11 - Final Project Report

Hemanth Kothapalli
es19btech11003@iith.ac.in

Mogatala Umesh Kumar Reddy
es19btech11005@iith.ac.in

Naveen
cs19bbtech11009@iith.ac.in

Nisha M
cs19btech11012@iith.ac.in

Harshitha Sagiraju
cs19btech11032@iith.ac.in

Abstract

Speech impairment is a disability which affects an individuals ability to communicate using speech and hearing. People who are affected by this use other media of communication such as sign language. Hand gestures are the most ubiquitous and important ways of communication in modern society. They can assist in the development of secure and comfortable user interfaces for a variety of applications. For hand gesture detection, many computer vision algorithms have used color and depth cameras, but effective classification of motions from different subjects remains a challenge. here we present a convolutional neural network-based approach for real-time hand gesture recognition (CNNs).

1. Introduction

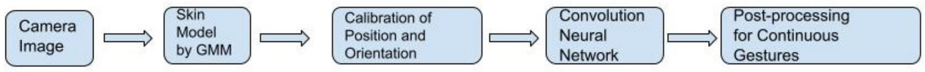
Hand gestures are one of the most popular ways to communicate, and they carry a lot of power. The major goal here is to improve people's quality of life by allowing them to do a larger range of daily duties more efficiently also For example, hand gesture is observed and recognized by surveillance cam- eras to prevent criminal behavior.Hand gesture recognition, in particular, has been acknowledged as a beneficial technique for a variety of applications, particularly in the field of Sign Language Recognition (SLR). Complex hand gestures are used in sign languages, and even little hand movements can have a range of meanings. As a result, various vision-based dynamic hand motion detection algorithms have been established in the previous decade.
Recently, classification with deep convolutional neural networks has been successful in various recognition challenges .Multi-column deep CNNs that employ multiple parallel networks have been shown to improve recognition rates of single networks by 30-80
This work is a CNN-based human hand gesture recognition system. CNN is a research branch of neural networks.Using a CNN to learn human gestures, there is no need to develop complicated algorithms to extract image features and learn them. Through the convolution and sub-sampling layers of a CNN, invariant features are allowed with little dislocation. To reduce the effect of various hand poses of a hand gesture type on the recognition accuracies, the principalaxis of the hand is found to calibrate the image in this work

2. Literature Review

2.1. Human hand gesture recognition using a convolution neural network

For a hand gesture recognition system because there are so many variables in image space and the size of an image space is so large, it's critical to extract information from it. To make it perform better Typically, numerous gestures are required, and these should be modelled.since the light condition seriously affects the skin color, here we adopt a Gaussian Mixture model (GMM) to train the skin model which is used to robustly filter out non-skin colors of an imagedesign a system for recognising human gestures based on a image.The skin colour is represented by a Convolution Neural Network (CNN).

Figure 1. Framework of the proposed human gesture recognition system.


Camera Image
Skin Model by GMM
Calibration of Position and Orientation
Convolution Neural Network
Post-processing for Continuous Gestures

Skin model for color segmentation

In GMM training, we divide the workspace into 15 sections. For every section, we sample the color within the yellow area around the center of the hand.

- The light situation has a significant impact on the performance of a visual system. In traditional methodologies, colour thresholds are commonly used to classify skin and non-skin., but color thresholds are not enough to describe the statistical properties of skin color under various light conditions. Oftentimes some pixels on the image are classified as non-skin pixels. In this paper, Gaussian Mixture mode(GMM) is used to solve this problem. A GMM is represented by K Gaussian components as

  $$P(x) = \sum_{k=1}^{K} P(k)P(x|k) \tag{1}$$

  where P (k): probability
  P(x| k) : conditional probability formulated as a Gaussian distribution

