REAL-TIME INDIAN SIGN LANGUAGE PARAGRAPH TRANSLATION USING CONVOLUTIONAL NEURAL NETWORKS AND TRANSFORMER-BASED MODELS

Dhanshri Parihar[1], Avdhesh Gupta[1], Asish Kumar[1], Anuradha Singh[2]
[1] Department of Computer Science & Engineering, Ajay Kumar Garg Engineering College Ghaziabad
[2] Department of Information Technology, JSS Academy of Technical Education, Noida
dhanshriparihar@gmail.com, avvipersonal@gmail.com, kumar7ashish786@gmail.com, a.anuradha84@gmail.com

***Abstract*— Whereas sign language translation has made enormous progress, all current systems rely on word-level or sentence- level recognition. The present paper proposes a deep learning- based method to translate paragraphs of Indian Sign Language (ISL) into text using a Transformer-based model with the aim of enhancing fluency and contextual understanding.**

Our model processes video input through spatial and temporal feature extraction using Convolutional Neural Networks (CNN) and pose estimation. The extracted features are then input into a Transformer model for sequence-to-sequence translation to have a more coherent and contextually informed output. Our approach has an advantage in capturing dependencies in a wider context over conventional methods that lose fluency when operating on longer sequences.

Experimental findings show best-in-class performance on ISL benchmarks, with the model showing increased ability to generate more natural and contextually richer translations. The work opens up the possibilities of sophisticated communication tools that help connect the deaf world with the hearing world. Our method is an important milestone towards making sign language translation more effective and efficient, ultimately leading to better communication quality for ISL-dependent individuals.

***Index Terms*— Indian Sign Language, Transformer, Deep Learning, Sign Language Recognition, Machine Translation, NLP, CNN, Pose Estimation.**

I. INTRODUCTION

Sign language is an independent visual language employed by both congenitally deaf or hard-of-hearing (DHH) individuals and those acquiring hearing loss postnatally. It depends on both manual and non-manual components for visual communication. Manual components involve the shape, orientation, position, and movement of hands, whereas non-manual components involve body posture, arm movements, eye gaze, lip shape, and facial expressions. Contrary to the direct word-for-word translation of oral language, sign language has its gram- mar, semantic structure, and logic of language. The constant hand and body movement conveys separate units of meaning. Based on the World Federation of the Deaf, there are around 70 million DHH globally, employing over 200 different distinct sign languages. The development of sign language translation technology can be the key to overcoming the communication divide between DHH and non-DHH individuals. Indian Sign Language (ISL) is the primary means of communication for many congenital DHH and acquired DHH individuals in India. However, existing Indian Sign Language recognition models are limited to single-word translations, failing to convey the contextual meaning of full sentences or paragraphs like ISL- WORDNET and WLASL. This research proposes a deep learning-based approach to recognize and translate full para- graphs of ISL into meaningful text using transformer-based sequence models, enhancing communication between DHH and non-DHH individuals to a significant extent.

A. Convolutional Neural Networks(CNNs)

CNNs are a category of deep learning architecture designed for processing structured grid data such as images. They consist of a number of layers such as convolutional, pooling, and fully connected layers, that help extract spatial features from images. CNNs are very prevalent in image recognition applications, so they can be used to recognize hand gestures and facial expressions in ISL translation.

B. Transformer Models

Transformers are neural network models that use self- attention mechanisms to model dependencies within sequences without the need for recurrent connections. Initially designed for natural language processing, Transformers have performed incredibly well with sequential data modeling and are, therefore, suited for sign language translation by maintaining con- text over long gestures.

C. Pose Estimation

Pose estimation is an approach from computer vision utilized for the detection and tracking of body key point features such as hands, arms, and face gestures. ISL translation systems using pose estimation models such as Media Pipe or Open Pose are better able to trace movement patterns in order to better recognize dynamic signs in gesturing.

