Proceedings of the Second International Conference on Artificial Intelligence and Smart Energy (ICAIS-2022)
IEEE Xplore Part Number: CFP22OAB-ART; ISBN: 978-1-6654-0052-7

Hand Gesture Recognition and voice, text conversion using CNN and ANN

Surekha P
Computer Science Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
prekha.572@gmail.com

Niharika Vitta
Computer Science Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
niharika6112000@gmail.com

Teja Sree Desani
Computer Science Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
desani.tejasree2000@gmail.com

Pranavi Duggirala
Computer Science Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
pranavi.reddy01@gmail.com

Venkata Surya Saranya Ambadipudi
Computer Science Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
a.suryasaranya@gmail.com

2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS) | 978-1-6654-0052-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAIS53314.2022.9743064

***Abstract*— Individuals primarily communicate with one another. Blind and deaf people use sign language to communicate with others. These individuals have difficulty communicating their message to ordinary people. Deaf and blind people believe they are unable to communicate because of a lack of communication skills, and as a result, they are unable to express their emotions. Because most individuals aren't educated in sign language, communicating in an emergency can be extremely challenging. As a consequence, the challenge may be solved by converting hand gestures into human-hearing sounds and text. Vision and non-vision approaches are two of the most commonly used methods for detecting hand movements or gestures. In a vision-based approach, a camera will be used for gesture detection, whereas sensors will be employed in a non-vision-based technique. In this study, a vision-based technique was used. This device detects and locates hand motions in order to keep a communication channel open with others. Using convolutional neural networks and artificial neural networks, this research develops a gesture recognition system. This study looks into the advantages and disadvantages of hand motion recognition.**

Keywords—Hand gesture, Gesture Recognition, Sign Language, Dumb and Deaf, Feature Extraction, Deep Learning, Webcam, Image Pre-processing.

I. INTRODUCTION

Sign language is becoming more popular as a technique to communicate with those who are unable to communicate verbally. It is a language in which hand motions are used to express alphabets and words. The vision technique has been the most extensively used method for sign recognition in recent decades. It's a technology that uses a camera to identify data transmitted by finger motions. It is the most commonly used visual-based method. Vision-based sign recognition systems have taken a lot of time and effort to develop all over the world. The two vision-based gesture recognition systems are direct and indirect. Previously, for the recognition of hand gestures, a vision-based approach was used. However, the ambient influence on the detected picture is significant in this approach. The hand motion is detected and converted into speech and text.

One of the most important challenges that this one-of-a-kind personality suffers from is the communication gap between a disabled person and an ordinary person. Due to a lack of communication, deaf and dumb people are unable to express their feelings. Hand Gesture Recognition and Voice Conversion (HGRVC) technology identifies and monitors the hand motions of the deaf and dumb, allowing them to converse with others. Webcams are used to detect hand movements. With the help of pre-processing, the images are then converted to normal size. The goal of this study is to create a system that can translate hand gesture into speech and text. Hand gesture is analysed as part of the identification. The technology provides text output, which helps deaf people, and also speech output, which helps blind people and humans communicate more effectively.

A. Problem Statement

To communicate with the wider public, deaf and blind people rely significantly on sign language. Those persons find it challenging to express their message to regular people. Due to a lack of communication, deaf and blind people are unable to transmit and express their feelings. Hand signals can be converted into human voice and writing to remedy the problem. People engage with one other mostly through communication. Dumb and blind persons use sign language to communicate with those who are not deaf or blind. Those individuals find it extremely difficult to communicate their message to the general public. They are uneasy about taking on such a massive task. Dumb and blind individuals believe they are unable to communicate because of a lack of communication, and as a result, they are unable to convey their emotions. Because most individuals aren't trained in sign language, communicating their message in an emergency is extremely challenging. As a result, the solution to this challenge is to transform hand movements into human hearing voice and text.

