Pakistani Word-level Sign Language Recognition Based on Deep Spatiotemporal Network

Shehryar Naeem[1], **Hanan Salam**[2], **Md Azher Uddin**[1]*

[1]School of Mathematical and Computer Sciences, Heriot-Watt University Dubai
Dubai, United Arab Emirates
[2]Computer Science Department, New York University Abu Dhabi
Abu Dhabi, United Arab Emirates
m.uddin@hw.ac.uk, hanan.salam@nyu.edu

Abstract

Sign language is crucial for the Deaf and Hard-of-Hearing community because it facilitates visual movement-based communication. Nevertheless, most are not familiar with it, rendering interactions with the hearing impaired complicated. While there has been significant work on languages, for instance, American and Chinese Sign Language, Pakistani Sign Language (PSL) at the word level has received less attention and has been studied based on static images. To address this, we introduce a deep spatiotemporal network for word-level PSL recognition from video. It commences by employing top-k frame extraction to enhance processing efficiency. Second, the ResNet-101 model is utilized for extracting deep spatial features from each frame. Subsequently, we introduce the Adaptive Motion Binary Pattern (AMBP), a new spatiotemporal feature descriptor that effectively extracts the spatiotemporal features. These spatial and spatiotemporal are fused and input into the transformer model that processes these representations for better recognition. Experimental evaluations confirm that our framework achieves state-of-the-art results.

Introduction

Sign language serves as a crucial mode of communication for the Deaf and Hard-of-Hearing (DHH) community worldwide (Scott and Dostal 2019). However, sign languages are not universal; different countries and regions have developed distinct linguistic structures, each tailored to their cultural and linguistic contexts (Wheatley and Pabsch 2010). For instance, American Sign Language (ASL) (Bantupalli and Xie 2018) is widely used in the United States, whereas British Sign Language (BSL) (Bird, Ekárt, and Faria 2020) follows a completely different grammatical structure. Other widely recognized sign languages include Indian Sign Language (ISL) (Attar, Goyal, and Goyal 2023), Arabic Sign Language (ArSL) (Al-Shamayleh, Ahmad, and Jomhari 2020), Korean Sign Language (KSL) (Shin et al. 2024), and Chinese Sign Language (CSL) (Jiang, Zhang, and Lei 2024), and Pakistani Sign Language (PSL) (Arooj et al. 2024). Despite this diversity, research on automated sign language recognition has primarily focused on well-established languages such as ASL and BSL, leaving many regional sign languages, such as PSL, relatively underexplored.

Pakistan has a large Deaf and Hard-of-Hearing population, yet computational research in PSL recognition has received minimal attention in computational research (Farooq et al. 2021). Most existing studies focus on static PSL recognition, where individual signs are identified from still images (Najib 2024). However, natural communication in PSL, like in other sign languages, often involves dynamic word-level recognition, which requires understanding sequential hand movements and temporal dependencies (Mujeeb et al. 2024). Unfortunately, very few studies have tackled word-based PSL recognition comprehensively. The ability to recognize full words, rather than isolated static gestures, is essential for building practical communication systems that more accurately represent natural sign language. In addition, dynamic texture descriptors have not been investigated well for word-level PSL recognition.

The emergence of Cyber-Physical Systems (CPS) (Gunes et al. 2014; Horvath and Gerritsen 2012) provide an opportunity to bridge the communication gap between the DHH community and modern technology by enabling Deaf individuals to interact seamlessly with computer-based technologies. CPS integrates physical and computational elements to enhance real-world interactions, making them highly relevant for assistive technologies. Sign language recognition, when integrated with CPS, can facilitate interactions between Deaf users and smart systems such as Amazon Alexa, Google Assistant, and other intelligent home automation systems (Ahmed et al. 2018). This integration can enhance accessibility, allowing Deaf individuals to communicate naturally with computer-based technologies in their everyday lives.

In this paper, we propose a deep spatiotemporal network for word-level Pakistani Sign Language (PSL) recognition from video data. Our end-to-end system effectively extracts deep spatial as well as dynamic spatiotemporal features for overcoming the previously mentioned challenges. Initially,

*Corresponding author: Md Azher Uddin.
This work is supported in part by the NYUAD Center for Artificial Intelligence and Robotics, funded by Tamkeen under the NYUAD Research Institute Award CG010.

we employ the ResNet-101 model (He et al. 2016) for extracting fine-grained spatial features from individual frames. Subsequently, we introduce a new dynamic feature descriptor, Adaptive Motion Binary Pattern (AMBP), which plays a crucial role in efficiently extracting temporal dynamics. These spatial and temporal features are later fused and processed through a Transformer model, which learns effectively a sequence of movements inherent in sign language. We assessed the performance of our framework on two diverse PSL datasets. The results of our experiments prove that our model outperforms state-of-the-art methods extensively, confirming its effectiveness for effectively recognizing PSL.

