Journal of Ubiquitous Computing and Communication Technologies (ISSN: 2582-337X)
www.irojournals.com/jucct/

Machine Learning based ISL Identification and Translation

Thirumahal R.[1], Aswath Harish Jayaprakash[2], Shiva Prakash P.[3], Yuvaraj Kesavan P.[4], Chirenjeevi M.[5], Siva M.[6]

[1]Assistant Professor, [2-6]Under Graduate, Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, India

E-mail: [1]trk.cse@psgtech.ac.in, [2]aswathharish03@gmail.com, [3]delhichandru@outlook.com, [4] yuvarajkesavan2003@gmail.com, [5]chirenjeevi415@gmail.com, [6]sivapsg2020@gmail.com

Abstract

Sign language is an essential means of communication for the deaf and hard-of-hearing community. However, effective communication between sign language users and those unfamiliar with sign language can be challenging. The primary goal is to utilize the machine learning to automatically identify sign language gestures and translate them into easily understandable formats. This research presents a comprehensive sign language detection system that captures sign language gestures, detects them, and provides output in text using LSTM (Long Short-Term Memory) and Transformers with an accuracy of 79%. This multimodal approach ensures the system helps in understanding the sign language.

Keywords: Sign Language Recognition, Indian Sign Language (ISL), Machine Learning, LSTM (Long Short-Term Memory), Transformers, Real-time Gesture Recognition, Feature Extraction, Computer Vision, ISL Dataset, Gesture-to-Text Translation, Human-Computer Interaction (HCI).

1. Introduction

The ability to communicate effectively is essential for transferring thoughts and feelings both among groups and between individuals. A skilful interchange of ideas encourages progress and positive thinking. Language is a vital instrument for communication

DOI: https://doi.org/10.36548/jucct.2024.4.003
Received: 11.11.2024, received in revised form: 04.12.2024, accepted: 21.12.2024, published: 30.12.2024

because it includes both verbal and nonverbal behaviors. To communicate with others, deaf and mute people use sign language, which involves body movements and hand gestures. Unfortunately, not everyone can understand sign language, which makes it difficult for people who are deaf, hearing-impaired, or speech-disabled to express themselves and communicate with others. As a result, this obstruction creates a barrier between those who are deaf or mute and those who are not. So, one effective way to overcome this obstacle is to create a sign language recognition system. There are now many studies being conducted in this area that will greatly help society. The world is always changing as a result of advances in the quickly changing technology landscape of today. Because of this, the interpreter fills an essential role in ensuring that everyone, regardless of disability, has access to equal opportunities.

There are numerous distinctive languages on our planet, each associated with a particular place. Similar to how spoken languages vary geographically, so do sign languages, according to regional languages. In the context of this research, communication takes place in English while the Indian Sign Language (ISL) is used. The system's function involves real-time recognition of hand gestures, captured through the camera. The system’s output will be in the form of text which makes the opposite person understand the sign language.

2. Related Work

In the last few years, several research have contributed to the field of machine learning approaches for sign language recognition and multimodal translation systems. Pragati Goel et al. [1] propose a desktop program that employs a webcam to record and decipher ASL movements in real-time, then uses Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models to translate the gestures into text and speech. People with speech or hearing difficulties can communicate more effectively thanks to this method. The computational demands of training CNN and LSTM models, acquiring a diverse dataset, and managing geographical differences in ASL are obstacles, too. Notwithstanding these challenges, the system demonstrates how AI might improve inclusion.

Mallika Garg et al [2] provide a system for recognizing hand gestures in which a convolutional residual network is used to first segment the hand region. The recognition network achieves 98.75% accuracy on upgraded data by combining look and shape information for gesture classification. Convolutional neural networks (CNN), Atrous Spatial Pyramid Pooling (ASPP), and data augmentation are important elements. By using these

strategies, the model is able to overcome data bias and overfitting, allowing for precise interpretation of human body language.

Kohsheen Tiku et al [3] suggests a Java-based OpenCV and Histogram of Oriented Gradients (HOG) for feature extraction in a real-time ASL recognition system. The model uses a Support Vector Machine (SVM) for classification, along with PCA for dimensionality reduction and a variety of parameters for image preprocessing, including skeletonization, Canny edges, and contour mask. With 98.75% accuracy, the system was trained on a condensed ASL Kaggle dataset (100 images per class). Effective image segmentation is one of the main issues addressed, where performance is improved using the Otsu approach for thresholding, which maximizes object-background separation.

