Sign Language Recognition and Translation Systems for Enhanced Communication for the Hearing Impaired

Kambhampati Sai Sindhu[1]
Department of Computer Science and Engineering
BVRIT HYDERABAD College of Engineering for Women
Hyderabad, India
21wh1a0584@bvrithyderabad.edu.in[1]

Mehnaaz[2]
Department of Computer Science and Engineering
BVRIT HYDERABAD College of Engineering for Women
Hyderabad, India
21wh1a0582@bvrithyderabad.edu.in[2]

Biradar Nikitha[3]
Department of Computer Science and Engineering
BVRIT HYDERABAD College of Engineering for Women
Hyderabad, India
21wh1a05C0@bvrithyderabad.edu.in[3]

Penumathsa Likhita Varma[4]
Department of Computer Science and Engineering
BVRIT HYDERABAD College of Engineering for Women
Hyderabad, India
21wh1a0578@bvrithyderabad.edu.in[4]

Chandrasekhar Uddagiri[5]
Department of Computer Science and Engineering
GITAM University, Hyderabad
Hyderabad, India
cuddagir@gitam.edu[5]

2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU) | 979-8-3503-6178-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/IC-CGU58078.2024.10530832

***Abstract*: This paper addresses the challenges in Sign Language Recognition Systems (SLR) and Sign Language Translation (SLT), focusing on the translation of sign language to text/speech and the reverse process. The unique grammatical structures of sign languages pose a core problem, prompting the development of computational models. The Sign Language Translation (SLT) module, translating recognized gloss into spoken language text, presents a formidable challenge due to grammatical and semantic intricacies. Additionally, the need for diverse datasets, including various sign language dialects in India, further complicates the development of robust SLR and SLT systems. The paper explores technological advancements, challenges, and solutions in both modules, contributing to inclusive communication tools.**

Keywords: Sign Language Recognition, Sign Language Translation, Reversible CNN, Multimodal Dynamic Sign Language Recognition, Gesture Recognition, Natural Language Processing, Indian Sign Language

I. Introduction

The challenge of bridging the communication gap between individuals with hearing and speech impairments and those with normal hearing has driven extensive research into the advancement of Sign Language Recognition Systems (SLR) and Sign Language Translation (SLT). This paper focuses on addressing the complex challenges inherent in two crucial modules vital for facilitating effective communication: the conversion of sign language into text/speech and the reciprocal process of translating text/speech into sign language.

Recognizing the unique grammatical structures and linguistic norms of sign languages as the core problem, researchers have endeavored to bridge this gap through the creation of computational models. One crucial module involves translating sign language expressions, known as "gloss," into textual or spoken language, facilitating communication for the hearing and speech impaired. Simultaneously, the reverse module seeks to interpret spoken or written language and convert it into comprehensible sign language expressions.

The most formidable challenge in this pursuit lies in Sign Language Translation (SLT), where recognized gloss must be accurately converted into spoken language text. This process demands a nuanced understanding of the grammatical and semantic intricacies specific to sign languages, including considerations of tense, order, direction, position, and repetition of signs. Addressing this challenge is critical for achieving seamless and meaningful communication between these two distinct modes of expression.

Beyond the overarching challenges, the development of robust SLR and SLT systems is further complicated by the need for extensive and diverse datasets. Training models capable of accurately interpreting the myriad expressions and gestures in sign languages necessitates a substantial and varied data pool. Moreover, the existence of various sign language dialects in India introduces an additional layer of complexity, requiring adaptable systems that can cater to linguistic diversity within the country.

This paper delves into the complexities associated with these two modules: sign language to text/speech and text/speech to sign language. By examining the technological advancements, challenges, and solutions in each module, the paper contributes to the ongoing efforts in creating inclusive communication tools for the hearing impaired and fostering effective interactions between individuals with diverse communication abilities.

II. Literature Survey:

Voice recognition is achieved through the utilization of RNN and CNN, translating human speech into sign language [4]. An enhanced Mel-DCT filter is applied for voice activity detection [1]. Mel-DCT, while effective, can be sensitive to background noise and non-speech sounds, potentially impacting the accuracy of voice activity detection and overall classification performance in noisy environments. An alternative method employs Multi-Layer Convolutional Neural Networks (ML-CNN) and an integrated encoder to extract hierarchical features from video sequences [2]. The reversible CNN model demonstrates superior accuracy with fewer model parameters, evaluated against existing G-CNN and VGG-11/16 models in both testing and training environments [3].