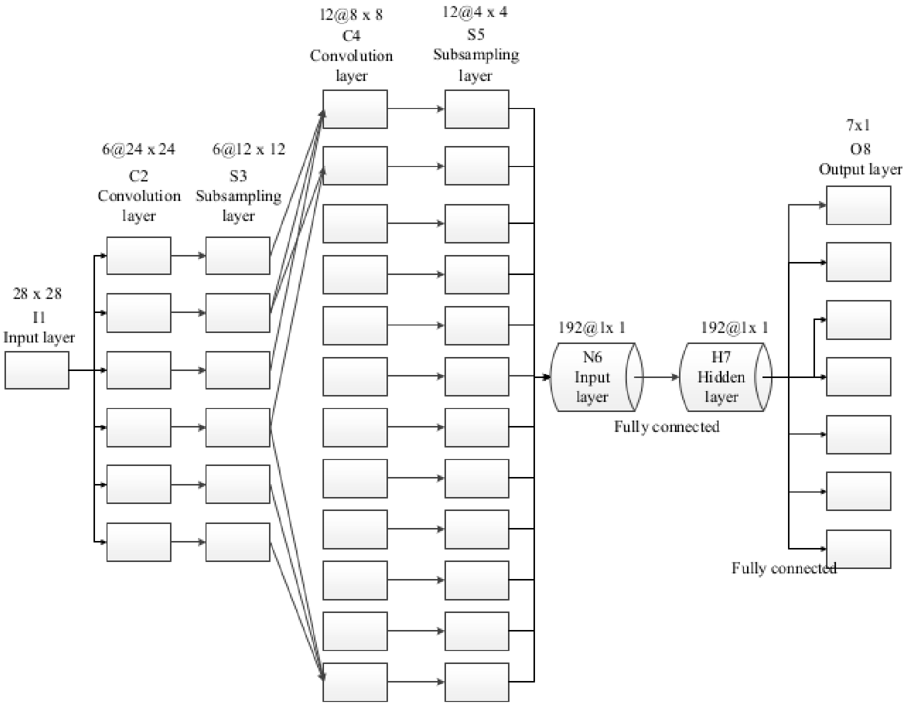
- In GMM training, we divide the workspace into 15 sections. For every section, we sample the color within the yellow area around the center of the hand.

- After Skin model for color segmentation it is followed by
  Calibration of hand position and orientation.
  here we derive the binarized image and find its center using mathematical formula. Also The angle of orientation is expressed by another formula.

- Next we have
  Gesture recognition by a convolution neural network:
  Comparatively for complicated image feature extraction, the CNN provides a robust and systematic methodology to classify the type of hand gesture. CNN Architecture here we have 8 layer of architecture.

- The best part of the system is that there is no need to build a model for every gesture using hand features such as fingertips and contours. To have robust performance, we applied a GMM to learn the skin model and segment the hand area for recognition.

Figure 2. Architecture of CNN



2.2. Human hand gesture recognition using a convolution neural network

- Hand gesture recognition being one of the valuable technologies for Sign Language Recognition has helped perform day to day activities more efficiently. Looking at the user's muscular activity is a natural method to build interfaces. Deep convolutional neural networks have recently proved effective in a variety of recognition difficulties .

- Multi-column deep CNNs that use several parallel networks have been demonstrated to enhance recognition rates of single networks by 30-80% for a variety of picture classification applications. The best results for large scale video classification were observed by merging CNNs trained on two distinct streams of original and spatially cropped video frames.The other techniques like data augmentation were limited to spatial variations.

- This paper introduces a hand gesture recognition system that detects hand components from images then learns and predicts them using 2D convolutional neural networks An efficient spatio-temporal data augmentation strategy for deforming the input volumes of hand gestures is provided to avoid possible overfitting and increase generalization of the gesture classifier. Existing spatial augmentation techniques are also included into the augmentation procedure.

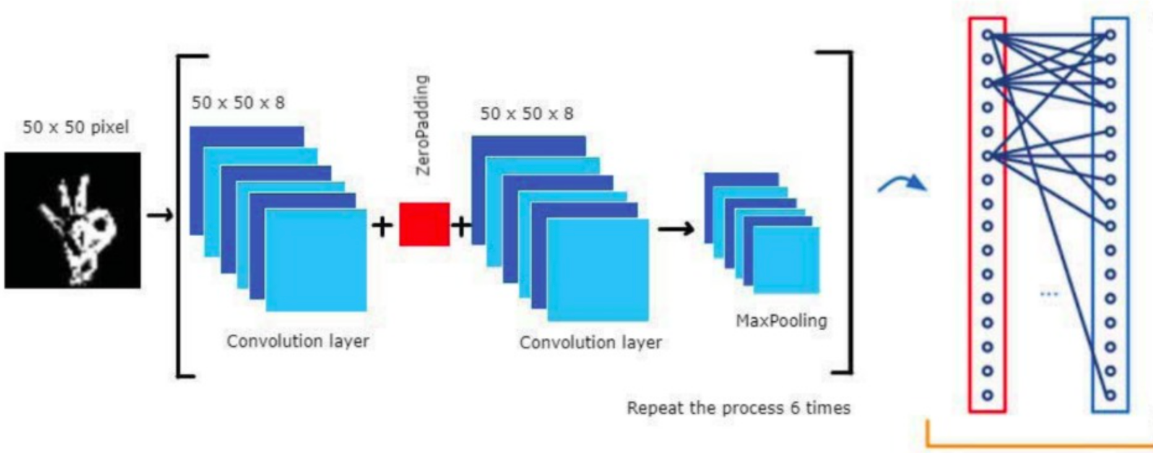
- The dataset used for training was additionally added with 4000 extra images after applying spatio - temporal data augmentation techniques (horizontal mirroring of the images to generate a new set of data) . Coming to the classifier, the network consists of six 2D convolution layers, each of which is followed by a max-pooling operator.

