IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 February 2025, accepted 4 March 2025, date of publication 25 March 2025, date of current version 11 April 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3554618

RESEARCH ARTICLE

Toward Real-Time Recognition of Continuous Indian Sign Language: A Multi-Modal Approach Using RGB and Pose

M. GEETHA[1], NEENA ALOYSIUS[2], DARSHIK A. SOMASUNDARAN[2], AMRITHA RAGHUNATH[2], AND PREMA NEDUNGADI[1]

[1]Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri, Kerala 690525, India
[2]AmritaCREATE, Amrita Vishwa Vidyapeetham, Amritapuri, Kerala 690525, India

Corresponding author: M. Geetha (geetham@am.amrita.edu)

This work was supported by the Ministry of Electronics and Information Technology (Meity), Government of India.

ABSTRACT Sign language recognition (SLR) involves translating visual gestures into meaningful text or speech, bridging the communication gap between signers and non-signers. However, real-time recognition remains a critical challenge due to variability in signing speed, subtle hand gestures, and the computational complexity of processing video data in real time. Existing methods often rely on encoding RGB videos into latent representations, making them unsuitable for real-time interaction. To address these challenges, we present SignFlow, a network for real-time recognition of continuous sign gestures. Our approach introduces a novel pre-processing technique to down-sample video streams, ensuring compatibility with varying frame rates across different devices. For the first time, the core components of our network are pre-trained on domain-specific Indian Sign Language (ISL) data. The CNN is pre-trained using ISL word videos, while the Transformer is pre-trained on Mediapipe pose estimates from ISL videos. This pretraining effectively captures the nuances of hand shapes and body movements unique to ISL, significantly enhancing sentence recognition. SignFlow combines a pre-trained CNN for feature extraction and a Transformer for learning the temporal dynamics of continuous signing. The framework is fine-tuned end-to-end using Connectionist Temporal Classification. For evaluating the real-time models, we have introduced a detection rate metric, which measures how accurately individual gestured words are recognized within a sequence, regardless of their order. SignFlow achieves a Word Error Rate (WER) of 19 on the Continuous ISL dataset, demonstrating its effectiveness for real-time ISL recognition. Additionally, it shows competitive performance on the German Phoenix 2014 and Phoenix 2014T datasets.

INDEX TERMS CNN, connectionist temporal classification, down sampling, Indian sign language, sign flow, sign language recognition, transformer.

I. INTRODUCTION

Sign language is a rich and expressive form of communication used by hard-of-hearing individuals to convey information, ideas, and emotions. Unlike spoken languages, sign languages rely on visual gestures, hand movements, facial expressions, and body postures to communicate meaning. Each country often has its own distinct sign language with unique grammar and vocabulary. For the deaf community, sign languages are integral to social, educational, and cultural interactions, fostering identity and belonging. In India, Indian Sign Language (ISL) is crucial for an estimated 18 million deaf individuals. However, despite its recognition as an official language, there remain significant accessibility challenges in government services for the deaf population. The lack of ISL awareness among officials often creates communication barriers, limiting access to healthcare, education, legal services, and government schemes. This results in the marginalization of the deaf community.

Additionally, public resources such as websites and announcements largely rely on spoken or written language, leaving the deaf community excluded. To bridge this gap,

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

real-time sign language recognition (SLR) technologies are essential for enabling equal access to services. However, despite the demand, there are few products available that perform real-time recognition of continuous sign language. Achieving real-time recognition is particularly challenging due to the complexity of interpreting continuous hand gestures, variations in signing speed, and the high computational demands of processing video data in real-time. Existing models rely on encoding RGB videos into latent representations, which are often unsuitable for real-time interaction due to latency and accuracy issues.

To address these challenges, we introduce SignFlow, a pioneering real-time continuous sign language recognition (CSLR) network designed specifically for ISL. SignFlow is the first system of its kind, leveraging pretraining on ISL-specific datasets, which captures the nuances of hand shapes and body movements unique to Indian Sign Language. Pretraining plays a critical role in improving recognition accuracy and reducing the model's reliance on general visual data. Our model utilizes a 3DResNet as a feature extractor, pre-trained on isolated ISL videos, and an 8-layer Transformer Encoder pre-trained on MediaPipe Pose and Hand estimates from the Continuous ISL dataset. This novel pretraining strategy, applied to both components, ensures domain-specific feature extraction and sequence learning.

A significant challenge in real-time recognition is the ability to process continuous signing across variable frame rates, signer variations, and gesture complexities. To overcome this, SignFlow introduces a novel preprocessing technique to downsample input videos dynamically, ensuring compatibility with different devices while selecting the most relevant frames for processing. Additionally, the model is trained end-to-end using Connectionist Temporal Classification (CTC) loss, optimizing performance without needing pre-segmented input data. Furthermore, SignFlow is pre-trained and fine-tuned specifically for ISL, making it the first framework to address the complexities of real-time ISL sentence recognition. By fine-tuning the feature extractor iteratively, we continuously improve recognition accuracy. To evaluate the real-time performance of the model, we introduce a detection rate metric, which measures how accurately individual gestured words are recognized in a sequence, regardless of their order. This project was funded by Ministry of Electronics and Information Technology (Government of India), towards developing ISL chatbot for the UMANG app [1] named eRaktkosh [2]. This framework marks the first time a pre-training strategy with ISL dataset has been adopted for developing a sign recognition model, where the datasets of the pretraining task and the downstream fine-tuning task are highly correlated. As there is currently no benchmark ISL CSLR dataset available, we took the initiative to collect the Continuous ISL dataset as well as videos of isolated ISL words for pretraining. Gloss-level annotation is also provided for the continuous dataset. Before using this data for model training and evaluation, we thoroughly cleaned and pre-processed it to ensure the highest quality. The dataset will be published in the near future.

This paper provides detailed information about all the test cases and results reported on the dataset, highlighting the effectiveness and performance of our proposed continuous sign language recognition framework for ISL. The key contributions of this work can be summarized as follows:

- Real-time CSLR for ISL: We introduce SignFlow, the first real-time continuous sign language recognition framework for ISL, addressing the challenges of real-time gesture processing.
- Pretraining on ISL-specific Data: The core components of SignFlow are pre-trained on ISL-specific data, with the CNN trained on isolated ISL videos and the Transformer trained on MediaPipe Pose and Hand estimates from the Continuous ISL dataset, using CTC loss
- Novel Video Pre-processing Technique: A dynamic video down-sampling technique ensures compatibility across devices with varying frame rates, increasing the diversity of training data and selecting highly correlated frames during inference.
- Innovative Training Strategy: SignFlow is the first to employ a Transformer model without positional encoding, improving word detection rates within the ISL vocabulary and enhancing real-time recognition accuracy.

II. RELATED WORKS

We conducted a thorough literature review in sign recognition before undertaking this work, [1], [2], [3]. Recent advancements in CSLR (Continuous Sign Language Recognition) have embraced Deep Learning methodologies structured around three core modules: the feature extractor, sequence learning, and alignment learning. The feature extraction components comprise the visual module for encoding short-term spatial details. CNNs [4] are commonly used for this purpose. The sequence-learning module is designed to encapsulate long-term contextual information. Classifiers can derive posterior probabilities essential for recognition using the features extracted from these modules. Moreover, the alignment module is critical in establishing precise synchronization between video clips and glosses. This synchronization is essential to ensure the accuracy and reliability of the training process. Recent work by Zuo et al. [5] introduced a Multi-Scale Temporal Network (MSTNet), which enhances temporal feature extraction by utilizing a multi-scale temporal block and transformers, leading to improved accuracy in CSLR without requiring prior knowledge. AdaSize [6] was proposed lately to tackle spatial redundancy in CSLR by dynamically adjusting frame resolution through a recurrent policy network, reducing computations and memory usage while maintaining state-of-the-art accuracy. With the introduction of the technique called feature disentanglement, Zuo et al. effectively removed the signer-specific information from backbone and thereby resulted in signer-independent recognition system [7]. The current state-of-the-art results are achieved with ConSignformer model [8], an innovative combination of the Conformer (transformer-variant) and S3D CNN. This architecture holds significant potential for real-world applications, as it has been specifically designed for practical purposes. Before this, TwoStream- SLT [9] by Chen et al. demonstrated exceptional performance on benchmark datasets. The sophisticated architecture features multiple auxiliary losses and Spatial Pyramid Networks, significantly reducing Word Error Rate (WER). The pretraining strategy employed in their previous SingleStream model [10] was also effectively utilized in the latter work. However, unlike ConSignformer, their architecture lacks a transformer backbone, which may cause it to fail when handling longer sequences in practical scenarios. However, the utilization of a multi-cue network with various fusion techniques [11], [12], [13], [14] introduces complexity and susceptibility to data noise. An approach that offers greater efficiency involves using a single cue while adopting cross-modal alignment [15]. CVT-SLR [15] stands out as an advanced single-cue framework that utilizes a Variation-AutoEncoder for sequence learning, a departure from traditional networks that often employ LSTMs or Transformers. Another remarkable contribution is the SMKD approach [16], which employs a single cue and combines 2D CNN-BiLSTM-CTC in its recognition network. Incorporating a novel three-stage optimization strategy in this work has reduced error rates. A similar kind of staged optimization was previously adopted by Cui et al. [17]. Iterative training is another strategy adopted for optimized learning to prevent overfitting issues [18], [19]. This kind of training strengthens the feature extractor. Visual Alignment Constraint (VAC) [20] proposed by Yuecong et al. further enhances the feature extractor with alignment supervision and adds a new perspective on the relationship between visual and alignment modules. In our prior study [21], which focused on single-cue methods, we explored various positioning schemes for the Transformer model [22]—a commonly employed architecture for contextual learning. In the present work, we enhance this foundational architecture by introducing innovative pretraining and pre-processing techniques to equip the network for real-time recognition.

