A Unified Communication System to Bridge the Gap for Inclusive Interaction among Individuals with and without Hearing Abilities

By Nandhini S

WORD COUNT 3109
TIME SUBMITTED 12-FEB-2024 12:35PM
PAPER ID 106833983

A Unified Communication System to Bridge the Gap for Inclusive Interaction among Individuals with and without Hearing Abilities

Abstract. To address the communication challenges faced by people with and without hearing abilities who want to communicate with each other. This solution utilizes real-time Computer Vision and Convolutional Neural Networks (CNN) to detect and translate sign languages into spoken languages. The system is designed to accurately recognize hand gestures and establish links to speech and broadcast messages, facilitating effective communication between people with and without hearing abilities. This solution caters to two distinct user groups: proactive communicators and those requiring translation assistance. By bridging the gap between speech and sign, this solution meets the communication and translation needs of the user, thereby enabling their active participation in various interactions. The effectiveness of the solution will be evaluated through rigorous testing, focusing on accuracy, real-time performance, and user satisfaction metrics.

I. INTRODUCTION

Communication is essential for everyone, and for individuals with hearing impairments, sign language serves as a crucial tool. It allows them to express themselves through hand gestures, facial expressions, and body movements. However, the sign language learners are mostly individuals with hearing impairment which makes it impossible for them to communicate with individuals with hearing abilities. In India alone, over 63 million people are estimated to have hearing-impairments, emphasizing the pressing need to address communication barriers within society. Limited understanding and proficiency in sign language among the general population further exacerbate these challenges.

Figure 1. Sign Language Hand Gestures for Alphabets and Digits, from Kaggle dataset.

To confront these obstacles, we propose a Sign Language Interpretation system that harnesses convolutional neural networks (CNNs). This system aims to accurately recognize hand gestures representing various elements of sign language, such as letters, words, phrases, and numbers. Beyond gesture recognition, our system is designed to generate complete sentences through finger spelling, enhancing comprehension for both hearing-impaired individuals and those without hearing impairments. This functionality facilitates seamless communication between these two groups, bridging the gap in understanding and promoting inclusivity.

The primary goal of our solution is to facilitate communication between individuals with and without hearing abilities. By promoting accessibility and inclusivity, our system strives to empower individuals with hearing impairments to fully participate in social interactions and everyday activities.

II. RELATED WORK

With the advancement of technology, there has been a growing interest in exploring hand gesture recognition techniques to facilitate communication for the hearing and speaking impaired community. Specifically, the recognition of Sign Language has emerged as a significant research area, driven by the urgent need to enhance communication between individuals with and without hearing abilities.

Recognition of Sign Language has emerged as a critical research area, driven by the requirement to enhance communication between individuals with hearing impairments and the general population. Various methodologies have been explored to address Sign Language recognition, encompassing deep learning techniques, convolutional neural networks (CNNs), and hidden Markov models (HMMs).

III. PROPOSED METHODOLOGY

The proposed methodology for the Sign Language interpretation System aims to develop a robust framework capable of recognizing hand gestures using CNNs, generating sentences through finger spelling, and converting speech into sign language gestures. Significant communication gap exists between ordinary individuals and those with hearing impairments due to a lack of understanding of sign language within the broader community. To address this gap, an effective sign language interpretation system is crucial. By recognizing hand gestures and converting them into speech, this system aims to bridge the communication barrier. Additionally, the proposed method integrates the individual recognition of hand gestures to construct complete sentences, while also incorporating features for speech-to-sign conversion. An Overview of the proposed Sign Language Interpretation System's process flow [Fig 2].