Related Works

Pakistani Sign Language (PSL), similar to other sign languages, involves a combination of hand shapes, movements, facial expressions, and spatial orientations to convey meaning. Research on PSL recognition has primarily focused on static (image-based) gesture recognition, particularly for alphabets, characters, and digits, while relatively few studies have explored dynamic (video-based) recognition, especially at the word and sentence levels. The following reviews related work on PSL in these two categories.

Static-Based PSL Recognition. A substantial number of studies have been conducted on static PSL gesture recognition, where each sign is represented as a still image. Researchers have used traditional machine learning techniques and deep learning models to classify PSL alphabets and numbers. For instance, (Ali, Hosseini, and Pervez 2025) evaluated seven machine learning models on a alphabet PSL recognition, identifying Random Forest as the most effective for PSL alphabet recognition on that dataset. Naseem et al. (Naseem et al. 2019) introduced a convolutional neural network-based web application that integrates hand tracking for real-time PSL recognition, achieving nearly perfect accuracy on static Pakistani alphabet signs, and highlighting the potential of deep learning in real-time PSL applications. (Arooj et al. 2024) proposed a hybrid PSL recognition model combining CNNs and SIFT-based feature extraction. Using Kinect sensor data, their system outperformed SVM-based and 3D-CNN models. Similarly, (Manzoor et al. 2024) employed CNNs in a bidirectional PSL and ASL sign language translation system, integrated in a real-time mobile application. Their system integrates CNN-based sign gesture recognition, NLP for text-to-sign conversion, and real-time hand tracking. Despite these advances, static PSL recognition systems are inherently limited as they do not account for motion, transitions between gestures, or real-world variations in signing speed and environment.

Word-Based and Dynamic PSL Recognition. While static PSL recognition has been extensively studied, dynamic PSL recognition (video-based) remains a relatively underdeveloped area. Recognizing PSL words and sentences involves temporal modeling of hand movements, which introduces additional challenges such as signing speed variations, hand trajectory tracking, and gesture coarticulation.

Some research has attempted to tackle this issue using Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, which are designed for sequential data. For instance, (Mujeeb et al. 2024) developed a real-time PSL web application that integrates dynamic PSL recognition into a browser-based environment. Their approach employs LSTM networks for recognizing video-based PSL gestures, using a dataset of 353 PSL videos. The dynamic recognition model achieved 100% accuracy in classifying three PSL dynamic gestures. However, despite its high accuracy, the number of dynamic PSL words recognized is very limited (only three words), which significantly restricts its usability for broader communication. Later on, (Javaid and Rizvi 2023) introduced a hybrid multimodal approach using Action Transformer Networks (SLATN) for PSL recognition. Their model extracts spatiotemporal features and leverages a Transformer-based attention mechanism to simultaneously track hand gestures, facial expressions, and body movements in video sequences. This approach outperforms traditional activity recognition methods and achieves a testing accuracy of 82.66% while maintaining high computational efficiency. Additionally, their work contributes a new dataset for Pakistani Sign Language (PkSLMNM), specifically designed to capture both manual and non-manual gestures. Furthermore, (Hamza and Wali 2023) proposed a PSL recognition system based on video data, leveraging Convolutional 3D (C3D), Inflated 3D ConvNet (I3D), and Temporal Shift Module (TSM) models. Due to the limited PSL dataset, the study introduced a data augmentation pipeline to improve model generalization. Results showed that rotation and translation-based augmentation significantly enhanced recognition accuracy, with C3D achieving 93.33% accuracy as the best-performing model. The study also highlighted the limitations of TSM for PSL recognition, as it struggled with movement similarities across different signs.

Proposed Framework

We propose an end-to-end spatio-temporal network for video-based word-level Pakistani Sign Language recognition. Figure 1 depicts the proposed framework. The framework starts with the pre-processing operations such as top-k frame extraction, frame resize and RGB to Grayscale conversion. Then, a pre-trained ResNet-101 model is used to extract the deep spatial features from each frame. Simultaneously, a novel dynamic feature descriptor named Adaptive Motion Binary Pattern (AMBP) is proposed to effectively extract the spatiotemporal features. Finally, these spatial and spatiotemporal extracted features are concatenated and fed into a transformer that learns the spatiotemporal representations to achieve better recognition.

Top-K frame extraction

Key-frame extraction is an important video analysis process that enables effective summarization by selecting the most representative frames. Instead of processing all the frames, which is computationally expensive and memory-intensive, key-frame extraction focuses on the identification of key frames that best capture the changes in the video content. The process is successfully applied to a number of

