RNN-driven integration of spatial, temporal, features for Indian sign language recognition and video captioning

Ajay Manohar Pol[1], Shrinivas A. Patil[2]
[1]Department of Electronics and Telecommunication, KIT's College of Engineering (Autonomous), Shivaji University, Kolhapur, India
[2]Department of Electronics and Telecommunication, DKTE Societies' Textile and Engineering Institute (Autonomous), Kolhapur, India

Article Info

Article history:

Received Apr 13, 2024
Revised Oct 30, 2024
Accepted Nov 11, 2024

Keywords:

BiLSTM
BLEU score evaluation
FFT-based feature extraction
Residual blocks
Sign language recognition

ABSTRACT

This paper presents a novel model that integrates spatial features from residual blocks and temporal features from FFT, alongside a sophisticated RNN architecture comprising BiLSTM, gated recurrent units (GRU) layers, and multi-head attention. Achieving nearly 99% accuracy on both WLASL and INCLUDE datasets, this model outperforms standard CNN pretrained models in feature extraction. Notably, the BiLSTM and GRU combination proves superior to other combinations such as LSTM and GRU. The BLEU score analysis further validates the model's efficacy, with scores of 0.51 and 0.54 on the WLASL and INCLUDE datasets, respectively. These results affirm the model's proficiency in capturing intricate spatial and temporal nuances inherent in sign language gestures, enhancing accessibility and communication for the deaf and hard-of-hearing communities. The comparison highlights the superiority of this paper's proposed model over standard approaches, emphasizing the significance of the integrated architecture. Continued refinement and optimization hold promise for further augmenting the model's performance and applicability in real-world scenarios, contributing to inclusive communication environments.

Corresponding Author:

Ajay Manohar Pol
Department of Electronics and Telecommunication, KIT's College of Engineering (Autonomous)
Shivaji University
Kolhapur, India
Email: kayajay2004@gmail.com

1. INTRODUCTION

For effective communication between normal and deaf people always hurdle is faced by normal people for interpretation of sign language as many remain unfamiliar with it. To minimize this gap of communication, technological advancement with the aid of sign language recognition (SLR) technique from video processing plays important role. As, deep learning are showing significant improvements and opening to new avenues, efficient SLR system development has become possible [1]. Gestures based SLR from video sequences is proposed in the work presented in this paper.

This study primarily focuses on the recognition of sign language gestures from video sequences. Video data is inherently dynamic and captures the temporal evolution of signs, presenting unique challenges compared to static image recognition. The objective is to create a resilient system that can precisely identify and interpret a broad spectrum of sign gestures into text. This endeavor holds great importance as it has the potential to enhance the accessibility of communication tools for the Deaf community. By pushing the boundaries of SLR in videos, this research strives to foster inclusivity, empower those with hearing impairments, and advocate for equal engagement in all facets of life [2]. The suggested system utilizes deep learning techniques to address the intricacies of SLR.

SLR, a crucial component of behavior identification, often harnesses machine learning, particularly delving into deep learning methodologies, necessitating considerable datasets for effective training. This process entails the intricate stages of gesture detection, tracking, and ultimate recognition, posing challenges that demand efficient feature extraction techniques.

In the realm of SLR, this study explores the generation of video-to-text descriptions, considering various approaches for processing both video and text data. Yousif and Al-Jammas [3] contribute a significant advancement with their state-of-the-art video captioning technique. Their method incorporates a deep reinforcement polishing network alongside word denoising and grammar checking networks, resulting in optimized performance, especially when handling lengthy video sequences. Rigorous evaluations underscore the efficacy of this approach, affirming its ability to accurately interpret and describe complex sign language gestures.

Another notable effort, as undertaken by Yasin *et al.* [4], strategically integrates word embedding techniques to enhance the focus on scene objects in sign language videos. Identifying similarities among words in descriptive text, this approach successfully captures the essence of the scenes depicted in sign language. The results obtained from this strategy are commendable, emphasizing the contextual information impacts. Also, for more effective video captioning within the context of SLR, sophisticated model architectures are important. SLR involves complex language representation. Xu et al.'s deep reinforcement learning and grammar checking networks advance video captioning, optimizing performance for sign language gestures. Yasin's use of word embedding techniques adds a linguistic perspective, enhancing scene object identification by considering context. These sophisticated model architectures reflect evolving machine learning trends, addressing multifaceted challenges.

Nabati and Behrad [5] introduced a state-of-the-art architecture characterized by parallel processing, LSTM networks, and iterative training methods reminiscent of AdaBoost. Through rigorous experimental tests, their approach showcases exceptional prowess, offering promising scalability, versatility, and enhanced text-image linkage facilitated by encoder-decoder models. The proposed model holds potential for broader applications, as it stands at the forefront of advancements in video captioning technology.

