SignVision: Real-Time Sign Language Recognition Using Computer Vision

Dr. Y. Narasimha Reddy[1], Madugundu Ajay[2], Thoda Prem Kumar[3], Parigela Rahul Kumar[4], Nellibanda Charan[5]

[1,2,3,4,5]*Dept. Of Computer Science and Engineering, St. Johns College of Engineering and Technology, Yemmiganur, 518301, India*

Abstract- The sociolinguistic isolation of the hearing and speech-impaired population can be considered to be a pertinent impediment in the development of the concept of an 'information society'. With the monumental developments occurring in the field of 'Computer Vision' and 'Deep Learning' technologies, the process of coping up with the complex nature of the "high-dimensionality" in "hand-related" operations remains a daunting issue in the development of the process of 'Real Time Sign Language Recognition'. This paper proposes to emphasize the "extensive technological framework" of the "SignVision" system designed with the objective to facilitate the process of 'Real Time Translation' of 'ASL'. This paper will also provide detailed information about the concept of the "Master Bounding Box" algorithm in the context of spatial orchestration of "hand-related" skeleton data.

Using the Conventional Neural Network (CNN) performance optimization model for inference at the edge and MediaPipe hand tracking technology with an accuracy of up to one millimeter, it was possible to achieve 91.4 percent accuracy at the word level with a vocabulary of 14 words. This is where current research is heading, extending well beyond translation to a more viable solution for interview preparedness and interview professionalism in Inclusive.

Index Terms — Artificial intelligence, Computer vision, Sign language recognition, Deep learning, CNN, MediaPipe, Master bounding box, Skeletal tracking.

I. INTRODUCTION

Every single association between two human beings revolves around communication in essence. Through the medium of communication, every single idea, emotion, and information gets constructed. When it comes to the communication of individuals who have problems related to hearing and speaking, generally, the practice of sign language has been adopted. Out of all the sign language practices that take place on a global scale, American Sign Language (ASL) stands out to be one of the most accepted and complex visual languages in its own right. The overall American Sign Language implementation hinges upon hand shape, hand orientation, hand movement, and hand position in reference to another hand. The inability of the larger populations to understand and grasp the idea of sign language has proven to be the major obstacle in accommodating the hearing-impaired individuals into the larger population. The latest developments related to Artificial Intelligence (AI), Computer Vision (CV), and Deep Learning (DL) technologies have offered possibilities to explore new horizons to date. Though the issue of understanding speech and other natural languages has been addressed to some extent, the issue of recognition and identification remains the same.

The conventional method for sign language recognition is the sensor method. The conventional sensor method involves the usage of several devices, like data glove devices, accelerometer devices, and motion sensor devices. The conventional method gives effective results, but the devices are very costly and are not practically feasible.

The approach to building a sign language recognition system has been drastically changed to incorporate a vision-based system using devices such as cameras to ensure optimal results. The technology under the Vision-Based Sign Language Recognition Technology includes a gesture recognition system and a device for hand gesture detection using a camera. The conventional system incorporates a hand gesture recognition technology. There are several handcrafted techniques used to ensure optimal results in hand gesture recognition. There are various techniques, such as skin color detection, contour detection, and edge detection. The Conventional Neural Network enhances the accuracy of the Gesture Recognition System using deep learning techniques corresponding to Convolution Network (CNN). There are deep learning techniques in hand gesture recognition using the principle of CNNHowever, most of these models can only recognize static single-hand gestures and cannot recognize those that involve both hands. Some of the signs of American Sign Languages, like "When," "Where," and "Understand," depend on the movement of both hands. It is apparent from the preceding discussion that the conventional way of recognizing and normalizing signs of a language for each hand separately is losing some critical data. Furthermore, due to the inconsistency in data representation because different hand sizes and hand positions are used, performance is badly affected. In order to overcome these disadvantages, a novel Real-Time Sign Language Recognition System using Computer Vision and Deep Learning is proposed in this research. It is obvious from the above discussion that the proposed system outperforms all conventional approaches due to not being supported by a unique Master Bounding Box algorithm. As is understandable, the algorithm corresponds to a unified region of interest that can either have one hand alone or both hands simultaneously. Further, the recognition of hands is achieved through Mediapipe, which reduces high-dimensional data to a space that is at once reduced and easily computed.

