LSTM-Based Recognition of Sign Language

Aman Bhadouria, A, Bhadouria
Department of Computer Science Engineering& Information Technology, Jaypee Institute of Information Technology Noida-62, Uttar Pradesh, India
amanbh0709@gmail.com

Parish Bindal, P, Bindal
Department of Computer Science Engineering& Information Technology, Jaypee Institute of Information Technology Noida-62, Uttar Pradesh, India
parishbindal@gmail.com

Naivedya Khare, N, Khare
Department of Computer Science Engineering& Information Technology, Jaypee Institute of Information Technology Noida-62, Uttar Pradesh, India
rohitkhare998@gmail.com

Deepti Singh, D, Singh
Department of Computer Science Engineering& Information Technology, Jaypee Institute of Information Technology Noida-62, Uttar Pradesh, India
deepti.singh@jiit.ac.in

Ankita Verma, A, Verma
Department of Computer Science Engineering& Information Technology, Jaypee Institute of Information Technology Noida-62, Uttar Pradesh, India
ankita.verma@mail.jiit.ac.in

ABSTRACT

Sign language recognition is a critical technology for enhancing communication accessibility for individuals with hearing impairments. In this paper, we present a robust and efficient system for sign language recognition using Long Short-Term Memory networks and Mediapipe, a versatile framework for machine learning solutions. Our approach leverages Mediapipe's pre-trained hand detection and tracking models to extract key hand landmarks from video sequences in real-time. These landmarks are then fed into an LSTM network, which is adept at capturing temporal dependencies, to classify various sign language gestures. There are learning aids available for those who are deaf or have trouble speaking or hearing, but they are rarely used. Real-time image processing would be used in the proposed system to handle live sign movements. After that, classifiers would be used to discriminate between different signs, and text would appear in the translated output. The existing systems can detect movements with a significant lag because they only use image processing. Our goal in our work is to develop a cognitive system that is trustworthy and sensitive enough for people with speech and hearing problems to use it in daily activities.

CCS CONCEPTS

• **Insert your first CCS term here**; • **Insert your second CCS term here**; • **Insert your third CCS term here**;

KEYWORDS

Long Short-Term Memory (LSTM), American Sign Language (ASL), sign gestures, hearing aid

ACM Reference Format:
Aman Bhadouria, A, Bhadouria, Parish Bindal, P, Bindal, Naivedya Khare, N, Khare, Deepti Singh, D, Singh, and Ankita Verma, A, Verma. 2024. LSTM-Based Recognition of Sign Language. In *2024 Sixteenth International Conference on Contemporary Computing (IC3-2024) (IC3 2024), August 08–10, 2024, Noida, India.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3675888.3676105

1 INTRODUCTION

Communicating with individuals who have hearing loss can present significant challenges due to a variety of factors that impact both the speaker and the listener. These challenges can include difficulty in clearly conveying and understanding spoken messages, which can lead to misunderstandings and frustration for both parties. It can be challenging for non-disabled people to grasp the hand movements used by the deaf and mute in sign language. In many places, American Sign Language (ASL) is the most commonly used sign language. For Deaf and Mute (D&M) individuals, sign language is the primary means of communication, as their specific handicap prevents them from using spoken languages. Communication involves the exchange of ideas and messages through speech, gestures, behavior, and visual cues. D&M individuals rely on their hands to create various gestures to convey their thoughts to others. Gestures transmit non-verbal messages that are interpreted visually. Sign language is this type of non-verbal communication used by Deaf and Mute individuals. Figure 1 shows different types of Gesture Movement that are usually there while performing Sign Language, like Fingerspelling in which words are broken down into alphabets and then shown in sign language. Additionally, sign language is a form of communication that integrates hand shapes, arm positioning and motion, facial expressions, and lip movements to convey concepts without relying on sound. Despite common belief, sign language is not universally understood. These differ depending on the locale [1].