For continuous sequences of gestures, a modified LSTM model is proposed, enabling the recognition of connected sign sequences. This LSTM model involves splitting continuous signs into sub-units and utilizing neural networks for modeling, eliminating the need to consider various subunit combinations during training [8][9]. Trained on a dataset comprising 35 isolated sign words using the Leap motion sensor for data input, the model achieves an average accuracy of 73.2%. Continuous improvement is crucial for robust and accurate recognition in diverse sign language contexts.

Addressing communication barriers, a movement in a video detection scheme extracts spatial and temporal features from each gesture. A neural network is trained for classification, utilizing a pre-trained Convolutional Neural Network (CNN) to classify gestures and extract frames from videos. However, its primary tuning for short gestures limits its applicability to medium and long gestures. While extending to medium gestures, addressing long gestures remains challenging.

Introducing a novel methodology for context-aware continuous sign language recognition, the Sign Language Recognition Generative Adversarial Network employs a generator for extracting spatial and temporal features from video sequences. It incorporates a discriminator that assesses prediction quality by integrating text information at both sentence and gloss levels [9].

In dynamic sign language recognition, the multimodal method named BLSTM-3D Residual Network (B3D ResNet) addresses drawbacks in existing methods. This approach utilizes a deep 3-dimensional residual ConvNet and bi-directional LSTM networks, focusing on recognizing complex hand gestures with improved accuracy [10]. The process involves object localization, spatiotemporal feature extraction, and video sequence classification.

III. Methodology :

The research process comprises four crucial stages :

1) Obtaining data
2) Preparing and Processing data
3) Implementing Classification.

Figure 1 illustrates a block diagram representing the envisioned mode.

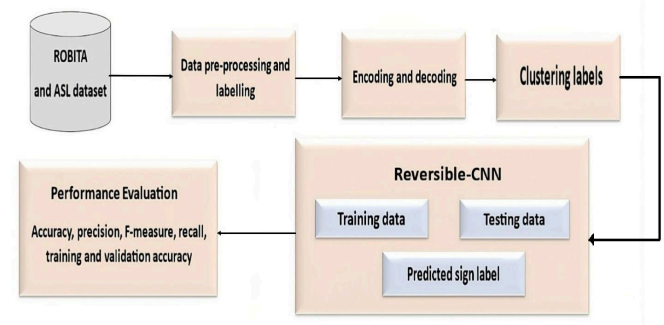


Figure. 1 Reversible CNN Block Diagram

In the initial phase of the technical process, a dataset of sign language videos is gathered, with specific mention of the Sign Language dataset, although alternative datasets are considered viable. Subsequently, the collected videos undergo pre-processing, involving tasks like cropping, resizing, and normalization, preparing them for training and testing a Convolutional Neural Network (CNN). The labor-intensive process of labeling each video with the corresponding sign labels follows. The labeled videos are then encoded into a numerical format, with each frame represented as a numerical array, facilitating comprehension by the CNN.

The technical methodology introduces the concept of a "reversible-CNN" for feature extraction, a CNN type that not only extracts features from the data but can also reconstruct the original data from these features. This unique characteristic proves advantageous for tasks such as anomaly detection and data compression. The extracted features are subsequently employed to cluster the data into distinct groups, aiding in the identification of patterns and ultimately enhancing CNN's accuracy.

Following this, the labeled data is divided into training and testing sets. The training data is utilized to train the CNN, while the testing data serves as a means to evaluate the CNN's performance. Performance evaluation is conducted using established metrics like accuracy, precision, recall, and F-measure, with attention paid to both training and validation accuracy. Finally, the trained CNN is applied to predict sign labels for new videos, showcasing the practical application of the entire technical process.

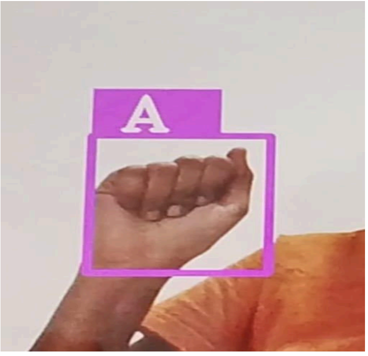
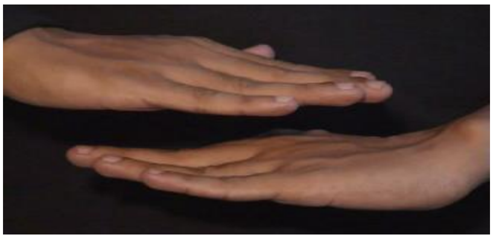


Fig.3 Output for characters

A. Dataset Description

We obtained two types of datasets: the ROBITA Indian Sign Language Gesture Dataset, and the other one we created our own dataset for English Alphabets.The dataset for Indian Sign Language (ISL) gestures consists of sequences of RGB frames corresponding to 23 isolated ISL gestures. The ISL gesture dataset is composed of sequences of RGB frames for 23 Isolated ISL gestures .Our custom dataset is composed of 24 alphabets (A-Y) and digits (0-9). The study utilized the ROBITA Indian Sign Language Gesture Database to gather real-time data, comprising both training and testing datasets labeled with counts. However, the dataset's limited size has been identified as a primary factor affecting prediction accuracy. There exists a direct correlation between the number of training samples and the predictive accuracy of the model.To process the real-time videos, they were initially transformed into a sequence of image frames



.