III. DATASET

Currently, there is no publicly available dataset on ISL to evaluate Continuous Sign Language Recognition (CSLR) approaches. We introduce UMANG-eRaktkosh-ISL-Continuous2023 and UMANG-eRaktkosh-ISL-Isolated2023 corpora. To enhance deaf accessibility to the chatbot in eRaktkosh, videos of sentences based on frequently asked questions (FAQs) from the app were collected as UMANG-eRaktkosh-ISL-Continuous2023 corpus. These videos were used to train a deep-learning model capable of real-time recognition. This effort represents the first benchmark ISL video dataset in the field of CSLR. Additionally, to adopt the latest pre-training strategy and improve the feature extractor in the end-to-end model, a video dataset consisting of isolated words from the dictionary of the services was also collected as UMANG-eRaktkosh-ISL-Isolated2023 corpus. The statistics of the datasets are given in tables 1 and 2.

TABLE 1. Statistics of UMANG-eRaktkosh-ISL-Continuous2023 corpus.

Vocabulary size	163
Unique ISL sentences count	226
Signers	13
Total videos	11,752
Studio environment videos	2,938
Home environment videos	8,814

TABLE 2. Statistics of UMANG-eRaktkosh-ISL-Isolated2023 corpus.

Vocabulary size	163
Signers	13
Total videos	8,476
Studio environment videos	2,119
Home environment videos	6,357

The recordings were conducted by certified ISL interpreters approved by the Rehabilitation Council of India (RCI), along with the assistance of deaf students under the interpreters' guidance. We did not impose any restrictions on the background, signing speed, recording device (laptop or mobile), mode (portrait or landscape), lighting conditions, signer's clothing, posture (sitting or standing), or facial expressions. The videos were captured in both studio and home environments, and we even included selfie videos to ensure a diverse range of scenarios. Compared to the current benchmark German datasets [23], [24], this makes the dataset challenging.

It is important to note that the gesturing in the dataset strictly adheres to the standards set by the Indian Sign Language Research and Training Centre (ISLRTC), which is the official body in India responsible for standardizing gestures across the nation. All valid variations of a word relevant to Eraktkosh were considered for training and thus became part of the video dataset. In cases where standardized signs as per ISLRTC did not exist for words in the service's dictionary, we considered widely used regional variations from across the country. Overall, the recordings involved 13 signers who contributed to this comprehensive dataset.

IV. DATA ENGINEERING

Unlike existing works, our proposed recognition model is developed to integrate with a video-based chatbot aimed at assisting the hard of hearing in India. Since the screen size of a chatbot is limited, the size of the video-capturing window is further restricted. It is challenging to ensure that the gestures performed by a signer are fully captured within the window frame, as the signer may not always be centered in the frame, especially in real-world scenarios. However, we address this challenge with our novel alignment-correction technique, which effectively centralizes the signer.

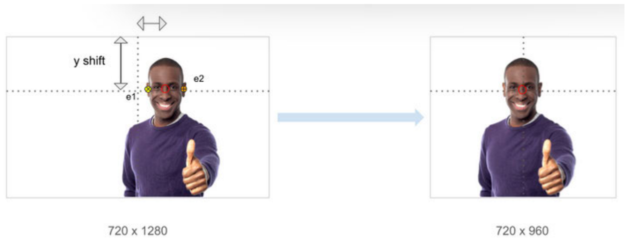


FIGURE 1. **Alignment-correction technique: Estimate ear midpoint and ear distance; evaluate the x-shift and y-shift from the center; scale value. Realign frame to the preferred center.**

In this method, the ear midpoint and ear distance are initially computed using MediaPipe pose estimates. The center of the line connecting these two midpoints corresponds to the signer's center. The exact center of the frame is determined directly from the image dimensions. The x-shift and y-shift, as indicated in the figure, are calculated from the actual center, and the signer is then realigned to the desired center. While maintaining the aspect ratio, the training videos were resized and cropped to a 4:3 aspect ratio with a resolution of 720 × 960 pixels. The methodology is graphically represented in figure 1. During live recognition, the videos are resized directly to 112 × 112 pixels, the size used for training the model.

V. SIGN FLOW NETWORK

In this section, we introduce our SignFlow Network for the recognition of continuous ISL gesturing. Initially, spatial embeddings from the alignment-corrected continuous sign videos were extracted using a mixed convolution 3DResNet model (mc3_18) that has been pre-trained on isolated sign videos. These embeddings are then fed into an 8-layer Transformer Encoder responsible for sequential learning. An effective pretraining was performed for the Transformer model as well. The self-attention layers within the Transformer Encoder are carefully designed to capture the relative information between the frames of a video. To ensure stable learning, each encoder operation is followed by residual addition and normalization.

Given that frame-level annotations are unavailable, we adopt a weaker supervised learning approach using the Connectionist Temporal Classification (CTC) loss function. By applying softmax activation to the spatio-temporal representation generated by the encoder, we derive frame-level gloss probabilities. CTC then models the probability p(Gloss/Input video) by calculating the expectation over all possible frames-to-gloss alignments. This comprehensive approach allows our SignFlow Network to effectively recognize continuous ISL gestures, bridging the gap in sign language recognition and advancing accessibility for the deaf community.

A. PRETRAINING OF CNN

Pre-training involves training a model on one task to help it develop parameters that can be applied to other tasks. We have done our own pretraining task on domain-specific ISL dataset which has helped in making the finetuning efficient. The pre-training task chosen was video classification or Isolated Sign Language Recognition (ISLR). In this study, ISLR refers to recognizing dynamic gestures corresponding to words in the vocabulary of eRaktkosh. Pre-training plays a crucial role in achieving a well-initialized feature extractor, which is implemented as a Convolutional Neural Network (CNN) as shown in Figure 2. Specifically, we selected the 3DResNet model from torchvision - MC3_18, which is already pre-trained on an action recognition dataset. For optimum results, it is essential to perform pretraining and fine-tuning tasks on datasets that exhibit some level of similarity. Consequently, we pre-trained the MC3_18 model using the isolated ISL dataset and saved its checkpoint, which was then further fine-tuned during the end-to-end CSLR training process.

B. PRETRAINING OF TRANSFORMER

The transformer is pre-trained on the Mediapipe Pose and Mediapipe Hand estimates of the Continuous ISL dataset with CTC loss. A detailed architecture diagram is illustrated in figure 4. 3D joint estimates contain X, Y, and Z values, where X and Y are coordinates and Z is the depth measurement. As shown in figure 3, only the upper body estimates were taken from Mediapipe Pose and features were generated based on it. Features taken are:

1) Raw joint estimates
2) Bone length - the Euclidean distance between the pose connection
3) Relative position of each joint with their adjacent or connected joints
4) Joint Euclidean distances - Left hand to right-hand palm, left hand to right-hand elbow, left palm to left shoulder, and right palm to right shoulder distances.

