Fine-Tuning TimeSformer for Sign Language Recognition

G. Charan Tej Paul[1], Vivek Kumar[1], Kunal Kashyap[1], and Senthil Murugan Balakrishnan[2]

[1] [1] Master of Computer Applications, Vellore Institute of Technology, Vellore, Tamil Nadu, India
[1] g.charantejpaul2024@vitstudent.ac.in, vivek.kumar2024@vitstudent.ac.in, kunal.kashyap2024@vitstudent.ac.in
[2] [2] Professor, SCORE, Vellore Institute of Technology, Vellore, Tamil Nadu, India
senthilmurugan.b@vit.ac.in

Abstract. Sign Language Recognition (SLR) has always been very important in bridging communication gaps between hard-of-hearing people and hearing people. Many have tried developing a robust sign language recognition system; However, existing SLR models often struggle with capturing long-range temporal dependencies, learning fine-grained spatial-temporal features, and handling large vocabulary variations. Traditional approaches, such as CNN-RNN hybrids or 3D CNNs, are limited by their inability to model global motion patterns effectively. The core idea of our proposed methodology is a system built upon the transformer architecture, leveraging TimeSformer, a pure attention-based architecture, to address these limitations. Unlike conventional models that rely on handcrafted features or optical flow, the TimeSformer model performs self-attention across both the spatial and temporal aspects of video, allowing an end-to-end learning process directly from raw RGB frames. In our study, to overcome the limitation of the lack of data, the TimeSformer model, which has been trained on the "Kinetics-400" dataset, is fine-tuned on a new dataset formed by merging two separate datasets, called MS-ASL [20] and WLASL [21], to recognise a specific set of 12 classes. Our approach eliminates the need for CNN-based feature extraction, improving efficiency while maintaining competitive accuracy.

Keywords: Sign Language Recognition · TimeSformer · fine-tune · Deep Learning · Transformer Models. · MS-ASL · WLASL

1 Introduction

Sign language serves as a primary medium of communication for the hard-of-hearing communities, both internally and externally. According to the WHO, there are around 70 million people with hearing disabilities; in retrospect, there are significantly fewer translators capable of translating sign language. Therefore, it can be irresponsible of us to solely depend on the capability of an interpreter. We need a bridge between hard-of-hearing people and hearing people for

the establishment of efficient communication. With the development of machine learning, we have been able to solve various problems, one of them being the development of SLR systems, but many of the models are incapable of recognising sign language with near-native accuracy. This is mainly due to the dynamic and variably rich nature of the hand gestures and the availability of different kinds of sign languages. The primary focus of our study is the development of an SLR system for the American Sign Language (ASL). Traditional SLR systems are mainly based on CNNs and RNNs or 3D CNN-based feature extractors to model spatial and temporal dependencies. However, these approaches suffer from limited receptive fields, difficulty in capturing long-term dependencies, and reliance on handcrafted features or optical flow. Recent advancements in Transformer-based architectures have demonstrated remarkable success in sequence modelling tasks, particularly in natural language processing and action recognition. Inspired by this, we explore the use of TimeSformer, a fully attention-based Transformer model designed for video understanding, for the task of SLR.

The TimeSformer architecture applies a self-attention mechanism across both spatial and temporal dimensions, enabling the model to learn much finer details about the sign being performed than a normal model. Due to the similarities between the "Kinetics-400" data and the sign language data, both of which are about the movement of human body parts, we will be fine-tuning a TimeSformer model trained on the "Kinetics-400" data, as we hope to train the model with fewer computational resources. This paper explores the application of TimeSformer for SLR. The contributions of this work are as follows:

- We have performed an analysis of whether a pure attention-based mechanism can be used for SLR.
- We present a comparative study of SLR approaches and highlight the advantages of transformer-based temporal modelling.
- We propose a framework leveraging TimeSformer to enhance recognition accuracy and real-time applicability in sign language translation systems.