- The output of the sixth convolution layer is given as input to a fully connected network having 9 layers.A sigmoid activation function is used in the output layer. Tanh activation function is used in the remaining eight layers. Additionally Batch Normalization is applied (before the non linearity) to address the overfitting issue. The training phase of CNN entails optimizing network parameters in order to minimize a cost function for the dataset. As the cost function, mean squared error was used. Optimization was performed via stochastic gradient descent and the NN parameters were updated with the Nesterov accelerated gradient at every iteration .

- BN was shown to yield faster training times whilst achieving better system accuracy and regularization. The classifier showed an accuracy of 98.74% on the test set.

- The proposed classifier utilizes spatio-temporal data augmentation to avoid overfitting.The combination of low and high resolution sub-networks improves classification accuracy considerably. It is further demonstrated that the proposed data augmentation technique plays an important role in achieving superior performance.

Figure 3. Architecture of CNN



2.3. Sign Language Recognition System using Convolutional Neural Network and Computer Vision

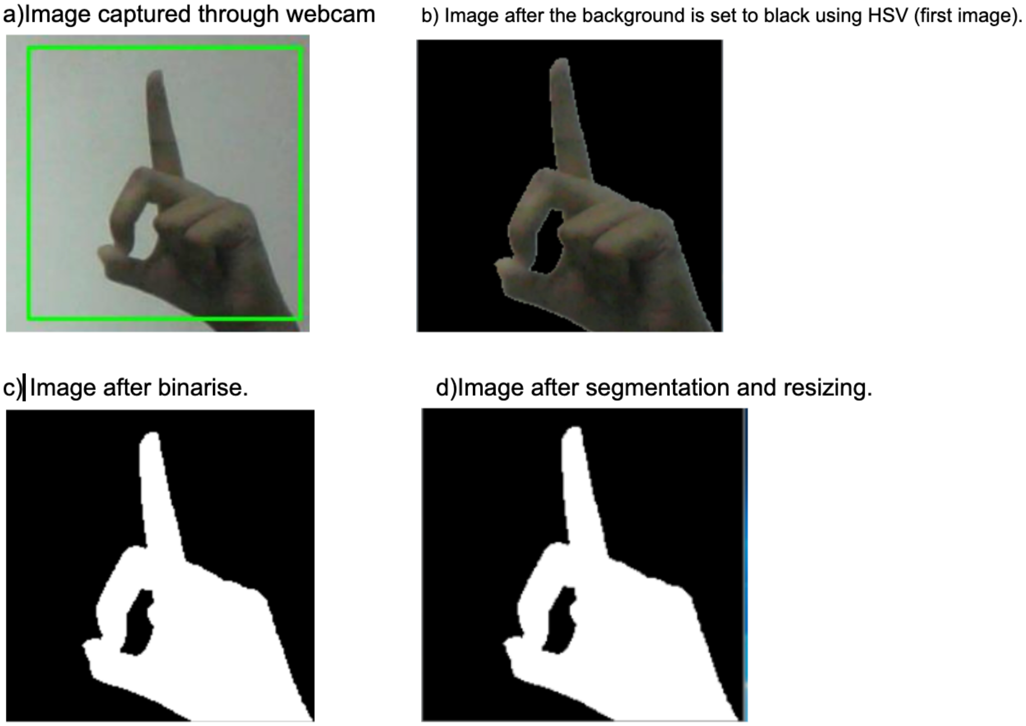
- In this paper, the images are captured through webcam. HSV colour algorithm is used to detect the hand gesture and set the background to black. Since the images obtained are in RGB colourspaces, it becomes more difficult to segment the hand gesture based on the skin colour only. So, therefore transform the images in HSV colourspace. It is a model which splits the colour of an image into 3 separate parts namely: Hue,Saturation and value. HSV is a powerful tool to improve stability of the images by setting apart brightness from the chromaticity . The Hue element is unaffected by any kind of illumination, shadows and shadings and can thus be considered for background removal.

