Recognition of Indian Sign Language by using Hand Gestures and Speech Generation

By Gaja Lakshmi J

WORD COUNT 5458
TIME SUBMITTED 03-MAY-2023 02:56PM
PAPER ID 99326501

Recognition of Indian Sign Language by using Hand Gestures and Speech Generation

Abstract - **Communicating with the hearing-impaired person is always a great challenge. People with hearing disability uses hand gestures as a medium to communicate with others. By these hand gestures it is not necessary for the person to communicate vocally. There are various sign languages developed and used according to their native language. In this study, a system has been proposed to recognize Indian Sign Language (ISL) from hand gestures and generates text for the same. It is then converted to speech for effective understanding. A CNN based model is developed to recognize the Indian Sign Language with the hand gestures through computer vision in real-time. The model is trained on a dataset of ISL hand gestures and can recognize a wide range of gestures with high accuracy. The speech generation module converts the recognized gestures into speech using text-to-speech technology. The proposed system has potential applications in assisting the hearing-impaired community in India by providing them with an efficient and easy-to-use tool for communication. In addition, this system provides the paradigm of generating sign language from text, which enables communication in two ways without the use of an interpreter.**

Keywords: Indian sign language (ISL), Hand gestures, Computer Vision, Convolutional Neural Networks (CNN), Voice Output, Two-way communication

I. INTRODUCTION

Sign language is a visual communication used primarily by people with hearing disabilities. Instead of using acoustic sound patterns, sign language relies on gestures such as hand gestures, hand orientation, and facial emotions. Sign language can be used to communicate in situations when verbal communication is not possible, such as between speakers of mutually incomprehensible languages or when one or more communicators would be hearing impaired. This form of language is not universal and has various patterns depending on the people. People utilize a variety of sign languages all, Australian, French, Spanish, Chinese, Japanese, Indian, and many more are among the most popular sign languages. It is becoming more and more difficult for hearing-impaired people to communicate without an interpreter because the majority of people in this world are not well versed with sign language, which makes them feel isolated. This recognition model for Sign Language has gained widespread acceptance as a means of communicating medium between people with hearing disabilities and ordinary people. Indian Sign Language (ISL) is a natural language used by the hearing-impaired community in India. It is a visual language that relies on hand gestures and facial expressions for communication. With the advancement of technology, many people show great interest in developing tools and systems that can aid in ISL recognition and communication.

Real-time methods for recognizing sign language are sensor-based and vision-based. The user of the sensor-based technique must put on gloves. Flex sensors, accelerometers, and motion sensors are built into the glove to track hand movement. A webcam is used in the vision-based system to collect photos of sign language, which are subsequently processed to identify the sign language. Many studies have been conducted in American Sign Language (ASL), while very few have been carried out in Indian Sign Language (ISL). With the help of Indian sign language (ISL), this proposed model tries to bridge the communication gap between hearing persons and the hearing impaired persons. This model employs a CNN to recognize hand gestures using computer vision to generate the corresponding characters, words, and sentences using finger spelling (used to spell words letter by letter). Latterly, CNN

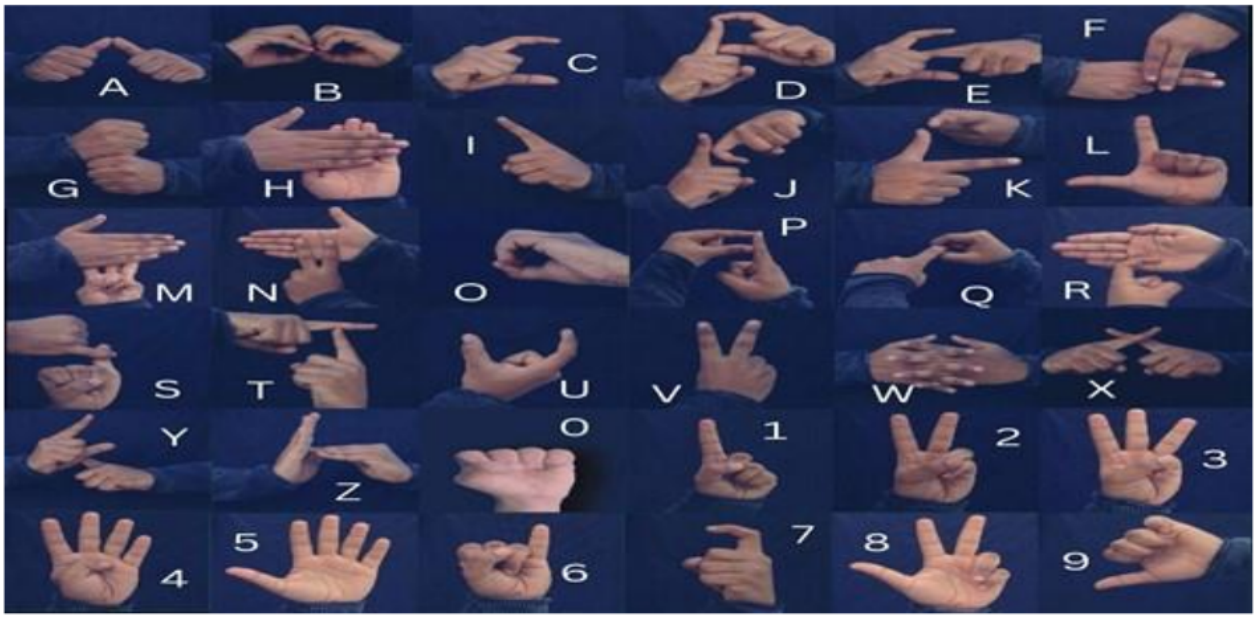

Figure1.Indian Sign Language Hand Gestures for Alphabets and Numbers

have shown great promise in image recognition tasks. By leveraging their ability to learn complex patterns and features from images, CNNs have been successfully applied to various recognition tasks. In this project, we propose an ISL recognition system based on CNN that can recognize a wide range of hand gestures with high accuracy. First, the sign images for ISL are captured using a webcam. The dataset consists of sign images for 26 alphabets, 0-9 digit and blank images for blank space as shown in figure 1. Second, the collected raw images are pre-processed and the feature is extracted. Third, the pre-processed image dataset of hand gestures is then trained using the proposed CNN based model to classify the images. Finally, the Indian sign language is recognized, and corresponding text is generated. To enable communication between hearing-impaired and blind persons, the created text is subsequently converted into speech to remove the communication barrier.

