JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 5, pp. 123-132, May 2026
https://doi.org/10.9708/jksci.2026.31.05.123

Real-Time Korean Sign Language Learning Recognition System Using MediaPipe and LSTM

Jieun Lee*, Young-Im Cho**

*Student, Dept. of Computer Engineering, Gachon University, Gyeonggi, Korea
**Professor, Dept. of Computer Engineering, Gachon University, Gyeonggi, Korea

[Abstract]

The need for sign language education to facilitate communication between the hearing-impaired and the general public is growing; however, limited access to professional institutions and the lack of real-time feedback hinder learning efficiency. This paper proposes a real-time Korean sign language recognition system that combines MediaPipe-based hand landmark extraction with an LSTM deep learning model. A dedicated dataset of 32 Korean sign language vocabularies was constructed under varied conditions, generating approximately 68,000 training sequences via a sliding window approach with an 80:20 train-test split. Experimental results demonstrate a recognition accuracy of 94.82% and a Marco F1-score of 0.95, with the LSTM model showing slightly higher accuracy and a more stable learning tendency compared to a GRU model. A cross-user evaluation achieved an average recognition rate of 83.3%, confirming practical generalization capability. The system is integrated with a web service to provide real-time feedback for improved accessibility in sign language education.

▸ **Key words**: Sign Language Recognition, Korean Sign Language, MediaPipe, LSTM, Deep Learning

[요 약]

청각장애인과 비장애인 간 의사소통을 위한 수어 교육의 필요성이 증가하고 있으나, 전문 교육 기관에 대한 접근 제한과 실시간 피드백 부재로 학습 효율이 저하되고 있다. 본 논문은 MediaPipe 기반 손 랜드마크 추출과 LSTM 모델을 결합한 실시간 한국수어 학습 인식 시스템을 제안한다. 32개 한국수어 어휘를 다양한 배경·조명·각도 조건에서 수집하여 슬라이딩 윈도우 방식으로 약 68,000개의 학습 시퀀스를 생성하고, 8:2 비율로 데이터를 분할하였다. 실험 결과, 인식 정확도 94.82%, Macro F1-score 0.95를 달성하였으며, 동일 조건의 GRU 모델 대비 소폭 높은 정확도와 안정적인 학습 경향을 보였다. 교차 사용자 실험에서 평균 83.3%의 인식률을 기록하여 일반화 가능성을 확인하였고, 웹 서비스 연동을 통해 실시간 피드백을 제공함으로써 수어 교육의 접근성 향상에 기여한다.

▸ **주제어**: 수어 인식, 한국수어, MediaPipe, LSTM, 딥러닝

• First Author: Jieun Lee, Corresponding Author: Young-Im Cho
*Jieun Lee (jinny2190@gachon.ac.kr), Dept. of Computer Engineering, Gachon University
**Young-Im Cho (yicho@gachon.ac.kr), Dept. of Computer Engineering, Gachon University
• Received: 2026. 03. 19, Revised: 2026. 04. 24, Accepted: 2026. 05. 13.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

세계보건기구(WHO)에 따르면 전 세계 인구의 5% 이상이 청각 장애를 가지고 있으며, 2050년까지 약 7억 명 이상이 청각 손실을 경험할 것으로 예측된다[1]. 청각장애인들은 수어를 주요 의사소통 수단으로 사용하지만, 수어를 이해하는 비장애인의 비율은 낮아 사회적 소통의 장벽이 존재한다. 또한, 후천적으로 청각장애를 가지게 된 사람들은 수어에 대한 사전 지식이 부족하여 의사소통에 어려움을 겪고 자막 번역과 같은 대안에 의존해야 하는 경우가 많다. 이를 해소하기 위한 수어 교육의 필요성이 높아지고 있으나, 전문 교육 기관 접근이 어렵거나 대면 학습 환경이 제한된 학습자에게는 여전히 제약이 많다. 현재 상용화된 수어 학습 앱 및 서비스의 대부분은 수어 사전 및 영상 제공에만 의존하며, 학습자가 자신의 동작이 올바른지 실시간으로 확인할 수 있는 피드백 기능은 제공되지 않고 있다. 이에 따라 언제 어디서나 접근 가능한 실시간 수어 학습 지원 시스템의 필요성이 대두된다.

최근 컴퓨터 비전과 딥러닝 기술의 발전으로 실시간 제스처 인식 시스템 연구가 활발히 진행되고 있다. 특히, Google의 MediaPipe 프레임워크는 단일 RGB 카메라만으로 손의 21개 랜드마크를 실시간으로 추출할 수 있어[2], 별도의 특수 장비 없이 수어 인식 시스템을 구현하는 데 적합한 기술로 주목받고 있다. 또한, LSTM 네트워크는 시계열 데이터의 장기 의존성 학습에 강점을 가져[3] 수어와 같이 시간적 동적 변화를 포함하는 제스처 인식에 효과적으로 활용된다.

그러나, 기존 연구들은 대부분 미국수어(ASL) 또는 인도수어(ISL) 등 외국 수어를 대상으로 하며[4], 한국수어를 대상으로 한 MediaPipe 기반 실시간 학습 서비스 연구는 미흡한 실정이다. 또한, AI Hub 등에서 제공하는 기존 한국수어 데이터셋은 대규모 영상(video) 기반으로 구성되어 있어, MediaPipe 랜드마크 좌표를 입력으로 하는 경량 실시간 인식 모델 학습에 직접 적용하기 어렵다는 한계가 있다.

본 연구는 기존 수어 인식 연구들과 아래와 같은 측면에서 차별화된다. 첫째, 기존 연구들이 주로 ASL·ISL 등 외국 수어를 대상으로 하거나 영상 기반 데이터셋을 활용하는 데 반해, 본 연구는 수어 학습 서비스 맥락에 특화된 32개 한국수어 어휘의 MediaPipe 랜드마크 기반 전용 데이터셋을 직접 구축한다. 둘째, 기존 수어 인식 연구들이 인식 정확도 제고에 초점을 맞추는 데 반해, 본 연구는 웹 서비스와 연동하여 학습자에게 즉각적인 실시간 피드백을 제공하는 수어 학습 보조 시스템으로의 활용을 목표로 한다. 셋째, 학습 데이터 수집에 참여하지 않은 피험자를 대상으로 교차 사용자 실험을 수행하여 시스템의 초기 수준의 사용자 독립 성능 가능성을 확인한다. 넷째, 동일 구조의 GRU 모델과의 비교를 통해 LSTM의 성능 특성과 학습 안정성 경향을 분석한다.

2장에서는 관련 연구를 살펴보고, 3장에서는 제안 시스템의 구조와 방법론을 설명한다. 4장에서는 실험 환경 및 결과를 분석하며, 5장에서 결론을 맺는다.

