Sign Language Detection Model Using Action Recognition LSTM and Deep Learning

Sourabh Pal[1], Tauseef Alam[2], Priyanshu Singh[3], Raju[4], Akash Chowdhury[5]
[1]Assistant Professor, Dept of CSE (Email id: iamtraj1@gmail.com)
[2]B. Tech. (CSE) Research Scholar, Department of CSE (Email id: mdtauseefalam123456@gmail.com)
IIMT College of Engineering, Greater Noida, UP, India

Abstract: *Sign language serves as a vital communication medium for the Deaf and Hard of Hearing (DHH) community, yet action recognition by computational systems remains challenging. It represents an approach for sign language detection using action recognition principles through a Long Short-Term Memory (LSTM) and deep learning model. By leveraging the temporal dynamics and sequential nature of sign language gestures, the LSTM model is trained to identify and classify signs from video data accurately. The proposed system processes the video sequences to extract key frame features, which are input into the LSTM network, designed to capture the temporal dependencies and nuanced movements characteristic of sign language. A comprehensive dataset comprising diverse sign language gestures is utilized to train and evaluate the model. Experimental results demonstrate that the LSTM-based approach achieves high accuracy in sign language detection, outperforming traditional static frame-based methods. The system's performance is evaluated through various metrics, including precision, recall, and F1-score, showcasing its robustness in real-world scenarios. In this project, our research aims to create a cognitive system that is sensitive and reliable so that persons with hearing and speech impairments may utilize it in day-to-day applications.*

Keywords: Image processing, sign motion, sensors.

Introduction

Sign language is an essential mode of communication for the Deaf and Hard of Hearing (DHH) community, facilitating interaction and expression in a visual language distinct from spoken languages. Despite its importance, the recognition and interpretation of sign language by computational systems pose significant challenges due to the complex, dynamic nature of its gestures. This thesis explores an innovative solution to this problem by employing action recognition principles within a Long Short-Term Memory (LSTM) deep learning framework. By focusing on the temporal and sequential aspects of sign language, the proposed LSTM model aims to enhance the accuracy and reliability of sign language detection from video data. This introduction outlines the motivation behind using advanced deep learning techniques, the methodology involving key frame feature extraction and LSTM network integration, and the anticipated impact of this research on improving sign language recognition systems.

Literature review

Sign Language Recognition (SLR), which tries to translate Sign Language (SL) into speech or text, aims to enhance communication between hearing-impaired people and able-bodied people. Because sign language is intricate and varies for different individuals, this problem is challenging and has a big social impact. [1]

Many different sign language recognition (SLR) algorithms have been developed by researchers, but they can only distinguish between distinct sign motions. In this research, we propose a modified long short-term memory (LSTM) model for continuous sequences of gestures, also known as continuous SLR, that may be able to recognize a collection of related gestures. [2].

Systems that comprehend sign language instantly translate signs in video feeds to text. Utilizing convolution neural networks (CNNs), feature pooling modules, and long short-term memory networks (LSTM), a novel isolated sign language recognition model is developed in this study. [3,4].

Deep learning methods can be applied to overcome communication barriers. To identify and detect words in a person's gestures, the model discussed in this paper makes use of deep learning. [5,6].

Hand and body gestures are used to symbolize the vocabulary of dynamic sign language. This approach uses a combination of RNN and long short-term memory networks (LSTM) models to address problems with dynamic sign language detection. We were able to determine the position, shape, and orientation of the objects by removing important hands, body, and facial parts. [7,8].

Using the sign language datasets and the human key points deduced from the face, hands, and other bodily parts, we develop a sign language recognition system. [9,10].

It is still challenging for non-sign language speakers to communicate with sign language users or signers, despite the fact that sign language has lately gained more popularity. Recent developments in deep learning and computer vision have led to promising success in the areas of motion and gesture recognition using deep learning and computer vision-based techniques. [10].