The objective is to propose a model to recognize hand gestures in real time and generate not only a single character but also a whole sentence based on finger spelling. After the text generation, it is converted to speech for better understanding. To enforce the two-way communication the hand gestures are also generated for the text or sentence,so that it will be useful for hearing impaired persons.

II. Related works

People who are hard of hearing exchange messages with one another using sign language. In recent times, a lot of explorations has been conducted on hand gesture recognition. Many studies have been conducted in ASL, but only a negligible amount of research has been conducted in ISL.

ISL recognition has been a valuable research topic in recent times due to the need to improve the conversation between ordinary and hearing-impaired peoples. Several approaches have been proposed for ISL recognition, including deep learning techniques, convolutional neural networks, and hidden Markov models.

Bambang KrismonoTriwijoyo [1] have used a deep learning method for sign language recognition that mainly focuses on image processing with background correction to improve model performance.

Shagun Katoch [2] have proposed a real-time recognition system for ISL using SURF with SVM and CNN with the constraint of delayed response to a live video stream.

In existing systems, much of the research on this topic uses both sensor-based and vision-based systems. Abey Abraham [3] proposed the use of flex sensors to predict sign language using back-propagation neural networks. The major problem with sensor-based systems is that to identify sign given off by hearing impaired people, users must attach expensive sensors to their gloves.

Mahesh Kumar N B [4] study demonstrates the usage of MATLAB to recognize 26 hand orientations and movements in ISL. The gesture was recognised using the Linear Discriminant Analysis (LDA) technique, and the recognized gesture was then transformed to text and speech format.

MehreenHurroo [5] trained only 10 alphabets to detect American Sign Language using 3 layers of CNN with TensorFlow library.

Ahmed Kasapbasi [6] proposed a CNN based human computer interface for ASL Recognition for hearing impaired persons.

Amrita Thakur [7] proposed a sign language recognition system in real time along with speech generation to recognize

26 alphabets using neural networks such as CNN and VGG model.

Rachana Patil [8] proposed an Indian Sign Language Recognition System using CNN to recognize only number digits (0-9) and the image processing is done using computer vision techniques such as Gray-scale, dilation, and mask operation.

Satwik Ram Kodandaram [9] proposed a SLR system that uses an ensemble model (LeNet-5, MobileNetV2, own CNN model) to recognize alphabets, numbers, and other dynamic gestures [9]. But it was unsuccessful to cover up other dynamic gestures.

Fayed F. M. Ghaleb [11] proposed a Computer Vision based Hand Gesture detection and Recognition Using CRF and SVM.

In this First, the hand is detected and segmented using a 3D depth map and the YCbCr colour space. The depth map's purpose is to mask complex background senses. Second, three orientation motion features are abstracted for the hand volume of dynamic affine invariants like elliptic Fourier and Zernike moments, together with 3D spatio-temporal features. Finally, the hand motion is identified using the discriminative Conditional Random Fields Model. As a result, a Support Vector Machine enforces a difficult view-invariant task by examining the shape of the hands at the start and finish of a meaningful gesture.

Mr. G. Sekhar Reddy [23] proposed deep neural networks which is used to develop a framework for SLR that directly translates signs into words. By utilising video sequences that include both temporal and spatial data. They trained the spatial and temporal features using two alternative models. They train the spatial features using the CNN model. RNN was used on the temporal features to train the model.

The development of Indian Sign Language recognition systems has great potential to improve the communication andquality of life for the hearing-impaired population in India.

III. Proposed Methodology

The proposed methodology for the ISL Recognition System aims to develop a comprehensive system capable of recognizing hand gestures by CNN, generating sentences using finger spelling as well as generating voice and convert text into sign language gestures. This system has great potential to improve conversation between Normal hearing and hearing impaired people in India. According to the WHO, there are around 63 million people with hearing impairment in India and sign language is the primary mode of communication for many of them. However, there is a significant gap in communication between ordinary and hearing-impaired people due to a lack of understanding of sign language in the community.This gap can be bridged by developing an effective sign language recognition system that can recognize hand gestures of the sign language and convert them into speech or sentences. The proposed method aims to fill this gap by combining individual recognized hand gesture to build sentence with speech generation and also additional feature of text-to-sign conversion. This system could have important applications in areas such as education, healthcare and media where effective communication are required. Therefore, the proposed systems have the ability to significant use for the hearing-impaired community in India and contribute to the overall economic and social development of the country.