Fig.2 Sample Sign Representation (ISL) for ÁBOVE

B. Preprocessing

In the domain of gesture recognition, pre-processing stands as a critical step aimed at elevating the quality of the dataset. A crucial stage in the preprocessing workflow involved meticulous background removal from the acquired images. The methods used for preprocessing the data are :

1. Gaussian Filter :

Applies a spatially-weighted convolution using a Gaussian kernel. Effectively reduces Gaussian noise but compromises edge sharpness and fine details.

2. Median Filter :

Non-linear, rank-based filtering. Replaces pixels with the median intensity of their neighborhood. Robust against impulsive noise but may smooth edges and introduce artifacts.

3. Bilateral Filter :

Combines Gaussian spatial weighting with intensity similarity weighting. Preserves edges and textures while smoothing noise. Requires careful parameter tuning for optimal performance.

C. Modules

1) Reversible CNN:

A 12-layer "reversible CNN" model, based on a known CNN architecture, tackles sign language recognition using voice and gesture inputs. This model goes beyond mere recognition, offering the potential to reconstruct the original input, with implications for enhanced analysis and communication possibilities (See fig.4).The concept of a reversible Convolutional Neural Network (CNN) for sign language to text conversion involves leveraging a specialized architecture that not only performs the recognition of sign language gestures but also facilitates the reconstruction of the original input. The reversible CNN model initially processed the input sign language gestures. It captured relevant features, patterns, and representations from the input images or video frames containing the gestures.The encoding phase had several layers of convolutional, pooling, and possibly recurrent or attention mechanisms, extracting hierarchical features from the sign language data(see Fig.5).

The CNN architecture created an intermediate representation of the sign language gestures. This representation captures the essential information required for accurate recognition and conversion These intermediate representations are in the form of feature vectors or tensors that hold crucial spatial and semantic information about the gestures.The reversible aspect of the CNN comes into play in the decoding phase. The model reconstructs or reverses the intermediate representation back into the original input format, aiming to regenerate the sign language gestures from the learned representations.

By reconstructing the input, the model ensures that the information vital for gesture recognition is retained within the reconstructed output.Alongside the reversible process, the CNN model connects the recognized gestures to a text generation or classification module. This component translates the reconstructed gestures or their representations into corresponding textual output, typically in the form of sentences or words that represent the signed content.During training, the reversible CNN model learns to optimize both the encoding and decoding processes simultaneously. It adjusts its parameters to ensure accurate reconstruction while maintaining the crucial information required for text conversion.Ultimately, the reversible CNN model generates the recognized text based on the reconstructed sign language gestures, providing a textual representation of the signed content.

This approach aims to not only recognize sign language gestures but also retain sufficient information to reconstruct the original input, allowing for bidirectional transformation between sign language and textual representations. The model's reversible nature ensures that the reconstruction process preserves the essential elements needed for accurate text generation.

2) Text to Sign Conversion:

Text-to-sign conversion has emerged as a powerful tool in this endeavor, offering accessibility and inclusivity for deaf and hard-of-hearing communities. NLP techniques dissect the text, understanding its meaning, structure, and grammar. Words are broken down, relationships between them are identified, and the overall sentence context is extracted. The system taps into a vast database of sign language gestures and their corresponding meanings. Matching the extracted concepts from the text to the database unlocks the appropriate signs(see Fig.4).

Bringing Signs to Life: 3D models or motion capture data come into play, animating the chosen signs with realistic hand movements, body posture, and Text-to-sign conversion typically involves a tech stack comprising essential components. Natural Language Processing (NLP) relies on libraries/frameworks like spaCy, TensorFlow, or PyTorch to analyze and understand written text. Speech-to-Text (STT) is achieved through Speech Recognition APIs such as Google Cloud Speech-to-Text or Microsoft Azure Speech SDK for transcribing spoken language. Sign language generation incorporates Computer Vision and Animation using OpenCV, OpenPose, and Blender to produce lifelike sign language animations. Frontend development employs standard web technologies like HTML, CSS, and JavaScript, often coupled with a frontend framework like React or Angular. Cloud services from platforms such as AWS, Google Cloud, or Azure are utilized for scalability and storage. This streamlined tech stack ensures accurate and efficient text-to-sign conversion, making information accessible for the hearing and speech impaired individuals.