Methodology

The method we propose can identify a variety of motions by recording video and converting it into distinct sign language labels. After being categorized and matched to a picture, manually created pixels are then compared to a trained model. Because of this, our system is very good at finding certain character labels. Our proposed system recognizes various actions in video recordings and separates them into discrete frames using sign language. Our method is quite tight in determining specific text labels for characters since the hand pixels are divided and matched to the generated picture before being transferred for comparison with a trained model. Collaborative Communication, which enables users to communicate successfully, is a feature of the suggested system.

The system described is designed to recognize sign language gestures through the use of **LSTM (Long Short-Term Memory) networks**, which are effective at processing sequential data and capturing temporal dependencies within gesture sequences. The architecture integrates action recognition principles to enhance performance, focusing on temporal modeling and motion analysis to improve the accuracy of sign language interpretation.

Several technologies are employed to ensure robust performance. **TensorFlow** is used for model building, training, and deployment, providing a flexible framework for deep learning applications. **Mediapipe** is leveraged for real-time hand and gesture tracking, allowing the system to process hand movements with high accuracy. **OpenCV-Python** is used for video processing, handling tasks like feature extraction from video frames. **NumPy** supports numerical operations and data manipulation, which is essential for handling large datasets and performing the necessary mathematical computations. Additionally, **Matplotlib** is utilized for visualizing data and performance metrics, aiding in model evaluation and adjustment.

For data collection, the system uses a diverse set of **sign language gesture datasets**, ensuring that it can handle a wide variety of gestures from different sign languages. This diversity is important for training the model to recognize gestures across different users and contexts. The datasets are accurately annotated, which is crucial for supervised learning. Preprocessing steps involve extracting key frame features, normalizing video sequences to maintain consistent quality and frame rates, and applying **data augmentation** techniques to improve the model's ability to generalize to new, unseen data.

The LSTM model is specifically designed to capture the sequential nature of sign language. It analyses the temporal relationships between gestures, ensuring that the system recognizes not only the individual signs but also how they relate

to each other over time. This is essential because sign language gestures often depend on previous movements for context and meaning. The model also incorporates key features such as hand position, speed, and motion direction, making the input richer and improving recognition accuracy.

Training the model involves using these diverse datasets, tuning hyperparameters for optimal performance, and validating the model using separate test sets. This approach ensures that the model generalizes well and does not overfit to specific data. Validation and testing are conducted to evaluate the system's robustness and reliability in different settings.

The performance of the system is measured using several key metrics, including **accuracy**, **precision**, **recall**, and **F1-score**. These metrics help assess the system's ability to correctly identify sign language gestures and its ability to capture all relevant instances. Benchmarking is done by comparing the LSTM-based approach to static frame-based methods, demonstrating the advantage of the LSTM model in capturing temporal dynamics. The model is also tested in real-world scenarios, ensuring that it performs well under diverse and unpredictable conditions.

In terms of development, the system is built for **real-time video input**, enabling immediate recognition of sign language gestures as they occur. A user-friendly interface is developed for seamless interaction, making the system accessible to a wide range of users. Scalability is achieved through **cloud deployment**, allowing the system to be accessed by users globally and handle large-scale data processing. The system is also designed for integration with **assistive devices**, making it a valuable tool for the Deaf and Hard of Hearing community. Additionally, it can be used as an **educational tool** to facilitate sign language learning, helping both learners and educators interact with and teach sign language more effectively.

In summary, the system combines advanced machine learning techniques, real-time processing capabilities, and robust data collection methods to provide an effective solution for sign language recognition. It is scalable, reliable, and adaptable to both assistive and educational settings.

Implementation

The figure 4.1 describes the system architecture employed in experimental in detail.

