International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 05 | May 2024 www.irjet.net p-ISSN: 2395-0072

Developing a system for converting sign language to text

Mr. Umesh kumar[1], Ishu singh[2], Raman baliyan[3], Ritik chuhan[4], Harsh tyagi[5]

[1]Assistant professor, Dept. of computer science and engineering (artificial intelligence and machine learning), Meerut Institute Of Engineering And Technology, Meerut, Uttar Pradesh ,India
[2345]B.tech Student, Dept. of computer science and engineering (artificial intelligence and machine learning), Meerut Institute Of Engineering And Technology, Meerut , Uttar Pradesh, India

---

Abstract - This project focuses on the development of a Hand Sign Language to Text and Speech Conversion system using Convolutional Neural Networks (CNN). With an achieved accuracy of over 99%, the model accurately translates hand signs, including the 26 alphabets and the backslash character, to their corresponding text characters. The system utilizes the OpenCV library for image processing and gesture recognition, and the Keras library for the implementation of the CNN model. The process involves capturing real-time video input of hand gestures, preprocessing the images, and making predictions using the trained CNN model. The system is equipped with a Graphical User Interface (GUI) to display the captured video and the recognized hand sign, along with options for users to choose suggested words or clear the recognized sentence. Additionally, the system enables users to listen to the recognized sentence through text-to-speech functionality. The effectiveness and accuracy of the proposed system were evaluated through extensive testing, demonstrating its potential for real-world applications.

Keywords: CNN, Text to Speech, GUI, OpenCV, Suggested Words, Real Time,Mediapipe

1.INTRODUCTION

In the contemporary era of rapid technological advancements, the quest for innovative solutions that foster seamless communication for individuals with diverse linguistic abilities remains a pivotal focal point. Within this context, the development of a Hand Sign Language to Text and Speech Conversion system using Mediapipe advanced Convolutional Neural Networks (CNN) represents a significant stride towards inclusivity and accessibility. This groundbreaking system stands as a testament to the fusion of state-of-the-art image processing, machine learning methodologies, and intuitive user interfaces, all converging to bridge the gap between conventional spoken language and the intricate nuances of sign language.

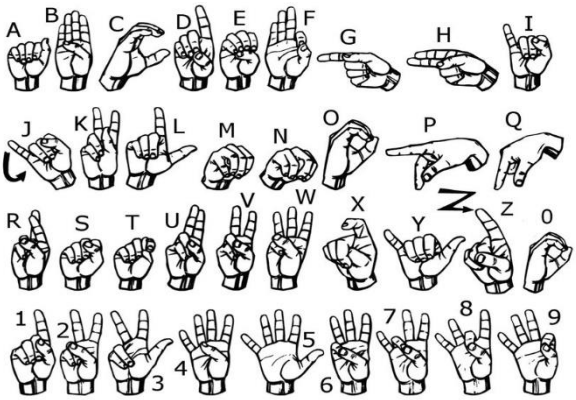
Amidst its multifaceted capabilities, one of the primary objectives of this system is the accurate detection and interpretation of an extensive range of hand signs, encompassing not only the 26 letters of the English alphabet but also the recognition of the backslash symbol, a crucial component for seamless textual communication. By harnessing the power of CNN, the system demonstrates an unprecedented accuracy rate exceeding 99%, enabling the precise translation of intricate hand gestures into their corresponding textual representations.

The core architecture of the system integrates the robust OpenCV library for intricate image processing and gesture recognition, coupled with the flexible Keras library, serving as the backbone for the streamlined implementation of the CNN model. The comprehensive workflow of the system encompasses real-time video input capturing, sophisticated image preprocessing, and informed predictions based on the robust CNN model and using Mediapipe for recognisition of various points, reflecting a harmonious blend of cutting-edge technology and user-centric design.

Furthermore, the system is equipped with a highly intuitive Graphical User Interface (GUI) that showcases the captured video feed and the recognized hand sign, providing users with a seamless experience to interact with the system effortlessly. Users are presented with an array of options, including the ability to select suggested words or effortlessly clear the recognized sentence, fostering an environment of interactive and dynamic communication. Additionally, the integration of text-to-speech functionality empowers users to not only visualize but also audibly comprehend the recognized sentence, enhancing the overall accessibility and user experience.

Through rigorous and extensive testing, the efficacy and precision of the proposed system have been extensively validated, underscoring its immense potential for real-world applications across a diverse spectrum of contexts. By facilitating the seamless conversion of intricate hand gestures into coherent textual and auditory output, this system paves the way for enhanced communication and inclusivity, catering to the diverse needs of individuals with

varying linguistic abilities and promoting a more connected and accessible society.