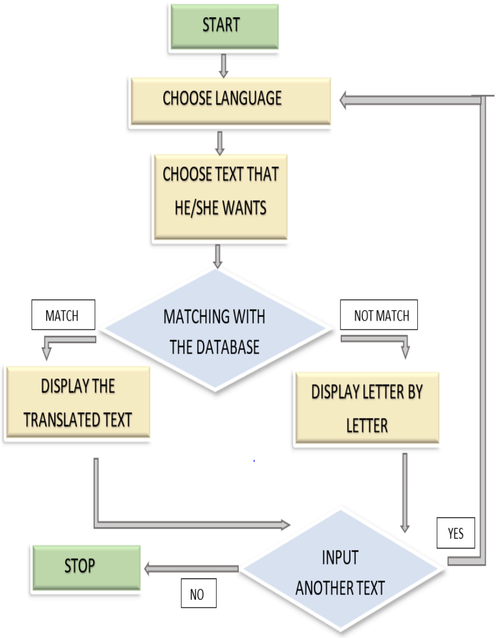


Fig.4 Text-to-Sign flowchart

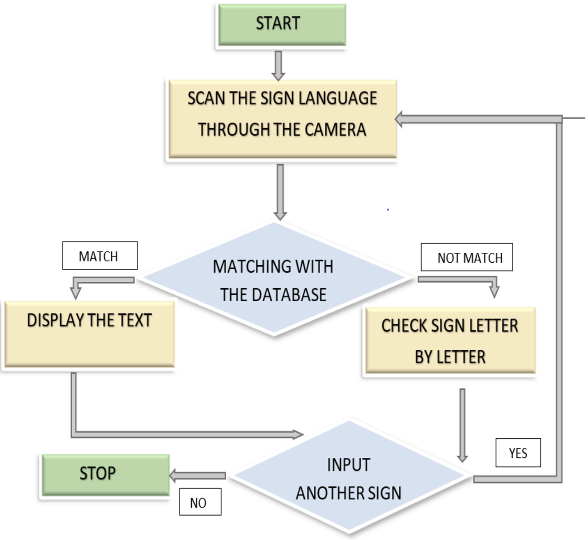

START
SCAN THE SIGN LANGUAGE THROUGH THE CAMERA
MATCH
MATCHING WITH THE DATABASE
NOT MATCH
DISPLAY THE TEXT
CHECK SIGN LETTER BY LETTER
YES
STOP
NO
INPUT ANOTHER SIGN

Fig. 5 Sign-to-Text flowchart

D. Potential Biases

Signers: The dataset has a limited number of gestures that could introduce demographic biases affecting generalizability.

Data Collection Settings: The controlled environment of data collection may not represent the complexities of real-world communication.

Data Selection: Potential biases could exist in the selection of specific words and variations included in the dataset.

D. Validity and Reliability:

1) **Validity:**

Internal Validity: The consistent background and controlled recording conditions contribute to internal validity.

External Validity: The limited vocabulary, static nature, and controlled environment raise concerns about the dataset's external validity and application to real-world scenarios.

2) Reliability:

Repeatability: The availability of multiple sequences for each gesture enhances the dataset's reliability.

Generalizability: The limitations mentioned above raise concerns about the reliability of the dataset for generalizing to broader sign language recognition tasks.

IV. Results & Discussion

This section delineates the process of sign language translation conducted on a frame-by-frame basis using input video sequences. The input data originates from the ROBITA Indian Sign Language Gesture Database (refer Fig. 6), and a representative output is depicted (refer to Fig. 7). Rigorous evaluation of the system's efficacy employed performance metrics, specifically precision, recall, F1 score, and the Kappa coefficient. The sign language to text conversion system yielded promising outcomes, achieving an overall accuracy of 70%. Precision, measured at 46.6428%, signifies the fraction of true positive (TP) samples among those classified as positive. Recall, registering at 75.5314%, denotes the fraction of TP samples captured among all actual positive instances. The F1-measure, representing the harmonic mean of recall and precision, stands at 49.4197%. The Kappa coefficient, a metric for assessing inter-rater reliability, recorded 40.3191%. In conclusion, these results substantiate the efficacy of the sign language to text conversion system and furnish valuable insights for prospective refinements and avenues for further research. The elucidation of Kappa as a measure of inter-rater reliability and the definition of F-measure as the harmonic mean of recall and precision contribute to a comprehensive technical understanding of the evaluation metrics employed.

