Design & Development of Hands Sign and Gesture Recognition

-Gaurav Kumar
-Anuj Kumar
-Rachit Sharma
(School of Computer Science and engineering)

Abstract—
The hand gesture is one of the non-verbal communication methods used in sign language. It is most used by people with speech or hearing impairments to communicate with non-disabled people and with each other. Numerous manufacturers worldwide have developed diverse sign language systems; however, they lack adaptability and affordability for end users. Thus, to help deaf and dumb people communicate with others more effectively, the "Hand sign and gesture recognition system software" proposed in this proposal provides a system prototype that can automatically understand sign language. In sites like "YouTube" videos where there is currently no feature for automatic text generation because of gestures, this approach can also be employed, likewise sign languages. Research on gesture recognition is still in its early stages. Hand gestures are vital to everyday life and play asignificant role in nonverbal communication. The software's goal is todemonstrate a real-time system for hand gesture and sign recognition by detecting certain shape-based features, such asorientation, the centroid of the Centre of Mass, the status of the fingers, and the thumb in positions where the hand's fingers are raised or folded. However, Convolutional Neural Networks (CNNs) will handle the feature extraction process entirely. Every frame of the video will be captured in this process, and each frame will be used to locate hands and clip them out so that our CNN may use them as input. We used ISL as a case study for our purposes. The back projection histogram approach was employed in this model to set the image's histogram.

We used CNNs for training and testing, and as a result, our test accuracy was 99.89%. Our model's independence from external hardware or devices is one of its benefits.

Keywords— Convolutional Neural Network (CNN)

I. INTRODUCTION

Our lives are growing more and more dependent on computers and other technological gadgets. As the market for these computing devices grew, so did the need for user-friendly and straightforward computer interfaces. Given its numerous potential applications in human-machine interaction, gesture recognition is becoming more and more popular in the research community because of the growing use of systems that rely on vision-based interface and control. Any vision-based interface is more comfortable, practical, and natural than a mouse and keyboard because gestures are so intuitive. Gesture recognition can be done mainly in three ways: with wearable gloves, with 3D hand key point placements, and with raw visual data.

The first method requires wearing a separate device with multiple cords, but it produces strong results in terms of speed and precision. The second, however, requires a manual extraction of important points, which is an extra step. To process the video stream, a sliding window technique with a stride of one is employed. The top graph shows the likelihood ratings for the detector, which is activated when a gesture starts and stays on until it ends. The second graph shows the categorization score for each class using a different colour. By using a weighted average, the third graph lessens the ambiguity between possible gesture candidates.

Filtering to raw classification results. On the bottom graph, which shows single-time activations, red arrows denote early detections and black ones, detections made after gestures had finished, respectively. additional processing time and expense. Finally, just an image capturing sensor—such as a camera, an infrared sensor, or a depth sensor—that is not user dependent is needed for (iii). This solution stands out as the most practical one since the user does not need to wear a cumbersome gadget to acquire an acceptable level of recognition accuracy and an adequate rate of computation. Any system for gesture recognition needs to have a workable infrastructure. After all, we want to apply knowledge to actual situations. We have created a vision-based gesture identification method in this work employing deep convolutional neural networks (CNNs) on raw video data to offer a workable solution. Presently, CNNs deliver the most cutting-edge results for tasks that use both images and videos, including gesture recognition, activity localization, and image based tasks like object detection, segmentation, and classification. There are numerous criteria that the system must meet in real-time gesture recognition applications: A single activation for each executed gesture, along with:

(i) An adequate classification accuracy,
(ii) Quick reaction time,
(iii) Resource efficiency, and
(iv) Single activation.

For a real-time vision-based gesture recognition program to be successful, each of these components is of vital importance. However, most of the earlier research ignores the remaining items and focuses solely on improving offline classification accuracy in gesture recognition. Due to the several deep CNNs that some proposed systems use on various input modalities and their inability to function in real-time, they are pushing the memory and power budgets to their limits. Convolutional Neural Networks (CNNs) have gained widespread popularity in recent years due to their impressive performance in a variety of computer vision tasks such as object recognition, image classification, and segmentation. CNNs have revolutionized the field of computer vision by enabling automated image analysis and interpretation with remarkable accuracy.

CNNs are a type of deep neural network that can learn hierarchical representations of images through a series of convolutional and pooling layers. These networks have been shown to be highly effective at learning features that are invariant to translation, rotation, and scaling, making them ideal for tasks such as hand and sign gesture detection.

One common approach to hand and sign gesture detection using CNNs involves training the network on a large dataset of

labeled images. The dataset typically includes a wide variety of hand and sign gestures, captured from different angles and under different lighting conditions. CNN is trained to learn discriminative features from these images, and to classify them into different categories based on the gesture being performed.