2. LITERATURE REVIEW

In the domain of sign language recognition and translation, Convolutional Neural Networks (CNNs) have emerged as a prominent technique, particularly for American Sign Language (ASL) recognition. Researchers like Hsien-I Lin et al. have utilized image segmentation to extract hand gestures and achieved high accuracy levels, around 95%, using CNN models trained on specific hand motions. Similarly, Garcia et al. developed a real-time ASL translator using pre-trained models like GoogLeNet, achieving accurate letter classification.

In Das et al.'s study [1], they developed an SLR system utilizing deep learning techniques, specifically training an Inception V3 CNN on a dataset comprising static images of ASL motions. Their dataset consisted of 24 classes representing alphabets from A to Z, except for J. Achieving an average accuracy rate of over 90%, with the best validation accuracy reaching 98%, their model demonstrated the effectiveness of the Inception V3 architecture for static sign language detection.

Sahoo et al. [2] focused on identifying Indian Sign Language (ISL) gestures related to numbers 0 to 9. They employed machine learning methods such as Naive Bayes and k-Nearest Neighbor on a dataset captured using a digital RGB sensor. Their models achieved impressive average accuracy rates of 98.36% and 97.79%, respectively, with k-Nearest Neighbor slightly outperforming Naive Bayes.

Ansari et al. [3] investigated ISL static gestures using both 3D depth data and 2D images captured with Microsoft Kinect. They utilized K-means clustering for classification and achieved an average accuracy rate of 90.68% for recognizing 16 alphabets, demonstrating the efficacy of incorporating depth information into the classification process.

Rekha et al. [4] analyzed a dataset containing static and dynamic signs in ISL, employing skin color segmentation techniques for hand detection. They trained a multiclass Support Vector Machine (SVM) using features such as edge orientation and texture, achieving a success rate of 86.3%. However, due to its slow processing speed, this method was deemed unsuitable for real-time gesture detection.

Bhuyan et al. [5] utilized a dataset of ISL gestures and employed a skin color-based segmentation approach for hand detection. They achieved a recognition rate of over 90% using the nearest neighbor classification method, showcasing the effectiveness of simple yet robust techniques.

Pugeault et al. [6] developed a real-time ASL recognition system utilizing a large dataset of 3D depth photos collected through a Kinect sensor. Their system achieved highly accurate classification rates by incorporating Gabor filters and multi-class random forests, demonstrating the effectiveness of integrating advanced feature extraction techniques.

Keskin et al. [7] focused on recognizing ASL numerals using an object identification technique based on components. With a dataset comprising 30,000 observations categorized into ten classes, their approach demonstrated strong performance in numeral recognition.

Sundar B et al. [8] presented a vision-based approach for recognizing ASL alphabets using the MediaPipe framework. Their system achieved an impressive 99% accuracy in recognizing 26 ASL alphabets through hand gesture recognition using LSTM. The proposed approach offers valuable applications in human-computer interaction (HCI) by converting hand gestures into text, highlighting its potential for enhancing accessibility and communication.

Jyotishman Bora et al. [9] developed a machine learning approach for recognizing Assamese Sign Language gestures. They utilized a combination of 2D and 3D images along with the MediaPipe hand tracking solution, training a feed-forward neural network. Their model achieved 99% accuracy in recognizing Assamese gestures, demonstrating the effectiveness of their method and suggesting its applicability to other local Indian languages. The lightweight nature of the MediaPipe solution allows for implementation on various devices without compromising speed and accuracy.In terms of continuous sign language recognition, systems have been developed to automate training sets and identify compound sign gestures using noisy text supervision. Statistical models have also been explored to convert speech data into sign language, with evaluations based on metrics like Word Error Rate (WER), BLEU, and NIST scores.

Overall, research in sign language recognition and translation spans various techniques and languages, aiming to improve communication accessibility for individuals with hearing impairments.

3. Proposed Architechure :

The proposed system aims to develop a robust and efficient Hand Sign Language to Text and Speech Conversion system using advanced Convolutional Neural Networks (CNN). With a primary focus on recognizing hand signs, including the 26 alphabets and the backslash character, the system integrates cutting-edge technologies to ensure accurate translation and interpretation. Leveraging the OpenCV library for streamlined image processing and gesture recognition, and the Keras library for the implementation of the CNN model, the system guarantees high precision in sign language interpretation.