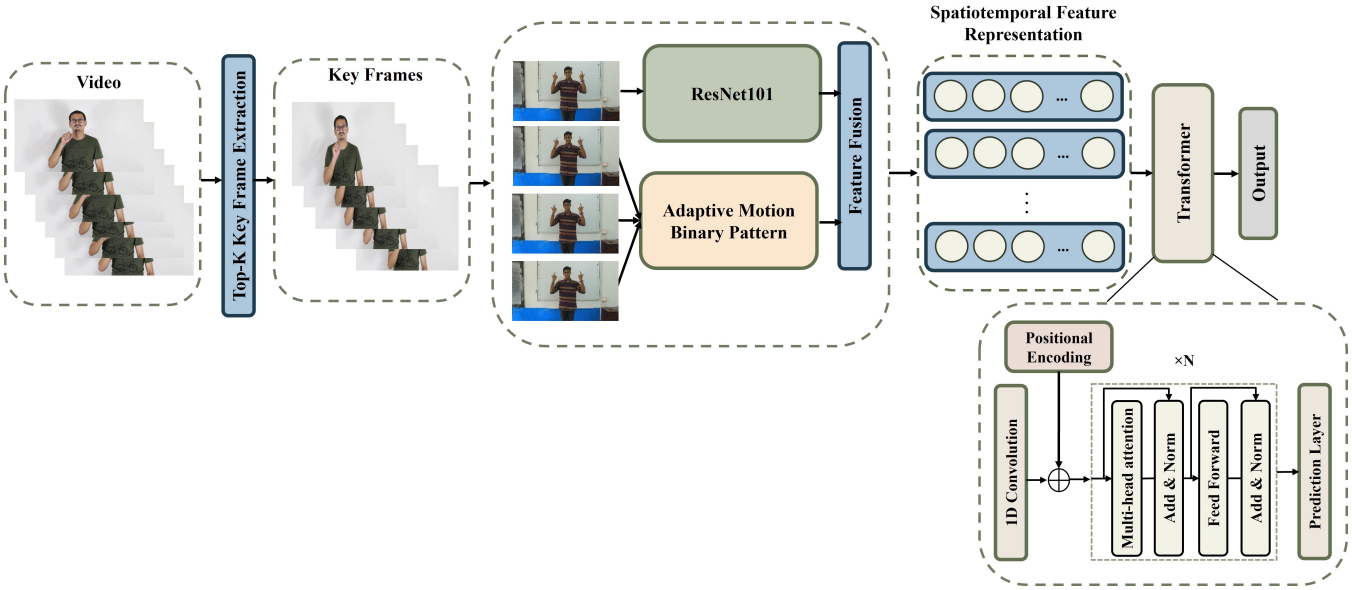

Figure 1: Proposed framework for word-level Pakistani sign language recognition from videos.

applications of computer vision such as video summarization (Apostolidis et al. 2021), and content-based retrieval (Spolaôr et al. 2020). In this paper, we extract the top-K key frames of videos which significantly improve the processing speed. The proposed method leverages a pre-trained deep neural network, ResNet-101 (He et al. 2016), to extract discriminative spatial features of individual frames. Through the comparison of consecutive frames, the method identifies key transitions and changes in the video. The contrast of the extracted features of two consecutive frames is computed to identify the change in the content. Frames with the maximum differences are the most informative and therefore selected as keyframes. The process effectively reduces redundancy without sacrificing the core visual information that allows the understanding of the video's content. Algorithm 1 1 presents the process of extracting the Top-K key frame. In our work, the K value is set to 30, which is empirically selected.

Algorithm 1: Top-K Key Frame Extraction

Input: Video File
Output: Top-K Key Frames
Procedure Key-Frame(video)
for $i \leftarrow 1$ to Number-of-Frames **do**
  $A \leftarrow$ ReadFrame(video, i)
  $B \leftarrow$ ReadFrame(video, $i + 1$)
  FeatureA $\leftarrow$ ExtractResNet101Feature(A)
  FeatureB $\leftarrow$ ExtractResNet101Feature(B)
  DifferenceAB $\leftarrow$ FeatureDifference(FeatureA, FeatureB)
  $X[i] \leftarrow$ DifferenceAB
end for
$[sortedX, sortingIndices] \leftarrow$ Sort(X, *'descend'*)
End Procedure

Deep Spatial Feature Extraction Using ResNet-101

The ResNet-101 model (He et al. 2016) is employed in extracting deep spatial features from video frames to provide robust feature representation for the recognition of Pakistani Sign Language (PSL). The Residual Network is a state-of-the-art deep learning model designed to ease the vanishing gradient issue by introducing residual connections to facilitate deeper network training while preserving essential feature information. The ResNet-101 network consists of multiple residual blocks, each of which comprises identity mappings and shortcut connections to enhance gradient flow. The model consists of four major stages, with each stage progressively extracting high-level spatial features through a series of convolutional, batch normalization, and ReLU activation layers. These layers capture complex spatial patterns that are required for discriminating between different sign gestures. One major advantage of using ResNet-101 for PSL recognition is that it can handle lighting, scale, and hand position variations because of its deep hierarchical feature representation. The convolutional layers act as spatial feature extractors, while global average pooling reduces dimensionality, keeping the most relevant information. In addition, batch normalization and ReLU activation introduce nonlinearity and accelerate training convergence. In this work, features are extracted from the GAP layer of ResNet-101, ensuring a compact and effective feature representation for PSL recognition.

