# 심층 학습을 이용한 실시간 한국 수화 인식 시스템*

강지연[O1] 육성현[2] 안이삭[3] 황정환[3] 이은규[3]
[1]서울시립대학교 컴퓨터과학부
[2]서울시립대학교 수학과
[3]인천대학교 정보통신공학과
kkkkk317@uos.ac.kr, chunhwa21@uos.ac.kr, isaac@inu.ac.kr, 0@inu.ac.kr, eklee@inu.ac.kr

# Real-Time Korean Sign Language Recognition Based on Deep Learning

Ji-Yeon Kang[O1] Sung-Hyun Yook[2] I-Saac Ahn[3] Jung-Hwan Hwang[3] Eun-Kyu Lee[3]
[1]Department of Computer Science and Engineering, University of Seoul
[2]Department of Mathematics, University of Seoul
[3]School of Information Technology, Incheon National University

**요 약**

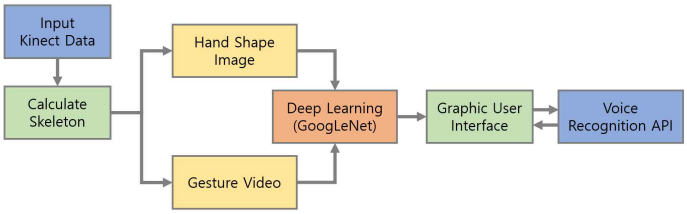
본 논문에서는 심층 학습(Deep Learning)을 이용한 실시간 수화 인식 시스템을 제안한다. 제안된 시스템은 데이터 저장, 수화 인식 그리고 음성 인식으로 구분된다. 데이터 저장은 키넥트의 관절 추출 기능으로 오른손과 상체 위치를 크롭(crop)하여 지화는 이미지로, 동작은 영상으로 저장한다. Matlab에서 GoogLeNet으로 심층학습을 진행하여 신경망을 생성한다. 수화 인식에서는 실시간 영상을 입력받고, 학습한 신경망을 통해 도출된 결과를 텍스트로 화면에 표시한다. 음성 인식에서는 음성 인식 API를 이용하여 텍스트를 음성으로 송출하고, 상대방의 음성을 사용자에게 텍스트로 보여준다. 지화 인식은 기존 방법인 기울기 히스토그램(Histogram Of Gradient)과 서포트 벡터 머신(Support Vector Machine)을 결합한 모델보다 인식 정확도가 41% 개선되었고, 동작 인식은 다양한 장소 및 각도로 촬영하면서 50%의 정확도가 증가하였다.

## 1. 서 론

수어로만 대화하는 농인이 의사소통에 어려움을 겪는 주된 이유는 수화통역사 수의 부족, 한글을 습득하지 않은 농인, 비장애인의 수화 습득의 한계이다. 본 연구에서는 청각장애인을 위한 IT 서비스를 개발하여 이런 요인을 해결하고, 일상에서 대화를 보조하고자 하였다. 따라서 수화를 번역해 음성과 텍스트로 출력하고, 한국 수화를 사용하는 농인이 새로운 수어를 직접 추가할 수 있는 시스템을 제작하였다.

수화는 크게 2가지로 구성된다. 음소를 나타내는 지화와 단어를 표현하는 동작이다. 수화 데이터 취득을 위해서 피부 색상과 화소를 이용하여 손을 검출한다[1]. 그러나 피부와 비슷한 색상의 옷이나 빛의 세기에 따라 검출 오차가 생기기에 키넥트의 관절 추출로 좌표를 저장한다.

본 논문에서는 실시간 수화 인식 시스템을 제안한다. [그림1]은 시스템 전체 흐름도로 데이터 저장, 수화 인식 그리고 음성 인식으로 구분한다. 데이터 저장에서는 키넥트의 관절 추출 기능으로 오른손과 상체 위치를 크롭(crop)한다. 지화는 이미지로 동작은 영상으로 저장하고, GoogLeNet[2]에서 심층 학습을 진행한다. 학습의 결과로 신경망이 생성된다. 이미지는 유방향 비순환 그래프 신경망(Directed Acyclic Graph Network) 객체를 생성한다. 영상은 각 프레임마다 추출한 특징을 장기간 단기 기억(Long Short-Term Memory)[3]으로 학습하여 시계열 신경망(Series Network) 객체를 생성한다. 수화 인식에서는 실시간 영상을 신경망에 입력해 가장 높은 확률값의 클래스를 화면에 표시한다. 마지막으로, 음성 인식 부분에서는 음성 인식 API를 이용해 화면에 표시된 텍스트를 음성으로 송출하고, 상대방의 음성을 텍스트로 나타내 사용자가 내용을 확인한다.




