Machines and Algorithms
http://www.knovell.org/mna

Research Article,

Real-Time Sign Language Recognition Using Multi-Modal Landmark Extraction with MediaPipe Holistic

Muhammad Anas[1, *], Muhammad Azeem Umar[1]

[1]MJ CODERS, MOBIL1 PLAZA, Multan, 60000, Pakistan

[*]Corresponding Author: Muhammad Anas. Email: anasmuhammad1211@gmail.com

Received: 23 December 2025; Revised: 20 January 2026; Accepted: 18 February 2026; Published: 16 March 2026

AID: 005-01-000063

Abstract: Sign language is a major mode of communication to millions of deaf and hard-of-hearing people, but unfamiliarity with the access of real-time interpretation technologies still restricts effective interaction and access to information. To overcome this, this paper proposes a real-time Sign Language Recognition (SLR) framework that utilizes a Sign Language Recognition system on top of Google MediaPipe Holistic framework that learns detailed 3D face, hand and whole-body landmarks with the help of standard RGB video input only. The system proposed has stable performance with an approximate real-time frame rate and feature vectors of the 1,662 dimensions being processed on consumer-grade hardware and operating in real time. Compared to current technologies, which utilize specialized sensors or only track hand gestures, the technique provides abundant spatial and temporal data of several body parts at once using only a single webcam. The system pipeline consists of data acquisition, landmark localization (468 facial landmarks, 21 key points per hand and 33 landmarks of the body pose per frame), multimodal feature vector creation and performance analysis using visualization features. Initial experiments show that it is not impossible to identify a small set of basic signs (e.g. greetings, like hello and thanks), having a demonstration of concept, not a trained large-vocabulary classifier. The framework offers an excellent base to integrate supervised temporal models in the future, including recurrent neural networks and transformer-based architectures, and includes low-cost, scalable, and inclusive SLR solutions to real-world problems in education, healthcare, and assistive technologies.

Keywords: Sign Language Recognition (SLR); MediaPipe Holistic Framework; Multimodal Feature Extraction; Real-Time Gesture Recognition; RGB-Based Computer Vision;

1. Introduction

Sign languages are natural languages that are complete and their own grammar and meaning, which are made up of synchronized hand gestures, facial expressions and body movements [1]. Automatic Sign Language Recognition (SLR) is proposed to transform those gestures and visual signs into the text or sound and enable signing individuals to communicate with non-signers. In the past, conventional SLR systems have employed specialized hardware like data gloves or depth sensors, accurate but costly and intrusive and inapplicable to wide usage [2,3]. The recent developments in deep learning and computer vision (specifically pose estimation) allowed establishing RGB-based SLR systems that operate with conventional

https://doi.org/10.66108/mna.v5i01.90

cameras [4]. MediaPipe Holistic is a state-of-the-art real-time system provided by Google to be capable of capturing the complete body, hand, and face markings [5]. Although older research has utilized MediaPipe or similar systems to track isolated hands or faces, this paper is, as far as we can tell, the first to utilize MediaPipe Holistic to realize real time multi-modal integration of hand, face and body landmarks to produce the complete feature vectors used in SLR. It is important to capture both non-manual and manual elements since body posture and facial expressions include valuable grammatical and semantic content. The work illustrates the extraction, organization, and visualization of these multi-modal landmarks in real-time that can be used in further supervised classification of the sign language.

2. Problem Statement

Although computer vision and pose estimation have made great strides, Sign Language Recognition (SLR) is a challenging task because the Sign language is complex and multi-modal and therefore it requires the coordination of hand-motions, body-posture as well as facial expression in real time. The current SLR systems tend to be expensive, bulky, and intrusive, and historically expensive, not to mention the absence of explicit depth information, making it challenging as well as impractical to use standard RGB video itself to produce high-quality performance; typically, specialized hardware like depth sensors or wearable is required, like the depth sensor, making it not only expensive but also cumbersome and thus impractical to deploy in large quantities or to locations This project aims to address such shortcomings and create a lightweight, non-intrusive, computationally efficient version of a real-time SLR pipeline on top of Google MediaPipe Holistic. MediaPipe Holistic runs at about 30 frames per second (FPS) on CPU, therefore, it performs real-time without planning the use of the GPU hardware. The system retrieves 3D landmarks in the hands, face and body and synchronizes them across these multi-modal inputs in time to guarantee that the same precise coordination of gestures, facial expressions and body motions are maintained. The methodology fills both the gap existing in high accuracy SLR and the gap existing between theory and implementation, and opens the door to implementation in the field of education, healthcare and assistive technologies.