Shiva Shankar Reddy et al [4] creates a web-based sign language recognition methodology that combines object observation for letter identification, deep learning models, and image preprocessing. The system shows promise for expanding to complete gesture translation and recognizes sign language characters with 95% accuracy. Using multiclass SVM models with linear kernels to increase accuracy for one-handed (56%) and two-handed (60%) alphabets is one of the challenges. The work emphasizes the value of strong classification techniques, which may find use in more general sign language recognition.

Ankit Ojha et al [5] creates a fingerspelling sign language translator with Convolutional Neural Networks (CNNs) that can translate ASL motions with 95% accuracy. The system's limitations include its need on ASL knowledge and the impracticability of gloves for users with different skin tones, despite its superiority in finger spelling. One of the system's strengths is its capacity to adapt to new sign languages through CNN retraining and dataset expansion, which encourages inclusivity for those with hearing disabilities. The difficulties, however, lie in improving accessibility for users who are not familiar with ASL signals and guaranteeing reliable performance in a variety of lighting scenarios and backgrounds.

With 31,000 ISL-English sentence/phrase pairs, AbhinavJoshi et al. [6] present ISLTranslate, a ground-breaking dataset for ISL translation that is the largest of its type. With its substantial potential to close communication gaps among the hard-of-hearing community in India and around the world, this dataset provides a solid basis for the creation of statistical sign language translation systems. By employing a transformer-based model to verify its effectiveness, the authors establish a standard for further study. The dataset's size and

extensive data cleaning are its strong points, but its dependence on a single pattern model and its aspirations for future model improvement and dataset growth are its weaknesses.

In contrast to American Sign Language (ASL), there is relatively less study on gesture recognition in Indian Sign Language (ISL), which is the focus of Rachana Patil et al. [7]. For image segmentation, the study uses methods like edge detection, skin-color detection, and context subtraction. In addition to noise reduction and morphological processes, gesture classification entails the extraction of features such as shape, contour, and color. Several classification methods are investigated, such as the Convolutional Neural Network (CNN), Principal Component Analysis (PCA), K-Nearest Neighbour, Support Vector Machine (SVM), Hidden Markov Model (HMM), and Artificial Neural Network (ANN). CNN achieves a remarkable 95% validation accuracy. The work draws attention to the difficulties in recognizing sign language, such as the necessity for real-world performance validation in a variety of circumstances and fluctuating loss and accuracy trends during training.

Konstantinos M et al [8] combines the WLASL and ASLLRP datasets. The empirical evaluation shows a 5% improvement in recognition rate when the forward and backward video streams are fused in the late stage. The incorporation of curricular learning, a dynamic technique for estimating difficulty, also improves recognition potential. These results demonstrate how this strategy holds promise for developing sign recognition technologies.

The goal of the sign language recognition system created by Parama Sridevi et al. [9] was to enhance communication for those who are hard of hearing. The model uses MATLAB to convert hand motions recorded by a webcam into alphabetic letters in real-time, with an average accuracy of 85.2%. Because of its small size and ability to promote seamless interactions and improve accessibility, this technology can be used on a variety of devices. But there are still issues like a small vocabulary, trouble with intricate hand gestures, and the linguistic distinctiveness of many sign languages. To improve its applicability across various sign languages, more training and customization are required.

Using hand-tracking and Artificial Neural Networks (ANN), Akshatha Rani K and Dr. N. Manjanaik et al. [10] created a system for recognizing sign language that achieved 74% accuracy in classifying the ASL alphabet. For accessibility, the device also has speech output, which helps people who are blind or hard of hearing. Although it works well in a variety of

settings, there are drawbacks, such as sensitivity to gesture quality and the possibility of overfitting if the model is not trained with a variety of data.

Sunitha Nandhini et al. [11] provide a sign language recognition system that makes use of convolutional neural networks (CNNs) to translate sign language into text and speech. The system uses image preprocessing methods including background removal and skin tone modification to increase accuracy. The technique uses CNN-based classification, cropping, and greyscale conversion to reach an average accuracy of 90%.