2 Literature Review

When translating sign language into natural spoken language, we first try to convert the gestures into the English alphabet. Since the alphabet can be represented with a single image, we use a simple CNN model to classify the alphabet and then form complete words. Singh et al. [1], in their work, created a custom dataset of 10,500 images of static hand gestures. Their model, based on a modified VGG16 CNN, accurately handles static hand signs that primarily represent the alphabet. Damaneh et al. [2] propose a system that combines CNNs with feature extraction to improve accuracy and reduce ambiguities, but extracting features is a tedious process. As a result, building real-time systems capable of recognising hand signs naturally is challenging, given the need for a feature extraction pipeline before making predictions. In a comparative study, Kolkur et al. [4] evaluated CNN, VGG16, and AlexNet architectures for their ability to translate sign language gestures into text, ultimately finding that a standard CNN,

enhanced by TensorFlow's object detection, performed the best. Kankariya et al. [5] applied the inception V3 algorithm to a dataset of 26 ASL alphabet images on top CNNs, and found that their model performed best for an 80:20 split.

Following the initial attempts at building an SLR system, further studies have been conducted, mainly dealing with dynamic signs. A glaring necessity of learning from dynamic signs is the need to remember long-range dependencies to predict the sign being performed. Daga et al. [6] proposed an LSTM (Long Short-Term Memory) network for efficiently remembering long-range dependencies. Building on this, Khartheesvar et al. [7] propose the use of MediaPipe holistic for extracting spatial features on top of LSTM to recognise hand-signs more effectively. Both the proposed models work on Indian Sign Language (ISL) [6,7]. Another study, Kushwaha et al. [3], proposes a model that integrates facial, head, arm, and hand motions, with an emphasis on hand gestures. This methodology uses Google's MediaPipe for continuous hand landmark detection and AlexNet for gesture classification in Indian Sign Language (ISL). An LSTM model only processes data in a single direction, often missing the context of the previous frames. To overcome this, we can use BiLSTM, a Bidirectional LSTM. Das et al. [8] propose a BiLSTM model along with the multi-featured deep network (MF-DNet). Shao et al. [11] presented the B3D network, which uses 3D CNNs and BiLSTM to capture spatiotemporal features, tested on the DEVISIGN-D and lab self-built SGR dataset, compared with other traditional methods such as Alex, C3D+LSTM. This model is noted to be more accurate.

Das et al. [9] introduced a system using ASL datasets with YOLO-v5, enhanced by a BOT control for versatile gesture applications. To improve accuracy further, Long et al. [10] combined hand skeleton feature extraction with facial expressions. Incorporating facial expression information further enhances the system's effectiveness. The method was tested on the Argentinian Sign Language dataset and SEU's Chinese Sign Language Recognition Database (SEUCSLRD), demonstrating improved performance. While many models focus on image data, building a Sign Language Recognition (SLR) system necessitates handling video data. For example, Yang et al. [12] proposed a two-stream 3D-ConvNet for recognising coal mine emergency gestures by decomposing gestures and using optical flow for feature extraction.

Recent advancements in transformer-based models have shown great promise in SLR systems; they are better at modelling sequences and can capture long-range dependencies more efficiently by leveraging the self-attention mechanism. Saproo and Aggarwal [13] employed deep learning models, including transformers, on the INCLUDE dataset, using pre-trained CNNs to extract embeddings from video frames before classification with a Transformer network. Agarwal et al. [14] addressed ISL interpretation with a Vision Transformer model using a self-created dataset of 72 words, while Francis [15] developed an advanced sign language translation system integrating a Vision Transformer with the PHI-1.5B language model for ISL. Singla et al. [16] present a unique approach to recognising ISL by using Keras, Visual Transformers (ViT), and advanced data augmentation techniques together.

We cannot always hope for our dataset to have pristine data; we can have data that is of low resolution, fast in motion, and the visible region might be small in scale. Wang et al. [18] propose a 3D-ResNet algorithm for sign language recognition, addressing these challenges. The proposed method enhances hand feature detection using an EfficientDet network, incorporating the Bi-FPN module and spatial attention to better detect small hand targets. Furthermore, the 3D-ResNet18 network is enhanced with a residual module to extract sign language features more effectively. It separates global, left-hand, and right-hand image sequences into distinct branches, allowing for improved feature fusion. Camgoz et al. [17] introduces a model for creating a system for SLR, which applies a joint space-time attention. A limitation of this study is the need to convert the raw video into an embedding through CNNs, which are later fed into the transformer network.

