2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA)
Pimpri Chinchwad College of Engineering (PCCOE), Pune, India. Aug 23-24, 2024

2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA) | 979-8-3503-9177-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCUBEA61740.2024.10774760

Sign Language Translation with fusion of Emotion Detection

Arpita Acharya
Information Technology
Ramrao Adik Institute Of Technology
Navi Mumbai, Maharashtra
helloarpita12@gmail.com

Navin Patil
Information Technology
Ramrao Adik Institute Of Technology
Navi Mumbai, Maharashtra
navinpatilwork@gmail.com

Utkarsh Pathak
Information Technology
Ramrao Adik Institute Of Technology
Navi Mumbai, Maharashtra
utkarsh.pathak@gmail.com

Sumedha Bhagwat
Information Technology
Ramrao Adik Institute Of Technology
Navi Mumbai, Maharashtra
bhagwatsumedha2019@gmail.com

***Abstract*—Sign languages play a crucial role in fostering natural interaction, breaking down communication barriers between the hearing impaired and society. However, recognizing words in sign language poses significant challenges, particularly when gestures for multiple words are similar. Additionally, the rapid transition between gestures during communication complicates the creation of coherent sentences based on recognized words. To address these challenges, we propose the Real-Time Sign Language Recognition and Emotion Detection (RTSLRED) model, leveraging MediaPipe's holistic pipeline and LSTM network. Specifically focusing on American Sign Language (ASL), our model aims to enhance accessibility and inclusivity for the hearing-impaired community. Our results indicate an impressive accuracy of 91.05%, showcasing the model's effectiveness in conveying emotions and sentences in sign language. This innovation holds promise for empowering the hearing impaired and promoting inclusive communication practices.**

***Index Terms*—ASL (American Sign Language), Holistic pipeline, CNN (Convolutional Neural Network), LSTM (Long Short Term Memory), RTSLRED (Real-Time Sign Language Recognition and Emotion Detection), SLR (Sign Language Recognition)**

I. Introduction

Sign language is an essential communication tool for the deaf and hard-of-hearing communities, yet existing automated translation systems often fail to capture the emotional nuances that are critical to fully understanding the conveyed message. This research seeks to bridge this gap by integrating emotion detection into sign language translation, thereby enhancing the accuracy and contextual relevance of the translations.

The primary objective of this study is to develop a hybrid model that combines sign language translation with emotion detection. By utilizing advanced computer vision techniques to capture both hand gestures and facial expressions, this model aims to produce translations that not only convey the literal meaning but also the emotional context, making the communication more effective and authentic.

This research represents a significant advancement in accessibility technologies, offering a more holistic approach to sign language translation. The remainder of this paper is structured as follows: a literature review (Section II), the proposed model and methodology (Section III), experimental results (Section IV), and conclusions, including implications and future work (Section V).

II. Literature Survey

Eeva A. Elliott et. al [1] proposed that differences in facial expressions and their similarities can be defined using three features semantic, iconic, and compositional. The features were inferred from their first working assumption; that some facial expressions are semiotic units(form-meaning pairings).However the paper fails to explain how children acquire facial expressions, rather they make a strong claim regarding what it is that children acquire: semiotic units and the knowledge of how to combine them into more complex semiotic units.

N. Praveen et. al [2] proposed a smart alternative of using a portable smart glove for interpreting sign language. A micro-controller (MSP430G2553) is used which converts analog voltage values of the gestured letter in digital samples and the ASCII code is then wirelessly transmitted using the ZigBee. Partial completion of project and conversion of text to audio output is a major gap for the deaf community.

M. Elmahgiubi et. al [3] focused on translation of sign language into normal text that can be read by anyone by developing a Data Acquisition and Control (DAC) system. The proposed methodology captures the gesture of the hand using a smart glove and converts these hand motions into readable

text. The challenge faced by the author was of ambiguity and 20 out of 26 letters detected was identified as gap .