In order to handle issues including motion, gesture size, background complexity, and lighting, Limaye, Heramba, and Shinde et al. [12] suggest a convolutional network-based sign language recognition system. The system uses OpenCV to handle camera data for hand identification and scaling using computer vision and machine learning. By translating sign language to English with an average accuracy of 83.76%, the model helps the deaf community and the general public communicate more effectively.

In order to recognize gestures in the Indonesian Sign Language System (ISLS), Khamid et al. [13] investigate integrating Leap Motion and Myo armband technologies. Leap Motion can track hand movements, but it has trouble with little gesture variations. By collecting finger positions, the Myo wristband, which monitors muscle activity, improves accuracy. The study classifies data from ten static and dynamic motions using Naïve Bayes. Accuracy is increased by combining the two modalities, indicating their potential for translating sign language. This study addresses issues including two-hand gestures and practitioner variability while highlighting the benefits of multimodal systems for the hearing impaired.

E. Rajalakshmi et al. [14] propose a hybrid deep neural network for recognizing gestures in Indian and Russian sign languages. The framework combines passive and non-passive joint tracking with spatial and temporal gesture feature extraction using a 3D neural network and attention-based Bi-LSTM. A modified autoencoder extracts discriminative features, enhancing gesture recognition. Experiments on an Indo-Russian dataset show the hybrid system outperforms existing models. Advantages include capturing manual co-articulations and nonmanual features, but challenges include the complexity of deep neural networks, limited generalization across languages, and difficulties in continuous recognition.

A desktop program created by Mrs. S. Chandra Gandhi et al. [15] records Indian Sign Language (ISL) motions with a webcam and converts them in real-time into speech and text. The research's main objectives are language translation and fingerspelling. It detects gestures using a Convolutional Neural Network (CNN), which makes feature detection effective without the need for explicit extraction. With the application's bidirectional translation feature, users can speak or make motions and get the corresponding outputs. The accuracy and performance of the system are improved by the CNN's efficacy and deep learning's advantages over conventional techniques. The FAST and SURF algorithms were used to successfully implement important techniques including image segmentation and feature detection by Victoria Adebimpe Akano and Adejoke O. Olamiti et al. [16]. Data collection using the KINECT sensor was the first step in the system's organized process, which was then followed by image segmentation, feature recognition, and extraction from Regions of Interest (ROIs). Text-to-speech (TTS) conversion was the final step in the procedure, which also involved supervised and unsupervised classification using K-Nearest Neighbour (KNN) algorithms. The successful integration of FAST and SURF with KNN produced both text and speech outputs, demonstrating the efficacy of unsupervised learning in precisely recognizing characteristics.

3. Proposed Method

3.1 Architecture

The training procedure begins with the careful preprocessing of the dataset, which includes the methodical grouping of video data into orderly folders, as shown in Figure 1. The next step is a laborious process that includes the accurate placement and ongoing monitoring of important points on people during the video to guarantee their accurate localization. By carefully mapping these key point sequences to particular actions or activities that take place in the videos, complete annotation is made possible. After obtaining the well-annotated dataset, the next step involves training the LSTM (Long Short-Term Memory) model. By taking advantage of this neural network architecture's natural affinity for sequence data, it is possible to identify and learn complex patterns from the mapped key point sequences, and then correlate those patterns with actions in the videos.