**[그림 1] 수화 인식 시스템 전체 흐름도**

지화에서는 기울기 히스토그램(Histogram Of Gradient)[4]과 서포트 벡터 머신(Support Vector Machine)을 결합한 기존 손 제스처 인식 모델과 비교하였을 때, 41%의 인식 정확성이 개선되었다. 동작 인식은 휴대성을 고려하여 여러 장소 및 다각도로 촬영한 데이터를 학습하였고, 단일 장소보다 50%의 성능 향상을 보였다.

## 2. 관련 연구

수화는 제스처의 일부로 볼 수 있다. 손 제스처 인식에 관한 연구에는 다양한 방법이 사용되었다. 이미지 특징 추출에서 가장 많이 사용되는 기법은 기울기 히스토그램(Histogram Of Gradient)이다. 이미지의 픽셀마다 기울기 방향을 표시하여 각 기울기의 개수를 히스토그램으로 표시한 것이다. 그러나 윤곽을 특징으로 표현하여 손가락 하나와 두 개를 붙인 지화를 분류하지 못한다. 움직임을 표현하는 기법에는 여러 히스토그램을 결합한 다각도 결합 히스토그램(Combined Angle Histogram)이 있다. 하지만 시간의 흐름에 따른 연관성을 특징에 포함하지 않는다.

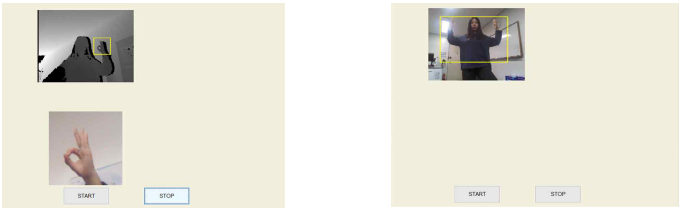
* 본 논문은 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 프로보노 ICT멘토링 프로젝트 결과물임.

데이터의 연속성을 특징으로 사용하는 인식 모델도 제안되었다. 은닉 마르코프 모델(Hidden Markov Model)[5]은 제스처의 시작과 끝을 검출하여 각 상태가 순서 구조를 갖는다. 동적 시간 정합(Dynamic Time Warping)[6]에서는 시계열 데이터의 유사도를 측정하며, 유클리디안 거리가 최소가 되도록 점을 매칭해 궤적을 분리한다. 하지만 이 두 모델은 모든 경우를 검사하기에 속도가 느리다는 단점이 있다.

## 3. 제안 방법

### 3.1 데이터 수집 단계

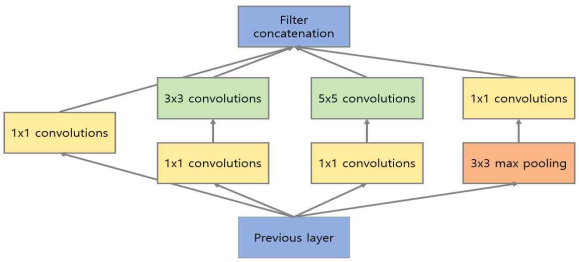
본 논문에서 사용된 입력 데이터는 오른손 이미지와 상체 영상이다. 지화 인식에서는 키넥트 센서로부터 획득한 오른손의 위치를 기준으로 이미지를 크롭(crop)한다. 오른손의 위치가 검출되면 0.25초 간격으로 이미지가 저장된다. 동작 인식에서는 키넥트 센서로부터 획득한 어깨의 중심의 위치를 기준으로 영상을 크롭한다. 상체 위치가 검출되면 3초 동안 30 프레임으로 영상 파일이 저장된다. 지화는 31개의 클래스가 존재하며 각각 150개씩 이미지를 저장하였다. 동작에서는 5개의 클래스가 있고, 각각 100개의 영상을 촬영하였다.