Our proposed model, TimeSformer, is a pure-attention-based mechanism that is free from convolution, allowing the system to truly learn from end-to-end, i.e., to learn features directly from the RGB frames, unlike the traditional methods that require either optical flows or hand-crafted features. This allows TimeSformer to learn spatial and temporal dynamics without long-winded pipelines. TimeSformer applies a divided space-time attention which, when compared to a model that applies joint space-time attention, is less computationally intensive. In this paper, we will be fine-tuning a TimeSformer model pre-trained on the Kinetics-400 dataset on a dataset of 12 classes, each with more than 50 videos to learn from, created from combining WLASL[20] and MS-ASL[21].

3 Proposed Methodology

This paper proposes the use of the TimeSformer architecture for the recognition of American Sign Language. The key innovation in our approach is the use of a pure attention-based mechanism instead of convolutional operations. This design allows the model to effectively capture complex spatial and temporal features of the sign being performed, and also the relationships between these features. Since we currently lack the necessary data and computational power to effectively train the TimeSformer model from ground up, we will be using a pre-trained TimeSformer model that has been trained on the "Kinetics-400" dataset. The overall workflow and the components of the proposed model are illustrated in Fig. 1.

3.1 Video Input

The input videos have been processed into `.npy` files, where each video is represented using 16 frames of resolution 224×224. Contrast Limited Adaptive Histogram Equalisation (CLAHE) has been used to improve the visibility of details, so that the model can learn from the data easily. By using the `.npy` we have reduced the time it takes the model to load videos through a pipeline, decreasing the training time. The Fig. 2 represents the input frames that have been fed into the model for training and validation purposes.

.

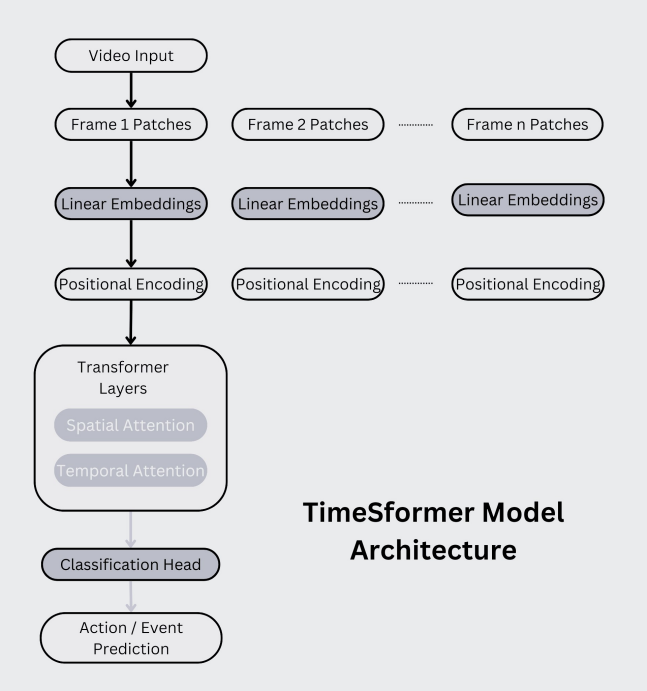

Fig. 1. *Architecture of TimeSformer model*

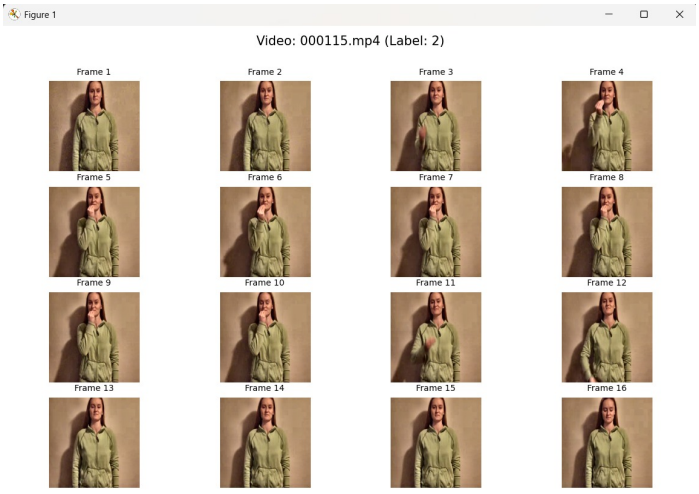

Fig. 2. *Video Input Frames*

3.2 Patch Embedding

As seen in Bertasius et al [19], each video frame is divided into fixed-size, non-overlapping patches. For a frame of size $H \times W$ split into patches of size $P \times P$,

the result is $\frac{H \times W}{P^2}$ patches per frame. This process converts a 2D frame into a sequence of tokens, analogous to words in natural language. Each patch captures local visual information (e.g., hand shapes, facial features), which the model processes sequentially.

