Grenze International Journal of Engineering and Technology, June Issue

Real-Time Sign Language Recognition using MediaPipe Landmark Pipeline and Deep Learning Classifiers

Anirudh Kumar[1], Shreya[2] and Deepak Singh[3]
[1-3]ABES Engineering College/CSE-DS, Ghaziabad, India
Email: arudh706@gmail.com, shreya3343@gmail.com, dsingh99906@gmail.com

***Abstract*— Communication between normal individuals and people with hearing and speech disabilities is often difficult due to the lack of an effective interaction method. This paper proposes a real-time sign language recognition system using a MediaPipe landmark pipeline and deep learning classifiers. The system uses a standard webcam to capture video frames and extracts 21 hand landmarks using MediaPipe Hands. These landmarks are converted into feature vectors and processed by a dual-classifier model consisting of a Multilayer Perceptron (MLP) for static gestures and a Long Short-Term Memory (LSTM) network for dynamic gestures. The proposed approach provides accurate and efficient real-time recognition of sign language gestures. Experimental evaluation on ASL and ISL datasets shows that the system achieves high accuracy and low latency, making it suitable for real-time communication applications.**

***Index Terms*— Sign Language Recognition, Mediapipe Hands, Hand Landmark Detection, Deep Learning, MLP, LSTM, Real-Time Gesture.**

I. Introduction

We already know the demand and continuously advancement in artificial intelligence and also in computer vision has opened to understand the sign language detection and understanding. In present time if someone wants to develop or make efficient and reliable sign language detection and recognition system. A system that works on high accuracy and maintains the consistent and reliable performance across different environment and varying conditions but we can say that the approaches we have in present time do not maintains all the criteria we mentioned. These systems often face challenges in low end conditions like in balancing speed, accessibility and precision at the time of deployment on standard computer systems. Our primary objective was to overcome these types of limitations. We are introducing our proposed system which can add the extension to the traditional sign language recognition systems. Our proposed system evaluated on standard datasets such as Indian Sign Language (ISL) and American Sign Language (ASL) these two are very famous datasets in the market. Our aim is to achieve the bestest real time sign language recognition with high accuracy and precision and also ensure the smooth operation work on basic general systems. The motivation behind this system arises from real life challenges faced by many individuals who belongs to the deaf and hearing communities. In our society, these problems represent the communication barrier between normal people and these people. This gap affects individuals in various critical situations.

Grenze ID: 01.GIJET.12.2.120_20

These situations are basic for normal people but for deaf and hearing communities, It is not that much simple. Let's take an example- If a normal person wants to describe the health symptoms to a doctor it can be easily done through verbal movement but for hearing and deaf people is a very crucial challenge to make the doctor understand their legit problem. Due to the lack of effective communication bridge between them, These everyday basic moments often lead to misunderstanding and social isolation. We want to overcome and minimize these barriers through our proposed system which consist of efficient, understandable and practical approach. Our system is not like the traditional sign language recognition system, Which works on complex sensors and high-end computer resources.

II. Literature Review

Classical Computer Vision: The mismatch between the problems of the data and the tooling to solve them drove the field to pursue a more scalable and non-intrusive tool: the camera. The first generation of vision systems used the methodology of classical computer vision. This was an improvement in that a conceptual engineer could manually design explicit features to isolate the hands from the background (e.g., via Skin Colour Segmentation) and to describe their shapes using feature descriptors like the Histogram of Oriented Gradients HOG. Unfortunately, these methods were also brittle. Al-Ani and Al-Jumaily [12] provided a comprehensive survey of skin color modeling and detection techniques that became foundational for early gesture recognition systems, on the other hand Pławiak and Rzecki [14] and A. Kumar et al. [1] demonstrated the use of Histogram of Oriented Gradients (HOG) combined with Support Vector Machines (SVM) for static hand gesture recognition in Indian Sign Language (ISL).

The Deep Learning Revolution with Raw Images The real revolution didn't occur until recently, when the deep learning revolution with Convolutional Neural Networks (CNNs) hit the tipping point. At this point, we no longer needed those fragile, hand-engineered features. We could let a CNN learn meaningful features from the raw pixels of an image directly. As a result, recognition accuracy of individual, static gestures - static, single signs, jumped several orders of magnitude. There was a huge downside to this though: cost.