II. Related Works

수어 인식 분야의 딥러닝 기반 연구는 크게 CNN 기반 정적 인식과 RNN 계열 기반 동적 인식으로 구분된다. Bagyammal et al.[4]은 MediaPipe와 LSTM을 결합하여 미국수어(ASL) 알파벳 26개를 인식하는 시스템을 제안하였으며, 99%의 인식 정확도를 달성하였다. 그러나, 이 연구는 정적 제스처만을 대상으로 하여 연속적인 동적 수어 표현 인식에는 한계가 있다.

Samaan et al.[5]은 MediaPipe와 RNN 계열 모델(GRU, LSTM, 양방향 LSTM)을 결합하여 동적 수어 인식 시스템을 비교 제안하였으며, 얼굴 키포인트 포함 여부에 따른 성능 차이를 분석하여 99% 이상의 정확도를 보고하였다. 해당 연구는 10개 어휘에 대한 자체 데이터셋을 구축하여 활용하였다는 점에서 본 연구와 유사한 접근 방식을 취하나, 본 연구는 32개 어휘로 범위를 확장하고 교차 사용자 검증을 추가로 수행하였다.

Wang et al.[6]은 MediaPipe를 특징 추출에 활용하고 LSA64 아르헨티나 수어 데이터셋(64개 어휘)에 대해 LSTM과 GRU 모델을 비교 실험하여 각각 94.06%, 94.53%의 정확도를 보고하였다. Tran et al.[7]은 MediaPipe 기반 CNN과 LSTM을 결합하여 연속 수어 인식에서 99.5%의 높은 정확도를 달성하였으며, Maheswara et al.[8]은 인도네시아 수어(BISINDO)를 대상으로 GRU와 LSTM 모델을 비교 분석하였다. 이상의 연구들은 MediaPipe와 RNN 계열 모델의 조합이 수어 인식에 효과적임을 보여주나, 한국수어를 대상으로 하여 수어 인식을 위한 전용 데이터셋 구축 및 연구가 충분히 이루어지지 않았다. 본 연구는 이러한 공백을 보완하고자 한다.

III. The Proposed System

3.1 System Overview

본 논문에서 제안하는 시스템은 데이터 수집, 랜드마크 추출, 시퀀스 구성, 모델 학습, 실시간 인식 및 웹 서비스 연동의 순서로 구성된다. 데이터 수집 단계에서는 OpenCV 기반 웹캠으로 한국수어 어휘 영상을 수집하고, 랜드마크 추출 단계에서는 MediaPipe Hands를 통해 각 프레임으로부터 21개 손 랜드마크(63차원)을 실시간으로 추출한다. 시퀀스 구성 단계에서는 슬라이딩 윈도우(크기 10프레임)를 적용하여 LSTM 입력 시퀀스를 생성하고, 모델 학습 단계에서는 LSTM 기반 분류 모델을 학습한다. 실시간 인식 및 웹 서비스 연동 단계에서는 학습된 모델을 웹 서비스에 탑재하여 학습자에게 즉각적인 수어 인식 피드백을 제공한다. 각 단계의 세부 구현은 3.2절~3.7절에서 순서대로 기술한다.

3.2 Data Collection and Dataset Composition

본 연구에서는 수어 학습 서비스에 필요한 32개 한국수어 어휘를 선정하였다. 선정 어휘는 일상 인사·감사·감정·장소·의문문 등 실제 학습 수요가 높은 표현으로 구성하였으며, Fig. 1에 전체 어휘 목록을 제시한다.

Fig. 1. Vocabulary List

기존의 대규모 수어 데이터셋(예: AI Hub 수어 영상 데이터)은 영상 (video) 기반으로 구성되어 MediaPipe 랜드마크 좌표를 입력으로 하는 경량 실시간 모델 학습에 직접 활용하기 어렵다. 이에 본 연구에서는 OpenCV를 활용한 웹캠 기반 데이터 수집 환경을 구축하고 단일 연구자가 직접 데이터를 수집하였다.

데이터 수집은 초당 약 10프레임(fps)으로 40초간 연속 촬영하는 방식으로 진행되었으며, 조명 조건, 배경 환경, 촬영 각도(정면/측면)를 달리하여 각 어휘당 3회씩 반복 수집하였다. 그 결과, 어휘당 720프레임 이상의 연속 이미지를 확보하였으며, 총 32개 어휘에 대해 약 23,000장 이상의 이미지 데이터셋을 구축하였다. 이후 슬라이딩 윈도우 방식(윈도우 크기 10)을 적용하여 최종적으로 약 68,000개 이상의 시퀀스 샘플을 생성하였고, 이를 8:2 비율로 분할하여 학습 데이터(약 54,400개)와 테스트 데이터(약 13,600개)로 구성하였다. 데이터셋 구성 정보를 Table 1에 요약하였다.

Table 1. Dataset Composition

Item	Details
Data Collection Participant	1 person
Number of Vocabulary Words	32 Korean Sign Language (KSL) vocabulary words
Collection Conditions	Combinations of lighting, background, and angle; 3 repetitions per vocabulary word
Frames per Vocabulary Word	Over 720 frames
Total Number of Images	23,000 frames or more
Total Sequences after Sliding Window	68,000 or more
Training / Test Data	80% / 20% (Random split)

본 연구의 학습·테스트 분할은 단일 수집자 데이터 내에서 무작위로 수행되었다. 따라서, 학습 및 테스트 데이터는 동일 사용자로부터 수집된 데이터를 공유한다는 점에 유의하여야 하며, 이를 보완하기 위해 학습 데이터 수집에 참여하지 않은 별도의 피험자 3명을 대상으로 교차 사용자 실험(4.5절)을 추가 수행하여 시스템의 사용자 독립적 일반화 성능을 별도로 검증하였다.

3.3 Landmark Extraction

MediaPipe Hands 모델[2,9]은 단일 RGB 카메라 입력으로부터 손의 21개 랜드마크를 실시간으로 추출한다. 각 랜드마크는 (x, y, z) 3차원 좌표로 표현되므로, 단일 프레임에서 추출되는 특징 벡터의 차원은 21 x 3 = 63차원이다. 이 방식은 원시 이미지를 직접 입력으로 사용하는 CNN 방식에 비해 계산 비용이 낮고, 배경이나 조명 변화

에 강인한 특성을 가진다. 실시간 환경에서는 두 손이 항상 동시에 안정적으로 검출되지 않으며, 손의 가림이나 프레임별 검출 순서 변화로 인해 특징 데이터의 일관성이 저하되는 문제를 고려하여 본 연구에서는 두 손의 랜드마크를 모두 사용하는 대신 주요 손(dominant hand)의 랜드마크만을 특징으로 사용하였다.

