MediaPipe를 활용한 수어 번역 시스템 개발

김경민[1], 송미화[2]
[1]세명대학교 정보통신학부, [2]세명대학교 스마트IT학부

Sign Language Translation System Development Using MediaPipe

Kyung-Min Kim[1], Mi-Hwa Song[2]
[1]School of Information and Communication Science, Semyung University,
[2]School of Smart IT, Semyung University

요 약

다양한 언어로 소통하고 있는 우리는 다른 언어와 교류하기 위해 번역, 통역의 존재가 필수가 되기도 한다. 하지만 음성언어를 사용하지 않는 즉, 손으로 언어를 표현하는 수어를 번역하는 통역의 존재는 아직 실현되지 않았다. 이에 본 논문에서는 MediaPipe와 OpenCV 라이브러리를 이용하여 손의 형태를 인식하고 CNN 알고리즘을 통한 텍스트 데이터화 하여 수어 동작을 학습시켜 이를 번역시켜주는 시스템을 연구한다. 이를 통해 공공기관을 이용함에 불편함을 줄이고, 농인의 의사를 보다 빠르게 파악할 수 있도록 도와주는 번역 시스템 제작하는 것에 목적이 있다.

1. 서론

세상에는 다양한 언어가 존재한다. 국가 내에서도 다양한 언어로 소통을 하며 교류하고 있는 세상을 살아가는 우리는 다른 언어를 모른다면 소통에 큰 어려움이 생길 것이다. 하지만 이를 완화 시키고 소통에 도움이 되는 편리한 방법이 있다. 바로 통역으로 다른 언어와 소통할 수 있는 수단이 된다. 하지만 통역의 존재를 음성 언어 만이 아닌 청각장애, 언어장애 등 농인들이 사용하는 수어의 통역을 제공하는 시스템을 제안한다.

행정안전부의 보건복지부에서 집계한 등록청각언어장애인은 2021년 약 435,000명으로 전체 등록장애인(약 2,588,000명)의 약 17%를 차지하고 있으며, 이는 전체 등록장애인 중 두 번째로 많은 수치이다. 해서 이번 프로젝트를 통해 수어를 통역하게 된다면 농인[1]이 공공기관을 이용하는데 있어 불편함을 줄여주고, 수어로 농인들의 의사를 빠르게 전달하지 못한다는 불편함을 통역의 존재로 편리함을 주고자 한다.

본 논문에서는 OpenCV 라이브러리와 MediaPipe, TensorFlow를 이용하여 손과 손가락을 트래킹하여 모션의 형태를 인식하며, CNN 기법을 이용하여 모션의 형태를 텍스트 형태의 데이터로 변환하여 TensorFlow를 통해 영상 학습과 텍스트 학습을 진행하여 농인과의 의사소통을 번역하는 시스템을 개발한다[1,2].

(그림 1) 연도별 등록장애인 추이

2. 관련 연구

2.1 OpenCV

OpenCV(Open Source Computer Vision)는 실시간 컴퓨터 비전을 목적으로 한 프로그래밍 라이브러리이다. 인텔이 개발한 실시간 이미지 프로세싱에 중점을 둔 오픈소스 라이브러리이다. 이번에 사용되는 MediaPipe와 TensorFlow의 딥러닝 프레임워크를 지원하며 기반이 될 라이브러리가 된다.

1) 농인 : 청각에 장애가 있어 소리를 듣지 못하는 사람. 주로 수어로 의사소통을 한다.

2.2 MediaPipe

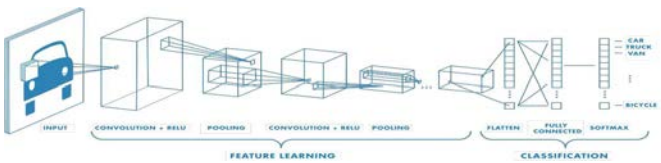
MediaPipe는 Google에서 제공하는 AI 프레임 워크이다. 비디오 형식 데이터를 이용한 다양한 비전 AI 기능을 파이프라인 형태로 손쉽게 사용할 수 있도록 제공된 오픈소스 크로스 플랫폼 프레임워크이다. 이번에 사용되는 것은 MediaPipe의 Hand Tracking 부분으로 ML을 사용하여 손에 21개의 Landmark를 추출하여 손을 트래킹 및 추출에 사용된다.

2.3 TensorFlow

TensorFlow는 다양한 작업에 대해 데이터 흐름 프로그래밍을 위한 오픈소스 소프트웨어 라이브러리로 심볼릭 수학 라이브러리이자, 인공 신경망 같은 기계 학습 응용프로그램 및 딥러닝에 사용된다. 본 연구에서는 MediaPipe에서 추출된 텍스트 데이터와 영상데이터의 학습을 통해 번역을 시키는 판단을 시킨다.