Adaptive Motion Binary Pattern Based Spatiotemporal Features Extraction

We propose Adaptive Motion Binary Pattern (AMBP) based Spatiotemporal feature descriptor that is able to effectively capture motion dynamics and spatial texture information from video frames. Unlike traditional texture-based descriptors such as Local Binary Pattern (LBP) (Ojala, Pietikainen,

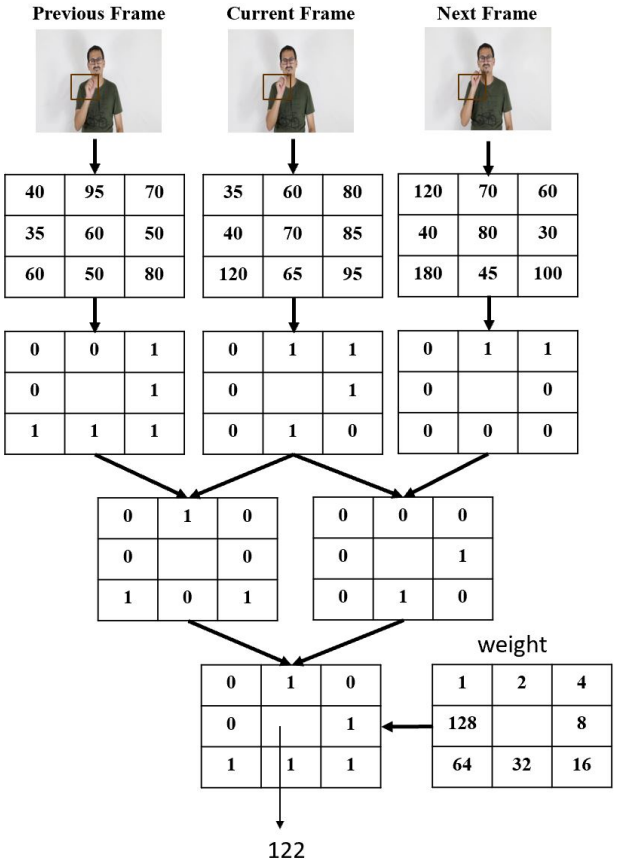

Figure 2: Spatiotemporal feature extraction using AMBP.

and Maenpaa 2002), which predominantly capture static textures, AMBP extends these methods by incorporating motion differences between three consecutive frames. In order to describe dynamic information from video data, researchers have already proposed techniques like Volume Local Binary Patterns (VLBP) (Zhao and Pietikainen 2007), Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) (Zhao and Pietikainen 2007), and adaptive local motion descriptor (ALMD) (Uddin et al. 2017). Nevertheless, these methods inherit the same drawbacks as conventional LBP, namely sensitivity to illumination changes and noise. To overcome these limitations, we present an adaptive threshold mechanism to improve adaptability. As a result, the proposed AMBP proves to be robust to illumination changes, noise, and motion.

AMBP derives motion features by analyzing pixel intensity differences among three continuous frames: previous frame, current frame, and next frame. First, every frame is divided into small local grids, where pixel intensity values are extracted. To give adaptability, a dynamic threshold (th) is derived from the absolute difference between the center frame center pixel C_c and neighboring pixels C_i, so that feature extraction is robust. The threshold is provided by:

$$th = \frac{1}{n}\sum_{i=1}^{n}|C_c - C_i| \tag{1}$$

Once the threshold is determined, a binary pattern for every pixel is computed through a comparison of the previous frame, current frame, and next frame neighboring pixels' intensity, with the current frame center pixel intensity and the threshold value. The intensity change is set to 1 if it is within the range of the threshold; otherwise, it is set to 0. This generates a binary motion map that highlights regions of significant pixel intensity differences, noting motion between frames. To encode motion features effectively, the binary maps from the previous, current, and next frames are processed using a bitwise XOR operation, given as:

$$AMBP_{n,r}(x_c, y_c) = \sum_{i=1}^{n}\Big(s(P_i - C_c) \oplus s(C_i - C_c) \oplus s(N_i - C_c)\Big) \times 2^i \tag{2}$$

$$s(a) = \begin{cases} 1, & \text{if } -th \le a \le th, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Where P_i, C_i, and N_i are neighboring pixel intensities in the previous, current, and next frames, respectively, and $\oplus$ denotes the XOR operation. This provides a compact but descriptive motion pattern representation. The resulting binary map is converted to a decimal number. Figure 2 illustrates the process of AMBP. AMBP presents several advantages over other feature descriptors, by incorporating adaptive thresholding, it can handle illumination variation and noise quite well, and it is robust to feature extraction under different environments. The combination of motion encoding and spatial texture representation enhances its ability to detect fine-grained movements in sign language recognition.

Feature Fusion

The pre-trained ResNet-101 model processes a $224 \times 224 \times 3$ RGB hand sign frame as input, capturing spatial details and generating a 1D feature vector with 2048 features. Meanwhile, the AMBP descriptor processes three consecutive frames (previous, current, and next), each sized 224×224, to extract spatiotemporal features, resulting in a 1D feature vector with 256 features. After extracting both deep and spatiotemporal features, we fuse them to produce a final feature vector of size 1×2304 per frame. Therefore, for the top-k key frames of a video, we obtain a $k \times 2304$ feature matrix. This feature matrix is then used as input for the transformer model.