Kappa: The Kappa coefficient is a statistical measure employed to assess the inter - rater reliability of different qualitative terms.

F-measure: It quantifies the harmonic mean between the recall and precision of models.

Precision: It is the ratio of true positive (TP) samples to the total number of samples classified as positive by the models.

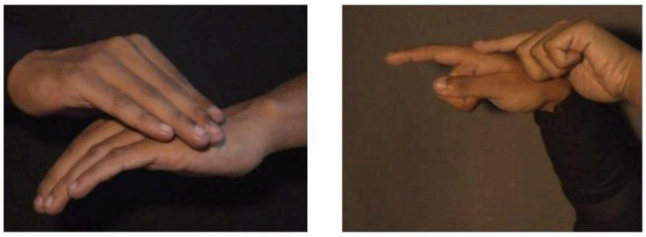


Fig. 6 Robita Dataset Samples

Reversible-CNN has high accuracy and potential for interpretability when compared to other techniques, but it may require more data.

CNN model is Robust and reliable, but less focused on temporal dynamics.

RNN is good for continuous gestures, but can be computationally expensive.

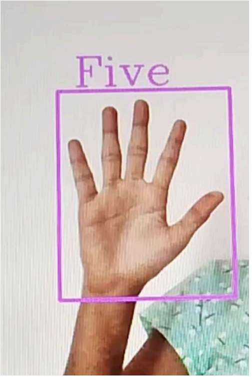


Fig. 7 Output (Digits)

HMM used are efficient with limited data, but for our project requires complex gestures where HMM fails to work.

Reversible CNN technique is preferred over other techniques.

V. Conclusions & Future Scope

The paper contributes insights into the complexities of sign language recognition and translation, providing an overview of challenges, methodologies, and evaluation metrics. The proposed Reversible CNN model and Text-to-Sign Conversion process demonstrate promising results, addressing key issues in bidirectional transformation and accessibility for the hearing impaired. The acknowledgment of biases and validity concerns highlights areas for further research and improvement in creating inclusive communication tools.Our future objectives involve amassing extensive Indian Sign Language (ISL) data for robust training datasets, with a focus on elevating translation capabilities to encompass sentence-level expressions. Additionally, we aim to extend the translated output beyond English to various regional languages, thereby enhancing the linguistic diversity and inclusivity of the system.

Acknowledgment

We thank IIT Hyderabad for funding this project as a part of BUILD Program.

References

[1] Arun Prasath G and Annapurani Panaiyappan k, " Design of an integrated learning approach to assist real-time deaf application using voice recognition system," *SCI.*

[2] G Arun Prasath and K Annapurani, "Prediction of sign . language recognition based on multi layered CNN," *SCI2.*

[3] G Arun Prasath and K Annapurani, "A Reversible Convolutional Neural Network Model for Sign Language Recognition," *inass Publication.*

[4] G. Arun Prasath and K. Annapurani, "A Review on Deaf and Dumb Communication System Based on Various Recognitions Aspect," *Review on ICDCI2021.*

[5] S. Shrenika and M. M. Bala, " Sign Language Recognition Using Template Matching Technique," *In:Proc. of International Conf. on Computer Science Engineering and Applications., pp*, 2020.

[6] M. R. I. J. S. Md. Abdur Rahim, "Non-Touch Sign Word Recognition Based on Dynamic Hand Gesture Using Hybrid Segmentation and CNN Feature Fusion," *Applied Sciences,* 2019.

[7] Angela C. Caliwag , Han-Jeong Hwang ,Sang-Ho Kim, Wansu Lim, "Movement-in-a-Video Detection Scheme for Sign Language Gesture Recognition Using Neural Network," *Applied Sciences,* 2022*,12,10542.*

[8] A. Mittal, P. Kumar, P. Pratim Roy, R. Balasubramanian, B.B. Chaudhuri, "A Modified LSTM Model for Continuous Sign Language Recognition Using Leap Motion," *IEEE Sensors Journal,* vol. 19, no. 16, pp. 7056-7063, 2019.

[9] Ilias Papastratis ,Kosmas Dimitropoulos, Petros Daras, "Continuous Sign Language Recognition through a Context-Aware Generative Adversarial Network," *Sensors (Basel),* vol. 7, no. 21, p. 2437, 2021

[10] Y. Liao, P. Xiong, W. Min, W. Min, J. Lu,"Dynamic Sign Language Recognition Based on Video Sequence With BLSTM-3D Residual Networks," *IEEE Access,* vol. 7, pp. 38044-38054, 2019.