ISSN 2383-6318(Print) / ISSN 2383-6326(Online)
KIISE Transactions on Computing Practices, Vol. 31, No. 9, pp. 407-414, 2025. 9
https://doi.org/10.5626/KTCP.2025.31.9.407

딥러닝과 은닉 마르코프 모델을 연계한 연속적인 손 제스처 적출에 관한 연구

(A Study on Continuous Hand Gesture Spotting Integrating Deep Learning and Hidden Markov Models)

이 현 규 †　　박 재 흥 ††
(Hyeonkyu Lee)　(Jaeheung Park)

요 약 본 논문은 실시간 제스처 인식기를 위해 딥러닝과 은닉 마르코프 모델(HMM)을 결합한 연속 손동작 기반 제스처 적출(spotting) 방법을 제안한다. 3차원 공간에서 손의 형상과 움직임은 MediaPipe Hands 알고리즘으로 추적되며, 이 데이터를 바탕으로 LSTM Autoencoder가 HMM 입력용 특징벡터를 생성한다. 시공간적 변이를 포함한 제스처 탐지는 Gaussian HMM을 통해 수행된다. 또한 입력패턴이 제스처와 얼마나 유사한지 판단하기 위해 비제스처 유사도를 기반으로 한 임계치 모델을 도입하였다. 실험 결과, 본 방법은 연속 손동작에서 제스처를 98.08%의 인식률과 5.60%의 단어오류율(WER)로 적출하여 높은 신뢰성을 입증하였다.

키워드: 인공지능, 머신러닝, 제스처 인식, 패턴인식, 은닉 마르코프 모델, 임계치 모델

Abstract This paper presents a method for spotting gestures from continuous hand movements by combining deep learning with Hidden Markov Models (HMM) for real-time gesture recognition. The MediaPipe Hands algorithm tracks hand shapes and trajectories in 3D space, and the resulting data is processed by an LSTM Autoencoder to generate feature vectors for HMM input. Gestures with spatiotemporal variations are identified using a Gaussian HMM. To evaluate the similarity between input patterns and gestures, a threshold model based on non-gesture similarity is introduced. Experimental results demonstrate that this method achieves a gesture recognition rate of 98.08% and a word error rate (WER) of 5.60%, indicating high reliability.

Keywords: artificial intelligence, gesture recognition, gesture spotting, pattern recognition, Hidden Markov Model, threshold model

· 이 논문은 인천대학교 2024년도 자체연구비 지원에 의하여 연구되었음.

† 종신회원 : 인천대학교 컴퓨터공학부 교수(Incheon Nat'l Univ.)
hyeonkyulee@inu.ac.kr
(Corresponding author임)
†† 학생회원 : 인천대학교 무역학부 학생
james2p@naver.com

논문접수 : 2025년 7월 22일
(Received 22 July 2025)
논문수정 : 2025년 8월 20일
(Revised 20 August 2025)
심사완료 : 2025년 8월 22일
(Accepted 22 August 2025)

정보과학회 컴퓨팅의 실제 논문지 제31권 제9호(2025. 9)

1. 서 론

최근 HCI 분야에서는 직관적이고 비언어적인 소통방식에 대한 관심이 급증하고 있으며[1,2], 특히 손 제스처 인식은 가상현실[3], 스마트홈[4], 수화통역 시스템[5] 등 다양한 응용에서 핵심기술로 부각되고 있다. 그러나 실시간 환경에서 연속 손동작에서 의미있는 제스처를 정확히 탐지하고 인식하는 적출(spotting) 문제는 여전히 기술적 과제로 남아 있다. 이는 손동작의 시공간적 변이가 크고, 제스처 간 경계가 모호하며, 일상적 손동작 중 비제스처가 실제 제스처보다 훨씬 많이 발생하는 등 복합적인 요인에 기인한다.

이러한 문제를 해결하기 위해 다양한 컴퓨터 비전 및 머신러닝 기반 접근법이 제안되어 왔다. 특히, 손의 위치와 형태를 정밀하게 추적할 수 있는 구글의 MediaPipe Hands[6,7] 같은 3차원 손추적 알고리즘이 등장하면서 실시간 손동작 인식을 위한 입력 데이터의 정확도가 크게 향상되었다. 또한, 시계열 데이터의 복잡한 패턴을 효과적으로 모델링하는 LSTM Autoencoder[8]와, 시공간 변이에 강한 은닉 마르코프 모델(HMM; Hidden Markov Model)[9,10]이 제스처 인식의 주요 기법으로 주목받고 있다.