Chohan *et al.* [6] contributed to the field by exploring image captioning techniques employing encoder-decoder models and attention mechanisms. Their work includes a comprehensive analysis, suggesting diverse applications across various domains such as medical, industry, agriculture, and entertainment. The exploration of encoder-decoder models and attention mechanisms not only enhances the understanding of image captioning processes but also opens avenues for innovative applications in different sectors, underlining the versatility of their proposed methodology.

Mun *et al.* [7] proposed a novel video captioning method that places emphasis on temporal features and coherent feature matching. Leveraging reinforcement learning with event-oriented sequences for training, their approach better performance on the ActivityNet captions dataset. By prioritizing temporal dynamics and coherent feature alignment, Mun et al.'s method contributes significantly to the improvement of video captioning precision. The incorporation of reinforcement learning further underscores the adaptability of their model to diverse video datasets and scenarios.

Xiao and Shi [8] delve into the realm of video captioning through the lens of deep learning, specifically exploring the integration of CNN models using a generative adversarial approach. Their innovative approach suggests a significant departure from traditional methodologies, indicating the potential of adversarial techniques in enhancing the generation of descriptive video captions. The utilization of deep learning techniques, particularly in the context of generative adversarial networks, reflects a commitment to pushing the boundaries of video captioning technology.

Guo *et al.* [9] contribute to the field by incorporating a semantic guidance network, focusing on key frames during the training of target text descriptions. The inclusion of semantic guidance adds a layer of precision to the captioning process, demonstrating an understanding of the importance of context and semantic relevance in generating accurate and meaningful video captions. By emphasizing key frames, their approach aligns with the selective attention mechanisms critical for effective video description generation.

The strategy of updating model with addition of new data vectors for improvement of performance of the model, Fujii *et al.* [10] proposed a method. Particularly within the framework of encoder-decoder-based models in supervised learning the strategy is evaluated. This forward-thinking approach allows the model to adapt and evolve with the inclusion of new information, ensuring a continuous improvement in the accuracy and relevance of video captions. The emphasis on supervised learning further underscores the commitment to refining models through meticulous training and data utilization.

In the captioning of videos, Zhang *et al.* [11] introduced the cross-modal commonsense reasoning (CMCR) model. This model incorporates a cross-modal module, commonsense reasoning, and an innovative event refactoring mechanism. By combining these elements, the CMCR model offers a comprehensive approach to video captioning, addressing not only the modality challenges but also infusing commonsense reasoning for a more nuanced understanding of video content. The introduction of event refactoring further distinguishes their approach, showcasing a holistic strategy for improved performance.

In the work by Tateno *et al*. [12] devised a reference pyramid network integrated with ResNet 152 for precise word recognition in sign language videos. This innovative model places particular emphasis on harnessing the characteristics of object motion vectors, enhancing the accuracy of sign language recognition. By fusing the robust capabilities of ResNet 152 with the strategic architecture of a reference pyramid network, Liu et al. create a synergistic framework that adeptly captures and interprets the intricate dynamics of sign language gestures. This approach not only contributes to heightened recognition accuracy but also signifies a noteworthy advancement in the field of sign language video processing.

One notable contribution comes from Li *et al*. [13], who employed a deep learning model utilizing a temporal features-based approach. This innovative method involves extracting graph-based temporal features to process sign language videos, providing a more nuanced understanding of the temporal dynamics inherent in sign language communication. Li *et al*. [13], approach recognizes the importance of temporal information in sign language, where the sequencing and duration of gestures play a crucial role in conveying meaning. By incorporating graph-based temporal features into their deep learning model, they enhance the system's ability to capture and interpret the subtle intricacies of sign language movements over time. This not only improves the accuracy of sign language recognition but also enables the model to better handle variations in signing speed and rhythm.

Building upon this foundation, Mahyoub *et al*. [14] adopted a hybrid model that combines VGG16 with gated recurrent units (GRU) for both transformer encoder and decoder modules. This integration of convolutional neural networks (CNNs) with recurrent neural networks (RNNs) allows the model to effectively capture spatial features from video frames using VGG16 while leveraging the temporal dependencies encoded by GRU. The transformer architecture further refines the representation of sign language gestures, enhancing the overall performance of the recognition system.

Additionally, Xu *et al*. [15] introduced the use of ResNet as a kernel choice in their research on sign language recognition. ResNet, known for its deep residual learning capabilities [16], [17], offers a robust framework for capturing hierarchical features in complex datasets. By incorporating ResNet as a kernel, Xu *et al*. [15] enhance the model's ability to learn and represent intricate patterns in sign language videos, contributing to improved recognition accuracy.

This paper evaluates sign language recognition by extracting local and global (spatial and temporal) features using transfer learning. Various CNN models, including VGGNets, ResNets, Inception, DenseNet, and MobileNet, are compared. The main objective is to analyze the temporal dynamics' interpretation capability for sign language words. Motion vector features' significance in video processing is highlighted. The proposed model integrates spatial features from residual blocks with temporal features from fast fourier transform (FFT), capturing both local and global patterns to enhance accuracy and robustness in Indian Sign Language recognition.