A. Kumar et. al [4] discussed on conversion of speech to signs along with sign language recognition. The algorithm proposed here was very much able of extracting signs from video sequences under clean and dynamic background and uses skin color segmentation. By distinguishing static and dynamic gestures, the methodology extracts the appropriate feature vector. Support Vector Machines(SVM) is used for classification. The Speech recognition was built upon standard module called Sphinx. But low system performance and favourable lighting conditions considered in proposed method including limited set of gestures proved as a challenge for target communities .

G. G. Nath et. al [5] recommended using ARM CORTEX A8 processor board using convex hull algorithm and template matching algorithm for SLR for deaf and dumb people. With inclusion to this the system might be able used to control devices like Robot, Car Audio Systems, home appliances. The proposed methodology uses Image Recognition Procedure and Recognition of Number Using Convex Hull Method. No major challenges were found in the paper apart from too many areas to cater.

G. Zhu et. al [6] introduced a method utilizing 3D-CNN and ConvLSTM for gesture recognition. However, challenges in recognizing gestures with fast movements, ambiguous static frames, and poor lighting persisted. Additionally, distinguishing similar gestures and separating hand movements from the background remained difficult, highlighting the need for improved methods in capturing and recognizing diverse gestures.

Lugaresi, C. et. al [7] presented MediaPipe as a robust framework for perceiving and processing reality. Despite its power, MediaPipe may face accuracy issues in complex environments and require substantial computational resources. These challenges could limit its applicability in certain scenarios, necessitating further optimization and refinement.

Y. Liao et. al [8] proposed a method for dynamic SLR using BLSTM-3D residual networks and MediaPipe. However, challenges in recognizing complex hand gestures, lower accuracy in dynamic SLR, and potential issues with training on larger video sequences were identified. Moreover, the focus on Chinese sign language may limit its applicability to other sign languages.

S. A. E. El-Din et. al [9] presented a five flex sensors, an alternative on ML, mounted on a glove. The paper proposed an interface with a control unit fixed on the arm, which converts both American Sign Language (ASL) and Arabic Sign Language (ArSL) in text and speech, and then projected on a GUI. The proposed methodology consists of three main components: input, processing and output units. Data is processed using Arduino Mega micro-controller board. Small size of glove, Sensor errors and human inability in performing signs, results in wrong recognition of words.

A. Sharma et. al [10] targets to build a Real-time Sign language to Speech translation. The paper proposed the using convolutional neural networks (CNN) along-with the Text-to-Speech translator. graph clustering is phenomenon part included in first phase. The proposed method required the compulsory camera access for live video inputs for live translations. The algorithm proposed does not consider context-recognition which is identified as challenge according to us.

Y. Grover and R. Aggarwal et. al [11] both focused on discussing already existing methodologies on Sign-Language Translation. Also the classification of translators and recognisers are done based on taxonomy. According the authors a complete sign language comprises of Finger-Spelling, Word-Level Vocabulary and Non-manual features. Two-way sign language system for indian sign language(ISL) was not observed during survey.

Aditi Chavan et. al [12] elaborates the sign language methodologies into indian sign language interpreter. A vision-based approach is employed that uses Machine Learning technique of 3D-CNN (3-Dimensional Convolutional Neural Network) alongside LSTM (Long Short-Term Memory) neural network to effectively map indian sign gestures with the designated words. The vision-based approach was proved to be better option compared to glove-based as there was no malfunciton and higher accuracy.

Kezar et. al [13]. indicated that there are two methods for studying facial expression and emotional relationships through probability and prediction.The methods discussed were: large corpus of natural signing which were manually annotated with facial features and lexical emotion datasets.There were no gaps in this paper as it only discussed few ways.