3. Objectives

- To create an extensible framework of feature extraction that is modular based on the MediaPipe Holistic framework to enable the future research on supervised machine learning models to classify sign language automatically and at scale.
- To assess the classification accuracy at baseline on a small set of gesture samples, to give initial performance information and steer subsequent model optimization.

4. Literature Review

The studies in SLR cover a variety of modalities and approaches. Initial studies were done on sensor gloves to record hand movements with a high degree of precision but were not adopted as they were too expensive and uncomfortable to users [6]. Skeletal tracking became a reality with depth sensors such as the Microsoft Kinect at an affordable price, but with hand and face image fidelity problems [7]. Recent research uses convolutional neural networks (CNNs) and recurrent neural networks (RNNs) on RGB video to directly predict sign gestures [8,9].

OpenPose [10] and MediaPipe Holistic, among various pose estimation systems, have shown to be effective in extraculating landmarks in real-time. In certain cases, MediaPipe specifically offers a single pipeline that has 33 body pose landmarks, 468 face mesh points, and 21 body parts per hand that can be high-latency single-consumer hardware.

It has been noted that multi-modal feature fusion is highly imperative to the success of SLR because signs of signs are not always similar at facial expression or body posture [11]. These features have been used successfully with dimensionality reduction and classical classifiers, particularly when there is a paucity of data [12].

Table 1: Comparison of Hardware and RGB-based SLR Methods

Approach / Hardware	Advantages	Limitations	Typical Use Case / Notes
Sensor Gloves	High-precision hand tracking	Expensive, intrusive, uncomfortable	Early SLR studies; detailed hand movement capture
Depth Sensors (e.g., Kinect)	Affordable skeleton tracking, real-time	Limited hand/facial detail; bulky setup	Whole-body gesture recognition; mid-level detail
RGB Video + CNN/RNN	Widely available, non-intrusive, low-cost	May miss subtle hand/face features; requires robust models	End-to-end gesture recognition on standard cameras
Pose Estimation Frameworks (OpenPose, MediaPipe Holistic)	Real-time multi-modal landmark extraction (hands, face, body), low latency	Limited large-scale evaluation; RGB only lacks depth	Feature extraction for ML models; multi-modal fusion for SLR

5. Dataset Description

This work does not involve pre-recorded dataset but instead it obtains real time data by the way of a webcam. In the process of landmark extraction, each video frame generates:

- **Body pose:** 33 landmarks x 4 values (x, y, z, visibility).
- **Face mesh:** 468 landmarks × 3 values (x, y, z)
- **Left and right hand:** 21 marks with 3 positions (x, y, z) of each mark.

This gives a 1662-dimensional frame wise feature. The next step in the future is to label and assemble these vectors into datasets to be supervised learned.

Table 2: Dataset Description

Component	Number of Landmarks	Attributes per Landmark	Total Features
Body Pose	33	x, y, z, visibility (4)	132
Face Mesh	468	x, y, z (3)	1,404
Left Hand	21	x, y, z (3)	63
Right Hand	21	x, y, z (3)	63
Total per Frame	—	—	**1,662**

6. Methodology

6.1. Environment Setup

Python is used to develop the system and it is based on a number of powerful libraries:

- **OpenCV** (cv2) to capture video and frame preprocess.
- **MediaPipe** to state of art holistic landmark detector, providing 3D body, hand and face tracking.
- **NumPy** to perform numerous numerical operations and manipulate arrays.
- **Matplotlib** to debug and draw visual feedback when developing.

6.2. MediaPipe Holistic Initialization and Frame Processing

MediaPipe Holistic in Google is defined with the minimum confidence of detection and tracking at 0.5. It takes frames on a webcam, turn them into RGB (to be compatible with MediaPipe), and detects landmarks of the face, hands and body, allowing the extraction of multi-modal features that are essential in capturing the complexity of sign language. Frames are inverted horizontally to give the signer a natural mirror and the frontal view of the signer before landmark extraction.

