JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 24, No. 1, pp. 111-119, Jan. 2023

Check for updates

미디어파이프와 장단기기억을 이용한 수화 동작인식 앱 개발

김 신 영[1] · 엄 서 정[1] · 유 선 영[1] · 김 수 정[1] · 이 경 미[2*]
[1]덕성여자대학교 컴퓨터공학과 학부과정
[2*]덕성여자대학교 컴퓨터공학과 교수

Application of sign language gesture recognition using Mediapipe and LSTM

Shin-Yong Kim[1] · Seo-Jung Urm[1] · Sun-Young Yoo[1] · Soo-Jeong Kim[1] · Kyoung-Mi Lee[2*]

[1]Bachelor's Course, Department of Computer Science, Duksung Women's University, Seoul 01369, Korea
[2*]Professor, Department of Computer Science, Duksung Women's University, Seoul 01369, Korea

[요 약]

청각 장애인들의 의사소통을 위해 수화 동작인식 연구가 활발히 진행되었다. 그러나 대부분의 수화 동작인식 연구는 정지된 화면에서 손의 형태를 인식하는 방식으로 이루어졌다. 본 논문에서는 동적으로 변화는 동작뿐만 아니라 얼굴의 표정까지 포함하는 비수지 수화 동작인식 방법을 제안한다. 카메라를 통해 들어오는 매 프레임마다 미디어파이프(Mediapipe) 프레임워크를 사용해서 얼굴과 손을 포함한 인체의 랜드마크를 추출하고 비수지 신호 학습에 필요한 입력 특징값을 계산한다. 동적 동작인식을 위해 한 동작을 30개의 프레임으로 정의하고 순차 데이터 인식을 위한 장단기기억(LSTM) 모델을 이용하여 동적 수화 동작을 학습시킨다. 제안하는 동적 비수지 동작인식 방법은 수화 동작을 배우려는 사용자들을 위한 앱을 개발하는데 적용했다.

[Abstract]

Researches on gesture recognition for sign language have been actively conducted to communicate with the hearing impaired. However, most of them focused on a way that recognizes the shape of hands on a still image. In this paper, we propose a method that includes not only dynamically changing gestures but also facial expressions. For every frame that comes in through the camera, the Mediapipe framework is used to extract human-body landmarks, including the face and hands, and the proposed method calculates the input features required for non-manual signal learning. For dynamic gesture recognition, one gesture is defined as 30 frames, and dynamic sign language gestures are learned using the LSTM model for sequential data recognition. The proposed dynamic non-manual gesture recognition method is applied to developing applications for users who want to learn sign language gestures.

색인어 : 수화, 비수지 신호, 동작인식, 미디어파이프, 장단기기억

Keyword : Sign language, Non-manual signal, Gesture recognition, Mediapipe, Long short term memory

http://dx.doi.org/10.9728/dcs.2023.24.1.111

Received 16 November 2022; **Revised** 29 December 2022
Accepted 10 January 2023

***Corresponding Author; Kyoung-Mi Lee**

Tel: +82-2-901-8348
E-mail: kmlee@duksung.ac.kr

Ⅰ. 서 론

보건복지부에서 해마다 조사하는 시. 도 장애인등록현황에 따르면 2021년도 청각 장애로 등록한 장애인 수는 43.5만 명으로 전체 등록 장애인 264.5만 명의 16.45%를 차지하고 있다[1]. 연도별로 청각 장애 등록인은 2015년도 26.9만 명, 10.8%에서부터 꾸준히 그 등록인 수와 비율이 증가하고 있다(그림 1). 더욱이 노령화로 인해 60세 이상 노령층에서의 청각 장애 등록인 수가 다른 연령대에 비해 뚜렷하게 증가하여, 2021년도에는 36.7만 명으로 전체 청각 장애 등록인의 84.51%를 차지한다.

청각장애인들은 일상생활에서 외부와의 의사소통을 위해 수화(sign language)를 사용한다. 수화는 손을 이용하여 의사를 표현하는 일종의 시각 언어로 요즘은 수어(手語)라고도 한다. 수화는 다시 손의 모양, 위치, 움직임으로 나타내는 수지 신호와 얼굴의 표정, 머리와 몸의 움직임 등으로 나타내는 비수지 신호로 나뉜다[2]. 수지 신호, 특히 손의 형태만으로 의사소통이 가능할 수 있지만, 같은 손 모양이라 하더라도 얼굴 표정에 따라 뜻이 달라질 수 있고 얼굴로 감정을 표현할 수 있으므로 충분한 의사소통을 위해서 비수지 신호를 많이 사용하게 된다.

청각장애인과 소통하기 위해 수화를 배우고 통역하는 것은 매우 중요하며, 이를 도와주는 수화 동작인식 연구가 꾸준히 진행되어 왔다[3,4]. 최근 들어 딥러닝을 적용한 수화 동작인식 연구가 활발히 이루어지고 있다[5]. 특히 카메라로부터 취득한 영상을 이용한 수화 인식 연구의 대부분은 정지 영상으로부터 CNN 또는 DNN과 같은 딥러닝 신경망을 이용하여 수화동작을 인식하거나[6,7] 이미 학습된 인체 검출 오픈 소스인 OpenPose 또는 Mediapipe로부터 인체 랜드마크를 추출한 후 기계학습 또는 신경망을 이용하여 수화 동작을 인식한다[2,8].

오랫동안 다양한 수화 동작인식 연구가 이루어졌음에도 불구하고, 여전히 수화 동작인식 연구는 수화자의 외모 차이, 동일 동작의 움직임 크기의 차이에도 강건하게 인식해야하는 인체 동작인식 기술의 어려움과 연속된 동영상 스트림에서 동작의 시작과 끝을 찾고 동일 동작의 움직임 속도의 차이를 처리해야하는 동영상 처리 기술의 어려움을 포함하고 있다.