Systems that can recognize different symptoms and explain them to the public are thus necessary. The goal of reducing the communication barrier between Deaf and Mute (D&M) individuals and those

IC3 2024, August 08–10, 2024, Noida, India

ACM ISBN 979-8-4007-0972-2/24/08
https://doi.org/10.1145/3675888.3676105

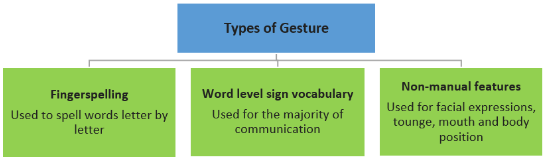

Types of Gesture
Fingerspelling
Used to spell words letter by letter
Word level sign vocabulary
Used for the majority of communication
Non-manual features
Used for facial expressions, tounge, mouth and body position

Figure 1: Types of Gesture

who are not is driven by the importance of effective communication for everyone. Sign language translation is a rapidly expanding field of study and offers the most innate form of communication for individuals who are deaf or have hearing impairments. Utilizing hand gesture recognition technology allows deaf individuals to verbally communicate with hearing individuals without requiring an interpreter [2]. The system is engineered to automatically translate American Sign Language (ASL) into text and speech. It is crucial to create specialized sign language applications for individuals who are deaf or mute, allowing them to communicate with those who do not understand sign language. Our primary goal in this project is to start closing the communication divide between hearing individuals and users of sign language who are deaf or mute. To achieve this, we have developed a vision-based system that can recognize sign language gestures in videos or action sequences. The temporal and spatial properties of the video sequences were employed as the method for identifying the sign language motions. By employing models, spatial and temporal attributes have been acquired. The LSTM [3] model of the recurrent neural network was trained using spatial data extracted from the video series. With good performance, the suggested approach offers an effective means of translating sign language to text. The technology has numerous uses, such as teaching young children sign language comprehension and introducing them to computers.

2 LITERATURE SURVEY

Many sign languages are utilized worldwide, such as American Sign Language (ASL), French Sign Language, British Sign Language (BSL), Indian Sign Language, Japanese Sign Language, among others, which have been studied extensively worldwide. Basic glove solutions, employed to monitor intricate hand movements, are insufficient for fulfilling signers' everyday communication requirements [4]. These solutions can only identify individual discrete gestures, such as numbers, letters, or words, rather than complete sentence. In a previous study [5], researchers proposed an AI-driven sign language identification and communication system utilizing a virtual reality interface, deep learning components, and sensor-equipped gloves. Their deep learning models, with and without segmentation assistance, demonstrated the ability to recognize 50 words and 20 sentences. The segmentation process, a key aspect, divides continuous sign signals into individual word units. Subsequently, the deep learning model deciphers each word unit, reconstructs the sentences in reverse order, and identifies them. In another work [6], the authors research different approaches to translate English-Sign language. For this initially they used a hardware glove with flex sensors, but it soon becomes clear that its precision is not very good. Accuracy was further enhanced by training a convolutional neural network model with an existing data set. The scope was limited and the data sets were not diverse, thus a new diversified data set was made and the model was further enhanced. The new model can anticipate practically every letter with a very high degree of accuracy. They have expanded the data set with several additional gestures that included both hands and facial elements. In [7], authors have designed the Indian Sign Language Recognition System (ISLR), a system specifically for users of Indian Sign Language (ISL) and used holistic landmarks and LSTM model to build the complete system.

In another work authors have proposed framework to create smart mobile devices that will enable the deaf and dumb community to communicate wherever they are, closing the gap in communication between normal, hearing, and speech-impaired persons. Smartphones serve as a translator between speech and sign language and vice versa. The numerous gesture-based indicators are automatically translated into speech while being captured. using a variety of image processing methods and deep learning models to process a gesture image. The study's conclusions show that this dynamic hand gesture detection model for smartphones provides both signers and non-signers with an accurate and practical way to interact [8]. In their work on Continuous Sign Language Recognition (CSLR), researchers in [9] developed a deep learning system named SignBERT to capture the core elements of sign languages and extract spatial features. SignBERT combines the bidirectional encoder representations from transformers (BERT) with the residual neural network (ResNet). The work has also been done to remove communication barriers with those who are hard of hearing and to offer educational resources for sign language, the goal is to use unsupervised feature learning to translate ASL signed hand motions into text and speech [10]. The automatic detection of human gestures from camera images is a compelling area of research for intelligent vision systems. In [11], the authors employ a convolutional neural network (CNN) [14] to recognize hand movements associated with human tasks from camera-captured images. They use a skin model and calibrate hand position and orientation to gather training and testing data, enabling the CNN to achieve robust performance.