3.3 Linear Embedding

The patches are flattened into one-dimensional vectors. If a patch is of size $P \times P \times C$ (with C being the number of channels, e.g., 3 for RGB), flattening produces a vector of length $P^2 \times C$. These vectors are then projected into a fixed-dimensional embedding space, enabling the transformer to leverage self-attention mechanisms for both local and global pattern recognition.

$$\mathbf{z}_{(p,t)}^{(0)} = E\mathbf{x}_{(p,t)} + \mathbf{e}_{(p,t)}^{pos} \tag{[19]}$$

3.4 Positional Encoding

The self-attention mechanism does not inherently process the order of the tokens; therefore, we need to add the positional data to the patch embeddings. The positional encoding $\mathbf{z}_{(p,t)}^{(0)}$ [19] is calculated as a weighted sum of the vectors, using the self-attention coefficients from the attention heads which act as the weights.

$$\mathbf{s}_{(p,t)}^{(\ell,a)} = \alpha_{(p,t),(0,0)}^{(\ell,a)} \mathbf{v}_{(0,0)}^{(\ell,a)} + \sum_{p'=1}^{N} \sum_{t'=1}^{F} \alpha_{(p,t),(p',t')}^{(\ell,a)} \mathbf{v}_{(p',t')}^{(\ell,a)} \tag{[19]}$$

Then, from all heads these vectors are passed through an MLP, using residual connection after each operation:

$$\mathbf{z'}_{(p,t)}^{(\ell)} = W_O \begin{bmatrix} \mathbf{s}_{(p,t)}^{(\ell,1)} \\ \vdots \\ \mathbf{s}_{(p,t)}^{(\ell,A)} \end{bmatrix} + \mathbf{z}_{(p,t)}^{(\ell-1)} \tag{[19]}$$

$$\mathbf{z}_{(p,t)}^{(\ell)} = \mathrm{MLP}\left(\mathrm{LN}\left(\mathbf{z'}_{(p,t)}^{(\ell)}\right)\right) + \mathbf{z'}_{(p,t)}^{(\ell)} \tag{[19]}$$

3.5 Spatial and Temporal Attention

Spatial Attention. Spatial attention captures relationships between patches within a single frame, enabling the model to understand details such as hand shapes, facial expressions, and textures.

$$\boldsymbol{\alpha}_{(p,t)}^{(\ell,a)\,\mathrm{spatial}} = \mathrm{SM}\left(\left(\frac{\mathbf{q}_{(p,t)}^{(\ell,a)}}{\sqrt{D_h}}\right)^{\top} \cdot \left[\mathbf{k}_{(0,0)}^{(\ell,a)} \left\{\mathbf{k}_{(p',t)}^{(\ell,a)}\right\}_{p'=1,\ldots,N}\right]\right) \tag{[19]}$$

Temporal Attention. Temporal attention captures the evolution of spatial representations across frames, modeling motion and transitions essential for recognising sequences in sign language videos.

$$\boldsymbol{\alpha}_{(p,t)}^{(\ell,a)\,\text{time}} = \text{SM}\left(\left(\frac{\mathbf{q}_{(p,t)}^{(\ell,a)}}{\sqrt{D_h}}\right)^{\top} \cdot \left[\mathbf{k}_{(0,0)}^{(\ell,a)} \left\{\mathbf{k}_{(p,t')}^{(\ell,a)}\right\}_{t'=1,\ldots,F}\right]\right) \qquad ([19])$$

Together, spatial and temporal attention allow TimeSformer to interpret both static visual content and dynamic video sequences effectively.

3.6 Classification Head

The aggregated final video representation is fed into a classification head. This head is typically a linear, i.e., a fully connected layer that maps the feature vector to the final number of output classes. This operation is described by the equation (1) below:

$$\text{logits} = Wz + b \tag{1}$$

Here, z is the aggregated feature vector.W is the weight matrix and b is the bias vector.