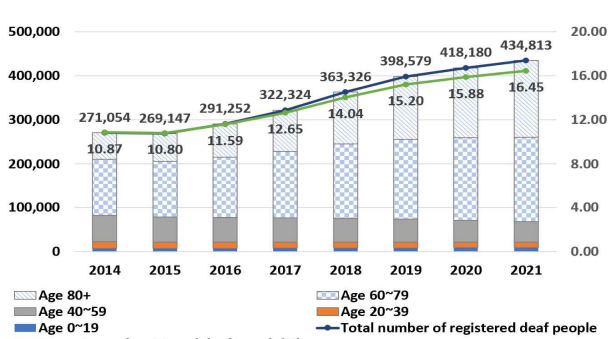


그림 1. 전국 청각 장애 연령별 등록 통계
Fig. 1. Statistics of registered deaf people nationwide

논문 [2]는 동적 비수지 신호에 대해 합성곱 신경망(CNN)으로 각 프레임 속 정지 동작을 인식하고 다수결 투표방식으로 동적 동작을 인식한다. 그러나 연속된 동영상 스트림이 아닌 분리된(isolated) 동영상을 대상으로 하므로 동영상으로 수화동작을 촬영하고 저장한 후에만 사용할 수 있어 실시간 사례에 적용하기 어렵다. 논문 [8]은 카메라를 통한 실시간 수화 동작인식이 가능하지만 각 프레임에서 정지된 손가락 형태만을 합성곱 신경망으로 인식할 뿐 얼굴 표정 등을 반영한 비수지 신호와 동적 수화 동작을 포함하지 않았다.

본 논문에서는 정적인 손가락 형태만 인식하는 수화 동작인식 연구의 한계를 극복하기 위해 연속된 동영상 스트림에서 동적으로 움직이는 손가락 동작과 얼굴 표정을 포함한 비수지 수화 동작을 인식하는 방법을 제안한다. 제안하는 방법은 미디어파이프로부터 얼굴 및 손, 상반신의 랜드마크를 추출하고 학습에 필요한 입력 특징값을 계산한다. 동적 동작인식을 위해 동작 시퀀스(sequence) 단위로 인식할 수 있는 장단기기억 모델을 학습시켜 수화 동작을 인식한다. 학습된 수화 동작 모델은 안드로이드 폰의 카메라로 들어오는 실시간 영상에서 수화 동작을 배우고 통역할 수 있는 앱 개발에 적용된다.

Ⅱ. 관련 연구

2-1 미디어파이프(Mediapipe)

미디어파이프는 비디오, 오디오 등과 같은 시계열 데이터를 처리하는 기계 학습 파이프라인을 구축하기 위한 프레임워크이다[9]. 최근에는 동영상 또는 실시간 영상으로부터 얼굴(Face Mesh), 동공(Iris), 손(Hands), 전신(Pose)을 각각 인식할 수 있을 뿐만 아니라 이 모두를 통합하여(Holistic) 인식할 수 있는 다양한 기계학습 솔루션을 제공한다. 미디어파이프는 이들 인식 솔루션을 안드로이드, iOS, C++, Python, JavaScript 등 여러 플랫폼에서 설치하고 추가로 사용자 정의(customize)하여 자신만의 인식 애플리케이션을 효율적으로 개발할 수 있도록 한다. 본 논문에서는 얼굴, 손, 상반신에 대한 정보가 필요하므로 통합하여 인식하는 미디어파이프 Holistic 솔루션을 이용한다.

2-2 장단기기억(Long Short-Term Memory) 모델

딥러닝에서 텍스트, 오디오 및 비디오와 같은 시계열 데이터를 학습시키는 대표적인 모델은 순환신경망(Recurrent Neural Network)이다. 순환신경망은 상위 층과 하위 층의 뉴런들 사이의 순환 연결과 선택적 자가 피드백 연결을 가지고 있다. 이런 피드백 연결을 통해 순환신경망은 이전 단계에서 현 단계로 데이터를 전파하는 방식으로 시계열 데이터의 메모리를 구축한다[10]. 그러나 순환신경망은 역전파되는 오류가

점점 줄어들어 학습에 거의 영향을 미치지 못하는 기울기 소멸 문제로 인해 5~10개 이상의 시간 단계를 연결할 수 없다[10].

이런 기울기 소멸 문제를 해결하기 위해 게이트 기능을 도입한 장단기기억 모델이 제안되었다[11]. 장단기기억 모델은 1,000개 이상의 개별 시간 단계의 최소 시간 지연을 연결하는 방법을 학습할 수 있다. 특수 셀 내에서 일정한 오류 흐름을 강제할 수 있으며, 셀에 대한 접근은 접근 권한을 부여할 시기를 학습하는 게이트 장치에 의해 처리된다. 본 논문에서는 장단기기억 모델을 이용하여 시계열 데이터인 비디오로부터 동적인 수화동작을 인식하는 방법을 제안한다.

Ⅲ. 제안하는 동적 비수지 수화 동작인식

3-1 대상 수화동작

본 논문에서는 '가족', '기차', '비빔밥', '안경', '얼굴', '여동생' 등의 명사와 '괜찮다', '만나다', '먹다', ' 앉다', '좋다' 등의 서술어로 이루어진 36개의 한국 수화 단어 동작을 대상으로 한다. 그림 2는 그 중 10개의 수화 동작 예시를 보여주고 있다[12]. 그림 2의 각 동작에 표시된 화살표는 움직이는 방향을 나타낸다.