수어는 손동작뿐만 아니라 얼굴 표정, 시선, 상체 동작 등 비수지적 표지가 의미 전달에 중요한 역할을 한다. 그러나, 본 연구에서는 다음과 같은 이유로 MediaPipe Hands 의 주요 손 랜드마크 21개, 즉 63차원 특징만을 사용하였다. 첫째, 본 시스템의 목적은 수어 완전 번역이 아닌 학습자를 위한 수어 동작 실시간 피드백으로, 수어 학습 초기 단계에서 가장 중요한 교정 대상은 손 형태와 이동 동선의 정확성이다. 둘째, 본 연구에서 구성한 32개 한국수어 어휘는 일상 인사·감사·감정·장소·의문문 표현을 중심으로 하며, 이들 어휘는 대체로 손의 형태와 이동 궤적에 의해 1차적으로 변별된다. 비수지 신호(얼굴 표정, 시선, 고개 움직임)에 의해서만 의미가 분화되는 동형 수어 쌍이 포함되지 않도록 어휘를 구성하여, 손 특징만으로도 어휘 간 분류가 가능한 조건을 확보하였다. 정적 수어 제스처의 경우 한 손의 형태만으로 의미 구분이 가능하므로, 단일 손 기반 특징을 사용하더라도 수어 인식 성능을 충분히 확보할 수 있다.

3.4 Sequence Construction

LSTM 모델의 입력으로 사용하기 위해 연속 촬영된 프레임 이미지로부터 추출된 랜드마크 데이터를 슬라이딩 윈도우 방식으로 시퀀스화하였다. 시퀀스 길이는 10프레임으로 설정하였으며, 이는 약 1초간의 수어 동작에 해당한다. 시퀀스 길이 설정을 위해 5프레임, 10프레임, 15프레임을 각각 예비 실험을 수행하였다. 5프레임(0.5초)의 경우, 연속 손동작의 전체 궤적을 충분히 포착하지 못하여 검증 정확도가 저조하였다. 15프레임(1.5초)의 경우, 시퀀스 생성에 소요되는 시간이 길어져 실시간 피드백 응답 속도가 지연되는 문제가 발생하였다. 10프레임은 대부분의 한국수어 어휘 동작 궤적을 충분히 포착하면서도 1초 이내의 실시간 응답이 가능하여 인식 성능과 응답 속도 간의 균형을 가장 효과적으로 달성하는 설정으로 최종 선정하였다. 슬라이딩 윈도우 방식을 적용함으로써 N개의 연속 프레임으로부터 N-9개의 시퀀스 샘플을 생성할 수 있어 학습 데이터를 효율적으로 증강하는 효과가 있다. 최종 입력 데이터 형태는 (samples, 10, 63)이다.

3.5 LSTM Model Architecture

제안하는 LSTM 모델의 구조는 Table 3과 같다. 모델은 2개의 LSTM 레이어(256유닛 → 128유닛)와 각 레이어 뒤에 배치된 BatchNormalization, Dropout으로 구성되며, 과적합 방지를 위해 L2 정규화(λ=0.005)를 적용하였다. 활성화 함수는 LSTM 레이어에 tanh, 출력 레이어에 Softmax를 사용하였고, 모델의 총 파라미터 수는 약 530,000개이다. 수어 동작은 손의 이동 경로, 속도 변화, 정지 동작 등 복잡한 시간적 패턴을 포함하므로 장기 의존성 포착 능력이 중요하다. LSTM은 3중 게이트(입력·망각·출력) 구조와 독립적인 셀 상태(cell state)를 통해 이러한 장기 의존성을 효과적으로 학습할 수 있어[3] 기본 모델로 선정하였다. 이는 Samaan et al.[5] 및 Maheswara et al.[8]이 동적 수어 인식에서 LSTM이 GRU 대비 안정적인 분류 성능을 보임을 보고한 선행 연구와도 일치한다.

Table 2. LSTM Model Architecture

Layer	Number of Units	Settings
LSTM (1)	256	return_sequences=True, L2=0.005
BatchNormalization	-	-
Dropout	-	rate=0.3
LSTM (2)	128	return_sequences=False, L2=0.005
BatchNormalization	-	-
Dropout	-	rate=0.3
Dense	32	Softmax, L2=0.005

1번 LSTM 레이어의 256유닛은 입력 특징 차원(63차원) 대비 충분한 표현 용량을 확보하면서도 과도한 파라미터 증가를 방지할 수 있도록 설정하였다. 이는 10프레임 시퀀스 내에서 랜드마크 간 공간적 상관관계 및 프레임 간 시간적 변화 패턴을 효과적으로 학습하기 위함이다. 시계열 분류 과제에서 은닉 유닛 수가 너무 작을 경우 복잡한 시간적 패턴을 충분히 학습하지 못하고, 너무 클 경우 학습 데이터 규모 대비 과적합 위험과 연산 비용이 증가한다. 실제로 예비 실험에서 128유닛 단층 구성은 32개 어휘의 손동작 패턴을 충분히 구분하지 못하는 경향을 보였다.

3.6 Training Configuration

학습에는 Adam 옵티마이저[10]를 사용하였으며, 손실 함수는 다중 클래스 분류에 적합한 categorical cross-entropy를 적용하였다. 각 하이퍼파라미터의 설정 근거는 다음과 같다.

- 학습률 (learning_rate = 0.0003): Adam의 기본값(0.001) 대비 낮게 설정하여 수렴 안정성을 높였다. 수어 동작 데이터는 클래스 간 손동작 패턴의 유사성이 높아, 학습률이 클 경우 파라미터 업데이트 폭이 과도하여 손실 함수가 진동하는 현상이 발생할 수 있다. 0.0003은 예비 실험(0.001, 0.0005, 0.0003, 0.0001)에서 검증 손실 수렴 속도와 최종 정확도의 균형이 가장 우수한 값으로 선정하였다.
- 배치 크기 (batch_size = 32): 배치 크기가 너무 클 경우 기울기 추정의 분산이 낮아져 지역 최솟값(local minima)에 수렴하기 쉽고, 너무 작을 경우 학습 시간이 증가하고 기울기 잡음이 커진다. 32는 순환 신경망 기반 시계열 분류 과제에서 범용적으로 사용되는 값으로, 본 데이터셋 규모(약 68,000 시퀀스)에서 학습 안정성과 속도 간 균형을 제공하였다.
- 최대 에폭 (max_epochs = 80) + EarlyStopping (patience=15): 최대 에폭은 충분한 학습 기회를 보장하기 위해 80으로 설정하였으나, 실제 학습은 EarlyStopping 콜백에 의해 검증 손실이 15 에폭 이상 개선되지 않을 경우 자동으로 조기 종료되어 과적합을 방지한다.
- ReduceLROnPlateau(factor=0.5, patience=7): 검증 손실이 7 에폭 이상 정체될 경우 학습률을 0.5배로 감소시켜 세밀한 수렴을 유도한다. 이는 EarlyStopping 이전 단계에서 최적점 근방에서의 탐색을 정교화한다.
- Dropout (rate = 0.3): 각 LSTM 레이어 뒤에 Dropout(0.3)을 적용하여 특정 뉴런에 대한 과의존을 방지하였다. 0.3은 Dropout 비율이 과도할 경우(>0.5) 학습 능력을 저해하고 부족할 경우(<0.2) 과적합 억제 효과가 미미한 점을 고려하여 선정하였다.
- L2 정규화 (λ = 0.005): LSTM 레이어 및 출력 Dense 레이어에 L2 정규화를 적용하여 가중치가 지나치게 커지는 것을 억제하였다. λ=0.005는 예비 실험에서 정규화 효과와 학습 용량 유지 간의 균형이 가장 우수한 값으로 선정하였다.