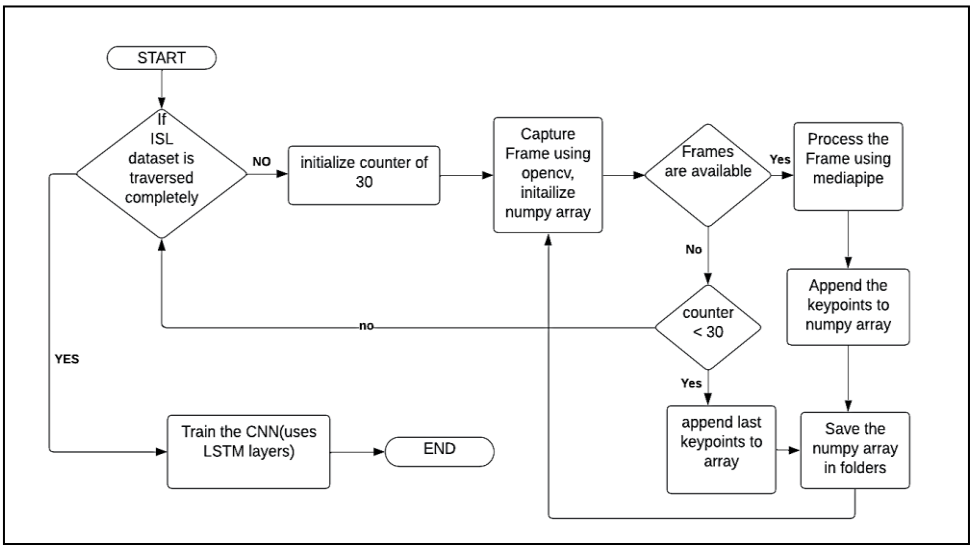

Figure 1. Feature Extraction and Training Architecture

Figure 2 explains the prediction process.

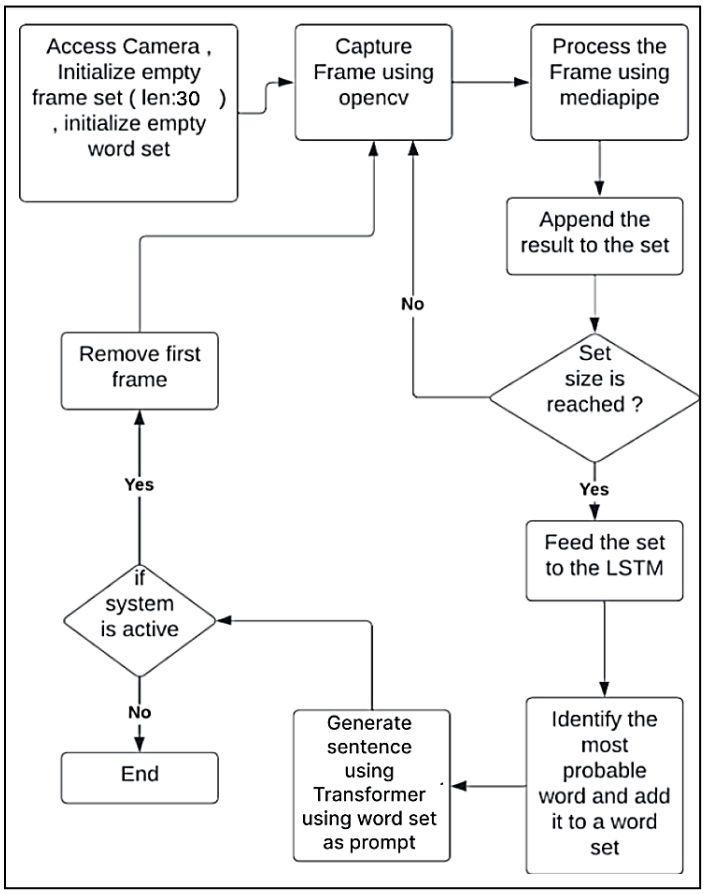

Figure 2. Prediction Architecture

First, the camera was initialized and set up to capture frames. An empty set was created to store the processed frames, and declare the size of the set. A frame was captured from the camera using the OpenCV library. The captured frame was processed using the MediaPipeHolistics. The processed frames are appended to the set. This step checks whether the set has reached its declared size. If it has not, the flow goes back to capture another frame. The set of processed frames are fed to the LSTM model. The LSTM model is a type of recurrent neural network that is well-suited for processing sequential data, such as video frames. The LSTM model outputs the most probable word that is represented by the sequence of processed frames. The flow ends if the system is not active. The first frame is removed and the new frame is captured.

3.2 Dataset Used

A valuable resource for the study of translation and communication in Indian Sign Language (ISL) is the ISLTranslate dataset. With a total of 558 words. This dataset aims to bridge the communication gap that exists between spoken language experts and the hard-of-hearing community. ISLTranslate is an invaluable resource for the natural language processing field since it provides thorough coverage of everyday communication vocabulary. The development of language processing systems for ISL could be greatly impacted by its availability, leading to improvements in accessibility, education, and sign language translation. Another approach attempted here is creating our own dataset to make it real-time using 900 videos.

