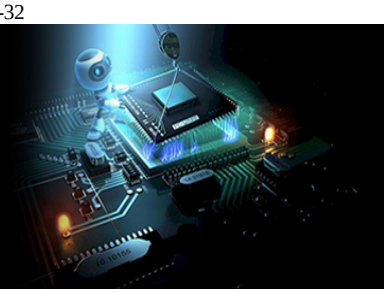
International Journal of Engineering in Computer Science



E-ISSN: 2663-3590
P-ISSN: 2663-3582
Impact Factor (RJIF): 5.52
www.computersciencejournals.com/ijecs
IJECS 2026; 8(2): 24-32
Received: 10-02-2026
Accepted: 12-03-2026

P Rajesh Assistant Professor,
Department of Computer Science and Engineering, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India

M Karthik
Department of Computer Science and Engineering, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India

P Deepthi Sai Sree
Department of Computer Science and Engineering, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India

M Praveen
Department of Computer Science and Engineering, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India

M Geethika Navya
Department of Computer Science and Engineering, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India

P Harshitha
Department of Computer Science and Engineering, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India

Corresponding Author:
P. Rajesh Assistant Professor,
Department of Computer Science and Engineering, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India

Enhancing sign language recognition with mediapipe-based hand detection

P Rajesh, M Karthik, P Deepthi Sai Sree, M Praveen, M Geethika Navya and P Harshitha

DOI: https://www.doi.org/10.33545/26633582.2026.v8.i2a.273

Abstract
The recognition of sign language among hitherto unheard-off people is also an issue that is not so easily tackled since the styles of signing, patterns of motions and the forms of hands are varied among different users. The performance of many current recognition systems has only been shown to be high when tested on signers that were part of the training set thus limiting the applicability of these systems to practice. The current paper puts forward a signer-neutral recognition system that can be assessed with the help of the AUTSL dataset comprising 226 isolated sign types delivered by several subjects. The system does not handle raw video frame, but instead, it uses gestures in sequences of hand and upper-body landmarks that are extracted on a frame-by-frame basis. Such landmark sequences are done using a normalization that minimizes appearance-dependent variability and modeled with a Transformer encoder with the capacity to reflect long range temporal relations. An attention-based temporal aggregation mechanism is employed to emphasize informative motion segments while suppressing redundant frames. Tests with a rigorous signer-disjoin evaluation program show validation and test accuracies of 85.45 and 83.52 suggesting high test precision in the case of signers who are not seen.

Keywords: Landmark-based recognition, sign language recognition, signer-independent learning, temporal sequence modeling

Introduction
Sign language recognition is a research field that has gained significance to the enhancement of communication technologies to individuals with hearing disabilities [1]. The latest technology in computer vision and deep learning has also allowed smart devices to process visual data and comprehend simple hand gestures and body language [2]. Nevertheless, it is still a significant challenge to develop systems that can generalize to users who have never been encountered before. Differences in the style of signing, the speed of movement and the sturdy form within the body bring whole models that have been trained on a limited sample of signers to fail when transferred to new signers and result in overoptimistic outcomes on signer-dependent tests. Signer-independent recognition methods that use motion information as opposed to visual appearance have been explored in recent research to enhance generalization. Here, hands and upper body skeletal landmarks can offer a concise and compact description of gesture dynamics, and lower sensitivity to background and lighting changes [4], [5]. But even though the concept of modeling the temporal laws on how landmark sequences interrelate is not easy, sign gestures can develop with time. This study contributes in the following ways. First, signer-independent recognition framework is built based on the normalised skeletal landmarks retrieved by MediaPipe Holistic with minimal variations of appearance between different backgrounds and differences in illuminations and signer peculiarities. Second, a Transformer-based temporal modeling approach is used to learn long-range correlations in landmark chains and model the intricate gesture dynamics [8]. Third, an attention-based temporal aggregation model is proposed which highlights informative frames and simplifies redundant motion segments [12]. Lastly, the proposed framework is tested on the AUTSL dataset in a strict signer-disjoint protocol to prove its suitability in large-vocabulary signer-independent recognition. Despite the fact that the separate entities involved in the structure, including landmark extraction and Transformer-based modeling, have been studied in prior research, their combination in a motion-oriented

Signer-independent pipeline has a number of positive issues. The proposed method is less sensitive to variations in visual appearance and background by dividing gestures with normal skeletal landmarks, rather than raw RGB frames. With Transformer-based temporal modelling and sequence aggregation with attention, this architecture will help the system to simultaneously encode fine-grained finger movements and long-range movement correlations, and achieve an effective and scalable large-vocabulary sign language recognition system.

Related work

The advancement of sign language recognition (SLR) studies has come to a large extent due to advances in data driven learning models and the increased access to annotated sign language datasets. Initial methods like deep learning usually directly fed raw RGB video sequences to the convolutional neural network with the aim of identifying spatial and motion-related cues directly out of the pixel values [1], [2]. These models had good results albeit with some promise. They were the results of regulated experiments prone to the change of recording conditions as including evolution in the appearance as it relied on the appearance. Other background difference, light, and have features that are unique variating performance was often the result of signers, and it incapacitated their utility in the field deployment environment, because elucidated upon by R. Rastgoo [3].