In addition to learning static images of hands, a CNN was spatially aware - it learned what a hand looked like. However, a CNN was lacking one thing inherent to the nature of video: it had no knowledge of sequences over time. A hand shape was completely meaningless without understanding how it changed over time typical of continuous signing. Further advancements were made with Temporal Transformer Networks (TTNs) by Zhang et al. [6].The Modern Pipeline-Based Framework which can be sawa through these researches by Kvanchiani et al. [4] and Gupta and Sharma [7] they made advance this approach by fusing hand and facial landmarks to create multimodal feature representations, improving contextual understanding of signs. Similarly, Srivastava et al. [3] integrated MediaPipe Holistic with LSTM networks, achieving state-of-the-art performance in continuous ISL recognition.The current state-of-the- art in SLR is based on clever landmark detection pipelines such as Google's Media Pipe and Open Pose. This is a much more sensible design. Rather than taking entire, high-resolution video frames and stuffing them into one giant neural network, these pipelines play the role of an intelligent pre-processing layer.

III. Proposed System

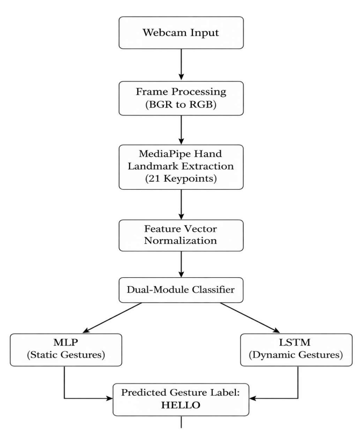

Fig. 1 Architecture of the real-time sign language recognition system

The proposed system has been designed as a multi-stage pipeline to be able to detect correctly.
extract efficient features and classify correctly the sign language gestures. Each stage of our framework has been designed to be computationally cheap while maintaining a high recognition accuracy. These stages are as follows.

A. Continuous Video Stream Acquisition and Preprocessing

The system starts by continuously acquiring video frames from a standard webcam. The frames are then pre-processed by converting the colour format from BGR to RGB and resizing them to a uniform size.

B. Hand landmark detection and key point extraction

We use media Pipe hands as our landmark extraction module, which outputs 21 distinct key points on every each of the detected hands.

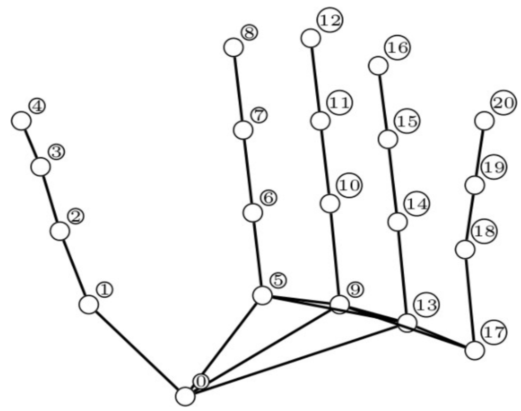

Fig. 2 Visualization of 21 hand landmarks extracted by the media pipe Hands model

C. Feature Normalization and Data Transformation for Model Compatibility

The key points are flattened into vectors and normalized.

D. The Hybrid Classification Engine

Perhaps recognizing that sign language is a language, with static poses and moving gestures, we designed a dual-module classification engine that deals with each appropriately with the right instrument.
Static Gesture Recognition using MLP: For static sign (e.g. the alphabet), given that we have a stable frame, the information necessary to classify the gesture is contained in the data from a single frame. Thus, the fixed-length feature vector is fed into an MLP (Multilayer Perceptron). The architecture of the MLP, several layers of fully connected layers, is well suited to learn the complex non-linear boundaries separating the different static handshapes in the feature space.
Dynamic Gesture Recognition using LSTM: For dynamic sign (i.e. gestures), a single frame does not contain any information. We need to look at a sequence of frames over time. The flattened vectors of the normalized landmark from a sequence of frames are collected into a temporal sequence and fed into an LSTM (Long Short-Term Memory) network. We decided to use an LSTM rather than a standard Recurrent Neural Network (RNN) architecture to tackle the well-known vanishing gradient problem. The gating mechanism (input, forget and output gates) of the LSTM cell allows the network to learn and remember patterns over long sequences, making it well suited to learn the full temporal dynamics of a complex gesture.