4. Implementation

4.1. Data Processing

4.1.1 Determining Frames to Skip

An examination of the frame sizes in the dataset is the first step in the preprocessing procedure. Finding the best method for skipping frames while preserving data quality above a predetermined threshold is the goal. By varying the number of frames to be skipped, we iteratively investigate different frame-skipping choices to accomplish this.

4.1.2. Filtering and Adjusting Frames

Every frame in the dataset is examined and handled separately. The number of frames to be skipped is determined by their size and the threshold set at the beginning of the process. Frames that don't meet the quality threshold are skipped over in the process and won't be analyzed further. If any frames are not skipped, their sizes are changed to match up to the skipping plan, guaranteeing that they fulfill the predetermined standards for quality.

4.2. Feature Extraction

In Figure 3, particular important spots on the hand that are monitored and examined serve as reference points for determining the posture of hand movements. In order to match and identify the sign language gestures in various movies, it is essential to comprehend the movements and positions of the hands. These fundamental ideas form the basis of further matching and recognition procedures.

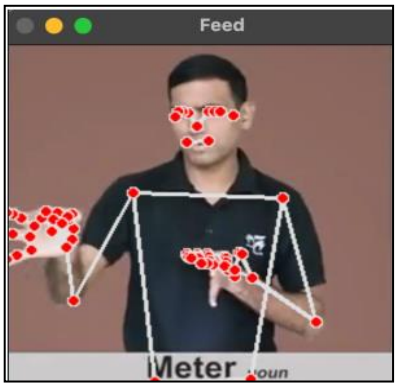

Figure 3. Keypoint Annotation during Feature Extraction

4.2.1. Extraction

During this process, particular features are extracted from the provided results data. The coordinates for each point are flattened into an array to record the locations of facial features if facial landmarks are found. In a similar vein, coordinates for any posture markers that are present are flattened along with visibility information. The left and right-hand landmarks follow the same process, with their corresponding coordinates being flattened into arrays. After that, these arrays are concatenated to create an extensive dataset that contains all of the pertinent facial, position, and hand landmark data.

4.2.2. Draw Postures

Visual representations of identified landmarks are superimposed over the provided image to improve it. The function draws particular items on the image by using the results data. Using tessellation patterns, facial landmarks are first drawn, defining the fine aspects of the face. The landmarks on the left and right hands are then illustrated, highlighting the relationships between them. Lastly, the function illustrates posture landmarks that call attention to the main points of the body and their relationships. The process of graphically mapping these landmarks onto the image yields a visualization that facilitates the comprehensive analysis of the input data by offering a clear and understandable depiction of the detected facial, hand, and body positions.

4.2.3. Detect Video

A pre-trained machine-learning model and an input image are used to start this process. The model uses complex algorithms to analyze the image and extract useful information, such as action recognition and object detection. After this examination, the function outputs two essential elements: the original image in its original form and the outcomes of the model's calculations. These outcomes capture the main ideas contained in the image, making it easier to analyze and use the information that has been extracted in the future.

