2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT) | 979-8-3503-7131-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/InCACCT61598.2024.10551000

Hand Gesture Recognition and Translation for International Sign Language Communication using Convolutional Neural Networks

Sivamohan S
St. Joseph's Institute of Technology
Chennai, India
sivamohan7@gmail.com

Anslam Sibi S
St. Joseph's Institute of Technology
Chennai, India
anslam.sibi@gmail.com

Divakar T R
St. Joseph's Institute of Technology
Chennai, India
trdiva2608@gmail.com

Jagan S
St. Joseph's Institute of Technology
Chennai, India
j2a7g0a2n2003@gmail.com

***Abstract*—A fundamental and natural form of communication, hand gestures are especially important for those who struggle with hearing loss. This Research provides a novel real-time technique for understanding hand motions used in sign language, specifically for International Sign Language (ISL), by using Convolutional Neural Networks (CNNs). This ground-breaking method aims to translate hand motions seen on a live video feed, making communication easier for those who use sign language. It's amazing how well the system recognizes ISL movements and converts them into audible speech, removing a major barrier to communication between those who use sign language and others who don't. This technology innovation promotes inclusivity by facilitating effortless communication for people with all linguistic and auditory skills. Extensive testing conducted under various settings highlights the system's resilience, exhibiting a remarkable accuracy rate of 97.85%. This accomplishment confirms the system's dependability in practical applications by highlighting its feasibility despite changing background intricacies and illumination subtleties.**

***Index Terms*—Gesture recognition, Speech synthesis, Machine learning, Sign language to speech conversion, People with language barriers.**

I. Introduction

Communication with hand gestures is among the oldest modes of human expression. Hence, it goes beyond linguistics by enriching the very fabric of human interaction across cultures. A fascinating division in computer vision technology is the automatic recognition of human gestures from live camera feeds. International Sign Language Communication refers to a standardized system of communication used by deaf individuals from different countries to communicate with each other. It's designed to bridge language barriers that might exist between different sign languages used around the world. International Sign is not a universal language but rather a set of gestures, signs, and expressions that are commonly understood among deaf individuals from various linguistic backgrounds. International Sign Language Communication means ensuring that your system can accurately interpret and translate not just one specific sign language (e.g., American Sign Language or British Sign Language) but also the common elements of International Sign that may be used by signers from different linguistic backgrounds. This could involve developing a robust recognition system that can identify and interpret signs that are common across various sign languages, as well as incorporating a translation component that can generate spoken language output based on the recognized signs. This study therefore aims at making a significant contribution to this area through the development of a robust and flexible system based on convolutional neural network (CNN) architecture for real-time identification of hand gestures associated with human activities. Hand gesture recognition comprises three types: fingerspelling, word-level sign vocabulary, and non-manual features[1]. The main purpose of this paper is to recognize and understand hand movements done by those people whose primary language is sign language among other applications that can be developed using this technology as well. This ingenious method goes beyond the confines of traditional gesture identification. Other than accurately recognizing ISL (International Sign Language) hand gestures, this also changes these signs into voice thus transcending the visual aspect. This system aims to open up communication barriers and create an inclusive environment where those who communicate through sign language can communicate with other people using speech or hearing ability without any problems. This system is a major step to ensure that everyone, irrespective of whether they can hear or speak, can communicate with each other.

979-8-3503-7131-4/24/$31.00 ©2024 IEEE

II. Related Works

A. Hand Gesture Recognition

The previous studies of hand gesture recognition have set a strong base for research. There have been several approaches including template-based methods, feature-based methods, and most recently, deep learning techniques. Template-based methods use pre-defined templates of hand gestures that are compared to input to recognize them. Feature-based methods extract multiple features of hand gestures using machine learning algorithms for classification. There were some sorts of classifications passed through like finger spelling, word-level sign vocabulary, and non-manual features as well. They have become a dominant method in hand gesture recognition due to their ability to learn significant features from raw image data automatically - Convolutional Neural Networks (CNNs). Researchers worldwide have employed CNNs in various sign language recognition systems from International Sign Language (ISL) to custom-made sign languages [1,2]. This has remarkably improved the field's current position.

B. Sign Language Translation

There have been attempts to create sign language translation systems to bridge the communication gap between deaf and hearing individuals. Some of the existing sign language recognition and translation systems range from wearable devices to those based on cameras[3,4]. The wearable ones are usually equipped with sensors that are used for tracking a user's hand movements by using accelerometers and then translating them into sign language [5]. Also, Kinect uses depth sensors in its camera-based solution to record body and hand movements, which help it recognize gestures and then convert them either into texts or sounds of speech.