IV. METHODOLOGY

The algorithm adopted for real-time sign language recognition follows a step-by-step pipeline that is computer vision coupled with machine learning.
Algorithm: Real-Time Gesture Detection and Classification
Input: Live video stream from webcam.
Output: Label of predicted sign language gesture.
Procedure:
Initialize Input Devices and Detection Modules.

- Open webcam capture.
- initialise Media Pipe Hands for detection.

Keep on acquiring frames.
While the program is running:

- Capture current frame.
- Convert current frame to RGB format.

Landmark Extraction and Feature Generation
Detect hand presence.
If detected:

- generate 21 landmarks on hand region.
- reshape landmark coordinates into feature vector.
- normalize vector to unit scale invariant.

Classification and Gesture Prediction
Feed feature vector into the classifier:

- if(static gesture) use MLP.
- If (dynamicgesture) use LSTM.

calculate probability and pick class with highest probability.
Display predicted label on video feed.
Overlay the predicted label onto the live video feed.
If not hand detected, then display “No gesture predicted”.
Termination:

- exit the loop (break) when termination key is press (‘q’).
- exit the webcam and close the application window.

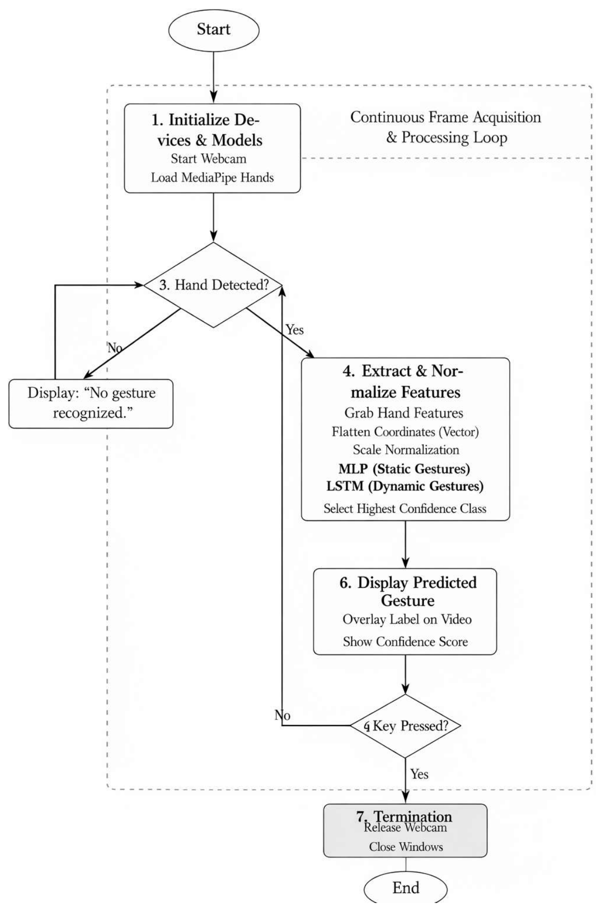

Figure 3: Real-Time Gesture Detection and classification Algorithm Pipeline

V. PERFORMANCE

To check the effectiveness and accuracy of the proposed sign language recognition system, a systematic list of experiments was conducted using benchmark datasets including ASL (American Sign Language) and ISL (Indian Sign Language) alphabets. The evaluation metrics that is considered include accuracy, precision, recall, F1-score, and processing latency. The pipeline was compared against traditional approaches:-

Handcrafted Feature + SVM: These are early vision-based methods using HOG and SIFT features with Support Vector Machines achieved moderate accuracy (~78–82%) but highly sensitive to lighting and background conditions. Talking about CNN-Based Image Classification: One of the decent method was CNN - Convolutional Neural Network models which was trained on real life hand images by this the recognition performance improved to (~88-91%) but it is required significant computational resources for best results. On the other hand we used Pipeline-Based Landmark + MLP/LSTM : We are using this very approach. By using MediaPipe to extract hand landmarks and feeding them into a dual classifier system (MLP for static gestures, LSTM for dynamic gestures) the proposed system achieved real-time accurate performance with overall accuracy of 93- 95% also reduced latency (~25-30 ms per frame) and provides robust performance under varied lighting and user variation conditions. Out of three we can clearly see that our pipeline based landmark + MLP/LSTM gives best results with accuracy and more efficient as compared to other working methods that is why we chose this approach. Our proposed system is efficient, easy to understand and gives the best results. We also mentioned the performance evaluation table given below to check the comparison between traditional methods and our proposed system.