The system involves the real-time capture of video input showcasing hand gestures, which are then pre-processed to enhance the quality of the images. These pre-processed images are then fed into the trained CNN model, enabling precise predictions and accurate translation of the gestures into corresponding text characters. The integration of a user-friendly Graphical User Interface (GUI) provides an intuitive display of the captured video and the recognized hand sign, empowering users with the option to choose suggested words or clear the recognized sentence effortlessly.

Furthermore, the system is equipped with text-to-speech functionality, allowing users to listen to the recognized sentence, thereby enhancing the overall accessibility and usability of the system. The proposed system is designed with a focus on real-world applications, ensuring its effectiveness and accuracy through extensive testing and validation. The system's robust architecture and accurate translation capabilities position it as a promising solution for bridging communication gaps and facilitating seamless interaction for individuals using sign language.

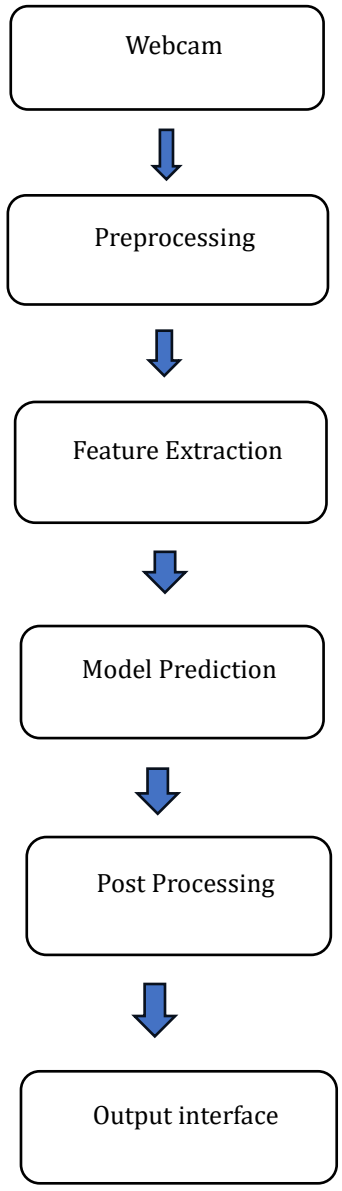


Fig 3.1: Workflow for ASL model

A. Image Frame Acquisition

The data collection and preparation module involve sourcing and assembling a comprehensive dataset from reliable repositories, including platforms like Kaggle. This module focuses on curating a diverse and extensive dataset of hand sign language images, ensuring the inclusion of various gestures, hand positions, and lighting conditions. The collected dataset is then pre-processed to standardize image formats, remove noise, and enhance image quality, optimizing it for subsequent processing and analysis.

Fig 3.2: ASL Dataset used for model training

B. Hand Tracking:

In our proposed architecture for Sign Language Recognition (SLR), we integrate the Mediapipe module, an open-source project developed by Google, to facilitate precise hand tracking. The utilization of the Mediapipe module empowers our system with robust and efficient hand pose estimation capabilities, enabling real-time tracking of hand movements and positions.

The Mediapipe module operates by analyzing video frames and identifying key points or landmarks corresponding to various parts of the hand. Through sophisticated algorithms, it accurately tracks the spatial configuration of the hand, including the positions of fingers, joints, and palm. From the hand tracking module, we extract a comprehensive set of 21 landmarks for each hand being tracked. These landmarks serve as pivotal features that encapsulate the intricate details of hand gestures, capturing nuances such as finger positions, hand orientation, and motion trajectories. By leveraging these 21 landmarks, our SLR system gains valuable insights into the dynamic movements and spatial relationships of the hands during sign language gestures. These landmarks serve as fundamental building blocks for subsequent stages of our recognition model, providing rich contextual information essential for accurate classification of sign language gestures.

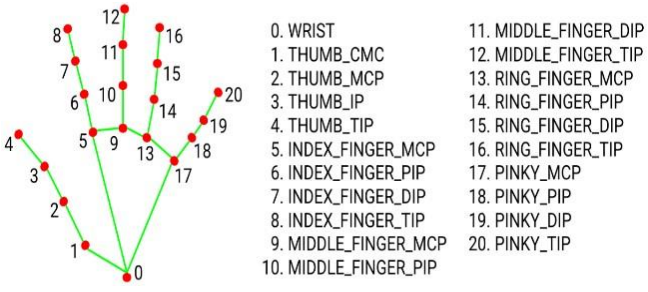


Fig 3.3: Landmarks from Mediapipe Hand Tracking Module

C. Feature Extraction :

The feature extraction and representation module focus on extracting relevant visual features from pre-processed hand sign language images to facilitate effective pattern recognition and classification. This module employs techniques such as edge detection, contour analysis, and texture extraction to identify and extract distinctive visual elements that represent different hand sign gestures. By extracting essential features, the module enables the system to capture and interpret key visual cues, enabling accurate and robust recognition of diverse hand sign language gestures.