본 연구는 특징추출에 강점을 가진 딥러닝과 시계열 패턴 인식에 적합한 HMM을 결합해, 연속 손동작 흐름에서 의미있는 제스처를 실시간으로 적출하는 하이브리드 손 제스처 적출기를 제안한다. 제안된 시스템은 MediaPipe Hands로 손의 3차원 움직임을 추적하고, LSTM Autoencoder로 손동작 시퀀스를 저차원 특징벡터로 변환한 후, 이를 Gaussian HMM[10,11]에 입력한다. 또한, 인식성능 향상을 위해 전체 손동작 중 비제스처 구간을 명확히 구분할 수 있도록 임계치 모델(threshold model)을 도입하여[12,13], 비제스처를 배제하고 입력패턴이 사전 정의된 제스처와 얼마나 유사한지를 평가한다.

본 논문의 주요 기여는 다음과 같다. 첫째, MediaPipe Hands 기반 손추적과 LSTM Autoencoder를 결합해 영상에서 안정적으로 특징을 추출하는 구조를 제안했다. 둘째, Gaussian HMM과 임계치 모델을 활용한 실시간 제스처 탐지방법을 제시했다. 셋째, 실제 카메라 데이터셋을 대상으로 98.08%의 제스처 인식률과 5.60%의 단어오류율(WER; Word Error Rate)을 달성해, 연속 손동작에서 의미있는 제스처를 신뢰성 있게 분리하고 식별하였다.

이 연구는 제스처 기반 사용자 인터페이스의 실용화를 촉진할 수 있으며, 향후 수어인식을 통한 수어번역 시스템[14]을 비롯한 다양한 실시간 HCI 시스템에 적용 가능한 기술적 기반을 제공할 것으로 기대된다.

2. 손 제스처 적출 연구현황

손 제스처 인식은 사용자의 의도나 명령을 직관적으로 전달하는 중요한 비언어 인터페이스로, 다양한 분야에서 활발히 연구되어 왔다. 특히 연속 손동작에서 의미있는 제스처를 자동으로 추출하고 인식하는 적출 문제는 제스처 인식의 핵심과제 중 하나로[12,13,15], 단순분류 문제보다 높은 난이도를 요구한다.

기존 연구는 시계열 모델 기반 접근과 딥러닝 기반 접근으로 나뉜다. 손 제스처는 본질적으로 시계열 특성을 가지므로, HMM이나 CRF(Conditional Random Field) 같은 통계적 시퀀스 모델이 주로 활용되어 왔다[12,16]. HMM은 상태 간 전이확률과 관측확률을 이용해 시공간 변이가 큰 입력 시퀀스를 효과적으로 모델링할 수 있으나, 입력으로 스칼라 값을 사용하므로 고차원 특징정보를 벡터양자화 과정에서 단순화해야 하며, 이로 인해 표현력이 제한되고 고차원 데이터의 직접 활용에는 한계가 있다.

이러한 한계를 극복하기 위해 최근에는 LSTM(Long Short-Term Memory)과 GRU(Gated Recurrent Unit) 기반의 딥러닝 시퀀스 모델이 도입되어 적출 성능이 향상되었다[17,18]. 특히 LSTM Autoencoder는 복잡한 시계열 입력의 잠재표현(latent representation)을 효과적으로 추출할 수 있으며, 이를 이용해 제스처 시퀀스를 비제스처로부터 분리하는 다양한 방법이 제안되었다[8]. 그러나 대부분의 연구는 입력구간이 제스처일 것이라는 가정을 전제로 하며, 연속 입력에서 정의된 제스처인지 여부를 판단하는 데에는 제약이 있다.

연속 손동작에는 의미없는 비제스처 구간이 많아 이를 명확히 구분하기 위해 임계치 모델이 활용된다[12,13]. 임계치 모델은 일반적으로 HMM 기반 제스처 모델과 함께 사용되며, 입력 시퀀스가 학습된 제스처 패턴과 얼마나 유사한지를 판단하는 기준 역할을 한다. 즉, 유사도가 사전에 정의된 임계치 이상일 경우에만 제스처 후보로 간주된다. 하지만 기존 연구들 대부분은 임계치 모델을 별도로 학습하지 않거나 단일 기준만 적용해, 제스처 경계 인식의 정교성과 신뢰도가 제한되는 경우가 많았다.

3. 본 연구에 사용된 주요 기술

본 연구는 연속적인 손동작 흐름에서 의미있는 제스처를 효과적으로 적출하기 위해, 컴퓨터 비전과 시계열 모델링 기법을 결합한 하이브리드 인식방법을 제안한다. 이를 위해 MediaPipe Hands, LSTM Autoencoder, Gaussian HMM, 임계치 모델의 네 가지 핵심기술을 도입하였다.