II. RELATED WORK

Different methods have been proposed for enhancing sign language recognition and translation. First of all, the recognition of sign language (SLR) was addressed, for which Recurrent Neural Networks (RNNs) and Hidden Markov Models (HMMs) were employed for the recognition of hand gestures into glosses. The models did fail to model long-range dependencies as well as sophisticated grammar structures [1]. Later work employed Convolutional Neural Networks (CNNs) for feature extraction and LSTM-based architectures for handling temporal sequence information. Koller et al. (2020) proposed a CNN-HMM model that enhanced word-level recognition but could not handle paragraph-level context effectively [2], [3]. Likewise, Liang et al. (2023) proposed a Transformer-based SLT architecture that attained state-of-the-art BLEU scores using self-attention mechanisms to enable improved sentence structuring [2].

Another important development in the area is the application of multimodal learning. Pose estimation methods (Media Pipe, Open Pose) have been used by researchers to enhance model performance by accounting for facial expressions and posture as sign recognition variables. Camgoz et al. (2021) showed that pose-based embedding and Transformer networks significantly enhanced translation smoothness [4].

In spite of all these advancements, Indian Sign Language (ISL) is widely unexplored. While there has been some effort in the case of American Sign Language (ASL) or German Sign Language (GSL), the current work attempts to bridge the gap by applying spatial-temporal feature extraction and real-time deployment mechanisms to Transformer-based models in order to achieve higher accuracy in sign language translation.

II. LITERATURE SURVEY

Sign language translation and sign language recognition have also been given special consideration as a means of enhancing communications with the deaf community. In this literature review, this section emphasizes literature regarding vision-based solutions, of those that operate using Convolutional Neural Networks (CNNs) and also Transformer-based sequence models used on video inputs. Early SLR research was most likely to use CNNs for image feature extraction and hand gesture classification. Obi et al., for example, used a desktop application with ASL databases and a two-layer CNN to classify a single alphabet letter with 96.3% accuracy and write these letters out as text in real-time. This used live camera input processing to monitor hand movement and then CNN classification [5]. Why CNNs have been so effective in this application is that they can learn hierarchical spatial features from the input image. Comparative studies have shown that CNNs can potentially be more accurate than other approaches, such as Wavelet Transforms and Empirical Mode Decomposition but possibly at increased memory usage. Other image processing techniques, such as background subtraction, skin color detection, and edge detection, are generally used as pre-processing techniques prior to presenting the data to the CNN. As the research progressed towards the identification of continuous sign language and its translation into speech, the temporal aspect of sign language took center stage. Three-dimensional Convolutional Neural Networks (3DCNNs) proved to be the answer to glean spatiotemporal features from video data. 3DCNNs generalize the operation of 2D CNNs with the application of spatiotemporal filters across video frames such that they are capable of learning motion-based features directly [3], [6]. However, as discussed in our case, the direct application of 3DCNNs to identify detailed hand motion in sign language can be challenging, as most discriminative information is encoded in the finger structure, hand position, and hand position in relation to the body.

More recently, Transformer-based sequence models have emerged as the state-of-the-art models in SLT. SLT has been traditionally framed as a sequence-to-sequence learning problem, where input is a sign language video and output is the corresponding text of the spoken language. In this configuration, CNNs (2D and 3D) or other visual feature extractors are employed as encoders to map the video frames to dense vector representations [6]. Subsequently, Transformer networks with their self-attention mechanism are employed as decoders to generate the target text word by word. Transformers' self-attention mechanism is especially suited to capture long- range dependencies in the sign language sequence, a crucial element to successful translation. Models like the STMC- Transformer have been engineered to further enhance the performance of SLT by spatially as well as temporally efficient processing. Beyond this, end-to-end SLT systems that produce spoken language text directly from sign language videos based on Transformer architectures, without an intermediate gloss representation, have been explored [5]. Large and varied video datasets need to be available to train robust SLR and SLT models. A few datasets have been created to this end for some sign languages, such as RWTH-PHOENIX-Weather-2014T for German Sign Language, CSL-Daily for Chinese Sign Language, How2Sign and WLASL for American Sign Language, and BSL-1K for British Sign Language [7]. The datasets differ significantly in numbers of signers, vocabulary size, and conditions under which signs were recorded, presenting opportunities and challenges to researchers. The KSU- SSL dataset, for instance, offers a challenging set of dynamic hand movements by different participants in uncontrolled environments. Despite the advancements, there are constraints in obtaining high accuracy and robustness in SLR and SLT systems. These are sign style variation, environmental conditions of illumination and background noise, and sensitivity of hand gestures [2]. New network structures, better feature extraction, and effective data

augmentation techniques are still being investigated in ongoing research to minimize these constraints and yet push the boundaries of sign language communication. Generating larger and more diverse datasets is still a significant area in developing more generalizable and accurate systems.