The later studies resorted to alleviating these limitations by adopting motion-based representations which separate the gesture visualization and visual dynamics. Pose based methods represent the signs as stereotypes of the joint position by a sequence of body posture, hand motions paths and skeletal keypoints. These kinds of representations enable models to be much more concerned with motion dynamics than those based solely on pixel-based visual cues as described in the previous works by Liu [4] and Huang [5] Large-scale datasets like AUTSL have been introduced, and made it possible to evaluate these representations in a systematic manner across a large variety of sign types and signers [6]. Even though landmark-based inputs enhanced the resistance to variability of appearances, subsequent research revealed that representation alone fails to address the generalization problem fully, with the anatomy and movement performance of different people remains posing a great deal of variance, according to Bragg [7].

Consequently, signer independent recognition is the corner stone in modern SLR studies. Stern separation of signers in experimental regimes of training and testing is a consistently recurring issue distinguishing significant performance loss when models are used to examine unknown persons and the hardship of cross-signer generalization has been demonstrated by N. C. Camgöz [8]. To mitigate the situation, several normalization methods have been given such as coordinate statistics, scale statistics, and motion normalizations that both aim at the reduction of identity biases and the conservation of fundamental structure of signing motion as suggested by Cheng [9].

A good temporal modeling is also another extremely important element of SLR systems since signs are not specific fixed arrangements but the sequence of movement. L. Some of the most commonly used recurrent neural networks, including LSTM and GRU, have been applied to process landmark sequences and learn temporal evolution L. Shi [10].

Recategorically of these shortcomings, more recent studies have been adopting Transformer-based designs wherein the temporal association between a complete sequences is inferred jointly by the use of self-attention, as proposed by H. Zhou [11]. This ability to think globally allows the establishment of stronger temporal reasoning and has proven to be less sensitive to noise than recurrent models in many attributes of sign language recognition as well as sign language translation in applications as shown by Cheng [12]. Other studies have also provided the methodological framework to follow the research, signer-independent visual gesture recognition, including preprocessing- algorithms, visual features representation, pattern-matching algorithms, and assistive computer vision systems[13][14] studies which has been reported previously in the same studies as far back as the classical image retrieval methods [15], and feature matching methods [16].Recently, more studies have expanded on the topic of the advanced spatial-temporal learning strategies that could enhance the gesture representation and recognition performance. As an example, multi-scale spatial-temporal feature modeling has been suggested so as to have more ability to capture complex motion dynamics in sign language sequences [17]. On the same note, the graph-based pose modelling methods have also been presented to display spatial associations among body joints and enhance recognition accuracy in skeleton-based gesture recognition tasks [18]. The developments underscore the increasing relevance of integrating designed skeletal representations and attention-based temporal modeling schemes. Based on these developments, the current paper combines normalized landmark representations with a Transformer-based system to obtain a higher signer-independent recognition rate.

Dataset overview

The given signer-independent sign language recognition framework is estimated on the basis of the popular Autism Sign Language data set, currently used to benchmark large-vocabulary independent isolated sign recognition under cross-signer framework. The data has video samples of 226 separate sign categories within which the participants performed, which gives a lot of fluctuation in the gesture, and the signer attributes. These samples are acted by 42 signers, which adds significant diversity to the signing style, movement, body shape, and conditions of recordings. The sample size is moderate with an equal representation of the different classes, with each sign group having about 70 to more than 200 samples. Each of the classes has approximately 150 samples on average, which seems large enough to offer sufficient representation without discarding natural variability by category of gesture. In particular video samples are divided into mutually exclusive training, validation and test subsets so that no one signer is represented in more than a subset. The dataset is separated into training, validation and testing sets following a signer-independent protocol. There are 26,154 samples used in training the model; the sample consists of 30 signers. The validation set comprises of 4,418 samples by 6 other signers and serves to

tune the hyperparameter and to provide intermediate performance results. The other 3,742 samples that were given by a different group of 6 signers constitute the test set that is used to test out the final model construction.

Methodology

The sign language recognition model with landmark preprocessing, time modelling structure and training plan, and a sign language signer-independent framework. The design aims at learning signer independent motion representations and reduces appearance sensitivity and identity specific sensitivity. Fig. 1 represents the general design of the proposed signer-independent system.