- Enhanced RNN architecture: The paper introduces a sophisticated recurrent neural network (RNN) architecture comprising bidirectional long short-term memory (BiLSTM), GRU layers, and multi-head attention mechanism. This architecture enables the model to effectively capture temporal dependencies and contextual information from video frame sequences, facilitating more accurate and contextually relevant sign language interpretation.
- Superior performance and evaluation metrics: Through extensive experimentation on benchmark datasets such as WLASL and INCLUDE, the proposed model achieves exceptional accuracy, surpassing standard CNN pretrained models in feature extraction. Additionally, the model's performance is rigorously evaluated using BLEU score analysis, providing quantitative insights into the alignment between generated captions and human references in video captioning tasks. These contributions advance the state-of-the-art in sign language recognition and video captioning, fostering improved accessibility and communication for the deaf and hard-of-hearing communities.

Furhter in the article, section 2 shows the proposed methodology, section 3 covers the results and analysis and section 4 concludes the paper.

2. PROPOSED METHOD

The proposed model combines FFT and residual block-based CNN for video feature extraction alongside BiLSTM and GRU models for sign language interpretation. In the depicted block diagram Figure 1, input videos with paired sign language interpretations undergo processing. Videos are fed into the CNN model to extract local and global features from frames. Simultaneously, text data is tokenized for preprocessing. The extracted features and tokenized text vectors serve as inputs for the BiLSTM and GRU models [18]. During testing, video features are similarly extracted, and the model predicts sign language interpretations. Predictions are compared with ground truth for evaluation. This comprehensive approach leverages both visual and textual cues for accurate sign language recognition, promising robust performance across various scenarios.

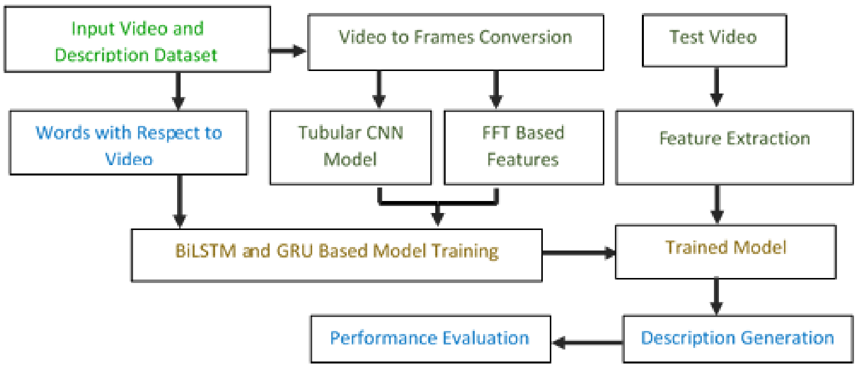

Input Video and Description Dataset
Video to Frames Conversion
Test Video
Words with Respect to Video
Tubular CNN Model
FFT Based Features
Feature Extraction
BiLSTM and GRU Based Model Training
Trained Model
Performance Evaluation
Description Generation

Figure 1. Stages involved in the proposed system framework

2.1. Video feature extraction

The proposed CNN architecture for extraction of features from video frames is shown in Figure 2. The model consists of 4 residual blocks in parallel to FFT features [19]. The combined features provide important features in terms of local and global. Let R_i be the output of residual block. Thus, in generalized view it can be represented as (1):

$$Ri = ResidualBlock\ (vi) \tag{1}$$

Where, v_i is the vector of pixels from i^{th} frame.
Similarly, in parallel branch of the model, the FFT features are extracted. The generalized equation for FFT features can be represented as (2):

$$F(vi) = FFT(vi) \tag{2}$$

The output obtained from residual block and FFT branch are then concatenated which combines both sptial and temporal features. Thus, combined features can be expressed as (3),

$$Combined\_Featuresi = Concatenate(R4, F(vi))\ or\ Combined\_Featuresi = Merge(R4, F(vi)) \tag{3}$$

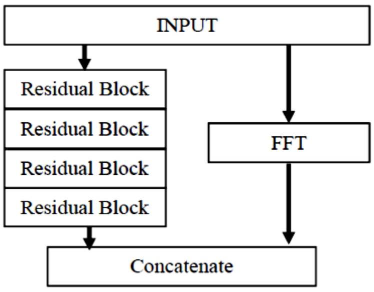


Figure 2. Video features extraction model

2.2. Training of model using video features and text

The training process involves a systematic approach to leverage video features and corresponding textual labels for SLR. This section outlines the main components and methodologies applied in training the model, which consists of three critical phases: feature extraction, temporal dependency modeling, and label prediction.