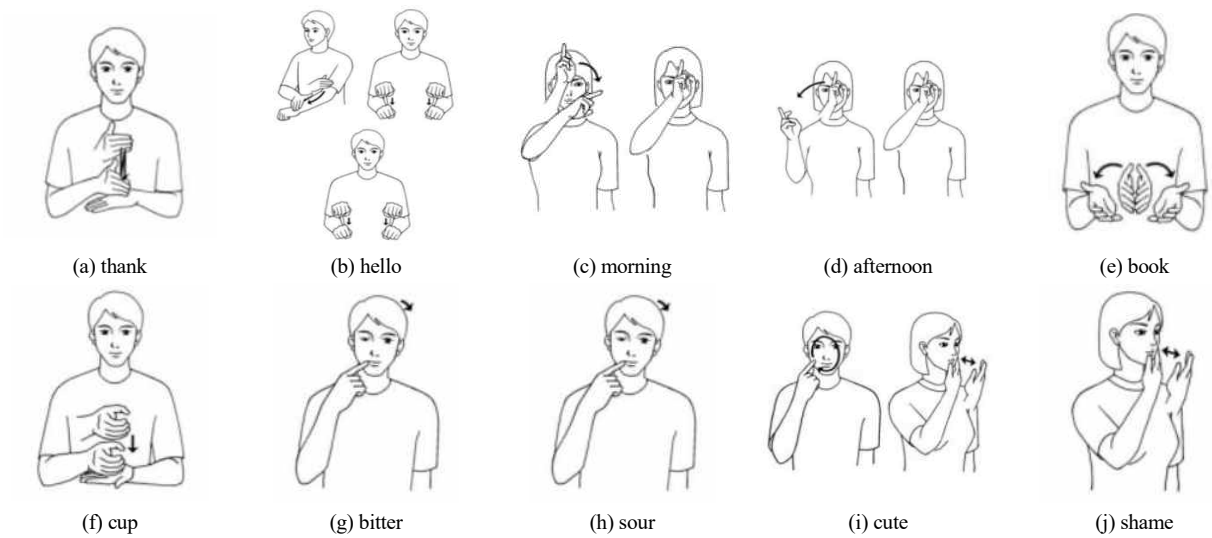


그림 2. 수어 예시[12]

Fig. 2. Examples of sign language[12]

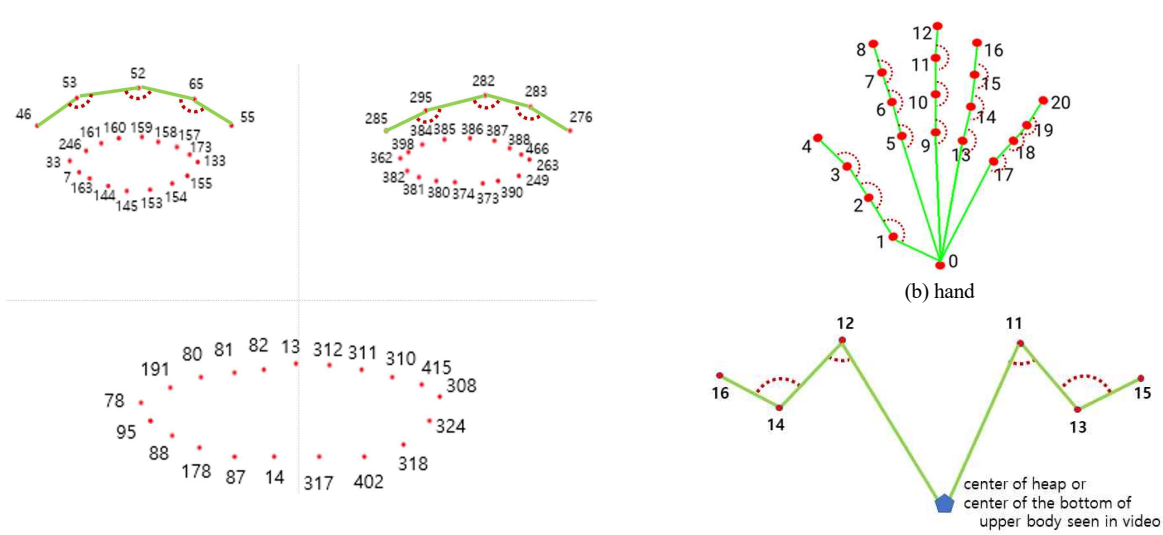


그림 3. 본 논문에서 사용된 랜드마크, 벡터, 그리고 관절 사이각. 점은 랜드마크를, 점의 숫자는 미디어파이프에서 배정된 번호를, 랜드마크 사이의 선은 벡터를, 점선 호는 관절 사이각을 의미한다.

Fig. 3. Landmarks, vectors, and joint angles used in this paper. Dots mean landmarks, numbers next to points mean ordinal numbers assigned by Mediapipe, lines between landmarks mean vectors, and dotted arcs mean joint angles.

그림 4. 동영상 프레임에서 랜드마크 추출 결과
Fig. 4. Results of detecting landmarks from video frames

그림 5. 30개의 프레임으로 구성된 '안녕하세요' 동영상 클립에서 랜드마크 추출 결과 예
Fig. 5. Example of landmark detection results from a 30-frame 'hello' video clip

그림 2(b) '안녕', 그림 2(c) '오전', 그림 2(d) '오후', 그림 2(i) '귀엽다'와 같이 연속된 손의 움직임이 하나의 수어 단어를 나타낼 수도 있다. 그림 2(g) '쓰다'와 그림 2(h) '시다'는 오른 주먹의 검지를 펴서 오른쪽 입가에 대며 머리를 왼쪽으로 약간 기울이는 동작은 동일하지만, '쓰다'는 양쪽 눈을 감는 식으로 찡그리고, '시다'는 오른쪽 눈을 살짝 감으므로 얼굴 표정에 차이가 있다. '귀엽다'는 그림 2(j) '아깝다'와 같이 오른 손가락을 다 펴고 왼쪽 볼 옆에 붙여서 두 번 두드리는 손동작으로 나타내는데, '귀엽다'는 웃는 표정으로, '아깝다'는 아쉬운 표정을 지음으로써 수화 단어를 구분한다[13].