- The images undergo a series of processing steps which include various Computer vision techniques such as the conversion to grayscale, dilation and mask operation. And the region of interest which, in our case is the hand gesture is segmented. The features extracted are the binary pixels of the images.

- The hand gesture is segmented firstly by taking out all the joined components in the image and secondly by letting only the part which is immensely connected, in our case is the hand gesture. The frame is resized to a size of 64 by 64 pixel. At the end of the segmentation process, binary images of size 64 by 64 are obtained where the area in white represents the hand gesture, and the black coloured area is the rest.

- A CNN model is used to extract features from the frames and to predict hand gestures. It is a multilayered feedforward neural network mostly used in image recognition. The architecture of CNN consists of some convolution layers, each comprising of a pooling layer, activation function, and batch normalisation which is optional. It also has a set of fully connected layers. As one of the images moves across the network, it gets reduced in size. This happens as a result of max pooling. The last layer gives us the prediction of the class probabilities.

Figure 4



a)Image captured through webcam

b) Image after the background is set to black using HSV (first image).

c)Image after binarise.

d)Image after segmentation and resizing.

3. Results

In our gesture recognition system we have included a total of ten american sign language hand gestures (A,B,C,D,F,G,K,O,P,Y).

We have implement Hand Gesture Recognition Using Background Elimination and CNN. Our first approach to create a gesture recognition system was through the method of background Elimination. Background Elimination, as the name suggests, is the process of separating foreground objects from the background in a sequence of video frames.We binarize this absolute difference image using a threshold. We trained the CNN model with 10*1000 binarized train images and validated with 10*100 binarized test.

Figure 5

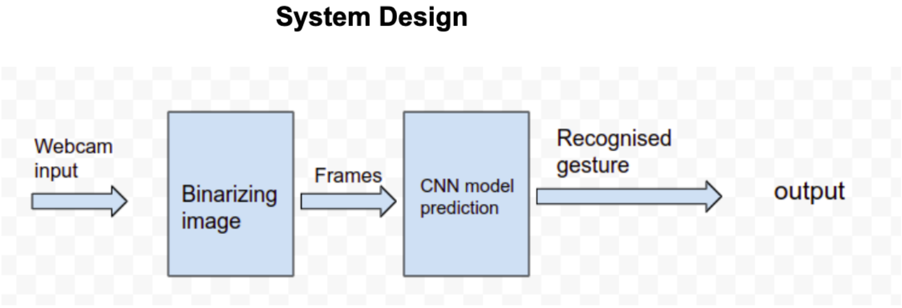
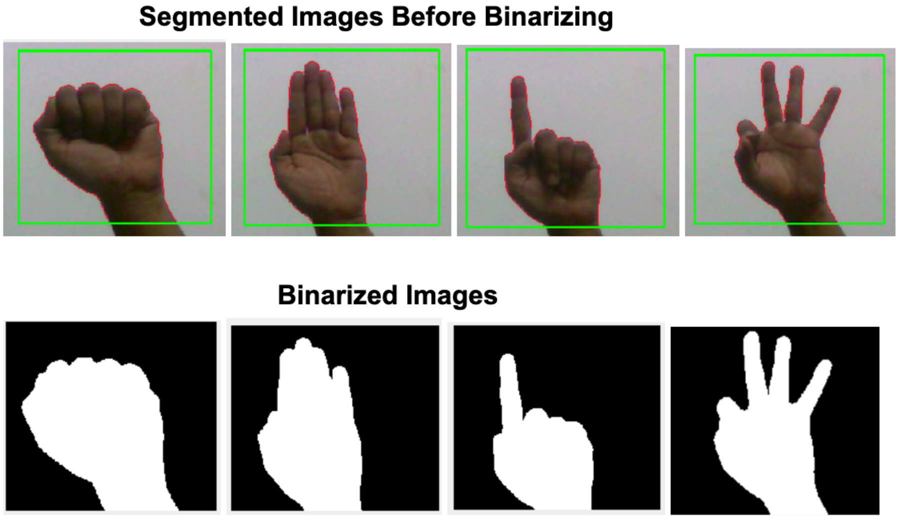