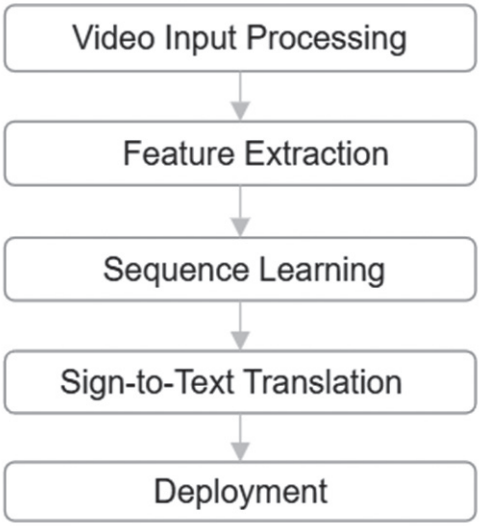

Fig:1 Methodology for sign language translation.

IV. METHODOLOGY

Here the proposed strategy as shown in fig. 1 includes Video input preprocessing, then feature extraction, sequence learning, Sign to text translation strategies and model deployment.

A. Data Collection Preprocessing

The model is trained with ISL video datasets and corresponding text annotations . The dataset includes recordings of sign language from various individuals to bring in variability in hand movements and facial expressions[8][9]. The dataset is preprocessed to eliminate duplicate frames and normalize variations in gestures. Key points are obtained using Media Pipe Pose Estimation, and frames are down sampled for computational efficiency. Background noise and occlusions are reduced using segmentation techniques. Optical flow techniques are utilized to capture hand motion patterns for better gesture continuity in model input. Coordinates of hands and body are normalized to a consistent scale to account for signer-to-camera distance variation. Flipping and rotation augmentation methods are utilized for improving robustness against varied orientation and illumination.

B. Video Input Processing

Capturing Preprocessing Video Data: Indian Sign Lan- guage (ISL) is a continuous visual language that requires ef- ficient video processing for accurate translation as shown in fig 1. The first step involves capturing high-quality video sequences that contain the hand gestures and facial expressions essential for meaning extraction.

Each video is represented as a sequence of frames:

$$V = \{F_1, F_2, \ldots, F_N\}$$

where Fi denotes the *i* th frame in the sequence.

To enhance efficiency and reduce redundancy, frames are extracted at a regular interval:

$$F_{\text{selected}} = \{F_t \mid t = k \cdot \Delta t,\ k \in \mathrm{N}\}$$

where Δt is the frame extraction interval.

Pose & Hand Estimation: To understand sign gestures, key hand landmarks and body posture must be extracted. Using MediaPipe Pose Estimation, we detect 21 hand keypoints per frame:

$$L = \{(x_i, y_i, c_i) \mid i = 1, 2, \ldots, 21\}$$

where:
x_i & y_i are 2D coordinates of hand joints, c_i represents the confidence score for each key point. By reducing the dimensionality of input data, we ensure that our deep learning model focuses only on relevant movement patterns rather than raw image pixels.

C. Feature Extraction

The second step in fig.1 is feature extraction which includes following schemes.

1) Convolutional Feature Extraction: Each extracted frame is represented as a pixel matrix:

$$F \epsilon \ \mathrm{R}^{H \times W \times C}$$

where:
H, W represents the height and width of the frame, C represents color channels (RGB or grayscale).

