IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Deep Gesture Interpretation For American Sign Language

[1]Shruti Sharma, [2] Sandeep Das, [3]Diya Tembhurne, [4]Krupali Dhawale

[1]Student, [2]Student, [3]Student, [4]Assistant Professor

[1]Department of Artificial Intelligence,

[1]G.H. Raisoni College of Engineering, Nagpur, India

***Abstract:*** Human interaction necessitates effective communication, and for those with speech impediment or Hard of hearing, sign language is a crucial communication tool. However, barriers between these individuals and the rest of the population are often caused by the absence of widespread awareness of sign language. To tackle this, a real-time system which utilizes deep convolutional neural networks for converting hand gestures in ASL into text and speech has been developed. Monitor, segment, gesture capture, extract feature, recognize gesture, and convert text-to-speech are some of the crucial steps in our method. The system can reliably identify and convert hand gestures into legible text and audible speech since it has been trained on a dynamic collection of hand gestures. The goal of this technology is to enable more seamless communication between people who are deaf and those who do not know sign language. This is a major step towards ensuring accessibility and inclusion because deep learning and CV are used to enhance SLR accuracy in addition to bridging communication gaps.

***Index Terms*** **-** speech or hearing impairments, deep CNNs, American sign language (ASL), real-time gesture recognition, hand gesture tracking, feature extraction, text-to-speech conversion, hand gesture dataset, Computer vision, Sign language accessibility, Inclusion through technology, Gesture segmentation, Communication gap bridging

## I. INTRODUCTION

Sign language is a special kind of communication that has been used by people with disabilities for centuries. Hand gestures have been used earlier than human civilization. A unique method of communication that often gets overlooked is sign language. Sign language is a way to convey letters, words, or sentences using various kinds of hand gestures. Its grammar and vocabulary are completely different from those of written and verbal languages. The ability to produce sounds that correspond to specific words and grammatical combinations is necessary for spoken languages in order to transmit meaningful information that is subsequently processed by the auditory faculties.

. The auditory faculties then receive the oratory components and process them appropriately. Spoken language does not use the visual capabilities like sign language does. Sign language involves various hand gestures, such as shape of hand, location of hand, and palm or fingers movement, as well as non-manual features like eye gazing, head nods, and facial expressions. Sign language allows those with poor hearing or no hearing and speech impairment to communicate, allowing them to express opinions more easily and close the communication gap.

Knowledge of sign language is required to effectively and accurately translate sign language to text or spoken language. The system uses neural networks to interpret American Sign Language (ASL), which consists of 22 signs corresponding to the 26 letters of the alphabet. Deep learning can help reduce communication, and the design includes tracking, classification, behavior detection, feature extraction,

gesture recognition, and text-to-speech conversion. The goal is to bridge the communication gap between the disabled and the public. Knowledge of sign language is required to effectively and accurately translate sign language to text or spoken language. The system uses neural networks to interpret American Sign Language (ASL), which consists of 22 signs corresponding to the 26 letters of the alphabet. Deep learning can help reduce communication, and the design includes tracking, classification, behavior detection, feature extraction, gesture recognition, and text-to-speech conversion. The goal is to bridge the communication gap between the disabled and the public Literature survey

Upon doing a comprehensive literature study, it is evident that considerable research has been done to address the problem of identifying signs in videos and photographs using a variety of methodologies. Siming He [1] created a system with 10,000 images in sign language and a dataset of 40 commonly utilised elements. Rapid classification has been rendered achievable through the use of Faster R-CNN with an embedded Region Proposal Network (RPN), which improved detection accuracy from 89.0% to 91.7% when when compared to Fast-RCNN. Using a common vocabulary dataset, the framework yields a 99% recognition rate by integrating Long Short-Term Memory (LSTM) networks for language sequence processing with 3D CNNs for feature extraction.

Centroid mapping was used in [11] to effectively recognise Sinhala sign gestures in photos with a green background, resulting in a 92% success rate for identification. The method turned out to be affordable.Using Support Vector Machines, B-Spline approximations were examined by M. Geetha and U. C. Manjusha [7] for Indian Sign Language identification, with 90% accuracy.
Pigou [8] developed a 2D CNN with the CLAP14 dataset, and the result was an accuracy of 91.7%. A 3D CNN was used by J. Huang [10] to develop a dataset based on Kinect, with an accuracy of 94.2%. Using transfer learning with pre-trained models, research by J. Carriera [10] achieved 98.0% accuracy on the UCF-101 dataset.