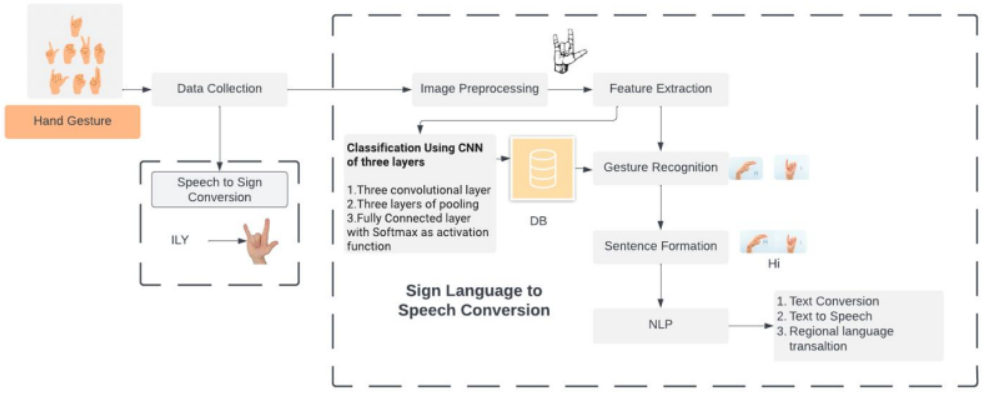

Figure 2. Proposed System Architecture

The first stage of the system involves data collection, where a diverse dataset of hand gesture images representing various sign languages from around the world is compiled. This dataset encompasses gestures for alphabets, numbers, and common phrases. Following data collection, the dataset undergoes preprocessing. This involves converting the raw images to grayscale and applying a Gaussian blur filter to enhance feature extraction capabilities. These pre-processed images serve as input for the subsequent stages of the system. The training phase utilizes a CNN-based model trained on the pre-processed dataset to categorize images based on hand gesture features.

The model architecture consists of multiple convolutional layers for feature extraction, pooling layers with activation functions for dimensionality reduction, and a fully connected layer for gesture classification. Gesture recognition is the next stage, where the system utilizes OpenCV to capture live camera feed input. This input is pre-

processed and fed into the trained model, which predicts the corresponding hand gesture in real-time. Once a hand gesture is recognized, the system proceeds to sentence formation based on the recognized individual gestures. This involves combining recognized gestures into intelligible sentences, representing the intended message. Subsequently, speech generation is initiated, where the formed sentence is converted into spoken language for better understanding by individuals with hearing impairments or those unfamiliar with sign language. Additionally, the proposed system incorporates a module for speech-to-sign conversion. This module allows users to input speech, which is then converted into equivalent hand gestures, providing an additional means of communication for users familiar with sign language.

A. Data Collection

The first step in creating a Sign Language interpretation system involves data collection, an important step given the lack of current language-level global sign language recognition data. To address this gap, we have collected data from various sources to provide services. For word and phrase recognition analysis, a comprehensive global dataset comprising 200 commonly used words and phrases were utilized. Each word is represented by several scenes shot from different angles and lighting to ensure the strength and breadth of the model. This paper promotes better communication and participation for individuals who use sign language through data collection and design.

Figure 3. Sample Dataset

B. Data Pre-Processing

Raw Image

Gray Scale

After Applying Gaussian Filter

Figure 4. Hand Gesture Pre-processing

The initial step involves converting the input image into grayscale, simplifying it by condensing colour information to a single channel. Subsequently, a Gaussian filter is applied to blur the grayscale image, reducing noise and smoothing hand gesture edges [Fig 4]. This is crucial for optimizing the hand recognition algorithm and minimizing false positives. Ultimately, the hand region is isolated from the background through thresholding. This process transforms the grayscale hand gesture image into a binary image, where pixels above a specified threshold turn white, and those below turn black. Various techniques can be employed to determine the threshold value by minimizing variance within the hand region.

The pre-processing steps of converting images to grayscale and applying Gaussian blur enhance image quality and aid in distinguishing hand gestures effectively for Sign Language Recognition models. Grayscale conversion removes colour information, enhancing contrast and edge clarity, while Gaussian blur smoothens images and reduces noise. The Gaussian filter formula is utilized for this purpose. These pre-processing methods streamline the training process, improve generalization, and accuracy by prioritizing essential features and mitigating over fitting. Various algorithms are compared for Sign Language Recognition dataset training. The Gaussian filter formula (1) is utilized for this purpose:

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

Where a and b represent location indices, and σ denotes the standard deviation distribution. The variance of the Gaussian distribution, controlled by σ, determines the blurring effect surrounding a pixel.

Proposed Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) serves as a pivotal tool in recognizing and classifying hand gestures for the proposed Sign Language Interpretation System. CNNs are expert at automatically learning relevant features from input image data, eliminating the need for manual feature extraction. The core concept behind CNNs [Fig 5] lies in the use of convolutional filters, which slide over the input image, computing dot products with pixel values in local regions to extract features. These features are then processed through subsequent layers for further analysis and classification.