Therefore, the total features count to 443 per frame (Hands + Pose). The feature vector is transformed to 512-dimensional by passing through a linear layer, and then as input to transformer. Isolated Sign Language Recognition (ISLR) is fixed as the pretraining task in this pose-based network. On convergence, the saved checkpoint corresponds to a pre-trained transformer model that is subsequently used for fine-tuning the proposed model.

C. SIGNFLOW NETWORK

The SignFlow Network comprises a Convolutional Neural Network (CNN) responsible for generating the required feature vectors. These vectors assist the Transformer in mapping spatio-temporal relations with the support of Connectionist Temporal Classification (CTC). The CNNs, which can be 2D or 3D, generate either frame-wise embeddings or divide the video into multiple overlapping clips, generating spatial embeddings for each of these clips. In this context, we utilize a 3D Convolutional Neural Network (3DCNN). Distinguishing itself from existing methodologies

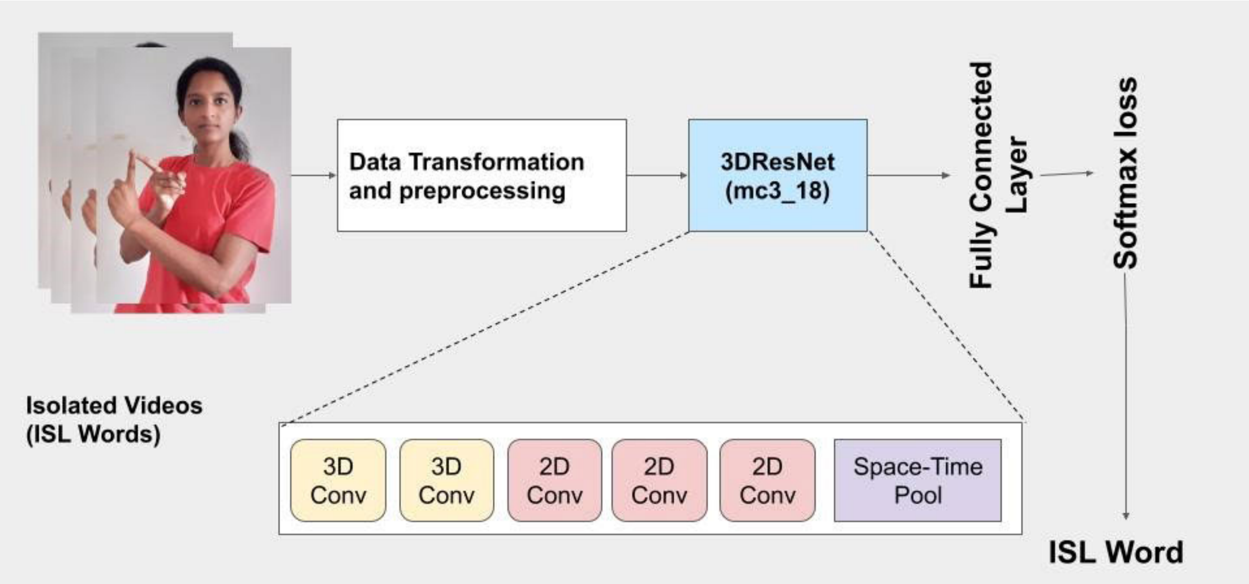


FIGURE 2. MC3_18 ResNet model pretrained on isolated RGB videos.

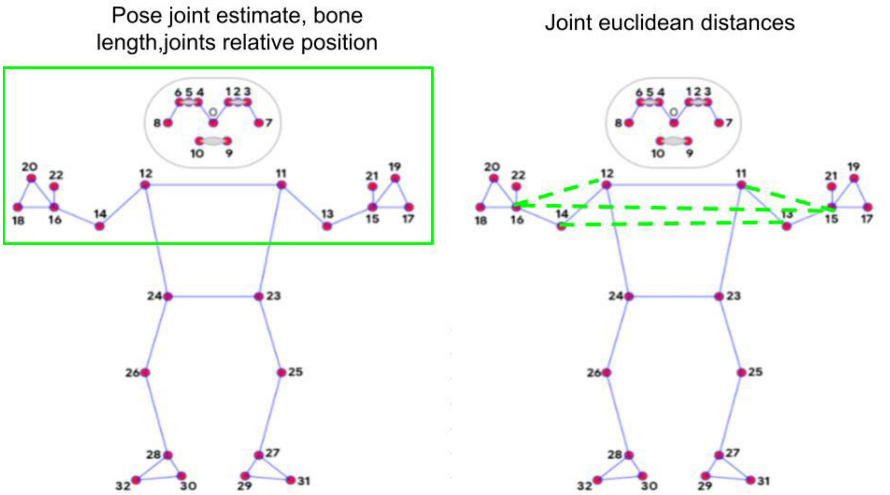

Pose joint estimate, bone length,joints relative position
Joint euclidean distances
6 5 4
1 2 3
8
7
10
9
20
22
12
11
21
19
18
16
14
13
15
17
24
23
26
25
28
27
32
30
29
31

FIGURE 3. Features from Mediapipe pose.

for sign recognition, this approach involves processing all video frames simultaneously within the CNN, yielding an embedding for each video clip. This implies that CNN analyzes the entire video as a cohesive entity, producing meaningful representations for smaller segments known as clips. To accomplish this, 2D pooling is applied to the final layers of the MC3_18 and R3D_18 CNN architectures. This pooling operation reduces the spatial dimensions of the data while retaining crucial features, thereby enabling a more concise and manageable representation.

The resulting feature vectors from the CNN are passed through an 8-layer Transformer encoder. The Transformer encoder represents a potent neural network architecture renowned for its effectiveness in processing sequential data. It excels at capturing long-range dependencies and interdependencies among elements within the vectors. During our experimentation, we explored both absolute and relative position embedding schemes. Remarkably, we also employed the Transformer without any position embeddings, marking the first instance of such usage. As a consequence of the positioning scheme employed within the Transformer, we trained two variations of our model:-

- CNN - Transformer with position embedding
- CNN - Transformer without position embedding

When compared to the current state-of-the-art models, this architecture appears to be relatively uncomplicated. This simplicity has been intentionally maintained to create a lightweight model. The aim is to enable easy deployment