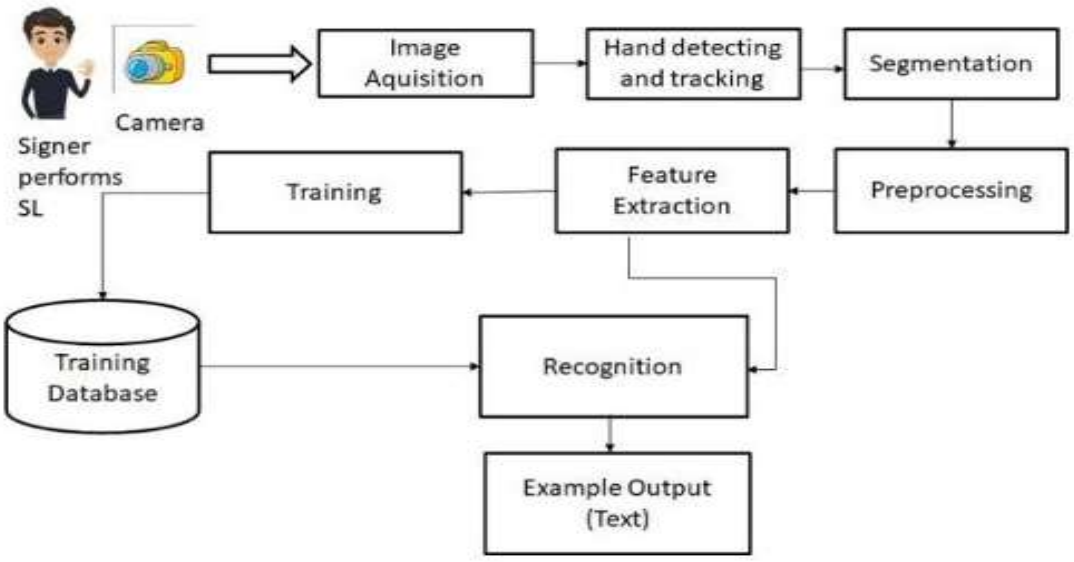


Fig 4.1 : System Architecture

Image Acquisition: It is the process of removing a picture from a source, usually one that is hardware-based, for image processing. The hardware-based source for our project is Web Camera. Due to the fact that no processing can be done

without a picture, it is the initial stage in the workflow sequence. The image that is obtained has not undergone any kind of processing.

Segmentation: It is a method of removing objects or other background details from a recorded image. The segmentation procedure makes use of edge detection, skin colour detection, and context subtraction. Order to recognize gestures, the motion and position of the hand must be classified as well as identified Edge Based Segmentation is used in this project to achieve Segmentation.

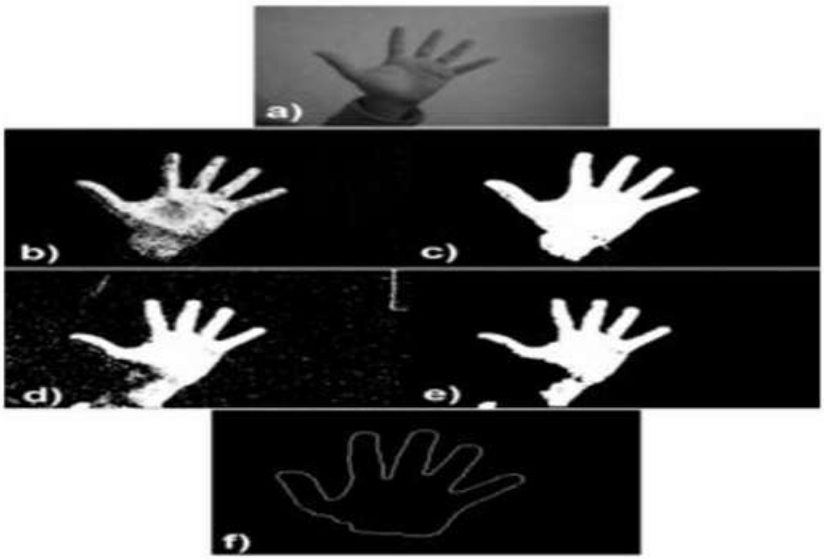


Fig 4.2: Segmentation process

Preprocessing Process: Images need to be processed before they can be used by models for inference and training. This includes, but is not limited to, changes in colour, size, and orientation. Additionally, preprocessing a model can shorten the training process and speed up inference. Shrinkage of extremely big input photos will greatly shorten the training period without significantly affecting model performance. The following are the stages of preprocessing:

- **Morpholical transform:** Morphological processes create an output image of the same size by using the structural characteristics of an input image. By comparing the matching pixel from the input image to its neighbour, each pixel in the output image is given a value. Dilation and erosion are two separate types of morphological changes.
- **Blurring:** One example is the blurring of a picture using a low-pass filter. The term "low-pass filter" in computer vision refers to a method for reducing noise from a photograph while keeping the majority of the image. The blur must be finished before tackling harder tasks, including edge detection.
- **Recognition:** Children who have hearing loss are at a disadvantage since they find it difficult to understand the lectures that are shown on the screen. The American Sign Language was developed to assist these kids in managing their schooling as well as to make daily life easier for them. To assist these kids in learning, we came up with a model that would let them make ASL motions to the camera, which would then interpret them and give feedback on what language was understood. To do this, we combined Media pipe Holistic with OpenCV to determine the essential indicators of the poser with all the values that needed to be collected and trained on the Long Short Term Memory.
- **Feature Extraction:** In order to extract preset properties from the already possessed images, such as shape, contour, geometrical feature (position, angle, distance, etc.), color feature, histogram, and others that are later utilized for categorizing or identifying signs. Feature extraction is a stage in the dimensionality reduction procedure that isolates and arranges a sizable collection of raw data. Class sizes were reduced to more manageable numbers. As a result, processing would be simpler. These massive data sets profusion of variables is their most

notable feature. These variables demand a large amount of CPU processing power. The best feature can be extracted from huge data sets via function extraction, which selects and combines variables into functions to reduce the amount of data. The Feature extraction is achieved by storing key point(face,pose, hand) values in an array using Media pipe.

- **Text Output:** Recognizing diverse postures and bodily gestures, as well as converting them into text, to better understand human behaviour.