2.4 CNN

CNN(Convolutional Neural Network)는 이미지를 분석하기 위해 패턴을 찾는데 유용한 알고리즘으로 데이터에서 이미지를 직접 학습하고 패턴을 사용해 이미지를 분류한다. CNN은 이미지의 공간정보를 유지하며 학습하며 고도로 정확한 인식 결과를 생성한다. 이번 ML에 사용되는 주 알고리즘이다[3].



(그림 2) CNN의 구조

2.5 수어

수어는 손짓을 포함한 다양한 시각적인 정보를 매개로 의사를 전달하는 언어이다. 손, 손가락 그리고 팔로 그리는 모양, 상황을 올바르게 전달하는 표정이나 입술의 움직임 등을 종합하여 뜻을 전달하는 의사소통의 수단이다. 본 논문에서는 영어 지문자 통역을 통해 시스템을 개발한다.

3. 시스템 설계 및 구현

3.1 대상 수어

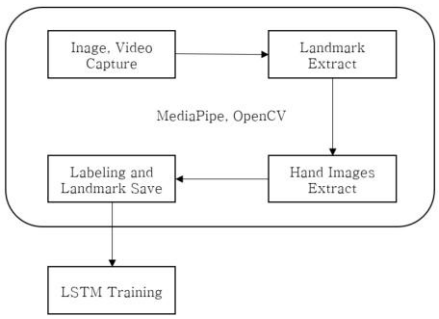
우리가 언어를 배우기 전에 자음과 모음을 배우는 것과 마찬가지로 수어를 배우기 위해서는 지문자를

먼저 배우게 된다. 지문자는 수어의 자음과 모음으로 본 연구에서는 영어 지문자 통역을 주를 이루어 진행된다[4, 5]

(그림 3) 영어 지문자

3.2 시스템 모델

기계학습을 위한 학습 모듈의 구조는 그림4와 같다.

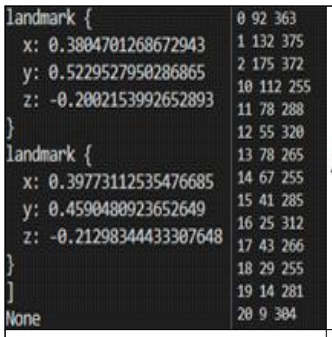


(그림 4) 학습 모델

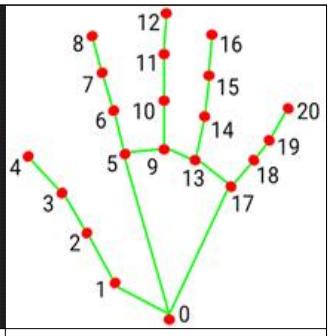
3.3 시스템 구성

카메라를 통한 이미지, 영상을 OpenCV를 통해 지속적으로 프레임 단위로 받아오게 된다. 이와 동시에 MediaPipe에서는 표 1의 코드를 통해 손바닥의 추적과 Landmark를 잡게 되며, Landmark는 MediaPipe의 내부에서 2D 형상 모듈 내에 설정된 계측 3D 공간을 통해 2D Landmark 화면 좌표가 미터법 3D 공간 좌표로 변환된다. Hand 변환 메트릭스는 둘 사이의 차이를 최소화하는 방식으로 runtime Hand 메트릭 Landmark 세트로 설정된 표준 메트릭 Landmark에서 견고한 선형 mapping으로 추정한다. 2D 메시는 runtime 2D 메트릭 Landmark를 정점 위치(x,y,z)로 사용하여 생성되며, 정점 텍스처 좌표(UV)와 삼각형 위상은 모두 표준 2D 모델에서 상속된다. 그 개념을 가지고 손의 landmark와 손 영역 데이터를 이용하여 모든 프레임에 대해서 반복적인 학습을 진행하며 텍스트 데이터 형태의 Dataset을 생성한다. 그림 5와 그림 6은 각 landmark 영역과 손가락의 데이터 값과 MediaPiped

의 손가락 포인트이다. 이를 통하여 알파벳 지문자 A ~ Y까지 학습을 시킨다.