[그림 2] 데이터 저장 화면
(좌: 자음 'ㅇ' 지화 이미지, 우: '덥다' 동작 영상)

### 3.2 모델 학습 단계

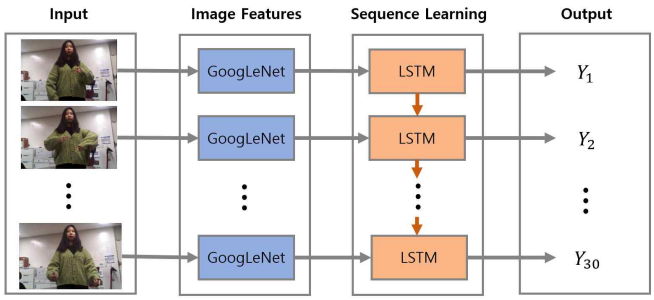
이미지의 다양한 특징을 얻을 수 있는 GoogLeNet[2]을 사용하였다. [그림 3]과 같은 구조로, 여러 크기(scale)의 합성곱 필터(Convolution filter)를 적용하여 서로 다른 크기의 특징을 얻는다. 또한, 1x1 합성곱이 차원을 줄여 망이 깊어져도 연산량이 적다는 장점이 있어 GoogLeNet을 모델로 결정하였다.




[그림 3] GoogLeNet의 Inception module

지화 학습을 위해, 이미지를 224 * 224 * 3 크기로 합성곱 인공신경망(Convolutional Neural Network)인 GoogLeNet에 입력한다. 새로운 클래스를 분류하도록 마지막 층을 변경하여 학습을 진행한다. 모든 과정이 완료되면, 유방향 비순환 그래프 신경망(Directed Acyclic Graph Network) 객체가 생성된다.

동작 데이터는 GoogLeNet과 장기간 단기 기억(Long Short-Term Memory)을 결합한 모델로 심층 학습을 한다[7]. 동영상의 각 프레임마다 특징을 추출하기 위해서 이미지 크기가 224 * 224 * 30 인 픽셀로 형성한다. 이를 GoogLeNet에 입력하고, 연속적인 벡터 구조로 변경한다. 해당 벡터를 장기간 단기 기억 네트워크에서 학습하여 출력층으로 전달하면, 시계열 신경망(Series Network) 객체가 생성된다.

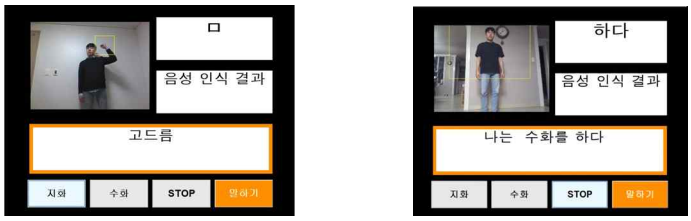



[그림 4] 영상 데이터의 심층 학습 모델 구조

### 3.3 실시간 수화 인식 단계

수화 인식 단계에서는 키넥트에서 실시간으로 받아온 영상으로 학습 결과의 정확성을 판단한다. 모델 학습 단계에서 생성된 신경망 객체에 테스트 데이터를 입력한다. 확률이 가장 높은 클래스를 화면의 하단에 텍스트로 표시한다.

지화 버튼을 클릭하면 오른손의 위치를 추출한다. 수화 버튼을 클릭하면 상체 위치를 인식하고, 3초간 촬영된다. 데이터를 저장할 때처럼 해당 부분을 크롭(crop)한 후, 생성한 신경망 객체의 입력값으로 넣어 출력값을 도출한다. 음소는 결과가 합쳐져 단어로 표현되고, 단어는 문장으로 구성된다.




[그림 5] 실시간 수화 인식 화면(좌: 지화, 우: 동작)

### 3.4 음성 인식(Voice Recognition) 단계

수화를 번역한 음소 및 단어를 음성으로 송출하기 위해 음성 합성 시스템(Text-to-Speech)을 사용하였다. [그림 5]처럼 지화와 동작을 인식하여 화면 하단에 텍스트로 결과가 표시되면, 음성 합성 시스템을 통해 기기에서 음성으로 송출한다.

