SignX: Real-Time Sign Language Action Recognition & Translation

Aman Saini
m24cse003@iitj.ac.in
IIT Jodhpur, India

Swaminath Bera
m24cse007@iitj.ac.in
IIT Jodhpur, India

Prathmesh Gosavi
m24csa023@iitj.ac.in
IIT Jodhpur, India

Abstract

Sign language serves as a crucial communication medium for the deaf and hard-of-hearing communities. However, the absence of real-time translation tools often leads to communication barriers in daily interactions. To address this challenge, we introduce **SignX**, a real-time system designed to recognize and translate sign language gestures into textual words. Leveraging the comprehensive Word-Level American Sign Language (WLASL) dataset, our approach focuses on accurately interpreting isolated sign gestures, facilitating seamless communication between sign language users and the broader community.

Our system employs a 3D Convolutional Neural Network (3D-CNN) architecture to effectively capture both spatial and temporal features inherent in sign language gestures. By processing sequences of video frames, the model learns to identify dynamic hand movements and transitions, enabling precise recognition of various signs. This methodology is inspired by advancements in spatiotemporal feature extraction techniques, which have demonstrated significant improvements in action recognition tasks. Through rigorous training and optimization, SignX achieves real-time performance, offering a practical solution to bridge the communication gap faced by the deaf community.

1 Introduction

For millions of deaf and hard-of-hearing people across the world, sign language is a vital communication medium. However, despite its importance, there are still many challenges in building a more inclusive society, one of which is the communication barrier between sign language users and those who are not familiar with it. This gap is bridged by the **SignX** system, which aims to develop a real-time sign language recognition system capable of accurately identifying gestures and translating them into text.

Recent advances in deep learning for video understanding and action recognition form the foundation of our approach. In particular, our methodology leverages 3D Convolutional Neural Networks (3D-CNNs), which have been shown to better capture spatiotemporal features from video data than traditional 2D CNNs. Unlike 2D CNNs, 3D CNNs perform convolutions across both spatial and temporal dimensions simultaneously, allowing them to model motion information across neighboring frames more effectively. Ji et al. (3) describe how stacking a set of contiguous frames into a cube and convolving with a 3D kernel enables the network to learn patterns of motion across time.

Building further on these techniques, Varol et al. (1) demonstrate that Long-term Temporal Convolutions (LTC) can improve action recognition by expanding the temporal receptive field of the model. They show that actions typically exhibit long-term temporal structure, and breaking them into short clips hinders recognition accuracy. Inspired by this, **SignX** proposes a hybrid architecture combining 3D CNNs with Temporal Convolutional Networks (TCNs) to jointly model short-term spatial features and long-term temporal dependencies in sign language gestures.

Preprint. Under review.

2 Literature Review

In recent years, research in computer vision has witnessed significant progress in the domain of action recognition from videos. Traditional approaches often relied on handcrafted features to learn spatiotemporal representations, but deep learning methods have recently emerged as powerful alternatives due to their ability to learn complex patterns directly from raw data.

Convolutional Neural Networks (CNNs) have achieved outstanding success in image recognition tasks. However, their extension to video analysis presents additional challenges due to the temporal dimension inherent in video data. Several methods have been proposed to integrate motion information into CNN-based models. One of the earliest and most influential efforts was presented by Ji et al. (3), who introduced 3D CNNs. Their model applied 3D convolutions to stacked video frames, enabling the network to capture spatiotemporal features effectively. They demonstrated that 3D CNNs outperformed traditional 2D CNNs in human action recognition tasks due to their capability to model motion across multiple contiguous frames.

Building on this foundation, Tran et al. (2) introduced the C3D architecture, which utilized compact 3D convolutional kernels of size $3 \times 3 \times 3$. Their investigation showed that such small kernels are optimal for capturing spatiotemporal dynamics and result in compact yet highly discriminative video descriptors. The C3D model achieved state-of-the-art performance on several action recognition benchmarks and proved valuable for other video analysis tasks as well.

Varol et al. (1) proposed Long-term Temporal Convolutions (LTC) to capture extended temporal dependencies in video sequences. By increasing the temporal receptive field of 3D CNNs from 16 frames up to 60–100 frames, they demonstrated improved recognition performance. Their findings highlighted the importance of modeling long-range temporal structures and cautioned against slicing actions into short clips, which may omit essential context. Despite their strengths, 3D CNNs are often computationally expensive and prone to overfitting, particularly when trained on small datasets. To address this, researchers have explored factorized 3D convolutions and hybrid 2D–3D architectures.