Learning Spatiotemporal Features Using Transformer Model

Initially, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks were employed to model spatiotemporal dynamics of the video data to perform various recognition tasks such as gesture and action recognition (Hu et al. 2018; Cifuentes et al. 2019). These methods

effectively encoded sequential information but suffer from the issues of vanishing gradients, parallel processing issues, and modeling long-term temporal dependencies. Transformers initially emerged in the domain of Natural Language Processing (NLP) applications (Gillioz et al. 2020) and later moved to computer vision applications (Han et al. 2022) to excel at modeling long-range interactions utilizing self-attention mechanisms. The Transformer-based models were recently employed in the community of time series to carry out forecasting and regression tasks (Wu et al. 2020; Born and Manica 2023) but their application to learn spatiotemporal features from the extracted features of the videos is still comparatively limited. To mitigate the drawbacks of RNN and LSTM-based models, our approach introduces a Transformer-based model designed to learn spatiotemporal representation from the extracted features of the videos to achieve effective word-level PSL recognition. The proposed Transformer model comprises four components: embedding layer, positional encoding, Transformer encoder, and prediction layer. At first, we employ a 1D convolution-based embedding layer with d_{model} filters on the input features from ResNet-101 model and AMBP. More specifically, given the input feature sequence $X \in \mathbb{R}^{k \times d}$, where k is the number of frames and d is the dimension of the features, the embedded representation X_{embed} is computed as:

$$X_{embed} = \text{Conv1D}(X) \tag{4}$$

Afterward, to incorporate temporal relationship crucial for sequential data modeling, positional encoding is added to the embeddings. Specifically, we utilize sinusoidal positional encoding defined by:

$$PosEnc_{(po,2i)} = \sin\left(\frac{po}{10000^{2i/d_{model}}}\right), \tag{5}$$

$$PosEnc_{(po,2i+1)} = \cos\left(\frac{po}{10000^{2i/d_{model}}}\right) \tag{6}$$

where p indicates the position of the feature in the sequence, i represents the dimension index, and d_{model} denotes the dimension of the embedding vector. This positional encoding enables the model to retain sequential ordering, addressing the permutation-invariance issue inherent to standard Transformer architectures. The Transformer encoder consists of multiple layers, each comprising a multi-head self-attention sub-layer followed by a position-wise feed-forward neural network. The multi-head self-attention mechanism computes attention weights between every pair of positions, effectively capturing global contextual dependencies within the entire sequence. Mathematically, the self-attention mechanism is expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

where queries Q, keys K, and values V are linear projections of the embedded input, and d_k is the dimensionality of the keys. The outputs from multiple attention heads are concatenated and subsequently projected linearly to obtain the final output. In this work, we employed 3 encoder layers, which is chosen empirically. Each Transformer encoder layer also contains a feed-forward neural network, which is composed of two fully-connected layers with a nonlinear activation, along with residual connections and layer normalization for stable training. Finally, extracted features from the Transformer encoder are fed into the prediction layer designed for multiclass classification. This prediction layer consists of a dense layer with a softmax activation function to output probability distributions over different hand sign classes. To reduce potential overfitting during training, we integrate a dropout layer with a dropout rate of 0.2 before the dense layer. The training procedure employs the categorical cross-entropy loss function, and the Adam as optimizer, with a batch size of 32, a momentum value of 0.9, and a learning rate of 1×10^{-4}.

Experiments

In this section, we analyze our proposed framework's performance. First, we provide a description of the datasets and experimental conditions. Next, we perform an ablation study for assessing the contribution of various components of our framework. Finally, we demonstrate a comparison of our approach's performance with state-of-the-art techniques.

Experimental Settings

We evaluated our framework using the PkSLMNM dataset (Javaid and Rizvi 2023) and the PSL dictionary dataset (Hamza and Wali 2023). The PkSLMNM dataset comprises 665 videos of 180 individuals. The dataset contains 7 classes of Pakistani hand signs, which include bad, best, sad, glad, scared, stiff, and surprise adjectives. In this dataset, 80% of data were used for training the model while 20% of data were used for testing. On the other hand, the PSL dictionary dataset consists of 160 samples for 80 words. The dataset is split into 80 training samples and 80 test samples. To evaluate the performance of the proposed model we used accuracy as an evaluation metric, which is represented as follows.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Samples}} \tag{8}$$

The experiment was conducted on a PC running Windows 10 with a 64-bit architecture, equipped with an Intel(R) Core(TM) i7-10750H CPU and 16GB of RAM.

Ablation Study

Figure 3 illustrates an ablation experiment on how each component affects the accuracy of our proposed deep spatiotemporal network. The experiment shows that incorporating Adaptive Motion Binary Pattern (AMBP) with ResNet-101 GAP features improves the accuracy. This notable improvement indicates the necessity of incorporating temporal dynamics along with spatial features, and it also attests to the effectiveness of our integrated approach for handling the complexities of sign language recognition.