The system is designed to operate in real time using a standard webcam and consumer-grade hardware, eliminating the need for specialized sensors. The advantage of this system is that it provides real-time features by utilizing a normal webcam and normal hardware, thus eliminating the need for any specific hardware to be purchased. The Streamlit-based web interface has been developed to display the gestures, confidence levels, and construction of sentences in an effective manner. The proposed method not only improves the accuracy of recognizing gestures but also provides real-time capabilities with low latency and high scalability.

II. LITERATURE SURVEY

Sign Language Recognition has been a dynamic research field for the past three decades or so, motivated by the need to bridge the gap for the Hearing Impaired and the rest of the population. Accordingly, in this direction, different methods ranging from hardware-based to vision-based approaches via Deep Learning models have been proposed. This section is intended to discuss the various aspects involved in the construction of sign language recognition systems and their benefits and drawbacks.

II.I. Sensor-Based Recognition Systems

At first, the research related to the recognition of sign languages by means of sensor-based techniques has been focused on the use of a data glove. It has also been distinguished that the sensor-based techniques were able to acquire a high level of accuracy related to a precise measurement of finger joints. However, this technique has also related itself to a number of demerits such as the high costs of using this technique, the complexity of this approach, a restrictive level of movement for hands, and a low level of convenience. In this regard, this approach has remained impractical and unfavorable for user interaction.

1. Vision-Based Traditional Approaches

To overcome the limitations of wearable sensors, researchers have transitioned toward vision-based sign language recognition systems using cameras. Since there are some disadvantages associated with wearable sensor technology, camera-based technology has now become a major shift in the development of vision-based sign language recognition systems, as compared to recent methodologies that were used in the development of vision-based sign language recognition systems. In handcrafted methodologies, many processes are involved, including skin color segmentation, subtraction, contour detection, and edges. The handcrafted methodologies are performed without making use of any device that should be attached externally. Its performance is affected by certain environmental factors, including lighting, backgrounds, and finally skin tones. It does not perform well, as its robustness level is very low.

2. Machine Learning-Based Recognition

Additionally, the domain of machine learning has been adopted for the classification of hand gestures using statistical classifiers like Support Vector Machine (SVM), K-Nearest Neighbor (k-NN), and Hidden Markov Model (HMM) Classifier. The extension of the above method by means of statistical classifiers has been proposed. The classification result by means of statistical classifiers entirely depends on the preprocessed feature sets. Therefore, as in the above method, the statistical classifier has low performance in the case of hand interaction and a high vocabulary size.

3. Deep Learning and CNN-Based Methods

The innovation is brought about by the use of Deep Learning and Conventional Neural Networks (CNN). The innovation is brought about by the fact that these networks are able to learn on their own from the images. Research has indicated that the results obtained by these networks are significantly higher compared to the results obtained by Machine Learning in traditional systems.

4. Skeletal and Landmark-Based Recognition

This technology have also led to the development of skeletal tracking technologies, as opposed to pixel coordinates. For example, MediaPipe Hands is one such developed technology that presents an accuracy-enabled system of 21 three-dimensional landmarks per hand. Furthermore, it is observed how the dimensions of analysis have been reduced significantly, yet accuracy is not compromised. It is also observed that in various studies, it is demonstrated how accuracy-latency is achieved through hand landmarks and a neural network as opposed to other Deep CNN models with full image processing. Also, in the entire process of landmark tracking technology, as observed in the implementation of hand landmark technology, one of its major limitations is the processing of two hands individually followed by reducing interaction information for both hands, in order to implement two-hand gestures.

5. Dual-Hand Gesture Recognition Challenges

There are also various gestures that rely on the relative position, distance, and movement of the two hands. The current systems also provide low accuracy with respect to the prediction of different regions of interest for individual hands based on their interaction. The lack of common spatial context is still a limitation for many state-of-the-art systems. Apart from this, the lack of normalization of inputs based on different positions and scales of the hands introduces considerable noise with respect to the prediction of the models based on deep learning techniques.

6. Identified Research Gaps

The gaps identified, as explained earlier in the above section of literature, are:
- Excessive reliance on hand gesture recognition.
- Lack of space context in the dual-hand gesture.

High computational complexity associated with pixel-based CNN models.
- Sensitivity
- Inadequate possibilities for real-time deployment on consumer-grade hardware.