3.1 MediaPipe Hands를 이용한 손 랜드마크 추출〔6,7〕

MediaPipe Hands는 Google에서 개발한 실시간 손 추적 프레임워크로, 단일 RGB 영상입력만으로도 고정밀 3차원 손 랜드마크(landmark)를 추출할 수 있는 경량 알고리즘이다. 이 프레임워크는 손바닥 검출과 손 관절위치 추정의 두 단계로 구성된다.

먼저 손바닥 검출단계에서는 BlazePalm 기반 객체 검출기를 통해 입력 프레임에서 손의 전체 위치를 빠르게 찾아 바운딩 박스를 생성한다. 이 박스는 이후 정밀한 랜드마크 추정의 초기영역으로 사용된다.

다음 단계에서는 해당 바운딩 박스를 입력으로 정밀 회귀모델을 적용하여, 손가락 관절과 손바닥 기준점을 포함한 21개의 랜드마크 좌표(x, y, z)를 정규화된 3차원 값으로 예측한다. MediaPipe Hands는 BlazePalm 기반 CNN 구조 덕분에 빠르고 안정적인 실시간 손 검출이 가능하며, 추출된 좌표는 해상도나 손크기 변화에 강인하다.

그림 1은 Mediapipe Hands를 이용하여 예측한 손 랜드마크를 시각화한 예이다.

본 연구에서는 매 프레임마다 추출된 21개의 랜드마크를 시간 순으로 배열해 손의 움직임과 형태의 변화를 시계열 데이터로 구성하였다. 이를 통해 손동작의 시공간적 정보를 효과적으로 표현할 수 있는 기반 데이터를 확보하였고, 이후 LSTM Autoencoder 입력에 적합한 고차원 벡터로 변환하였다. 또한 손가락 간 상대 거리와 방향 정보를 일관되게 유지하기 위해 중심정렬, 크기 정규화, 저대역 필터링 등의 전처리를 수행하였다.

3.2 LSTM Autoencoder를 이용한 손동작 시퀀스의 특징추출〔8〕

LSTM Autoencoder는 순환신경망(RNN)의 확장구조로, 시계열 데이터의 장기 의존성(long-term dependency)을 학습할 수 있다. 일반적인 Autoencoder는 입력 데이터를 저차원 공간에 압축한 뒤 이를 복원하는 구조이며, LSTM Autoencoder는 이를 시계열 데이터에 특화한 형태이다.

그림 1 Mediapipe Hands를 이용한 손 랜드마크 추출
Fig. 1 Visualization of the extracted landmarks

본 연구에서는 다층 LSTM 인코더와 디코더로 구성된 모델을 사용하였다. 인코더는 손 랜드마크 시퀀스를 고정차원의 잠재벡터(latent vector)로 압축하고, 디코더는 이 벡터를 기반으로 원래 시퀀스를 재구성한다.

학습 시에는 손동작 시퀀스를 입력하여 복원된 시퀀스와의 MSE(Mean Squared Error)를 최소화하도록 파라미터를 최적화하였다. 학습 완료 후 인코더의 출력 벡터는 손동작 시퀀스의 시공간 정보를 함축하는 의미 있는 특징으로 간주되며, 이후 HMM의 입력으로 사용된다.

3.3 Gaussian HMM을 이용한 제스처 모델링〔9-11〕

HMM은 관측 시퀀스가 내부적으로 전이하는 숨겨진 상태의 확률적 전개에 의해 생성된다는 가정에 기반한 시계열 모델이다. 본 연구에서는 각 상태의 출력분포를 다변량 Gaussian 분포로 설정해, LSTM Autoencoder로부터 생성된 고차원 연속 특징벡터를 효과적으로 모델링하였다.

HMM은 다음 네 요소로 정의된다:

- $Q=\{q_1,q_2,\cdots,q_N\}$: 숨겨진 상태집합
- $\pi=\{\pi_i\}$: 초기상태 확률, $\pi_i=P(s_1=q_i)$
- $A=\{a_{ij}\}$: 상태전이 확률, $a_{ij}=P(s_{t+1}=q_j|s_t=q_i)$
- $B=\{b_i(x)\}$: 관측 확률분포. 여기서는 Gaussian 분포 $N(x;\mu_i,\Sigma_i)$

관측 시퀀스 $X=(x_1,x_2,\cdots,x_T)$에 대해 전체 확률은 다음과 같이 계산된다:

$$P(X,S)=\pi_{s_1}\cdot b_{s_1}(x_1)\cdot\prod_{t=2}^{T}a_{s_{t-1},s_t}\cdot b_{s_t}(x_t)$$

각 제스처 클래스 G_k에 대해 독립적인 HMM 모델 $\lambda_k=(\pi_k,A_k,B_k)$을 학습하며, Baum-Welch 알고리즘[9]으로 파라미터를 최적화하였다. 상태 수 N, 반복 횟수, 수렴조건 등은 사전 실험을 통해 조정하였다.