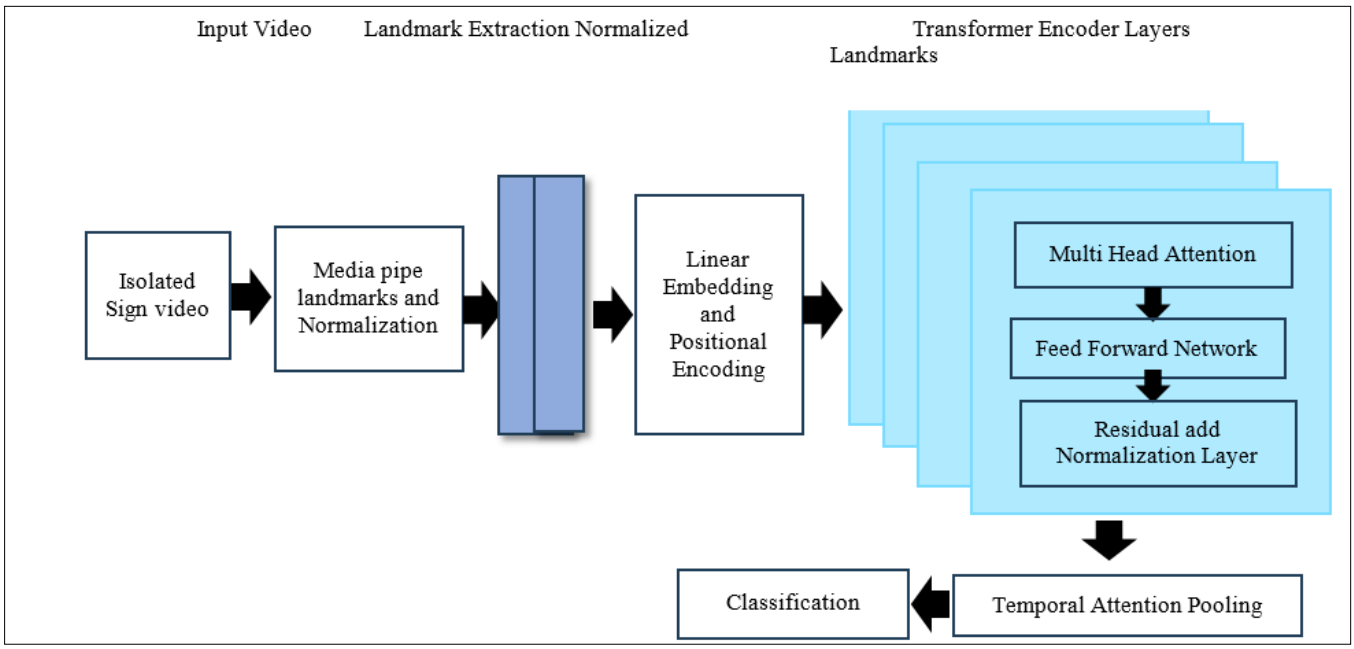


Fig. 1: Architecture of proposed signer-independent Transformer model for sign language recognition

Landmark extraction and temporal alignment

In this work, each input sign video is first converted into a sequence of skeletal landmarks using the MediaPipe Holistic framework. In doing so, this strategy helps in ensuring that the system does detect small-scale finger gestures among others and general. Motions of the upper-body, necessary to sign-accuracy. Recognition. There are three sets of keypoints in every frame. Extracted: left hand landmarks right hand landmarks and upper-body pose joints. The hand has 21 keypoints each. When 12 joints of the upper body (shoulders, elbow, wrists, and torso) give contextual motion data. Since each key point is represented using three-dimensional coordinates(x, y, z), a single frame is encoded as a 162-dimensional feature vector $((21 + 21 + 12) * 3))$as in (1):

$$x_t \in \mathbb{R}^{162} \quad (1)$$

Where $x_t$represents the feature vector at time step t.
Thus, a sign video can be represented as a temporal sequence as in (2):

$$X = \{x_1, x_2, \dots, x_{T_{\text{orig}}}\} \quad (2)$$

Where T_{orig} denotes the original sequence length.
The model can distinguish merely similar signs using local finger articulation and global arm movement, thus effectively capturing these movements. Since sign videos have inherent variable length, all sequences are uniformly scaled in time to a constant length of T=48frames as in (3) and (4):

$$t_i = \left\lfloor \frac{i \cdot (T_{\text{orig}} - 1)}{T - 1} \right\rfloor, \quad i = 0,1, \dots, T - 1 \quad (3)$$

$$x_i' = x_{t_i} \quad (4)$$

Where t_i represents the sampled frame index.
This keeps the size of input constant as well as maintaining the total size dynamics of the gesture movement. Sign videos differ as a result of the dissimilarity length of hand gesture varies according to the complexity of gesture and the treason would be the sign speed of the performer which is the landmark. Sequences are of various lengths. In order to give a regular contribution. Forming of the temporal model, the sequences are all turned into a definite duration of 48 frames by the even temporal sampling. On this process, the chosen frames are evenly chosen interspersed intervals between the sequences, which enables the main Strategy of gesture movement to be kept during making the time dimension normal. A review has been made on the most isolated signs, as indicated by AUTSL dataset, are usually across the board roughly 30 to 70 frames. The length of a sequence was chosen as 48 frames thus represents a sensible tradeoff between memorizing valuable motion moves and preventing

unwarranted frames that raise the cost of computing. In addition, there is temporal augmentation that is made when removal on a randomly chosen number of frames in training before resampling. This promotes learning amongst the model motion patterns, which do not change despite gesture length of time or rate of signing is different in individuals.

Sample products of the process of landmark extraction using MediaPipe are shown in Fig. 2 in which hand keypoints have been detected and the joints of the upper bodies are superimposed on the input frames. This illustration shows the functioning of both articulate finger capturing movements and global body posture is done. Subsequent processing.