C. Challenges in Existing Systems

These moves are commendable but have their restrictions. These techniques often get stuck because of a lack of flexibility. [5] Wearable systems, while portable, maybe less reachable and may not be adjusted to various sign languages. In most occasions, camera-based systems that guarantee adaptability require appropriate lighting conditions within a controlled environment to recognize accurately. The problem of translating sign language in real-time into speech is still wide open for betterment.

III. Methodology

To achieve in-air hand gesture recognition and generate the voice of any recognized sign, This system employs advanced methodologies in computer vision and deep learning. The proposed system is based on International Sign Language (ISL) for real-time hand gesture recognition and the translation of these gestures into audible voice. Thus, this proposed system uses Finger spelling type which is one of hand gesture classification. These methods include data acquisition, preprocessing and feature extraction, gesture classification by CNNs and text-to-speech conversion[4,6].

A. Data Procurement

The Data Procurement module employs vision-based techniques to effectively acquire hand gesture information via the computer webcam as the only input device. With a camera frame rate per second, individual frames are captured, making it possible for fast data acquisition hence enabling real video frames that are extracted and processed in a looped fashion. The idea of using this vision-based technique is optimized for responsiveness and establishes a natural and cost-effective interaction between users and computers [6]. This gives way to a subsequent feature extraction module that processes input data in terms of images, thus enabling relevant hand gestures to be extracted. It is noteworthy that this approach focuses on one hand without using gloves to avoid potential conflicts in extracting features from multiple hands[6]. With this straightforward approach, prediction accuracy is increased for more productive user-computer interaction.

B. Data Normalization and Feature Selection

In order to locate hands in processed frames, this module mainly uses the HandDetector class from the cvzone module. The HandDetector uses aspect ratio and filters in an advanced algorithm to improve hand detection accuracy. Using the bounding box coordinates of the detected hand, the module extracts the Region Of Interest (ROI) after a successful detection. This is necessary to isolate the hand region for additional processing. Another important thing to remember is to make hand landmarks visible at runtime on a white background canvas. A graphical depiction of the hand move that was identified is produced by using this canvas as a rendering area for hand landmarks[7]. These twenty-one landmarks, each representing one of the twenty-one attributes, are connected by lines to form a complex representation of the placement and movement of hands. Aspect ratio and additional filters are used to improve the accuracy of hand gesture identification. The software creates a comprehensive depiction of the hand's anatomy and attitude by creating an entire image of the 21 hand landmarks on canvas (Fig. 1). The aspect ratio, filters, and this thorough representation ensure that the hand motions that are displayed are accurate enough for additional examination.

C. Gesture Interpretation

A proposed neural network architecture that is based on a well-designed Convolutional Neural Network (CNN) specifically optimized for image classification tasks. The initial convolutional layer has 32 filters with a 3x3 kernel size and generates feature maps. By using 32 filters in this way, followed by max-pooling and generating feature maps, the next one in this sequence also involves two other convolutional layers that use 16 filters each before additional max-pooling (Fig. 2). The last dense layer provides an output prediction where each node corresponds to a class in the classification task. This subpart helps the model to differentiate between nearly similar alphabets.

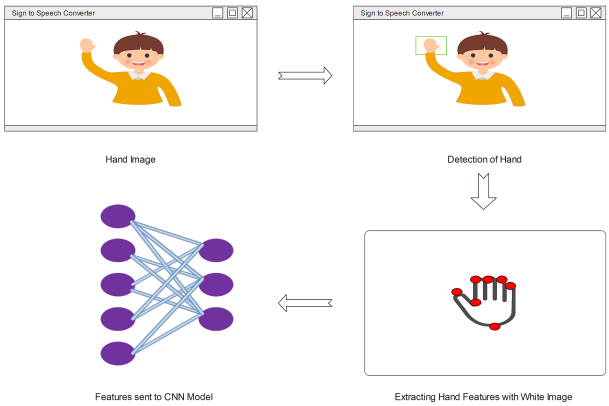

Fig. 1: Feature Extraction

D. Speech Synthesis

The system goes beyond simply recognizing sign language gestures. Additionally, it incorporates a text-to-speech module that uses PyTtsx3 to convert the motions that are identified into spoken language. With this ground-breaking technology, sign language users can now effectively communicate not just with others, but also with people unfamiliar with sign language (like ISL)[8]. This feature promotes inclusivity across a range of contexts, including social interactions, healthcare, and education, by facilitating easy communication between people of different language backgrounds.