테스트 단계에서는 입력 시퀀스에 대해 각 모델의 로그 유사도 $\log P(X|\lambda_k)$를 계산하고, 가장 가능성이 높은 제스처 클래스를 선택하였다:

$$\hat{k}=\arg\max_k \log P(X|\lambda_k)$$

로그 유사도는 Forward 알고리즘[9]으로 계산하였으며, 이는 시퀀스의 정확한 상태열 복원보다는 적출 성능에 중점을 둔 것이다.

Gaussian HMM은 분류뿐 아니라 연속 손동작 시퀀스에서 제스처의 시작과 종료를 탐지하는 데도 활용된다. 이를 위해 슬라이딩 윈도우 방식으로 구간별 유사도를 평가하고, 유사도가 급상승하는 구간에서 제스처의 존재를 식별하였다. 이후 단계에서는 오탐지를 줄이기 위해 임계치 모델을 적용하여 비제스처 구간을 추가로 필터링하였다.

3.4 임계치 모델을 이용한 비제스처 구간 필터링 [12,13]

연속적인 손동작에는 무의미하거나 과도기적인 비제스처 구간이 포함되며, 이는 제스처 적출 과정에서 오탐지를 유발하는 주요 원인이다. 본 연구는 이 문제를 해결하기 위해 임계치 모델 기반의 필터링 기법을 도입하였다. 임계치 모델은 입력 시퀀스가 특정 제스처에 해당하는지를 유사도에 로그를 취한 값인 로그 유사도를 기준으로 판단한다.

각 제스처 모델 λ_k의 로그 유사도 $\log P(X|\lambda_k)$를 사전에 정의된 임계치 θ_k와 비교하며, 다음과 같이 평가한다:

$$\text{is\_gesture}(X,\lambda_k) = \begin{cases} 1 & \text{if } \log P(X \mid \lambda_k) \geq \theta_k \\ 0 & otherwise \end{cases}$$

임계치 θ_k는 다음과 같이 계산된다:

$$\theta_k = \mu_k - \alpha \cdot \sigma_k$$

여기서 μ_k는 학습 데이터의 평균 로그 유사도, σ_k는 표준편차, α는 신뢰구간 조정 상수 (본 연구에서는 1.0)이다.

그림 2는 입력된 손동작 시퀀스가 제스처 모델 및 임계치 모델과 각각 얼마나 잘 일치하는지를 나타낸 로그 유사도(유사도의 로그값)를 시각화한 그래프이다. 이를 통해 해당 시퀀스가 제스처로 인식 가능한지를 판단할 수 있다.

본 연구에서는 HMM의 내재적 분할(internal segmentation) 특성을 활용하여, 상태 전이 순서가 달라도 일정 유사도를 유지하는 ergodic 모델을 임계치 모델로 구성하였다. 이는 제스처의 다양한 부패턴 조합에 유연하게 대응하며, 특정 제스처 모델보다 낮은 유사도를 출력해 적응형 임계치로 기능한다.

제스처 적출 단계에서는 슬라이딩 윈도우 기반으로 로그 유사도를 계산하고, 유사도가 기준 이상인 구간만 제스처 후보로 판단한다. 이 방식은 오탐지를 줄이고 시스템의 신뢰도를 향상시킨다. 정의 불가한 비제스처 패턴을 학습하는 garbage 모델과 달리, 임계치 모델은 특정 제스처에 대한 절대 유사도 기준으로 작동하여 적출의 안정성을 높여준다. 본 방법은 사전정의되지 않은 제스처나 비제스처 입력에 대해서도 유연하게 대응하며, 향후 다양한 입력유형으로의 구조 확장이 가능하다.

마지막으로, 본 연구에서는 모든 제스처 모델과 임계치 모델을 하나의 제스처 적출 네트워크(Gesture Spotting Network)로 통합하여, 연속적인 손동작 입력에서 제스처를 지속적으로 적출할 수 있도록 하였다. 이 네트워크는 가상의 시작 상태 S에서 출발하여, 모든 제스처 모델과 임계치 모델을 하나의 흐름으로 연결한 통합 구조로, 무한히 입력되는 손동작 시퀀스 중 사전 정의된 제스처를 효과적으로 탐지할 수 있도록 구성되었다. 그림 3은 본 연구에서 구축한 제스처 적출 네트워크의 구조를 보여준다.