Diagram

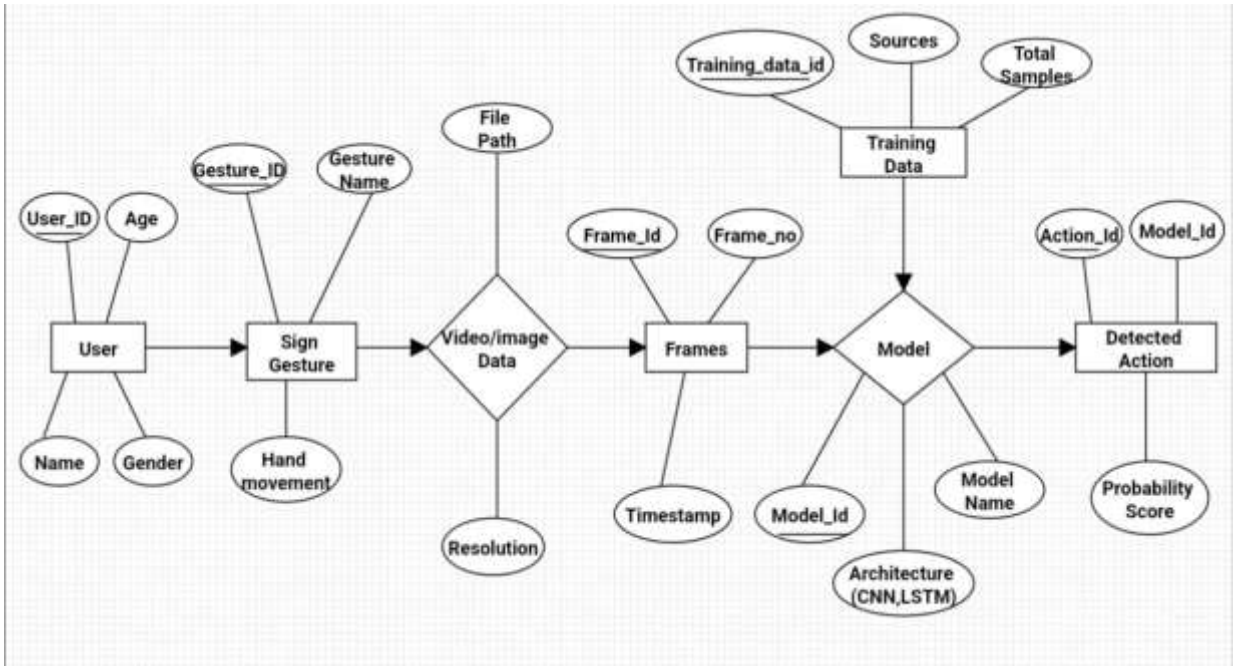


Fig 4.3 ER Diagram

Key Challenges

The challenges highlighted in this section revolve around several fundamental aspects that impact the development and performance of sign language recognition systems, especially those based on deep learning models such as LSTM networks. These challenges need to be carefully addressed to ensure the system's robustness, scalability, and effectiveness in real-world applications.

One of the key issues is data scarcity. Sign language datasets are often limited, especially when it comes to representing a wide range of languages, signers, and contexts. The lack of diverse data poses a significant challenge because the models trained on limited datasets may not generalize well to new or unseen data. This results in models that perform well on training data but fail to accurately recognize gestures in real-world scenarios. To overcome this, it's essential to create larger and more diverse datasets that encompass a broader range of sign languages and individual signing styles. Moreover, techniques like data augmentation can be used to artificially expand the training data by introducing variations such as different signing speeds, hand orientations, and backgrounds, which can help the model become more generalized and robust to unseen situations.

Another major challenge is user variability. Sign language communication is inherently personal, with significant variations in signing style, speed, and even facial expressions that contribute to the meaning of the gestures. These individual differences make it difficult for a model to recognize signs from different users, as it needs to be flexible enough to accommodate these variations. Additionally, the context in which sign language is used—such as the physical environment or social setting—can affect how gestures are performed. Therefore, models need to be designed to not only recognize the gestures but also adapt to these variations in signing style and context. To address this, it's crucial to incorporate more comprehensive data that includes not just the gestures but also features that account for the signer's unique characteristics and the environment in which the communication occurs.

Another key challenge in sign language recognition is temporal modeling. Sign language is a dynamic and continuous form of communication where the meaning is conveyed through a sequence of movements and facial expressions that are often performed fluidly and without clear breaks. While LSTM networks are particularly suited for capturing temporal dependencies in sequential data, they still face difficulties in handling the complex, often ambiguous nature of continuous sign language. Gestures and expressions in sign language often blend together, making it hard for the model to distinguish one gesture from another, especially when the boundaries between gestures are not clearly defined. To enhance the model's ability to handle this dynamic and continuous nature of sign language, further improvements in temporal modelling are needed. This could involve the integration of attention mechanisms, which allow the model to focus on the most relevant parts of the input sequence, or the adoption of more sophisticated LSTM architectures, such as bidirectional LSTMs or attention-based LSTMs, that can better handle long-range dependencies in the data.

Finally, computational cost and efficiency are significant concerns when deploying deep learning models, particularly those involving LSTM layers, which can be computationally expensive. Training and deploying deep learning models, especially on large datasets, can require substantial computational resources and time. This becomes even more critical when the system needs to operate in real-time or on devices with limited processing power. For instance, running sign language recognition models on mobile devices or edge devices can be a challenge due to limited resources. To mitigate this, researchers and developers need to focus on improving the efficiency of the models. This could involve developing lighter models that consume less computational power, or using techniques like model compression and quantization to reduce the size and complexity of the models without sacrificing their performance.