음성을 사용자에게 텍스트로 보여주기 위해 음성 인식 시스템(Speech-to-Text)을 사용하였다. 상대방이 음성으로 대화를 할 시에 말하기 버튼을 클릭한다. 기기에 대고 말하면, 음성 인식 시스템을 통해 음성이 '음성 인식 결과' 부분에 텍스트로 나타난다.


[그림 6] 음성 인식 결과 화면

## 4. 실험 결과

### 4.1 실험 환경(키넥트와 Matlab)

키넥트는 내부 카메라 모듈(RGB, 깊이, 적외선)을 통해 관절 20개의 위치를 검출한다. Kinect for Windows v1을 사용하였기에 Kinect for Windows SDK v1.8을 설치한다.

Matlab에서 총 8개의 부가기능을 설치한다. 키넥트를 이용하기 위해 Image Acquisition Toolbox Support Package for Kinect for Windows Sensor, Image Acquisition Toolbox, Computer Vision Toolbox를 설치한다. GoogLeNet을 사용하기 위해 Deep Learning Toolbox, Deep Learning Toolbox Model for GoogLeNet Network을 설치한다. 마지막으로, 음성 인식을 위해서 Audio Toolbox, speech2text, text2speech를 설치한다.

### 4.2 실험 결과

제안하는 실시간 수화 인식 시스템의 성능을 입증하기 위해 기존 방식과 인식 정확도를 비교하였다. 해당 정확도는 수화자 4명이 데이터에 포함되지 않은 장소 2군데에서 수화를 10번 수행한 결과의 평균이다.

**[표 1] 지화 인식 정확도 결과**

| 음소 | ㄱ | ㄴ | ㄷ | ㄹ | ㅁ | ㅂ | ㅅ | ㅇ | ㅈ | ㅊ | ㅋ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HOG +SVM | 0.96 | 0.05 | 0.56 | 0.97 | 0.86 | 0.52 | 0.30 | 0.52 | 0.36 | 0.08 | 0.97 |
| Goog LeNet | 1.00 | 1.00 | 1.00 | 0.90 | 1.00 | 1.00 | 1.00 | 1.00 | 0.75 | 1.00 | 1.00 |

| 음소 | ㅌ | ㅍ | ㅎ | ㅏ | ㅑ | ㅓ | ㅕ | ㅗ | ㅛ | ㅜ | ㅠ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HOG +SVM | 0.96 | 0.09 | 0.82 | 0.96 | 0.75 | 0.07 | 0.64 | 0.47 | 0.50 | 0.30 | 0.28 |
| Goog LeNet | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.15 | 1.00 | 1.00 | 1.00 | 0.80 |

| 음소 | ㅡ | ㅣ | ㅐ | ㅒ | ㅔ | ㅖ | ㅢ | ㅚ | ㅟ | 평균 |
|---|---|---|---|---|---|---|---|---|---|---|
| HOG +SVM | 0.92 | 0.95 | 0.72 | 0.34 | 0.66 | 0.23 | 0.48 | 0.33 | 0.39 | **0.54** |
| Goog LeNet | 1.00 | 1.00 | 1.00 | 1.00 | 0.75 | 0.95 | 1.00 | 1.00 | 1.00 | **0.95** |

지화 인식에서는 손 제스처 인식에서 사용되는 기울기 히스토그램(Histogram Of Gradient)과 서포트 벡터 머신(Support Vector Machine)을 결합한 모델을 기존 방식으로 선정하였다. [표 1]에서 기존 방식에 비해 지화 인식 정확성이 평균 41%가 개선되었으며, 각 음소에서도 대부분 인식률이 증가하였다. GoogLeNet이 다양한 특성을 추출했기 때문이라고 볼 수 있다.