B. Theoretical Background

The invention of glove-based control interfaces paved the way for hand gesture recognition for computer control. Researchers discovered that basic orders to a computer interface may be given using motions inspired by sign language. This has progressively evolved with the development of more precise accelerometers, infrared cameras, and even fiber optic bend-sensors. Some of these developments in glove-based systems enable computer vision-based recognition without the need for any sensors on the glove. These are the colorful or multicolored gloves that are utilized in computer vision-based gesture recognition for finger tracking.

[4] J RajaLaxmi and P Kumar suggested a Gesture Recognition System based on a deep neural network that combines the Convolution Neural Network (CNN) with the Recurrent Neural Network (RNN) (RNN). [12] Image processing may be divided into two categories: analogue and digital. Hard copies, such as printouts and pictures, can benefit from analogue image processing. When applying these visual tools, image analysts use a variety of interpretive basics. The use of computers to manipulate digital photographs is aided by digital image processing techniques. Pre-processing, augmentation, and presentation, and information extraction are the three main processes that all sorts of data must go through when employing digital techniques. [3] Handwriting recognition computers are believed to be able to acquire and detect characters on paper, photographs, touch screens, and other sources and convert them into machine-encoded form. [5] B Vivekanandam presented a technique for preserving acquired photographs at several periods that comprises of a camera and a cloud storage device. Through an internet connection, the photographs may be downloaded to a local data storage device. [6] Speech recognition is a critical component of many signal processing applications. It is a method for converting verbal content voice signals into a machine-readable format known as "Text.". [1] In order to construct smart cities, municipal wireless networks employing Wi-Fi connectivity are being developed. Wi-Fi is available in tens of thousands of towns around the world. [2] Tripathi Milan's proposed deep learning-based image classification approaches are evaluated using a variety of classifiers in order to achieve a better result. The image data augmentation technique is used to increase the size of the training dataset in order to improve the model's performance and generalization capacity. [10] In the current situation, most communication is done through vocal sounds and body language gestures. While vocal sound plays a vital part in communication, diverse body language gestures give even more weight to the conversation. [11] In target recognition and object detection, the directional gradient histogram (HOG) has strong geometric and optical invariance, making it one of the best features for extracting edge and contour information. [7] Vision-based hand gesture recognition systems that were specifically built to recognize sign language. The current problem is that most people are unable to comprehend hand signals or convert them into spoken language quickly enough for the listener to understand. Furthermore, utilizing sign language to communicate is tough. [8] Sign language is required for deaf people to communicate with non-deaf people. Silent people have a difficult time communicating their message to non-mute people. Most people are unable to communicate successfully because they have never been trained how to utilize hand signs. [9] P.V.Krishna Rao and V.Niharika presented a system display with flex sensors, that can see sign language, allowing ordinary individuals to communicate more effectively with persons who are unable to speak. [13] A gesture is a bodily movement of the head, face, arms, hand, or body that conveys a message. A computing device's mathematical interpretation of a human motion is known as gesture recognition. [14] A hand gesture that might be used as an educator as well as a means of communication amongst deaf, dumb, and mute persons. Individuals who are deaf or dumb have difficulty communicating with regular people and expressing themselves. [15] The majority of today's human computer interaction interfaces need explicit user instructions in the form of keyboard taps or mouse clicks.

C. Automated Gesture Recognition

The research mentioned in this system aims to develop a system that can assist deaf-dumb people by translating their voices into sign language. This approach converts the verbal signal into American Sign Language (ASL). The sign from the American Sign Language Lexicon is shown in a prerecorded ASL. On the monitor of a laptop computer, an ASL representation of a sign is displayed. In real life, Deaf people do this on their terms without the use of specialized signs, such as proper names. A Hidden Markov Model (HMM) is used to identify audio signals from the user and transform them into cue symbols for those who have trouble speaking. The proposed challenge is a supplement to the ongoing study into identifying a vocally handicapped person's finger movement as a speech signal known as "Boltay Haath". The suggested AISR system, when combined with the Boltay Haath system, has the ability to bridge the communication gap between the general public and persons with vocal difficulties.