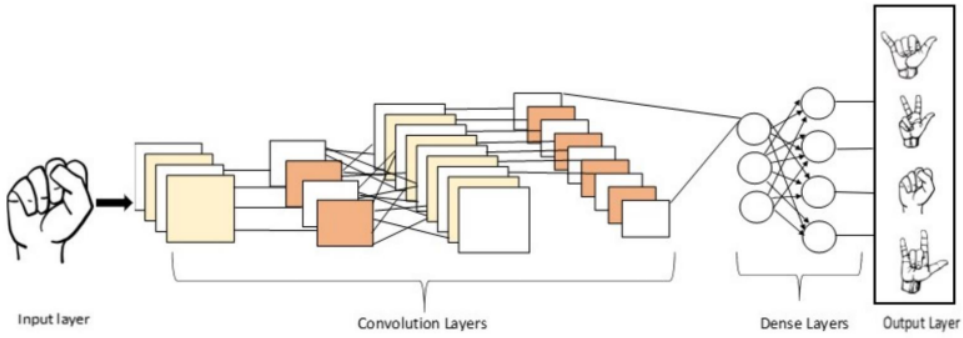

Figure 5. Proposed CNN based Architecture.

Each layer of CNN architecture serves a distinct function in the feature extraction and classification process:

- Input Layer: Receives the input image for self-generated dataset and frames from global dataset.

- Convolutional Layers: Extract significant features such as edges, shapes, and lines from the input images and frames.

- ReLU Activation Function: Introduces non-linearity to enhance model accuracy.

The CNN model comprises three convolutional layers with 64 filters each, followed by max pooling layers and ReLU activation functions. The third max pooling layer's output is flattened and fed into a dense layer with 128 units, followed by a fully connected layer with SoftMax activation. The final output layer has 36 units representing each class. The ReLU activation function is defined by equation (2).

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

- Pooling Layers: Down sample the output, reducing dimensionality while preserving crucial features.

- Dropout Layer: Prevents overfitting and promotes generalization by randomly removing neurons during training.

- Fully Connected Layers: Apply linear transformations to learned features, followed by SoftMax activation functions for classification.

In the proposed SoftMax activation function, denoted as r(t), the input is represented by t. This function yields a piecewise linear behavior, remaining linear for positive input values and outputting 0 for negative input values. Specifically, it returns t for positive inputs and 0 for negative inputs. The formula for the SoftMax activation function is depicted as equation (3).

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

- Output Layer: Produces the final prediction for the recognized hand gesture.

The proposed CNN architecture comprises multiple convolutional layers with ReLU activation functions, followed by max pooling layers and a flatten layer to simplify the model and create a one-dimensional vector for classification. The final fully connected layer utilizes an SoftMax activation function to generate a probability distribution over the classes, enabling predictions.

C. Classification

After identifying the Region of Interest (ROI) that is hand, deep learning techniques are utilized to train the image dataset for classification purposes. This step is crucial in the Sign Language interpretation system, as it entails recognizing the user's specific gesture and associating it with the corresponding sign language word or phrase. Convolutional Neural Networks (CNNs) are favoured for gesture recognition due to their ability to learn features from hand gesture images and attain impressive accuracy levels.

D. Gesture Recognition

This method entails applying a colour filter to the input image and isolating pixels within a specified colour range. To distinguish the hand region from the background, a threshold is applied to the resulting image.

The CNN model was trained for 100 epochs to strike a balance between learning and generalization. An average training value of 0.00019 was used to prevent overshooting and guide the model towards the correct solution. With 36 categories for Indian languages, the SoftMax function was applied to the output to calculate predicted probabilities for each class. The binary cross-entropy loss function (Equation 4) was employed to measure the disparity between predicted and actual classes for each sample in the dataset. This loss function is advantageous due to its well-defined nature and suitability for optimization using gradient descent-based algorithms. The Adam optimizer was utilized in this training process. The binary cross-entropy model is shown in Equation (4).

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

E. Finger Spelling Recognition and Sentence Formation

Figure 6. Sentence Formation using Finger Spelling Method

Finger spelling recognition plays a vital role in the proposed Sign Language Interpretation System, enabling the identification of words and phrases not represented by specific hand gestures. Fingerspelling involves spelling out

words letter-by-letter through hand gestures and serves as a common mode of communication for individuals with hearing impairments.

