IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SYSTEM FOR CONVERSION OF HAND GESTURES TO SPEECH AND TEXT

[1]Neha S Bharadwaj, [2]Smitha S M, [3]Prema K N, [4] Roopa B S

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

ECE Department,
JNNCE, Shivamogga,India

Abstract: Communication is the act of passing ideas from one group to another using mutually recognized signs, symbols and semiotic rules. It is the only medium through which people can exchange their thoughts or convey messages. Deaf and mute people have their own set of signs. Hand gestures are a form of non-verbal communication used by deaf and mute people to communicate with each other and with the outside world. However, the problem with the current world is that most people are not knowledgeable enough to interpret hand gestures or translate them into a language that normal people can understand. Thus, to bridge the communication gap between deaf-mute and normal people, it is important to develop a system that can help them translate gestures into text and/or speech. Creating a robust communication system for the deaf community will help them become more independent and confident. Robust hand gesture recognition has played a significant role in the field of human-computer interaction for a long time, but it is still full of challenges due to many acceptances such as blurred background and hand self-occlusion. With the help of depth information, depth-based methods perform better, but depth cameras are not as widely used and affordable as color cameras. Therefore, we propose two-stage deep convolutional neural network (CNN) architecture for accurate color-based hand gesture recognition. The first phase performs the generation of pseudo-depth hand images from color images, and the second phase recognizes hand gesture classes using the color image and its pseudo-depth hand image. The architecture of the generation phase is based on a picture-to-picture translation network. In the recognition stage, a two-stream CNN architecture with a color image and its pseudo-depth image is proposed to improve the performance of color image-based recognition. Experiments show that our approach significantly improves hand gesture recognition performance in RGB only. The proposed system is user-friendly as it is easy to use and able to create effective and efficient human-computer interaction. A gestural type of communication called sign language that is observed among deaf groups around the world. We designed a deaf-mute verbal exchange device that translates hand gestures into sound massage as an interpreter.

Key word **– CNN**

I. INTRODUCTION

Hand gestures are an aspect of body language that can be conveyed by the center of the palm, the position of the finger and the shape made by the hand. Hand gestures offer an inspiring field of research because they can facilitate communication and provide a natural means of interaction that can be used in a variety of applications. Previously, hand gesture recognition was achieved by wearable sensors attached directly to the gloved hand. These sensors detected a physical response based on hand movements or finger bends. The collected data was then processed using a computer connected to the glove via a wire. Although the above techniques have given good results, they have various limitations that make them unsuitable for the elderly, who may experience discomfort and confusion due to wire connection issues. In addition, elderly people suffering from chronic medical conditions that lead to loss of muscle function may not be able to put on and take off gloves, causing them discomfort and limitations if they are used for long periods of time. These sensors can also cause skin damage, infection, or adverse reactions in people with sensitive skin or those with burns. In addition, some sensors are quite expensive. These drawbacks led to the development of promising and cost-effective techniques that did not require the wearing of cumbersome gloves. These techniques are called camera vision based sensor technologies. With the evolution of open source software libraries, it is easier than ever to detect hand gestures that can be used in a wide range of applications, such as sign language for clinical operations, virtual environments for controlling robots, home automation for personal computers, and gaming on tablets. These techniques essentially involve replacing an instrumented glove with a camera. In this work, image processing technique is used to detect and recognize hand gesture in real time.

II. THEORITICAL BACKGROUND

Extracting image features to model a visual appearance, such as a hand, and comparing these parameters to features extracted from input image frames. 3D model-based recognition, depends on 3D kinematics and a model that has a large degree of freedom, is a new model for understanding the interactions between 3D hands and an object using a single RGB image, where one image is trained end-to-end using a neural network and display joint estimation of hand and object position in 3D[1].

[2] explains different types of hand recognition system. In the sensor-based approach, two-way communication has been provided using a data glove, which is interpreted by a microcontroller and converted into a message and audio output via a Bluetooth speaker. This lacks a quick response in case of misinterpretation, making real-time use difficult. In the vision-based approach, a red-green-blue (RGB) input image is taken and skin segmentation is performed. The region of interest is extracted using morphological operations.

An electronic system to help mute people exchange their ideas with a normal person in emergency situations. The system consists of a glove that the subject can wear that converts hand gestures into speech and text. The displayed message will also help the deaf to understand their thoughts[6.]