4.3. Model

Figure 4 explains the layers of the LSTM model which is used here to deal with sequential data. Three stacked LSTM layers make up the model; the first two layers have 128 units each and return sequences to represent temporal dependencies. With 64 units, the third LSTM layer improves upon the feature representations. The network can learn complicated patterns from the LSTM outputs by introducing non-linearity through a dense layer with 256 units and a ReLU activation function. Depending on the particular application, a probability distribution over 55 classes for different categories is produced by the final dense layer with 55 units and a softmax activation function.

```
Model: "sequential_4"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm_9 (LSTM)               (None, 30, 128)           198144

 lstm_10 (LSTM)              (None, 30, 128)           131584

 lstm_11 (LSTM)              (None, 64)                49408

 dense_6 (Dense)             (None, 256)               16640

 dense_7 (Dense)             (None, 55)                14135

=================================================================
Total params: 409911 (1.56 MB)
Trainable params: 409911 (1.56 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 4. Layers of the Model

4.4. Video – based Prediction

Figure 5 and 6, displays how the system records and examines footage from a camera continuously, concentrating on the motions of human bodies. It uses the model to recognize gestures and records key point data from these movements. A recognized gesture is immediately shown in the live video feed if it satisfies a predetermined confidence level. It allows for real-time gesture-based engagement and continues until the user decides to stop it.

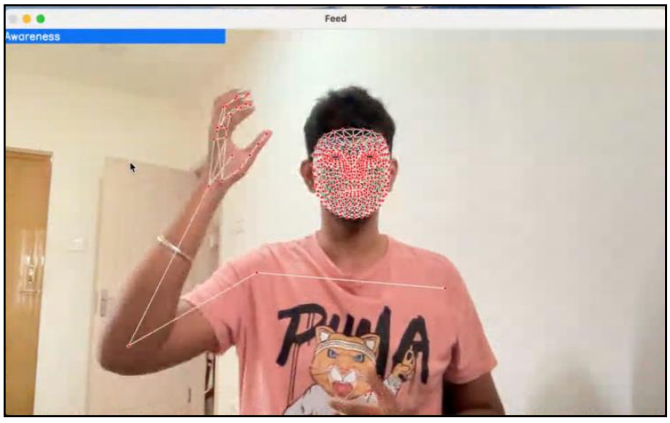

Figure 5. Word Prediction for Awareness

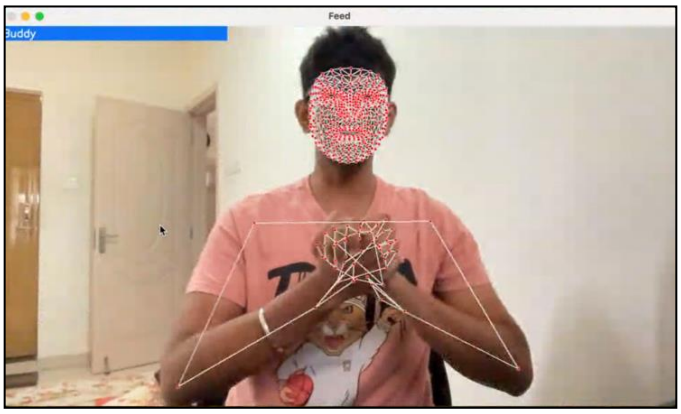

Figure 6. Word Prediction for Buddy

4.5. Sentence Generation

In Figure 7, the system uses keypoint detection for analyzing human body movements in live or recorded videos. It maintains a sequence of key information, continuously updates keypoint data, and makes predictions. Once a certain sequence length is reached. If the model's confidence in recognizing gestures exceeds a set threshold, it incorporates these gestures into an ongoing conversation. This system generates real-time dialogues with user prompts and displays responses on the video feed, offering an interactive dialogue platform that runs until the user decides to stop it.

Figure 7. Sentence Generation Output

5. Conclusion

This research helps in real-time sign language detection and also helps in converting them to text, by predicting the words first, and then generating phrases from them. The proposed system can recognize any type of Indian sign language using the LSTM model. This research is aimed at a societal contribution. We hope this ISL Dataset will create excitement in the sign language research community and have an impact on it. The dataset can also be used for further research making this system even more efficient. Words are predicted at an accuracy of 79%. This system helps in detecting and predicting sign language. To sum up, the creation of a sign language recognition system is a significant stride in addressing the communication challenges experienced by individuals who are deaf, have hearing impairments, or face speech-related limitations. This technology's ability to instantly identify hand gestures and convert them into text has the potential to enhance the ability of sign language users to convey their thoughts effectively, enhancing meaningful connections with the wider community. As the world continues to evolve, the adoption of such innovative solutions ensures that equal opportunities and inclusivity are provided to everyone, irrespective of language disparities or disabilities

References