동작 인식에서는 손의 위치 변화와 연속된 정보를 이용하기 위해 기울기 히스토그램을 사용하지 않았다. 그러나 농인이 어떤 상황에서 수화를 하더라도 인식할 수 있도록 고려해야 하였다. 그리하여 단일 장소에서 촬영한 영상으로 학습한 신경망을 기존 방식으로 선정하였다. 이를 4군데 장소 및 키넥트 각도를 변경하여 촬영한 데이터와 성능을 비교하였다.

**[표 2] 동작 인식 정확도 결과**

| 단어 | 안녕하세요 | 가다 | 먹다 | 달리다 | 덥다 | 평균 |
|---|---|---|---|---|---|---|
| 단일 장소 | 0.50 | 0.00 | 0.80 | 0.60 | 0.10 | **0.40** |
| 장소 및 각도 변경 | 0.80 | 0.90 | 0.80 | 1.00 | 1.00 | **0.90** |

[표 2]에서 동작 인식 정확도가 기존 방식에 비해 평균적으로 50% 증가한 것을 볼 수 있다. '먹다'를 제외한 모든 단어에서 인식률이 높아졌으며, 다양한 조건을 포함한 데이터로 학습된 신경망의 오류 비율이 낮다는 것을 알 수 있다.

## 5. 결론 및 향후 연구

본 논문에서는 심층 학습(Deep Learning)을 이용한 실시간 수화 인식 시스템을 제안하였다. 지화에서는 기울기 히스토그램(Histogram Of Gradient)과 서포트 벡터 머신(Support Vector Machine)을 결합한 모델과 비교하였을 때, 41%의 인식 정확도가 개선되었다. 동작에서는 단일 장소보다 장소와 각도를 변경해 데이터를 생성한 결과, 50%의 성능 향상을 보였다. 이 결과는 기존 손 제스처 인식 모델보다 심층 학습이, 단일보다 다양한 조건의 데이터 확보가 수화 인식에 효율적이라는 사실을 보여준다. 본 연구는 수화를 사용하는 청각장애인과 비장애인의 소통 문제 해결에 중요한 지표가 될 것으로 기대된다.

추후 해결해야할 과제로는 첫째, 'ㅕ', 'ㅖ'와 같이 키넥트 정면을 가리키는 음소에서 편향된 인식률을 보인 것이다. 따라서 정면에서 손 모양을 구분할 수 있는 새로운 알고리즘이 연구되어야 할 것이다. 둘째, 동작 인식 정확도를 높이기 위해서는 더 많은 수화 데이터를 확보하여야 한다. 셋째, 한 사용자가 신경망을 학습하면 다른 사용자의 신경망 객체가 자동으로 업데이트(update)되지 않는다는 점이다. 학습한 신경망을 전달하면 다른 사용자도 새로운 클래스를 분류할 수 있다는 장점이 있으나 수작업이 필요하다. 그러므로 생성한 신경망을 클라우드(cloud)에 연결하여 데이터 학습을 진행해야 할 것이다.

## 참 고 문 헌


[1] H. Park, C. Bae, "Face and Hand Tracking Algorithm for Sign Language Recognition," *The Journal of Korean Institute of Communications and Information Sciences*, Vol.31, No.11C, pp. 1071-1076, 2016. (in Korean)

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going Deeper With Convolutions," *The IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 1-9, 2015.

[3] S. Hochreiter, J. Schmidhuber, "Long short-term memory," *Neural computation*, Vol.9, No.3, pp.1735-1780, 1997.

[4] S. Cho, H. Byun, H. Lee and J. Cha, "Hand Gesture Recognition from Kinect Sensor Data," *Journal of Broadcast Engineering*, Vol.17, No.3, pp.447-457, 2015. (in Korean)

[5] H. Yang, A. Park, and S. Lee, "Gesture Spotting and Recognition for Human-Robot Interaction," *IEEE Trans. on Robotics*, Vol.23, No.2, pp.256-270, 2007.

[6] P. Doliotis, A. Stefan, C. McMurrough, D. Eckhard, and V. Athitsos, "Comparing Gesture Recognition Accuracy Using Color and Depth Information," *PETRA*, Vol.4, No.20, pp.1-7, 2011.

[7] J. Donahue, L. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko and T. Darrell, "Long-term Recurrent Convolutional Networks for Visual Recognition and Description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.39, No.4, pp.677-691, 2017.