3 METHODOLOGY

The two main components of the project are a gesture recognition system and a translation system. The gesture recognition system will use a camera to capture the ASL movements, and it will then process and recognise the gesture using machine learning methods. After the gesture has been recognised, text/speech translation will be carried out using techniques for natural language processing. In Phase-I, we built a framework to convert text to sign language. This can be challenging given that sign language is utilised in numerous distinctive regional languages all over the world. Furthermore, sign language does not directly translate spoken words; instead, it makes use of body language, facial expressions, and hand gestures. The system must be able to understand the complexities of sign language in order to successfully translate the text into the appropriate sign language. For Phase-II, the system must be capable of precisely translating sign language back to text. A recurrent neural network (RNN) architecture called LSTM (Long Short-Term Memory) [13] can be used to recognize and decipher sign language motions and movements. A series of video frames capturing the hand and body motions of a signer while they use sign language would be the input for an LSTM network for sign language translation. In order to recall and understand the patterns and connections between the signs being performed, the LSTM network would process each frame in the sequence one at a time. The LSTM network adjusts its internal state as it analyzes each frame in the sequence, producing an output that is consistent with the sign's most likely interpretation. Additionally, the fact that regional and cultural variations in sign language might make this task challenging. Before converting sign language into text, the system needs to accurately identify and understand the gestures, facial expressions, and body language inherent in sign language, as it is primarily a visual language. A functional web application that could instantaneously translate text and voice to sign language was what we aimed to create in Phase-I. The most frequently used ASL words were required for this, and they will be documented and mapped. We have used www.handspeak.com website to choose the most-used words. The algorithm used in Phase-I is mentioned below:

Algorithm 1 Phase-I Algorithm

Step 1: Reviewing the English word.
Preprocessing is a method of looking at the grammatical structure of a sentence. The conjugation must be obtained to model English phrases using syntactic grammar, such as the directive technique. It was essential to translate the English word's grammatical structures into sentences that reflected the English word's linguistic form.
Step 2: Guidelines for converting a sentence from English into ASL
With the aid of the data, a statement in English is converted into ASL using 20 various verb tenses. All 20 verb tense patterns are taken into consideration while translating an English sentence to ASL.
Step 3: Discarding offensive language
Keywords consist of American Sign Language (ASL) phrases, and any unnecessary terms should be excluded. These grammatical structures are not utilized in ASL translation. Therefore, the module was designed to eliminate all words that cannot be converted into ASL.
Step 4: Lemmatization and synonym replacement
Popular nonverbal communication techniques include ASL. An auxiliary verb, an inflexion, or a suffix are not permitted for any of the nouns on the list. The ASL sentences' non-root words are lemmatized by the recognition system [15].
Step 5: A sign's visual representation
The algorithm will then scan the database for the ASL speech signal after completing the preceding steps. The similarities between the different expressions. It will be built on a core string-matching technique and used for both video classification and text corpus processing.

Figure 2, represents the flowchart of Phase-I.

The system must be able to accurately translate sign language back to text for Phase-II. This task may also be difficult due to regional and cultural differences in sign language. Being a visual language, sign language requires the system to accurately detect and interpret the gestures, facial expressions, and body language before it can be translated into type. The algorithm used in Phase-II is mentioned below:

Algorithm 2 Phase-II Algorithm

Step 1: Extraction of Keypoints.
One of the models in the Mediapipe machine learning pipeline detects palms by analyzing the entire image.
The palm detection approach is employed only when the landmark model fails to localize the hand. The machine learning pipeline utilizes the hand landmarks detected in the preceding frame to crop the hand.
To enhance comprehension of using Mediapipe for hand gesture detection, hand movements are analyzed using the K nearest neighbor algorithm in conjunction with Mediapipe.
Step 2: LSTM-Based Recognition of Sign Language
The LSTM model [13] must then be trained using the dataset.
To forecast gestures in real time, a saved version of the machine learning model is used.
The algorithm is set up to run for 2000 iterations.
The system triggers the rectified linear activation function (ReLU), which outputs the input directly if it is positive and 0 otherwise [16].
The system utilizes the Adam Optimizer method [17], which is an optimization strategy for updating network weights based on the stochastic gradient descent method.

Figure 3, represents the flowchart of Phase-II.

Dataset Collection: Using the open-source MediaPipe [12] package, the dataset needed for the proposed algorithm's experimental analysis is produced. In order to train the model, it records the hand motions for 60 frames and saves them as a NumPy array. Thirty distinct alphabet permutations can be found in the dataset. We categorized our data since we planned to split it into two categories: gesture and static. We recorded videos of mostly alphabets (a-z) and numbers (0-9) first, and then we recorded videos of words and phrases that needed motions. Figure 4 shows the glimpse of videos recorded by our team member.