Fig. 2: Sample visualization of MediaPipe-based landmark extraction showing hand keypoints and upper-body pose joints used for feature representation.

Spatial normalization and sequence standardization

In order to achieve the model that is robust to various individuals, structured normalization pipeline is used to minimize the variations due to body size, pose, and recording conditions. To ensure that the model is robust to various people, a higher-level pipeline of normalization becomes part of the structuring to minimize changes due to body sizes, pose, and recording parameters as in (5):

$$c_t = \frac{P_{\text{shoulder left}} + P_{\text{shoulder right}}}{2}, \quad x_t^{(1)} = x_t - c_t \ (5)$$

Where c_t denotes the center point between shoulders.
Next, Scale normalization is then carried out with the help of the average. Distance between the shoulders as in (6):

$$d_s = \frac{1}{T}\sum_{t=1}^{T} |P_{\text{shoulder left}} - P_{\text{shoulder right}}|, \quad x_t^{(2)} = \frac{x_t^{(1)}}{d_s} \ (6)$$

Where d_s represents the scale normalization factor.
In order to further decrease structural differences between signers, a further normalization stage which depends on average bone. Magnitude is applied as in (7):

$$d_b = \frac{1}{T}\sum_{t=1}^{T} |x_t^{(2)}|, \quad x_t^{(3)} = \frac{x_t^{(2)}}{d_b} \ (7)$$

Where d_b denotes the bone normalization factor
Finally, each sequence is standardized as in (8):

$$\widehat{x_t} = \frac{x_t^{(3)} - \mu}{\sigma} \ (8)$$

Where μ and σ denote mean and standard deviation respectively.
This chain of conversions is such that the representation is also made translation, scale-invariant, etc. individual body proportions. As a result, the model is promised to emphasize the patterns of motion and not signer. Specific characteristics.

Temporal augmentation

To consider the differences in speed of signing and performance. Style, temporal augmentation is created during training. In this, a random sample of frames is deleted, and the sequence is then resampled to its constant length as in (9):

$$\bar{X} = \text{Resample}\big(\text{DropFrames}(X')\big) \ (9)$$

Transformer-based temporal feature modeling

A normalized landmark sequence is thereafter fed through the Architecture to model temporal transformer-based dependencies. This is projected into a higher frame before each frame. Dimensional embedding space as in (10):

$$z_t = W_e \overline{x_t} + b_e, \quad z_t \in \mathbb{R}^{256} \ (10)$$

Where W_e and b_e are learnable parameters.
In order to store information about the order over time, sinusoidal positional. Encoding is added as in (11):

$$\widetilde{z_t} = z_t + PE(t) \ (11)$$

The resulting sequence is processed using stacked Transformer encoder layers. Each layer employs multi-head

self-attention as in (12) and (13):

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (12)$$

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\dots,\text{head}_H)W^O \quad (13)$$

In this design, the model is able to win the relationship between every frame at the same time. Unlike sequential models the Transformer has the ability to model long sequences such as LSTMs range dependencies and hence especially effective with replicating the movement of complex gestures. The temporal dynamics of the landmark sequences are modeled using a Transformer-based encoder network. This architecture enables the system to capture relationships between frames across the entire gesture sequence. At each time step, the 162-dimensional landmark vector is projected into a 256-dimensional embedding space using a linear embedding layer. To preserve the temporal structure of the input sequence, sinusoidal positional encodings are incorporated into the feature embeddings, enabling the model to differentiate gestures occurring at different time steps. The resulting position-aware representations are processed by a Transformer encoder consisting of four stacked layers. Each encoder layer applies multi-head self-attention with eight attention heads to learn relationships between frames across the entire sequence. The attention mechanism is followed by a feed-forward network with 384 hidden units that further refines the temporal features. To enhance generalization, a dropout rate of 0.35 is applied within the encoder layers. By jointly analyzing all frames in the sequence, the Transformer architecture effectively captures long-range temporal dependencies between hand gestures and upper-body movements.

Attention-based temporal pooling

To convert frame-level features into a compact sequence representation, an attention-based temporal pooling mechanism is employed. Let $h^t \in R^d$ denote the feature embedding of the t^{th} frame produced by the Transformer encoder, where T represents the sequence length. A learnable attention layer computes an importance score for each frame as in (14):

$$s_t = W^T h_t \quad (14)$$

Where W is a trainable weight vector. The attention scores are converted into attention weights by applying the SoftMax function as in (15):

$$\mathrm{a_t} = \frac{\exp(\mathrm{s_t})}{\sum_{\mathrm{i=1}}^{\mathrm{T}} \exp(\mathrm{s_i})} \quad (15)$$