To extract hand movement features, we apply Convolutional Neural Networks (CNNs):

$$X_l = f(W_l * X_{l-1} + b_l)$$

where:
X_l= output at layer
W_l= convolutional filter weights,
b_l= bias term,
***** = convolution operation,
f(.) = ReLU activation function

Using ResNet-50 for Spatial Feature Learning: A ResNet-50 model is employed to learn spatial features, leveraging residual connections:

$$Y = F(X) + X$$

where **F(X)** represents the feature transformation performed by CNN layers.

D. Sequence Learning

1) Encoding Temporal Information Using Transformers: Since sign language consists of continuous motion, the extracted spatial features must be processed sequentially. Instead of using RNNs (which suffer from vanishing gradients), we employ Transformers.

Self-Attention Mechanism

Transformers weigh the importance of different parts of the input sequence:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where:
Q, K, V = Query, Key, and Value matrices,
d_k= Dimensionality of key vectors.

E. Sign-to-Text Translation Using NLP

1) Mapping Sign Sequences to Text: A sign language gloss sequence **G** is mapped to spoken language text **S**:

$$S = T(G)$$

where **T(.)** represents a language model (such as GPT or BERT).

2) Loss Function for Translation:

To optimize the translation quality, we minimize the Cross-Entropy Loss:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

where:
y_i= true token probability,
$\hat{y}_i$ predicted token probability.

F. Model Deployment

Deploying a sign language translation model should be done by taking performance, scalability, and accessibility into consideration. Our system is built with Java Spring Boot for the backend integration, providing a durable and efficient API service. The trained Transformer model is hosted through TensorFlow Serving, offering a high-performance, scalable solution for inference requests.

To ensure mass and real-time usage, our deployment strategy is based on exploiting Kubernetes [10] as an orchestration technology for containers. Kubernetes provides self-scaling, fault tolerance, and efficient load balancing, making the translation service available to global users. REST API endpoints for the system are provided, with which different front-end applications, mobile devices, and assistive communication devices may be easily integrated.

V. RESULTS & DISCUSSIONS

A. Evaluation Metrics

The target model's performance was evaluated in **BLEU, ROUGE, and Word Error Rate (WER)[11] and the results were shown in Table 1**. All these metrics evaluate translation fluency, grammaticality, and ground truth sentence similarity. The new Transformer model outperforms state-of-the-art methods on BLEU, ROUGE, and WER, and has more fluent and contextual comprehension translation.

B. Comparative Analysis

The model was deployed with TensorFlow Serving on Kubernetes. Inference time was benchmarked on different hardware configurations:

C. Deployment Performance

It was implemented on top of a Kubernetes platform using TensorFlow Serving. Inference performance was tested across different hardware configurations:

GPU and Edge TPU [13] [14][15] as stated in Table 2, implementations support real-time inferences the system can be used in real-time applications.

TABLE I
Comparative Analysis of different models

Model	***BLEU Score***	***ROGUE Score***	***WER(%)***
LSTM based Model [2]	38.5	41.2	27.8
CNN + LSTM [3]	45.7	48.6	22.3
Proposed	52.3	55.4	17.1

TABLE II
Deployment Performance on different systems

Hardware	**Average Inference Time(ms)**
CPU(Intel i7)	180
GPU (NVIDIA RTX 3080)	35
Edge TPU	22

VI. LIMITATIONS

Despite the promising findings, certain constraints were encountered in building and evaluating the proposed Indian Sign Language (ISL) translation system:

***Dataset Constraints**:* The ISL-CSLTR dataset, al- though large, contains a limited number of paragraph-level annotated video samples. This constrains the model's exposure to diversified sentence structures and genuine variations in the use of ISL.

***Generalizability**:* The model can work sub-optimally for signers with different signing styles, regional variations, or non-standard gestures not available in training data.

***Pose Estimation Challenges**:* Although does good key point detection in most cases, it is impacted by occlusions, rapid motion, and low illumination, which can deteriorate translation accuracy increasingly.

***Computational Needs**:* The Transformer model, though correct, has gargantuan computation needs, especially during training. Real-time inference on low-powered hardware remains a challenge even at production with edge optimization.