Here is an overview of the proposed ISLR system process flow. The data collection process is the first stage of the proposed system. The proposed system's dataset consists of hand gesture images for alphabets and numbers. The following stage is data pre-processing. It involves converting the raw dataset to a grayscale image and applying a Gaussian blur filter to that image to extract its features during the data

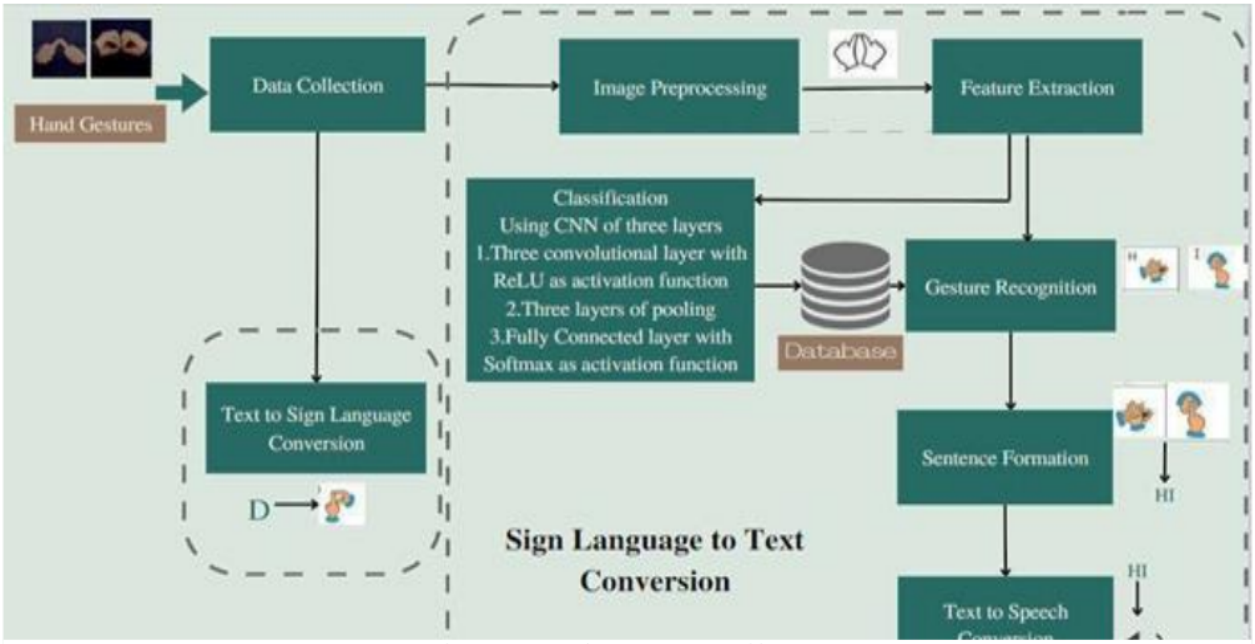

Figure 2.Proposed ISLR System Architecture

pre-processing stage. The next stage is the training phase. The proposed CNN-based model is trained to categorize images using hand gesture features extracted during the pre-processing stage.The proposed model includes three convolutional layers to extract high-level features from hand gestures, three pooling layers with ReLU activation functions to reduce dimensionality while preserving the most important feature, and a fully connected layer to classify the hand gestures.The following stage is gesture recognition. In this stage, the system uses OpenCV to capture the live camera feed, then pre-processes the shown gesture, and finally predicts the hand gesture using the trained proposed model. The next stage is sentence formation based on the recognized individual hand gesture. Next, the speech is generated for the sentence formed . The proposed ISLR system includes an additional module for converting text to sign that converts the text into equivalent hand gesture.

A. Data Collection

The first step is the data collection for Indian Sign Language Recognition System (ISLR). Because of a lack of study in this area, an appropriate data set for ISL is not yet readily available. To fill in the gaps left by unavailable data; we have generated the own dataset using data from a variety of sources. The dataset consists of images for alphabets (A-Z), numbers (0-9) and blank images for space in a sentence.

Therefore, the total collection has 37 categories. Each category comprises of 1200 hand gesture images along with the labels corresponding to them. Figure 3 shows sample dataset for the proposed ISL system.

Figure3. Sample Dataset

B. Data Pre-processing

The input image must first be converted to grayscale. Grayscale conversion simplifies the image by reducing the colour information to a single channel, making it easier to extract the important features of the hand gesture. Next, the grayscale image is then blurred using a Gaussian filter to lessen noise and smoothen the edges of the hand region as seen in figure 4. This step is required to improve the efficiency of the hand recognition algorithm and reduce false positives.

Finally, the hand region is then separated from the background using thresholding. Thresholding converts the grayscale hand gesture image into a binary image, where pixels with values above a certain threshold are set to white, and those below the threshold are set to black. The threshold value can be determined using various techniques such as Otsu's method, which selects the threshold value that minimizes the variance within the hand region.

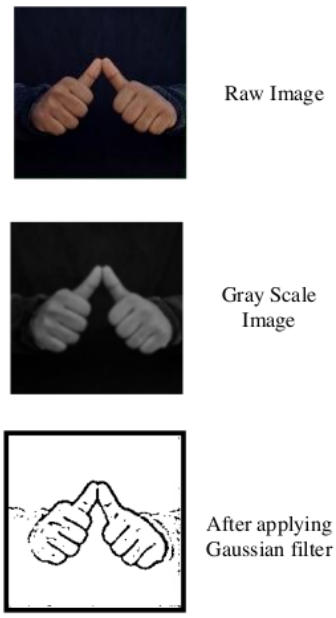

Figure 4. Hand Gesture Pre-processing

C. Classification

Once the ROI has been identified, we then use the deep learning techniques for training the image dataset for classification. Classification of hand gestures is an important step in the Indian Sign Language recognition system. It involves identifying a specific gesture made by the user and matching that gesture with the corresponding word or phrase in sign language. CNNs are a popular technique for gesture recognition since they can learn complex features from the input images of hand gestures and achieve high accuracy.

Proposed Convolutional Neural Network (CNN)

A CNN is a type of neural network that is generally used for classifying the images, recognizing the objects, and other computer vision tasks. Without the necessity of manual feature extraction, CNNs are built to learn relevant features on their own from the input image data.

The key idea behind a CNN is the use of convolutional filters, which are small matrices that slide over the source image and compute dot products with the pixel values in each local region. The output of the convolutional filter is a new set of values, which represent the presence or absence of certain features in the input image. These features can then be fed into subsequent layers of the CNN for further processing and classification.