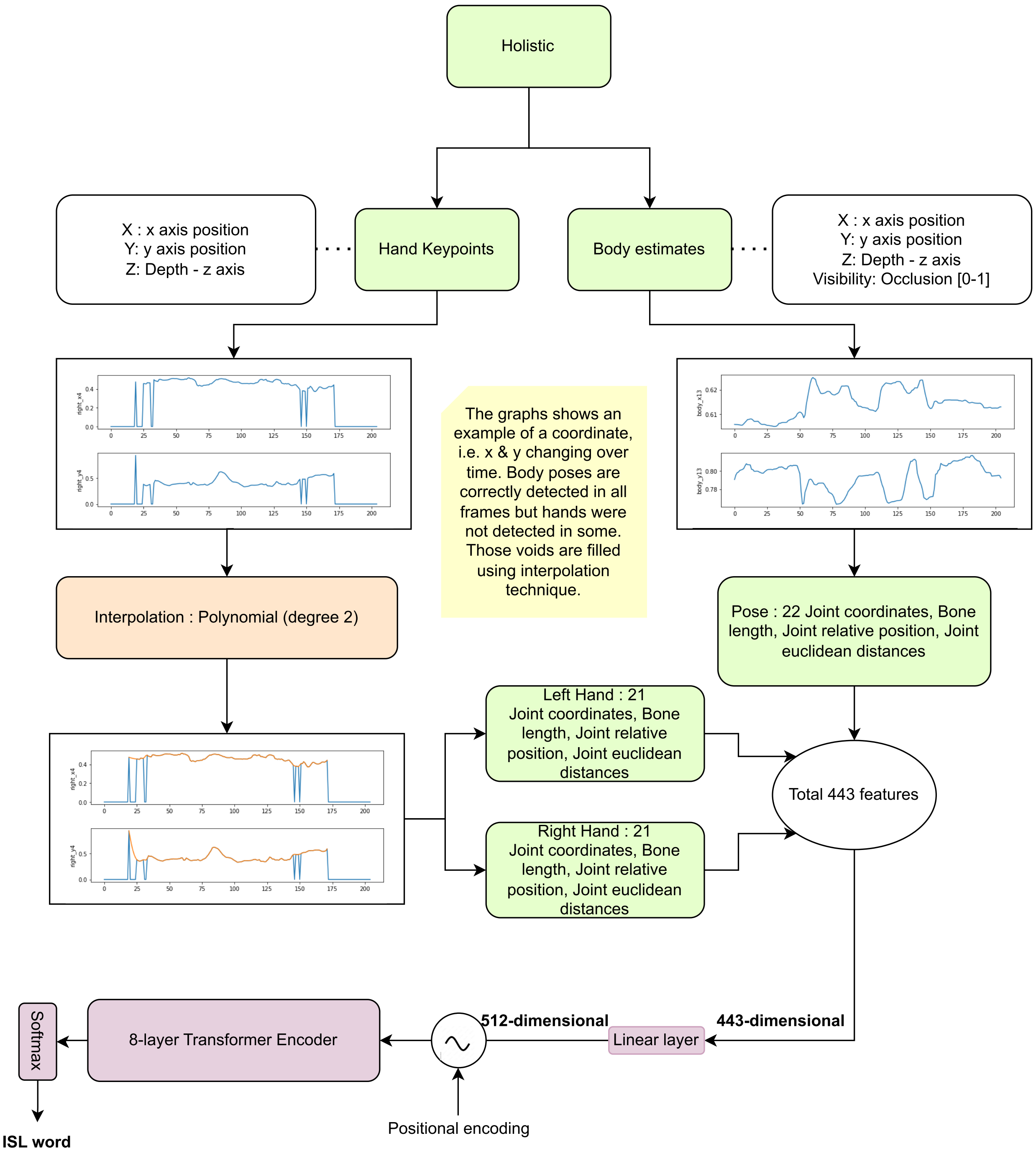

Holistic
Hand Keypoints
Body estimates
X : x axis position
Y: y axis position
Z: Depth - z axis
X : x axis position
Y: y axis position
Z: Depth - z axis
Visibility: Occlusion [0-1]
right_x4
right_y4
body_x13
body_y13
The graphs shows an example of a coordinate, i.e. x & y changing over time. Body poses are correctly detected in all frames but hands were not detected in some. Those voids are filled using interpolation technique.
Interpolation : Polynomial (degree 2)
Pose : 22 Joint coordinates, Bone length, Joint relative position, Joint euclidean distances
Left Hand : 21 Joint coordinates, Bone length, Joint relative position, Joint euclidean distances
Right Hand : 21 Joint coordinates, Bone length, Joint relative position, Joint euclidean distances
Total 443 features
443-dimensional
Linear layer
512-dimensional
Positional encoding
8-layer Transformer Encoder
Softmax
ISL word

FIGURE 4. Transformer pretrained on continuous pose videos.

on the web, facilitating real-time inference. This design choice balances model performance and complexity, ensuring practicality in the context of our ISL chatbot.

Figure 5 illustrates the preprocessing steps applied to the video during real-time inference. The videos, sourced from a cloud database (currently Firebase), undergo alignment correction and resizing as initial measures. Subsequently, using a similarity metric, the video is downsampled to 6fps. This downsampled video is then input to the model for decoding the gesturing actions. Additional processing steps are explained in Figure 6. The preprocessed and downsampled video initially undergoes CNN processing, followed by an 8-layer Transformer encoder, resulting in a spatio-temporal embedding. The CTC beam search decoder is then employed to decode the embedding, producing a sequence of glosses. This prediction is subsequently

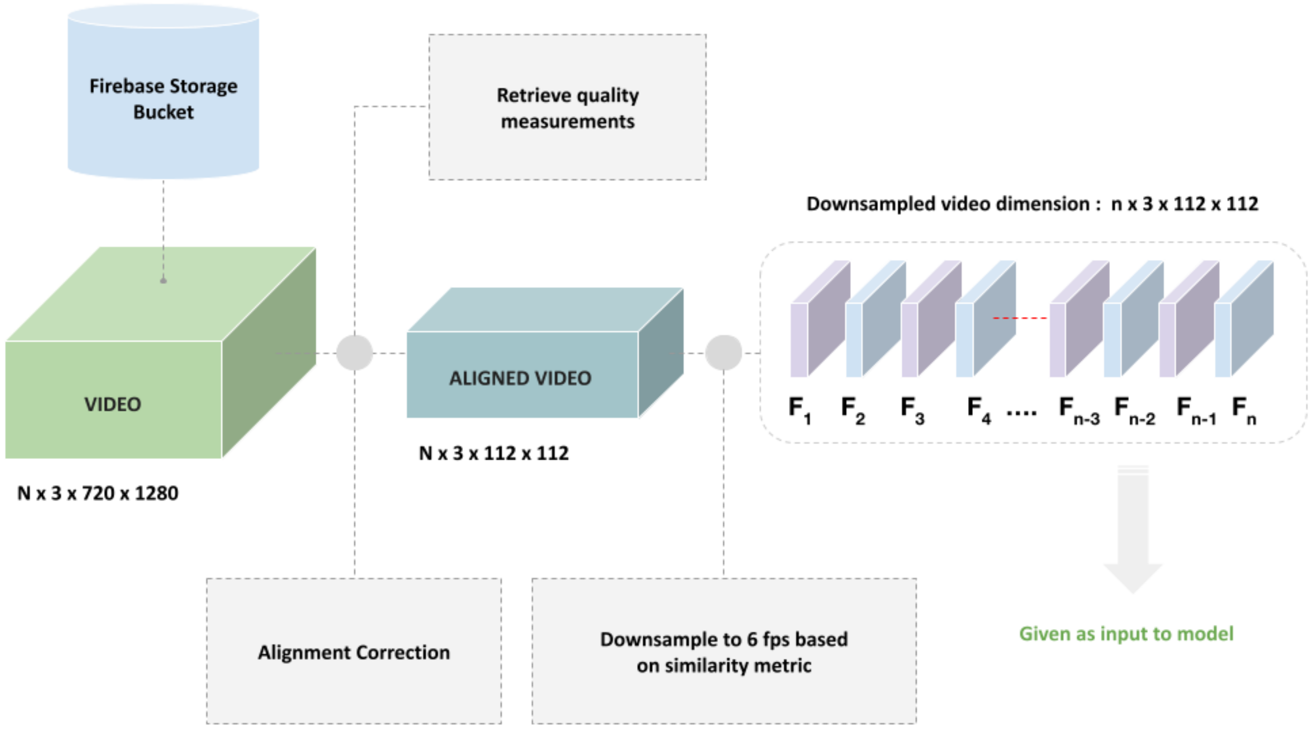

Firebase Storage Bucket
Retrieve quality measurements
Downsampled video dimension : n x 3 x 112 x 112
VIDEO
N x 3 x 720 x 1280
ALIGNED VIDEO
N x 3 x 112 x 112
F1 F2 F3 F4 .... Fn-3 Fn-2 Fn-1 Fn
Given as input to model
Alignment Correction
Downsample to 6 fps based on similarity metric

FIGURE 5. Preprocessing applied during real-time inference. Alignment correction, Video resizing, and Downsampling to 6fps are the important steps in the process. If the input video is less than 6fps, which means a low-quality video, prediction is not done. The user is requested for a retake from another device or network.

forwarded to the next module within the chatbot for further post-processing.

1) LOSS FUNCTION

The loss function is computed at the final layer of the network which is the CTC layer. For a given input sequence X, CTC gives us an output distribution over all possible output sequences Y. This distribution could be used either to infer a likely output or to assess the probability of a given output. This is the alignment proposal and based on the proposal, CTC loss and finally, the decoding, is performed.

a: CTC LOSS

- For a given input, the model is trained to maximize the probability it assigns to the right answer. To do this, the conditional probability $p(Y \mid X)$ is efficiently computed.

b: CTC DECODE

- After training the model, it is used to infer a likely Y (ISL gloss) given an X (ISL video). This means solving

$$Y^* = \arg\max_y p(Y|X) \tag{1}$$

VI. EXPERIMENTS

A. EVALUATION METRICS

To evaluate the CSLR performance of the recognition models, the prevalent metric Word Error Rate (WER) was used [23]. WER takes into account substitutions, insertions, and deletions of words. It is generally used to assess the overall accuracy of the recognition system. WER is calculated as

$$WER = \frac{\#insertions + \#deletions + \#substitutions}{\#words\ in\ reference}. \tag{2}$$

Specific to our work, another metric called Detection Rate was also calculated as,

$$Detection\ Rate = \frac{\#Words\ correctly\ recognized}{\#Words\ in\ reference}. \tag{3}$$

Unlike WER, the detection rate measures how accurately a gesture recognition system identifies and classifies individual gestured words within a sequence, without considering their order. This could be useful in scenarios where the order of gestures doesn't matter, and the goal is to detect and recognize specific gestures regardless of their position in the sequence.