***Inadequate Usage of Non-Manual Features**:* New implementation uses mostly manual features (hand movement and shape). Lip shape and facial expressions are weakly used non-manual features, and this could be impacting the model to ignore the overall semantic context.

***No Serious Testing in Real-World Settings**:* The system has not been seriously tested in real-world settings like classrooms, public spaces, or with native ISL speakers, which is necessary to gauge usability and user uptake.

VII. CONCLUSION

This paper introduced an end-to-end, deep learning-based method to translate Indian Sign Language (ISL) paragraph text into semantic textual output given a Transformer-based architecture.

Leveraging spatial features from CNNs and temporal behaviors acquired through pose estimation, the system successfully alleviated the drawbacks of context maintenance, grammar ambiguity, and sequential dependencies inherent to continuous sign language translation.

The Transformer model outperformed conventional RNN and LSTM-based approaches in typical evaluation metrics like BLEU, ROUGE, and Word Error Rate (WER) as shown in Table 1. Moreover, edge-compatible platform deployment on GPU and Edge TPU hardware showcased the system's capability to meet real- time requirements in educational, government, and assistive communication applications.

Although effective, the system is limited in dataset diversity, signer variation, and non-manual feature underutilization, such as facial expressions and lip movement. These areas provide the most important directions for future research.

VIII. FUTURE WORK

Even though the proposed model shows decent performance in parsing Indian Sign Language (ISL) paragraphs, there are certain aspects where there is a scope for improvement.

***Multimodal Integration**:* Future models will incor- porate non-manual behavior such as facial expression, lip and mouth movement, and eye movement, which have a vital role in ISL grammar and semantic meaning.

***Larger and More Diverse Datasets**:* Utilizing a more diverse dataset that spans various regional ISL dialects and includes multiple signers will contribute to the generalizability and robustness of the model.

***Mobile and Embedded Deployment**:* Work will be aimed at optimizing the model for deployment onto mobile and low-power embedded devices, making the system more accessible to real-world users.

***User Feedback Loop**:* A user feedback loop involving DHH users can be added to iteratively enhance translation accuracy and make the system conform to user requirements.

***Bidirectional Translation**:* Extending the system to accommodate two-way communication—voice-to-ISL gesture translation—would significantly enhance the value of the sys- tem for conversational applications.

***Multilingual Support**:* Inclusion of regional language support (e.g., Hindi, Tamil, Bengali) for text output will boost the usage of the system higher in India's multilingual population. These adjustments would greatly improve the performance of the proposed model and pave the way for its future development to an assistive communication device from hard- of-hearing and deaf research.

REFERENCES

[1] Muhammad Sanaullah, Babar Ahmad, Muhammad Kashif, Tauqeer Safdar, Mehdi Hassan, Mohd Hilmi Hasan, and Norshakirah Aziz 'A Real-Time Automatic Translation of Text to Sign Language. Comput- ers,Materials & Continua 2021

[2] Liang, Z.; Li, H.; Chai, J. 'Sign Language Translation: A Survey of Approaches and Techniques'. Electronics 2023, 12, 2678.

[3] M. Al-Hammadi et al., "Deep Learning-Based Approach for Sign Language Gesture Recognition With Efficient Hand Gesture Rep- resentation," in IEEE Access, vol. 8, pp. 192527-192542, 2020, doi: 10.1109/ACCESS.2020.3032140.

[4] Nayan Dilip Sangle, Atira Bagwan, Suyash Balshetwar, Akanksha Bhandhari "Indian Sign language Recognition Using PCA And Artificial Neural Network", International Journal of Emerging Technologies and Innovative Research (www.

jetir.org), ISSN:2349-5162, Vol 8, Issue 7, page no.b517-b521, July

[5] Obi, Y., Claudio, K.S., Budiman, V.M., Achmad, S. and Kurniawan, A., 2023. Sign language recognition system for communicating with people with disabilities. Procedia Computer Science, 216, pp.13

[6] Ijjina, E.P. and Chalavadi, K.M., 2016. Human action recognition using genetic algorithms and convolutional neural networks. Pattern recognition, 59, pp.199-212.