Figure 4 represents the result of varying the number of keyframes on the recognition of sign language. Accuracy peaks with 30 keyframes, demonstrating an optimal balance between detail capture and computational efficiency. However, increasing to 35 keyframes shows a slight decrease in accuracy. This experiment underscores the importance of

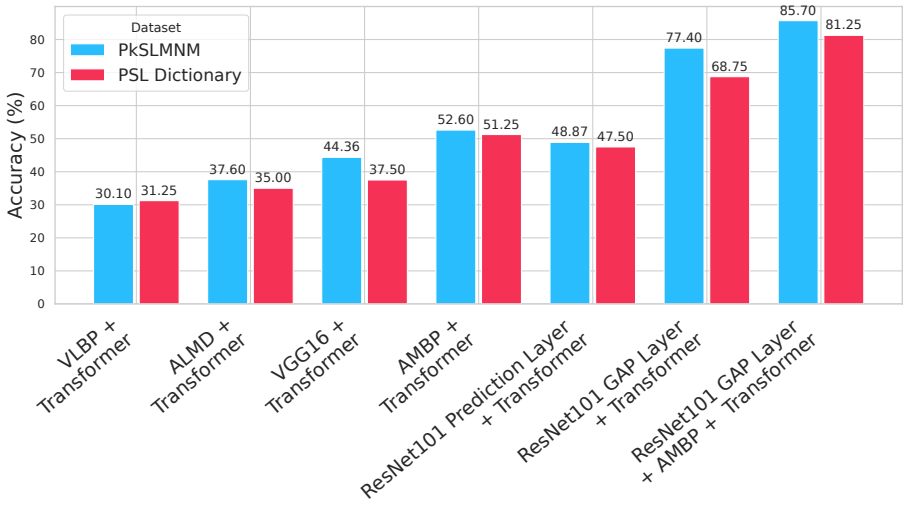

Figure 3: Assessing the effectiveness of different components in the proposed framework.

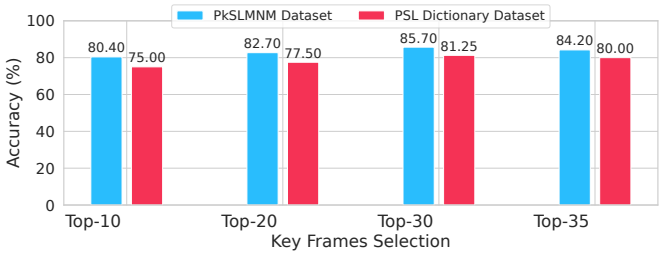

Figure 4: Investigating the accuracy with different key frames.

optimal key frame selection for efficient and effective sign language recognition.

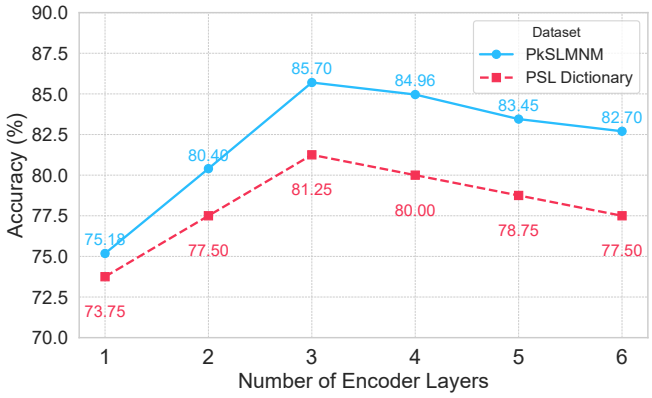

Figure 5: Investigating the accuracy of the transformer model by varying the number of encoder layers.

Then, we experimented with varying the number of layers for the encoder layer of the Transformer architecture,

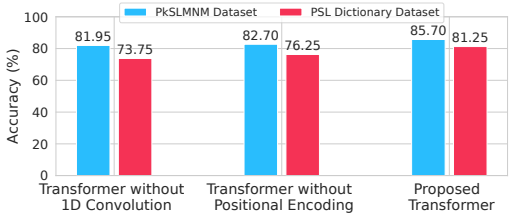

Figure 6: Investigating the accuracy of the transformer model with different component.

as depicted in Figure 5. The model's accuracy is best with three layers for the encoder, reaching a peak accuracy of 85.7% for PkSLMNM dataset and 81.25% for PSL dictionary dataset, respectively.

Figure 6 shows a comparison between three configurations of Transformer models on PkSLMNM and PSL dictionary datasets. Notably, the proposed Transformer model exhibits superior performance. This indicates that the integration of both 1D convolutional layers and positional encoding is crucial for extracting the complexities of Pakistani sign language.

Fig. 7 illustrates a comparison of our proposed transformer model's accuracy with other traditional machine learning methods, such as SVM, Random Forest, and Ada Boost (Uddin, Denny, and Joolee 2022), and with various deep learning networks, such as 1D CNN (Mendez, Uddin, and Joolee 2022), BiLSTM (Uddin, Joolee, and Lee 2020), GRU (Subramanian et al. 2022), and Encoder-Decoder Networks (Uddin, Denny, and Joolee 2022). From the experiment, it is clear that our proposed model outperforms all these alternatives with a huge margin, highlighting its better capability for extracting and processing features necessary for correctly identifying word-level Pakistani Sign Language.