[1] Goel, Pragati, Ashutosh Sharma, Vikas Goel, and Vikas Jain. "Real-time sign language to text and speech translation and hand gesture recognition using the LSTM model." In 2022 3rd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT),Ghaziabad, India. IEEE, 2022. 1-6.

[2] Mallika, Garg, Debashis Ghosh, and Pyari Mohan Pradhan. "A two-stage convolutional neural network for hand gesture recognition." In Proceedings of the 6th International Conference on Advance Computing and Intelligent Engineering: ICACIE 2021, Singapore: Springer Nature Singapore, 2022. 383-392.

[3] Tiku, Kohsheen, Jayshree Maloo, Aishwarya Ramesh, and R. Indra. "Real-time conversion of sign language to text and speech." In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India. IEEE, 2020. 346-351.

[4] Srininvas, Lokavarapu V., Chitri Raminaidu, Devareddi Ravibabu, and Shiva Shankar Reddy. "A framework to recognize the sign language system for deaf and dumb using mining techniques." Indonesian Journal of Electrical Engineering and Computer Science 29, no. 2 (2023): 1006-1016.

[5] Ojha, Ankit, Ayush Pandey, Shubham Maurya, Abhishek Thakur, and P. Dayananda. "Sign language to text and speech translation in real time using convolutional neural network." International Journal of Engineering Research & Tecnology.(IJERT) 8, no. 15 (2020): 191-196.

[6] Joshi, Abhinav, Susmit Agrawal, and Ashutosh Modi. "ISLTranslate: Dataset for Translating Indian Sign Language." arXiv preprint arXiv:2307.05440 (2023).

[7] Patil, Rachana, Vivek Patil, Abhishek Bahuguna, and Gaurav Datkhile. "Indian sign language recognition using convolutional neural network." In ITM web of conferences, vol. 40, p. 03004. EDP Sciences, 2021.

[8] Dafnis, Konstantinos M., Evgenia Chroni, Carol Neidle, and Dimitris N. Metaxas. "Isolated sign recognition using ASL datasets with consistent text-based gloss labeling and curriculum learning." In Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology (STLAT 7), LREC 2022. European Language Resources Association (ELRA), 2022.

[9] Sridevi, Parama, Tahmida Islam, Urmi Debnath, Noor A. Nazia, Rajat Chakraborty, and Celia Shahnaz. "Sign Language recognition for Speech and Hearing Impaired by Image processing in matlab." In 2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC),Malambe, Sri Lanka. IEEE, 2018. 1-4.

[10] Akshatha Rani K and Dr. N Manjanaik, "Sign Language to Text-Speech Translator Using Machine Learning", International Journal of Emerging Trends in Engineering Research, Vol. 9, no. 7, 2021.

[11] Nandhini, A. Sunitha, D. Shiva Roopan, S. Shiyaam, and S. Yogesh. "Sign language recognition using convolutional neural network." In Journal of Physics: Conference Series, vol. 1916, no. 1, p. 012091. IOP Publishing, 2021.

[12] Limaye, Heramba, Shraddha Shinde, Anurag Bapat, and Nimish Samant. "Sign Language Recognition using Convolutional Neural Network with Customization." Available at SSRN 4169172 (2022).

[13] Wibawa, Adhi Dharma, and Surya Sumpeno. "Gesture recognition for Indonesian Sign Language Systems (ISLS) using multimodal sensor leap motion and myo armband controllers based-on naïve bayes classifier." In 2017 International Conference on Soft Computing, Intelligent System and Information Technology (ICSIIT), Changa, India pp. 1-6. IEEE, 2017.

[14] Rajalakshmi, E., R. Elakkiya, V. Subramaniyaswamy, L. Prikhodko Alexey, Grif Mikhail, Maxim Bakaev, Ketan Kotecha, Lubna Abdelkareim Gabralla, and Ajith Abraham. "Multi-semantic discriminative feature learning for sign gesture recognition using hybrid deep neural architecture." *IEEE Access* 11 (2023): 2226-2238.

[15] Mrs. S Chandra Gandhi, Aakash raj R, Muhammed Shamil ML, and Akhil S, "Real Time Translation Of Sign Language To Speech And Text", International Advance Research Journal in Science, Engineering, and Technology, Vol. 8, no 4, 2021. 56-58

[16] Victoria Adebimpe Akano and Adejoke O Olamiti, Conversion of Sign Language To Text And Speech Using Machine Learning Techniques", Journal Of Research And Review In Science, Vol. 5, 58-65, 2018.