3.7 Real-time Recognition

학습된 모델은 웹 서비스와 연동되어 실시간 수어 학습 피드백 기능을 제공할 것이다. 본 시스템의 전형적인 사용 시나리오는 다음과 같다. 학습자가 웹 브라우저를 통해 서비스에 접속하면 웹캠 영상이 클라이언트에서 서버로 실시간 전송된다. 서버 측에서는 MediaPipe를 통해 매 프레임마다 손 랜드마크를 추출하고, 슬라이딩 윈도우 방식으로 최근 10프레임의 랜드마크 시퀀스를 구성하여 LSTM 모델에 입력한다. 모델의 분류 결과는 JSON 형식으로 클라이언트에 반환되며, 학습자 화면에 인식된 수어 어휘가 즉시 표시된다. 전체 파이프라인의 응답 지연은 1초 이내로 유지되어(MediaPipe 약 30.3 FPS, LSTM 추론 약 70ms), 학습자는 자신의 수어 동작에 대한 실시간 피드백을 받으며 반복 연습이 가능하다. 이러한 즉각적 교정 피드백 구조는 전문 강사 없이도 자기주도적 수어 학습 환경을 제공한다는 점에서 수어 교육 접근성 향상에 기여할 수 있다.

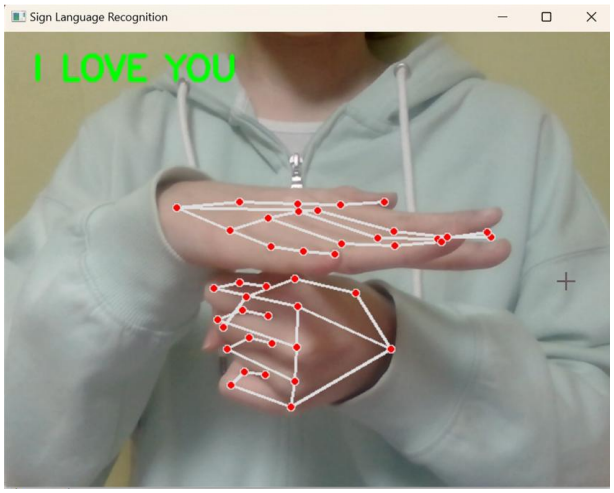


Fig. 2. Example of Real-time Recognition for the "I Love You" (사랑해) Sign

그러나, 인식 결과를 화면에 출력하는 과정에서 OpenCV 라이브러리의 기본 텍스트 출력 함수인 cv2.putText()가 한글을 지원하지 않는 문제가 발생하였다. 해당 함수는 ASCII 기반의 영문 폰트만 지원하기 때문에 한글을 직접 출력할 경우 문자 깨짐 현상이 발생하거나 물음표 형태로 표시되는 문제가 있었다. 본 연구에서는 시스템의 실시간 처리 성능과 구현의 단순성을 고려하여 인식 결과를 영어 라벨로 변환하여 출력하는 방식을 채택하였다. 구체적으로, 수어 데이터셋에 포함된 한국어 수어 라벨을 영어로 매핑한 라벨 리스트를 구성하고, 모델의 예측 결과를 해당 영어 라벨로 변환하여 화면에 출력하였다. 이를 통해 OpenCV의 기본 텍스트 출력 기능을 그대로 활용하면서도 실시간 수어 인식 결과를 안정적으로 시각화할 수 있었다.

처리 속도 측면에서, MediaPipe 랜드마크 추출은 약 30.3 FPS로 동작하며, 10프레임 슬라이딩 윈도우 기반 LSTM 모델의 평균 추론 지연(latency)은 약 70.01ms로 측정되었다. 이에 전체 인식 응답 시간은 1초 이내로 유지

되며, 실시간 학습 피드백 시스템으로서 충분한 처리 속도를 확보하고 있음을 확인하였다.

IV. Experiments

4.1 Experimental Environment

실험은 Table 3에 제시된 환경에서 수행되었다.

Table 3. Experimental Environment

Item	Specification
CPU	11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40GHz
RAM	8.0 GB
GPU	Intel(R) Iris(R) Xe Graphics
Operating System	Windows 10
Programming Language	Python 3.12.4
Deep Learning Framework	TensorFlow 2.18.0
MediaPipe Version	0.10.14
OpenCV Version	4.10.0
Dataset Class	32 KSL vocabulary words
Training / Test Split	80% / 20%

4.2 Comparative Experiment: LSTM vs. GRU

본 연구에서 비교 모델로 GRU를 선택한 이유는 다음과 같다. GRU는 LSTM과 동일한 RNN 계열의 게이트 순환 구조를 기반으로 하면서도, 게이트 수가 적어 파라미터 수가 약 25% 적은 경량 대안 모델이다[11]. 수어 인식 분야의 선행 연구들에서 LSTM과 GRU를 동일한 실험 조건에서 비교한 사례가 보고되고 있어, 본 연구에서도 두 모델의 성능 차이를 정량적으로 비교함으로써 학습 특성과 성능 경향 차이를 분석하고자 하였다. 일반 RNN은 시계열 학습 시 기울기 소실 문제가 발생하여 적합하지 않으며, 양방향 LSTM은 전체 시퀀스를 사전에 필요로 하므로 실시간 스트리밍 환경과 호환되지 않아 비교 대상에서 제외하였다. 공정한 비교를 위해 GRU 모델은 LSTM 모델과 동일한 레이어 구조(256->128), 동일한 하이퍼파라미터(학습률, 배치 크기, 에폭, 정규화 계수), 동일한 학습 데이터를 사용하였다. Table 4에 두 모델의 비교 결과를 제시한다.