To put it briefly, we tracked and identified the hands' landmarks, extracted them, and saved them in a sequence of datasets for each hand gesture. This is how we developed and gathered the dataset. In the next step we performed preprocessing data and labeled each gesture. Understanding hand shape and motion can improve user experience across a variety of technical platforms and professions. A high-precision approach for tracking hands and fingers is Media Pipe Hands. Unlike the state-of-the-art approaches used today, which often require powerful desktop workstations for inference, our technology provides many hands scale real-time performance. Figure 5 shows hand landmarks used and Figure 6 shows types of hand landmarks.

4 RESULT ANALYSIS

We were able to translate text to sign language video during the project's first phase. To do this, a web-based application was created that allowed users to enter text, which the system subsequently processed to create a sign language video. The thoughtful design

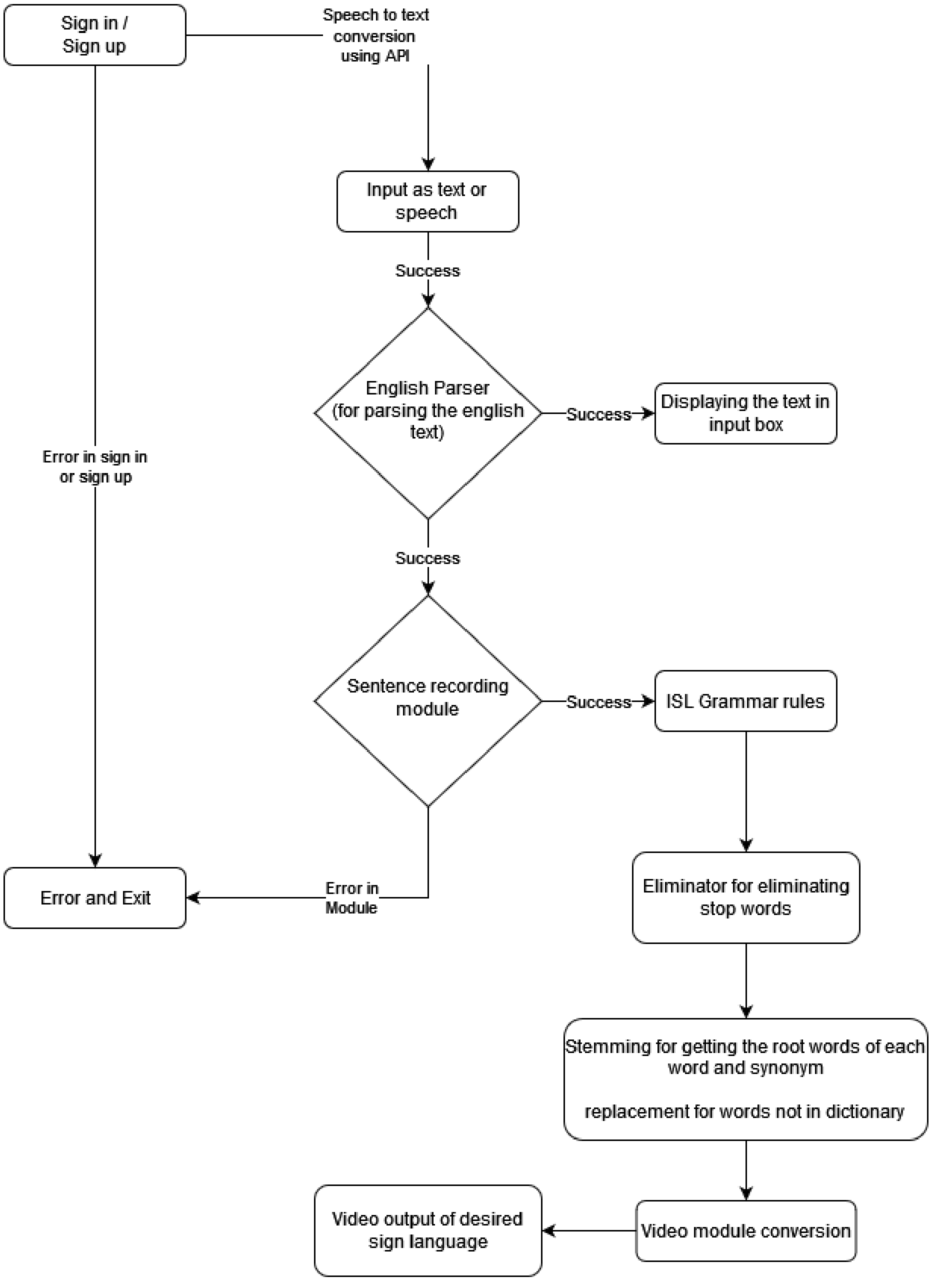

Sign in /
Sign up
Speech to text conversion using API
Input as text or speech
Success
English Parser (for parsing the english text)
Success
Displaying the text in input box
Error in sign in or sign up
Success
Sentence recording module
Success
ISL Grammar rules
Error and Exit
Error in Module
Eliminator for eliminating stop words
Stemming for getting the root words of each word and synonym
replacement for words not in dictionary
Video output of desired sign language
Video module conversion