The final sequence representation is then computed as a weighted aggregation of frame embeddings as in (16):

$$H = \sum_{t=1}^{T} a_t h_t \quad (16)$$

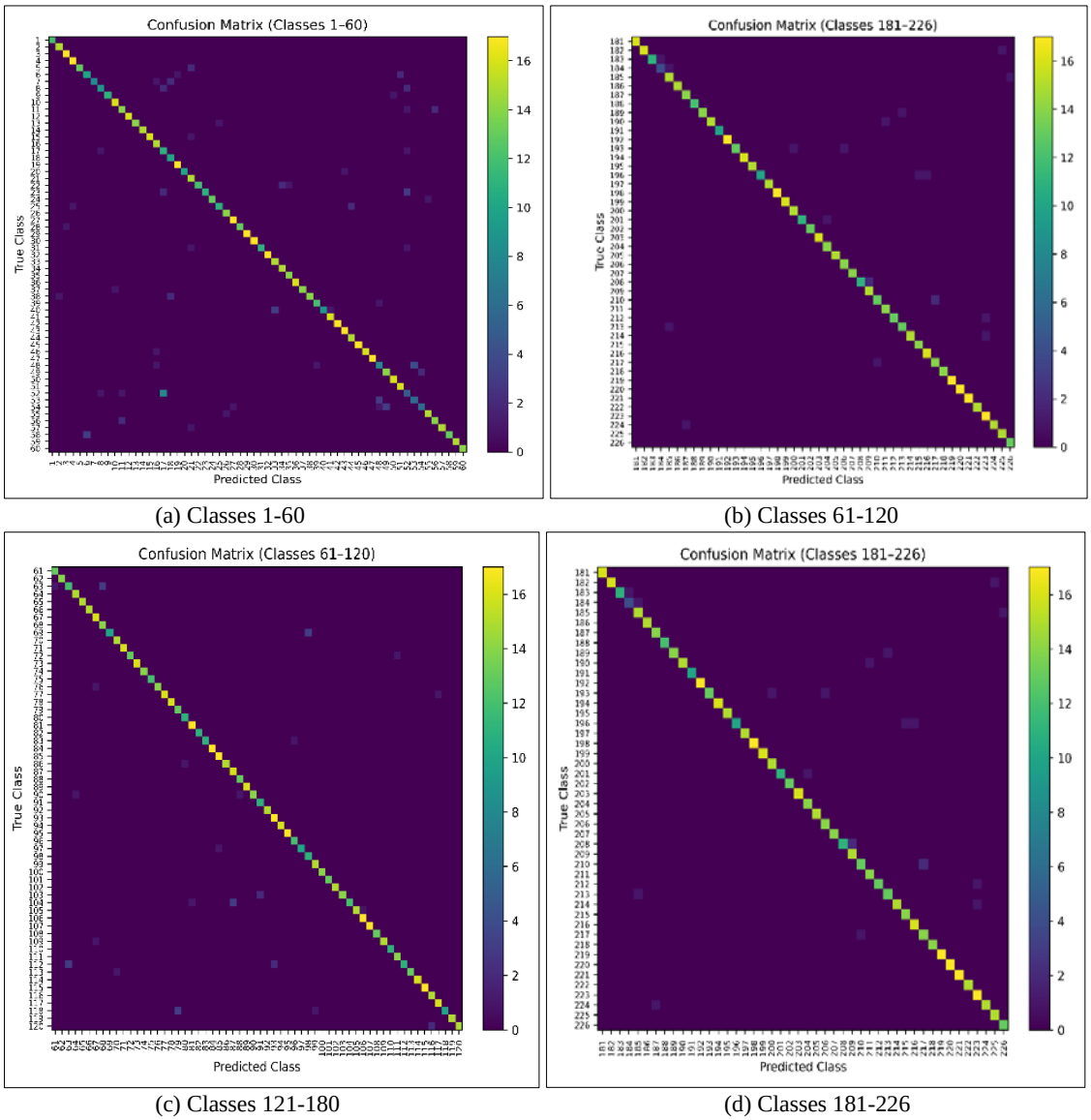
The process of this adaptive pooling strategy enables the model to focus attention upon informative gesture segments and diminish the role of less significant frames and creates a compact model of the sign sequence.

Classification and training strategy

The sequence features are then aggregated, and then regularized with the help of a dropout layer and then ran through a fully-connected classifier, which predicts the corresponding sign label. The model is also resolved through cross-entropy loss by training in an end-to-end fashion with label-smoothing being 0.1 which prevents prediction over-confidence and enhanced generalization. The optimization Problem is solved with AdamW optimizer with an initial learning rate of 0.0002 and a weight decay of 1×10^{-4}. To reduce overfitting, the dropout rate of 0.35 is used in the Transformer encoder and an extra dropout layer with the dropout rate 0.4 is added before the final classification layer. The training is performed in mini-batches of 32 samples and approximately 40 epochs are applied to ensure the stabilization of the optimization process. A cosine annealing learning rate schedule is employed. To further prevent overfitting, validation performance is monitored during training and early stopping is applied when improvement plateaus. All experiments follow a strict signer-independent protocol, ensuring that signers present in the test set are not included in the training data.