Input Video Features: A video sequence of sign language gestures is processed using a hybrid model combining CNN and FFT. This hybrid approach, represented as $f_{CNN-FFT}$, extracts meaningful spatial and frequency-domain features from the video, resulting in a feature matrix X. The combination of CNN and FFT allows the model to capture both spatial patterns and spectral information, which are essential for recognizing dynamic gestures in sign language.

BiLSTM and GRU Models: The extracted feature matrix X is fed into BiLSTM and GRU architectures to capture temporal dependencies in the video sequence. These architectures process the features bidirectionally to account for both past and future context, resulting in output representations denoted as H_{BiLSTM} and H_{GRU}, respectively:

$$H_{BiLSTM} = BiLSTM(X) \quad (4)$$

$$H_{GRU} = GRU(X) \quad (5)$$

These outputs serve as high-level temporal representations of the video data.

- SLR Labels: The extracted features and temporal representations are mapped to corresponding sign language labels, denoted as Y. These labels represent the ground truth for sign gestures in the video dataset.
- Training: The training phase involves optimizing model parameters for the CNN-FFT hybrid, BiLSTM, and GRU architectures jointly. A combined loss function, such as categorical cross-entropy, is minimized over the training dataset using stochastic gradient descent (SGD) or the Adam optimizer. The SLR predictions, denoted as $g(H_{BiLSTM}, H_{GRU}, Y)$, are compared against the ground truth labels to calculate the loss:

$$L = \sum_{n=1}^{N} \text{loss(Predictions, Ground Truth)} \quad (7)$$

The hybrid CNN-FFT model ensures robust feature extraction, while the BiLSTM and GRU layers effectively model temporal dependencies. The combined architecture is designed to maximize recognition accuracy by leveraging complementary strengths of spatial, spectral, and temporal modeling. N is the number of samples in the training dataset. Loss is the chosen loss function, such as categorical cross-entropy loss. Figure 3 shows the proposed model architecture.

The model is trained for 200 epochs using the Adam optimizer with a learning rate of 0.001. Each epoch involved processing video sequences to extract features using the CNN-FFT hybrid model, followed by temporal modelling through BiLSTM and GRU layers. The training minimized a combined categorical cross-entropy loss function over the predictions and ground truth labels. Batch size was set to 32, and early stopping monitored validation loss to prevent over fitting. Training achieved robust feature learning and effective temporal dependency modelling.

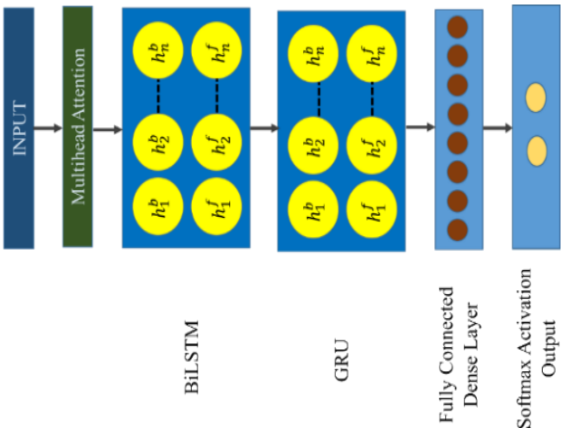


Figure 3. Proposed system architecture

3. RESULTS AND DISCUSSION

The evaluation of SLR performance involves preparing the dataset and employing various standard CNN models for extracting video features. The proposed model, integrating BiLSTM and GRU layers, is assessed using appropriate performance metrics.

3.1. Dataset preparation

The two-phase evaluation compares datasets for American sign language (ASL) and Indian sign language (ISL). The word-level American sign language (WLASL) dataset [20] showcases 2000+ ASL words by 100+ signers. The INCLUDE dataset [21], previously Indian lexicon sign language dataset, is tailored for ISL tasks, with 0.27 million frames across 4,287 videos. It covers 263 unique ISL signs categorized into 15-word groups, reflecting diverse linguistic concepts. This facilitates research in ISL recognition, offering a broad spectrum of ISL expressions and vocabulary.

3.2. Performance analysis

For the single n-gram, the BLEU score is calculated using (8).

$$BLEU_n = \frac{\min(c,r)}{\max(c,r)} \tag{8}$$

Where c and r the n-gram counts from inference output and reference ground truth words respectively. If their multiple n-grams, then weighted geometric mean is calculated as in (6) with maximum size N.

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N}(w_n \log(BLEU_n)\right) \tag{9}$$

In case of shorter candidate generations, BP is used to penalize. BP is the brevity penalty, given by (10).

$$BP = \left(1 - \frac{r}{c}\right) \tag{10}$$

Classification performance analysis utilizes accuracy, specificity, sensitivity, and F1 score parameters, detailed in Table 1. Video frame feature extraction employs standard CNN models. Figure 4 illustrates the comparative performance analysis of these models. This nuanced evaluation of SLR includes accuracy, specificity, sensitivity, and F1 score, crucial for identifying areas needing improvement. Low sensitivity for a gesture indicates recognition enhancement is necessary. Balancing these metrics ensures accuracy and effectiveness in capturing sign language nuances. Figures 5 and 6 show the comparative analysis of the proposed model with other standard CNN models-based feature extraction of video sequence and combinations of different types of layers in RNN.