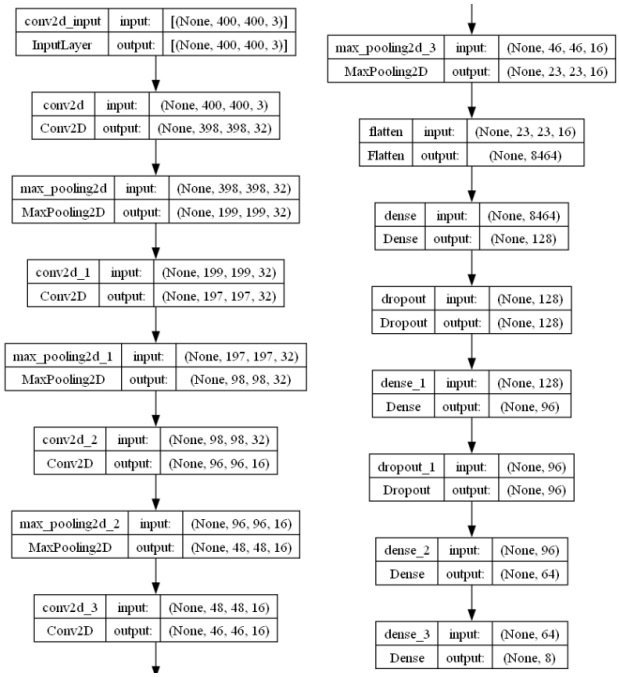

Fig. 2: The Proposed CNN Model

IV. System Architecture

A flexible architecture made up of interconnected modules is incorporated in the suggested architecture for real-time International Sign Language (ISL) gesture recognition and translation. As seen in Fig. 3, these modules are painstakingly designed to guarantee smooth connection and effective data processing across the system. Data Procurement, Normalization and Feature Selection, Gesture Understanding, and Speech Synthesis are the four main components that make up the architecture. The entry point is the Data Procurement module, which is in charge of obtaining input data. Typically, this data is in the form of video or picture feeds that record ISL gestures. The collected data is then subjected to feature selection and normalization in the next module. In this case, preprocessing is done on the input data to normalize it and extract pertinent features that are necessary for gesture identification. The Gesture Understanding module receives the processed data after which it applies advanced algorithms to the derived features in order to decipher the meaning of the ISL gestures. To identify and comprehend the user's gestures in real time, this module is essential. After the ISL motions are understood, the Speech Synthesis module receives the interpreted data from the Gesture Understanding module. Since the recognized motions are translated into spoken language output, this enables the deaf user to converse with hearing individuals intelligibly. The modules are intimately integrated throughout the architecture, which makes it easier for data to move smoothly from one step to the next. This interconnectivity guarantees the system's smooth operation and permits precise real-time identification and spoken language translation of ISL motions.

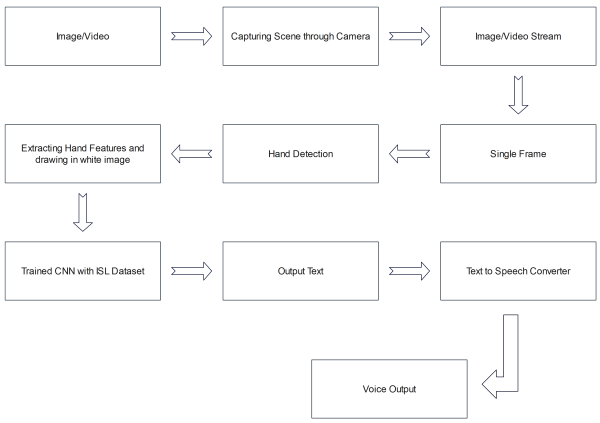

Fig. 3: Architecture Diagram

V. Implementation

Three main modules make up the system that was put into use: speech synthesis, translation, and gesture recognition (Fig. 4). The system uses a webcam for gesture capture implemented with a Convolutional Neural Network (CNN) to provide real-time hand gesture recognition[9]. Convolutional