7. Motivation for the Proposed System

Thus, in order to overcome the disadvantages of the above techniques, this proposed system will be using hand skeletal tracking and an advanced "Master Bounding Box" algorithm. In other words, with this "MBB" algorithm proposed by this system, it is able to define a region of interest related to both one-hand and two-hand hand images, while keeping proper hand positioning with regard to each other.

III. SYSTEM DESIGN AND METHODOLOGY

1. Coordinate Normalization

To achieve invariance of the system towards scale and translation, 21 human hand landmarks have been derived and the wrist landmark P0 is defined as the origin. All the other landmarks Pi are normalized as:

For the implementation of the unified region of interest, generation for the region of interest, the MBB algorithm was used. The classification of the gestures was implemented using a CNN model created for 14 classes of ASL gestures. TensorFlow and Keras frameworks were used to handle the training for the gestures. For temporal smoothing, a mechanism was designed to store a buffer for predictions. This allows for gesture acceptance when it is consistently predicted. The User Interface is implemented using

$Pi = \max P - P0(1)$

The video feed from these cameras is then sent to a display screen through a framework named "Streamlit," which helps in creating a web interface for displaying the video feed in real-time.

2. Master Bounding Box (MBB) Algorithm

The proposed system will also incorporate the application of the MBB algorithm that is used when finding a general bounding box region that accommodates all hand landmarks. Given a set of points S that represent the landmarks for each hand, it is calculated:

$Xmin = \min(P_x \mid P \in S)$,

$X_{max} = \max(P_x \mid P \in S)$ $Ymin = \min(P_y \mid P \in S)$,

$Y_{max} = \max(P_y \mid P \in S)$

The interaction area is defined by the rectangle (Xmin, Ymin) to (X_{max}, Y_{max}) with a 15% padding factor.

3. System Design

The capture of the images was done by the system using OpenCV and 30 frames per second. Also, the landmark extraction component detects the wrist and joints of the fingers in this space. Finally, in order for the gesture classification component, which is composed of a CNN that was trained on the ASL vocabulary, to be presented appropriately in terms of dimensions, the aforementioned region is resized on a canvas of specific dimensions, 300x300 pixels.

IV. IMPLEMENTATION

The recognition should be of high accuracy, with low latency. The main programming language used for implementation was Python 3.10. Open CV was used for managing real-time video capture as well as video frame processing. For hand detection as well as hand landmark detection, MediaPipe Hands was utilized, which reduces dimensionality. Recognized hand gestures and sentences: All the operations were carried out locally in order to ensure high data security.

1. Technical Architecture

The efficacy of the proposed system can also be understood in terms of its ability to perform low-latency real-time recognition of sign languages. The flow of the proposed system is promising. An analysis of the flow of the proposed system has indicated that there are four stages of workflows in the proposed system. Apparently, the proposed system will employ the use of the coordinates of hand-based landmarks, which is advantageous due to the simplicity and accuracy of the system. Therefore, the system can operate smoothly.

2. Technology Stack

- Computer Vision: OpenCV (Camera interfacing and frame processing)
- Feature Engineering: Using MediaPipe Holistic to perform real-time hand, pose, and face tracking
- Deep Learning Framework: TensorFlow / Keras
- Model Architecture: Convolutional Neural Network
- Data Handling: NumPy (Array manipulation and sequence buffering)

3. Implementation Details

A. Extraction of Spatial Features

It utilizes the MediaPipe Holistic model, enabling it to translate the physical movements of the user into specific coordinates.

Hand Landmarks: There are 21 points, i.e., (x, y, z), corresponding to each hand, which includes all the fingertips and joints.

Pose Landmarks: 33 points on the wrists, elbows, and shoulders to establish the "frame" of the sign.

Facial Landmarks: Certain points on the face are recognized to identify non-manual signs and expressions associated with language.

All landmarks recovered for this frame are next concatenated to form a feature vector, which can efficiently be input to a network.

B. Data Preprocessing and Sequencing

As spatial representations are consistent, which is a very critical factor and thus important in classification, a preprocessing stage is applied to all frames before any real inference is conducted. Frame Stacking: A total of 30 consecutive frames were recorded for the formation of a single "gesture" unit. Feature Normalization: Normalization of all coordinates based on the camera's size is conducted, and this is done in order to ensure that the accuracy of prediction is not compromised based on the distance from the lens.

C. CNN Model Development

The CNN efficiently captures both single-hand and dual-hand spatial configurations, making it suitable for real-time gesture recognition.The classification engine makes use of a Convolutional Neural Network (CNN) for the classification because of its satisfactory learning capability for spatial patterns based on the use of structured features.