Figures 5(a) and 5(b), Figueres 6(a) and 6(b) depict comparisons of pretrained models and attention-based models using WLASL and INCLUDE datasets. This work demonstrates the improvement in BLEU score with respect to combinations of the models in which proposed model shows highest performance over others. The comparative study with other existing methods demonstrated in the Table 2 shows that, the proposed model outperforms over the other methods.

Table 1. Classification performance parameters

Parameter	Formula
Accuracy	TP+TN/(TP+TN+FP+FN)
Specificity	TN/(TN+FP)
Sensitivity/Recall	TP/(TP+FN)
Precision	TP/(TP+FP)
F1 Score	2*(Recall*Precision)/(Recall Precision)

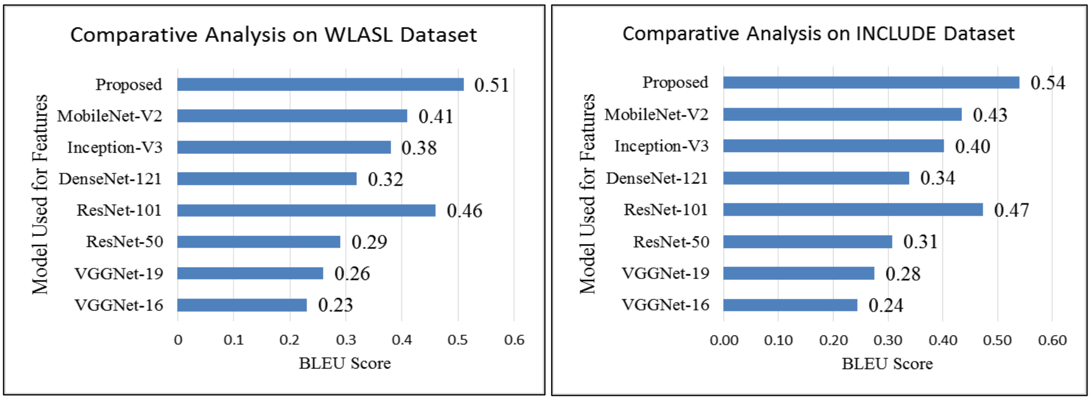
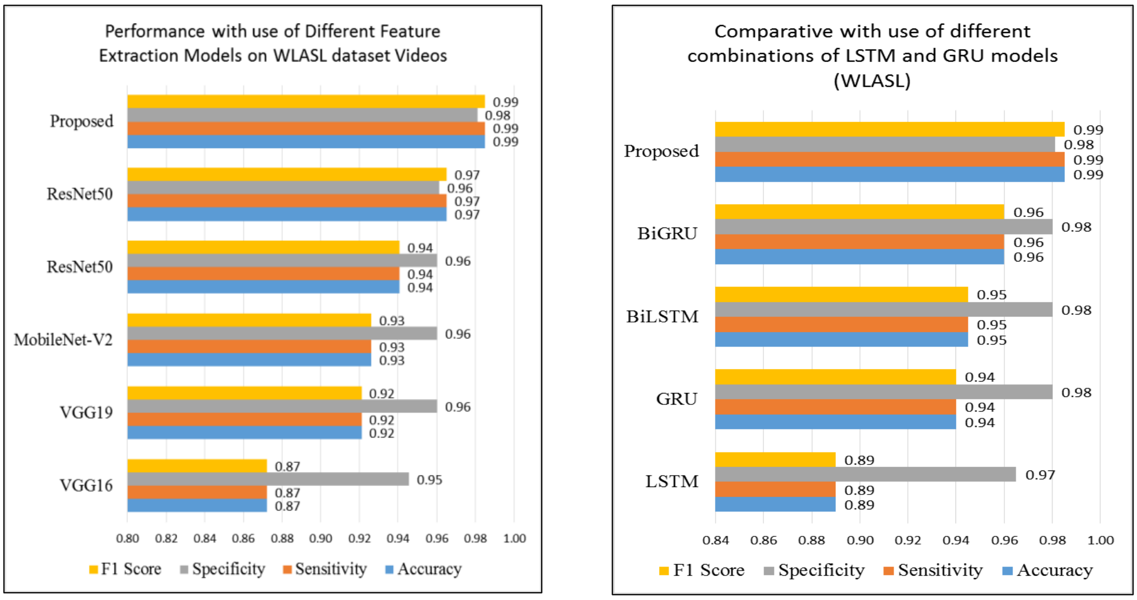
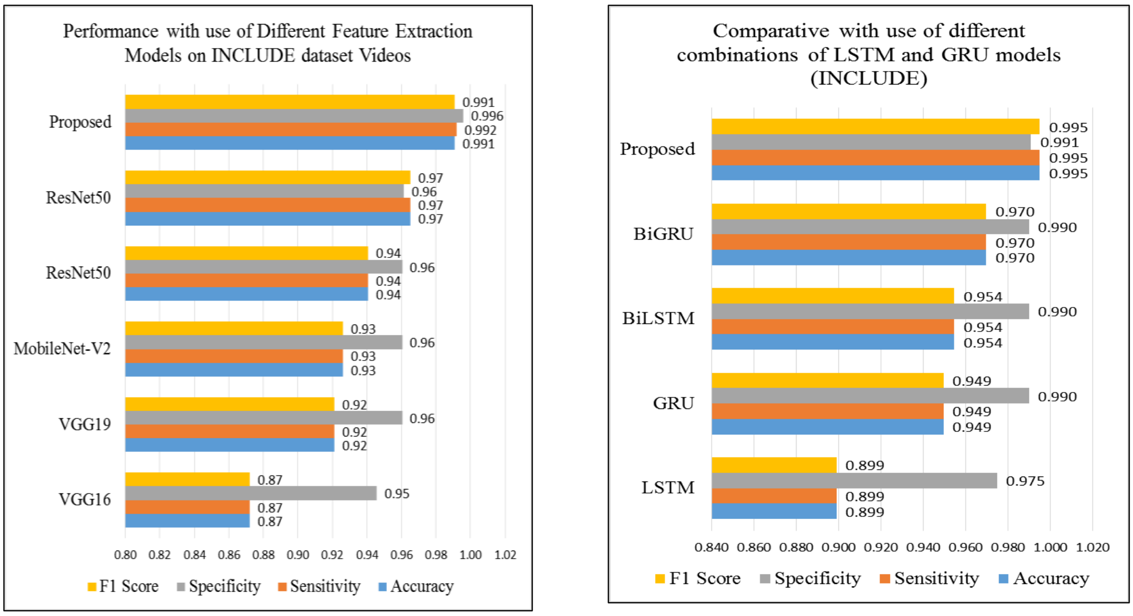