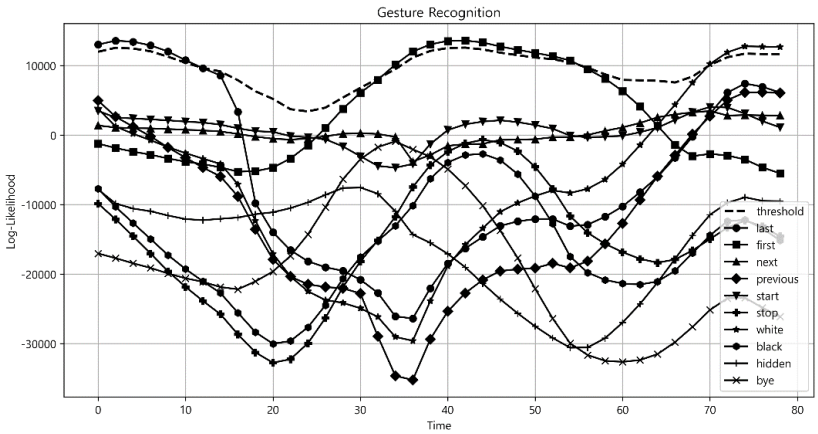


그림 2 제스처 모델과 임계치 모델의 로그 유사도 그래프
Fig. 2 Log-likelihood graph of gesture models and threshold model

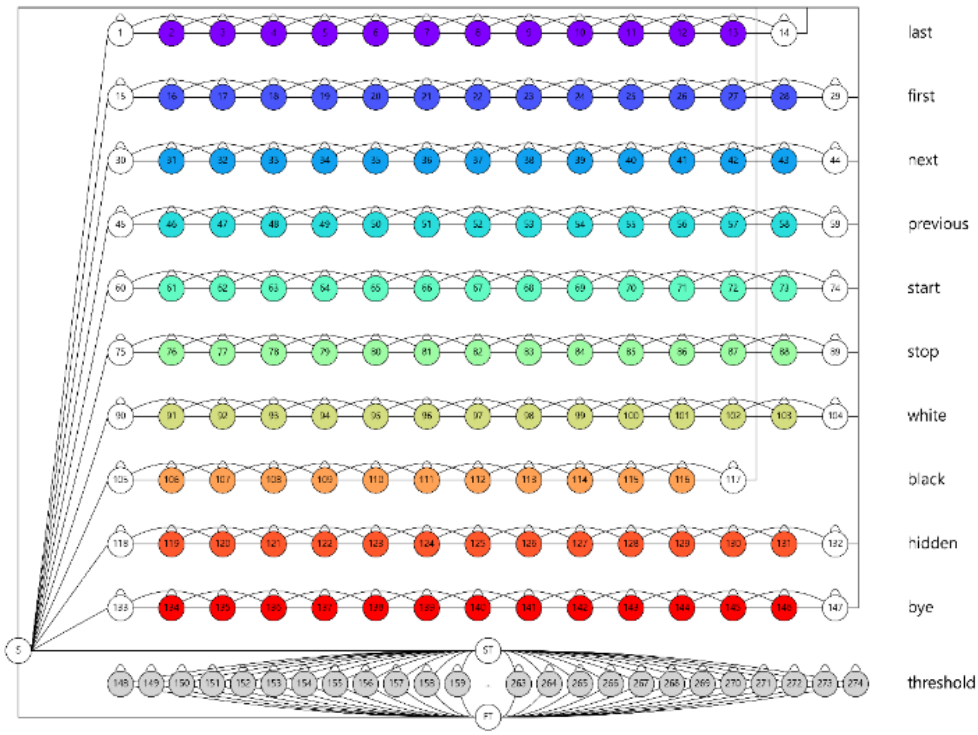


그림 3 제스처 적출 네트워크의 구조
Fig. 3 Structure of the gesture spotting network

4. 실험 및 결과

과거 제스처 적출 연구에서는 손의 깊이(카메라에서의 거리) 변화나 손가락 움직임을 탐지하지 못해, 손 전체의 움직임을 2차원 평면에 투영하고 궤적 정보만을 활용하는 방식이 주로 사용되었다[12,13]. 그러나 본 연구는 딥러닝 기반 MediaPipe Hands 프레임워크를 적용함으로써 3차원 손동작을 정밀하게 추적할 수 있었고, 이를 통해 보다 다양하고 풍부한 형태의 제스처 정의가 가능해졌다.

본 연구에서는 Microsoft PowerPoint 제어를 위한 10개의 명령어 제스처를 정의하였다. 선정된 제스처는 단순한 상하좌우 이동뿐만 아니라 손의 깊이방향 움직임과 손가락의 세부동작까지 포함한다. 손동작의 수행 용이성과 명령 전달의 명확성을 고려하여, 오른손만 사용하는 단순한 수어 동작 중에서 선택하였다. 실험에 사용된 제스처 목록은 표 1에 제시하였으며, 각 제스처 모델은 100개의 학습 데이터를 기반으로 학습되었다.