A. Akandeh et. al [14] proposed a Sentence-Level Sign Language Recognition Framework. A bi-solution was presented to sentence-level SLR in this paper. The paper also discussed continuous sign language recognition (CSLR) in which temporal boundaries of the words in the sentences are not defined. The proposed framework/models was evaluated on the basis of RWTH-PHOENIX Weather dataset . Non-manual elements including the eye gaze and mouth shape alongwith facial expression w.r.t to Sentence-Level SLR were not discussed in paper.

D. Gandhi et. al [15] provided a break-through in Sign Language Recognition Systems. A dynamic sign language recognition and emotion detection (DSLRED) model that uses MediaPipe's holistic pipeline along with an LSTM network which also classifies the emotion on a person's face into one of seven categories, using deep convolutional neural networks

was proposed by author. The proposed model was observed to acheived the highest accuracy of 98.95 percent among all papers.But the proposed system was not trained on existing datasets, it was built to classify only seven sets of hand gestures.

III. PROBLEM STATEMENT

The goal of this paper is to develop a Real-time Sign Language Recognition system based on persons' emotions that takes into account facial expressions, to interpret and translate sign languages. Our system tackles various challenges such as implementing LSTM Neural Network that is effectively capable of learning and remembering words over time while the MediaPipe Holistic model ensuring that the recognized gestures are trustworthy and reflective of genuine user conversations and ensuring that the system can handle low lighting backgrounds around users while maintaining responsiveness and accuracy along with approximate emotions/contexts.

IV. PROPOSED METHODOLOGY

We have implemented Media-Pipe holistic model in our Real-time Sign Language Recognition System. It plays a crucial role in a SLR System for several reasons. Firstly, it enables us to build-in fast ML inference and video processing, which includes GPU, CPU, or TPU. Secondly, The all-in-one framework of MediaPipe is suitable for Android, iOS, desktop, cloud, web, and IoT platforms. Thirdly, inference AI models and other reusable components make MediaPipe our choice for this Project. The MediePipe Holistic model is combined with LSTM Neural Networks.

The MediaPipe's holistic pipeline and the LSTM (Long Short-Term Memory) network is appropriate for this application. Reasons behind using this combination is because MediaPipe's holistic pipeline provides a robust framework for capturing and processing multi-modal data, including hand gestures, body poses, and facial expressions. This makes it particularly suitable for sign language recognition, where such data is crucial. LSTM networks are effective in handling sequential data and can model the temporal dependencies in sign language sequences, making them well-suited for this task. However, the paper could strengthen its justification by comparing this approach to other potential methods or by providing more empirical evidence of its effectiveness in this context.

1) Preprocessing - The initial step encompasses preprocessing the input image to ready it for sign detection and recognition. This entails tasks like frame reading, keypoint detection, and landmark/keypoint extraction. Keypoints play a pivotal role as they serve as essential attributes for identifying and aligning objects in images, encompassing corners, edges, or other distinctive traits of traffic signs.

2) Sign Detection and Recognition - After the preprocessing stage, the system advances to sign