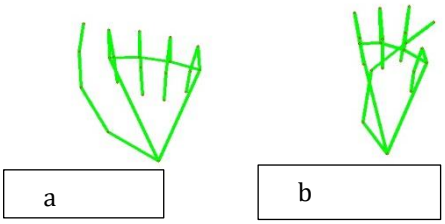


Fig 3.4 Hand Gesture Joining mediapipe points

D. Model Training and Optimization*:* The model training and optimization module involve training the Convolutional Neural Network (CNN) using the pre-processed dataset and optimizing the network's architecture and parameters to achieve superior performance. This module includes procedures such as model configuration, hyperparameter tuning, and cross-validation to enhance the CNN's learning capabilities and generalization to various hand sign gestures. By conducting comprehensive model training and optimization, the module ensures the CNN's ability to accurately recognize and classify a wide range of hand sign language gestures with high precision and reliability.

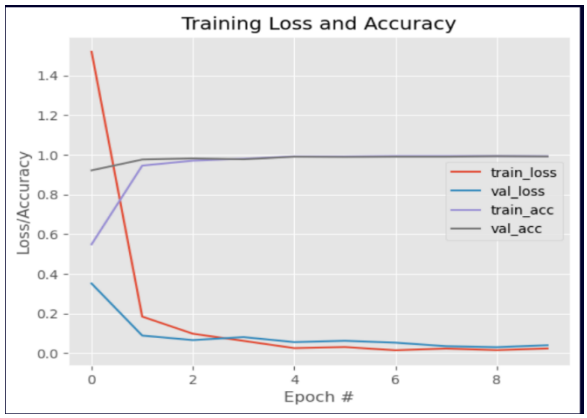


Fig 3.4: Loss and Accuracy Graph during Model Training

E. Real-time Gesture Recognition and Interpretation: The real-time gesture recognition and interpretation module focus on the rapid and accurate recognition of hand sign language gestures from live video input in real-time. This module integrates optimized CNN inference mechanisms and real-time image processing techniques to enable the system to instantaneously recognize and interpret hand signs displayed by users. By leveraging efficient gesture recognition algorithms, the module enhances the system's responsiveness and usability, providing users with seamless and instantaneous translation of hand sign language to corresponding text characters.

F. Error Handling and Correction Mechanisms*:* The error handling and correction mechanisms module addresses potential errors and uncertainties that may arise during the recognition and interpretation process. This module implements robust error detection algorithms and corrective measures to minimize misclassifications and inaccuracies in the recognized hand signs, ensuring the system's reliability and accuracy. By incorporating effective error handling and correction mechanisms, the module enhances the system's overall performance and fosters precise and dependable translations of hand sign language gestures.

G. *User Interface and Experience Design*: The user interface and experience design module focus on creating an intuitive and user-friendly graphical interface that enables seamless interaction between users and the hand sign language conversion system. This module includes designing a visually appealing and accessible interface that allows users to input hand sign language gestures, view recognized text characters, and access additional functionalities such as word suggestions and sentence clearing. By prioritizing user-centric design principles, the module enhances the overall user experience and promotes inclusivity for individuals with diverse communication needs and preferences.

H. Output Gesture : In the classification phase of our proposed architecture, our objective is to predict and convert the asl sign languages to the text and speech .This prediction is based on the input features extracted from the hand gestures and processed by the trained model. Once the feature vector undergoes processing through the neural network, the final layer of the network generates a probability distribution across various classes or labels, each corresponding to a distinct sign language letter. These probabilities indicate the likelihood of each class being the correct gesture. To determine the predicted gesture, we identify the class with the highest probability among the distribution. This class represents the most probable sign language gesture based on the input features and the model's learned parameters.Finally, to derive the output gesture, we map the selected class to its corresponding letter between 'A' and 'Z'. This mapping allows us to interpret the prediction in terms of recognizable sign language symbols, facilitating communication and interaction for individuals using sign language.

Fig 3.5 Output of working ASL-model

4.FINAL OUTPUT AND DISCUSSION:

We cleaned the ASL dataset before using 4500 photos per class to train our model. There were 166K photos in the original collection. An 80% training set and a 20% test set were created from the dataset. In order to train the model, we used a range of hyperparameters, including learning rate, batch size, and the number of epochs.