According to Wadhawan [13], the majority of studies on camera systems for static, isolated, and single-handed signs have been carried out in the area of slr. Their study aims to provide a road map for next research and improvements in the field's expertise. In-depth research on popular deep neural network patterns for SLR has been performed by Adaloglou [14], who offered a comparative evaluation based on large-scale testing with three publicly available datasets.A low-complexity model that combines deep learning methods with singular value decomposition (SVD) for RGB video processing is presented by Rastgoo [15] in their real-time system for isolated hand sign language recognition.

Technologies for people with hearing impairments continue to face hurdles despite developments in SLR. The current evaluation emphasises the continuous advancements and the demand for further developments in deep learning methodologies for improving the accuracy and efficiency of recognition.

.

## II. METHODOLOGY

Hand identification, picture pre-processing, and model training and prediction are the three main parts of the sign language recognition system's technique. The cvzone.HandTrackingModule is used by the system for recognizing and real-time tracking of hand movement. OpenCV is used to record a video feed, and the hand detector analyses each frame to find hands in the scene. The hand images are cropped for further processing based on the bounding box coordinates of the detected hands. The collected region is scaled to a standard size of 300x300 pixels while keeping the aspect ratio whenever a hand is detected. To make sure the resized hand image fits inside it properly, a white canvas is prepared. When the user hits a pre-set key, the processed photos are saved, making it easier to gather training data.

Built around the EfficientNet80 architecture, the model is enhanced with pictures from the ASL dataset. In order to improve the training dataset, data augmentation methods including rotation, shifting, and zooming are used. Appropriate callbacks are used during model training to avoid overfitting. Using a history of previous predictions to stabilise output labels, a smoothed prediction is calculated during inference to increase accuracy.

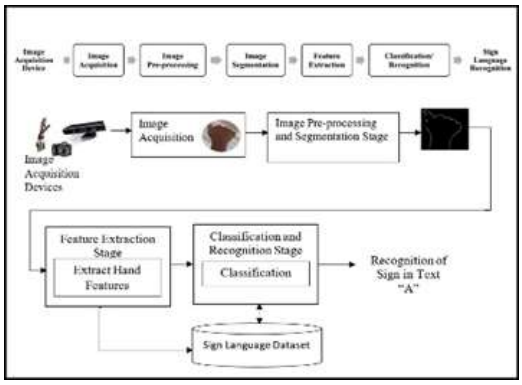


Fig. 1: Model Architecture

## III. DATASET

Fig. 2: ASL Alphabet Recognition

Images that accurately depict the American Sign Language alphabet can be found in the ASL Alphabet dataset on Kaggle. This dataset includes thousands of images, each labeled with the corresponding ASL letter, making it an ideal resource for training machine learning models in sign language recognition. The dataset contains 29 classes - 26 for the alphabets and the remaining three for "nothing," "space," and "delete." The images are captured in various lighting conditions and backgrounds, providing a diverse set of examples that can help improve the robustness of the models. The deaf and hard-of-hearing community will benefit from improved communication accessibility as a result of researchers and developers using this dataset to build applications that correctly interpret ASL gestures.

## IV. DATA PROCESSING

### 4.1Population and Sample

Noise Removal: One of the most important preprocessing steps is to eliminate extraneous data points, such as background noise, erroneous labels, or improperly identified gestures.

Outlier Detection: Finding and eliminating outliers, such as hand motions that deviate from the typical hand position or movement, is known as outlier detection.

### 4.2Data Augmentation

Rotation and Scaling: To make sure that models adapt well to various hand orientations and sizes, hand photos or gesture data are frequently rotated, flipped, or scaled.

Cropping and Resizing: Pictures are scaled or cropped to just highlight the hand area, eliminating extraneous background details that can lead the model astray.
Flipping: To replicate various gesture orientations, flip in a horizontal or vertical direction.

### 4.3Normalisation

Pixel Value Scaling: To enable quicker and more stable convergence during training, pixel values are normalised (e.g., scaled between 0 and 1 or -1 and 1) for image-based gesture recognition.

Pose Normalisation: To guarantee that movements are constant across different hand sizes or camera views, keypoints such as finger locations or hand orientation are normalised.

### 4.4Extraction of Keypoints

MediaPipe or OpenPose Keypoints from hands (finger joints, wrist positions) are extracted using trained models such as MediaPipe or OpenPose, which helps condense the input space and narrows the model's emphasis to important aspects for recognition.