3-2 미디어파이프를 이용한 랜드마크 추출

먼저 동영상이 입력되면 각 프레임마다 미디어파이프의 Holistic 솔루션을 이용하여 얼굴 468개, 양손 각각 21개씩, 포즈(전신) 33개로 총 543개의 인체 관절점, 즉 랜드마크를 추출한다. 본 논문에서 제안하는 비수지 수화 동작인식 방법은 543개의 랜드마크를 모두 사용하는 대신, 수화 동작인식에 상대적으로 덜 중요한 얼굴의 코와 얼굴 윤곽, 포즈의 하반신 등의 랜드마크를 제외한다. 그림 3은 본 논문에서 사용하는 랜드마크를 보여주는데, 얼굴은 눈, 눈썹, 입의 랜드마크로 62개(그림 3(a)), 양손은 각각 21개(그림 3(b)), 어깨와 양팔의 포즈는 6개(그림 3(c))로 총 110개의 랜드마크를 사용한다.

그림 4는 각 수화에 대해 동영상에서 검출된 랜드마크 결과를 보여준다. 수화 동작에 필요한 손의 랜드마크 뿐만 아니라, '쓰다'와 '시다', '귀엽다'와 '아깝다'와 같은 비수지 수화 동작 인식에 필요한 눈과 입의 랜드마크의 검출 결과를 보여준다.

3-3 입력 특징 계산

3-2절에서 추출된 랜드마크는 인체 관절점의 위치 정보만 담고 있다 ($l=(x,y,z,v)$). 동적 동작을 보다 정확하게 인식하기 위해서 관절점의 위치와 방향에 불변한 관절의 사이각 정보를 이용한다. 그림 3의 랜드마크인 인체 관절점은 그들 사이의 위치 관계를 이용하여 선분으로 연결될 수 있고, 선분과 인접한 선분 사이의 관절 사이각은 점선 호로 표시된다.

표 1은 랜드마크(l_1,l_2,l_3)로부터 연결된 선분을 벡터(v_1,v_2)로 계산하고 벡터 사이의 관절 사이각($jointangle$)을 계산하는 알고리즘을 보여준다. 관절 사이각은 두 벡터를 정규화하여 내적(doc product)을 계산한 후 아크코사인(arccos)을 적용하여 계산한다.

랜드마크를 이용하여 관절 사이각을 계산하는 것과 달리, 어깨 관절의 사이각을 계산하기 위해서는 상체의 아래 부분을 의미하는 왼쪽 엉덩이와 오른쪽 엉덩이의 랜드마크 대신 두 랜드마크의 중심값으로 계산되는 엉덩이의 중앙 위치를 사용한다(그림 3(c)).

표 1. 랜드마크로부터 관절 사이각 계산 알고리즘
Table 1. Algorithm for calculating joint angles from landmarks

l_1, l_2, l_3 : landmarks

$v_1 = l_2 - l_1$
$v_2 = l_3 - l_2$
$v_1 = v_1$ / np.linalg.norm(v_1, axis=1)[:, np.newaxis]
$v_2 = v_2$ / np.linalg.norm(v_2, axis=1)[:, np.newaxis]
$jointangle$ = np.arccos(np.einsum('nt,nt->n', v_1, v_2))
$jointangle$ = np.degrees($jointangle$)

만약 수화를 위해 상반신만 비추는 카메라에서 어깨와 팔은 인식되지만, 엉덩이 부분이 포함되지 않아 랜드마크가 인식되지 않는 경우엔 카메라 하단의 중심 좌표로 대신한다. 이렇게 계산한 상반신 하단 중심의 위치와 어깨 랜드마크, 팔꿈치 랜드마크를 이용하여 표 1과 같이 어깨 사이각을 계산한다.

3-4 LSTM 모델 및 학습

동적 수화 동작을 인식하기 위해서 한 동작이 이루어지는 모든 프레임에서의 입력 특징을 계산해야 한다. 본 논문에서는 하나의 동작을 구성하는 비디오 클립을 30개의 프레임으로 정의한다. 그림 5는 30개의 프레임으로 구성된 '안녕하세요' 수화 동작에 대해 0, 12, 15, 18, 24 번째 프레임에서 랜드마크가 검출된 결과를 보여주고 있다. 한 동작에 대해 각 프레임에서 미디어파이프가 검출한 랜드마크와 3.3 절에서 계산한 관절 사이각을 이용하여 한 프레임에서 M개의 입력 특징을 계산한 후 $30 \times M$의 크기를 가진 입력 벡터를 구성한다. 본 논문에서는 적절한 M을 찾기 위해 그림 3에서 선분과 점선 호로 표시된 눈썹, 양손, 어깨의 입력 특징값을 필수로 포함시키면서 다수의 랜드마크로 구성된 눈과 입에서 일부 또는 전부를 포함하는 방식으로 M을 정한다.

본 논문에서 사용하는 동작인식 학습 모델은 표 2와 같이 1개의 장단기기억 층과 4개의 완전연결층(Dense)을 쌓아서 구성된다. 입력벡터의 각 프레임에 대한 M개 특징은 30개까지 순차적으로 512개의 메모리 단위가 있는 장단기기억 층의 입력으로 등장한다. 완전연결층은 각각 1024, 256, 256, N개의 출력을 가진다. 마지막 완전연결층의 출력의 개수인 N은 인식하려는 수화동작의 개수를 의미하고 본 논문에서는 36이다. 또한 드랍아웃(Dropout)은 0.3을 사용하고, 마지막 층의 활성화 함수로는 softmax를 사용한다.

3-5 실험 결과

본 논문에서 제안하는 동적 비수지 수화 동작인식을 위해 사용한 데이터는 한국 수화인식을 위한 데이터 셋[14], AIHub에서 제공하는 수어 영상 데이터[15] 함께 직접 촬영한 동영상 클립(video clip)들로 이루어졌다.