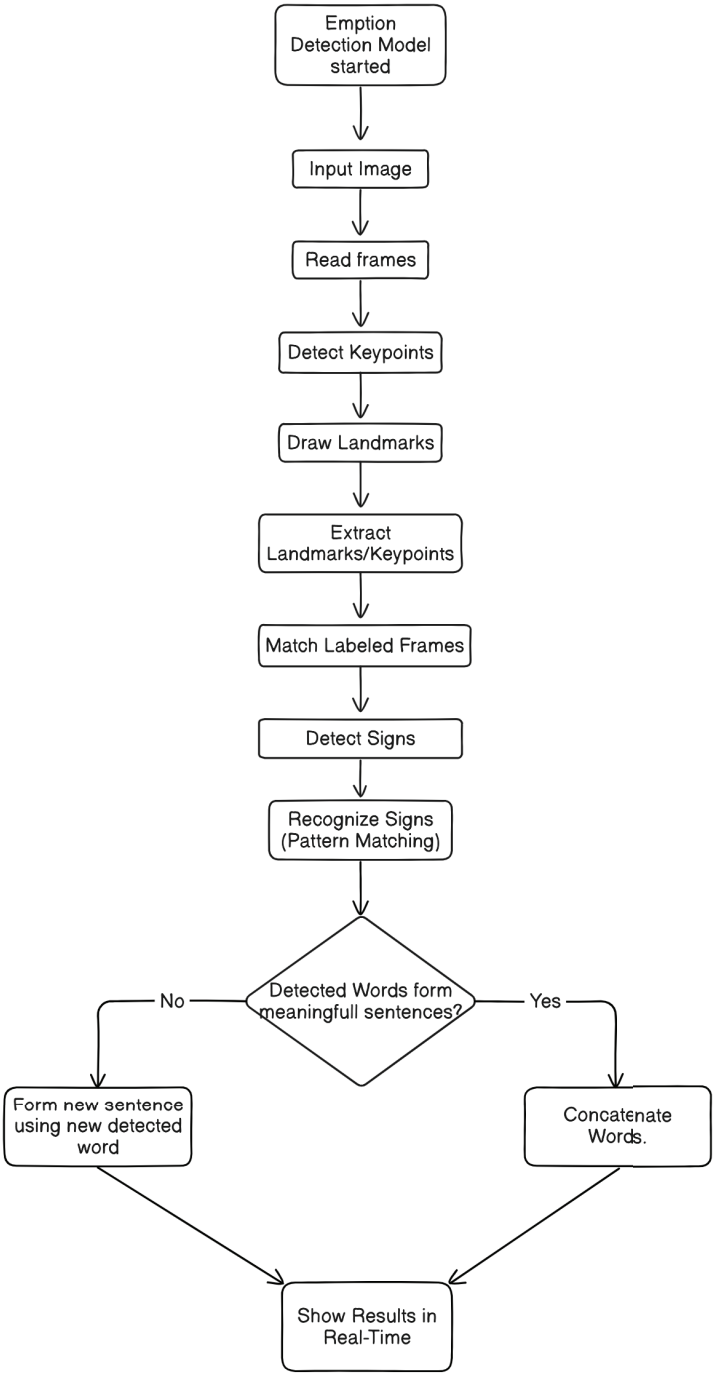

Fig. 1: Working of Sign-Language Model

detection within the image and the subsequent identification of the detected sign type. This stage entails matching the extracted landmarks/keypoints with a library of labeled frames. Upon finding a suitable match, the system employs pattern-matching techniques, as shown in Figure 1 for sign recognition. Pattern matching facilitates a comparison between the extracted features of the unidentified sign and those of known signs in the library. The sign with the closest resemblance to the unidentified one is then recognized as the most probable sign.

3) Output - In the final stage, the system promptly presents the detected sign alongside its corresponding label in real-time. This instant feedback equips drivers with timely information regarding the traffic signs in their proximity, fostering heightened situational awareness and facilitating well-informed decision-making.

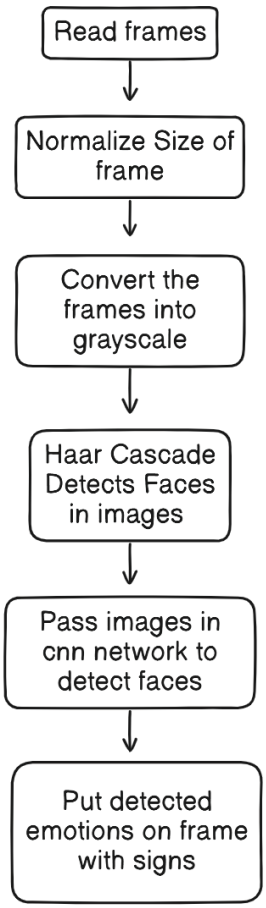

Fig. 2: Working Of Emotion Detection Model

1) **Read frames:** Video frames are sequentially captured and processed by the system to analyze motion.
2) **Normalize frame size:** Frames are resized to a standardized dimension for efficient processing.
3) **Convert to grayscale:** Frames are converted to grayscale to streamline facial detection, reducing computational load.
4) **Facial detection with Haar cascade:** A Haar cascade classifier identifies facial features in grayscale images, leveraging machine learning algorithms.
5) **CNN for refined detection:** Following initial detection, Convolutional Neural Networks (CNNs) may further enhance facial recognition accuracy.
6) **Emotion detection:** Detected faces are analyzed to identify emotions, with corresponding labels displayed alongside the faces.