CNNs are commonly used for ISLR using hand gesture[13]. The architecture of CNN basically consists of a series of convolutional and pooling layers, preceded by fully connected layers. An image of the hand gesture serves as the network's input, which is pre-processed to ensure that it is suitable for training.

Input Layer - The input image is taken in by this layer, which then sends it to the following layer for processing.

Convolutional Layer - In order to extract significant characteristics from the input image, such as edges, lines, and shapes, this layer performs convolution operations on the image using a variety of learnable filters. From the input image, this layer aids in the identification of local patterns and features.

ReLU [9]yer - The convolutional layer's output is given to this layer's Rectified Linear Unit (ReLU) activation function, which provides non-linearity to the model and aids in increasing its accuracy.

Pooling Layer - This layer takes the maximum or average value within a particular window or region to down sample the output of the previous layers. This helps in simplifying the model and lowering its complexity by reducing the output's dimensionality.

Dropout Layer - In order to avoid overfitting and increase the model's capacity to generalise, this layer randomly removes some of the neurons from the previous layer during training.

Fully Connected Layer - This layer applies a linear transformation to the output of the previous layers, followed by a non-linear activation function. The learned features are mapped to the output class labels with the use of this layer.

Output Layer - The predicted class label of the input image is produced by this layer, which provides the final output of this model.

Three convolutional layers with ReLU activation

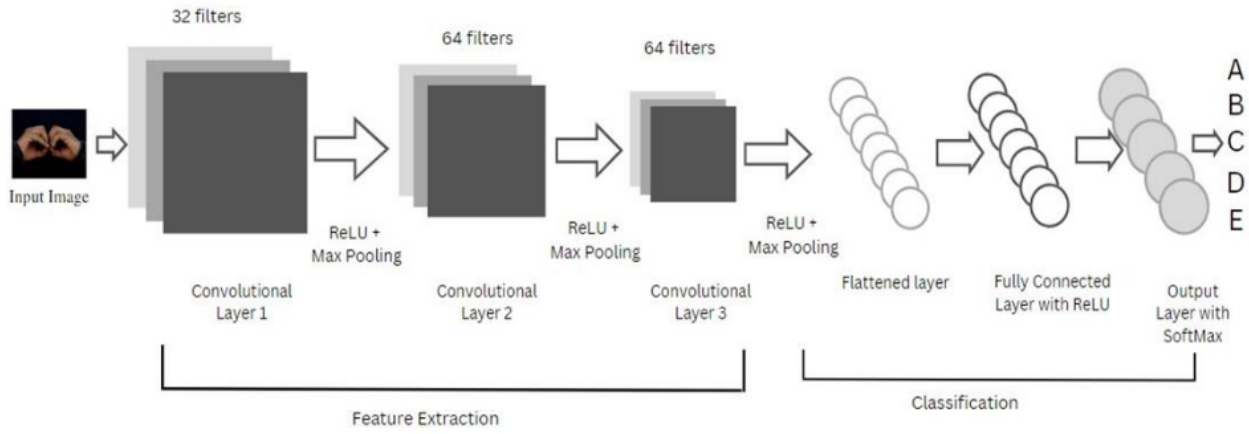

Figure 5. Proposed CNN based Architecture.

functions, followed by max pooling layers, a flatten layer, and a fully connected layer with a SoftMax activation function make up the proposed CNN architecture in figure 5 for ISLR utilizing hand gestures.

Three convolutional layers, three maximum pooling layers, a flatten layer that turns the output of the final pooling layer into a 1D array, and two fully linked layers make up the CNN architecture. The input picture shape is (64, 64, 1), indicating that the images are grayscale and have 64-pixel height and width.

64 filters with a kernel size of (3, 3) and a ReLU activation function make up the first convolutional layer. The max pooling layer is applied to the output feature maps with a pool size of (2, 2), which reduces the output's spatial dimensions in half.

64 filters with a kernel size of (3, 3) and a ReLU activation function make up the second convolutional layer. Yet again, the output feature maps are processed through a layer called Max Pooling with a pool size of (2, 2).

A ReLU activation function and 64 filters with a kernel size of (3, 3) are featured in the third convolutional layer. And again, the output feature maps are processed through a layer called Max Pooling with a pool size of (2, 2).

After being flattened into a 1D array, the output of the last pooling layer is fed into a fully linked layer with 128 units and a ReLU activation function. Finally, a fully connected layer with 37 units and the SoftMax activation function is sent

through the dense layer's output to provide the probability distribution for the 37 classes.

High-level features from the input image are extracted by the three convolutional layers. A ReLU activation function that provides non-linearity into the network and enables the network to learn more complicated features comes after each convolutional layer. A max pooling layer is applied after each convolutional layer to decrease the feature maps' spatial dimensions while retaining the most crucial characteristics. In order to feed the fully connected layer for classification, the output of the last max pooling layer must first be flattened into a one-dimensional vector. The fully connected layer features a SoftMax activation function that provides a probability distribution over the various classes and normalizes the layer's output.

Overall, this CNN architecture is well-suited for the proposed system of Indian sign language recognition using hand gestures. The multiple convolutional layers with ReLU activation functions allow the network to learn complex features, while the max pooling layers and flatten layer lessen the dimensionality of the feature maps and produce a one-dimensional vector for classification. Ultimately, the SoftMax activation function in the fully connected layer creates a probability distribution over the various classes, enabling the network to produce accurate predictions.

Convolutional layers aid in the identification and extraction of significant characteristics such as edges, shapes and lines from the input image of hand gestures.

ReLU activation function introduces the non-linearity to the proposed model, which helps in improving the model accuracy. Pooling layers makes the model less complex and easier to train by lowering the output's dimensionality. Dropout layers helps in avoiding overfitting and enhancing the model's capacity for generalization. The output class labels are mapped to the learned features using fully connected layers. Output layers produce the final prediction of the model for recognized hand gesture, making it suitable for classification tasks.