표 2. 본 논문에서 동적 수화동작 인식을 위해 사용한 장단기기억 모델
Table 2. LSTM model used for dynamic sign language recognition in this paper

N : number of sign language gestures to recognize
M : number of input features computed in one frame

```
model = tf.keras.models.Sequential([
 tf.keras.layers.Input(shape=(30, M), name='input'),
 tf.keras.layers.LSTM(512, time_major=False,
                      return_sequences=True),
 tf.keras.layers.Dense(1024, activation=tf.nn.relu),
 tf.keras.layers.Dropout(0.3),
 tf.keras.layers.Dense(256, activation=tf.nn.relu),
 tf.keras.layers.Dense(256, activation=tf.nn.relu),
 tf.keras.layers.Flatten(),
 tf.keras.layers.Dense(N, activation=tf.nn.softmax,
                       name='output')
])
```

각 동영상 클립의 길이는 1~5초이다. 각 수화 단어 당 1,200여개, 전체 43,200여개의 동영상 클립 데이터는 80%를 학습 데이터로, 10%를 검증(validation) 데이터로, 10%를 테스트 데이터로 사용했다.

수화 동작인식을 위한 학습은 운영체제는 윈도우 10, CPU는 Intel i5, GPU는 NVIDIA GeForce RTX 2060인 컴퓨터에서 이루어졌다. 표 2의 학습 모델을 사용하는데 필요한 하이퍼파라미터(hyper-parameter)는 학습률이 0.00005, 모멘텀(momentum)이 0.9, 가중치 감소(weight decay)가 e^{-6}이다. 손실 함수는 sparse categorical cross entropy를 사용한다.

그림 6은 학습 과정에서 각 에포크(epoch)마다 측정한 학습 데이터와 검증 데이터의 정확도와 손실률을 보여주고 있다. 학습을 마쳤을 때 학습 데이터의 최종 인식 정확도는 98.1%, 검증 데이터의 정확도는 89.7%이다. 학습데이터의 손실률은 0.2, 검증데이터의 손실률은 0.7이다.

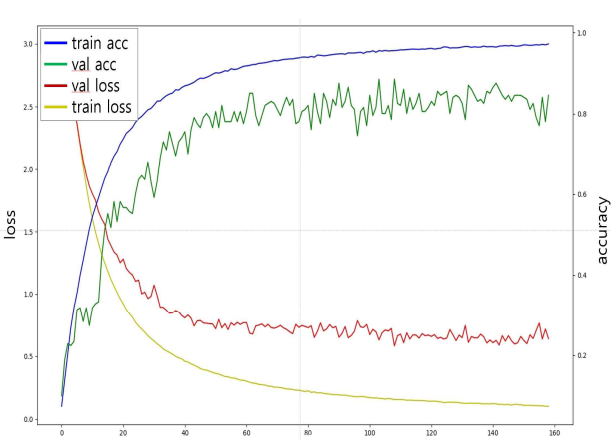


그림 6. 학습 정확도와 손실률
Fig. 6. Training accuracy and loss

표 3. 수화 동작인식 정확도(%)
Table 3. Gesture recognition accuracies(%) for sign languages

Model						Accuracy(%)
LRCN						71.12
proposed method	M = 480	position of 110 landmarks	40 jointangles of eyebrows, hands and a shoulder	×	×	74.39
	M = 532	position of 110 landmarks	40 jointangles of eyebrows, hands and a shoulder	all(52) of jointangles of eye and a mouth	×	81.46
	M=92	×	40 jointangles of eyebrows, hands and a shoulder	all(52) of jointangles of eye and a mouth	×	83.32
	M=56	×	40 jointangles of eyebrows, hands and a shoulder	some(16) of jointangles of eye and a mouth	×	83.29
	M=58	×	40 jointangles of eyebrows, hands and a shoulder	some(16) of jointangles of eye and a mouth	2 jointangles of wrists	84.57

표 4. 수화 동작인식 실험에 대한 비교
Table 4. Comparison of reported experiments on gesture recognition for sign languages

	Video data	Landmark detection	Gesture recognition	# of gesture signals	Recognition rate(%)
[2]	isolated video	OpenPose	CNN + majority voting	10 signals = 1 static manual + 8 dynamic manual + 1 dynamic non-manual signals	81%
[8]	continuous stream	Mediapipe	CNN	33 static manual signals(digits and korean alphabets)	x
proposed method	continuous stream	Mediapipe	LSTM	36 signals = 2 static manual + 28 dynamic manual + 6 dynamic non-manual signals	84.57%

제안하는 동적 비수지 수화 동작인식 방법의 우수성을 증명하기 위해 Long-term Recurrent Convolution Networks(LRCN)와 비교실험을 실행했다[14]. LRCN은 CNN과 LSTM이 결합된 구조로 각 프레임의 RGB 데이터를 학습시켰다. 표 3은 수화 동작인식 결과를 보여주는데, LRCN의 71.12%에 비해 전체적으로 제안하는 방법의 인식 정확도가 더 높음을 보여준다.

또한, 적절한 M을 찾기 위해 다양한 입력 특징을 이용한 비교실험을 진행했다. 랜드마크의 위치값을 포함하면, 그렇지 않은 경우보다 입력 특징의 개수가 늘어남에도 불구하고 인식 정확도는 오히려 좋지 않음을 보여준다. 눈과 입의 관절각은 전부 또는 일부를 포함하는 것이 정확도에 크게 영향을 미치지 않았다. 반면, 그림 3(b)의 손 랜드마크 9번을 그림 3(c)의 양팔 랜드마크 15, 16번과 각각 연결하여 계산한 관절 15, 16의 사이각을 추가하는 경우 가장 높은 정확도를 보여준다.