D. Finger Detection

Computer recognition of sign language is a crucial study topic for hearing-impaired people to communicate. The number of fingers open in an American Sign Language gesture indicating an alphabet is determined using the efficient and quick method outlined in this study. To recognize fingers, the concepts of border tracing and fingertip detection are applied. The system does not require the use of any special markers or input gloves, nor does it require the hand to be correctly oriented to the camera. Boundary tracing, computer access for the disabled, finger detection, picture processing, and sign language recognition are all terms that can be found in the index.

E. Existing System

In existing system, Anchal Sood and Anju Mishra have suggested a sign recognition system based on the Harris algorithm for feature extraction, in which the feature is

extracted and stored in the Nx2 matrix after the picture pre-processing stage. The image from the database is then matched using this matrix. The system does have certain limitations. Because they are regarded in the range value for skin segmentation, the very light brown to fairly dark brown backdrop causes inaccuracy. However, the outcomes are effective.

II. PROPOSED APPROACH

The proposed approach is developed on Windows-10 Operating System with 4GB RAM in Python programming language. However, it can also be developed in Windows-7 or above and Linux operating systems.

Deep Learning methods employed in the proposed system include Convolutional Neural Networks and Artificial Neural Networks. Deep learning is a method to machine learning and artificial intelligence (AI) that is based on the human learning process. Deep learning is a significant component of data science, which includes statistics and predictive modeling. For data scientists who must collect, analyze, and interpret massive amounts of data, deep learning is particularly useful since it speeds up and simplifies the process. At its most basic level, deep learning may be thought of as a way to automate predictive analytics.

CNN: A CNN (Convolutional Neural Network) is a Deep Learning technology that may take an image as input and prioritize or discriminate between numerous attributes or objects in the picture. A CNN requires much less pre-processing than other classification methods. Simple procedures are used to manually engineer filters. If given enough training, CNN can learn these filters and attributes.

The structure of a CNN was influenced by the architecture of the visual cortex, which is comparable to the connection pattern of neurons in the human brain. Individual neurons can only respond to stimuli that fall inside the receptive field, which is a small fraction of the visual field. A number of comparable fields can be stacked on top of one another to cover the whole visual field.

ANN: The phrase "Artificial Neural Network" comes from biological neural networks, which define the structure of the human brain. Artificial neural networks, like the human brain, include neurons that are coupled to one another in various levels of the networks. Nodes are the name for these neurons. Artificial neurons are a set of linked units or nodes in an ANN that loosely replicate the neurons in a biological brain. Each link may send a signal to other neurons, just like synapses in a human brain. An artificial neuron receives a signal, analyses it, and then sends signals to neurons it is linked to.

A non-linear function of a neuron's inputs determines its output, and the "signal" at a connection is a real number. Edges are the terms used to describe connections. As learning continues, the weight of neurons and edges fluctuates often. The weight influences the signal strength at a connection. Neurons may have a signal threshold over which they may only transmit if the total signal exceeds it. Layers of neurons are frequent in the brain. The inputs of various levels can be subjected to various modifications. Before reaching the output layer, signals may pass through the input layer multiple times

A. Proposed Modules

The proposed approach consists of the following modules:

1) Upload Hand Gesture Dataset: This project takes advantage of Kaggle's hand gesture recognition dataset. Various hand gesture images are included in this dataset. The data collection includes different hand movements.

2) Preprocess Dataset: Data pre-processing is a technique for transforming raw data into clean data. When data is acquired from several sources, it is in raw format, making analysis impossible.

3) Model Generation: In machine learning, model generation is an iterative process in which the machine learning models are constantly trained and tested to find the optimal one for a particular job. There are a variety of machine learning models to choose from, and the one to choose is mostly determined by the task at hand. In this system, Convolutional Neural Networks and Artificial Neural Networks are employed.

4) Train CNN & ANN Gesture Images: We're utilising deep learning techniques like CNN and ANN to train the gesture photos.