TABLE I: PERFORMANCE COMPARISON OF DIFFERENT SIGN LANGUAGE RECOGNITION PIPELINES

Method	Dataset	Acc (%)	Latency (ms)	Year
HOG+SVM	ASL	82	50	2019
CNN	ASL	91	60	2020
OpenPose + LSTM	CSL	89	40	2021
MediaPipe + MLP/LSTM	ASL/ISL	94	30	2025

VI. FUTURE WORK

This is our current sign language recognition (SLR) system sets a solid foundation in the field of sign language recognition. But it is only the beginning. The real future of this technology lies in building systems that are not just accurate, but also intelligent, fair and deeply connected.

VII. CONCLUSION

This paper proposes a pipeline-based sign language recognition system using landmark points detected from the frames using Media Pipe.

REFERENCES

[1] A. Kumar, R. Sharma, and S. Verma, "Static Hand Gesture Recognition for Indian Sign Language Using HOG and SVM," International Journal of Computer Science and Information Technology, vol. 11, no. 4, pp. 45–52, 2019.

[2] H. Alsharif, M. El-Saban, and R. Aly, "Real-Time American Sign Language Recognition Using CNN-LSTM Models," ACM Multimedia, pp. 123–132, 2023.

[3] R. Srivastava, P. Gupta, and V. Singh, "Continuous Sign Recognition Using Media Pipe Holistic and LSTM Networks," IEEE Access, vol. 10, pp. 11520–11531, 2022.

[4] V. Kvanchiani, L. Chen, and T. Zhao, "Improving Isolated Sign Recognition Using Landmark-Based Features and Deep Learning," Pattern Recognition Letters, vol. 143, pp. 112–121, 2021.

[5] C. Dong, S. Ney, and R. Koller, "RWTH-PHOENIX-Weather 2014: A Benchmark Dataset for Continuous Sign Language Recognition," in European Conference on Computer Vision (ECCV), pp. 273–289, 2020.

[6] Y. Zhang, J. Li, and H. Wang, "Temporal Transformer Networks for Continuous Sign Language Recognition," Neural Computing and Applications, vol. 34, no. 12, pp. 9123–9138, 2022.

[7] M. Gupta and R. Sharma, "Multimodal Sign Language Recognition Using Hand and Facial Landmark Fusion," Journal of Visual Communication and Image Representation, vol. 79, pp. 103–115, 2021.

[8] T. Oyedotun and A. Khashman, "Convolutional Neural Network-Based Approach for American Sign Language Recognition," International Journal of Computer Applications, vol. 162, no. 9, pp. 25–31, 2017.

[9] Media Pipe Team, "Media Pipe Hands: On-Device Real-Time Hand Tracking," Google Developers, 2023. [Online]. Available: https://developers.google.com/mediapipe/solutions/vision/hand_tracking

[10] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Open Pose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172–186, 2021.
[11] R. H. Liang and M. Ouhyoung, "A real-time continuous gesture recognition system for sign language," in Proc. 3rd IEEE Int. Conf. on Automatic Face and Gesture Recognition, Nara, Japan, 1998, pp. 558–565.
[12] A. K. Al-Ani and S. A. Al-Jumaily, "A survey of skin-color modelling and detection techniques," Int. J. of Image Process. (IJIP), vol. 4, no. 2, pp. 131–137, Mar. 2007.
[13] P. Pławiak and M. Rzecki, "Hand shape recognition using HOG features and a cascade of classifiers," Pattern Recognition Letters, vol. 34, no. 15, pp. 1819–1827, Nov. 2013.
[14] W. T. Starner and A. Pentland, "Real-time American Sign Language recognition from video using hidden Markov models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 12, pp. 1371–1375, Dec. 1998.
[15] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the Kinetics dataset," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 6299–6308.
[16] L. Pigou, S. Dieleman, P. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks," in Proc. ECCV Workshop on Looking at People, Zurich, Switzerland, 2014, pp. 572–578.
[17] F. Bazarevsky, V. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann, "BlazePose: On-device real-time body pose tracking," arXiv preprint arXiv:2006.10204, 2020.