표 4는 기존 수화 동작인식 연구와 제안하는 방법의 동작인식 방법, 대상 수화 신호, 인식 결과를 비교하여 보여주고 있다. 제안하는 방법은 카메라를 통해 들어오는 연속 동영상 스트림을 입력받아 미디어파이프와 장단기기억 모델을 사용하여 다양한 동적 수지 신호와 비수지 신호를 비교적 더 높게 인식하고 있음을 보여주고 있다.

Ⅳ. 수화 교육 앱에 적용

4-1 제안하는 수화 교육 앱 설계

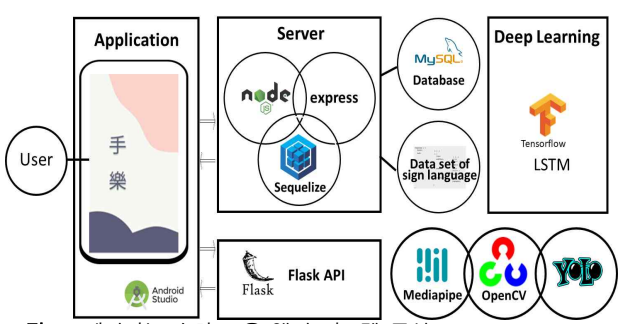


그림 7. 제안하는 수화 교육 앱의 시스템 구성도
Fig. 7. System diagram of the proposed sign language application

동적 비수지 수화 동작인식을 적용하기 위해 본 논문에서는 수화 교육 앱을 제안한다. 그림 7은 제안하는 수화 교육 앱의 시스템 구성도를 보여준다. 제안하는 수화 앱은 Node.js와 Flask를 이용하여 서버를 구축한다. Node.js 서버는 Express와 Sequelize를 사용하여 MySQL 데이터베이스에 접근한다. Node.js 서버의 데이터베이스는 사용자 정보와 수화 단어 교육에 필요한 단어 정보, 즉 단어, 단어 그림, 수화 동영상 등이 저장되어있다. Flask 서버는 3-3 절에서 설명한 학습에 필요한 입력 특징을 계산하여 json의 형태로 안드로이드 폰과 데이터를 주고받는다.

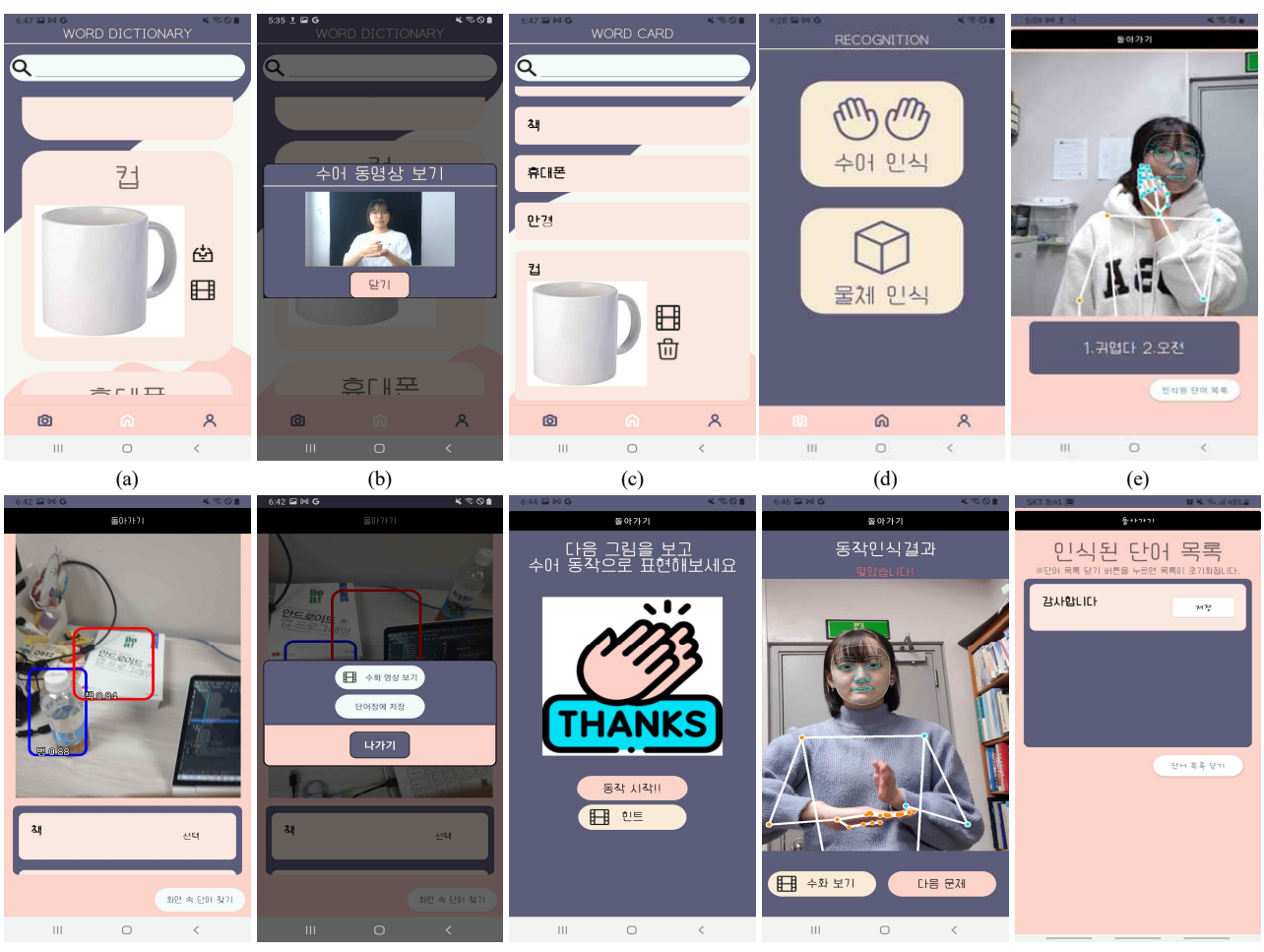
제안하는 수화 앱이 설치된 안드로이드 폰에는 수화 동작인식과 물체인식 기능을 위하여 3장에서 설명한 장단기기억 학습 모델과 물체인식을 위한 YOLO 모델[16]가 설치된다. 안드로이드 폰의 카메라를 통해 들어오는 동영상은 OpenCV