Figure 4. Average BLEU analysis for 10 videos



(a) (b)

Figure 5. Comparative analysis on WLASL dataset (a) different feature extraction models and (b) combinations of LSTM and GRU



(a) (b)

Figure 6. Comparative analysis on INCLUDE dataset and (a) different feature extraction models and (b) combinations of LSTM and GRU

Table 2. Comparative analysis with other existing methods

Method	Dataset	Performance
Customized design of CNN model [22]	ISL ROBITA Dataset	Accuracy 87.5%
Modifications in standard CNN model [23]	Baby Sign Language	Accuracy 89%
A module that extract global and local features (GLR) [9]	LSA64, INCLUDE and WLASL	Accuracy 91%
sign language translation network (SLTN) [24]	CSL-daily, SLR-100, RWTH	Accuracy 92%
Proposed	WLASL INCLUDE	Accuracy 99% Accuracy 99.1%

The proposed model, with 4 residual blocks and FFT for feature extraction plus BiLSTM and GRU for text processing, excels on WLASL and INCLUDE datasets. Residual blocks capture spatial features, FFT adds frequency data, and BiLSTM with GRU handles temporal dependencies, optimizing sign recognition. Dataset-specific tuning enhances accuracy, achieving 99% on WLASL and 99.1% on INCLUDE, significantly outperforming previous methods with accuracies between 87.5% and 92%, marking a notable improvement across datasets.

The proposed model significantly outperforms existing approaches, such as the customized CNN design [16], which achieved 87.5% accuracy on the ISL ROBITA dataset, and the Modified CNN Model [25], which reached 89% accuracy on the Baby Sign Language dataset. The proposed model's advanced feature extraction and attention mechanisms lead to a substantial performance boost, surpassing the Global and local feature extraction (GLR) Module [17], which attained 91% accuracy on datasets like LSA64, INCLUDE, and WLASL. Moreover, the model outperforms the state-of-the-art sign language translation network (SLTN) [19], which achieved 92% accuracy, by effectively capturing and translating sign language, as evidenced by the proposed model's superior accuracy of 99-99.1% on the WLASL and INCLUDE datasets. While the proposed model demonstrates exceptional performance on the WLASL and INCLUDE datasets, further testing across additional datasets would help validate its robustness and generalizability. Additionally, assessing the model's computational efficiency and inference speed is crucial for ensuring its suitability for real-time applications, particularly in resource-constrained environments.