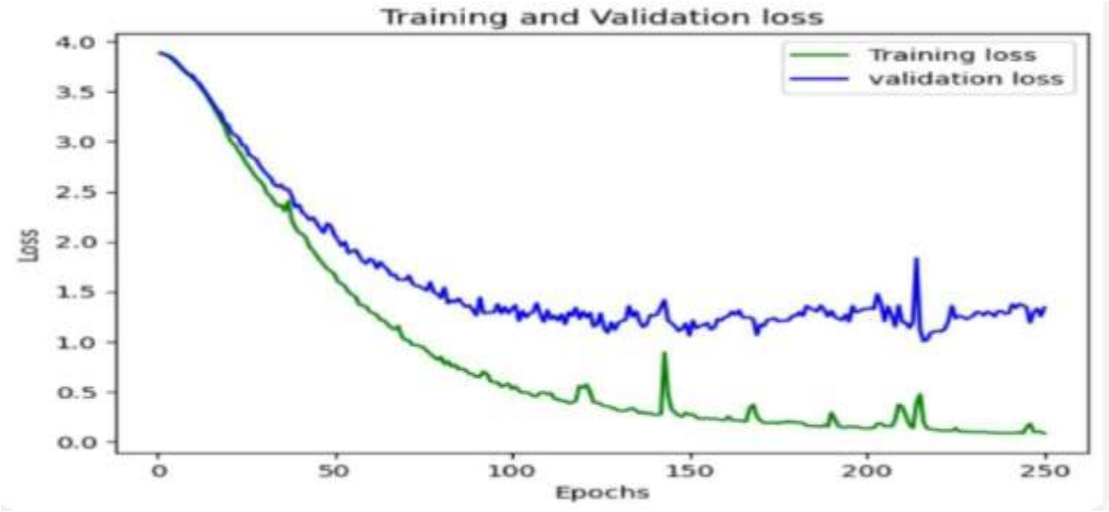
In summary, while deep learning models like LSTMs offer powerful capabilities for sign language recognition, several challenges remain. These challenges include data scarcity, user variability, temporal complexity of continuous signs, and computational inefficiency. Addressing these issues requires a multifaceted approach, including expanding datasets, improving model generalization, enhancing temporal modelling techniques, and optimizing the computational efficiency of the models. By tackling these challenges, we can develop more robust, adaptable, and efficient sign language recognition systems that can be used in real-world applications.

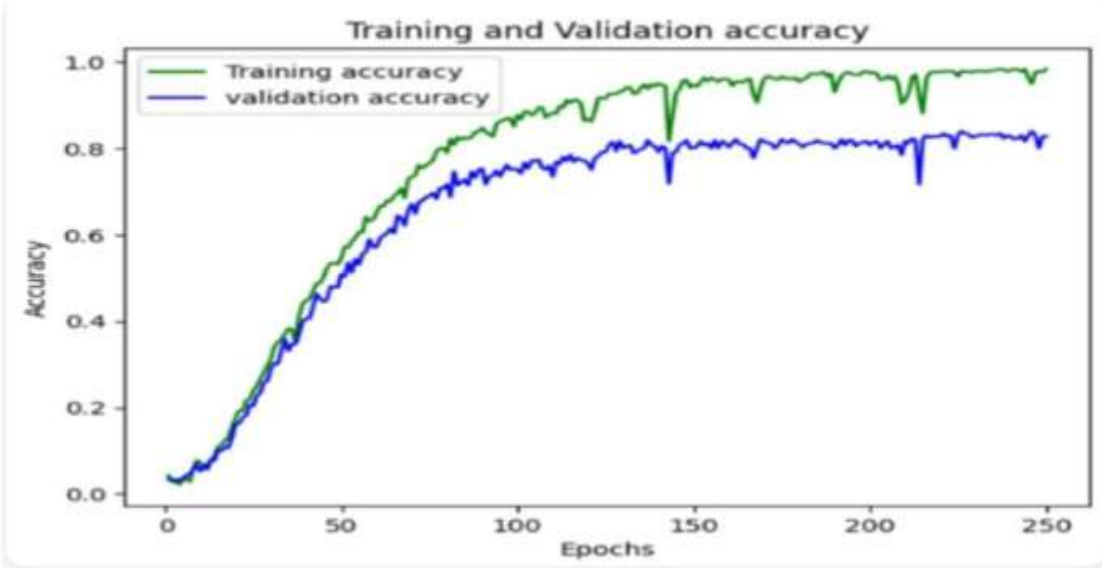
Result and Discussion

The dataset was then split into training and testing sets to train the LSTM-based model and evaluate its performance. The proposed system yielded promising results during the validation process, achieving a training accuracy of 96% and a test accuracy of 87% for ISL recognition. These results outperformed previous approaches in the field. The system's ability to effectively detect and recognize actions from dynamic ISL gestures, facilitated by the deep learning-based approach utilizing LSTM networks, demonstrates the potential for more accurate and robust sign language recognition systems.