Figure 2: Flowchart of Phase-I

and development of the user interface, which was simple and easy to use, was responsible for this phase's success. The system was also designed to manage a big dataset of terms and expressions that are often used in everyday speech. We were only able to test a small number of words and phrases during this stage due to the laptop's constrained capabilities, which included less GPU compute power and storage capacity. Figure 7, shows the snapshot of website developed to translate text and voice to sign language.

For Phase-II, the system must be capable of precisely translating sign language back to text. Additionally, the fact that regional and cultural variations in sign language might make this task challenging. The system must also accurately recognize and interpret the gestures, facial expressions, and body language used in sign language because it is a visual language before it can convert it into text. 7 8, shows real time feature extraction we have done.

Using the LSTM and MediaPipe approach, we were able to completely accurately translate three words from sign language to text during the project's second phase. This is a big accomplishment for the project because effective communication with the deaf and

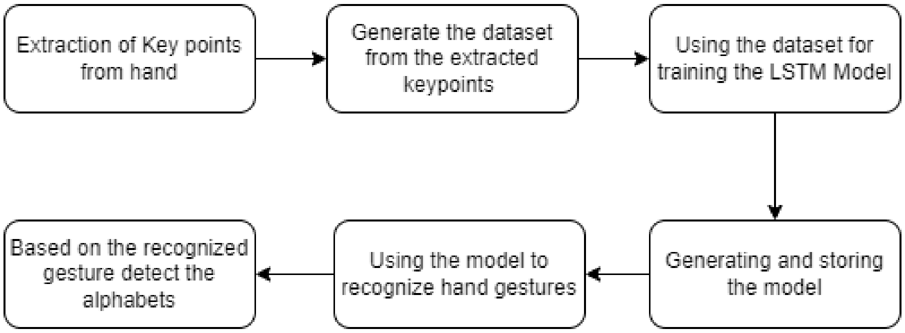

Extraction of Key points from hand
Generate the dataset from the extracted keypoints
Using the dataset for training the LSTM Model
Generating and storing the model
Using the model to recognize hand gestures
Based on the recognized gesture detect the alphabets