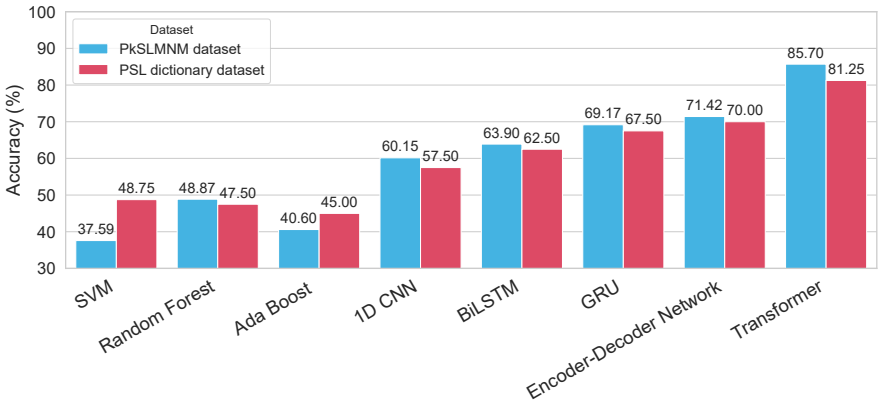

Figure 7: Comparison of different machine learning and deep learning models with the proposed Transformer.

Comparison with State-of-the-art Approaches

Table 1 shows a comparative analysis of our proposed approach with various state-of-the-art models for PkSLMNM Dataset and PSL Dictionary Dataset. The results indicate that our proposed approach achieves the best accuracy, 85.7% on PkSLMNM Dataset and 81.25% on PSL Dictionary Dataset, compared with models such as C3D (Hamza and Wali 2023), TSM (Hamza and Wali 2023), I3D (Hamza and Wali 2023), and SLATN (Javaid and Rizvi 2023). Notably, even though models like I3D achieve 77.50% accuracy on the PSL Dictionary Dataset, they are still below our approach. Such a significant improvement in performance indicates the potential of our framework for recognizing word-level Pakistani Sign Language as a superior choice for real-time applications for sign language interpretation.

	Dataset (%)	
Method	**PkSLMNM**	**PSL Dictionary**
C3D	78.9	66.67
TSM	75.9	33.75
I3D	78.2	77.50
SLATN	82.66	–
Ours	**85.7**	**81.25**

Table 1: Performance comparison in terms of Accuracy (%) between proposed framework and other State-of-the-art models.

Conclusion

This work significantly contributes to Pakistani Sign Language recognition with a robust deep spatiotemporal network capable of interpreting dynamic sign language effectively. Our proposed approach achieves better accuracy than state-of-the-art models on comprehensive PSL datasets. These findings prove the effectiveness of combining deep spatial and dynamic temporal features. It can revolutionize tools for communication among the Deaf and Hard-of-Hearing, and bring them closer to society through greater interactions with digital and cyber-physical devices. Our future work will focus on enhancing the model's efficiency and exploring its potential for real-time PSL translating devices.

References

Ahmed, M. A.; Zaidan, B. B.; Zaidan, A. A.; and Salih, M. M. 2018. A review on systems-based sensory gloves for sign language recognition: State of the art between 2007 and 2017. *Sensors*, 18(7): 2208.

Al-Shamayleh, A. S.; Ahmad, R.; and Jomhari, N. 2020. Automatic Arabic sign language recognition: A review, taxonomy, open challenges, research roadmap and future directions. *Malaysian Journal of Computer Science*.

Ali, M.; Hosseini, S. E.; and Pervez, S. 2025. Assessment and Enhancement of Real-Time Recognition of Sign Language Alphabets Through Diverse Machine Learning Techniques. In *2025 8th International Conference on Data Science and Machine Learning Applications (CDMA)*, 85–90. IEEE.

Apostolidis, E.; Adamantidou, E.; Metsai, A. I.; Mezaris, V.; and Patras, I. 2021. Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, 109(11): 1838–1863.

Arooj, S.; Altaf, S.; Ahmad, S.; Mahmoud, H.; and Mohamed, A. S. N. 2024. Enhancing sign language recognition using CNN and SIFT: A case study on Pakistan sign language. *Journal of King Saud University-Computer and Information Sciences*, 36(2): 101934.

Attar, R. K.; Goyal, V.; and Goyal, L. 2023. State of the art of automation in Sign Language: a systematic review. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Bantupalli, K.; and Xie, Y. 2018. American Sign Language Recognition using Deep Learning and Computer Vision. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE.

Bird, J. J.; Ekárt, A.; and Faria, D. R. 2020. British Sign Language Recognition via Late Fusion of Computer Vision

and Leap Motion with Transfer Learning to American Sign Language. *Sensors*, 20(18): 5151.

Born, J.; and Manica, M. 2023. Regression transformer enables concurrent sequence regression and generation for molecular language modelling. *Nature Machine Intelligence*, 5(4): 432–444.