Neural Networks (CNNs) are a powerful force in computer vision, altering the way this system analyzes images and movies. The secret to their amazing power is the meaningful visual elements they can extract, which gives them the ability to ”see” and comprehend an image’s information in a way that older approaches frequently find difficult. The specific layers in CNN design called ”convolutional layers” are the source of this extraordinary power. Consider these layers as painstaking filters that carefully examine the input image, multiplying each element by its associated pixel value. By identifying particular features in the image, CNNs iteratively construct a hierarchy of knowledge over time. While subsequent layers learn to identify increasingly intricate patterns and combinations of these traits, initial layers usually concentrate on identifying basic shapes and edges. To address overfitting and manage data size, pooling layers are employed, downsampling the data while retaining crucial information[10]. Finally, fully connected layers process the refined image representations to perform classification or prediction tasks. This distinctive architecture equips CNNs with several noteworthy advantages. One critical benefit lies in automated feature extraction, eliminating the traditionally laborious and time-consuming process of manual feature engineering. Additionally, CNNs exhibit translation invariance, meaning they can identify objects or patterns regardless of their location within the image[11].

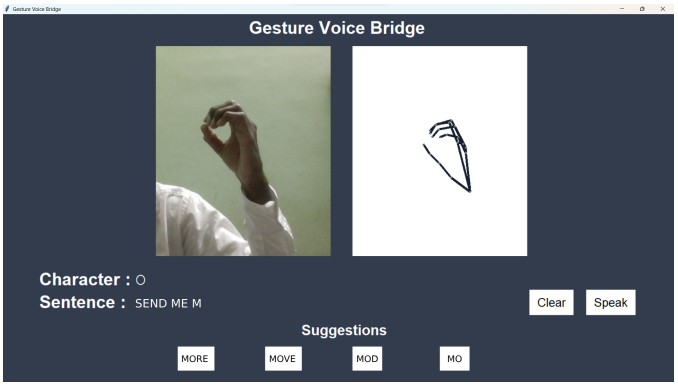

Fig. 4: Sample Conversion Module Output

VI. Experimental Results

A. Dataset and Training

To experiment with the system, a wide array of ISL movements have been used through which each hand’s landmark over all 26 alphabets was put into use to ensure that there is enough input variety aimed at improving the accuracy (Fig. 5). The dataset was collected from Kaggle, but we transformed and modified the datasets for our model training and testing modules. The dataset covers different lighting, backgrounds, and some user-related specific characteristics to make it close to reality. For training purposes, employed CNN Structured Algorithm setting up an epoch value of 60. Besides, a test on gloveless datasets consisting of hand landmarks that do not represent actual hands and the datasets were split into two such as train and test set for checking the accuracy of the model (Fig 6).

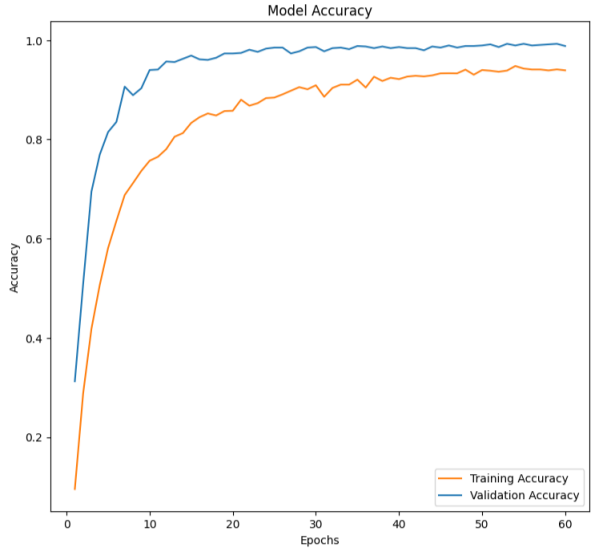

Fig. 5: Training and Validation Accuracy

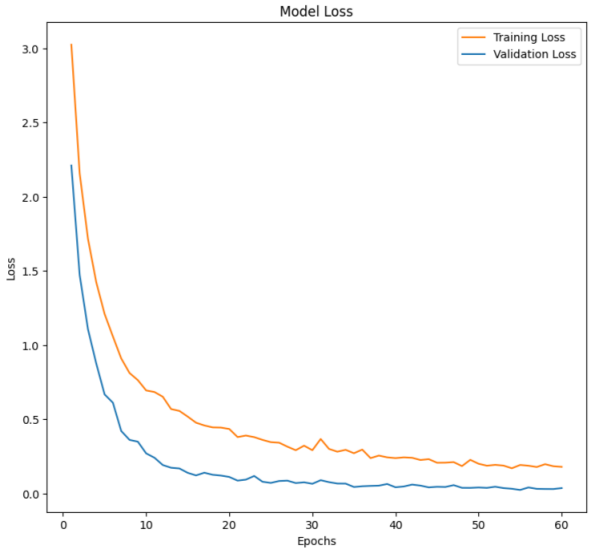

Fig. 6: Training and Validation Loss

B. Accuracy and Classification

Concerning the CNN model, the primary performance metric for the system is gesture recognition and classification accuracy, where it reaches an impressive 97.85%. The confusion matrix (Fig. 7) represents the comparison of these predictions to ground truth labels and the calculation of this accuracy is based on twenty-six alphabets which shows how strong this system is when interpreting ISL gestures. HandDetector class from cvzone library has its accuracy dependent on light