3.7 Computational Complexity

Let N be the number of patches per frame and D the hidden dimension:

- Spatial Attention: $O(T \cdot N^2 \cdot D)$
- Temporal Attention: $O(N \cdot T^2 \cdot D)$

This factorization is more efficient than joint attention: $O((T \cdot N)^2 \cdot D)$

4 Experimental Work

This section gives a detailed report of our proposed TimeSformer model, where we evaluate it's performance on the combined MS-ASL and WLASL dataset. We have used some of the standard classification and evaluation metrics to evaluate our model.

4.1 Classification Report

The Table 1 summarises the evaluation results of our model on the test set and gives a visual analysis of the model's performance and error patterns.

Table 1. Classification Report of the proposed TimeSformer model

Videos	Precision	Recall	F1-Score	
000051.mp4	0.85	0.85	0.85	
000102.mp4	0.73	0.85	0.79	
000155.mp4	0.75	0.92	0.83	
000206.mp4	0.92	0.92	0.92	
000257.mp4	1.00	0.77	0.87	
000312.mp4	0.86	0.86	0.86	
000362.mp4	1.00	1.00	1.00	
000417.mp4	0.89	0.57	0.70	
000471.mp4	1.00	1.00	1.00	
000528.mp4	1.00	1.00	1.00	
000578.mp4	0.79	0.92	0.85	
000633.mp4	0.87	0.93	0.90	
Accuracy			0.88	159
Macro Avg	0.89	0.88	0.88	159
Weighted Avg	0.89	0.88	0.88	159

4.2 Confusion Matrix

The confusion matrix in Fig. 3 shows correct predictions along the diagonal and misclassifications off-diagonal.

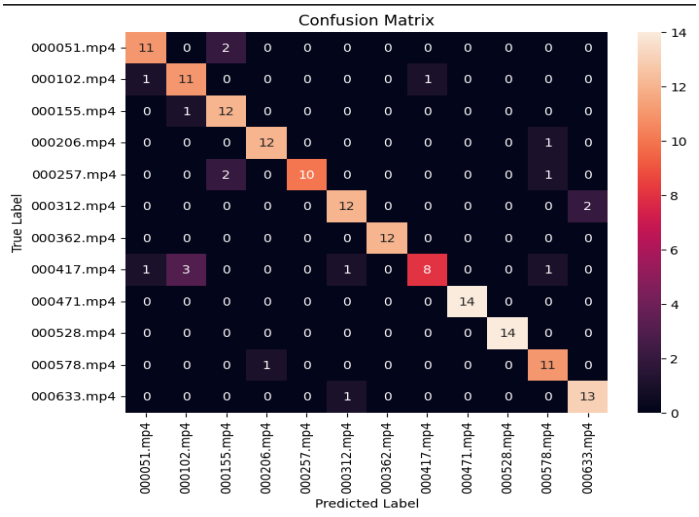

Fig. 3. *Confusion matrix of TimeSformer on test set*

5 Result Analysis

The performance of the proposed model was rigorously evaluated, and the results provide strong evidence for the effectiveness our proposed model for CSLR system.

5.1 Training vs Validation (Loss curve)

The loss curve, depicted in the Fig. 4 has two lines.The training loss with dashed blue line which shows a consistent decline, eventually stabilizes indicating that the model has successfully minimized the error on the training data. The validation loss with solid orange line decreases sharply initially, but then becomes stable reaching a minimum but then rises slightly in the epochs to come.

We have used the early stopping startegy to select the model from the epoch with lowest validation loss, which ensures an optimal generalization, making sure to choose the best performing model.