D. Gesture Recognition

In the proposed model, the hand gesture is recognized using ROI which is a popular technique for accurately identifying and classifying hand gestures. ROI refers to a specific region of the image where the hand gesture is expected to occur. By focusing on the ROI, the system can improve the accuracy of gesture recognition while reducing computational complexity.

The ROI is identified using the skin colour detection technique. Skin colour detection is a common technique used to identify the ROI since the hand is typically a different colour than the background. This technique involves applying a colour filter to the input image and extracting the pixels that fall within a certain colour range. In order to separate the hand region from the background, threshold is applied to the generated image.

E. Finger Spelling Recognition and Sentence Formation

Finger spelling recognition is an important component of Indian Sign Language (ISL) recognition, as it enables the system to recognize words and phrases that do not have specific hand gestures. Fingerspelling is the process of spelling out words letter-by-letter using hand gestures, and it is a common method of communication for the hearing-impaired community.

In proposed system, a letter is added to the current text if its count reaches a certain threshold number, and no other letters are comparable to it. The system displays alternative characters to add to the existing text if comparable letters are found. If the system notices a blank screen without a hand signal, it prints the space. The phrase is finally constructed by letter-by-letter recognition of the alphabetic hand gesture as seen in figure 6.

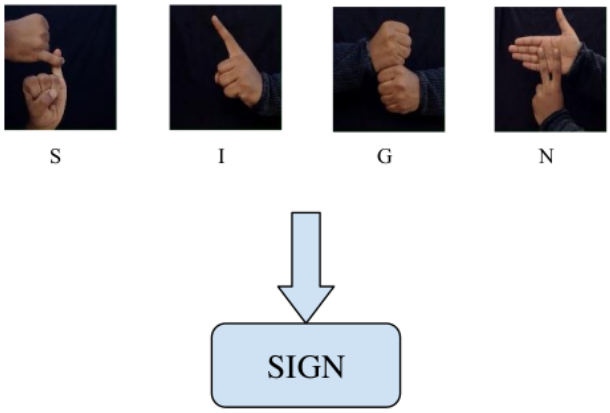

Figure 6. Sentence Formation using Finger Spelling Method

F. Speech Generation

In the proposed system, the text-to-speech conversion process enables the system to provide spoken language output for the sentence formed by detecting sign language gestures.

Text to speech systems work by analyzing the text input and breaking it down into individual phonemes, or units of sound. The system then uses a speech synthesis engine to generate the corresponding speech waveform for each phoneme, which is then combined to produce the final spoken output. The text to speech system uses python text to speech library (pyttsx3) to generate speech from text. Using the pyttsx library, developers can generate speech output in a variety of languages and accents, and customize the speed, volume, and pitch of the generated speech. The library also supports call-backs for events such as the start and end of speech, making it easy to integrate with other Python libraries and applications.

This can be especially useful for communication between the people with hearing impairment and the blind, as it enables them to communicate more effectively. Overall, text-to-speech

conversion is an essential component of SLR systems, as it enables the system to offer hearing-impaired people a systematic means of communication.

G. Text to Sign Conversion

The system enables two-way communication by providing the additional feature of converting text into the corresponding hand gesture as shown in figure 7. Text-to-sign conversion is a crucial component of Indian Sign Language (ISL) recognition systems as it enables the system to convert text into sign language gestures. The process involves converting the text input into sign language gestures by searching and retrieving the hand gesture from the classified dataset.

Step 1: Text Preprocessing

The first step is to preprocess the input text to remove any special characters, punctuations, and other unwanted characters. The input text is then converted to lowercase, and any stop words are removed. This step helps in reducing the complexity of the input and makes it easier for the system to recognize the gesture.

Step 2: Gesture Mapping

The second step involves mapping the preprocessed text to a corresponding hand gesture in Indian sign language. This is done using a gesture mapping dictionary, which contains the mapping of each word to its corresponding gesture. The mapping is done based on the standard Indian sign language dictionary, which contains a list of all the gestures used in Indian sign language.

Step 3: Image Rendering

The final step is to render the hand gesture images and display them to the user. The images are displayed ontothe screen.

Step 4: Voice Output for displayed gesture

The system also provides audio output to the user to help them understand the meaning of the gesture. The system will give voice output for each of the gesture it displays on the screen to make it understandable for learners and also for the users of ISLR system.

Overall, converting text into images of equivalent hand gestures in Indian sign language recognition involves a combination of computer vision, gesture recognition and text to voice generation techniques. The system uses a gesture mapping dictionary to map the text to a corresponding gesture and then displays the hand gesture images using computer vision. Finally, the images are rendered and displayed to the user, providing a seamless and intuitive way to communicate using sign language.

H. System Integration

Finally Integrate the hand gesture recognition system, finger spelling recognition system, sentence generation system, speech generation system, and text-to-sign conversion system to create a unified Indian Sign Language recognition system.

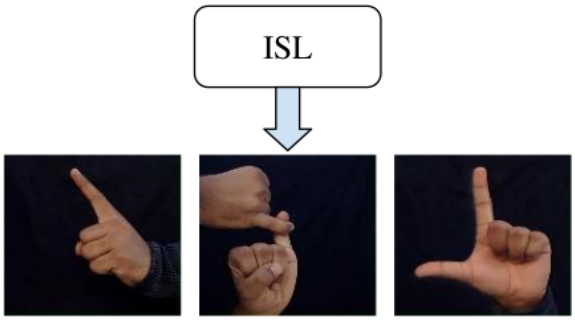

Figure 7. Text to Sign Language Conversion

In conclusion, this proposed methodology aims to develop a comprehensive ISLR system that can recognize hand gestures using CNN, generate sentences using finger spelling methods along with speech generation, and also the texts are converted to sign language gestures. This system can have significant applications in improving communication between the ordinary and people with hearing disability in the country.