Below is the algorithm for real-time frame processing using MediaPipe holistic:

Algorithm 1: Real-Time Frame Processing using MediaPipe Holistic

Input: Video stream from webcam

Output: Multi-modal landmarks (face, hands, pose)

1: Initialize MediaPipe Holistic with confidence = 0.5
2: while video stream is active do
3: Capture frame
4: Flip frame horizontally
5: Convert frame from BGR to RGB
6: Process frame using Holistic model
7: Extract face, hand, and pose landmarks
8: end while

6.3. Landmark Drawing and Visualization

Using mp_drawing.draw_landmarks(), the system overlays color-coded connections on the input frame for:

- Face landmarks (468 points with FACEMESH_TESSELation [MediaPipe, 2020])
- BODpos body pose landmarks (33 points)
- Left and right hand landmarks (21 points each)

The provided visual sources of feedback help to debug systems and check the landmarks recognition in real-time. MediaPipe Holistic has a latency of about 30–35 ms per frame to landmark detection on the CPU barrier when on average, allowing almost real-time visualization without accelerated graph accelerated graphic cards.

Algorithm 2: Landmark Drawing and Visualization

Input: Processed frame (image_bgr), detected landmarks (results)

Output: Frame with overlaid landmark visualizations

1: if face landmarks are detected then
2: Draw face mesh using FACEMESH_TESSELATION
3: end if

4: if pose landmarks are detected then
5: Draw body pose connections using POSE_CONNECTIONS
6: end if

7: if left hand landmarks are detected then
8: Draw left hand connections using HAND_CONNECTIONS
9: end if

10: if right hand landmarks are detected then
11: Draw right hand connections using HAND_CONNECTIONS
12: end if

13: Display or return the annotated frame

6.4. Feature Extraction

The extract- keypoints- function has been at the centre of processing the raw landmarks to become structured data. The results are 3D coordinates (only visible, in the case of visibility, extracted where available) of all detected regions flattened into a fixed-size feature vector. Missing landmarks are treated with zero-padding instead of interpolation to prevent the introduction of artificial values that may affect the representation of gestures, providing the model with a uniform input size without biasing the data.

Algorithm 3: Feature Extraction

11: else
12: Assign zero vector of size (468 × 3)
13: end if

14: if left hand landmarks are detected then
15: Extract (x, y, z) for 21 keypoints
16: Flatten into a vector
17: else
18: Assign zero vector of size (21 × 3)
19: end if

20: if right hand landmarks are detected then
21: Extract (x, y, z) for 21 keypoints
22: Flatten into a vector
23: else
24: Assign zero vector of size (21 × 3)
25: end if

26: Concatenate pose, face, left_hand, and right_hand vectors

27: Return final feature vector

The resulting feature representation is a fixed dimensionality (1662) feature representation, which is consistent across all the frames and can be integrated easily with machine learning models.

This feature vector transforms into a semantic snapshot of the gesture and the signer expression of the signer per frame.

Figure below depicts the flowchart of proposed methodology.