B. IMPLEMENTATION DETAILS

1) DATA AUGMENTATION

Augmentation is crucial in preparing the video data for the proposed model's training and inference. The following paragraphs describe the various augmentation techniques applied on the training videos:

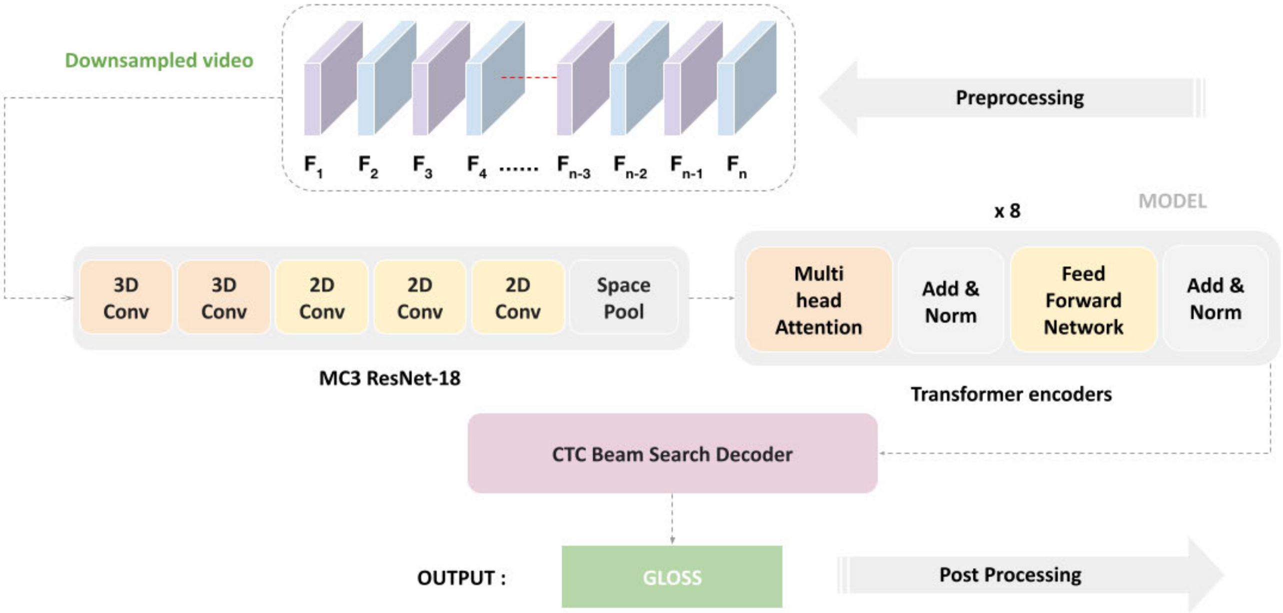


FIGURE 6. Model prediction using the downsampled video. MC3 ResNet-18 does the feature extraction; the 8-layer Transformer encoder does the sequence learning; Spatiotemporal features from the Transformer are decoded using a CTC decoder.

a: RESOLUTION AND DOWNSAMPLING

To ensure uniformity and reduce computational complexity, all videos are resized to a resolution of 112 × 112 pixels. Additionally, to introduce temporal variations and enhance the model's ability to handle diverse video sequences, the videos are randomly downsampled at different sample rates, typically within the range of [5, 8]. This downsampling process creates different perspectives of the same video, providing the model with exposure to varying frame rates.

b: SEGMENTATION AND FRAME SELECTION

Before feeding the videos into the model, they are divided into segments based on the actual and required frame rates. Within each segment, one frame is randomly selected for training. However, during inference (when using the model to make predictions), the most similar frame is selected using a similarity metric instead of choosing a random frame. This ensures that the model focuses on the most relevant frames during testing, enhancing its predictive accuracy.

c: SPATIAL/PIXEL LEVEL AUGMENTATIONS

To generate more spatial variations in the training data and create a robust model capable of handling various challenges, several augmentation techniques are applied. These techniques include random brightness, contrast and hue saturation adjustments, which simulate different video lighting conditions. Random rotation is also applied, slightly altering the orientation of frames (-4 to 4 degrees), introducing minor shifts in the video's perspective. Coarsedropout creates occlusions in the frames, mimicking real-world scenarios where objects might be partially hidden. Random sun flare introduces extreme light exposure in the background, making the model resilient to such variations. Finally, pixel dropout induces spatial variations, helping prevent the model from overfitting to specific pixel patterns.

d: COLLATING AND PADDING

The downsampled videos are collated and padded to a maximum sequence length of 250 frames. This step standardizes the video length, allowing the model to efficiently handle videos of different durations. Padding ensures shorter videos are extended with dummy frames to match the maximum sequence length, maintaining consistency during batch processing.

By incorporating these augmentations, the proposed model is equipped with a diverse and extended dataset that facilitates learning from different spatial and temporal variations, leading to improved generalization and robustness in predicting various challenges that might arise during video analysis tasks.

2) BACKBONES AND HYPER-PARAMETERS

The CNN model utilized for video feature extraction is the Mixed Convolution 3D ResNet. Our Transformers consist of 8 encoder layers, each with a dimensionality of 512 and 8 heads per layer. These networks underwent separate pre-training using isolated word videos for a classification task. To mitigate overfitting, Xavier initialization and a 0.1 dropout rate were applied across all layers and embeddings.

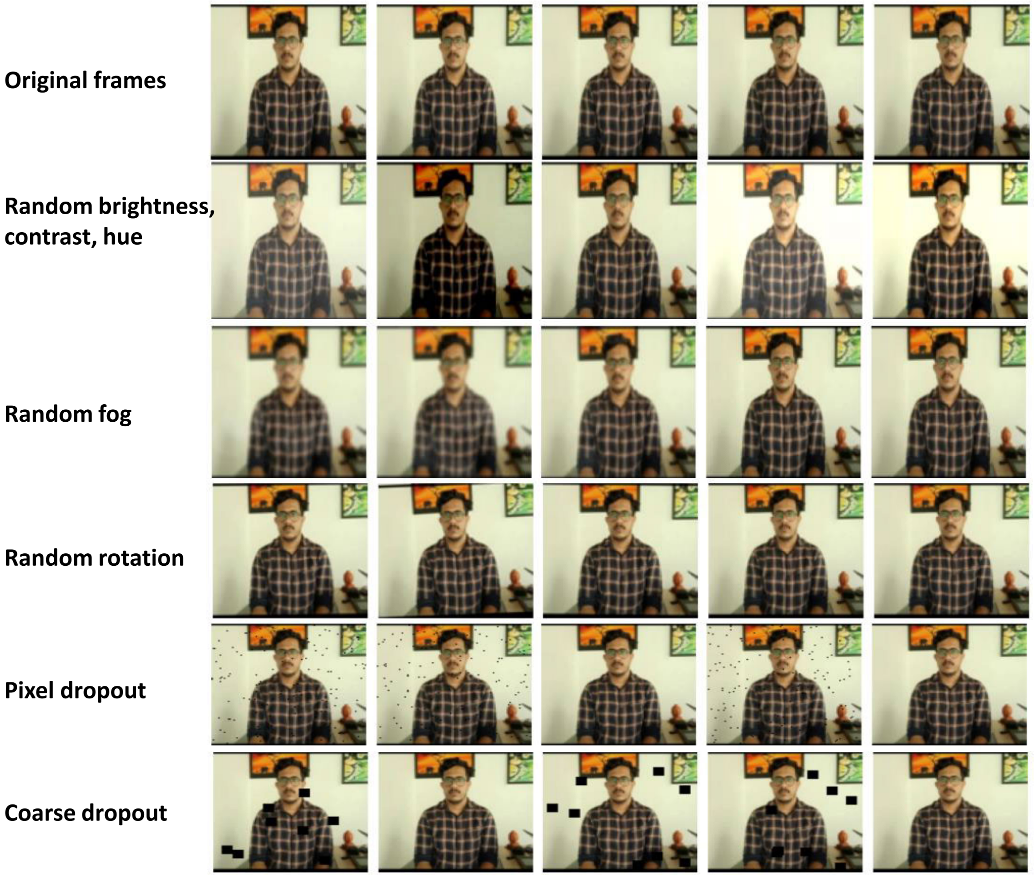


FIGURE 7. Augmentations applied to training videos.