IV. Experiments and Result

Images of hand movements corresponding to the alphabets and numeric symbols used in Indian Sign Language make up the dataset utilized in the proposed ISLR System. The collected data is intended to aid in the development of a proposed CNN-based model that is capable of effectively identifying and interpreting these gestures.

The dataset includes a total of 37 classes, representing the 26 alphabets, 10 digits and a blank image for blank space. The dataset is captured using a high-resolution camera that records the hand gestures at 30 frames per second, providing a rich and diverse set of datasets of hand gestures for training and testing the proposed model. The images and videos are captured in RGB format, providing colour information that can be used to improve the efficiency of the recognition model.

In data pre-processing, the process of Grayscale conversion removes the colour information from the images, resulting in a higher contrast and clearer edges, which can help the model to better distinguish between the different hand gestures. Gaussian blur can help to smooth out the images and reduce noise, making it easier for the model to identify the important features in the images. The formula for Gaussian filter is given in the below equation (1).

$$G(a,b) = \frac{1}{2\Pi\sigma^2} e^{-\frac{a^2+b^2}{2\sigma^2}} \quad (1)$$

where a and b are the location indices and σ is the standard deviation distribution. The Gaussian distribution's variance, which determines the value of the blurring effect surrounding a pixel, is controlled by the value of σ.

By reducing the noise and improving the quality of the images, pre-processing techniques like grayscale conversion and Gaussian blur can make it easier and faster for the proposed model to train on the dataset and to make accurate predictions.

Pre-processing the dataset in this way helps to reduce overfitting by removing unnecessary details from the images and focusing on the most important features, which leads to better generalization and more accurate predictions on new, unseen data.

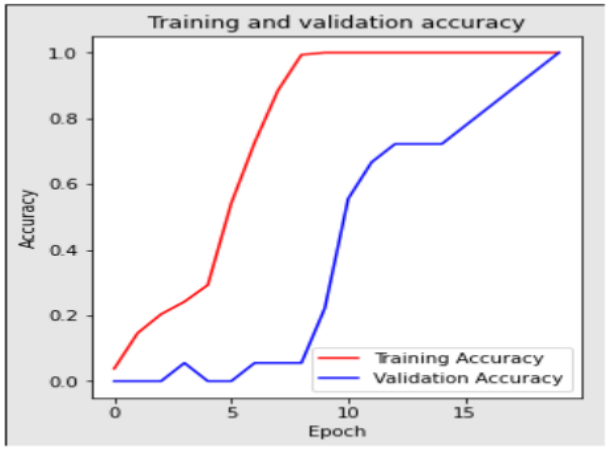

Figure8. Training and validation accuracy

The use of hand gestures in Sign Language Recognition employs a variety of algorithms to train the dataset. Here are some comparisons of popular algorithms.

With an accuracy of 85.45%,Sharvani Srivastava[12] proposed a TensorFlow Object Detection API method for Sign Language Recognition System . Jie Huang[13] proposed a Video-Based Sign Language Recognition without Temporal Segmentation using the method of LS-HAN and 3D CNN with an accuracy rate of 82.7%. Mandeep Kaur Ahuja[15] suggested a Principal Component Analysis method for Hand Gesture Recognition with accuracy of 91.43%. Neel Kamal Bhagat[25] propose a Indian Sign Language Gesture Recognition using Image Processing and Deep Learning and a method of Convolutional Neural Network , LSTMs , Microsoft Kinect with an accuracy of 78.3%. Tulay Karayilan[22] proposed a Sign Language Recognition using a method ANN(Backpropagation Algorithm) with accuracy rate of 85.7%. Fay F.M. Ghaleb[11] using a CRF and SVM method for Vision-Based Hand Gesture Spotting and Recognition with an accuracy rate of 92.50%.

Numerous algorithms and techniques are used in the current systems to recognize sign language. After researching various algorithm, it is concluded to propose a new Indian Sign Language Recognition using proposed CNN based model.

Table 1.Accuracy comparisons of various method with the proposed method

Method	Accuracy
TensorFlow Object Detection API [12]	85.45%
LS-HAN, 3D CNN [13]	82.7%
Principal Component Analysis (PCA) [15]	91.43%