### 4.5Reduction of Dimension

Principal Component Analysis, or PCA: In order to eliminate superfluous or unnecessary data, several projects employ PCA to minimise the amount of features or keypoints.

Feature Selection: Selecting significant features can help to improve accuracy and reduce computational complexity. Examples of such features are particular keypoints associated with ASL hand gestures.

### 4.6Segmentation

Hand Segmentation: To separate the hand region from complicated backgrounds, use techniques like OpenCV for background subtraction and skin colour recognition. Better training data and crisper gesture detection are aided by this.

Bounding Box Detection: Improves gesture localisation and cropping by automatically generating bounding boxes around hands.

### 4.7Temporal Alignment (for Continuous Gestures)

Frame Padding or Resampling: To improve temporal alignment in the model, gesture sequences may need to have their frames padded or resampled to ensure that every gesture in the dataset has a consistent length.

Time Normalisation: Time normalisation is the process of lining up several gesture sequences chronologically so that classification is unaffected by gesture speed or sequence length.

### 4.8Label Encoding

One-Hot Encoding: Labels (such as A, B, C, or word gestures) are encoded into one-hot vectors that are then supplied into the model to be trained in the gesture classification process.

Class Balancing: Using methods such as SMOTE (Synthetic Minority Over-sampling Technique) or oversampling under-represented gestures, one can address any imbalance in gesture classes.

By ensuring cleaner and more relevant input data for deep learning model training, these preprocessing techniques help to enhance the performance of ASL recognition.

## V. SYSTEM ARCHITECTURE

The three major layers of the SLRsystem's architecture are the input layer, processing layer, and output la yer.Each layer improves the system's overall function.

### 5.1Input Layer

The input layer's function is to acquire data from the user in real-time:

Video Capture: The system continuously captures video frames from a webcam by using OpenCV. This is the primary source for hand gesture detection.

Hand Detection Module: Utilising the cvzone.HandTrackingModule for the hand detection, it identifies hands by analysing every video frame. Bounding box coordinates are results, which describe the size and location of any hands that are detected. The current solution is mostly focused on one hand at a time, although it is capable of detecting many hands.

**5.2Processing Layer**

The two essential sub-components of this layer are the deep learning model and image processing.

5.2.1 Processing of Images:

1. Hand Image Cropping: Using the bounding box coordinates supplied by the hand detection module, the system crops the relevant portion of the video frame whenever it detects a hand.

2. Normalisation and Resizing: The aspect ratio is maintained when the hands' cropped photos are resized to a standard dimension of 300 by 300 pixels. During training, the resized photographs are put onto a white canvas to guarantee consistency.

3. Data Storage: The processed photos are saved to a specified directory for future training when the user presses a specific key. With the use of this functionality, users can compile a wide range of sign gesture examples.

5.2.2 Deep Learning Model

1. Model Architecture: EfficientNetB0, a sophisticated convolutional neural network (CNN) renowned for its effectiveness and performance, serves as the foundation for the model. It is made up of multiple layers:

i) Base Model: To leverage transfer learning, EfficientNetB0, the central component for feature extraction, has been pre-trained on the ImageNet dataset.
ii) Layer of Global Average Pooling: This layer makes feature maps smaller in space, allowing for a more condensed depiction.
iii) Dense Layers: To reduce overfitting and improve generalisation, a fully linked layer, Batch Normalisation, and Dropout are used.
iv) O/P Layer: The softmax activation function is used in final layer to provide class probabilities for every sign gesture.

2. Data Augmentation: To make the model more resilient during training, techniques like rotation, shifting, shear transformations, and zooming are applied. This enhances the model's capacity to generalise across many variations of hand gestures.

**5.3Output Layer**

The system provides the user with feedback at the output layer.

Real-time Predictions: The model makes predictions about the related sign language gesture in real-time by processing the preprocessed hand photos. Class probabilities are generated as predictions, and the gesture with the highest confidence score is chosen.

Smoothing Mechanism: A smoothing technique is used to lower variability and increase accuracy. The algorithm filters out erratic predictions caused by variations in hand location or movement speed by keeping track of the last few guesses and using a majority vote to determine the most likely current gesture.

Visual Feedback:

1. On-screen Display: The identified sign and the prediction's degree of confidence are shown on the video feed. User engagement is improved with a bounding box that surrounds the identified hand and offers visual confirmation.