During training, we configured hyperparameters as detailed in Table 3. If there is no improvement in the optimized metric for 8 consecutive evaluation steps, the learning rate is reduced by a factor of 0.7. Training ceases if the learning rate falls below 10^{-6}. On average, it took 120-130 epochs for all models to converge.

For decoding glosses, a greedy search decode [25] was employed during training and validation. During inference, beam search decoding [26] was utilized, with a variable beam size ranging from 0 to 10. A length penalty, following the approach in [27], was applied with α values ranging from 0 to 2. The optimal combination of α and beam size was determined by evaluating the model on the development set and subsequently on the test set.

Implementation of SignFlow was carried out using the PyTorch framework [28]. For CTC beam search decoding, we utilized the Tensorflow [29] implementation. The training process was conducted on an NVIDIA A100 machine equipped with 40 GB RAM.

TABLE 3. Configurations.

Hyper-parameter	Value
Number of encoders	8
Number of heads	8
Dropout	0.1
Learning rate (CNN)	0.0001
Learning rate (Transformer)	0.001
Number of epochs	140
Optimizer	Adam
Batch size	4

VII. RESULTS

In this section, we assess how well the proposed method performs on ISL continuous glosses in the context of eRaktkosh. Following that, we conduct ablation studies to dissect its components' impacts. For this purpose we have trained four models -

- CNN-Transformer with position encoding
- CNN-Transformer without position encoding
- CNN-Transformer pre-trained on pose with-position-encoding
- CNN-Transformer pre-trained on pose without-position-encoding

Lastly, we present the outcomes obtained from applying the method to benchmark German datasets.

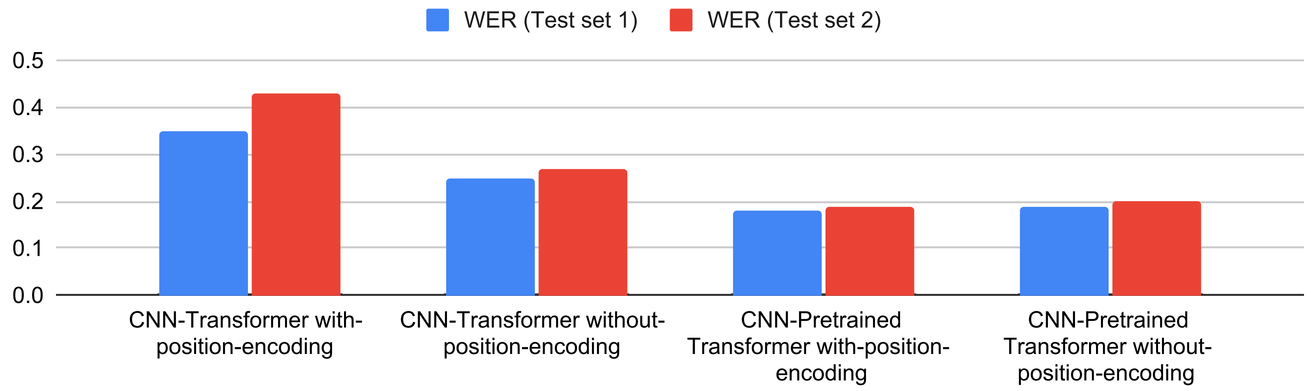


FIGURE 8. Performance comparison of all the models trained on ISL dataset, with challenging test glosses in real-time via chatbot.

A. PERFORMANCE EVALUATION ON ISL DATA

Testing is performed in real time using the chatbot. Given that the chatbot is designed for answering FAQs, deaf users can present their inquiries about eRaktkosh in ISL. Additionally, sample ISL FAQs have been crafted specifically for testing purposes. To evaluate the recognition rates of the four models, we generated two sets of ISL gloss sequences to query the chatbot with:

- Test set 1 - all glosses are within the vocabulary
- Test set 2 - out-of-vocabulary (OOV) glosses also included

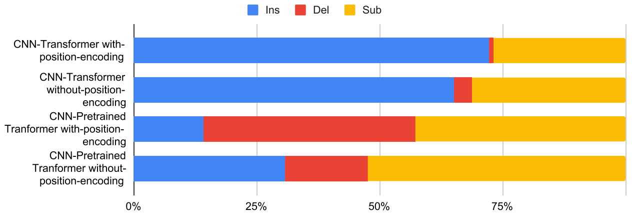


FIGURE 9. Insertion/Deletion/Substitution error rates comparison of all the models trained on ISL dataset, with challenging test glosses in real-time via chatbot.

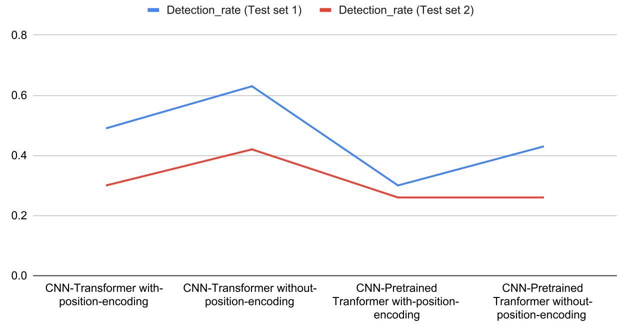
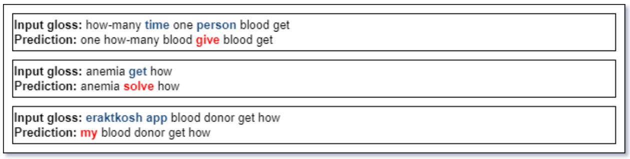


FIGURE 10. Comparison of detection rates of the keywords, for all the models trained on ISL dataset, with challenging test glosses in real-time via chatbot.

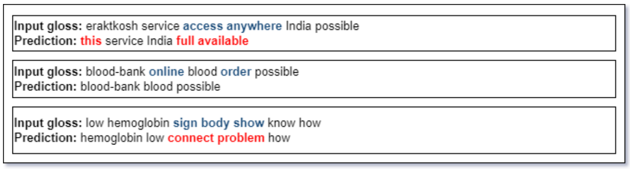


Input gloss: how-many time one person blood get
Prediction: one how-many blood give blood get

Input gloss: anemia get how
Prediction: anemia solve how

Input gloss: eraktkosh app blood donor get how
Prediction: my blood donor get how

FIGURE 11. Examples CNN-Transformer predictions from Test set 1. Color change in input gloss indicates words that were not predicted. In the predicted gloss sequence, red indicates insertion errors.



Input gloss: eraktkosh service access anywhere India possible
Prediction: this service India full available

Input gloss: blood-bank online blood order possible
Prediction: blood-bank blood possible

Input gloss: low hemoglobin sign body show know how
Prediction: hemoglobin low connect problem how

FIGURE 12. Examples CNN-Transformer predictions from Test set 2. Color change in input gloss indicates words that are OOV. In the predicted gloss sequence, red indicates insertion errors.

Figure 8 presents the WERs for all the trained models on the two test sets. The models with pre trained Transformers exhibit minimal overall WER. Detailed insertion, substitution, and deletion error rates are provided in Figure 9. Notably, the deletion rates significantly increase for CNN-pre-trained Transformer models, despite their comparatively lower WERs. However, the combined reduction in insertion and deletion errors contributes to the pre-trained models' ability to maintain a minimal overall WER.

The detection rate is a critical metric in our case for selecting the benchmark model to integrate into the chatbot. The graph illustrating the variation in the detection rate is plotted in Figure 10. Across both test sets, the CNN-Transformer without position encoding demonstrates outstanding performance in comparison to the other models. As a result, we have integrated this model into our chatbot, as recognizing the maximum number of words in the gestured glosses is essential for the post-processing task that aims to retrieve answers based on the detected glosses.

B. ABLATION STUDIES

1) POSITION ENCODING OF TRANSFORMERS

The experimental results for ablating the position encoding of the Transformer and the pretraining applied to the Transformer are provided in Table 4. The reported error

TABLE 4. Experimental results for ablating position encoding and Transformer pretraining.

Pretraining	Position Encoding	WER
No	No	27
Yes	No	21
No	Yes	43
Yes	Yes	19

rates demonstrate that pretraining has greatly contributed to a significant enhancement of the recognition system. However, the WERs resulting from the ablation of the position encoding of the Transformer indicate that its benefits are noticeable only when applied after pretraining. Otherwise, a model without position encoding seems to be more suitable for our tasks.