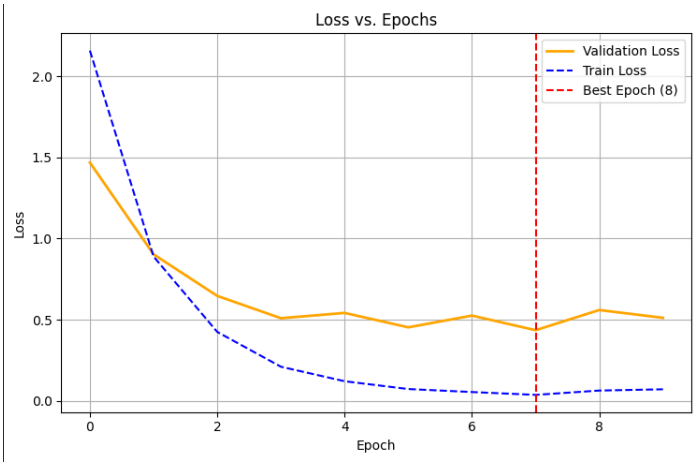

Fig. 4. *Loss Curve of TimeSformer on test set*

5.2 Training vs Validation (Accuracy curve)

The accuracy curve in the Fig. 5 gives an overview of the proposed model's classification performance. The training accuracy with dashed blue line rises rapidly and eventually approaches 99% stabilizing there. The validation accuracy with solid orange line, rises to a peak performance of 89.13% and then fluctuates in subsequent epochs. With the early stopping we have selected the peak performing model.

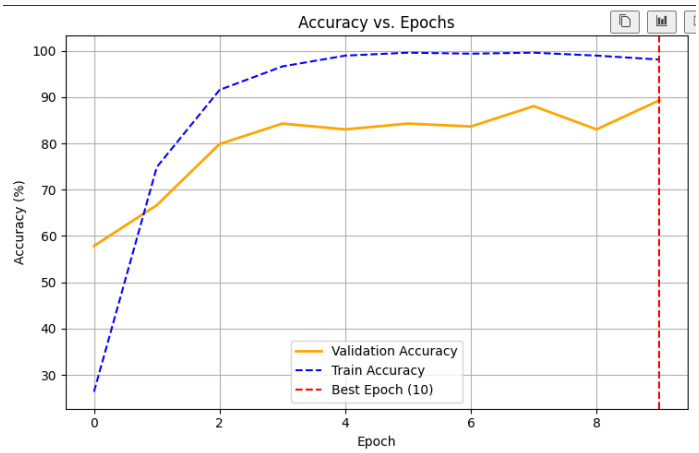

Fig. 5. *Accuracy Curve of TimeSformer on test set*

5.3 Comparison with Existing Models

The table 2 compares out proposed model's accuracy with some of the other models mentioned in the literature survey.

Table 2. Comparison of TimeSformer with existing CSLR approaches

Model	Dataset	Accuracy (%)
LSTM+MediaPipe [6]	6-Emergency Signs	88.88
B3D Net [11]	SGR_Dataset	88.7
MobileNet V2+Bert encoder [13]	INLCUDE	88.11
TimeSformer (ours)	MS-ASL+WLASL(12 classes)	**89.13**

As we can see, our model has achieved a peak accuracy of 89.13 by using a 75:25 split of the dataset, where 25% was used for validating and 75% for training.

6 Conclusion

To summarise, our study shows the effectiveness of the TimeSformer model in comparison to the traditional models in capturing long-range temporal dependencies and predicting the hand-sign with high accuracy. This methodology of fine-tuning a pretrained timesformer model not only tackles the problem of insufficient data but also allows the model to learn a representation quite well. As compared to the joint-space time attention [17], the divided space-time attention is computationally less intensive. Overall, these findings validate the use of TimeSformer for the task of sign language recognition, and this model can

be used as a foundation for the development of more accurate models for the betterment of SLR systems.

7 Future work

While this study provides a solid baseline for creating a complete SLR system, it does not address the problem of seamless translation of multiple hand-signs to effectively translate the hand-signs to naturally spoken language. Therefore, the main direction for future work is to expand the system's capabilities from recognising individual signs to translating entire sentences. To do so, we propose the use of Large Language Model (LLMs) to convert a sequence of signs to complete sentences.