Figure 3: Flowchart of Phase-II

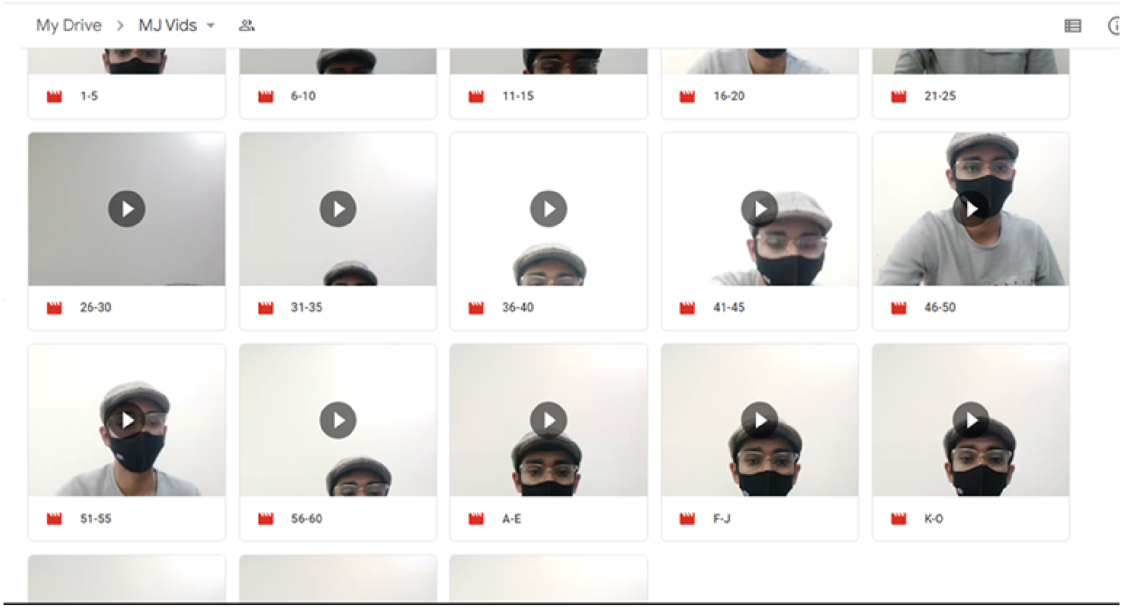


Figure 4: Dataset creation

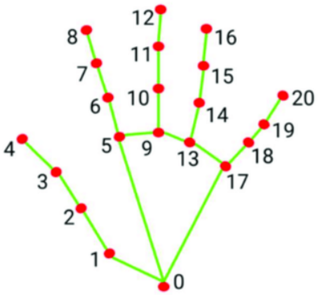
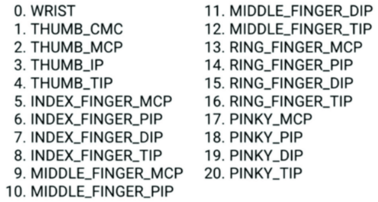


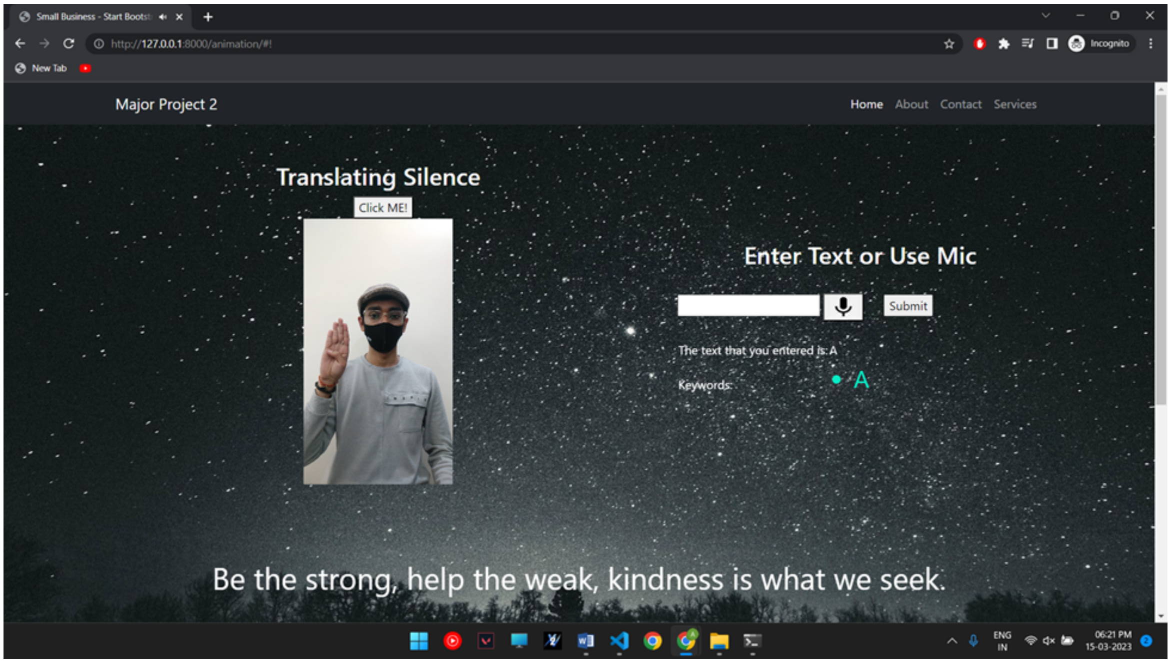
Figure 5: Hand Landmarks



0. WRIST
1. THUMB_CMC
2. THUMB_MCP
3. THUMB_IP
4. THUMB_TIP
5. INDEX_FINGER_MCP
6. INDEX_FINGER_PIP
7. INDEX_FINGER_DIP
8. INDEX_FINGER_TIP
9. MIDDLE_FINGER_MCP
10. MIDDLE_FINGER_PIP
11. MIDDLE_FINGER_DIP
12. MIDDLE_FINGER_TIP
13. RING_FINGER_MCP
14. RING_FINGER_PIP
15. RING_FINGER_DIP
16. RING_FINGER_TIP
17. PINKY_MCP
18. PINKY_PIP
19. PINKY_DIP
20. PINKY_TIP