[3] uses edge detection and the database contains many contained features that are adapted to predict the moment of the hand. It is easy to use and a less expensive method for accurate sign language identification is established. The expectation of achieving the desired result is 95 percent, which is used on a large scale for its intended purpose.

A convolutional neural network (CNN) is a type of artificial neural network that uses perceptron learning rules along with supervised learning to analyze data. CNN is used for image processing, natural language processing and other kinds of cognitive tasks. Like other kinds of artificial neural networks, a convolutional neural network contains an input layer, an output layer, and various hidden layers. Some of these layers are convolutional and use a mathematical model to pass results to subsequent layers. This simulates some of the actions in the human visual cortex. CNN is a basic example of a deep learning algorithm [3].

III. DESIGN AND IMPLEMENTATION

The first step of the proposed system is data collection. Many researchers use sensors or cameras to capture hand movements. For the implementation, we used a web camera to capture hand gestures. Images go through a series of pre-processing operations where backgrounds are detected and eliminated. Segmentation is then performed to find out the gesture region. Using morphological operations, a mask is applied to the images and a series of dilations and erosions are performed using elliptical kernels. With the resume open, the captured images are scaled to the same size, so there is no difference between images of different gestures. Binary pixels are extracted from each image frame and a convolutional neural network is used for training and classification. The model is then evaluated and the system would then be able to predict alphabets (Sentences).

The block diagram depicts how this proposed system works and methodology. It is implemented in three modules, that is, image input module, pre-processing and classification module. Taking pictures from webcam is the input module, detecting and tracking, segmenting, masking and extracting features come in pre-processing module. The rest come under classification.