시스템 성능평가는 두 단계로 나누어 진행하였다. 첫 번째 단계에서는 단일 제스처만을 포함된 평가 데이터를 활용해 제스처 모델의 학습효과와 임계치 모델의 유사도 기준이 적절히 작동하는지를 검증하였다(제스처 변별력 평가). 두 번째 단계에서는 복수의 제스처가 포함된 데이터를 이용해 연속적인 손동작 흐름에서의 제스처 적출(탐지 및 인식) 성능을 평가하였다(제스처 적출 평가). 제스처 적출 평가는 연속 입력흐름에서 제스처 존재구간을 탐지하는 정확도와, 각 구간 내 제스처 인식 정확도를 동시에 측정하는 방식으로 이루어졌다.

4.1 제스처 변별력 평가

제스처 변별력 평가는 단일 제스처 데이터를 이용해 제스처 모델과 임계치 모델의 분류 성능 및 변별력을 검증하는 단계이다. 본 연구에서는 각 제스처별로 100개의 학습 데이터와 20개의 독립된 평가 데이터를 사용하였으며, 전체 데이터 구성은 표 2와 같다.

평가결과는 표 3에 요약되어 있으며, 유사한 제스처가 다수인 어려운 조건에서도 전체 인식율은 98.50%로 높은 수준을 기록하였다. 대부분의 오인식은 제스처의 길이가 짧고(20~25 프레임) 동작간 유사성으로 인해 발생

표 1 실험에 사용한 제스처
Table 1 Gestures used in the experiment

Gesture	Hand Shape	Action	Description
last			With the index and middle fingers extended and abducted to form a V-shape, the fingertips are positioned adjacent to the eyes and subsequently lowered in a downward motion.
first			With the right hand slightly flexed, the fingertips are placed on the chest and then rotated in a full circular motion toward the left.
next			With the fingertips of the slightly flexed right hand placed on the chest, the hand performs a complete circular motion toward the left.
previous			With the hand open, the lateral side of the extended index finger is placed vertically against the chin and moved slightly side to side.
start			With the thumb and index fingertips initially joined, they are gradually separated during the downward motion, then brought closer together, and finally rejoined at the end.
stop			With the hand lightly clenched and facing forward, it moves quickly from the right side of the face to the front as the index and middle fingers extend into a V-shape.
white			With the thumb and index fingertips joined to form a circular shape, the hand is placed below the nose and then lowered while the fingers are extended.
black			With the thumb and index fingertips brought together, the hand moves downward in a single smooth motion.
hidden			With the fingers held together and pointing upward, the hand repeatedly opens and closes twice.
bye			With the thumb, index, and middle fingers extended and the dorsum of the hand facing outward, the wrist is gently shaken downward.

표 2 제스처 모델 학습 및 평가용 데이터
Table 2 Training and test data for the experiment

Gesture	Train Data	Test Data
last	100	20
first	100	20
next	100	20
previous	100	20
start	100	20
stop	100	20
white	100	20
black	100	20
hidden	100	20
bye	100	20
Sum	1,000	200

표 3 제스처 변별력 평가결과
Table 3 Gesture recognition rates from the experiment

Gesture	Test Data	Correct	Recognition Rate
last	20	20	100.00%
first	20	20	100.00%
next	20	20	100.00%
previous	20	20	100.00%
start	20	20	100.00%
stop	20	20	100.00%
white	20	18	90.00%
black	20	20	100.00%
hidden	20	19	95.00%
bye	20	20	100.00%
Sum	200	197	98.50%

하였다. 예를 들어, white는 주먹을 살짝 쥔 상태에서 코 근처에 대었다가 떼는 동작이며, hidden은 주먹을 쥔 상태에서 손가락을 펴는 동작으로, 두 동작이 유사하게 시작되어 간섭이 발생하였고, 이로 인해 제스처 모델의 유사도가 임계치 기준을 넘지 못하였다.

4.2 제스처 적출 평가

제스처 적출 평가는 연속적인 손동작 흐름에서의 제스처 탐지 및 인식 성능을 종합적으로 평가하기 위해 설계되었다. 총 121개의 평가 데이터에는 3~4개의 제스처가 포함되어 있다. 제스처 존재구간 탐지 성능은 Segmental F1-score[19], 구간 내 인식성능은 인식률과 WER [20]로 평가하였다.

Segmental F1-score 평가결과는 표 4에 요약되어 있다.