Cifuentes, J.; Boulanger, P.; Pham, M. T.; Prieto, F.; and Moreau, R. 2019. Gesture classification using LSTM recurrent neural networks. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 6864–6867. IEEE.

Farooq, U.; Rahim, M. S. M.; Abid, A.; and Khan, N. S. 2021. A Process to Develop Sign Language Corpus using Crowdsourcing. *International Journal of Innovative Technology and Exploring Engineering*.

Gillioz, A.; Casas, J.; Mugellini, E.; and Abou Khaled, O. 2020. Overview of the Transformer-based Models for NLP Tasks. In *2020 15th Conference on computer science and information systems (FedCSIS)*, 179–183. IEEE.

Gunes, V.; Peter, S.; Givargis, T.; and Vahid, F. 2014. A survey on concepts, applications, and challenges in cyber-physical systems. *Journal of Computing and Information Systems*.

Hamza, H. M.; and Wali, A. 2023. Pakistan sign language recognition: leveraging deep learning models with limited dataset. *Machine Vision and Applications*, 34(5): 71.

Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1): 87–110.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Horvath, I.; and Gerritsen, B. H. M. 2012. Cyber-physical systems: Concepts, technologies, and implementation principles. *Proceedings of TMCE*.

Hu, Y.; Wong, Y.; Wei, W.; Du, Y.; Kankanhalli, M.; and Geng, W. 2018. A novel attention-based hybrid CNN-RNN architecture for sEMG-based gesture recognition. *PloS one*, 13(10): e0206049.

Javaid, S.; and Rizvi, S. 2023. A Novel Action Transformer Network for Hybrid Multimodal Sign Language Recognition. *Computers, Materials & Continua*, 75(1).

Jiang, X.; Zhang, Y.; and Lei, J. 2024. A Survey on Chinese Sign Language Recognition: From Traditional Methods to Artificial Intelligence. *Computer Modeling in Engineering & Sciences*.

Manzoor, S.; Abbas, Z.; Chhabra, G.; Kaushik, K.; Zehra, M.; Haider, Z.; and Khan, I. U. 2024. Voice of Hearing and Speech Impaired People. In *2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)*, 1891–1897. IEEE.

Mendez, J.; Uddin, M. A.; and Joolee, J. B. 2022. Spontaneous Macro and Micro Facial Expression Recognition Using ResNet50 and VLDSP. In *International Conference on Information Technology and Applications*, 159–170. Springer.

Mujeeb, A. A.; Khan, A. H.; Khalid, S.; Mirza, M. S.; and Khan, S. J. 2024. A neural-network based web application on real-time recognition of Pakistani sign language. *Engineering Applications of Artificial Intelligence*, 135: 108761.

Najib, F. M. 2024. A multi-lingual sign language recognition system using machine learning. *Multimedia Tools and Applications*.

Naseem, M.; Sarafraz, S.; Abbas, A.; and Haider, A. 2019. Developing a prototype to translate pakistan sign language into text and speech while using convolutional neural networking. *Journal of Education and Practice*, 10(15): 10–7176.

Ojala, T.; Pietikainen, M.; and Maenpaa, T. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7): 971–987.

Scott, J. A.; and Dostal, H. M. 2019. Language Development and Deaf/Hard of Hearing Children. *Education Sciences*, 9(2): 135.

Shin, J.; Miah, A. S. M.; Akiba, Y.; Hirooka, K.; and Hassan, N. 2024. Korean sign language alphabet recognition through the integration of handcrafted and deep learning-based two-stream feature extraction approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Spolaôr, N.; Lee, H. D.; Takaki, W. S. R.; Ensina, L. A.; Coy, C. S. R.; and Wu, F. C. 2020. A systematic review on content-based video retrieval. *Engineering Applications of Artificial Intelligence*, 90: 103557.

Subramanian, B.; Olimov, B.; Naik, S. M.; Kim, S.; Park, K.-H.; and Kim, J. 2022. An integrated mediapipe-optimized GRU model for Indian sign language recognition. *Scientific Reports*, 12(1): 11964.

Uddin, M. A.; Denny, R.; and Joolee, J. B. 2022. Deep Spatiotemporal Network Based Indian Sign Language Recognition from Videos. In *International Conference on Information Technology and Applications*, 171–181. Springer.

Uddin, M. A.; Joolee, J. B.; Alam, A.; and Lee, Y.-K. 2017. Human action recognition using adaptive local motion descriptor in spark. *IEEE Access*, 5: 21157–21167.

Uddin, M. A.; Joolee, J. B.; and Lee, Y.-K. 2020. Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm. *IEEE Transactions on Affective Computing*, 13(2): 864–870.

Wheatley, M.; and Pabsch, A. 2010. Sign Language in Europe. In *Proceedings of sign-lang@LREC 2010*.

Wu, S.; Xiao, X.; Ding, Q.; Zhao, P.; Wei, Y.; and Huang, J. 2020. Adversarial sparse transformer for time series forecasting. *Advances in neural information processing systems*, 33: 17105–17115.

Zhao, G.; and Pietikainen, M. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6): 915–928.