4. CONCLUSION

The proposed model, integrating spatial features from residual blocks and temporal features from FFT, along with a sophisticated RNN architecture comprising BiLSTM, GRU layers, and multi-head attention, demonstrates exceptional performance. Achieving nearly 99% accuracy on both the WLASL and INCLUDE datasets, this model outperforms standard CNN pretrained models in feature extraction. Notably, the BiLSTM and GRU combination proves superior to other combinations such as LSTM and GRU. The BLEU score analysis further validates the model's efficacy, with scores of 0.51 and 0.54 on the WLASL and INCLUDE datasets, respectively. These results affirm the model's proficiency in capturing intricate spatial and temporal nuances inherent in sign language gestures, thereby enhancing accessibility and communication for the deaf and hard-of-hearing communities. The comparison highlights the superiority of our model over standard approaches, emphasizing the significance of the integrated architecture. Moving forward, continued refinement and optimization hold promise for further augmenting the model's performance and applicability in real-world scenarios. The findings underscore the potential of our approach to advance SLR and video captioning technologies, contributing to inclusive communication environments.

REFERENCES

[1] M. Li, Y. Jiang, Y. Zhang, and H. Zhu, "Medical image analysis using deep learning algorithms," *Front. Public Heal.*, vol. 11, 2023, doi: 10.3389/FPUBH.2023.1273253.

[2] M. Papatsimouli, P. Sarigiannidis, and G. F. Fragulis, "A Survey of Advancements in Real-Time Sign Language Translators: Integration with IoT Technology," *Technol. 2023, Vol. 11, Page 83*, vol. 11, no. 4, p. 83, Jun. 2023, doi: 10.3390/TECHNOLOGIES11040083.

[3] A. J. Yousif and M. H. Al-Jammas, "Exploring deep learning approaches for video captioning: A comprehensive review," *e-Prime - Adv. Electr. Eng. Electron. Energy*, vol. 6, p. 100372, Dec. 2023, doi: 10.1016/J.PRIME.2023.100372.

[4] D. Yasin, A. Sohail, and I. Siddiqi, "Semantic Video Retrieval using Deep Learning Techniques," *Proc. 2020 17th Int. Bhurban Conf. Appl. Sci. Technol. IBCAST 2020*, pp. 338–343, Jan. 2020, doi: 10.1109/IBCAST47879.2020.9044601.

[5] M. Nabati and A. Behrad, "Video captioning using boosted and parallel Long Short-Term Memory networks," *Comput. Vis. Image Underst.*, vol. 190, p. 102840, Jan. 2020, doi: 10.1016/J.CVIU.2019.102840.

[6] M. Chohan, A. Khan, M. S. Mahar, S. Hassan, A. Ghafoor, and M. Khan, "Image Captioning using Deep Learning: A Systematic Literature Review," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 5, pp. 278–286, 2020, doi: 10.14569/IJACSA.2020.0110537.

[7] J. Mun, L. Yang, Z. Ren, N. Xu, and B. Han, "Streamlined Dense Video Captioning," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 6581–6590, Apr. 2019, doi: 10.1109/CVPR.2019.00675.

[8] H. Xiao and J. Shi, "Video captioning with text-based dynamic attention and step-by-step learning," *Pattern Recognit. Lett.*, vol. 133, pp. 305–312, May 2020, doi: 10.1016/J.PATREC.2020.03.001.

[9] Z. Guo, Y. Hou, and W. Li, "Sign language recognition via dimensional global–local shift and cross-scale aggregation," *Neural Comput. Appl.*, pp. 1–13, Mar. 2023, doi: 10.1007/S00521-023-08380-9/METRICS.

[10] T. Fujii, Y. Sei, Y. Tahara, R. Orihara, and A. Ohsuga, "'Never fry carrots without cutting.' Cooking Recipe Generation from Videos Using Deep Learning Considering Previous Process," *Proc. - 2019 IEEE/ACIS 4th Int. Conf. Big Data, Cloud Comput. Data Sci. BCD 2019*, pp. 124–129, May 2019, doi: 10.1109/BCD.2019.8885222.

[11] X. Zhang, F. Zhang, and C. Xu, "Explicit Cross-Modal Representation Learning for Visual Commonsense Reasoning," *IEEE Trans. Multimed.*, vol. 24, pp. 2986–2997, 2022, doi: 10.1109/TMM.2021.3091882.

[12] S. Tateno, H. Liu, and J. Ou, "Development of Sign Language Motion Recognition System for Hearing-Impaired People Using Electromyography Signal," *Sensors (Basel).*, vol. 20, no. 20, pp. 1–22, Oct. 2020, doi: 10.3390/S20205807.

[13] D. Li, C. R. Opazo, X. Yu, and H. Li, "Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison," *Proc. - 2020 IEEE Winter Conf. Appl. Comput. Vision, WACV 2020*, pp. 1448–1458, Oct. 2019, doi: 10.1109/WACV45572.2020.9093512.

[14] M. Mahyoub, F. Natalia, S. Sudirman, and J. Mustafina, "Sign Language Recognition using Deep Learning," *Proc. - Int. Conf. Dev. eSystems Eng. DeSE*, vol. 2023-January, pp. 184–189, 2023, doi: 10.1109/DESE58274.2023.10100055.