Table 4. LSTM vs. GRU

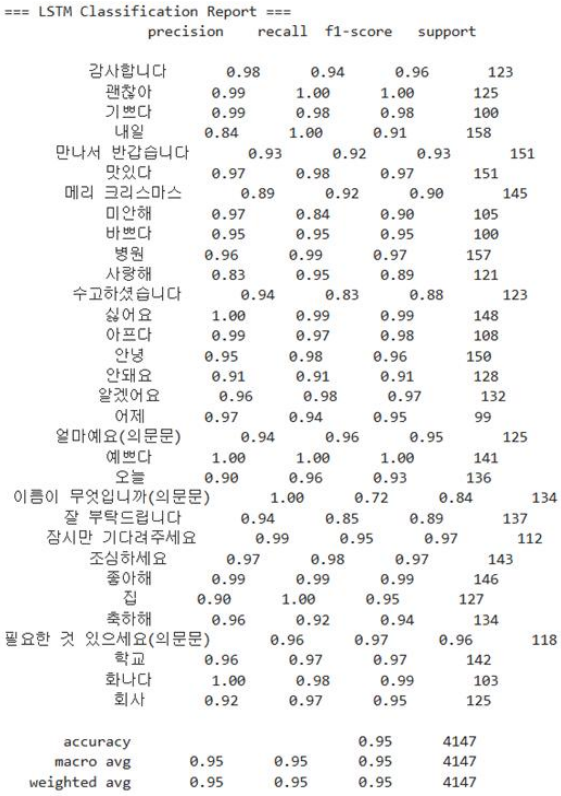
Model	Test Accuracy	Macro F1-score	Macro Precision / Recall
GRU	94.09%	0.94	0.94 / 0.94
LSTM	94.82%	0.95	0.95 / 0.95

제안한 LSTM 모델은 테스트 데이터셋에서 정확도 94.82%, Macro F1-score 0.95를 달성하였으며, 동일 구조의 GRU 모델(정확도 94.09%, Macro F1-score 0.94) 대비 정확도 기준 소폭 높은 성능 경향을 나타냈다.

LSTM이 GRU보다 높은 성능을 보인 주된 이유는 게이트 구조의 차이에서 기인한다. LSTM의 망각 게이트는 이전 정보를 선택적으로 삭제하고, 입력 게이트는 새로운 정보의 반영 정도를 독립적으로 제어하며, 셀 상태를 통해 장기 기억을 별도로 유지한다. 반면, GRU는 업데이트 게이트 하나가 이전 정보 유지와 새 정보 반영을 동시에 담당하므로, 복잡한 시간적 패턴 포착에는 상대적으로 제약이 있다. Fig. 3과 Fig. 4에서 한국수어 어휘 중 '수고하셨습니다'와 같이 다수의 연속적인 손동작과 복잡한 시간적 의존성을 요구하는 어휘에서 GRU 모델의 재현율이 0.72로 크게 저조한 반면, LSTM 모델은 0.83의 재현율을 보여 F1-score 기준 0.11의 차이가 발생한 것은 이를 구체적으로 뒷받침한다. '이름이 무엇입니까(의문문)'의 경우 두 모델 모두 재현율이 0.72~0.75 수준으로 저조하였는데, 이는 게이트 구조와 무관하게 의문문 형태 어휘 간 손동작 패턴의 유사성에 기인한 공통 한계로 분석된다.

4.3 Classification Performance Analysis

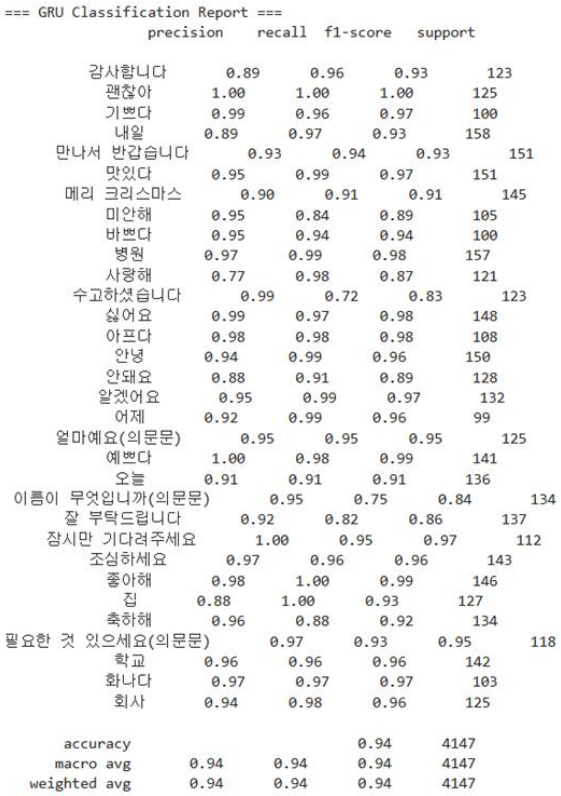
Fig. 3은 제안 모델(LSTM)의 혼동 행렬(Confusion Matrix)을, Fig. 4는 GRU 모델의 혼동 행렬을 나타낸다. 분류 보고서 분석 결과, 전체 32개 어휘 중 '괜찮아' 어휘는 두 모델 모두에서 F1-score 1.00을 달성하였으며, '좋아해'(0.99), '아프다'(0.98), '병원'(0.97) 등 대다수 어휘에서 F1-score 0.97 이상의 높은 분류 성능을 보였다.



=== LSTM Classification Report ===

	precision	recall	f1-score	support
감사합니다	0.98	0.94	0.96	123
괜찮아	0.99	1.00	1.00	125
기쁘다	0.99	0.98	0.98	100
내일	0.84	1.00	0.91	158
만나서 반갑습니다	0.93	0.92	0.93	151
맛있다	0.97	0.98	0.97	151
메리 크리스마스	0.89	0.92	0.90	145
미안해	0.97	0.84	0.90	105
바쁘다	0.95	0.95	0.95	100
병원	0.96	0.99	0.97	157
사랑해	0.83	0.95	0.89	121
수고하셨습니다	0.94	0.83	0.88	123
싫어요	1.00	0.99	0.99	148
아프다	0.99	0.97	0.98	108
안녕	0.95	0.98	0.96	150
안돼요	0.91	0.91	0.91	128
알겠어요	0.96	0.98	0.97	132
어제	0.97	0.94	0.95	99
얼마예요(의문문)	0.94	0.96	0.95	125
예쁘다	1.00	1.00	1.00	141
오늘	0.90	0.96	0.93	136
이름이 무엇입니까(의문문)	1.00	0.72	0.84	134
잘 부탁드립니다	0.94	0.85	0.89	137
잠시만 기다려주세요	0.99	0.95	0.97	112
조심하세요	0.97	0.98	0.97	143
좋아해	0.99	0.99	0.99	146
집	0.90	1.00	0.95	127
축하해	0.96	0.92	0.94	134
필요한 것 있으세요(의문문)	0.96	0.97	0.96	118
학교	0.96	0.97	0.97	142
화나다	1.00	0.98	0.99	103
회사	0.92	0.97	0.95	125
accuracy			0.95	4147
macro avg	0.95	0.95	0.95	4147
weighted avg	0.95	0.95	0.95	4147