2) BACKGROUND COMPLEXITY

We conducted an initial ablation study focused on background complexity. The same test glosses were shared with different signers, who were asked to gesture in both plain and cluttered backgrounds. A detailed analysis of the predictions revealed that the model's performance was independent of the background, demonstrating that the pretraining of the CNN and Transformer components enabled the final model to accurately identify salient regions in the video. Despite the model's strong performance in visually complex backgrounds, ensuring accurate recognition in real-world environments remains a priority. To address potential challenges posed by the presence of multiple individuals in the frame, our system includes an initial check to detect if more than one person is present. If multiple persons are identified, a warning message is triggered, allowing the user to adjust the environment for optimal recognition accuracy. This step enhances the model's robustness in diverse and dynamic settings.

3) SIGNING SPEED

This ablation study underscores the importance of handling variability in signing speed, especially in real-time applications. To study the effect of signing speed on the model's recognition rate, the same test glosses were given to different signers, who were asked to perform gestures at slow, medium, and fast paces. The model's predictions revealed that slow and medium signing produced reliable results, while the error rate increased significantly with fast signing. This is primarily due to an increase in deletion errors, as rapid signing causes short-duration signs to be heavily influenced by the coarticulation effect, making them more likely to be missed or misrecognized.

4) HAND DOMINANCE

We conducted an ablation study to assess the impact of hand dominance on the model's recognition performance. To test this, both left-dominant and right-dominant signers were asked to perform the same set of glosses. The analysis revealed that the model's predictions were consistent, regardless of whether the signer used their right or left hand. This outcome is attributed to the horizontal flip augmentation applied during training, which effectively trained the model to recognize signs independently of hand dominance. By introducing this augmentation, the model learned to focus on the spatial and temporal characteristics of the gestures, rather than the specific hand used, making it robust to variations in hand preference across different signers.

5) CAPTURING MODE

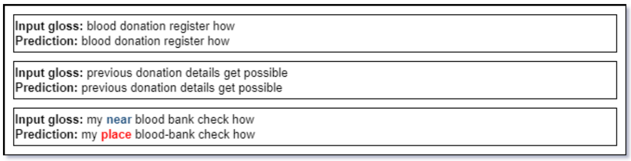
We conducted an ablation study to assess the impact of capturing mode—portrait versus landscape—on the model's recognition performance. The same test glosses were recorded in both portrait and landscape orientations. Our analysis revealed that landscape mode consistently provided more reliable results, while portrait mode often led to issues with hand visibility. In portrait recordings, the signer's hands frequently moved out of the frame, making the gesturing trajectory incomplete and difficult for the model to track. As a result, portrait mode is not recommended for real-world applications where full hand and upper body visibility is critical for accurate sign recognition. Therefore, we recommend using landscape mode for capturing, as it ensures that the full gesturing trajectory is captured and processed effectively.

6) COMPARISON WITH THE BENCHMARK DATASET

Contrasted with the German dataset, the ISL continuous dataset presents greater challenges due to its complexity in terms of varying backgrounds, diverse camera angles, and the attire worn by signers. Furthermore, the primary objective of the ISL dataset is to formulate a model for real-time recognition. As a result, the testing scenarios encompass novel signers, glosses that were not part of the training set, and the potential inclusion of out-of-vocabulary (OOV) elements. This approach aims to simulate real-world conditions, rendering it considerably more demanding compared to the German dataset, which was recorded within a controlled studio environment against a plain backdrop.

Nevertheless, as part of our standard procedure, we have also presented the results of the CNN-Transformer architecture using the German dataset, Phoenix-Weather-2014T corpus. These findings were previously published in our work [21], which primarily delves into the incorporation of position encoding within Transformers. The significance of position encoding in Transformers cannot be understated, as even our most successful model (with a lower WER) for the ISL dataset utilizes position encoding. However, when considering the practical application of our AI model in a chatbot scenario, we have opted for the model that exhibits the highest detection rate.

Unlike other works in this area, this is the first attempt towards real-time recognition. Our previously published work [21] emphasized the significance of position encoding in Transformers. Conversely, this study has illuminated that when the training data exhibits constrained context and the necessity leans towards heightened word-level recognition, as opposed to the conventional context assimilation seen in tasks like Neural Machine Translation, the positioning mechanism of Transformers diminishes in relevance. We reached this conclusion through a series of thorough experiments and rigorous testing, culminating in the development of a model tailored to our specific requirements. Our best-case test set consists of test glosses with words having a detection rate above 50%. Even though our system is trained with 13-signers, testing done with new signers using trained gloss sequences using shows 100% recognition accuracy and around 95% accuracy using the best case test set (no OOV). A few examples of this bestcase scenario are shown in Figure 13. We are thus successful in developing a signer-independent system. Both left-hand and right-hand dominant signers were treated equally due to the application of flip augmentation during the training process.


Input gloss: blood donation register how
Prediction: blood donation register how

Input gloss: previous donation details get possible
Prediction: previous donation details get possible

Input gloss: my near blood bank check how
Prediction: my place blood-bank check how

FIGURE 13. **Bestcase scenario: Test gloss sequences with all words in-vocab and having detection rate above 50%. Words in red indicate insertion errors.**

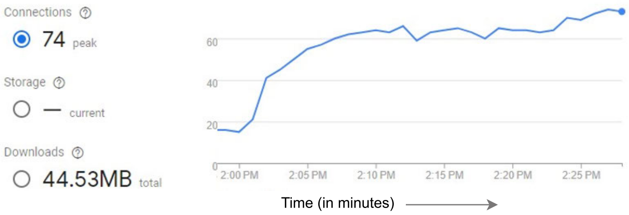


FIGURE 14. **Load testing.**

The model consistently displayed outstanding performance across all these scenarios. To assess the model's performance across various devices and networks simultaneously, we executed a load-testing session within a 30-minute timeframe. This load was managed by a single A6000 GPU card equipped with 40GB of memory. The corresponding graph is presented in Figure 14, where the peak signifies the presence of 74 concurrent users.

VIII. DISCUSSION

This study presents SignFlow, a novel framework designed for real-time recognition of continuous Indian Sign Language (ISL) gestures. The findings demonstrate the efficacy of leveraging domain-specific pretraining and hybrid architectures to address the unique challenges of ISL recognition. By pretraining the CNN on isolated ISL videos and the Transformer on Mediapipe pose estimates, the model effectively captured the nuances of hand shapes and body movements specific to ISL. This domain-specific approach significantly enhanced recognition accuracy while reducing reliance on general-purpose datasets.

A key innovation of this study was the dynamic downsampling technique, which ensured compatibility across devices with varying frame rates. This step proved crucial for real-time applications, as it maintained recognition accuracy while optimizing computational efficiency. Extensive testing revealed that downsampling videos to 6 fps provided an optimal balance between accuracy and processing speed, making it a practical choice for real-time deployment.

The hybrid architecture combining a 3DResNet CNN and an 8-layer Transformer Encoder effectively captured both spatial and temporal features. Notably, the Transformer performed well even without positional encoding, highlighting its adaptability for constrained contexts such as continuous ISL recognition. The model's performance, reflected in a Word Error Rate (WER) of 19 on the Continuous ISL dataset, underscores its competitive edge in real-time gesture recognition. The introduction of a detection rate metric further enriched the evaluation process, emphasizing the importance of recognizing individual words within sequences, irrespective of their order.

Experimental analyses provided valuable insights into the model's robustness and limitations. Ablation studies confirmed that pretraining played a critical role in enhancing generalization, enabling the model to handle diverse signing styles and environmental conditions. While positional encoding improved recognition in certain scenarios, its absence did not hinder performance for constrained contexts, indicating the architecture's flexibility. The model exhibited resilience to variations in background complexity, signing speed, and hand dominance, thanks to robust data augmentation techniques. However, challenges remain with rapid signing, which led to increased deletion errors, and with out-of-vocabulary (OOV) glosses, which affected recognition accuracy. These findings underscore the importance of expanding the dataset to include a broader vocabulary and more diverse signer profiles.

In real-world applications, the system demonstrated strong performance, making it suitable for deployment in tools like the eRaktkosh chatbot. Despite its successes, there are areas for further improvement. Incorporating additional training samples from fast signers could mitigate errors caused by rapid signing. Expanding the dataset and exploring advanced sequence-to-sequence techniques, such as attention-based models, may further enhance alignment and recognition accuracy.