제스처 적출과정에서 발생하는 에러는 삽입, 삭제, 대체의 세 가지 형태로 분류된다. 이 중 삽입에러는

표 4 연속 제스처 Spotting 실험의 F1-score
Table 4 F1-scores from the gesture spotting experiment

Gesture	True Positive	True Negative	False Positive	False Negative	Precision	Recall	F1-score
last	31	3	0	0	1.000	1.000	1.000
first	32	0	0	1	1.000	0.970	0.985
next	32	0	0	1	1.000	0.970	0.985
previous	40	0	1	0	0.976	1.000	0.985
start	43	0	0	0	1.000	1.000	1.000
stop	35	0	0	0	1.000	1.000	1.000
white	37	0	1	0	0.974	1.000	0.987
black	39	1	0	0	1.000	1.000	1.000
hidden	31	0	0	0	1.000	1.000	1.000
bye	37	1	0	0	1.000	1.000	1.000

표 5 연속 제스처 Spotting 실험의 인식률과 WER
Table 5 Recognition rates and WER from the gesture spotting experiment

Gesture	Test Data	Insertion Error	Deletion Error	Substitution Error	Correct	Recognition Rate	WER
last	34	2	3	0	31	91.18%	16.13%
first	33	2	0	1	32	96.97%	9.38%
next	33	1	0	1	32	96.97%	6.25%
previous	40	2	0	0	40	100.00%	5.00%
start	43	0	0	0	43	100.00%	0.00%
stop	35	1	0	0	35	100.00%	2.86%
white	37	3	0	0	37	100.00%	8.11%
black	40	1	1	0	39	97.50%	5.13%
hidden	31	0	0	0	31	100.00%	0.00%
bye	38	1	1	0	37	97.37%	5.41%
Sum	364	13	5	2	357	98.08%	5.60%

F1-score 계산에는 포함되지 않지만, 실제 제스처 일부 또는 전체가 잘못 제거되는 결과를 유발해 시스템 성능 저하로 이어질 수 있다. 본 연구에서는 이를 보완하기 위해 WER을 종합 성능지표로 활용하였다. WER 계산식은 다음과 같으며, 평가결과는 표 5에 정리하였다.

$$\mathrm{WER} = \frac{\mathrm{Insertion\ Error} + \mathrm{Deletion\ Error} + \mathrm{Substitution\ Error}}{\mathrm{Correct}}$$

평가과정에서 last와 first는 동작방식은 다르지만 모두 검지와 중지를 사용하는 제스처로서, 준비동작의 형태가 유사하다. 이로 인해, 상대적으로 길이가 짧은 last는 연결동작을 last로 오인식하는 삽입에러와 연결동작과 실제 제스처를 하나의 제스처로 오인식하는 삭제에러가 발생하였다. 반면, first는 손가락 움직임이 더 복잡한 동작으로서 다른 제스처가 first로 오인식되는 삽입에러 및 다른 제스처와 혼동되는 대체에러가 관찰되었다. 또한, 동작이 단순한 white의 경우에도 연결동작이 white로 오인식되는 삽입에러가 일부 테스트 데이터에서 발생하였다.

이처럼 제스처 간 유사한 시작 동작과 짧은 동작 길이(약 20~25 프레임)는 로그 유사도 기반의 임계치 판단에 영향을 주며, WER 성능 저하의 주요 원인이 되었다.

5. 결 론

실험 결과, 제안된 방법은 평균 98.08%의 높은 인식률과 5.60%의 낮은 WER을 기록하여, 연속 손동작 흐름에서 뛰어난 제스처 추출 및 인식 성능을 보였다. 이는 기존 연구에 비해 제스처 경계 탐지와 인식의 신뢰도가 크게 향상된 결과로 평가된다.

또한, 본 연구에서 적용한 임계치 모델은 구조가 단순함에도 입력패턴이 제스처와 얼마나 유사한지를 효과적으로 판단하는 기능을 보여주었다. 다만, 제스처 모델 수가 증가하면 임계치 모델의 상태 수도 함께 증가하여 적출속도가 저하되는 문제가 발생하였다. 이에 따라, 향후에는 모델 수에 무관하게 임계치 모델의 상태 수를 일정하게 유지할 수 있는 경량화 기법 연구가 필요할 것이다.

References

[1] J. Urakami and K. Seaborn, "Nonverbal cues in human-robot interaction: A communication studies perspective", *ACM Transactions on Human Robot Interaction,* Vol. 12, Article 1-21, 2023.

[2] R. Z. Khan and N. A. Ibraheem, "Hand Gesture Recognition: A Literature Review", *International Journal of Artificial Intelligence & Applications,* Vol. 3, Issue 4, pp. 161-174, 2012.

[3] G. Buckingham, "Hand Tracking for Immersive Virtual Reality: Opportunities and Challenges", *Frontiers in Virtual Reality,* Vol. 2, Article 728461, 2021.