와 미디어파이프에 의해 사용자의 신체를 인식하고 상반신(얼굴, 손, 포즈)의 랜드마크를 추출한다. 추출된 랜드마크는 Flask 서버에 전달되고 서버에서 계산된 학습에 필요한 입력 특징은 다시 안드로이드 폰으로 반환된다. 반환된 특징값은 안드로이드 내부에 설치된 장단기기억 학습 모델에 입력되어 수화 동작인식을 수행한다. 물체 인식도 안드로이드 폰의 카메라로 들어오는 동영상에서 YOLO를 이용하여 물체를 인식하고 그 결과를 사용자에게 보여준다.

4-2 제안하는 수화 교육 앱 개발 결과

그림 8은 본 논문에서 제안하는 동적 비수지 수화 동작인식을 이용한 수화 교육 앱의 개발 결과를 보여주고 있다. 로그인한 사용자는 '단어 사전', '단어 카드', '인식', '단어 퀴즈' 기능을 사용할 수 있다. '단어 사전'을 통해 제안하는 앱이 제공하는 모든 수화 단어 정보를 확인할 수 있다(그림 8(a)). 각 단어는 텍스트, 이미지, '내려받기', '수화 동영상'으로 구성되어 있다. 사용자는 각 단어의 수화 동작을 동영상으로 확인하고(그림 8(b)), 반복적으로 교육하고 싶은 단어들은 '단어 카드'에 저장해서 단어를 쉽게 찾을 수 있도록 한다(그림 8(c)).

'인식(그림 8(d))' 기능은 수화 인식과 물체 인식 두 가지가 가능하다. 수화인식을 위해 안드로이드 폰 카메라로 들어오는 동영상에 대해 미디어파이프를 이용하여 인체정보를 계속 획득한다. 실시간 동영상에서 동작인식은 동작의 시작을 찾는 것이 중요한데, 본 논문에서는 양 손이 인식되고 일정 크기의 움직임이 검출되면 동작의 시작으로 판단하여 동작인식을 위한 30개의 프레임을 추출하고 Flask 서버로부터 입력 특징을 계산한다. 계산된 입력 특징은 벡터 형태로 폰에 설치된 학습 모델에 입력되어 수화 동작인식을 수행한다. 인식결과는 그림 8(e)과 같이 바로 확인할 수 있다.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

* These figures are screen captures of the developed application, and this application is made for Korean users.

그림 8. 제안하는 수화 교육 앱 화면 : (a) 단어 사전, (b) 수화 동영상, (c) 단어 카드, (d) 인식, (e) 동작인식(예: 귀엽다), (f) 물체 인식, (g) 인식된 물체 단어 확인, (h) 단어 퀴즈, (i) 동작인식 결과(예: 감사합니다), (j) 인식된 동작 단어 저장

Fig. 8. Screenshots of the proposed application : (a) word dictionary, (b) video clip of sign language, (c) word card, (d) recognition, (e) gesture recognition(e.g. cute), (f) object recognition, (g) confirm the recognized object, (h) word quiz, (i) result of gesture recognition(e.g. thank) and (j) save the recognized gesture to word list

그림 8(f)은 YOLO를 이용한 물체인식 화면으로, 안드로이드 폰 카메라를 통해 주변의 물체를 비추는 비디오가 들어오면 폰에 설치된 YOLO 모델로 물체를 인식한다. 인식된 물체의 단어를 선택하면 그림 8(b)과 같은 동영상을 통해 수화 동작을 시청하며 교육할 수 있고, '단어 카드'에 저장하여 이후 반복 교육을 할 수 있다 (그림 8(g)).

제안하는 수화 교육 앱은 배운 내용을 '단어 퀴즈'를 통해 확인할 수 있다. 그림 8(h)과 같이 단어가 제시되면, '힌트'로 그림 8(c)과 같은 수화 동영상을 미리 보거나, '동작 시작' 버튼을 눌러 동작인식을 수행한다. 동작인식 수행방식은 '인식'의 수화 인식과 같으며, 사용자가 제시된 단어의 수화 동작을 올바르게 취하면 수화가 인식되는 화면이 보이며(그림 8(i)), 물체인식 단어와 마찬가지로 수화 동영상을 직접 확인하거나 '단어 카드'에 저장할 수 있다 (그림 8(j)).

V. 결 론

대부분의 수화 동작인식이 손의 형태를 인식하여 제한된 단어만 인식할 수 있는 한계를 극복하고자, 본 논문에서는 동적으로 변화는 동작 뿐만 아니라 얼굴 표정까지 포함하는 비수지 수화 동작을 인식하는 방법을 제안한다. 미디어파이프 프레임워크를 이용하여 얼굴과 손, 상반신의 관절점 정보를 획득하고, 동작인식에 필요한 관절 사이각을 계산하여 입력 벡터를 구성한다. 동적인 동작인식을 위해 한 동작을 30개의 프레임으로 정의하고 각 프레임에서 계산한 입력 벡터를 순차적으로 장단기기억 모델에 입력하여 동적인 수화 동작을 학습시킨다. 제안하는 방법은 비교 실험한 LRCN에 비해 더 높은 인식 정확도를 보여주고 있다.

제안하는 동작인식 방법은 수화 교육이 가능한 모바일 앱에 적용되었다. 사용자는 단어의 수화 동작을 동영상을 통해 언제든지 보고 배울 수 있을 뿐만 아니라 퀴즈 또는 인식 기능을 이용해 안드로이드 폰 카메라 앞에서 직접 동작을 취하고 인식 결과를 확인할 수 있다.