Figure 6: Types of hand landmarks

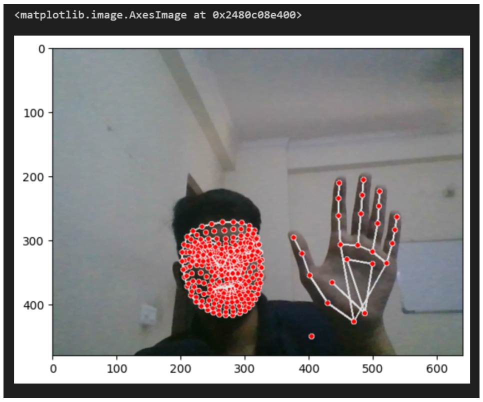
hard-of-hearing people depends on precise recognition of sign language motions. However, it is crucial to highlight that we were only able to test and translate a tiny collection of three words because of the laptop's constrained capabilities, including poor camera quality and fewer GPU compute capability. We were unable to scale up the system to handle a bigger dataset or more intricate sign language movements due to the laptop's low capacity. Figure 8, shows the LSTM parameters we have used in our model implementation. During the second phase of the study, we were able to totally and accurately translate three sign language words into text using the LSTM and MediaPipe method. This is a significant achievement



The text that you entered is: A

Be the strong, help the weak, kindness is what we seek.

Figure 7: Snapshot of website instantly translating text and voice to sign language



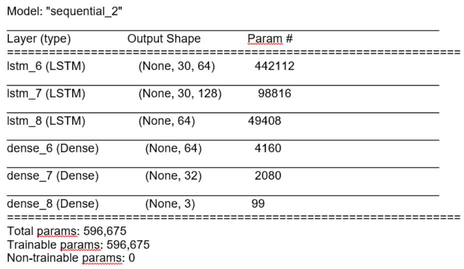
```
<matplotlib.image.AxesImage at 0x2480c08e400>
```

Figure 8: Real Time Feature Extraction

since accurate understanding of sign language gestures is essential for efficient communication with the deaf and hard-of-hearing. Figure 9 shows the sign detection process for the three words.

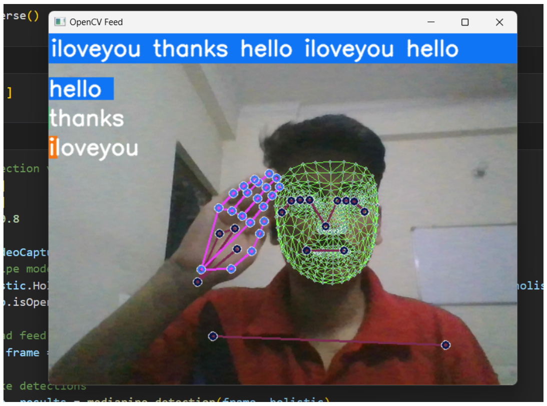
5 CONCLUSION & FUTURE WORK

The present work created a system that can translate American Sign Language into text or voice and the other way around. The technology would offer a reliable method of communication, closing the gap in understanding between the hearing community and the deaf and mute population. A system was developed using recurrent and convolutional neural networks for gesture translation and recognition. The system's user interface was designed to be straightforward and welcoming. We were able to receive an accuracy of 1.0 for the 3 words classification for this LSTM Model.



```
Model: "sequential_2"
_________________________________________________________________
Layer (type)              Output Shape            Param #
=================================================================
lstm_6 (LSTM)               (None, 30, 64)          442112
_________________________________________________________________
lstm_7 (LSTM)               (None, 30, 128)          98816
_________________________________________________________________
lstm_8 (LSTM)               (None, 64)            49408
_________________________________________________________________
dense_6 (Dense)              (None, 64)             4160
_________________________________________________________________
dense_7 (Dense)              (None, 32)             2080
_________________________________________________________________
dense_8 (Dense)              (None, 3)             99
=================================================================
Total params: 596,675
Trainable params: 596,675
Non-trainable params: 0
```