2. User Interaction: The system offers an easy-to-use interface for gathering data by enabling users to take and store pictures of hand motions. With a few keystrokes, users may quickly save photographs or close the application.

Real-time gesture identification is made possible by the architecture of the sign language recognition system, which skilfully blends computer vision and deep learning. The processing layer incorporates sophisticated picture handling and a strong deep learning model, the output layer provides users with instantaneous visual feedback, and the input layer records and processes video frames. This all-encompassing architecture facilitates a productive and enjoyable sign language recognition experience.

## VI. EVALUATION


Fig.2: ASL Alphabet Recognition

Table 6.1: Comparision From Other Research Papers

| Sr. no | Sign Language | Classification method | Recognition Rate |
|---|---|---|---|
| 1 | ASL | Dynamic Time Wrapping Static Gestures Dynamic Gestures | 92.82 |
| 2 | ASL | Boostmap Embedding | brute-force approach but 800 times faster. |
| 3 | PSL | ETPL(k) Graph Parsing Model | 94.31 |
| 4 | ASL | ANN(feed-forward BPN) without canny threshold with canny threshold(0.15) with canny threshold(0.25) | 77.72<br>91.53<br>92.33 |
| 5 | Face data | Recognition Method: Pca, Mpca-Ml, Mpca-Lv, Mpca-Js, Mpca-Ps, Pca+Lda, Pca+Lpp | 74.79, 74.04(83.08) 75.10, 92.79, 87.79, 89.89, 88.94 |
| 6 | ASL | Pseudo two-dimensional hidden markov models (p2-dhmms) | 98 |

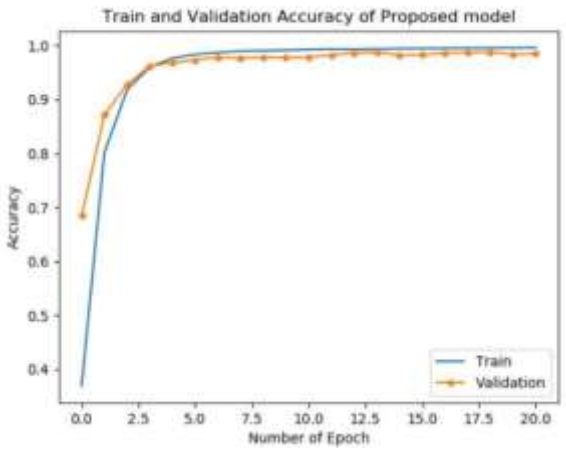


Fig.3: Trained and Validation Accuracy of Proposed Model

## VII. CONCLUSION

The majority of the signs used in current systems are static and consist of manual signs, alphabets, and digits. A thorough vocabulary database is desperately needed, as there aren't many standardised datasets accessible for use across different nations, continents, and languages. Nonverbal communication techniques and continuous or dynamic indicators should be given priority in future research. SLR systems need to be made to make it easier to collect data outside of the lab, in a variety of circumstances. These systems should also be able to concurrently recognise various body parts, hand movements (both left and right), and facial

expressions. Recognition procedures that are both practical and effective are crucial. The goal of the suggested SLR system is to improve communication with non-hearing people by translating hearing-impaired people's input expressions into text. Our program aims to serve as a real-time hand gesture detection system, helping non-governmental organisations and organisations that assist people with special needs by utilising deep learning and image processing techniques. System response times can be further decreased by enhancing camera and graphics capabilities.

## VIII. LIMITATIONS

Systems for recognising sign language face many important obstacles that have a large impact on their functionality and usefulness. The differences in skin tones and lighting might cause irregularities in gesture recognition, which is one of the main challenges. Since these systems usually use color-based segmentation algorithms, variations in illumination or the presence of different skin tones can cause disparities in the interpretation of motions, which will decrease accuracy. Furthermore, the work of real-time communication is made more difficult by the similarity of movements that indicate distinct alphabets, which increases the possibility of misclassification.

The unavailability of high-quality, diversified dataset for testing and training is another serious problem. The wide range of sign language used in various contexts and cultures is not well represented in many of the current datasets, which limits how well the model can generalise to new data. The sensitivity of these models to changes in skin tone and lighting exacerbates this problem and can negatively affect their overall accuracy. Furthermore, model overfitting—a condition in which a system performs remarkably well on training data but is unable to recognise indications in real-world scenarios—is frequently caused by small training datasets.