The analysis of the training loss and validation loss over the number of epochs provides valuable insights into the performance of the machine learning model. When both losses decrease together, it indicates that the model is effectively learning from the training data and generalizing well to new, unseen data. This is a positive sign as it suggests that the

model is not overfitting the training data and has the potential to make accurate predictions on new data points.





Conclusion and Future Work

In conclusion, the technique of hand motion recognition using vision-based methods offers significant advantages in sign language recognition systems. By utilizing models that analyze both spatial and temporal data, the system can classify gestures effectively. The use of **LSTM (Long Short-Term Memory)** networks to process these attributes has proven effective in recognizing sign language gestures, which is a complex task given the multitude of possible gesture combinations that a system must understand and translate accurately. However, considering the broad scope of this challenge, it is often beneficial to break it down into more manageable subproblems. The approach presented in this system serves as a potential solution to one of these subproblems—creating a **first-person sign language translation system** using simple cameras and **convolutional neural networks** (CNNs). This simplified approach can offer substantial progress in developing a practical system for real-time sign language recognition.

The main goal of **sign language recognition (SLR) systems** is to **accurately detect and interpret sign language gestures**, capturing the intricate details of not just the hand movements, but also the arm motions and facial expressions that contribute to the full meaning of a sign. The system discussed here achieves this by leveraging state-of-the-art machine learning techniques that can identify these subtle features, which is key for real-world applications. Such systems offer clear benefits, particularly in improving **communication accessibility** for the Deaf and hard-of-hearing communities.

They allow individuals to interact more seamlessly with technology and people, bridging a gap that has traditionally been challenging.

Looking ahead, there are several opportunities for future development. One significant direction is improving the system's ability to recognize **changes in temporal space**. Since sign language involves continuous, dynamic movements, recognizing the timing and sequencing of gestures is crucial. The system will need to be enhanced to detect these **temporal changes** more accurately. Another important development is creating a **comprehensive solution** that could facilitate **two-way communication**: not only translating **sign language into text or speech** but also converting spoken language into **sign language**. This requires refining the **image processing component** of the system to handle more complex tasks, such as recognizing spoken language and converting it into corresponding sign gestures.

Future work will also focus on identifying **motion-related cues** in the gesture sequences, as these are crucial for improving the system's recognition accuracy. Additionally, translating the sequence of movements into **text** or **words and sentences**, and then converting that into audible speech, will be a critical step toward making the system a truly **interactive communication tool**. By addressing these challenges, the system can evolve to bridge the communication gap for the Deaf and hard-of-hearing communities more effectively, fostering inclusivity and accessibility in society.

Refrences

[1] Hochreiter, S. & Schmidhuber, J. (2024). "Long Short-Term Memory." Neural Computation, Introduces the LSTM network, essential for sequence modelling tasks like sign language detection.

[2] Graves, A. (2023). Supervised Sequence Labelling with Recurrent Neural Networks. Springer. Covers the application of RNNs and LSTMs in sequence labelling, relevant for sign language recognition.

[3] Zisserman, A. (2022). "Two-Stream Convolutional Networks for Action Recognition in Videos." Advances in Neural Information Processing Systems (NIPS), 568-576. Introduces a two-stream convolutional network for action recognition, foundational for detecting dynamic gestures in sign language.

[4] Carreira, J. , & Zisserman, A. (2021). "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. " Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6299-6308.

[5] P.K. Singh, (2020). "Sign Language Detection Using Artificial Intelligence".

[6] Pigou, L. Dieleman, S. (2019). "Real-Time Continuous Sign Language Recognition Using LSTM Networks".

[7] Molchanov, P., Yang, X., (2018). "Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition".

[8] Y., Yan, S., Xiong, (2017). "Skeleton-Based Recognition with Temporal Graph Networks".

[9] Oyedotun, O.K., Khashman, (2016). "Sign Language Recognition Using Deep Convolutional Neural Networks".

[10] Nandakumar, R. Kinsy, (2015). "Continuous Gesture Recognition".