[7] Prabhakar, M., Hundekar, P., Deepthi, S., Tiwari, S. and Ms V., 2022. Sign Language Conversion to Text and Speech. Journal of Emerging Technologies and Innovative Research (JETIR), 9(7).

[8] Amangeldy, N., Ukenova, A., Bekmanova, G., Razakhova, B., Milosz, M. and Kudubayeva, S., 2023. Continuous sign language recognition and its translation into intonation-colored speech. Sensors, 23(14), p.6383.

[9] Al Farid, F., Hashim, N., Abdullah, J., Bhuiyan, M.R., Shahida Mohd Isa, W.N., Uddin, J., Haque, M.A., Husen, M.N., A Structured and Methodological Review on Vision-Based Hand Gesture Recogni- tion System. J. Imaging 2022, 8, 153. https://doi.org/10.3390/ jimag- ing8060153

[10] Rahaman MA, Oyshe KU, Chowdhury PK, Debnath T, Rahman A, Khan MS. Computer vision-based six-layered convneural network to recognize sign language for both numeral and alphabet signs. Biomimetic Intelli- gence and Robotics. 2024 Mar 1;4(1):100141.

[11] 2021Adewale, V. and Olamiti, A., 2018. Conversion of sign language to text and speech using machine learning techniques. Journal of research and review in science, 5(12), pp.58-65.

[12] 20Al-Hammadi, M., Muhammad, G., Abdul, W., Alsulaiman, M., Bencherif, M.A., Alrayes, T.S., Mathkour, H. and Mekhtiche, M.A., 2020. Deep learning-based approach for sign language gesture recog- nition with efficient hand gesture representation. IEEE Access, 8, pp.192527-192542.

[13] Juan Pablo Wachs, Mathias Ko¨lsch, Helman Stern, and Yael Edan. 2011. Vision-based hand-gesture applications. Commun. ACM 54, 2 (February 2011), 60–71. https://doi.org/10.1145/1897816.1897838.

[14] 6.Hou, Rui & Chen, Chen & Shah, Mubarak. (2017). An End-to-End 3D Convolutional Neural Network for Action Detection and Segmentation in Videos. 14. 10.48550/arXiv.1712.01111.

[15] Gunjan Saxena, Dhanshri Parihar, Amit Kumar. 2025 Prevention of Vision health from smart Phone light, https://www.akgec.ac.in/wp-content/uploads/2025/05/Paper-9.pdf.

ABOUT THE AUTHORS

Er. Dhanshri Parihar Assistant Professor in CSE at AKGEC, Ghaziabad since 2023. She Is having 11 years of teaching Experience. She has completed M.Tech and B.Tech in CSE From Himachal Pradesh University, H.P. Currently pursuing her PhD in Computer Science and Engineering from Galgotias University, Greater Noida U.P. She has Published 10 Research Paper. Her area of interest includes cybersecurity, Machine learning & deep learning.

Dr. Avdhesh Gupta, a seasoned academic with over 24 years of experience in Computer Science and Information Technology, is currently working as Professor and Professor-In charge at Ajay Kumar Garg Engineering College, Ghaziabad. With expertise in accreditation processes, quality assurance, and academic program development, he focuses on computational and mathematical modelling, data mining, network Security, and image processing.

Er. Ashish Kumar is having 12 years of teaching Experience. He has completed M.Tech and B.Tech in CSE and B.Tech in CSE From BPUT, Orissa. He has Published 6 Research Papers. His Subject Specialization area is Networking, Artificial intelligence, Machine learning and Cloud Computing.

Er. Anuradha Singh is working as Assistant Professor in the department of Information Technology at JSS Academy of Technical Education, Noida. She has completed M.Tech and B.Tech in IT from Guru Gobind Singh Indraprastha University, Delhi. Currently pursuing a PhD in Computer Science and Engineering from Dr. A.P.J. Abdul Kalam Technical University, Lucknow. She has over 11 years of academic experience. Her area of interest includes Cyber Security, Machine Learning and Deep Learning.