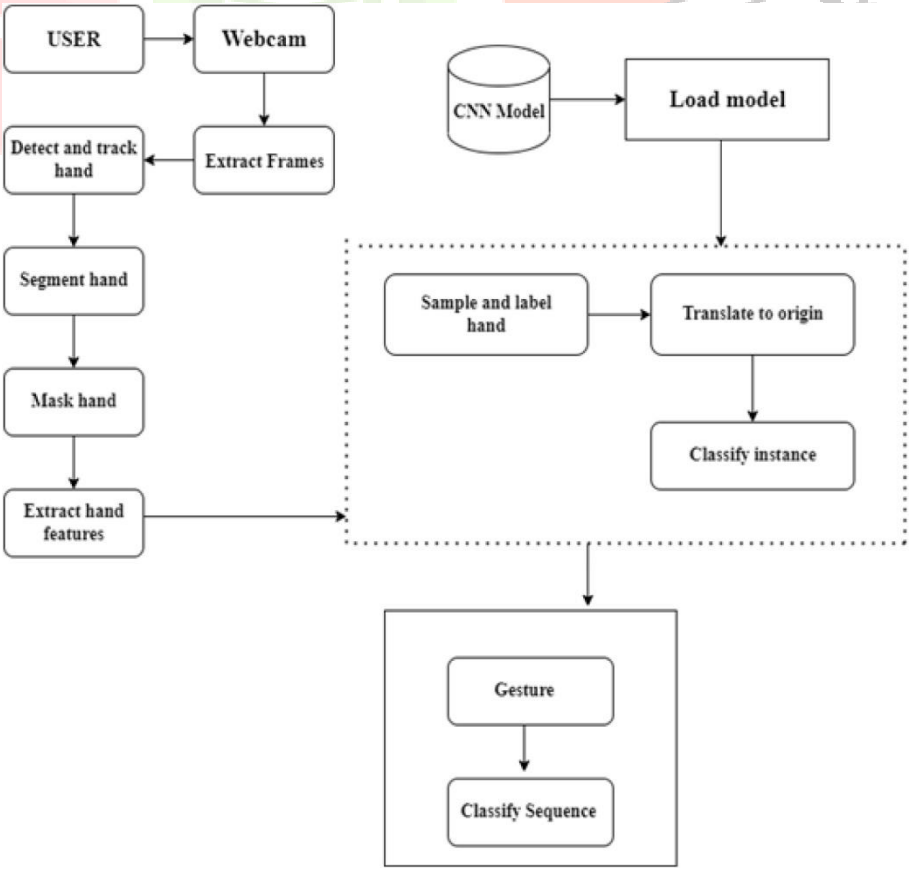

Figure 3.1: Block diagram of the proposed system

From the user, webcam will capture the image, we can also use Raspberry Pi camera. By default it has 24 fps(frame per second). In one second it will capture 24 images consecutively. From the output it has to recognize our hand. For that we have to set histogram, by looking at the histogram it can identify the skin tone of our hand. 50 photos of each gesture is taken with webcam in the plain background and same lighting. After capturing images, image pre-processing takes place. Image pre-processing involves: Cropping each image in order to make the size of all the images uniform. Applying various filters for the images, all the photos/dataset is used to train the CNN model. CNN model is trained using Keras library. Later, histogram of the hand is set, so that the system correctly recognizes the hand colour. Now, in the detect and track hand section it will only track the skin tone part. Then it will segment only hand part, it will crop the required part and processes further. In the masking hand, it will make the segmented part as white color so processing will be faster. Then it will extract only the shape of hand, it has to find the gesture according to shape then it will be compared with trained data set.

For one gesture we will put 1200 images and each set will be having label for eg. One gesture with "I need water", another gesture with "Sorry", so on, here we can replace sentence by letter "A", "B" etc. So comparison will be easier. If gesture matches with the trained data set it will recognize the particular gesture, otherwise it won't detect. So in this process we are using CNN model (convolution neural network), CNN takes the image's raw pixel data, trains the model, then extracts the features automatically for better classification.

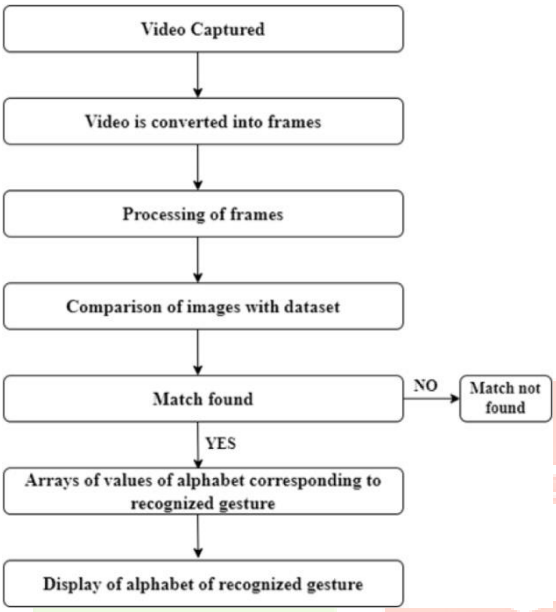

Figure 3.2 : Flow chart

First when we show any gesture, a video is captured. A video is a sequence of frames which is nothing but images. Hence that video is converted into frames since it would be easier to detect gesture form an image. Then the processing of frames will be done like converting to pixels etc, and then that image will be compared with the pretrained dataset. If the match is found, it will give the corresponding output, otherwise it will show 'Match not found'.

IV. RESULTS AND DISCUSSION

Dataset of different gestures was taken using phone camera. Later the captured images are pre-processed (cropping). That means all the images are resized and made uniform. Then when we test the system, we will get corresponding output according to the pre- trained dataset. Here all the gestures would have respected results in both text and voice. This result will be shown on the laptop screen and it can also be heard via speakers. Below are some examples of the trained gestures taken from phone.

1200 photos of each gesture is captured. Then all the captured photos are stored in a separate folder with respective id and names. All the id and names are stored in the database. In this method, we got noise in camera while contour formation. Since the result is based on contour formation, the result that we got was not that much accurate. We would get more accurate results if we use high end camera. Later we used raw images for training and testing the system.50 photos of each gesture were taken in white background and moderate lighting, also without shadow for training.