Fig. 3. LSTM Classification Report



=== GRU Classification Report ===

	precision	recall	f1-score	support
감사합니다	0.89	0.96	0.93	123
괜찮아	1.00	1.00	1.00	125
기쁘다	0.99	0.96	0.97	100
내일	0.89	0.97	0.93	158
만나서 반갑습니다	0.93	0.94	0.93	151
맛있다	0.95	0.99	0.97	151
메리 크리스마스	0.90	0.91	0.91	145
미안해	0.95	0.84	0.89	105
바쁘다	0.95	0.94	0.94	100
병원	0.97	0.99	0.98	157
사랑해	0.77	0.98	0.87	121
수고하셨습니다	0.99	0.72	0.83	123
싫어요	0.99	0.97	0.98	148
아프다	0.98	0.98	0.98	108
안녕	0.94	0.99	0.96	150
안돼요	0.88	0.91	0.89	128
알겠어요	0.95	0.99	0.97	132
어제	0.92	0.99	0.96	99
얼마예요(의문문)	0.95	0.95	0.95	125
예쁘다	1.00	0.98	0.99	141
오늘	0.91	0.91	0.91	136
이름이 무엇입니까(의문문)	0.95	0.75	0.84	134
잘 부탁드립니다	0.92	0.82	0.86	137
잠시만 기다려주세요	1.00	0.95	0.97	112
조심하세요	0.97	0.96	0.96	143
좋아해	0.98	1.00	0.99	146
집	0.88	1.00	0.93	127
축하해	0.96	0.88	0.92	134
필요한 것 있으세요(의문문)	0.97	0.93	0.95	118
학교	0.96	0.96	0.96	142
화나다	0.97	0.97	0.97	103
회사	0.94	0.98	0.96	125
accuracy			0.94	4147
macro avg	0.94	0.94	0.94	4147
weighted avg	0.94	0.94	0.94	4147

Fig. 4. GRU Classification Report

이처럼 대다수 어휘에서 높은 성능이 나타난 이유는, 해당 어휘들의 수어 동작이 손의 이동 방향, 속도, 정지 위치 등에서 타 어휘와 뚜렷하게 구별되기 때문이다. 예를 들어, '괜찮아'는 턱에 소지를 두드리거나 '예쁘다'는 검지로 볼을 찌르는 듯한 독특한 동작 패턴을 가지고 있어 혼동이 거의 발생하지 않은 것으로 판단된다.

반면 성능이 낮은 어휘를 분석하면 다음과 같다. '이름이 무엇입니까(의문문)'(F1 0.84)는 동일한 의문문 범주의 '얼마예요','필요한 것 있으세요' 등과 손동작의 말단 자세 및 이동 경로가 유사하여 혼동 행렬에서 해당 어휘 간 오분류가 집중적으로 관찰된다. 세 어휘 모두 손을 앞으로 내밀거나 상하 방향으로 움직이는 유사한 말단 동작을 공유하기 때문에 10프레임 내에서 변별적인 시간적 특징을 포착하기 어렵다는 구조적 원인이 있다. '수고하셨습니다'(F1 0.88)는 왼손을 가슴 앞에 눕혀주고 오른손을 사용하는 동작이 '잘 부탁드립니다', '감사합니다' 등 인사 계열 어휘와 유사하여 오분류가 발생하는 것으로 분석된다. 이러한 동작 유사성 문제는 어텐션 메커니즘 도입이나 손목·손가락 각도 기반의 추가 특징 활용을 통해 개선 가능할 것으로 판단된다.

4.4 Graph Analysis

Fig. 5와 Fig. 6은 각각 LSTM 및 GRU 모델의 학습 과정에서 정확도와 손실의 변화를 나타낸다.

LSTM 모델은 학습 초반부터 검증 정확도가 안정적으로 상승하는 양상을 보인 반면, GRU 모델은 0~35 epoch 구간에서 검증 정확도의 진동이 크게 나타났으며 손실 곡선에서도 spike가 반복적으로 관찰되었다. 이는 LSTM의 3중 게이트 구조와 독립적인 셀 상태가 학습 초기 파라미터 업데이트의 분산을 안정적으로 제어하는 반면, GRU는 업데이트 게이트 하나가 기억 유지와 새 정보 반영을 동시에 담당함에 따라 복잡한 시계열 패턴을 학습하는 초기 단계에서 최적화 경로가 불안정해지는 현상으로 해석된다. EarlyStopping 콜백에 의해 두 모델 모두 과적합이 제어되었으며, 최종 테스트 정확도는 LSTM 94.82%, GRU 94.09%로 LSTM은 GRU 대비 소폭 높은 정확도와 보다 안정적인 학습 곡선을 보이는 경향을 나타냈다.