Figure 2 shows our Emotion Detection model in a cascading waterfall style, illustrating the step-by-step transformation of input data into emotion predictions. The model uses Convolutional Neural Networks (CNN) and Keras to accurately analyze subtle facial expressions. This structured approach ensures efficient and effective emotion analysis, enhancing understanding of expressions in sign language gestures.

V. RESULTS AND ANALYSIS

Confusion Matrix 1

$$\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$$

Confusion Matrix 2

$$\begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}$$

Confusion Matrix is in form of:

1) TN = True Negative
2) FP = False Positive
3) FN = False Negative
4) TP = True Positive

A novel approach combines MediaPipe and LSTM models to enhance communication with the hearing-impaired, focusing on seven sign language word classes. The dataset includes fifty videos, each with thirty frames per word class. Leveraging LSTM's ability to understand sequential video context, the model effectively captures sign language nuances. Figure 4 illustrates the similarity between signs and their outputs.

Key actions with similar gestures, such as "please" and "sorry," were carefully selected, highlighting the subtle differences like palm configuration. The research includes a comparative analysis, demonstrating the effectiveness of our approach.

The system's versatility allows users to construct diverse sentences using identified words while conveying emotions. Notably, the system recognizes actions even if the training data, such as "how," "you," and "I am good," was created by different signers, enhancing its applicability.

Emotion detection, integrated with gesture recognition, enriches communication by adding emotional context. Using CNNs through Keras, our model accurately identifies emotions like happiness and sadness, as shown in Figure 3. Real-time processing ensures fluid and accurate emotion detection, improving interaction.

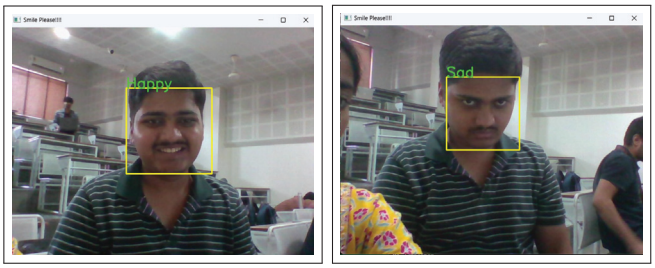

Fig. 3: Emotion Detection using CNN and Keras

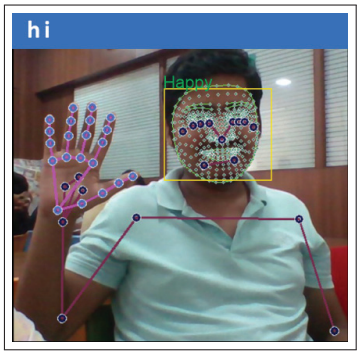

Fig. 4: The Combine Output

Figure 4 depicts the amalgamation of our sign language recognition and emotion detection system. This integrated visual represents the harmonious blending of technologies for emotion context identification and sign language interpretation. By combining robust algorithms for sign language recognition (MediaPipe+LSTM) and emotion identification (CNN), our system comprehensively understands communication subtleties. Users can effectively convey both linguistic content and emotional nuances, bridging the communication gap between individuals with speech or hearing impairments and the wider community.