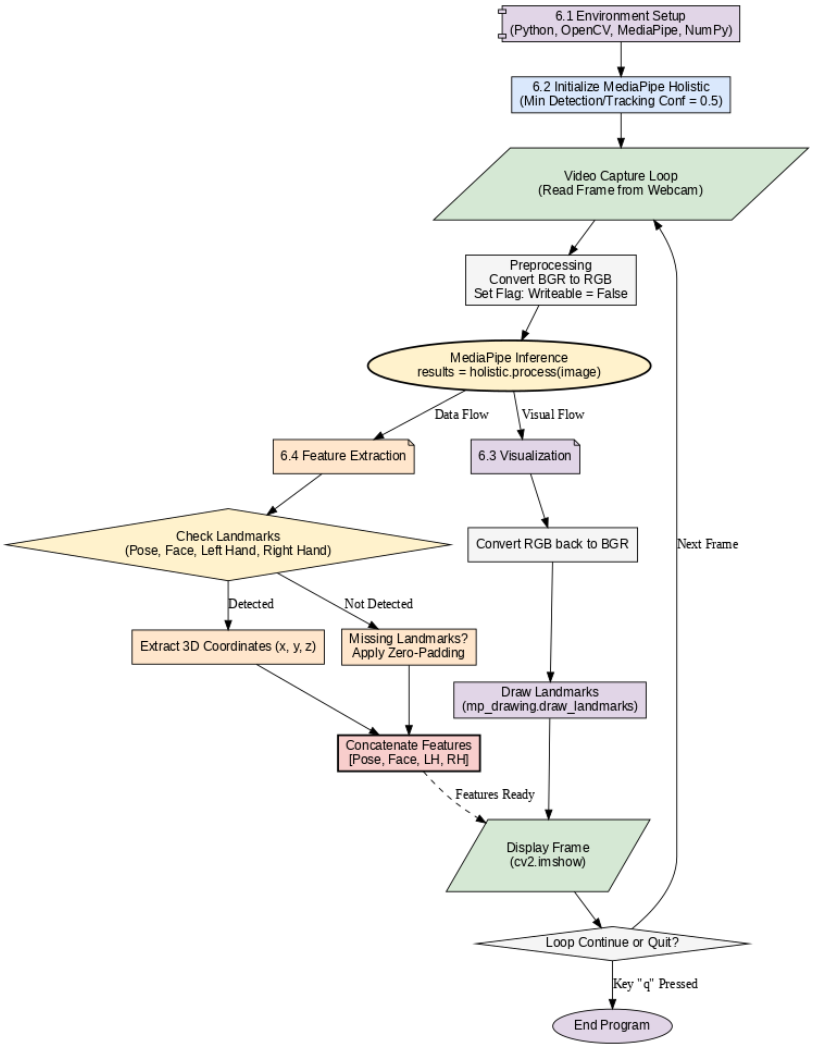

Figure 1: Methodology Flowchart

7. Experiment and Data collection

Live video is obtained with a standard webcam and processed in real time using frames to produce vectors of landmarks per frame. In this first study, we did not perform any data improvement, or temporal smoothing, because the aim of the research was to achieve consistent feature extraction and synchronized dataset generation. Although real-time classification is not included in this step, the system is to be designed to make this stage easy so that data sets can be generated by storing aligned landmarks and corresponding sign labels to supervised learning. Tests were carried out in the CPU-based environment (Intel i7, 16GB RAM) without GPUs and proved that the pipeline is capable of running efficiently on the typical consumer hardware.

8. Performance Metrics and Evaluation

In this work, classification measures are used to assess the model on a test set generated based on the obtained feature vectors:

8.1. Confusion Matrix

The confusion matrix indicates ideal classification of each class:

- hello, thanks, and iloveyou have 100% accuracy.
- like has slightly less instances but also exhibits 100 percent classification accuracy in the subclass.