[4] S. Dewangga, M. Subianto, and W. Swastika, "Implementation of Hand Gesture Recognition as Smart Home Devices Controller", *INSYST: Journal of Intelligent System and Computation,* Vol. 6, Issue 2, pp. 63-68, 2024.

[5] B. Fang, J. Co and M. Zhang, "DeepASL: Enabling Ubiquitous and Non-intrusive Word and Sentence-level Sign Language Translation", *Proc. of the 15th ACM Conference on Embedded Network Sensor Systems,* pp. 1-13, 2018.

[6] C. Lugaresi, J. Tang, et al., "Mediapipe: A Framework for Building Perception Pipelines", *arXiv preprint arXiv:1906.08172,* 2019.

[7] F. Zhang, V. Bazarevsky, et al., "Mediapipe Hands: On-device Real-time Hand Tracking", *arXiv preprint arXiv:2006.10214,* 2020.

[8] P. Malhotra, A. Ramakrishnan, et al., "LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection", *arXiv preprint arXiv:1607.00148,* 2016.

[9] X. D. Huang, Y. Ariki, and M. A. Jack, *Hidden Markov Models for Speech Recognition,* Edinburgh Univ. Press, 1990.

[10] C. Gruhl and B. Sick, "Variational Bayesian Inference for Hidden Markov Models With Multivariate Gaussian Output Distributions", *arXiv preprint arXiv:1605.08618,* 2016.

[11] K. Zheng, D. Shi, and L. Shi, "Learning Hidden Markov Models for Linear Gaussian Systems with Applications to Event-based State Estimation", *Automatica,* Vol. 128, Article 109560, 2021.

[12] H. K. Lee and J. H. Kim, "An HMM-Based Threshold Model Approach for Gesture Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI),* Vol. 21, No. 10, pp. 961-973, 1999.

[13] H. K. Lee, H. Y. Kim, and J. H. Kim, "A Study on the Threshold Model for Spotting Hand Gestures Based on Hidden Markov Model", *Journal of KIISE(B),* Vol. 25, No. 1, pp. 150-159, 1998. (in Korean)

[14] A. Sinha, C. Choi, and K. Ramani, "DeepHand: Robust Hand Shape Classification with CNN for Real-time Hand Gesture Recognition", *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 4150-4158, 2016.

[15] P. Neto, D. Pereira, et al., "Real-time and Continuous Hand Gesture Spotting: An Approach Based on Artificial Neural Networks", *Proc. of the IEEE International Conference on Robotics and Automation (ICRA),* pp. 178-183, 2013.

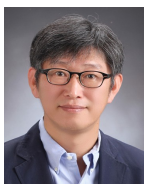
[16] S. B. Wang, A. Quattoni, et al., "Hidden Conditional Random Fields for Gesture Recognition", *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR),* Vol. 2, pp. 1521-1527, 2006.

[17] G. Chen, J. Chen, et al., "FLGR: Fixed Length Gists Representation Learning for RNN-HMM Hybrid-based Neuromorphic Continuous Gesture Recognition", *Frontiers in Neuroscience,* Vol. 13, Article 73, 2019.

[18] K. Cho, Y. Bengio, et al., "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation", *Proc. of 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* pp. 1724-1734, 2014.

[19] K. Renz, N. Stache, et al., "Sign Language Segmentation with Temporal Convolutional Networks", *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 2135-2139, 2021.

[20] F. Koller, O. Zargaran, and H. Ney, "Enabling Robust Statistical Continuous Sign Language Recognition: A CNN-HMM hybrid with end to end training", *International Journal of Computer Vision,* Vol. 127, Issue 8, pp. 1097-1116, 2019.



이 현 규

1985년 서울대학교 컴퓨터공학과 졸업(학사) 1987년 KAIST 전산학과 졸업(석사) 1998년 KAIST 전산학과 졸업(박사) 1991~1999년 핸디소프트 기술이사 2003~2007년 아이크로스테크놀로지 대표이사 2007~2010년 네이버 모바일센터장/이사 2011~2014년 KT 오픈플랫폼개발본부장/상무, 2018~2020년 KAIST 스마트에너지인공지능연구센터 교수, 2020~2024년 IITP 과기정통부 인공지능/빅데이터 PM, 2024~현재 인천대학교 컴퓨터공학부 교수. 관심분야는 인공지능, 생성형 AI, 패턴인식, 컴퓨터 비전, 제스처 인식

박 재 홍

2019년~현재 인천대학교 무역학부(주전공)/컴퓨터공학부(복수전공) 재학 중. 관심분야는 인공지능, 컴퓨터 비전, 제스처 인식, 공급망 관리