5) Sign Language Recognition from Webcam: Vision-based and non-vision-based approaches are the two types of gesture recognition techniques. Vision-based techniques employ a web camera and markers to identify signs. Vision-based technique is used for the gesture recognition in this system.

6) Extract image from Webcam: Input is taken through a webcam in this system, and then segments the palm and fingers to identify and recognize motions.

7) Convert image to binary or grey format and background removal: It's critical to separate items from their surroundings and convert them to binary images. In pattern recognition, binary images are utilised as inputs to the feature extraction process and play a crucial part in the development of unique features that may be used to discriminate between different classes.

8) Extract features from image: Parts or patterns of an item in a picture that assist in identifying it are known as "features." Feature extraction is a step in the dimensionality reduction process, which involves dividing and condensing a large amount of raw data into smaller groupings.

9) Recognition and play audio: Finally, the hand motions that have been identified are transformed into voice and text.

B. System architecture

Fig.1 illustrates the preferred method's system architecture. It denotes the system's whole physical installation. A camera records the user's input. After taking a frame of a picture obtained with the webcam, the image is preprocessed and the background noise is eliminated. The

image has now been converted to binary format, which allows features to be retrieved and identified. The obtained features are compared with the data set. The gesture with the highest matching rate is taken into account and transformed into text and voice.

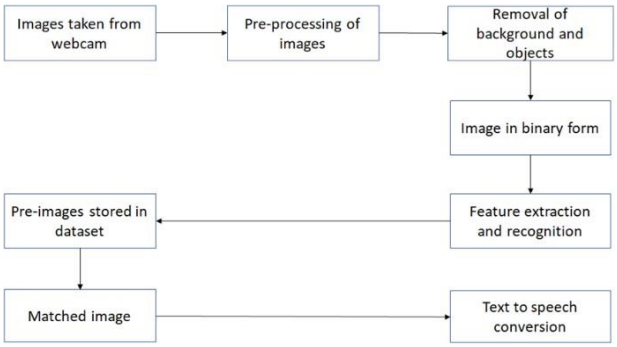

Fig. 1. System Architecture

C. Algorithm

Step-1: Start.
Step-2: Execute the python file in the terminal.
Step-3: Click on Hand Gesture Dataset button.
Step-4: Select the path for the Dataset and upload the dataset.
Step-5: Pre-process the dataset.
Step-6: Click on Train CNN & ANN Gesture Images button.
Step-7: Click on Hand Gesture Recognition from Webcam button.
Step-8: Image Extraction from Webcam.
Step-9: Convert image into Binary or gray format.
Step-10: Background Removal of image.
Step-11: Feature extraction from image.
Step-12: Result is displayed on the screen in text.
Step-13: Text is converted into audio.
Step-14: Goto step-8 until the window is closed.
Step-15: End.

III. Implementation and Testing

A. Implementation

Implementation of proposed approach consists of following 4 steps:

1) *Pre-process the Hand Gesture Dataset:*
In the first step, a function was written that saved the path to the hand gestures dataset in a variable and loaded all image-containing folders into arrays. To remove the noise, the mean filter method is utilized. In this simple sliding window, the average of the window's pixel values replaces the center value. For segmentation, the threshold method is used. It divides pixels in an image by comparing each pixel's intensity to a preset value. A contour is just a line that links all sections of the same color or intensity.

2) *Modeling and Training Gesture Images:*
Algorithms like CNN and ANN are employed to model the gesture images. The basic methods employed by CNN are: convolution, maximum pooling, and flattening. Convolution is the technique of altering an image by applying a kernel over the entire picture to each pixel and its local neighbors. Max Pooling is a convolution method in which the kernel convolutions the region with the greatest value. The process of converting data into a one-dimensional array for usage in the following layer is known as flattening.

3) *Webcam Image Extraction and Feature detection:*
At this stage, picture frames are read from the camera and image processing algorithms are applied to them, after which characteristics are retrieved from the image frames and motions are detected by comparing them to those in the dataset.

4) *Voice and text translation of hand gestures:*
Both text and speech versions of the predicted gesture are produced. The text is shown on the screen and also fed into Google Voice to provide an audio output.