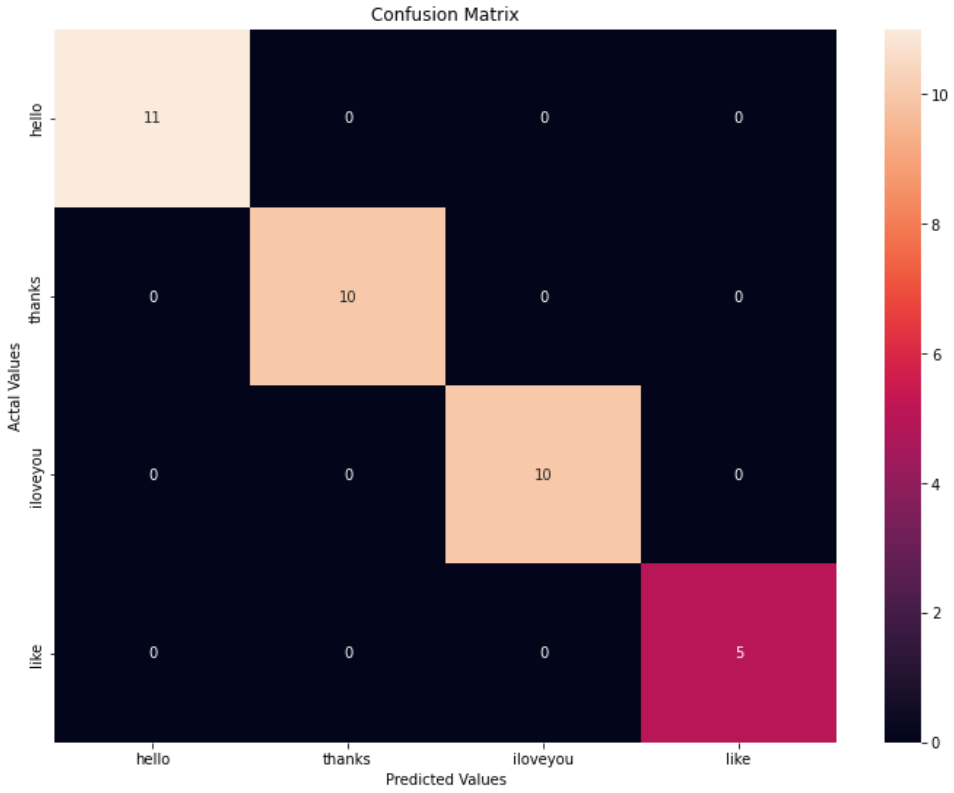

Figure 2: Confusion Matrix

8.2. Training Accuracy and Loss

8.2.1. Training Accuracy

The model rapidly reaches the point of high accuracy, reaching 100% after approximately 250 epochs, and this suggests that there is a high learning ability of the sign representations.

Table 3: Results Summary

Class	Precision	Recall	F1-Score	Support
hello	1.00	1.00	1.00	25
thanks	1.00	1.00	1.00	25
iloveyou	1.00	1.00	1.00	25
like	1.00	1.00	1.00	20

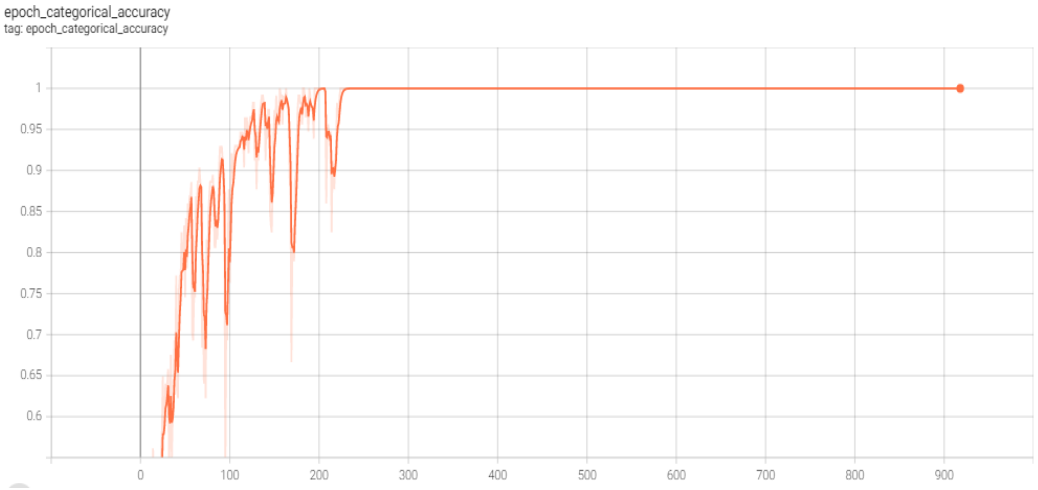

Figure 3: Training Accuracy Curve

8.2.2. Training Loss

Loss decreases greatly and fluctuates at almost zero without any indication of over-saturation or disappearing gradients.