To further evaluate our model's performance, we calculated precision, recall, and F1 score metrics. As shown in Figure 5, our model achieved a precision of 0.803, recall of 0.818, and an F1 score of 0.782. These metrics demonstrate the model's

```
# Precision, Recall, and F1 Score
from sklearn.metrics import precision_recall_fscore_support

# Compute Precision, Recall, F1 Score
precision, recall, f1, _ =
precision_recall_fscore_support(Ytrue, Yhat, average='weighted')

print(f"Precision: {precision}")
print(f"Recall: {recall}")
print(f"F1 Score: {f1}")
```

```
Precision: 0.8030303030303029
Recall: 0.8181818181818182
F1 Score: 0.7818181818181817
```

Fig. 5: The Precision, Recall, and F1 Score

```
[69]: accuracy_score(Ytrue, Yhat)

[69]: 0.9090909090909091
```

Fig. 6: Model Accuracy

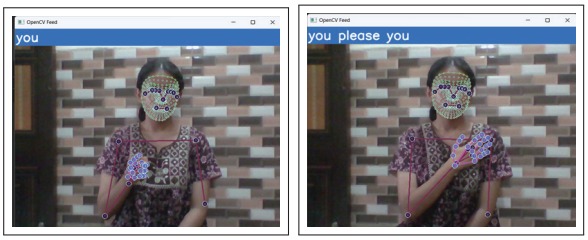

Fig. 7: Similarity between signs and their confidence

ability to balance between correctly identifying sign language gestures (precision) and capturing a high proportion of actual gestures (recall). The F1 score, being the harmonic mean of precision and recall, provides a single metric that captures the model's overall performance, indicating strong and balanced predictive capabilities.

With an impressive accuracy score of around 0.99, our project demonstrates the exceptional performance of our machine learning model. Our journey involved meticulous data preprocessing, careful model selection, and rigorous testing. As we conclude this phase, we eagerly anticipate leveraging this powerful tool for real-world impact.

VI. Misinterpretation

The model sometimes struggles to distinguish between similar signs in Sign Language, as illustrated in Figure 8. This challenge stems from the inherent ambiguity of Sign Language, where a single sign can convey multiple meanings based on subtle nuances and context. Accurately interpreting these variations is difficult for automated systems. However, continued refinement and the integration of advanced deep learning models can improve the system's ability to handle these complexities.

VII. Future Scope

To enhance accuracy and efficiency, integrating Attention Mechanisms with the LSTM network could help the model focus on the most relevant parts of the input sequence, improving both recognition and emotion detection. Adopting a Transformer-based architecture might further enhance performance by better capturing complex relationships between signs and emotions. Transfer Learning with pre-trained models on large datasets could also refine accuracy.

However, several challenges remain. Real-time processing with LSTM networks can be computationally intensive, potentially causing latency issues. The availability of large, annotated sign language datasets, particularly for emotion detection, is limited. Ensuring the model generalizes well across different sign languages and dialects will require extensive training data. Additionally, integrating this model with

existing communication systems might present compatibility and processing challenges. To ensure broader applicability, optimizing the model for lower-end devices and addressing potential hardware requirements would be beneficial.

VIII. CONCLUSION

In conclusion, our research has demonstrated the efficacy of the Real-Time Sign Language Recognition and Emotion Detection (RTSLRED) model in addressing the communication challenges faced by the hearing impaired community. By leveraging MediaPipe's holistic pipeline and LSTM network, we have achieved significant advancements in recognizing sign language gestures and detecting emotions with high accuracy. Our model not only enhances accessibility for individuals using American Sign Language (ASL) but also promotes inclusivity by facilitating seamless communication of emotions and sentences in sign language. Moving forward, further refinement and validation of the RTSLRED model in diverse real-world scenarios will be essential to its widespread adoption and impact. Ultimately, this research contributes to the broader goal of empowering the hearing impaired and fostering a more inclusive society through innovative technology solutions.