In this system, a letter is appended to the existing text when its occurrence frequency surpasses a predetermined threshold, and no other letters exhibit comparable frequencies. The system offers alternative character options if similar letters are detected, ensuring accuracy in text formation. Additionally, the system detects blank screens without hand signals, representing spaces between words in the constructed phrases. Ultimately, the phrase is assembled through sequential recognition of alphabetic hand gestures, facilitating effective communication for users of sign language.

F. Speech Generation

In the envisioned system, the process of converting text to speech plays a crucial role, allowing the system to articulate spoken language corresponding to the recognized sign language gestures.

Text-to-speech systems operate by analysing textual input and breaking it down into individual word elements, representing sound units. A speech synthesis engine is then employed to generate speech waveforms for each word element, and these are combined to form the final spoken output. The system utilizes the pyttsx3 Python library for text-to-speech conversion, providing flexibility in language, accent, and customizable parameters such as speed, volume, and pitch. This library supports event callbacks, enhancing integration with other Python applications.

G. Speech to Sign Conversion

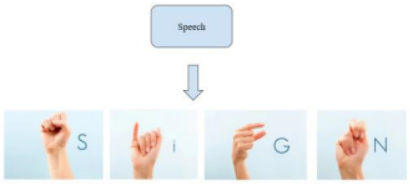

Figure 7. Text to Sign Language Conversion

The system facilitates two-way communication by incorporating a speech-to-sign conversion feature, demonstrated in [Fig 7]. This functionality is fundamental in Sign Language Interpretation Systems, enabling the translation of text into corresponding hand gestures for communication.

Step 1: Speech to Text and Preprocessing

The initial step involves preprocessing the input speech which is initially converted into text. The text is then converted to lowercase, and stop words are removed to streamline the input for gesture recognition.

Step 2: Gesture Mapping

Pre-processed text is mapped to appropriate hand gestures from the global dataset. Aligning each word with its corresponding gesture based on established sign language dictionaries.

Step 3: Image Rendering

Identified hand gesture images are rendered and presented to the user for communication.

Step 4: Voice Output for Displayed Gesture

System provides voice output accompanying each displayed gesture, aiding learners and users.

H. System Integration

The final step involves integrating the hand gesture recognition, finger spelling recognition, sentence generation, speech generation, and speech-to-sign conversion components to develop a unified system for recognizing Sign Language.

In conclusion, this proposed methodology strives to build a holistic Sign Language Interpretation system capable of hand gesture recognition using CNN, sentence generation through finger spelling techniques, speech generation, and speech-to-sign conversion. The system holds immense potential in enhancing communication between individuals with and without hearing disabilities on a global scale.

IV. EXPERIMENTS AND RESULT

In this study, we utilize self-produced datasets containing captured hand gestures portraying alphabets and numerals. These datasets are carefully selected and organized to ensure that they are accurate and directly relevant to our research on sign language recognition. For word and phrase recognition analysis, we utilize a comprehensive global dataset comprising 200 commonly used words and phrases. Through the integration of our self-generated datasets with this extensive global dataset. Our study underscores the significance of inclusive and varied datasets in the development of efficient communication solutions for individuals with hearing impairments.

In sign language, communication relies on various parameters to convey messages effectively. Handshape, movement, location, orientation, non-manual markers, timing/rhythm, symmetry/asymmetry, contact, expression, and space all play crucial roles in this complex system. Handshape refers to the configuration of the hands and fingers to create distinct signs, while movement adds dynamism and expression. Location indicates where signs are made in relation to the signer's body or the surrounding space, providing context. Orientation influences interpretation by indicating the direction of the hand or palm. Non-manual markers, like facial expressions and body language, convey emotions and emphasis. Timing and rhythm establish the flow of sign language, while symmetry/asymmetry and contact add variety and meaning. Expression, encompassing facial expressions and body language, gives signs depth and context. Space utilization uses the area around the signer to convey information such as locations or relationships between objects. Together, these elements form a rich and nuanced communication system, facilitating meaningful interactions in both deaf and hearing communities.

By enhancing image quality and minimizing noise, pre-processing methods such as grayscale conversion and Gaussian blur streamline the training process for the proposed model, facilitating quicker and accurate predictions on the dataset. This approach aids in mitigating overfitting by eliminating unnecessary image details and prioritizing essential features, resulting in improved generalization and accuracy when predicting on new, unseen data. Sign Language Recognition leverages diverse algorithms for dataset training. Below are comparisons of well-known algorithms in this domain.