Figure 9: LSTM Parameters



```
iloveyou thanks hello iloveyou hello
hello
thanks
iloveyou
```

Figure 10: Sign Detection

Convolutional neural networks were utilized for gesture identification and recurrent neural networks for translation to recognise and translate ASL gestures with excellent accuracy. The system's utility was increased by the realization that all users could use its user interface. The system created can act as a springboard for additional study and creation in assistive technology. We can build a more accessible and inclusive society where everyone has an equal chance to communicate and succeed with continuous efforts in this area,

The system may be expanded to recognize and translate new sign languages in the future, and its accuracy and speed can be increased. The system can also be integrated with other communication technologies, including video conferencing, to give the deaf and mute community a more comprehensive communication alternative. Several enhancements are possible in the current work like -enhancing the precision and utility of the sign language translation system, enhancing the quality of the input data, which is crucial for precise gesture identification and mapping. To do this, more sign language motions will need to be gathered, and the input quality will need to be improved utilizing better resolution cameras and more sophisticated processing methods. The system's performance, particularly in terms of processing speed and accuracy, will be another area of focus. To increase the effectiveness and accuracy of the system, this will entail investigating various machine learning methods and optimization strategies. Another improvement can be done to enhance the quality of the input data, which is essential for accurate gesture identification and mapping. More sign language gestures must be collected to accomplish this, and the input quality must be enhanced using higher-resolution cameras and more advanced processing techniques. To ensure that the system is created with the requirements and preferences of the deaf and hard of hearing community in mind, we also intend to work with them. To do this, it will be necessary to work closely with community members to comprehend their communication needs and incorporate their comments into the system's design and development.

REFERENCES

[1] Stokoe Jr, W. C. (2005). Sign language structure: An outline of the visual communication systems of the American deaf. Journal of deaf studies and deaf education, 10(1), 3-37.

[2] Nikam, A. S., & Ambekar, A. G. (2016, November). Sign language recognition using image based hand gesture recognition techniques. In 2016 online international conference on green engineering and technologies (IC-GET) (pp. 1-5). IEEE.

[3] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306.

[4] Fenlon, J., & Wilkinson, E. (2015). Sign languages in the world. Sociolinguistics and deaf communities, 1, 5-28.

[5] Wen, F., Zhang, Z., He, T., & Lee, C. (2021). AI enabled sign language recognition and VR space bidirectional communication using triboelectric smart glove. Nature communications, 12(1), 5378.

[6] Fernandes, L., Dalvi, P., Junnarkar, A., & Bansode, M. (2020, August). Convolutional neural network based bidirectional sign language translation system. In 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 769-775). IEEE.

[7] Sharma, K., Aaryan, K. A., Dhangar, U., Sharma, R., & Taneja, S. (2022, November). Automated indian sign language recognition system using lstm models. In 2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) (pp. 461-466). IEEE.

[8] Sageengrana, S., Kumar, V. R., & Kumaresan, N. (2022, January). Smart Phone Based Bidirectional Communication System for Hearing & Speaking Impaired Community and Normal Society. In 2022 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-7). IEEE.

[9] Zhou, Z., Tam, V. W., & Lam, E. Y. (2021). SignBERT: a BERT-based deep learning framework for continuous sign language recognition. IEEE Access, 9, 161669-161682.

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

[11] Lin, H. I., Hsu, M. H., & Chen, W. K. (2014, August). Human hand gesture recognition using a convolution neural network. In 2014 IEEE International Conference on Automation Science and Engineering (CASE) (pp. 1038-1043). IEEE.

[12] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., ... & Grundmann, M. MediaPipe: A Framework for Building Perception Pipelines.

[13] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306.

[14] Wu, J. (2017). Introduction to convolutional neural networks. National Key Lab for Novel Software Technology. Nanjing University. China, 5(23), 495.

[15] Korenius, T., Laurikkala, J., Järvelin, K., & Juhola, M. (2004, November). Stemming and lemmatization in the clustering of finnish text documents. In Proceedings of the thirteenth ACM international conference on Information and knowledge management (pp. 625-633).

[16] Agarap, A. F. M. Deep Learning using Rectified Linear Units (ReLU).

[17] Diederik, P. K. (2014). Adam: A method for stochastic optimization.