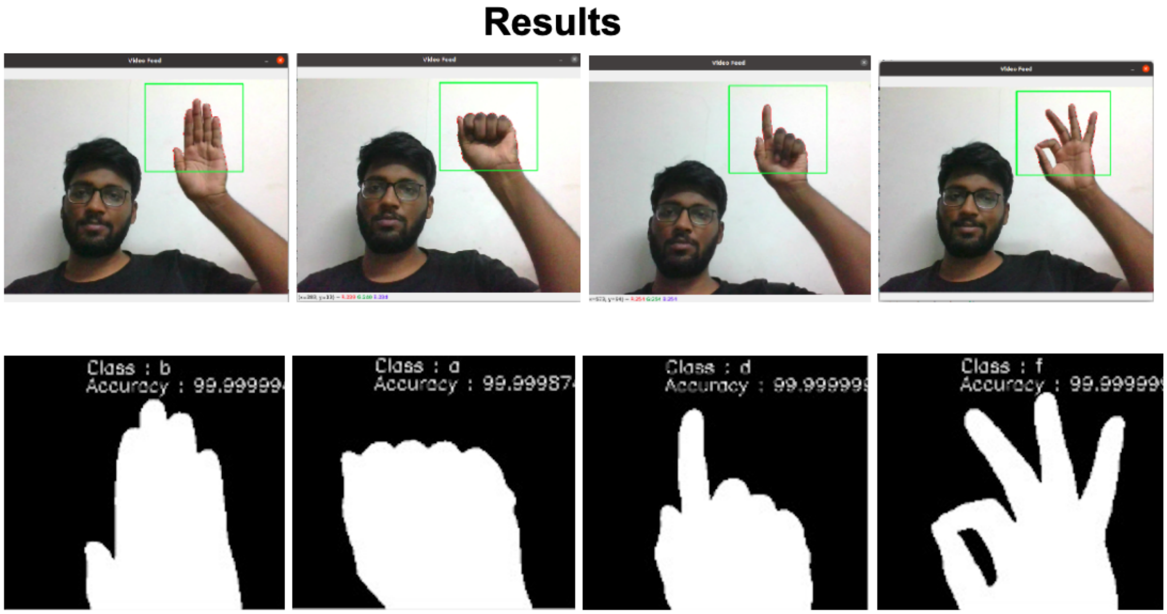
Figure 6



We faced various limitations and accuracy concerns when constructing the recognition system utilizing background subtraction. Background subtraction is unable to cope with abrupt, severe illumination changes, resulting in a slew of discrepancies. When used against a plain background, the gesture detection system proved to be reliable and accurate.The items in the background revealed to be irregularities in the image capture process in circumstances when the background was not plain, resulting in inaccurate outputs.After evaluating the outcomes of the gesture recognition system in various environments, it is advised that this system be utilized with a plain background to achieve the best potential results and high accuracy.

We build using a CNN model which trained on train data set of all the available training binarised hand gestures present and are later being tested on the test dataset and helps in finding accuracy of our developed CNN model. Our CNN model consists of 7 layers where the kernel size increases from 32 to 256 in multiples of 2 and again comes back to 64(so total 7) . We'll use the Adam Optimizer for our CNN model initializing the learning rate equals 0..001. Considering these parameters our model is trained on the respective train images.

Figure 7



4. Conclusion and future work:

Conclusion and future work: The proposed method's performance is significantly dependent on the results of hand detection. If there are moving objects with a colour that is similar to that of the skin, the objects are detected as a result of the hand detection and reduce the hand gesture recognition performance. Machine learning algorithms, on the other hand, can distinguish the hand from the backdrop. ToF cameras provide depth information that can help enhance hand detection performance. To address the complicated background problem and improve the robustness of hand identification, machine learning algorithms and ToF cameras may be applied in future research. We were successful in developing a reliable gesture recognition system. We would like to increase the accuracy of this gesture recognition system in the future by implementing the latest model training algorithms, as well as launch programmes and open certain popular websites, and add new gestures to implement more functions which would inturn help the people in need .We also want to expand our domain scenarios and integrate our tracking mechanism into a range of gear, such as digital television and mobile phones. We also want to make this process accessible to a wider group of users, including disabled people.

References

[1] felix zhan, "Hand gesture recognition with convolution neural networks," 2019.

[2] a. W.-k. C. Hesin-I Lin, MIng-Hsiang Hsu, "Human hand gesture recognition with convolution neural networks," 2014.

[3] M. H. Mohammad Elham Walizad, "Sign language recognition system using convolutional neural network and computer visions," 2020.