Another direction for temporal modeling is the use of Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks. Donahue et al. (7) introduced Long-term Recurrent Convolutional Networks (LRCN), which combine CNNs for spatial feature extraction with LSTMs for temporal sequence modeling. Additionally, input modality plays a crucial role in video understanding. While RGB frames capture appearance information, optical flow captures motion, and combining them—as done in Two-Stream Networks by Simonyan and Zisserman (6)—has proven effective.

More recently, architectures like SlowFast networks have been developed to enhance both accuracy and efficiency. These models utilize two separate pathways operating at different temporal resolutions, enabling the model to process both slow and fast motion patterns simultaneously. In summary, while 3D CNNs and their extensions have significantly advanced video-based action recognition, challenges remain in computational efficiency, long-term temporal modeling, and multi-modal integration. Future work is expected to further refine these areas and lead to more robust video understanding systems.

3 Problem Statement

SignX is an innovative system designed to recognize sign language actions in real-time. It leverages two cutting-edge deep learning architectures: 3D Convolutional Neural Networks (3D CNNs) and Temporal Convolutional Networks (TCNs). This project addresses the critical challenge of accurately identifying and translating sign language gestures on the fly, which is essential for enhancing communication accessibility for the deaf and hard-of-hearing community.

The problem involves modeling both spatial and temporal aspects of sign gestures from raw video input in a way that is robust, fast, and scalable. Unlike conventional systems that may struggle with frame-level limitations or handcrafted features, SignX seeks to learn these representations directly from data. The ultimate objective is to develop a system capable of delivering real-time, accurate gesture-to-text translation that can be deployed in practical assistive technology applications.

4 Proposed Method

The **SignX** system utilizes a hybrid deep learning architecture that integrates 3D Convolutional Neural Networks (3D CNNs) and Temporal Convolutional Networks (TCNs) to perform accurate and efficient sign language recognition in real time. This combination leverages the strengths of both architectures: 3D CNNs effectively capture spatial and short-term temporal patterns from video sequences, while TCNs are designed to model long-range temporal dependencies, thereby enabling the system to learn richer and more contextualized gesture representations.

4.1 Data Preprocessing and Input Pipeline

The input to our system consists of video sequences containing isolated sign language gestures. Before feeding these sequences into the model, we apply several preprocessing steps to prepare consistent and informative inputs:

- **Video Segmentation:** Each raw video is segmented into fixed-length clips, typically containing 16 to 60 consecutive frames, depending on the experimental setup.
- **Spatial Preprocessing:** All frames are resized to a resolution of 112×112 pixels and normalized to zero mean and unit variance.
- **Multimodal Input Preparation:** We extract three input modalities:
  - RGB frames capturing appearance
  - Optical flow (Brox method) for motion
  - Skeletal tracking data for joint positions
- **Data Augmentation:** We use random cropping, flipping, temporal clipping, and multiscale resizing.

4.2 Hybrid Architecture Integration

- **Feature Extraction Pipeline:** The 3D CNN extracts spatiotemporal features, which are passed to the TCN for capturing long-term dependencies.
- **Multi-stage Training Strategy:**
  - Train 3D CNN on 16-frame clips
  - Extend to 60–100 frame sequences
  - Train TCN on extracted features
- **Loss Function:** We use cross-entropy loss with L_2 regularization, dropout, and batch normalization.

5 Experiment & Results

In this section, we present the per-class accuracy metrics of our sign language recognition model. The table below summarizes the top-1, top-5, and top-10 average per-class accuracy values obtained during the evaluation process.

We performed the experiments on validation set of size 2845 videos.

Table 1: Per-Class Accuracy Metrics

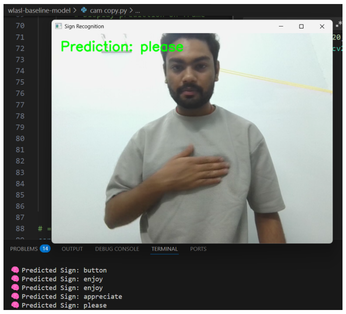
Per-Class Accuracy	**Value**
Top-1 Average Per Class	0.4270
Top-5 Average Per Class	0.7621
Top-10 Average Per Class	0.8380

6 Citations, Figures, Tables, and References

In this section, we present various visual examples that demonstrate the webcam output of different sign language gestures recognized in real-time. The following figures show signs such as "please", "sleepy", "sorry", "restroom", and "pray", which were captured through our system's webcam input. These images are part of our experimental setup to analyze the performance of our sign language recognition system. Additionally, we will provide the necessary citations and references to support our methodology and the models used.