Moreover, continuous sign language gestures are often absent from databases, despite their critical role in accurately capturing the nuances and fluidity of real-world discussions. The majority of existing systems tend to emphasise static gestures, which limits their usefulness in real-world scenarios where interactive and dynamic communication is essential. This emphasises how urgently more extensive algorithms are needed, ones that can identify both static and dynamic movements and take contextual clues into account that improve comprehension. Addressing these issues is essential to developing more dependable and user-friendly sign language recognition systems that can facilitate successful communication for the hearing impaired in a range of contexts.

## IX. FUTURE SCOPE

Several tactics can be used to improve the American sign language recognition system's functionality and accuracy.

Real-time Predictions: The model makes predictions about the related sign language gesture in real-time by processing the preprocessed hand photos. Class probabilities are generated as predictions, and the gesture with the highest confidence score is chosen.

Boost Hardware to Get Better Real-Time Performance: Purchasing sophisticated hardware, including sharper cameras and stronger GPUs, can greatly improve the system's real-time video processing capabilities. This will facilitate more fluid gesture recognition by lowering latency and enhancing user experience generally.

Expand the Dataset to Promote More Diversity: Increasing the number of hand gestures in the dataset can aid in the model's learning process. By adding a variety of samples that reflect various user demographics, hand sizes, and environmental circumstances, the model will become more resilient and scenario-adaptable.

Create Dynamic Gesture Recognition Models: The system will be able to understand movement-based gestures, like waving or signing language, with accuracy if dynamic gesture recognition models are included. To extract temporal information from video sequences, this may include utilising LSTM networks or RNNs.

Make Use of More Typical Datasets: The accuracy of the model will increase with the use of datasets that more accurately reflect the target population and context. The system will perform better in real-world applications if it has access to datasets covering a range of demographics, lighting settings, and backgrounds. These datasets will also guarantee that the system generalises well across diverse users and scenarios.

## REFERENCES


**[1]**Siming, H. Research of a sign language translation system based on deep learning.

**[2]** Mehreen, H. and Mohammad, E. W. Sign language recognition system using convolutional neural network and computer vision.

**[3]** Ashok, K. S., Gouri, S. M. and Kiran, K. R. Sign language recognition: state of the art.

**[4]** Pandit, S. R., Pawar, J., Pawar, R., Pote, P., and Sangale, V. Sign language recognition using deep learning.

**[5]** Bheda, V. and Radpour, N. D. Using deep convolutional networks for gesture recognition in American sign language.

**[6]** Malhotra, M. American sign language recognition using deep learning.

**[7]** Geetha, M. and Manjusha, U. C. 2012. A vision based recognition of Indian sign language alphabets and numerals using B-Spline approximation. International Journal on Computer Science and Engineering (IJCSE), 4(3): 406-415.

**[8]** Pigou, L., Dieleman, S., Kindermans, P. J., and Schrauwen, B. 2015. Sign language recognition using convolutional neural networks. In: Agapito L., Bronstein M., Rother C. (eds) Computer Vision - ECCV 2014 Workshops. ECCV 2014. Lecture Notes in Computer Science, 8925.

**[9]** Huang, J., Zhou, W., and Li, H. 2015. Sign language recognition using 3D convolutional neural networks. IEEE International Conference on Multimedia and Expo (ICME): 1-6.

**[10]** Carriera, J. and Zisserman, A. 2018. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference: 4724-4733.

**[11]** Herath, H. C. M., Kumari, W. A. L. V., Senevirathne, W. A. P. B., and Dissanayake, M. 2013. Image based sign language recognition system for Sinhala sign language.

**[12]** Nagaraj, A. and Khan, A. American Sign Language to Speech Application**.**

**[13]** Wadhawan, A. and Kumar, P. 2020. Deep learning-based sign language recognition system for static signs. Neural Computing and Applications, 32(12): 7957-7968.

**[14]** Adaloglou, N., Chatzis, T., Papastratis, I., Stergioulas, A., Papadopoulos, G. T., Zacharopoulou, V., Xydopoulos, G. J., Atzakas, K., Papazachariou, D., and Daras, P. 2021. A comprehensive study on deep learning-based methods for sign language recognition. IEEE Transactions on Multimedia, 24: 1750-1762**.**

**[15]** Rastgoo, R., Kiani, K., and Escalera, S. 2022. Real-time isolated hand sign language recognition using deep networks and svd. Journal of Ambient Intelligence and Humanized Computing, 13(1): 591-611.