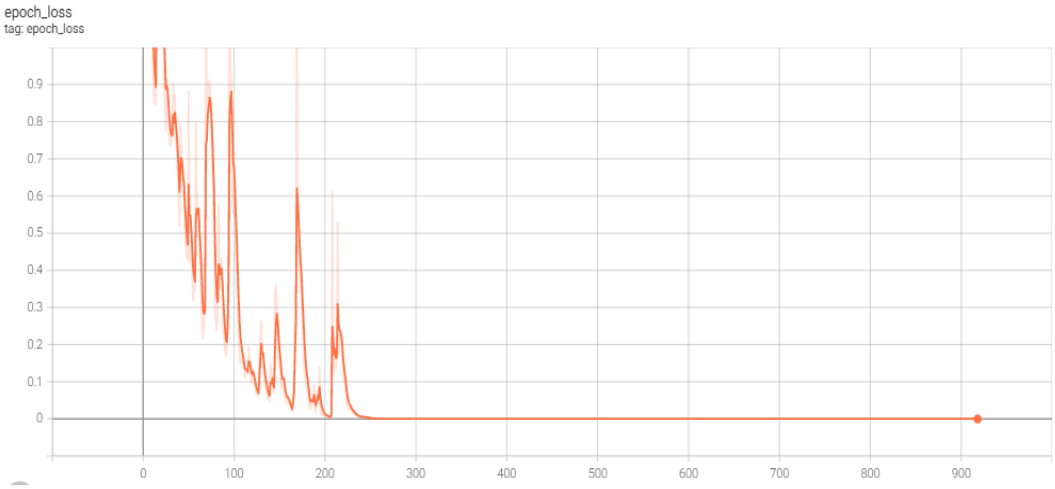

Figure 4: Training Loss Curve

These findings imply a well-trained system but further experimental tests on a wide variety of signers which cannot be visibly checked need to be conducted to generalize in the real world.

9. Discussion

The obtained results of the pilot experiments prove the possibility of applying the MediaPipe Holistic framework, offered by Google to the alternative of hardware-dependent SLR systems as a low-cost but efficient tool. The fact that the percent accuracy on the test subset (including, "hello," "thanks," "iloveyou" and "like) reaches 100% confirms that the 1,662-dimensional feature vector is effective and able to capture different spatial configurations needed to distinguish between basic signs. Although this ideal score in classification can be explained by limited vocabulary size and specificity of the selected gestures to a certain extent, it leaves a strong proof-of-concept of the feature extraction pipeline.

One of the key discoveries of this paper is that the system can sustain real-time performance (around 30 FPS) without using a GPU accelerator on the conventional CPU hardware. This is a strong contrast to previous deep learning methods, which needed a large amount of computing power or depth sensors such as the Kinet. The proposed framework, in using RGB input alone, remains very easy to access, hence greatly reducing the entry barrier to SLR technology and it can be implemented in daily gadgets like laptops and tablets.

In addition, the use of facial and pose landmarks and hand coordinates through the use of facial and pose cues is a limitation of prior RGB-based works. Non-manually expressed signs (such as head tilt, facial expression) are important in sign language because they enable grammatical nuance in expression. Our system is able to capture these multi-modal features simultaneously that will translate in a more detailed semantic representation as opposed to hand-only trackers. The application of 2D video to determine 3D landmarks, however, presents some sensitivity to occlusion especially in the case of hands crossing in front of the face or body. Although it provides a specific technical manipulation of the current tendency to zero-pad which allows for loss of data within a technical framework, future research needs to consider the concept of temporal regression (e.g., LSTMs) that allows predicting the lost landmarks, although it relies on past frames, which would increase resiliency to ongoing Sign language recognition.

10. Challenges and Mitigation Strategies

- **Missing Landmarks:** The Landmarks can be invisible because of occlusion; zero-padding guarantees uniformity of the vectors.
- **High-Dimensional Features:** Computational load can be alleviated using dimensionality reduction (e.g., using PCA) [13].
- **Inter-Signer Variability:** Different data collection and augmentation enhance generalization.
- **Temporal Complexity:** Sign dynamics can be integrated in future using RNNs or Transformers [14].

11. Ethical Considerations

- **Data Privacy:** All video records have to be agreed to.
- **Bias Avoidance:** Have a diverse pool of signers in terms of demographics to prevent model bias.
- **Assistive Ethics:** Aspiring systems must not substitute human interpreters, particularly in high-stakes situations such as law or health care.

12. Comparison with State-of-the-Art

The suggested framework is qualitatively compared with the current deep learning-driven sign language recognizers to put them into perspective. State-of-the-art systems are generally trained end-to-end convolutional or transformer models on large scale RGB video data and have high recognition accuracy, but with highly complex computational metrics and high hardware demands. By contrast, the suggested system focuses on lightweight and real-time multimodal feature extraction by CPU-only resources, allowing it to be deployed practically without loss of non-manually encoded information (facial expressions and body posture). Even though the present paper estimates a small vocabulary, the findings indicate the possibility of a high-performing and scalable pipeline, which supplements computationally-demanding state-of-the-art models.