B. Results

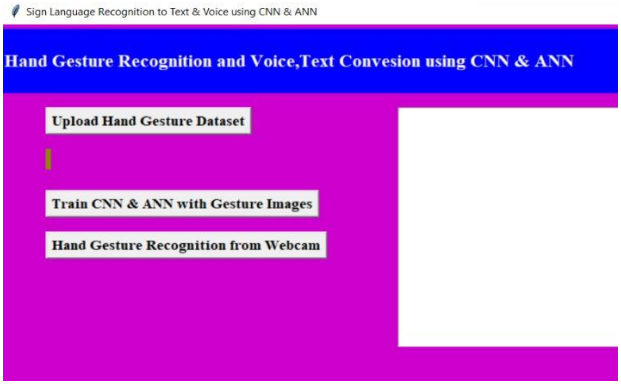

Fig. 2. User Interface

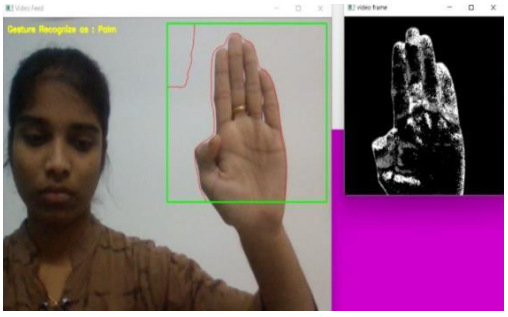

Fig. 3. Palm Gesture

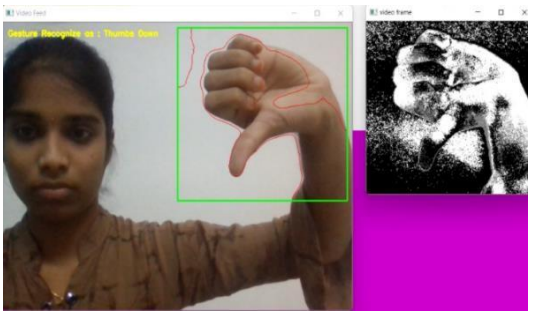

Fig. 4. Thumbs Down Gesture

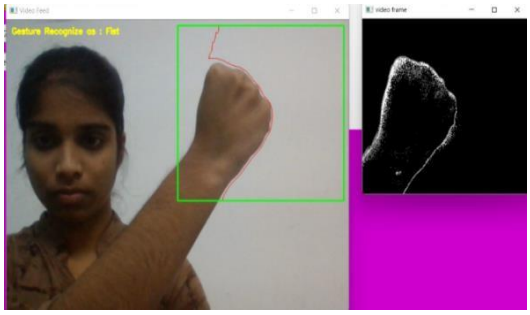

Fig. 5. Fist Gesture

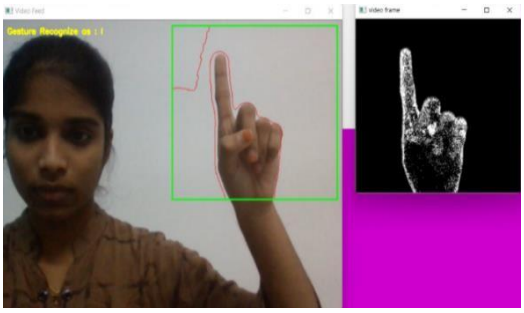

Fig. 6. I Gesture

The image is captured from the web camera. Following the capture of an image, the image is identified and classified using the CNN and ANN algorithms. The gesture is matched to the dataset, and the gesture with the highest matching rate is categorized. The static texts, as well as the audio format, are then presented on the screen. The ANN algorithm is 86% accurate, whereas CNN is 98% accurate.