References

1. Singh, D. K., Kumar, A., & Ansari, M. A. (2021, February). Robust modelling of static hand gestures using deep convolutional network for sign language translation. In 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) (pp. 487-492). IEEE.
2. Damaneh, M. M., Mohanna, F., & Jafari, P. (2023). Static hand gesture recognition in sign language based on convolutional neural network with feature extraction method using ORB descriptor and Gabor filter. Expert Systems with Applications, 211, 118559.
3. Kushwaha, R., Kaur, G., & Kumar, M. (2023, May). Hand Gesture Based Sign Language Recognition Using Deep Learning. In 2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC) (pp. 293-297). IEEE.
4. Kolkur, A., Yattinmalgi, A., Korimath, G., Chikkamath, S., Nirmala, S. R., & Budihal, S. (2024, March). Deep Learning based Indian Sign Language Recognition for People with Speech and Hearing Impairment. In 2024 IEEE International Conference on Contemporary Computing and Communications (InC4) (Vol. 1, pp. 1-5). IEEE.
5. Kankariya, S., Thakre, K., Solanki, U., Mali, S., & Chunawale, A. (2024, March). Sign language gestures recognition using CNN and inception V3. In 2024 International Conference on Emerging Smart Computing and Informatics (ESCI) (pp. 1-6). IEEE.
6. Daga, S., Dusane, A., & Bobby, D. (2024, March). With You-Indian Sign Language Detection and Alert System. In 2024 International Conference on Emerging Smart Computing and Informatics (ESCI) (pp. 1-5). IEEE.
7. Khartheesvar, G., Kumar, M., Yadav, A. K., & Yadav, D. (2024). Automatic Indian sign language recognition using MediaPipe holistic and LSTM network. Multimedia Tools and Applications, 83(20), 58329-58348.
8. Das, S., Biswas, S. K., & Purkayastha, B. (2024). Occlusion robust sign language recognition system for indian sign language using CNN and pose features. Multimedia Tools and Applications, 83(36), 84141-84160.
9. Das, A., Maitra, K., Roy, S., Ganguly, B., Sengupta, M., & Biswas, S. (2023, November). Development of a Real Time Vision-Based Hand Gesture Recognition System for Human-Computer Interaction. In 2023 IEEE 3rd Applied Signal Processing Conference (ASPCON) (pp. 294-299). IEEE.

10. Long, Z., Liu, X., Qiao, J., & Li, Z. (2023, August). Sign Language Recognition Based On Facial Expression and Hand Skeleton. In 2023 38th Youth Academic Annual Conference of Chinese Association of Automation (YAC) (pp. 237-241). IEEE.
11. Shao, L., Chao, X., & Lirong, Z. (2020, October). Gesture Recognition Based on Video Sequence with B3D Convolutional Neural Network. In 2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP) (pp. 26-31). IEEE.
12. Yang, F., Hou, Z., & Zhang, M. (2023, November). Video based Gesture Recognition Method using Two-Stream 3D-ConvNet Network for Coal Mine Emergency Rescue Robot. In 2023 China Automation Congress (CAC) (pp. 5403-5407). IEEE.
13. Saproo, V., & Aggarwal, R. K. (2024, August). A Transformer Based Indian Signed Language Recognition. In 2024 First International Conference on Pioneering Developments in Computer Science & Digital Technologies (IC2SDT) (pp. 170-174). IEEE.
14. Agarwal, A., Sreemathy, R., Turuk, M., Jagdale, J., & Kumar, V. (2023, December). Indian Sign Language Recognition using Skin Segmentation and Vision Transformer. In 2023 IEEE 20th India Council International Conference (INDICON) (pp. 857-862). IEEE.
15. Francis, R. (2024, August). Sign2Text: Deep Learning-based Sign Language Translation System Using Vision Transformers and PHI-1.5 B. In 2024 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET) (pp. 282-287). IEEE.
16. Singla, V., Bawa, S., & Singh, J. (2024). Enhancing Indian sign language recognition through data augmentation and visual transformer. Neural Computing and Applications, 36(24), 15103-15116.
17. Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10023-10033).
18. Wang, S., Wang, K., Yang, T., Li, Y., & Fan, D. (2022). Improved 3D-ResNet sign language recognition algorithm with enhanced hand features. Scientific Reports, 12(1), 17812.
19. Bertasius, G., Wang, H., & Torresani, L. (2021, July). Is space-time attention all you need for video understanding?. In ICML (Vol. 2, No. 3, p. 4).
20. Li, D., Rodriguez, C., Yu, X., & Li, H. (2020). Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. In The IEEE Winter Conference on Applications of Computer Vision (pp. 1459–1469).
21. Vaezi Joze, H. R., & Koller, O. (2019, September). MS-ASL: A large-scale data set and benchmark for understanding American Sign Language. The British Machine Vision Conference (BMVC).