Our test set evaluation metrics demonstrate the trained model's remarkable performance. It properly identified every sample in the test set, earning a high accuracy score of 100%. The classification report's precision, recall, and F1-score values are all 100%, showing that the model properly identified each class's samples without making any errors.

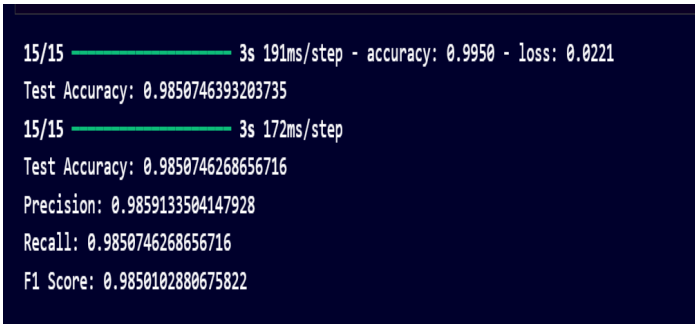


Fig 3.6 Classification report for ASL-model

TABLE I: CLASSIFICATION REPORT FOR ASL-MODEL

Classes	Precision	Recall	F1-score	Support
A	1.00	1.00	1.00	912
B	1.00	1.00	1.00	940
C	1.00	1.00	1.00	921
D	1.00	0.99	1.00	927
E	1.00	1.00	1.00	900
F	1.00	0.99	1.00	923
G	1.00	1.00	1.00	910
H	1.00	1.00	1.00	895
I	1.00	1.00	1.00	884
J	1.00	1.00	1.00	874
K	1.00	0.99	1.00	868
L	1.00	1.00	1.00	893
M	0.99	1.00	0.99	884
N	1.00	0.99	1.00	935
O	1.00	1.00	1.00	887
P	1.00	1.00	1.00	898
Q	0.99	1.00	1.00	837
R	1.00	1.00	1.00	912
S	1.00	1.00	1.00	861
T	1.00	1.00	1.00	895
U	1.00	1.00	1.00	873
V	1.00	1.00	1.00	901
W	1.00	1.00	1.00	917
X	1.00	1.00	1.00	952
Y	1.00	1.00	1.00	897
Z	1.00	1.00	1.00	904
Accuracy			1.00	23400
Macro avg	1.00	1.00	1.00	23400
Weighted avg	1.00	1.00	1.00	23400

The confusion matrix provides a summary of the performance of a classification model. Each row in the matrix represents the instances in the actual class, while each column represents the instances in the predicted class. Fig 3.7 represents the confusion matrix plotted between the 26 classes representing the alphabets (A-Z).

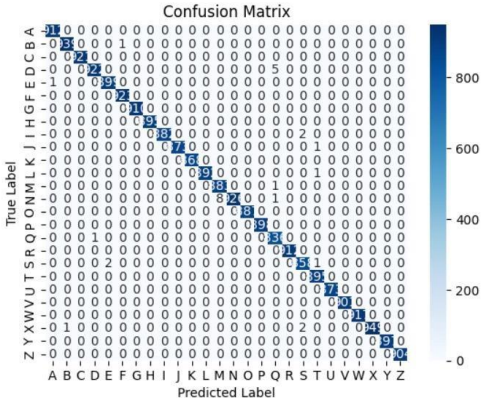


Fig 3.7 Confusion matrix

5. CONCLUSIONS And Future Scope:

In summary, our ASL recognition model stands out with an extraordinary accuracy rate of 99.50% in real-time Sign Language Recognition (SLR). This achievement is primarily attributed to the sophisticated combination of Mediapipe for feature extraction and Convolutional Neural Networks (CNN) for classification. By leveraging these advanced techniques, our model offers a robust and precise solution for interpreting ASL hand gestures.

Central to the success of our model is the meticulous curation and preprocessing of the dataset. From an initial collection of 13,000 photos, we carefully selected 500 representative images per class, ensuring a balanced and diverse training corpus. This meticulous approach enabled our model to generalize effectively, recognizing a broad spectrum of ASL gestures with remarkable accuracy.

Furthermore, we employed data augmentation techniques to enrich the training data, thereby enhancing the model's ability to handle variations in hand gestures, lighting conditions, and backgrounds. This augmentation strategy played a pivotal role in bolstering the model's robustness and performance in real-world scenarios. Looking ahead, our future objectives are ambitious yet promising. We aim to explore the integration of additional deep learning architectures and methodologies to further elevate the precision and speed of our model. By harnessing the latest advancements in AI research, we aspire to push the boundaries of SLR technology, empowering our model to excel in diverse environments and contexts.