[15] X. Xu, K. Meng, C. Chen, and L. Lu, "Isolated Word Sign Language Recognition Based on Improved SKResNet-TCN Network," *J. Sensors*, vol. 2023, pp. 1–10, Jul. 2023, doi: 10.1155/2023/9503961.

[16] J. Rastus Shane and V. Vanitha, "Sign Language Detection Using Faster RCNN Resnet," *2nd Int. Conf. Adv. Electr. Electron. Commun. Comput. Autom. ICAECA 2023*, 2023, doi: 10.1109/ICAECA56562.2023.10200987.

[17] S. Wang, K. Wang, T. Yang, Y. Li, and D. Fan, "Improved 3D-ResNet sign language recognition algorithm with enhanced hand features," *Sci. Reports 2022 121*, vol. 12, no. 1, pp. 1–19, Oct. 2022, doi: 10.1038/s41598-022-21636-z.

[18] S. Vashisht, P. Kumar, and M. C. Trivedi, "Enhanced GRU-BiLSTM Technique for Crop Yield Prediction," *Multimed. Tools Appl.*, pp. 1–26, Mar. 2024, doi: 10.1007/S11042-024-18898-2/METRICS.

[19] N. N. H. Van, P. H. Do, V. N. Hoang, A. Borodko, and T. D. Le, "Leveraging FFT and Hybrid EfficientNet for Enhanced Action Recognition in Video Sequences," *ACM Int. Conf. Proceeding Ser.*, pp. 32–39, Dec. 2023, doi: 10.1145/3628797.3628827.

[20] D. Li, C. R. Opazo, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," *Proc. - 2020 IEEE Winter Conf. Appl. Comput. Vision, WACV 2020*, pp. 1448–1458, Mar. 2020, doi: 10.1109/WACV45572.2020.9093512.

[21] A. Sridhar, R. G. Ganesan, P. Kumar, and M. Khapra, "INCLUDE: A Large Scale Dataset for Indian Sign Language Recognition," *MM 2020 - Proc. 28th ACM Int. Conf. Multimed.*, pp. 1366–1375, Oct. 2020, doi: 10.1145/3394171.3413528.

[22] G. Arun Prasath and K. Annapurani, "Prediction of sign language recognition based on multi layered CNN," *Multimed. Tools Appl.*, pp. 1–21, Mar. 2023, doi: 10.1007/S11042-023-14548-1/METRICS.

[23] V. Enireddy, J. Anitha, N. Mahendra, and G. Kishore, "An optimized automated recognition of infant sign language using enhanced convolution neural network and deep LSTM," *Multimed. Tools Appl.*, pp. 1–23, Feb. 2023, doi: 10.1007/S11042-023-14428-8/METRICS.

[24] R. Li and L. Meng, "Sign language recognition and translation network based on multi-view data," *Appl. Intell.*, vol. 52, no. 13, pp. 14624–14638, Jul. 2022, doi: 10.1007/S10489-022-03407-5/METRICS.

[25] H. Chao, W. Fenhua, and Z. Ran, "Sign language recognition based on CBAM-RESNET," *ACM Int. Conf. Proceeding Ser.*, Oct. 2019, doi: 10.1145/3358331.3358379.

BIOGRAPHIES OF AUTHORS

Ajay Manohar Pol received the Bachelor of Engineering in Electronics Engineering from Shivaji University Kolhapur, India and holds a M. Tech. degree in Electronics Engineering with specialization in digital systems from Savitribai Phule Pune University Pune, India. Currently he is working as an Assistant Professor at Kolhapur Institute Technology's (KIT's) College of Engineering, Kolhapur, Shivaji University of Kolhapur, India. He is having 17 years of teaching. His research areas are image/signal processing, image analysis, pattern recognition, and artificial intelligence and machine learning. He used to hold administrative posts as Deputy Registrar of Examination and Evaluation in KIT's College of Engineering, Shivaji University, Kolhapur, India, from 2022 to present. He can be contacted at email: kayajay2004@gmail.com, pol.ajay@kitcoek.in.

Shrinivas A. Patil received the B. E. degree in Electronics Engineering from KIT's, College of Engineering, Kolhapur, M. Tech. in Bio-Medical Engineering. from IIT, Bombay, and PhD in Electronics from Shivaji University, Kolhapur. Currently he is working as a Professor and Head of Electronics and Telecommunication Engineering. at DKTE's, Textile & Engineering. Institute, Ichalkaranji, Maharashtra State, India. He is having 33 years of teaching and one-year industrial experience. He has supervised and co-supervised more than 15 M. Tech. and 8 Ph.D. students. He has authored, coauthored, and presented more than 80 publications in peer reviewed journals and international conferences. His research interests include image processing, artificial intelligence and machine learning, embedded and VLSI system design. He can be contacted at email: sapatil@dkte.ac.in.