With an accuracy of 90.45% Jayesh, Nilkanth, Sunil [1] proposed MRS CRF algorithm for dynamic hand gesture recognition. Sai Bharat et all [4] proposed CNN algorithm of accuracy 95% to extract spatial features from image. Navya, Ayushi et all [6] proposed HMM algorithm of accuracy 100% to convert speech to text. Saurdi, Chitra [7] proposed CNN algorithm of accuracy 62.63% to find the finger spelling. Kaur, Krishna [8] proposed SIFT algorithm of accuracy 99.43% to extract features and optimize it. Vivek, Dianna [10] proposed CNN algorithm of accuracy 82.5% to recognize alphabet and 97% to recognize digits and only 67% to 70% for self-generated dataset. Manikandan et all [11] proposed OpenCV tool with accuracy of 75% to 85% for color analysis and feature extraction. Hauda, Juan, Kalika [13] proposed OpenPose and MediaPipe tool of accuracy 75% to 84% for 2D and 3D normalization. Katoch et all [15] proposed SVM algorithm of 99.14% and CNN of 94% to 99% for feature extraction and recognition.

The proposed CNN architecture comprises 36 classes, encompassing alphabets from A to Z, numbers from 0 to 9, and a blank image for spacing. The dataset comprises a total of 44,400 images, with each class containing approximately 1,200 images.

Adam (Adaptive Moment Estimation) stands as an optimization algorithm that updates model parameters by computing a moving average of both the gradient and the squared gradient. Additionally, it incorporates bias correction to mitigate biases arising from the initial estimates of the moving averages. Adam proves efficient for optimizing the proposed model, particularly in training deep learning models and handling sparse gradients.

Method	Accuracy
MRS CRF algorithm [1]	90.45%
Open CV [11]	75%-85%
MediaPipe, OpenPose [13]	75%-84%
TensorFlow Object Detection API [16]	85.45%
LS-HAN, 3D CNN [17]	82.7%
Convolutional Neural Networks, LSTMs, Microsoft Kinect [18]	78.3%
Artificial Neural Network (Backpropagation Algorithm) [19]	85.7%
Conditional Random Fields, Support Vector Machine [20]	92.5%
ResNet50	93%
VGG16	92%
Proposed CNN Model	95%

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

After evaluating various techniques as presented in Table 1, we are expecting that the proposed CNN-based model will be expected to achieve a 95% accuracy rate. Additionally, the system facilitates sentence formation from gesture recognition and integrates text-to-speech (TTS) conversion to produce spoken output from the formed sentences. Utilizing a Python text-to-speech library, the system generates speech from text, enabling the conversion of recognized gestures into spoken output.

V. CONCLUSION

The experiments conducted on sign language recognition showcase the efficacy of our model in identifying gestures and converting them into spoken language. However, additional research is imperative to refine the accuracy and resilience of these systems. Moreover, there's a pressing need to advance speech generation systems customized for sign language recognition. While current recognition systems primarily center on hand gesture recognition and text-to-sign language conversion, there's room to integrate other modalities for enhanced sign language recognition. Future research could delve into implementing multi-modal recognition systems that amalgamate hand gesture recognition with other modalities.

A Unified Communication System to Bridge the Gap for Inclusive Interaction among Individuals with and without Hearing Abilities

ORIGINALITY REPORT

3%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.researchsquare.com Internet	18 words — 1%
2	Gopalakrishnan Srinivasan, Kaushik Roy. "ReStoCNet: Residual Stochastic Binary Convolutional Spiking Neural Network for Memory-Efficient Neuromorphic Computing", Frontiers in Neuroscience, 2019 Crossref	15 words — < 1%
3	raw.githubusercontent.com Internet	12 words — < 1%
4	academic.oup.com Internet	11 words — < 1%
5	www.techtarget.com Internet	10 words — < 1%
6	Adithya, V., P. R. Vinod, and Usha Gopalakrishnan. "Artificial neural network based method for Indian sign language recognition", 2013 IEEE CONFERENCE ON INFORMATION AND COMMUNICATION TECHNOLOGIES, 2013. Crossref	9 words — < 1%
7	adayinourshoes.com Internet	9 words — < 1%

8 arxiv.org
Internet

9 words — < 1%

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