IV. Conclusion and Future Scope

A. Conclusion

The CNN model is utilized for hand gesture recognition since its accuracy is greater than that of ANN. The practical adaptation of the interface solution for visually impaired and blind people is constrained by its simplicity and applicability in real-world circumstances. Hand gestures to speech and text translation have been employed in this approach to allow the reduction of hardware components as a simple and practical technique to establish human-computer interaction. Overall, the approach tries to help people in need while maintaining societal significance. The system's user-friendly design assures that it can be used by anybody without difficulty or complication. The application is low-cost and does not require the use of pricey technologies. The communication gap between blind and deaf persons is bridged using this technology.

B. Future Scope

To boost user engagement and make the system more robust, the app may be linked with other mobile and IoT devices. The goal of developing the solution as a commercially viable product for users is to aid the global community of visually impaired people.

References

[1] Raj, Jennifer S., and Mr C. Vijesh Joe. "Wi-Fi Network Profiling and QoS Assessment for Real Time Video Streaming." IRO Journal on Sustainable Wireless Systems 3, no. 1 (2021): 21-30.

[2] Tripathi, Milan. "Analysis of Convolutional Neural Network based Image Classification Techniques." Journal of Innovative Image Processing (JIIP) 3, no. 02 (2021): 100-117.

[3] Hamdan, Yasir Babiker. "Construction of Statistical SVM based Recognition Model for Handwritten Character Recognition." Journal of Information Technology 3, no. 02 (2021): 92-107.

[4] J. Rajalakshmi, P. Kumar. "Hand Gesture Recognition using CNN and RNN", International Journal of Recent Technology and Engineering (IJRTE), 2020.

[5] Vivekanandam, B. "Evaluation of Activity Monitoring Algorithm based on Smart Approaches." Journal of Electronics 2, no. 03 (2020): 175-181.

[6] Manoharan, Samuel, and Narain Ponraj. "Analysis of Complex Non-Linear Environment Exploration in Speech Recognition by Hybrid Learning Technique." Journal of Innovative Image Processing (JIIP) 2, no. 04 (2020): 202-209.

[7] Syed Raquib, Shareef, Mohammed, Mannan Hussain, Akash Gupta, Hakeem Aejaz Aslam."Hand Gesture Recognition System for Deaf and Dumb", International Journal of Multidisciplinary and Current Educational Research (IJMCER) 2020.

[8] S. Vigneshwaran, M. Shifa Fathima, V. Vijay Sagar, R. Sree Arshika. "Hand Gesture Recognition and Voice Conversion System for Dumb People", 5th International Conference on Advanced Computing & Communication Systems (ICACCS), 2019.

[9] Mr.P.V.Krishna Rao, V.Niharika, V.Prashanthi, V.Akhila, V.Gayathri. "Hand Gesture Recognition and Voice Conversion System for Dumb and Deaf People".Journal of Emerging Technologies and Innovative Research (JETIR) April 2019.

[10] Prema Sharma, Naman Sharma." Gesture Recognition System", 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU), 2019.

[11] V.Swaminathan, B.Sakthivel, Rohan.K Vishwanath. "Hand Gesture Based Recognition and Voice Conversion System for Physically Disabled People".International Journal of Engineering and Science Vol.9, Issue 8 (August 2019).

[12] Rupesh Prajapati, Vedant Pandey, Nupur Jamindar, Neeraj Yadav, Prof. Neelam Phadnis. "Hand Gesture Recognition and Voice Conversion for Deaf and Dumb", International Research Journal of Engineering and Technology (IRJET), 2018.

[13] Shinde, Shweta S., Rajesh M. Autee, and Vitthal K. Bhosale. "Real time two way communication approaches for hearing impaired and dumb person based on image processing." Computational Intelligence and Computing Research (ICCIC), 2016 IEEE International Conference on. IEEE, 2016.

[14] Sood, Anchal, and Anju Mishra. "AAWAAZ: A communication system for deaf and dumb." Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2016 5th International Conference on. IEEE, 2016.

[15] Rishabh Agarwal, Nikita Gupta. "Real Time Hand Gesture Recognition for Human Computer Interaction", IEEE 6th International Conference on Advanced Computing (IACC), 2016.