Performance evaluation

The proposed signer-independent sign language recognition model based on a Transformer architecture was developed using the PyTorch framework and tested on the AUTSL dataset, which includes 226 isolated sign categories. The model training was done in 40 epochs with AdamW optimization algorithm and initial learning rate was 0.0002 with a cosine annealing schedule to encourage a smooth and steady optimization. A cross-entropy loss with a label smoothing was used to enhance the generalization performance and reduce the overconfidence in predictions. In these training conditions, the model was able to reach a peak validation accuracy of 85.45% which means that the model learned discriminative motion representations and was also resistant to signer variation. In a separate test set that was not observed during training, the trained model was further tested on the generalization ability where it obtained an accuracy of 83.52. The minor variation in validation and test performance can be accounted by the presence of consistent learning behaviour and it implies that overfitting is under appropriate control even when a large sign vocabulary and a significant range between gesture classes are used. Besides the overall accuracy, various multi-class evaluation measures were also calculated to give a more detailed evaluation on the model. The proposed model attained 84.10 precision, 83.52 recall and 83.81 F1-score on the test set, showing consistent classification performance with a large variety of sign categories. In order to further evaluate the performance of classification, 226 x 226 confusion matrix was produced using the test dataset as shown in Fig. 3(a)-(d). The confusion matrix has a high degree of diagonal dominance, which proves that most samples were correctly labeled in all categories of signs. It is worth noting that no zero recall of every 226 classes was obtained confirming that the model can recognize the entire vocabulary without class collapse. The most frequent error in misclassifications is that of similar signs visually or semantically and is a manifestation of intrinsic ambiguity in a few motions of the hand, and not deficiency in the model structure. A further look at the confusion matrix shows that the errors of misinterpretations happen mostly between signs having a similar shape of the hand or movement pattern. As an illustration, gestures with similar patterns of finger articulation or minor variations in hand orientation are more likely to cause a higher rate of confusion. Nonetheless, most of the categories of signs are properly placed, which suggests that the proposed landmark-based framework is effective towards the representation of discriminative gesture behavior.



(a) Classes 1-60

(b) Classes 61-120

(c) Classes 121-180

(d) Classes 181-226

Fig. 3: Block-wise confusion matrices of the proposed model across 226 sign classes

Due to the large number of sign classes (226), the confusion matrix is divided into four sub-figures corresponding to different class ranges: (a) classes 1-60, (b) classes 61-120, (c) classes 121-180, and (d) classes 181-226. The strong diagonal dominance across all sub-matrices indicates that the proposed model correctly classifies most sign categories, while misclassifications mainly occur between visually similar gestures.