One of the key challenges in hand and sign gesture detection is dealing with the variability in hand shapes and orientations. To address this challenge, researchers have developed a range of CNN architectures that are designed to be robust to variations in hand pose and appearance. For example, some architectures incorporate recurrent neural networks (RNNs) to capture the temporal dynamics of hand movements, while others use attention mechanisms to focus on specific parts of the hand.

Another important consideration in hand and sign gesture detection is real-time performance. CNNs can be computationally intensive and may require significant processing power to run in real-time on a mobile or embedded device. To address this issue, researchers have explored various techniques such as model compression, pruning, and quantization to reduce the computational requirements of CNNs.

II. LITERATURE SURVEY

To improve communication between deaf communities and others, gesture-based sign language recognition systems are crucial. Convolutional Neural Network (CNN) experiments have been done to recognize gestures after some pre-processing of input data from input devices. Yet, in those experiments, the complexity and diversity of hand gestures had a significant impact on the accuracy and identification rates. In their research work, the authors (Md Abdur Rahim, Jungpil Shin, and Md Rashedul Islam) present a successful solution to this issue: hand gesture detection using CNN with improved data pre-processing, such as feature fusion CNN, RGB colour input to YCbCr, binarization, erosion, and hole filling. Instead of concentrating solely on data preprocessing as other recent research papers at that time did, the authors of this study (Md. Rashedul Islam, Rasel Ahmed Bhuiyan, Ummey Kulsum Mitu, and Jungpil Shin) came up with a novel way to categorise them using Multi-class Support Vector Machine (MCSVM), which increases productivity and efficiency as the number of specimens or categories rises (types of gestures). After subdividing the multiclass problem into smaller problems, all of which are binary classification problems, the general idea of SVM is used in MCSVM. As demonstrated in the research report, this strategy improves accuracy, however data pre-processing is also employed to achieve high accuracy. They removed background images from the submitted photographs, leaving only those that showed the Area of Interest (ROI), in this case, the hands. Nevertheless, because they utilized a very ineffective technique to eliminate background images—taking two photos, one with ROI and the other without—the speed of prediction was significantly reduced, costing the model an approximate average accuracy of 95%. Afterwards, background removal, filtering, noise reduction, and grayscale conversion were used. [1].

Patel & et al. developed a static hand gesture identification method for American Sign Language using deep convolutional neural network. The system architecture weighs little, making it easy to deploy and transfer the system around. to attain high accuracy in real-time conditions, a variety of image processing methods are used to assist with

background reduction and frame segmentation. The approach emphasizes mobility, straightforward deployment at no cost, and no computational overhead. The approaches used in this are feature extraction, the Hand Segmentation Approach (HSA), and glove-based hand motion detection. During image processing and frame segmentation, the model is implemented using a Gaussian Mixture-based background segmentation technique (FS). The two main types of noise that were found were salt and pepper noise associated with illumination and spatial noise associated with motion. The spatial noise in the subtracted image was removed using low pass spatial filtering with a kernel size of 3, and the other forms of noise were removed using morphological opening with a structuring element of size 5. The image must then be converted to grayscale as the last stage. This eliminates any prejudice brought on by the user's skin tone or foreground lighting during recognition. They use picture recognition with the help of a convolutional neural network (CNN). In recent research papers about gesture recognition systems using CNN and Deep Learning, there were only a small number of papers that focused on Region of Interest (ROI) segmentation from image, which made the accuracy lower and especially for hand signs and gestures that have nearlymore than 3000 different signs and to classify them accurately on the basis of different features that have little variation compared to others, so the image segmentation is a very crucial point that was overlooked in many of the studies. They used the Histogram Back-Projection technique tosegment images and create datasets toimprove the quality of the training materials.The classes created from these datasets are subsequently put into a CNN.

I. METHODOLOGY

Our project deals with the OpenCV python module to record the live footage from a device camera and it does all processing without any artificial devices for this project to run. This is what we call a video-based project, thus increasing scalability.

Raw video footage will act as a input for our software input. Raw video footage is now broken down to frames.

Then these frames will be then sent to Haar_Cascade model to filter out the Region of Interest (ROl), in our case it is hands. Then these ROI are now cropped out offrames and sent to CNN model which will classify these images.

In the traditional system a different model is used to extract region of interest, but in our system, we implemented the use of Haar-Cascade algorithm to extract it.

The cascaded image is then sent to CNN model where the given image is classified.