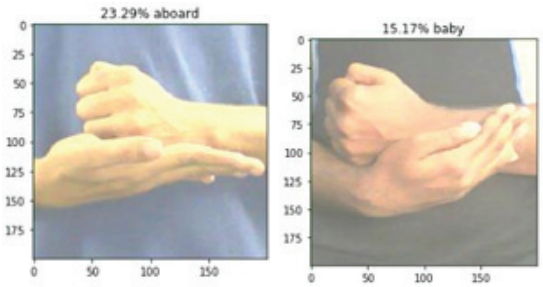

Fig. 8: Similarity between signs and their confidence

REFERENCES

[1] E. Elliott and A. Jacobs, "Facial expressions, emotions, and sign languages," *Frontiers in psychology*, vol. 4, p. 115, Mar. 2013. DOI: 10.3389/fpsyg.2013.00115.

[2] N. Praveen, N. Karanth, and M. S. Megha, "Sign language interpreter using a smart glove," in *2014 International Conference on Advances in Electronics Computers and Communications*, 2014, pp. 1–5. DOI: 10.1109/ICAECC.2014.7002401.

[3] M. Elmahgiubi, M. Ennajar, N. Drawil, and M. S. Elbuni, "Sign language translator and gesture recognition," in *2015 Global Summit on Computer Information Technology (GSCIT)*, 2015, pp. 1–6. DOI: 10.1109/GSCIT.2015.7353332.

[4] A. Kumar, K. Thankachan, and M. M. Dominic, "Sign language recognition," in *2016 3rd International Conference on Recent Advances in Information Technology (RAIT)*, 2016, pp. 422–428. DOI: 10.1109/RAIT.2016.7507939.

[5] G. G. Nath and C. S. Arun, "Real time sign language interpreter," in *2017 IEEE International Conference on Electrical, Instrumentation and Communication Engineering (ICEICE)*, 2017, pp. 1–5. DOI: 10.1109/ICEICE.2017.8191869.

[6] G. Zhu, L. Zhang, P. Shen, and J. Song, "Multimodal gesture recognition using 3-d convolution and convolutional lstm," *IEEE Access*, vol. 5, pp. 4517–4524, 2017. DOI: 10.1109/ACCESS.2017.2684186.

[7] C. Lugaresi, J. Tang, H. Nash, *et al.*, "Mediapipe: A framework for perceiving and processing reality," in *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019. [Online]. Available: https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf.

[8] Y. Liao, P. Xiong, W. Min, W. Min, and J. Lu, "Dynamic sign language recognition based on video sequence with blstm-3d residual networks," *IEEE Access*, vol. 7, pp. 38 044–38 054, 2019. DOI: 10.1109/ACCESS.2019.2904749.

[9] S. A. E. El-Din and M. A. A. El-Ghany, "Sign language interpreter system: An alternative system for machine learning," in *2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, 2020, pp. 332–337. DOI: 10.1109/NILES50944.2020.9257958.

[10] A. Sharma, S. Panda, and S. Verma, "Sign language to speech translation," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, pp. 1–8. DOI: 10.1109/ICCCNT49239.2020.9225422.

[11] Y. Grover, R. Aggarwal, D. Sharma, and P. K. Gupta, "Sign language translation systems for hearing/speech impaired people: A review," in *2021 International Conference on Innovative Practices in Technology and Management (ICIPTM)*, 2021, pp. 10–14. DOI: 10.1109/ICIPTM52218.2021.9388330.

[12] A. Chavan, "Indian sign language interpreter for deaf and mute people," in *International Journal of Creative Research Thoughts (IJCRT)*, vol. 9, 2021, pp. 1–5.

[13] L. Kezar and P. Zhou, "The role of facial expressions and emotion in asl," Jan. 2022.

[14] A. Akandeh, "Sentence-level sign language recognition framework," in *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2022, pp. 1436–1441. DOI: 10.1109/CSCI58124.2022.00256.

[15] D. Gandhi, K. Shah, and M. Chandane, "Dynamic sign language recognition and emotion detection using mediapipe and deep learning," in *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2022, pp. 1–7. DOI: 10.1109/ICCCNT54827.2022.9984592.