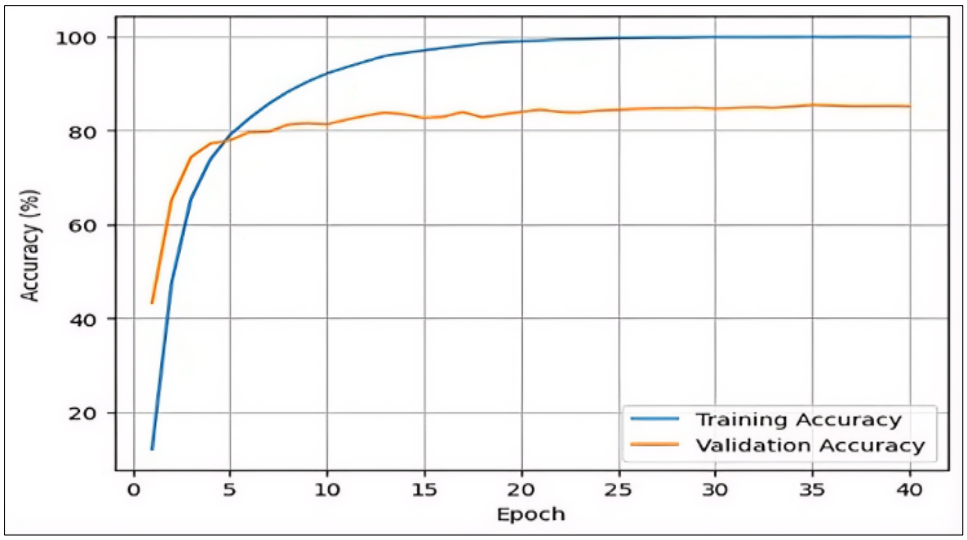


Fig. 4: Training and validation accuracy trends across 40 epochs

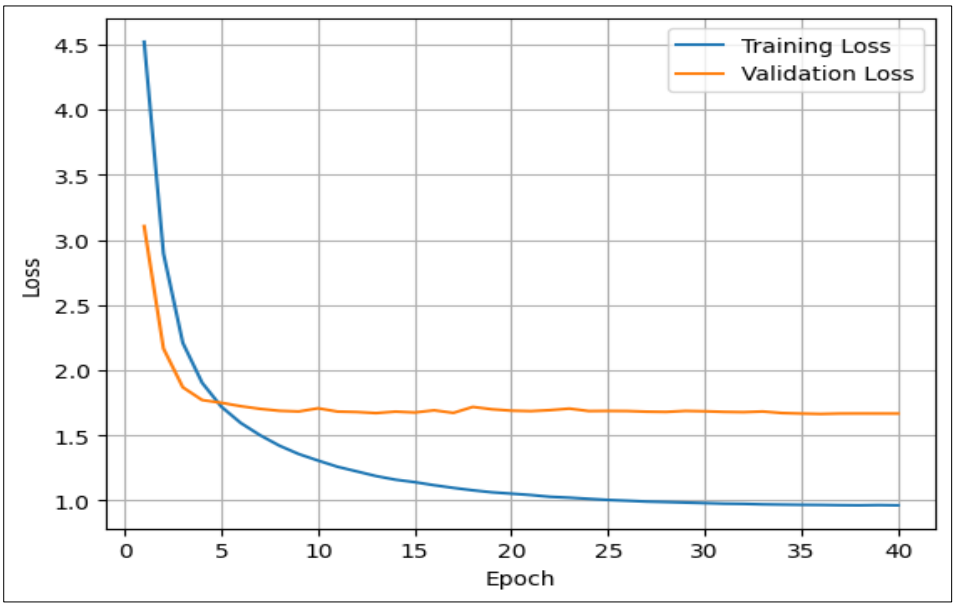


Fig. 5: Training and validation loss trends across 40 epochs

The learning curves shown in Fig. 4 and Fig. 5 indicate stable convergence across the 40 training epochs. During the early stages of training, both training and validation accuracy gradually improve before reaching a stable level, while the corresponding loss values steadily decrease. The relatively small gap between training and validation performance suggests good generalization ability and minimal overfitting, indicating that the applied regularization techniques and learning rate scheduling strategy are effective.

Results

The proposed signer-independent Transformer-based sign language recognizer system was trained and evaluated on the AUTSL-dataset that consists of 226 isolated sign categories coded in skeletal landmark sequences. In order to make the assessment objective and unbiased, the dataset was split into training, validation, and test ones, based on the official AUTSL signer-independent protocol. All the experimental assessments were performed in a PyTorch implementation to achieve hardware acceleration to facilitate the efficient model optimization and consistent training dynamics.

Confusion matrix analysis

The classification performance of the proposed model was examined using a 226 × 226 confusion matrix, as shown in Fig. 3(a)-(d). The model received a maximum validation actual correctness of 85.45% and test accuracy of 83.52, denoting high generalization during unseen signer and sign instances. The confusion matrix has definite diagonal dominance, or displaying that a large fraction of sign samples is properly assigned to most of the classes.

Quantitative analysis

Through its quantitative study, it can be shown that the suggested technique is generalizable to new signers who have never been observed. The low difference between the validation and the test accuracy (85.45 and 83.52 respectively) points to the stable learning pattern and the performance even under the conditions other than the training data. The Transformer-based temporal modeling algorithm, along with the attention process, enables the system to learn long-range motion relationships of the signing sequence along with the use of skeletal landmarks exclusively, which makes the system robust to the differences between the signature styles between people. Combined, with these design choices, these offer a scalable and robust sign language recognition system that attempts large vocabulary (signer-independent).All the findings explained during the experiments allow concluding that the proposed framework is very appropriate to practical SLR applications and has high potential in real-world implementation.

Comparative evaluation

To better assess the performance of the proposed signer-independent Transformer-based model, a comparative analysis is conducted against representative deep learning approaches reported in prior work using RGB-D data. Table 1 summarizes the recognition rates achieved by different CNN-LSTM-based architectures under the same dataset.

Conventional CNN + LSTM models have a rather poor recognition capacity with Top-1 scores under 40 percent on test set. The consideration of feature pyramid modules (FPM) and attention mechanisms are great enhancements to performance and it seems that indicating the significance of multi-scale feature extraction and time attention to dynamic sign recognition. One of these approaches, CNN + FPM + BLSTM + Attention model attains the best reported achieving a Top-1 accuracy of 62.02% on the balanced test set and 59.24% on the imbalanced test set, indicating the effectiveness of a two-way time model and focusing on fusion attention. To ensure similarity, all the compared models are reported in terms of recognition performance in terms of Top-1 accuracy as a percentage.

Table 1: Performance Comparison of Different Models

Method	Input Modality	Accuracy
CNN + FPM + LSTM + Attention [6]	RGB Isolated Sign video	60.02%
CNN + FPM + BLSTM + Attention [6]	RGB Isolated Sign video	62.02%
Proposed Signer-Independent Transformer	RGB Isolated Sign video	83.52%

As shown in Table 1, the model worked comparatively well, in a signer-independent evaluation environment that is strictly imposed. Even though the obtained level of accuracy is somewhat lower than that of certain existing methods, it should be taken into account that there are also certain differences in the evaluation protocols. Several of the previous approaches are claimed to perform better in less demanding settings, including signer-dependent or mixed-signer settings, in which the model can implicitly acquire identity-specific behavior. The offered method, in its turn, is aimed at generalization between invisible signers and this results in a great deal of variability; variability of hand shape, style of motion and the speed of execution. It is more challenging to learn invariant representations in these conditions, and thus it is natural to expect that a slight decrease in peak accuracy will be observed. Nonetheless, the design option enhances the strength of the model as used in practice.

