그림 13〉은 ELAN이라는 동영상에서 특정한 형태의 이미지 혹은 제스처 위치를 찾아 주는 소프트웨어의 예이다.



〈그림 13〉 ELAN 화면의 예[24]

### IV. 제스처 인식 기술의 연구 이슈

제스처 인식 기술은 음성 인식 기술에 비해 아직 기술의 완성도가 높지가 않다. 제스처 인식 기술이 대중화, 보편화되려면 아래와 같은 몇 가지 사항을 해결해야 한다.

#### ① 제품 사용의 편리성

접촉식 방식은 사용자에게 사용의 불편함을 주기 때문에 이를 대체할 수 있는 비접촉식 또는 웨어러블 센서의 발전이 필요

#### ② 제스처 인식 기술의 정확성

제스처 인식률의 정확도를 높여 제스처 인식 기술의 적용 분야를 다양화할 수 있도록 노력



〈그림 14〉 NUI 통합 플랫폼의 예[25]

## 이 필요

### ③ 다양한 환경에서의 적용 가능

모션 데이터 생성, 스마트 폰 터치 등 일부 제약적인 환경에서 적용 가능한 것을 다양한 환경에서 사용할 수 있도록 기술 개발이 필요

### ④ 멀티 모달 간 융합

다양한 센서간의 융합 및 다양한 모션 인식 기술 간의 융합, 음성 등 다른 인터페이스 방식과의 융합이 필요

### ⑤ NUI를 위한 융합 플랫폼의 발전

현재의 제스처 인식 기술 개발을 위해서는 센서의 개발사에서 제공하고 있는 SDK를 이용하는데 제조사간 사용하는 방식이 다른 문제가 발생하기 때문에 이에 대한 통합된 플랫폼이 필요하며, 〈그림 14〉는 통합 NUI 플랫폼의 예를 보여줌

## V. 결론

제스처는 사람이 무의식적으로 사용하는 의사전달 도구이다. 사람이 놀라거나 특별한 상황에 접하면 음성 및 제스처로 전달되기 때문이다. 그만큼 제스처는 인간의 자연스러운 행위이며 훌륭한 의사전달 수단이다. 따라서, 제스처를 공연 및 방송 분야에서도 사용하는 사례가 늘어나고 있으며, 이를 활용하고자 하는 노력도 지속적으로 증가할 것으로 기대된다.

제스처 인식 기술의 정확도를 향상시키는 것과 함께 이를 활용할 수 있는 자연스러운 인터페이스를 설계하는 것도 매우 중요하지만, 이를 활용할 수 있는 서비스를 개발하는 것도 중요하다. 현재 제스처 인식 기술이 가장 활발히 적용되고 있는 게임 또는 기기 제어 분야를 벗어나 수술실에서 의사의 손동작을 인식하여 수술을 대신하는 분야 등 그 응용 분야는 무궁무진할 것으로 보인다.

### 참고문헌

- [1] <http://www.apple.com/kr/ios/siri/>
- [2] <https://atap.google.com/soli/>
- [3] 유문우, “동작 인식을 위한 3D 센싱 기술”, 계장기술, No. 7, pp. 102-111, 2016
- [4] 정현태, “웨어러블 디바이스를 이용한 제스처 인식 기술 동향”, 전자공학회지, 42(6), pp. 56-62, 2015
- [5] 안세영, “3D 센서 기술(동작인식)”, ETRI 기술보고서, 2013.
- [6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097-1105, 2012.
- [7] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lars Wolf. Deep-face: Closing the gap to human-level performance in face verification. In Computer Vision and Pattern Recognition (CVPR), 2014
- [8] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In Advances in Neural Information Processing Systems, pages 1799-1807, 2014.
- [9] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pages 1653-1660. IEEE, 2014.
- [9] <http://en.wikipedia.org/wiki/Wii>
- [10] <https://namu.wiki/w/3D%20터치>
- [11] <https://www.leapmotion.com>
- [12] <http://www.intel.com/content/www/us/en/architecture-and-technology/realsense-overview.html>
- [13] <http://www.tobii.com>
- [14] <https://www.myo.com>
- [15] 류한석, “동작인식 기술 트렌트 및 시사점”, 디지에코 보고서, 2015
- [16] 조성식, 석희일, 이성환, “로봇 비전 기반 제스처 인식 연구 동향”, 로봇학회지, 7(1), pp. 16-23, 2010.
- [17] [http://www.bmw.com/com/en/newvehicles/7series/sedan/2015/showroom/innovative\\_functionality.html](http://www.bmw.com/com/en/newvehicles/7series/sedan/2015/showroom/innovative_functionality.html)
- [18] <http://cats.gatech.edu/content/copycat>
- [19] <http://www.exile.at/ko/installations.html>
- [20] <http://www.mogencelab.com>
- [21] <http://www.xsens.com>
- [22] <http://cyberith.com/>
- [23] 김지섭, 남장군, 장병탁, “딥러닝 기반 비디오 분석 기술”, 정보과학회지, 33(9), pp. 21-31, 2015.
- [24] <https://tla mpi.nl/tools/tla-tools/elan/>
- [25] <http://www.slideshare.net/JonghoonSeo/20-24091955>

### 필자소개



#### 양희덕

- 1998년 : 충남대학교 컴퓨터과학과 이학사
- 2003년 : 고려대학교 컴퓨터학과 이학석사
- 2008년 : 고려대학교 컴퓨터학과 이학박사
- 2006년 : Boston University Department of Computer Science 방문연구원
- 2012년 : EPFL Institute of Bioengineering 방문교수
- 현재 : 조선대학교 전자정보공과대학 컴퓨터공학과 부교수
- 주관심분야 : Human Computer Interaction, 영상검색, 의료영상처리