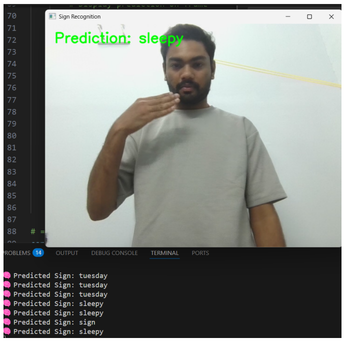
6.1 Figures

The following images represent different signs captured during the real-time recognition process:



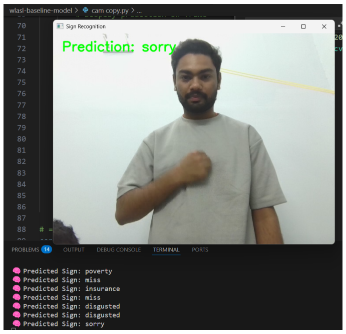
```
🧠 Predicted Sign: button
🧠 Predicted Sign: enjoy
🧠 Predicted Sign: enjoy
🧠 Predicted Sign: appreciate
🧠 Predicted Sign: please
```

Figure 1: Sample webcam output showing sign "please".



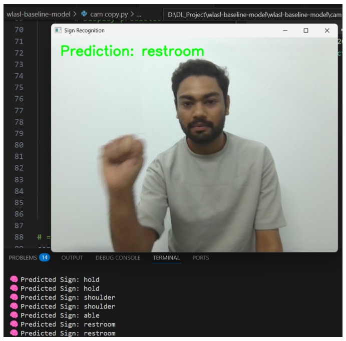
```
🧠 Predicted Sign: tuesday
🧠 Predicted Sign: tuesday
🧠 Predicted Sign: tuesday
🧠 Predicted Sign: sleepy
🧠 Predicted Sign: sleepy
🧠 Predicted Sign: sign
🧠 Predicted Sign: sleepy
```

Figure 2: Sample webcam output showing sign "sleepy".



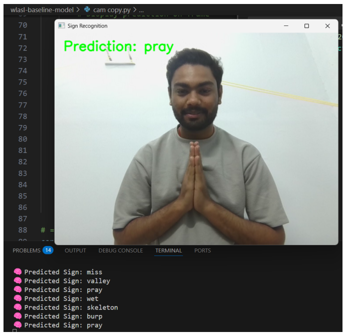
```
🧠 Predicted Sign: poverty
🧠 Predicted Sign: miss
🧠 Predicted Sign: insurance
🧠 Predicted Sign: miss
🧠 Predicted Sign: disgusted
🧠 Predicted Sign: disgusted
🧠 Predicted Sign: sorry
```

Figure 3: Sample webcam output showing sign "sorry".


wlasl-baseline-model > cam copy.py > ...
D:\DL_Project\wlasl-baseline-model\wlasl-baseline-model\cam
Sign Recognition
Prediction: restroom
PROBLEMS 14
OUTPUT
DEBUG CONSOLE
TERMINAL
PORTS

```
🧠 Predicted Sign: hold
🧠 Predicted Sign: hold
🧠 Predicted Sign: shoulder
🧠 Predicted Sign: shoulder
🧠 Predicted Sign: able
🧠 Predicted Sign: restroom
🧠 Predicted Sign: restroom
```

Figure 4: Sample webcam output showing sign "restroom".



```
🧠 Predicted Sign: miss
🧠 Predicted Sign: valley
🧠 Predicted Sign: pray
🧠 Predicted Sign: wet
🧠 Predicted Sign: skeleton
🧠 Predicted Sign: burp
🧠 Predicted Sign: pray
```

Figure 5: Sample webcam output showing sign "pray".

References

[1] G. Varol, I. Laptev, and C. Schmid. Long-term temporal convolutions for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1510–1517, 2018.

[2] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[3] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013.

[4] D. Li, C. Rodriguez, X. Yu, and H. Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1469, 2020.

[5] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017.

[6] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, pages 568–576, 2014.

[7] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2625–2634, 2015.