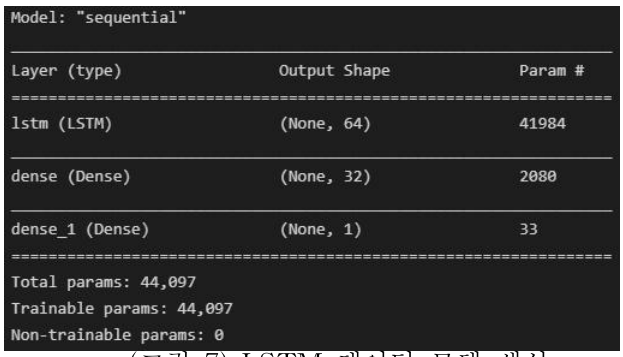


(그림 5) Landmark 영역



(그림 6) Hand Landmarks

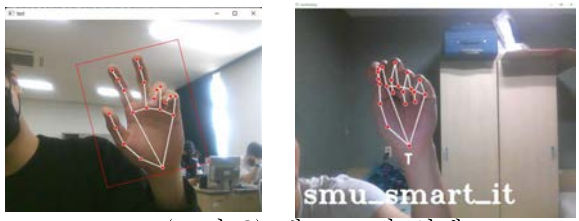
지문자 Z는 검지를 통하여 알파벳 Z의 형태로 그려나가기 때문에 MediaPipe의 고정된 좌표값을 가지고 학습시키기는 방식에 옳지 않다. 움직이는 좌표에 대한 학습은 LSTM을 통하여 학습을 진행한다. 이 방식은 30초 동안의 행동을 MediaPipe의 텍스트 데이터를 통해 포인터 값을 이어 붙여 학습을 진행하게 된다. 그림 7은 위와 같은 방식으로 100개의 데이터를 통해 LSTM 학습을 진행하였다.



(그림 7) LSTM 데이터 모델 생성

3.4 테스트 및 실행

생성된 데이터 5000개를 Dataset에 저장하여 실시간 Video 프레임과 비교하며 영어 지문자 수어에 대한 테스트를 진행한다. 해당되는 영어 지문자에 대한 수어를 표시하면 화면에 출력하며 번역방식을 확인할 수 있다.



(그림 8) 테스트 및 실행

4. 결론

본 논문에서는 MediaPipe를 활용하여 수어를 번역하는 시스템을 개발하였다. 수어에 대한 관심도가 높아짐에 따라 수어의 인식이 많이 개선되었지만, 아직 수어를 사용하는 사람과 대화를 원활하게 하기에는 어렵다. 이러한 불편한 점을 번역 도구로써 사용되기를 원하여 번역 시스템을 제시하고자 하였다. MediaPipe를 통해 CNN 알고리즘을 이용해 손에 대한 Landmark를 추출하고, 시각화하여 텍스트 데이터를 학습 및 판단하여 번역해줌으로써 의미있는 결과를 도출하였다.

향후 연구에서는 다음 사항을 고려한 추가 연구 및 보완이 필요할 것이다. 보완점은 데이터 학습의 불균형으로 더 큰 동작을 통해 데이터 전처리를 진행해야 한다. 추가 연구로는 상반신의 전체 인식과 얼굴의 표정까지 학습을 시켜 실제로 많이 사용되는 단어를 표기하여 하나의 문장을 만드는 시스템을 개발하여 통역에 좀 더 가까운 형태로 진화할 예정이다.

참고문헌

[1] 김진영(Jin-Young Kim),and 심현(Hyun Sim). "청각장애인의 수어 교육을 위한 MediaPipe 활용 수어 학습 보조 시스템 개발." 한국전자통신학회 논문지 16.6 (2021): 1355-1361.

[2] 박형모. "로봇팔을 이용한 머신러닝 기반의 스팟용접부 크랙탐지." 국내석사학위논문 인천대학교 일반대학원, 2021. 인천

[3] 허이룬. "합성곱 신경망(CNN)기반 이미지 처리 시스템." 국내박사학위논문 배재대학교, 2018. 대전

[4] Paweł Rutkowski,and Sylwia Łozińska. "Argument Linearization in a Three-Dimensional Grammar: A Typological Perspective on Word Order in Polish Sign Language (PJM)." Journal of Universal Language 17.1 (2016): 109-134.

[5] Rahib H Abiyev, Murat Arslan, and John Bush Idoko, "Sign Language Translation Using Deep Convolutional Neural Networks." KSII Transactions on Internet and Information Systems(TIIS) 14.2 (2020): 631-653.

[6] Yejin Kwon, Dongho Kim.(2022). Real-Time Workout Posture Correction using OpenCV and MediaPipe. 한국정보기술학회논문지, 20(1), 199-208.

[7] Google, “MediaPipe Hands”, © GOOGLE LLC https://google.github.io/mediapipe/solutions/hands#example-apps