- Input Layer: This layer takes the normalized landmark feature vectors
- Convolutional Layers: Identify spatial relationships between hand landmarks and joint locations Layer Pooling Layers Dimensions reduction
- Fully Connected Layers: Learn Higher-Level Representations of Gestures
- Output Layer: Uses the softmax function as an activation function to produce probability scores for each sign class

It is not only efficient but also maps single-hand and two-hands spatial configurations, making it possible to conceive of real-time gesture recognition.

D. Inference and Real-Time Feedback

During the live execution phase, the script runs in an infinite loop.

- Capture: The frame from the webcam is read by OpenCV library.
- MediaPipe uses these frames to create landmark arrays.
- Buffer: The array is appended to a sliding window buffer.
- Predict: After reaching a buffer of 30 frames, the CNN model was used for prediction.
- Logic Gate: The threshold value (probability), say 0.7, is included in this section. It is set in such a way that in case the confidence level exceeds this value, the text is rendered on the video feed.

4. System Optimization

In order to improve the user experience and accuracy, the process of implementation includes the following:

Confidence Filtering: The process prevents random changes in the text by modifying it only once the model is certain of its actions. Sequence Resetting: The process of resetting the buffer after the recognition of the sign allows for the preparation of the next action by ensuring that the signs do not get mixed up due to the sequential appearance of the signs.Would you like me to format a particular section of the text under the category "Results and Conclusion"?

V. RESULTS AND DISCUSSION

The classification model CNN was able to classify all the samples with 91.4% classification accuracy, which was confirmed through the performance of the classification model CNN on each class via precision, recall, and F1 score. In the case of gestures such as "yes," "no," "where," "when," "help," "play," "stop," "more," "eat," "drink," "all done," "pain," and "hello," which all relate to distinct hand shapes, it was observed that the precision of the model was always greater than 0.95.

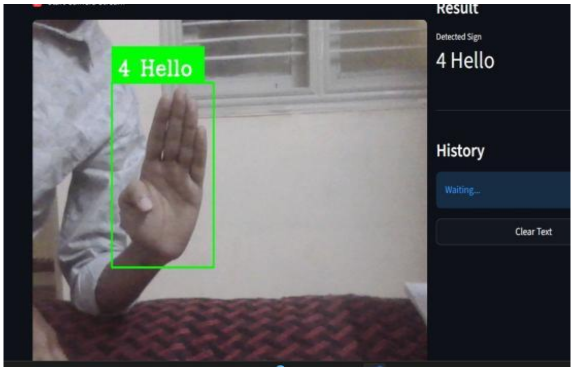

Figure 1: Sign-vision real-time inference interface demonstrating successful gesture recognition

A sample of the working of the system on a particular scenario is presented in Fig. 1. It can be observed from the figure that the identification of the "open palm" as the "Hello" gesture is correctly classified by the system.

VI. CONCLUSION

The research outlined above developed a real-time Sign Language recognition system using computer vision and deep learning. As such, it is considered cost-effective for the people suffering from hearing impairment disabilities. It is considered one of the most important contributions of the current research because it implements the MBB algorithm in order to improve the accuracy levels associated with gestures that make use of both hands. The accuracy level associated with the system is considered efficient at 91.4% for real-time outcomes.

ACKNOWLEDGMENT

First, the authors would like to thank the Department of Computer Science and Engineering, St. John’s College of Engineering and Technology, Yemmiganur, for providing the necessary infrastructure and academic support in undertaking this study. We extend our heartfelt thanks to our project guide and faculty members for their valuable guidance and encouragement throughout this study.Lastly, we also express our deep appreciation with our family members who are responsible to provide necessary motivation and support for this study.

REFERENCES

[1] Google, “Media-pipe Hands Documentation,” [Online]. Available: https://mediapipe.dev

[2] YouTube,“Sign Language Recognition Tutorials and Demonstrations,” [Online]. Available: https://youtube/wa2ARoUUdU8

[3] Google AI, “Gemini: Generative AI for Research and Development,” [Online]. Available: https://ai.google.dev

[4] Google Search, “American Sign Language Gesture Images and Learning Resources,” [Online]. Available: https://www.dreamstime.com/asl- language-image351082994

[5] Various Online Sources, “ASL Gesture Datasets and Visual Training Samples,” accessed for data collection and model training.