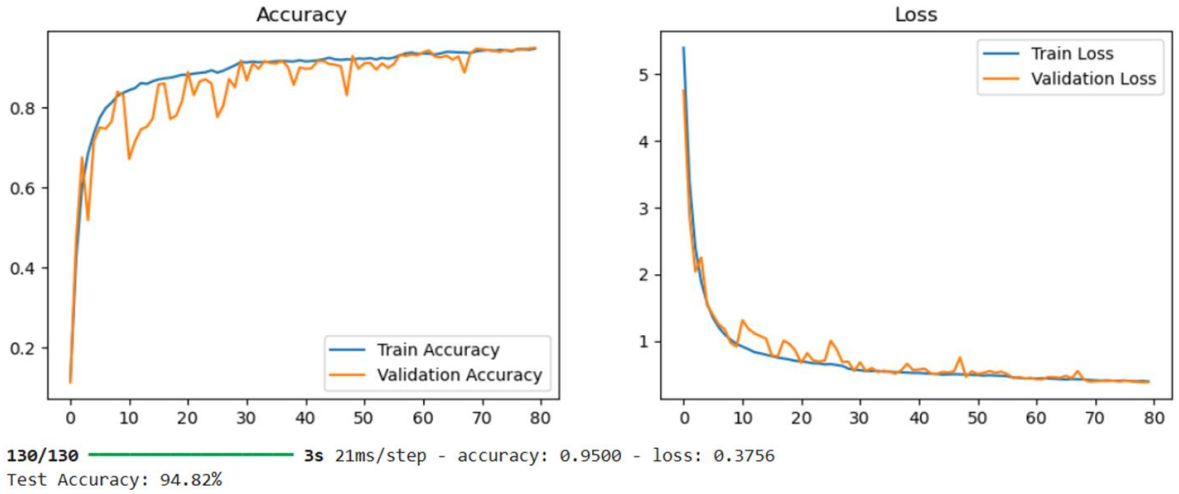


Fig. 5. LSTM Training Accuracy and Loss Curves

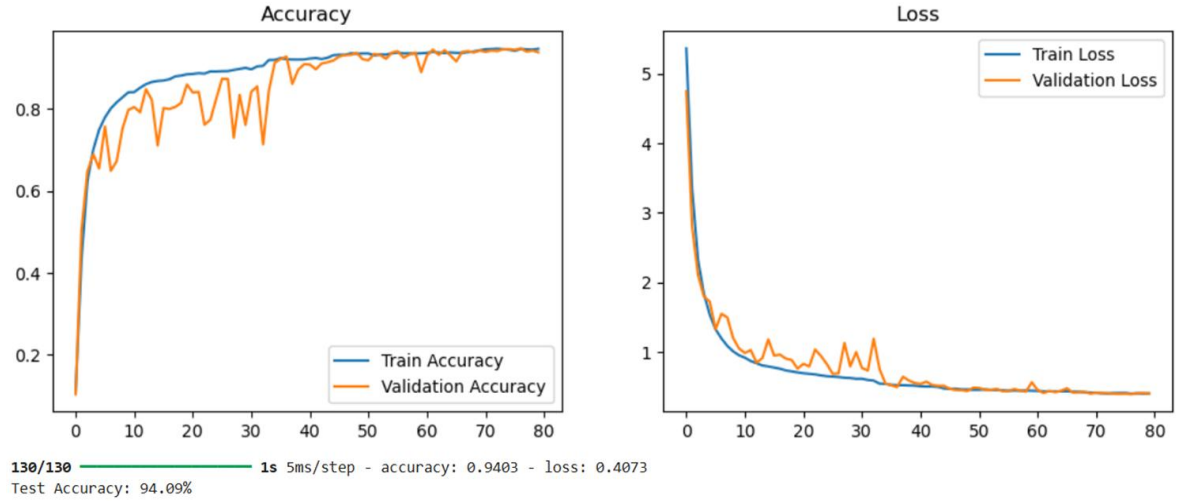


Fig. 6. GRU Training Accuracy and Loss Curves

4.5 Cross-User Evaluation

앞서 보고된 테스트 정확도(94.82%)는 학습 데이터와 동일 사용자의 데이터를 기반으로 한 사용자 종속 성능이므로, 실제 활용 환경에서의 사용자 독립적 일반화 성능을 별도로 확인할 필요가 있다. 이를 위해 학습 데이터 수집에 참여하지 않은 총 3명의 피험자를 대상으로 추가 인식 실험을 수행하였다. 본 실험은 무제한 규모의 사용자 평가를 목표로 하는 것이 아니라, 학습 데이터 미수집 사용자에 대해 제안 시스템이 실용적 수준의 일반화 성능을 보유하는지를 초기 검증하는 것을 목적으로 한다. 본 연구에서는 학습 데이터 수집에 참여하지 않은 총 3명의 피험자를 대상으로 실험을 수행하였으며, 각 피험자는 전체 32개 어휘 중 12개를 선정하여 실시간 인식 환경에서 동작을 수행하였다. 테스트 어휘 12개는 다음 기준에 따라 선정하였다. 첫째, 인사·감사·감정·장소·의문문 등 전체 어휘 범주를 균형 있게 포함하도록 하였다. 둘째, 분류 성능이 높은 어휘($F1 \geq 0.97$)와 상대적으로 낮은 어휘($F1 \leq 0.90$)를 모두 포함하여 일반화 성능을 다각도로 평가할 수 있도록 하였다. 셋째, 피험자당 실험 소요 시간을 적절히 유지하기 위해 전체 32개 어휘 중 37.5%에 해당하는 12개로 범위를 한정하였다.

3명의 평균 인식 성공률은 12개 중 10개(83.3%)로 나타났다. 이는 학습 데이터 기반 테스트 정확도(94.82%) 대비 약 11.5% 낮은 수치로, 이러한 성능 차이는 학습 데이터에 포함되지 않은 사용자의 개인별 수어 동작 편차, 특히 손 크기·동작 속도·서명 스타일의 차이에 기인한 것으로 분석된다. 오인식이 집중된 어휘는 '이름이 무엇입니까', '수고하셨습니다', '잠시만 기다려주세요', '사랑해' 등으로, 4.3절에서 식별된 클래스별 취약 어휘와 일치한다. 이는 해당 어휘들의 동작 유사성 문제가 사용자 간 편차가 가중될 경우 더욱 두드러지게 나타남을 시사한다. 본 예비 실험은 피험자 수(3명) 및 평가 어휘 수(12개)의 제한으로

Table 5. Cross-User Evaluation

Subject	Number of Test Vocabulary Words	Number of Successful Recognitions	Misrecognized Vocabulary
A	12	11	수고하셨습니다
B	12	10	이름이 무엇입니까, 사랑해
C	12	9	이름이 무엇입니까, 잠시만 기다려주세요, 내일

인해 통계적으로 대표성 있는 사용자 독립 검증으로서는 한계를 가진다. 그럼에도 불구하고, 평균 83.3%의 인식률은 제안 모델이 특정 사용자에 과적합되지 않고 미학습 사용자에 대해서도 실용적 수준의 인식 성능을 보유하고 있음을 초기 단계에서 확인하는 근거로 해석할 수 있다. 더 엄밀한 사용자 독립 검증을 위해서는 다양한 신체적 특성을 가진 더 많은 피험자와 전체 어휘를 대상으로 한 확장 실험이 후속 연구에서 필요하다.