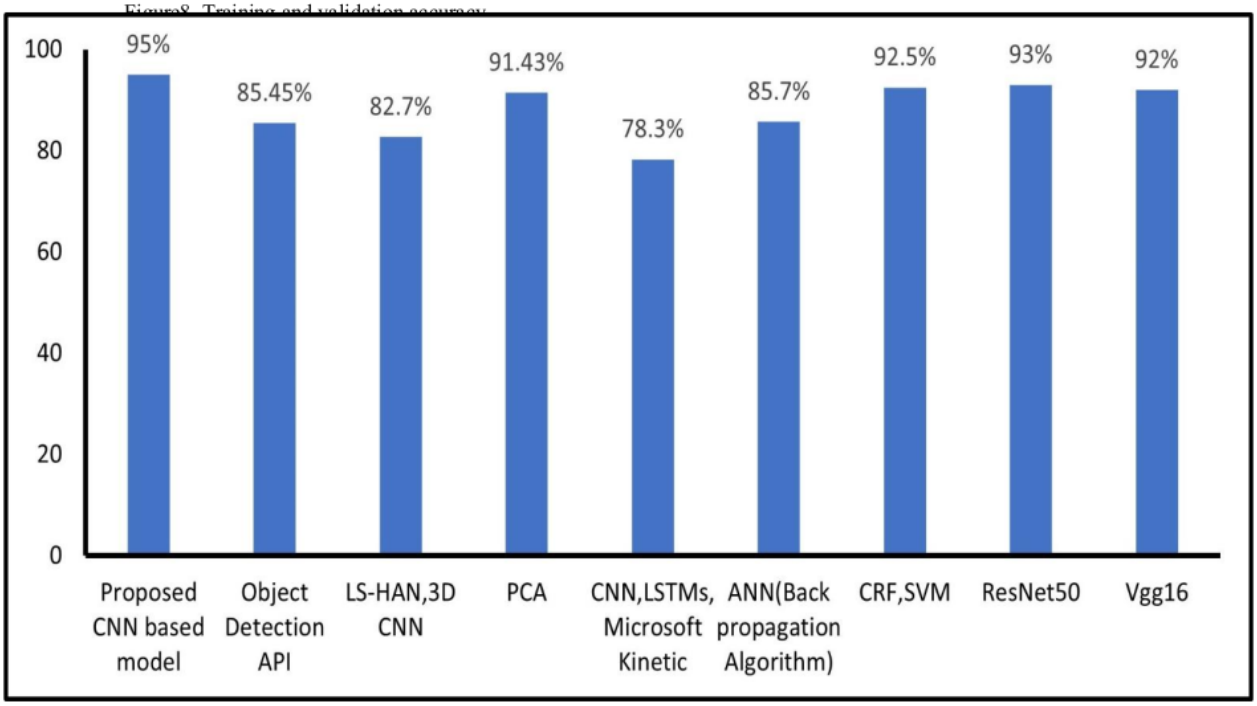

Figure 9. Model Accuracy (in %) Comparisons from Table 1

Convolutional Neural Networks, LSTMs, Microsoft Kinect [25]	78.3%
Artificial Neural Network (Backpropagation Algorithm) [22]	85.7%
Conditional Random Fields, Support Vector Machine [11]	92.50%
ResNet50	93%
VGG16	92%
Proposed CNN based Model	95%

The proposed CNN model consists of 37 classes including alphabets from A to Z and Numbers from 0 to 9 and blank image for space. The dataset contains a total of 44400, with each class having approximately 1200 images.

This technique makes use of a CNN architecture with three convolutional layers with 32, 64 filters, respectively. Each convolutional layer is followed by max pooling layer that have ReLU activation functions. The output of the third max pooling layer is flattened and fed into a dense layer with 128 units, and then fully connected layer that has a SoftMax activation function. The final output layer has 37 units, one for each class. The formula for the ReLU activation function is given as equation (2).

$$r(t) = \max(0, t) \qquad (2)$$

where the function's input is t, and its output is r(t). The activation function produces a piecewise linear function that is linear for positive input values and 0 for negative input values, returning t for positive input values and 0 for negative input values. The formula for SoftMax activation function is given as equation (3).

$$Softmax(v_x) = {e^{v_x}}/{\sum_y e^{v_y}} \qquad (3)$$

where v is a vector of real-valued scores and x and y are indices that range over the K classes. This ensures that the outputs of the SoftMax function are between 0 and 1 and that they sum to 1, which makes them suitable for representing probabilities over classes.

The following parameters were used in the training process of proposed CNN model. The model was trained for 100 epochs. This is a moderate number of epochs for training the proposed model, which allows the model to learn patterns and generalize well without overfitting to the training data. A small learning rate of 0.00019 was used. This small learning rate is useful because it helps in preventing overshooting the optimal solution and allows the model to converge to a more accurate solution.In the proposed Indian Sign Language Recognition model, there are 37 classes (hand gestures). The goal of the model is to predict the exact class for each input image. The predicted probabilities for each class are computed using the SoftMax activation function in the output layer of model.The loss function used in the proposed model is binary cross entropy.The binary cross-entropy loss function calculates the difference between the predicted probabilities and the actual class labels for each example in the training data. The formula for the binary cross entropy is given in equation(4).

$$Binary\ Cross\ Entrophy = \frac{1}{N}\sum_{i}^{N}\sum_{j}^{M} y_{ij}\log(p_{ij}) \qquad (4)$$

where N is the number of rows, M is the number of classes. The main benefit of using the binary cross-entropy loss function is that it is a well-defined and can be easily optimized using gradient descent-based algorithm.Here Adam optimizer is used in this training process.

Adam (Adaptive Moment Estimation) is an optimisation algorithm that updates model parameters by taking a moving average of the gradient and the squared gradient. It also employs bias correction to correct for the bias caused by the moving averages' initial estimates. Adam is used for the proposed model because it is an efficient optimisation algorithm for training the deep learning model and it can handle sparse gradients.

After training the proposed model for 100 epochs, the proposed model achieved an accuracy of 95% on the validation set as seen in figure 8.

Overall, the model was trained on the ISL image dataset

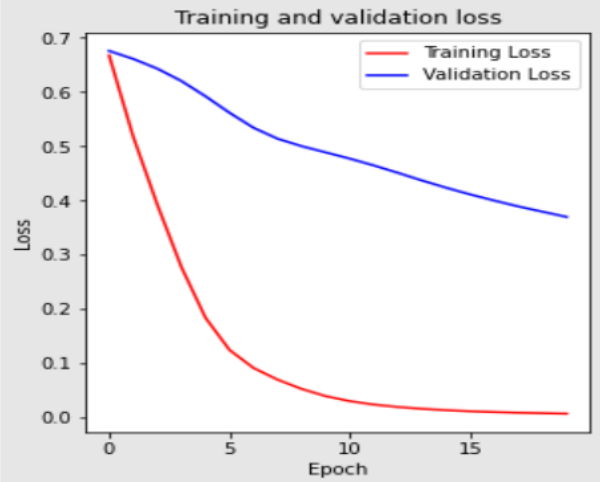

Figure10.Training and validation loss

of hand gestures using data enhancement techniques to increase the amount of training data. It is anticipated that the experiment, which uses the proposed CNN based model with three convolutional layers, three pooling layers, and a fully connected layer with SoftMax activation function and Adam optimizer, will provide findings with an accuracy rate of about 95%.