본 논문에서 제안하는 수화 동작 인식이 명사 또는 서술어로 구성된 단어 위주였다면, 향후 여러 개의 단어로 구성된 문장을 인식할 수 있도록 확장할 계획이다. 연속으로 수행하는 손동작과 얼굴 표정에서 의미 있는 수화 단어를 분리하고 단어를 인식하여 문장으로 해석할 수 있는 방안을 연구할 계획이다. 또한 얼굴 표정을 활용하여 종결형 문장과 의문형 문장을 구분할 수 있도록 추가 학습할 계획이다.

감사의 글

본 연구는 2021년도 덕성여자대학교 교내연구비 지원에 의해 이루어졌습니다.

참고문헌

[1] Korean Statistical Information Service, Number of registered disabled people by type and gender nationwide [Internet]. Available: https://kosis.kr/statHtml/statHtml.do?orgId=117&tblId=DT_11761_N001&conn_path=I2.

[2] I.H. Kim and I.H. Jung, "A study on Korean sign language motion recognition using OpenPose based deep learning," *Journal of Digital Contents Society*, Vol. 22, No, 4, pp. 681-687, April 2021. https://doi.org/10.9728/dcs.2021.22.4.681

[3] R. Rastgoo, K. Kiani, and S. Escalera, "Sign language recognition: A deep survey," *Expert systems with applications,* Vol. 164, 113794, February 2021. https://doi.org/10.1016/j.eswa.2020.113794

[4] N. Adaloglou, *et al*. "A comprehensive study on sign language recognition methods," *arXiv preprint arXiv*:2007.12530, July 2020. https://doi.org/10.48550/arXiv.2007.12530

[5] M. Al-Qurishi, T. Khalid and R. Souissi, "Deep learning for sign language recognition: Current techniques, benchmarks, and open issues," *IEEE Access*, Vol. 9, pp. 126917-126951, December 2021. https://doi.org/10.1109/ACCESS.2021.3110912

[6] G. Latif, *et al*. "An automatic Arabic sign language recognition system based on deep CNN: an assistive system for the deaf and hard of hearing," *International Journal of Computing and Digital Systems*, Vol. 9, No. 4, pp. 715-724, July 2020. http://dx.doi.org/10.12785/ijcds/090418

[7] A. Chaikaew; K. Somkuan, and T. Yuyen, "Thai sign language recognition: an application of deep neural network," in *Proceedings of 2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering*, Chiang Mai, pp. 128-131, March 2021. https://doi.org/10.1109/ECTIDAMTNCON51128.2021.9425711

[8] J.Y. Kim and H. Sim, "Development of a sign language learning assistance system using Mediapipe for sign language education of deaf-mutility," *Journal of Korea Institute of Electronic Communication Science*, Vol. 16, No. 6, pp. 1355-1362, December 2021. https://doi.org/10.13067/JKIECS.2021.16.6.135

[9] C. Lugaresi, *et al*. "Mediapipe: A framework for building perception pipelines," in *Proceedings of Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition*, Long Beach, June 2019.

https://doi.org/10.48550/arXiv.1906.08172
[10] R.C. Staudemeyer and E.R. Morris, "Understanding LSTM - a tutorial into long short-term memory recurrent neural networks," *arXiv preprint arXiv*:1909.09586, September 2019. https://doi.org/10.48550/arXiv.1909.09586
[11] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, Vol. 9, No. 8, pp. 1735-1780, November 1997. https://doi.org/10.1162/neco.1997.9.8.1735
[12] Sign Language Dictionary, Seoul Welfare Portal [Internet]. Available: https://wis.seoul.go.kr/handicap/understand/signLanguageDictionary.do.
[13] Gyeonggi-do sign language education center, Korean sign language – Beginners [Internet]. Available: https://www.gg.go.kr/cmmn/download.do?idx=566825.
[14] S.H. Yang, S.J. Jung, H.K. Kang and C.I. Kim, “The Korean sign language dataset for action recognition,” in *Proceedings of 26th International Conference on MultiMedia Modeling*, pp. 532 - 542, January 2020. https://doi.org/10.1007/978-3-030-37731-1_43
[15] Sign language video clips, AIHub [Internet]. https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=116&topMenu=100&aihubDataSe=ty&dataSetSn=103
[16] J. Redmon and A. Farhadi, “YOLO v3: An incremental improvement”, *arXiv preprint arXiv*:1804.02767v1, April 2018. https://doi.org/10.48550/arXiv.1804.02767

김신영(Shin-Yong Kim)

2019년～현 재: 덕성여자대학교 컴퓨터공학과 학사과정
※관심분야 : 안드로이드, 백엔드, 인공지능

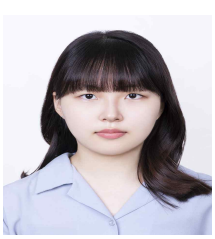
엄서정(Seo-Jung Urm)

2019년～현 재: 덕성여자대학교 컴퓨터공학과 학사과정
※관심분야 : 보안, 안드로이드, 백엔드

유선영(Sun-Young Yoo)

2019년～현 재: 덕성여자대학교 컴퓨터공학과 학사과정
※관심분야 : 안드로이드, 웹, 프론트엔드

김수정(Soo-Jeong Kim)



2019년～현 재: 덕성여자대학교 컴퓨터공학과 학사과정
※관심분야 : 웹, 프론트엔드, UX/UI

이경미(Kyoung-Mi Lee)

1993년 : 덕성여자대학교 전산학과 (이학사)
1996년 : 연세대학교 전산과학과 (이학석사)
2001년 : 아이오와 대학교 전산학과 (전산학박사-지능형 멀티미디어)

2003년～현 재: 덕성여자대학교 컴퓨터공학과 교수
※관심분야 : 영상처리, 컴퓨터비전, 패턴인식