Figure 4.1 : Trained Gestures

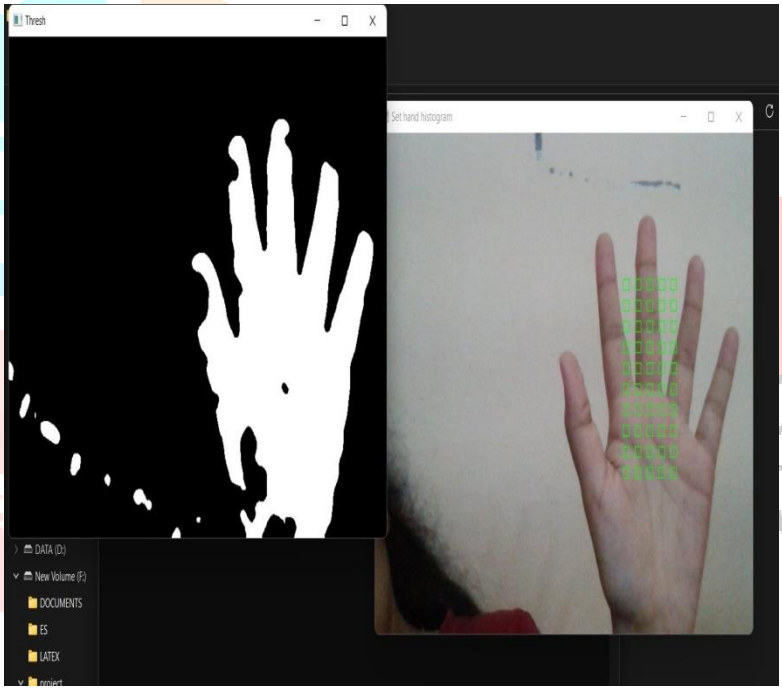

Figure 4.2 : Noise while contour formation

The Gestures corresponding to the sentences are indicating below:

Gesture 1: HOW ARE YOU?
Gesture 2: I WANT WATER?
Gesture 3: I WANT GO OUT?
Gesture 4: SORRY
Gesture 5: THANK YOU
Gesture 6: EMERGENCY
Gesture 7: I AM HAPPY
Gesture 8: I AM HUNGRY
Gesture 9: HELLO

Figure 4.3 : Sample of Predicted Gesture

IV. CONCLUSION

A sign language recognition system is feasible for mute people because they can communicate with themselves or others through this system. The system can capture hand gestures and navigate words as in text format and also provides voice output, it will be used for mute people to see these words and understand sentences. The implementation uses Python and Open CV. Vision-based gesture recognition requires gesture recognition in a noise-free environment, which is difficult to set up every time communication needs to be made. They are usually expensive and require high computing power to recognize gestures in real time. The accuracy of the system can be increased by using neural networks. Open CV functions and defocusing techniques can be used to reduce noise. Controlling things by hand is more natural, easier, more flexible and cheaper, and there is no need to deal with the problems caused by hardware devices because no hardware components are needed. The proposed system is more efficient when we use high-end cameras for calculations.

REFERENCES

[1] J. P. Singh, A. Gupta et al., "Scientific exploration of hand gesture recognition to text," in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE, 2020, pp. 363–367.
[2] K. Manikandan, A. Patidar, P. Walia, and A. B. Roy, "Hand gesture detection and conver- sion to speech and text," arXiv preprint arXiv:1811.11997, 2018.
[3] O. M. Foong, T. J. Low, and S. Wibowo, "Hand gesture recognition: sign to voice system (s2v)," International Journal of Computer and Information Engineering, vol. 2, no. 6, pp. 1794–1798, 2008.
[4] D. Xu, "A neural network approach for hand gesture recognition in virtual reality driv- ing training system of spg," in 18th International Conference on Pattern Recognition (ICPR'06), vol. 3. IEEE, 2006, pp. 519–522.
[5] Q. Chen, N. D. Georganas, and E. M. Petriu, "Real-time vision-based hand gesture recog- nition using haar-like features," in 2007 IEEE instrumentation & measurement technology conference IMTC 2007. IEEE, 2007, pp. 1–6.
[6] D. K. Sarji, "Handtalk: Assistive technology for the deaf," Computer, vol. 41, no. 7, pp.84–86, 2008.
[7] M. Maebatake, I. Suzuki, M. Nishida, Y. Horiuchi, and S. Kuroiwa, "Sign language recog- nition based on position and movement using multi-stream hmm," in 2008 Second Inter- national Symposium on Universal Communication. IEEE, 2008, pp. 478–481.
[8] J. Liu, K. Furusawa, T. Tateyama, Y. Iwamoto, and Y.-W. Chen, "An improved hand gesture recognition with two-stage convolution neural networks using a hand color image and its pseudo-depth image," in 2019 IEEE international conference on image processing (ICIP). IEEE, 2019, pp. 375–379.