After comparisons with various techniques as seen in Table 1, it is predicted that the suggested CNN-based model would attain an accuracy rate of about 95%. Proposed system also includes sentence formation from gesture recognition, the system also used text-to-speech (TTS) conversion to generate spoken output from the sentence formed. The text to speech system uses python text to speech library to generate speech from text, which is then used to convert the recognized gestures into spoken output.

The proposed ISLR system's overall performance is effective in recognizing ISL gestures and generating spoken output.

CONCLUSION

These experiments for Indian Sign Language Recognition proves the effectiveness of the proposed model for recognizing ISL gestures and generating spoken output from the recognized gestures and the feature of text to sign language conversion. However, there is still a need for further research to improve the accuracy and robustness of these systems, as well as to develop more advanced speech generation systems for sign language recognition. While current ISL recognition systems focus primarily on hand gesture recognition and conversion of text to sign language, there is potential to integrate other modalities, such as real time handwritten text to sign language conversion to improve the ISLR System. Future research could explore the use of multi-modal recognition systems that combine hand gesture recognition with other modalities.

Recognition of Indian Sign Language by using Hand Gestures and Speech Generation

ORIGINALITY REPORT

8%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.researchgate.net Internet	52 words — 1%
2	Lecture Notes in Computer Science, 2015. Crossref	41 words — 1%
3	Khalid Manzoor, M. Arif Wani, Saduf Afzal. "A Deep Neural Network for the Detection of COVID-19 from Chest X-ray Images", 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom), 2022 Crossref	40 words — 1%
4	Luigi D'Arco, Haiying Wang, Huiru Zheng. "DeepHAR: a deep feed-forward neural network algorithm for smart insole-based human activity recognition", Neural Computing and Applications, 2023 Crossref	32 words — 1%
5	repository.tudelft.nl Internet	26 words — < 1%
6	escholarship.org Internet	20 words — < 1%
7	link.springer.com Internet	20 words — < 1%

8	mdpi-res.com Internet	20 words — < 1%
9	www.frontiersin.org Internet	19 words — < 1%
10	P. Paulachan, J. Siegert, I. Wiesler, Roland Brunner. "An End-to-End Convolutional Neural Network for Automated Failure Localisation and Characterisation of 3D Interconnects", Research Square Platform LLC, 2023 Crossref Posted Content	16 words — < 1%
11	www.astesj.com Internet	16 words — < 1%
12	"Bio-inspired Computing: Theories and Applications", Springer Science and Business Media LLC, 2020 Crossref	15 words — < 1%
13	"Proceedings of International Conference on Intelligent Vision and Computing (ICIVC 2022)", Springer Science and Business Media LLC, 2023 Crossref	15 words — < 1%
14	Fayed F. M. Ghaleb, Ebrahim A. Youness, Mahmoud Elmezain, Fatma Sh. Dewdar. "Vision-Based Hand Gesture Spotting and Recognition Using CRF and SVM", Journal of Software Engineering and Applications, 2015 Crossref	14 words — < 1%
15	Suhail Muhammad Kamal, Yidong Chen, Shaozi Li, Xiaodong Shi, Jiangbin Zheng. "Technical Approaches to Chinese Sign Language Processing: A Review", IEEE Access, 2019 Crossref	13 words — < 1%

16	www.hkmh.org Internet	12 words — < 1%
17	www.rsisinternational.org Internet	12 words — < 1%
18	Neena Aloysius, Geetha M, Prema Nedungadi. "Incorporating Relative Position Information in Transformer-Based Sign Language Recognition and Translation", IEEE Access, 2021 Crossref	11 words — < 1%
19	www.jetir.org Internet	11 words — < 1%
20	insideaiml.com Internet	9 words — < 1%
21	"From Bioinspired Systems and Biomedical Applications to Machine Learning", Springer Science and Business Media LLC, 2019 Crossref	8 words — < 1%
22	Enas M.F. El Houby, Nisreen I.R. Yassin. "Malignant and nonmalignant classification of breast lesions in mammograms using convolutional neural networks", Biomedical Signal Processing and Control, 2021 Crossref	8 words — < 1%
23	Futane, Pravin R., and Rajiv V. Dharaskar. "Video gestures identification and recognition using Fourier descriptor and general fuzzy minmax neural network for subset of Indian sign language", 2012 12th International Conference on Hybrid Intelligent Systems (HIS), 2012. Crossref	8 words — < 1%

24 Haimin Yang, Zhisong Pan, Qing Tao. "Robust and Adaptive Online Time Series Prediction with Long Short-Term Memory", Computational Intelligence and Neuroscience, 2017
Crossref
8 words — < 1%

25 SOUMEN DAS, Saroj kr. Biswas, Biswajit Purkayastha. "Occlusion Robust Sign Language Recognition System for Indian Sign Language Using CNN and Pose Features", Research Square Platform LLC, 2023
Crossref Posted Content
8 words — < 1%

26 Safayet Anowar Shurid, Khandaker Habibul Amin, Md. Shahnawaz Mirbahar, Dolan Karmaker et al. "Bangla Sign Language Recognition and Sentence Building Using Deep Learning", 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020
Crossref
8 words — < 1%

27 www.ncbi.nlm.nih.gov
Internet
8 words — < 1%

28 Rohan Singh, Mahesh Jangid. "Indian Sign language Recognition Using Color Space Model and Thresholding", 2021 Asian Conference on Innovation in Technology (ASIANCON), 2021
Crossref
7 words — < 1%

29 Agrawal, Subhash Chand, Anand Singh Jalal, and Charul Bhatnagar. "Redundancy removal for isolated gesture in Indian sign language and recognition using multi-class support vector machine", International Journal of Computational Vision and Robotics, 2014.
Crossref
6 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	OFF	EXCLUDE MATCHES	OFF