Comparatively, the signer-independent proposed signer-free Transformer-based model records 83.52% as the Top-1 recognition rate, which is significantly higher than the CNN-LSTM based approaches of Table 1. The fact that transformer architectures excel in capturing long-range temporal dependencies over the recurrent networks is used to highlight the effectiveness of models. In contrast to LSTM-based models, the self-attention mechanism in the proposed model allows explicit communication between all the time frames, leading to more discriminative motion representatives. This proposed model was able to reach a precision of 84.10%, recall of 83.52%and an F1-score of 83.81%.

Overall, the comparison clearly indicates that transformer-based temporal modeling combined with landmark-level normalization provides a powerful alternative to conventional CNN-LSTM pipelines. Not only the proposed model is better than the available RGB-based approaches with references to Top-1 recognition rate but it is also more robust and scalability to real world sign language recognition applications.

Conclusion and future scope

The Proposed independent sign language recognition model that trains on skeletal landmark sequence with a transformer-based model. The suggested approach relies on motion patterns as the input of information to achieve the reliable identification with a vast scope of signs in a vocabulary but also tolerant of variations in signing style. Analysis of the AUTSL data through experimental testing reveals that there is high recognition rate such that it is highly validated and tested accuracy and validation of the efficacy of temporal modeling as related to attention in dynamic sign gestures. Moreover, the existing system is programmed to work with individual signs and is not yet set to work with continuous recognition of sign language on sentence basis.

This work is opening a few promising directions in the future. The model can be extended here in future with finer cues of motion or more modalities to reflect more expressive dynamic gestures. The system will also be generalized into the creation of persistent identification and sentence-scope perceiving, which will bring the automatic sign language translation to the realm of usefulness. The specified approach can be made the contributor to the inclusion and availability of the communication technologies with further optimization associated with the deployment in real-time.

References

1. Li D, Rodriguez C, Yu X, Li H. Word-level deep sign language recognition from video. IEEE Trans Multimedia. 2020; 22(10):2561-2573.
2. Huang J, Zhou W, Li H, Li W. Sign language recognition using 3D convolutional neural networks. IEEE Trans Multimedia. 2020; 22(11):2938-2949.
3. Rastgoo R, Kiani K, Escalera S. Sign language recognition: A deep survey. IEEE Trans Pattern Anal Mach Intell. 2021; 43(11):3765-3788.
4. Liu J, Wang X, Chen Y, Liu S, Gao Z. Skeleton-based human action recognition with deep learning: A survey. IEEE Trans Pattern Anal Mach Intell. 2021; 43(9):3047-3066.
5. Huang J, Zhou W, Li H. Skeleton-based sign language recognition using sequence learning. IEEE Access. 2021; 9:16806-16815.
6. Sincan O, Keles HY. AUTSL: A large-scale multi-modal Turkish sign language dataset and baseline methods. IEEE Access. 2020; 8:181340-181352.
7. Bragg D, Koller O, Bellard M, Berke L, Boudreault P, Braffort A, et al. Sign language recognition, generation, and translation: An interdisciplinary perspective. IEEE Computer. 2021; 54(3):16-26.
8. Camgöz NC, Hadfield S, Koller O, Bowden R. 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 (CVPR); 2020. p. 10023-10033.
9. Cheng Y, Yang X, Q islands Li, Chen X. Improving signer-independent sign language recognition using *pose normalization* and *attention mechanisms*. IEEE Trans Circuits Syst Video Technol. 2023; 33(6):3124-3136.
10. Shi L, Zhang Y, Cheng J, Lu H. Skeleton-based action recognition with directed graph neural networks. IEEE Trans Pattern Anal Mach Intell. 2022; 44(6):3323-3339.
11. Zhou H, Zhou W, Li H, Wang W. *Spatial-temporal* multi-cue network for continuous sign language recognition. IEEE Trans Multimedia. 2021; 23:1521-1534.
12. Cheng J, Li P, Wang M. Attention-based *temporal aggregation* for skeleton sequence representation. IEEE Access. 2021; 9:104832-104843.
13. Mekala VK, Rao MB, Reddy S. Visual DUX: A low cost wearable device for guiding the blind. Turkish J Comput Math Educ. 2021; 12(8):2731-2738.
14. Davuluri RV, Ragupathy R. Improved classification model using *CNN* for detection of *Alzheimer's disease*. Journal of Computer Science. 2022; 18(5):415-425.
15. Babu Rao M, Raghava Rao K, Kishore PVV. Content

based image retrieval based on dominant color, scan pattern co-occurrence matrix of a motif and shape. In: Proceedings of the International Conference on Computational Intelligence and Information Technology (CIIT). LNCS-CCIS, vol. 250. Springer; 2011. p. 353-357.
16. Babu Rao M, Raghava Rao K, Kishore PVV. Content based image retrieval using an integrated matching technique based on the most similar highest priority principle on color and texture features of image sub-blocks. In: Proceedings of the International Conference on Advances in Information Technology and Mobile Communication (AIM). LNCS-CCIS, vol. 147. Springer; 2011. p. 399-402.
17. Naz N, Akram MA, Khan SA. Sign Graph: An efficient *pose-based* graph convolution approach toward sign language recognition. IEEE Access. 2023; 11:19135-19147.