variation and background complexity, which gives it more than 95% in the case of good lighting. In situations where backgrounds are complex, however, the HandDetector class achieves about a fig less by 50%. Hence training takes place in a properly lit environment with minimal interference from the surroundings since this will lead to higher precision in terms of these two factors such as good light and less complex backdrops.

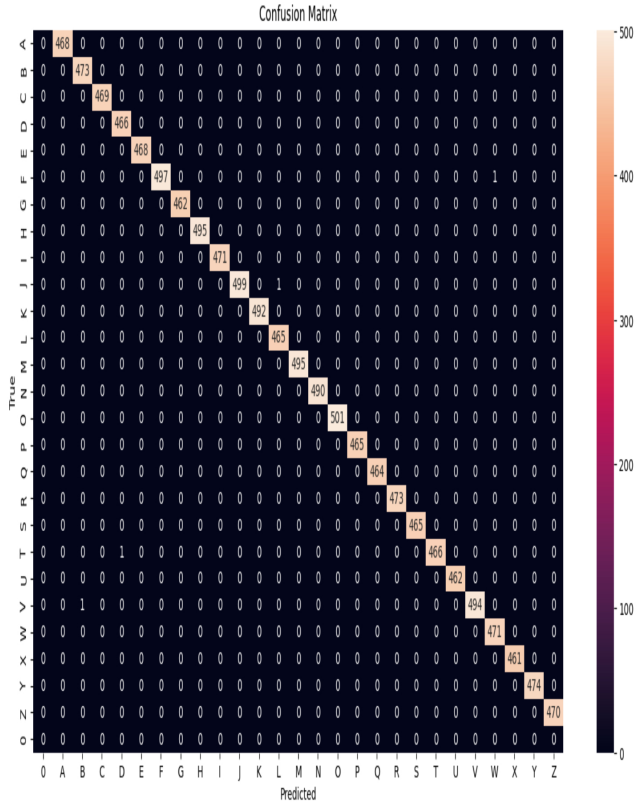

Fig. 7: Confusion Matrix

C. Real-Time Recognition

The system depends on the live operation performance given. It is a system that works in real-time, processes incoming image data, and almost instantly provides results of gesture recognition for all frame rates predicted by video. The success of experiments showed that milliseconds were enough for us to perform this type of recognition smoothly, so it was possible to generate an interactive communication environment. Since frame rate is a key parameter in performance, choose a 30fps camera which they can use to minimize latency in communication. This makes it perfect for interactive and inclusive communications since its processing ability is in real-time while supported by an i5 11th-generation processor. This synergy between cutting-edge technology and real-time processing capabilities positions the system as a versatile and effective tool.

D. Quality of Sign Language Translation

Furthermore, sign language translation quality was assessed by comparing the textual representation's accuracy produced by its output with its synthetic voice output naturalness and intelligibility. The system usually generated precise text representations for identified gestures. For proper communication, it is necessary to have a good and reliable speaker for synthesizing speech. The voice output from the text-to-speech conversion module that employed the pyttsx3 library was clear, expressive, and easy to understand. This high level of translation ensures effective communication and comprehension. Moreover, the enchant library is employed in this software to provide word suggestions that facilitate fast writing. Other libraries such as Hunspell are also recommended depending on the user operating system type. Furthermore, this application smoothly deals with human errors whereby correct spellings are suggested to enhance better communication.