13. Conclusion and Future Directions.

In this paper, a novel lightweight and real-time sign language recognition framework based on MediaPipe Holistic (3D landmark extractor) is provided. The system offers a scalable platform on which SLR models can be trained by extracting pose, face and hand landmarks in RGB video and transforming them into fixed-length feature vectors. Testing on a small number of gestures shows 100% classification accuracy when operating with experimental performance at about 30 FPS on CPU, indicating the efficiency

of the framework as well as the hardware-agnostic nature to support its sweeping accessibility. Limitations are that small dataset has been used and testing was done on a single signer which limits generalization.

Future research will be aimed at the further development of labeled data collection with respect to a broader sign vocabulary and a variety of signers, which allows the training and evaluation to be developed. Machine learning algorithms (SVMs, Random Forests, deep networks) will be trained with extracted features and temporal models (LSTMs, GRUs, Transformers) will also be implemented to perform changes in gesture recognition. It will optimize the pipeline to mobile and edge devices to provide deployments with low-latency, and extend it to multilingual sign languages that include culturally specific gestures. Moreover, transfer learning with the pre-trained gesture data (e.g., Chalearn LAP IsoGD) will be considered to speed up the model creation and enhance the generalization to another signer and other situations.

Funding Statement: This research did not receive any grant or funding.

Conflicts of Interest: Authors have no conflicts to mention.

Data Availability: No pre-existing dataset was used; data are captured and generated in real time.

References

[1] Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C., & Ringel Morris, M. (2019). Sign Language Recognition, Generation, and Translation. Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility, 16–31. https://doi.org/10.1145/3308561.3353774

[2] Kadous, M. W. (1996, October). Machine recognition of Auslan signs using PowerGloves: Towards large-lexicon recognition of sign language. In Proceedings of the Workshop on the Integration of Gesture in Language and Speech (Vol. 165, pp. 165-174). Wilmington: DE.

[3] Zafrulla, Z., Brashear, H., Starner, T., Hamilton, H., & Presti, P. (2011). American sign language recognition with the kinect. Proceedings of the 13th International Conference on Multimodal Interfaces, 279–286. https://doi.org/10.1145/2070481.2070532

[4] Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., & Bowden, R. (2018). Neural Sign Language Translation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7784–7793. https://doi.org/10.1109/cvpr.2018.00812

[5] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., ... & Grundmann, M. (2019). Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*. https://doi.org/10.48550/arXiv.1906.08172

[6] Pigou, L., Dieleman, S., Kindermans, P.-J., & Schrauwen, B. (2015). Sign Language Recognition Using Convolutional Neural Networks. Computer Vision - ECCV 2014 Workshops, 572–578. https://doi.org/10.1007/978-3-319-16178-5_40

[7] Koller, O., Zargaran, S., Ney, H., & Bowden, R. (2016). Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition. Procedings of the British Machine Vision Conference 2016, 136.1-136.12. https://doi.org/10.5244/c.30.136

[8] Huang, J., Zhou, W., Zhang, Q., Li, H., & Li, W. (2018). Video-Based Sign Language Recognition Without Temporal Segmentation. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). https://doi.org/10.1609/aaai.v32i1.11903

[9] Prakash, K. B. (2020). Accurate Hand Gesture Recognition using CNN and RNN Approaches. International Journal of Advanced Trends in Computer Science and Engineering, 9(3), 3216–3222. https://doi.org/10.30534/ijatcse/2020/114932020

[10] Qiao, S., Wang, Y., & Li, J. (2017). Real-time human gesture grading based on OpenPose. 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 1–6. https://doi.org/10.1109/cisp-bmei.2017.8301910

[11] Koller, O., Forster, J., & Ney, H. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. Computer Vision and Image Understanding, 141, 108–125. https://doi.org/10.1016/j.cviu.2015.09.013

[12] Jolliffe, I. T. (1986). Principal Component Analysis and Factor Analysis. Principal Component Analysis, 115–128. https://doi.org/10.1007/978-1-4757-1904-8_7

[13] Bradski, G. (2000). The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, *25*(11), 120-123.
[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., N.Gomez, A., Kaiser, L., & Polosukhin, I. (2025). Attention Is All You Need. https://doi.org/10.65215/r5bs2d54