In conclusion, SignFlow represents a significant advancement in real-time ISL recognition, offering a scalable and efficient solution for bridging communication gaps in the deaf community. The insights gained from this study pave the way for future developments in continuous sign language recognition systems, aligning with the need to enhance accessibility and inclusivity [30]

IX. LIMITATIONS

After conducting several rounds of pilot testing with deaf users nationwide, we have identified a few limitations of our chatbot. Although the model is capable of recognizing new test sequences, if one or more gestures are OOV, accurate recognition becomes difficult. Additionally, the presence of OOV gestures, along with in-vocabulary words, affects the accurate detection of words within the vocabulary. In specific instances, signing speed plays a significant role. For example, some deaf individuals sign at a rapid pace, and in such scenarios, our model tends to generate inaccurate predictions. However, addressing this issue could involve incorporating additional training samples from individuals who sign quickly and subsequently updating the model. This approach may offer a solution to the problem in the future.

X. FUTURE WORK

In the context of real-time application, we did not select the model solely based on achieving the lowest WER, as elaborated upon earlier. The pivotal criterion centers around the detection rate of keywords, which subsequently facilitates the retrieval of answers within the chatbot, following the model's predictions. This aspect provides an avenue for ongoing model enhancement, focusing on improving the detection rate of keywords.

ACKNOWLEDGMENT

This project is inspired and guided by Amrita University's Chancellor Amma. Special thanks to their collaborator, the Centre for Development of Advanced Computing (CDAC), for their valuable contributions to the research especially in the dataset collection. Lastly, the authors are grateful to all those who have contributed to the success of this project in various ways.

REFERENCES

[1] A. Othman, "Sign language varieties around the world," in *Sign Language Processing: From Gesture to Meaning*. Cham, Switzerland: Springer, 2024, pp. 41–56.

[2] S. Renjith and R. Manazhy, "Sign language : A systematic review on classification and recognition," *Multimedia Tools Appl.*, vol. 83, no. 31, pp. 77077–77127, Feb. 2024.

[3] N. Aloysius and M. Geetha, "Understanding vision-based continuous sign language recognition," *Multimedia Tools Appl.*, vol. 79, nos. 31–32, pp. 22177–22209, Aug. 2020.

[4] N. Aloysius and M. Geetha, "A review on deep convolutional neural networks," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Apr. 2017, pp. 588–592.

[5] Q. Zhu, J. Li, F. Yuan, and Q. Gan, "Multiscale temporal network for continuous sign language recognition," *J. Electron. Imag.*, vol. 33, no. 2, Apr. 2024, Art. no. 023059.

[6] L. Hu, L. Gao, Z. Liu, and W. Feng, "Scalable frame resolution for efficient continuous sign language recognition," *Pattern Recognit.*, vol. 145, Jan. 2024, Art. no. 109903.

[7] R. Zuo and B. Mak, "Improving continuous sign language recognition with consistency constraints and signer removal," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 20, no. 6, pp. 1–25, Jun. 2024.

[8] N. Aloysius, M. Geetha, and P. Nedungadi, "Continuous sign language recognition with adapted conformer via unsupervised pretraining," 2024, *arXiv:2405.12018*.

[9] Y. Chen, R. Zuo, F. Wei, Y. Wu, S. Liu, and B. Mak, "Two-stream network for sign language recognition and translation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, Jan. 2022, pp. 17043–17056.

[10] Y. Chen, F. Wei, X. Sun, Z. Wu, and S. Lin, "A simple multi-modality transfer learning baseline for sign language translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5110–5120.

[11] O. Koller, N. C. Camgoz, H. Ney, and R. Bowden, "Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2306–2320, Sep. 2020.

[12] O. Koller, S. Zargaran, and H. Ney, "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3416–3424.

[13] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial–temporal multi-cue network for continuous sign language recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, Apr. 2020, pp. 13009–13016.

[14] R. Zuo and B. Mak, "C2SLR: Consistency-enhanced continuous sign language recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5121–5130.

[15] J. Zheng, Y. Wang, C. Tan, S. Li, G. Wang, J. Xia, Y. Chen, and S. Z. Li, "CVT-SLR: Contrastive visual-textual transformation for sign language recognition with variational alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23141–23150.

[16] A. Hao, Y. Min, and X. Chen, "Self-mutual distillation learning for continuous sign language recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11283–11292.

[17] R. Cui, H. Liu, and C. Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1610–1618.

[18] R. Cui, H. Liu, and C. Zhang, "A deep neural framework for continuous sign language recognition by iterative training," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1880–1891, Jul. 2019.

[19] J. Pu, W. Zhou, and H. Li, "Iterative alignment network for continuous sign language recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4160–4169.

[20] Y. Min, A. Hao, X. Chai, and X. Chen, "Visual alignment constraint for continuous sign language recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11522–11531.

[21] N. Aloysius, M. Geetha, and P. Nedungadi, "Incorporating relative position information in transformer-based sign language recognition and translation," *IEEE Access*, vol. 9, pp. 145929–145942, 2021.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[23] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Comput. Vis. Image Understand.*, vol. 141, pp. 108–125, Dec. 2015.

[24] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7784–7793.

[25] A. Graves, "Sequence transduction with recurrent neural networks," in *Proc. Workshop Represent. Learn.*, Nov. 2012, pp. 1–9.

[26] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Audio chord recognition with recurrent neural networks," in *Proc. ISMIR*, Jan. 2013, pp. 335–340.

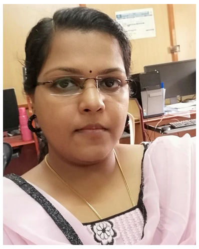
[27] Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*.

[28] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Proc. NIPS Workshop*, 2017, pp. 1–12.

[29] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in *Proc. 12th Symp. Operating Syst. Design Implement.*, Nov. 2016, pp. 265–283.

[30] R. Raman, D. Pattnaik, L. Hughes, and P. Nedungadi, "Unveiling the dynamics of AI applications: A review of reviews using scientometrics and BERTopic modeling," *J. Innov. Knowl.*, vol. 9, no. 3, Jul. 2024, Art. no. 100517.

M. GEETHA received the Ph.D. degree from Amrita Vishwa Vidyapeetham, Amritapuri. She has been involved in teaching and research, since 2003. She is currently the Chairperson and an Associate Professor with the Department of Computer Science Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham. Her research interest includes the area of video analytics, machine learning, deep learning for edge devices, and computer vision. She has funded projects and patents in the area of video analytics. Her research interests include computer vision, machine learning, deep learning, and pattern recognition.



NEENA ALOYSIUS received the Ph.D. degree in computer science and engineering from Amrita Vishwa Vidyapeetham, India. She was a Research Associate with the Amrita Center for Research in Analytics, Technologies, and Education (AmritaCREATE) for the past four years. Previously, she gained more than five years of IT experience as a Systems Engineer with Infosys Technologies Ltd. Her research interests include deep learning, computer vision, and pattern recognition.

DARSHIK A. SOMASUNDARAN received the bachelor's degree in computer science and engineering from Amrita Vishwa Vidyapeetham, India. He received the Ontario College Graduate Certificates in business analytics and applied artificial intelligence and machine learning from Conestoga College, Canada. He is currently a Machine Learning Engineer with a strong foundation in AI and data-driven solutions. He has contributed to projects in deep learning optimization, sign language recognition for accessibility, and predictive analytics in retail and telecom. His research interests include python programming, computer vision, and generative AI, with a focus on solving real-world challenges and improving operational efficiency.

AMRITHA RAGHUNATH received the B.Tech. degree in computer science and engineering and the M.Tech. degree in artificial intelligence from Amrita Vishwa Vidyapeetham, in 2018 and 2021, respectively. She was a Machine Learning Engineer with the Amrita Center for Research in Analytics, Technologies and Education (AmritaCREATE), Amritapuri Campus. She was also a Student with the University of Paderborn briefly for an exchange semester.

PREMA NEDUNGADI received the Ph.D. degree in computer science and engineering from Amrita Vishwa Vidyapeetham, India. She is currently the Director of the Amrita Center for Research in Analytics, Technologies, and Education (AmritaCREATE), Amrita University, and a Professor with the Amrita School of Computing, Amrita Vishwa Vidyapeetham, India. She was a recipient of the Digital India Award from the Ministry of Electronics and Information Technology, India, in the category of digital empowerment. She was a Finalist in U.S. $7 million Barbara Bush Foundation Adult Literacy XPRIZE Competition.