V. Conclusions

본 논문에서는 수어 학습 서비스에 특화된 32개 한국수어 어휘 전용 데이터셋을 직접 구축하고, MediaPipe 손 랜드마크 추출과 LSTM 딥러닝 모델을 결합한 실시간 한국수어 학습 인식 시스템을 제안하였다. 다양한 환경에서 수집한 약 68,000개 이상의 시퀀스 데이터를 기반으로 슬라이딩 윈도우 방식의 시퀀스 구성 및 LSTM 분류 모델을 통해 94.82%의 테스트 정확도와 Macro F1-score 0.95를 달성하였다. GRU 모델(정확도 94.09%, Macro F1-score 0.94)과의 비교 실험을 통해 LSTM이 GRU 대비 보다 안정적인 학습 곡선과 소폭 높은 성능 경향을 보였으며, 교차 사용자 실험에서 평균 83.3%의 인식률을 달성하여 초기 수준의 사용자 독립 성능 가능성을 확인하였다. 본 시스템은 웹 서비스와 연동하여 실시간 수어 학습 피드백을 제공함으로써 수어 교육의 접근성 향상에 기여할 수 있다. 본 연구는 인식 정확도 향상 자체에 초점을 맞춘 기존 연구들과 달리, 학습 지원 시스템 구현이라는 실용적 관점에서 수어 교육 서비스로의 실제 적용 가능성을 제시하였다는 점에서 차별화된 의미를 가진다. 최근 시계열 처리 분야에서는 Attention Mechanism 및 Transformer 계열 모델이 수어 인식에 활발히 적용되고 있으며, Attention 기법은 정확도를 추가적으로 4-7% 향상시킬 수 있는 것으로 보고된다. 다만, 이러한 모델들은 대규모 학습 데이터와 높은 연산 비용을 요구하는 한계가 있어, 소규모 단일 수집자 환경 및 경량 실시간 배포를 목표로 하는 본 연구에서는 LSTM을 채택하였으며, 데이터셋 확장 이후 Attention 강화 LSTM 또는 경량 Transformer와의 비교 실험을 후속 과제로 설정한다.

그러나, 본 연구는 다음과 같은 한계를 가진다. 첫째, 학습 데이터를 단일 연구자가 수집하여 데이터셋의 사용자 다양성이 제한적이며, 이로 인해 사용자 간 수어 동작 편차에 대한 강건성이 충분하지 않다. 둘째, 슬라이딩 윈도우 기반 데이터 구성으로 인해 인접 시퀀스 간 높은 상관성이 존재할 가능성이 있으며, 이는 실제 일반화 성능 대비 테스트 정확도를 높게 형성했을 가능성이 있다. 셋째, 교차 사용자 검증에 참여한 피험자가 3명에 불과하여 실험 결과의 통계적 대표성이 제한적이다. 넷째, 인식 어휘 수가 32개로 실제 수어 의사소통에서 활용되는 다양한 어휘를 포괄하지 못한다. 마지막으로, 현재 시스템은 개별 어휘 단위의 인식만을 지원하며 문장 단위 연속 수어 인식에는 적용되지 않는다.

이러한 한계를 바탕으로 향후 연구 방향을 다음과 같이 제시한다. 우선, 다양한 사용자로부터 데이터를 수집하여 데이터셋의 다양성을 확보하고, 교차 사용자 환경에서의 인식 성능을 향상시킬 필요가 있다. 현재 슬라이딩 윈도우 시퀀스 단위로 수행된 학습·테스트 분할을 향후 녹화 회차 단위 분할(recording-level split) 또는 k-fold cross-validation 방식으로 전환하고, 다중 사용자 데이터를 확보하여 보다 엄밀한 사용자 독립적 평가 체계를 구축할 계획이다. 또한, 양손 인식 및 얼굴 표정을 포함한 다중 랜드마크 특징을 활용하여 보다 풍부한 수어 정보를 반영할 수 있을 것이다. 더 나아가, 인식 가능한 어휘 수를 확장하고, 어텐션 메커니즘 기반 유사 동작 어휘 변별력 강화 또는 개인화 학습을 통한 사용자 적응형 인식 모델로의 발전을 통해 문장 단위 연속 수어 인식 시스템으로 확장할 수 있을 것으로 기대된다.

ACKNOWLEDGEMENT

This paper was supported by the National Standard Technology Enhancement Program in 2022 under Project No. RS-2022-KT220734.

REFERENCES

[1] WHO, "Deafness and hearing loss," World Health Organization, https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss, 2023.

[2] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C. Chang, and M. Grundmann, "MediaPipe Hands: On-device Real-time Hand Tracking," arXiv preprint arXiv:2006.10214, Jun. 2020.

[3] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. DOI: 10.1162/neco.1997.9.8.1735

[4] B. Sundar and T. Bagyammal, "American Sign Language Recognition for Alphabets Using MediaPipe and LSTM," Procedia Computer Science, vol. 215, pp. 642–651, Dec. 2022. DOI: 10.1016/j.procs.2022.12.066

[5] G. H. Samaan, A. R. Wadie, A. K. Attia, A. M. Asaad, and A. E. Kamel, "MediaPipe's Landmarks with RNN for Dynamic Sign Language Recognition," Electronics, vol. 11, no. 19, p. 3228, Oct. 2022. DOI: 10.3390/electronics11193228

[6] Y. Wang, R. Li, and G. Li, "Sign Language Recognition Using MediaPipe," in Proc. SPIE 12604, International Conference on Computer Graphics, Artificial Intelligence, and Data Processing (ICCAID 2022), 1260434, May 2023. DOI: 10.1117/12.2674613

[7] K. B. Tran, U. D. Nguyen, and Q. T. Huynh, "Continuous Sign Language Recognition Using MediaPipe," in Proc. 2023 International Conference on Advanced Technologies for Communications (ATC), Da Nang, Vietnam, Oct. 2023, pp. 493–498. DOI: 10.1109/ATC58710.2023.10318855

[8] Y. D. Maheswara, K. Afifah, P. A. Wicaksono, M. A. Al-Sulthon, and N. Prihatiningrum, "Real-Time BISINDO Sign Language Recognition: A Dynamic Approach with GRU and LSTM Models Leveraging MediaPipe," in Proc. 2023 6th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), pp. 226-232, Batam, Indonesia, Dec. 2023. DOI: 10.1109/ISRITI60336.2023.10467586

[9] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. G. Yong, J. Lee, W. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A Framework for Building Perception Pipelines," arXiv preprint arXiv:1906.08172, Jun. 2019.

[10] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proc. 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, May 2015. arXiv preprint arXiv:1412.6980

[11] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," in Proc. EMNLP 2014, pp. 1724–1734, Doha, Qatar, Oct. 2014, DOI: 10.3115/v1/D14-1179

Authors

Jieun Lee is currently an undergraduate student in the Department of Computer Engineering at Gachon University, Korea, and in her fourth year of the B.S. program. Jieun Lee has been studying as an undergraduate researcher at the AI & Smart city Lab in the Department of Computer Engineering at Gachon University, Korea, since 2021. Her research interests include computer vision, autonomous driving, robotics, and deep learning.

Young-Im Cho received the B.S., M.S., and Ph.D. degrees in Computer Science and Engineering from Korea University, Seoul, Republic of Korea, in 1989, 1991, and 2004, respectively. She was a Postdoctoral Researcher at the University of Massachusetts Amherst, Amherst, MA, USA, from 1999 to 2000. Professor Cho is currently a Professor in the Department of Computer Engineering at Gachon University, Korea. She has been the Director of the AI and Smart City Laboratory at Gachon University since 2015. Her research interests include artificial intelligence, multi-agent systems, lightweight AI, smart cities, multimodal systems, brain–computer interface (BCI) systems, and international AI standards.