VII. Conclusion and Future Work

This paper introduces an innovative real-time International Sign Language (ISL) gesture recognition and translation system that uses Convolutional Neural Networks (CNNs). The system's excellence is in its ability to recognize ISL gestures with high accuracy 97.85%, making it indeed a reliable tool for people who rely on sign language. By transforming the recognized ISL gestures into audible voice, the technology transcends the visual medium and bridges communication gaps between deaf persons and those using spoken language. The point about this technology is that it can potentially create an all-inclusive and communicative world where language would be borderless. The model was trained achieving a training accuracy of 94.10% and a validation accuracy of 99.25%. The corresponding training loss was 0.1834 while the validation loss was 0.0297. Therefore, these results demonstrate that the model effectively learned underlying patterns in the training data as well as showed high generalization performance on the validation set by utilizing CNN layers having a 3X3 Kernel Size for filtering. It must undergo rigorous testing and refinement based on user feedback to ensure usability and effectiveness. Rigorous testing and refinement driven by user feedback guarantee the usability and efficacy of its use. The adaptability plus robustness of this system makes it applicable in real-world scenarios. This shows that this system can be used in real-life scenarios since it is adaptable as well as robust enough. The future holds great promise in the areas of sign language identification and translation. The research provides a foundation for further investigations into expanding vocabulary, identifying more sign languages and dialects, and improving flexibility in various real-life situations. Such technology will continue to evolve through constant collaboration with the ISL community as well as user testing. This study can be improved through the expansion of the dataset to include a more comprehensive array of International Sign Language (ISL) letters and words. By augmenting the dataset, there is a potential to enhance the accuracy of the recognition system

while simultaneously reducing loss. Furthermore, incorporating additional words and phrases may strengthen the system's capability to predict complete expressions. Subsequently, the integration of a text-to-speech engine could facilitate the conversion of predicted expressions into audible speech. In future research, efforts may be directed toward developing methods to translate ISL gestures into commands understandable by robots or machines. This advancement would enable individuals to interact with robots using ISL gestures as commands, thereby contributing to enhanced accessibility and usability in human-robot interaction scenarios.

VIII. REFERENCES

[1] A. S. Nandhini, D. Shiva Roopan, S. Shiyaam, and S. Yogesh, "Retraction: Sign Language Recognition Using Convolutional Neural Network," Journal of Physics: Conference Series, vol. 1916, no. 1. IOP Publishing Ltd, May 27, 2021. doi: 10.1088/1742-6596/1916/1/012091.

[2] Y. Dhamecha, R. Pawar, A. Waghmare, and S. Ghosh, "Sign Language Conversion using Hand Gesture Recognition," 2023 2nd International Conference for Innovation in Technology, INOCON 2023, 2023, doi: 10.1109/INOCON57975.2023.10101099.

[3] S. Shivdikar, J. Thakur, and A. Agarwal, "Hand Gesture Recognition and Translation Application," International Research Journal of Engineering and Technology, 2022, [Online]. Available: www.irjet.net.

[4] A. Rathi, S. Pasari, and S. Sheoran, "Live Sign Language Recognition: Using Convolution Neural Networks," 8th International Conference on Advanced Computing and Communication Systems, ICACCS 2022, pp. 502–505, 2022, doi: 10.1109/ICACCS54159.2022.9785357.

[5] H. Limaye, S. Shinde, A. Bapat, and N. Samant, "Sign Language Recognition using Convolutional Neural Network with Customization," SSRN Electronic Journal, Jul. 2022, doi: 10.2139/SSRN.4169172.

[6] M. Naveenkumar, S. Srithar, G. R. Kalyan, E. Vetrimani, and S. Alagumuthukrishnan, "Hand Sign Recognition using Deep Convolutional Neural Network," 4th International Conference on Inventive Research in Computing Applications, ICIRCA 2022 - Proceedings, pp. 1159–1164, 2022, doi: 10.1109/ICIRCA54612.2022.9985691.

[7] K. K. Dutta and S. A. S. Bellary, "Machine Learning Techniques for Indian Sign Language Recognition," International Conference on Current Trends in Computer, Electrical, Electronics and Communication, CTCEEC 2017, pp. 333–336, Sep. 2018, doi: 10.1109/CTCEEC.2017.8454988.

[8] P. Rai, A. Alva, G. K. Mahale, J. S. Shetty, and M. A. N, "International Journal of Computer Science and Mobile Computing GESTURE RECOGNITION SYSTEM," 2018. [Online]. Available: www.ijcsmc.com

[9]] J. L. Crowley, ACM Digital Library., and ACM Special Interest Group on Computer-Human Interaction., Proceedings of the 2009 international conference on Multimodal interfaces. ACM, 2009.

[10] E. A. Kalsh and N. S. Garewal, "Sign Language Recognition System." [Online]. Available: www.ijceronline.com.

[11] S. Bele, A. Shinde, K. Sharma, and A. Shinde, "SIGN LANGUAGE RECOGNITION SYSTEM USING MACHINE LEARNING." [Online]. Available: www.ijirmps.org