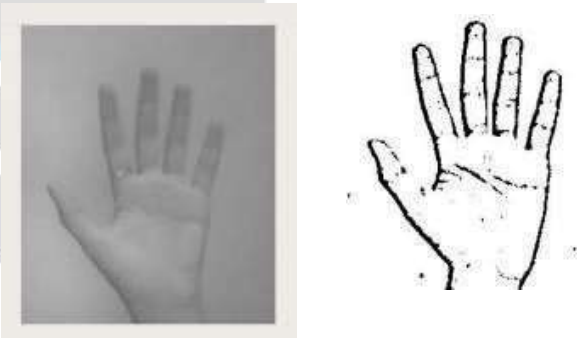
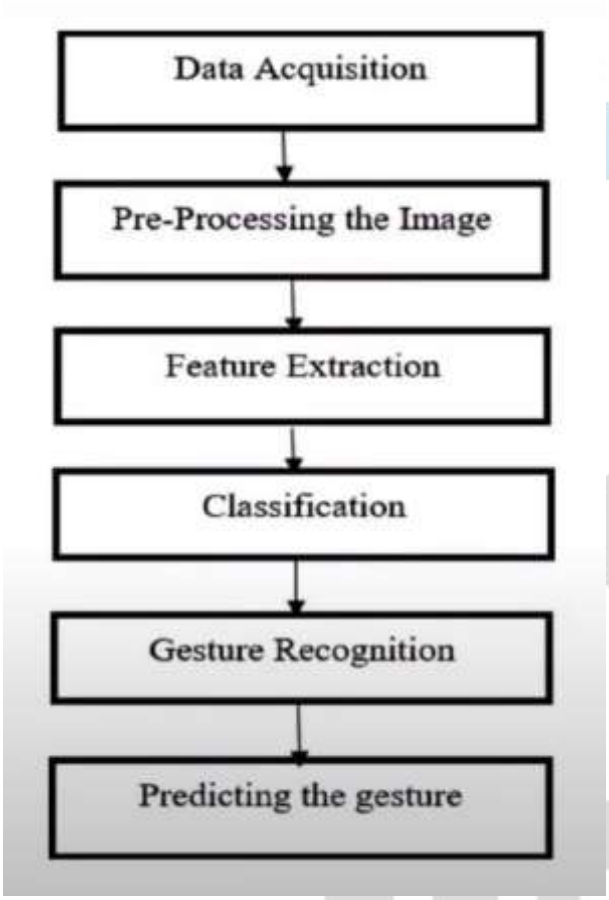


Figure1: This is image pre-processing for easier recognition and feature outlining. a) actual image b) Region of Interest

FLOW DIAGRAM



1. Data Acquisition

Data acquisition is the process of sampling signals that measure real - world physical phenomena and converting them into a digital form that can be manipulated by a computer and software.

2. Pre-Processing the Image

A preliminary processing of data in order to prepare it for the primary processing or for further analysis. The term can be applied to any first or preparatory processing stage when there are several steps required to prepare data for the user.

3. Feature Extraction

Feature extraction plays a crucial role in sign language recognition, enabling accurate interpretation of manual gestures and movements used in sign languages.

4. Classification

Classification involves the identification and categorization of different sign language gestures.

5. Gesture Recognition

Gesture recognition facilitates effective communication between individuals who are deaf or hard of hearing and the hearing community.

6. Predicting the gestures

Predicting gestures in sign language recognition is a challenging task that has garnered significant attention in recent years.

Sign language is a visual-spatial language used by individuals with hearing impairments to communicate.

Steps:

- Importing necessary packages for project.

```
import cv2
import numpy as np
import mediapipe as mp
import tensorflow as tf
from tensorflow.keras.models import load_model
```

- Initialize Mediapipe class object to importthe Haar-Cascade for detecting Region of Interest.

```
# initialize mediapipe
mpHands = mp.solutions.hands
hands = mpHands.Hands(max_num_hands=1,
                      min_detection_confidence=0.7)
mpDraw = mp.solutions.drawing_utils
```

- Loading the gesture model along with the last layer nodes label for recognition and initializing the web cam.

```
# Load the gesture recognizer model
model = load_model('mp_hand_gesture')

# Load class names
f = open('gesture.names', 'r')
classNames = f.read().split('\n')
f.close()
print(classNames)

# Initialize the webcam
cap = cv2.VideoCapture(0)
```

```
# Read each frame from the webcam
_, frame = cap.read()
x, y, c = frame.shape

# Flip the frame vertically
frame = cv2.flip(frame, 1)
framergb = cv2.cvtColor(frame, cv2.COLOR_BGR2RGB)

# Get hand landmark prediction
result = hands.process(framergb)

# print(result)
className = ''
```

```
if result.multi_hand_landmarks:
    landmarks = []
    for handslms in result.multi_hand_landmarks:
        for lm in handslms.landmark:
            # print(id, lm)
            lmx = int(lm.x * x)
            lmy = int(lm.y * y)
            landmarks.append([lmx, lmy])
```

- Now drawing landmark on the screen along with prediction and displaying the text

```
        prediction = model.predict([landmarks])
        # print(prediction)
        classID = np.argmax(prediction)
        className = classNames[classID]
# show the prediction on the frame
cv2.putText(frame, className, (10, 50),
            cv2.FONT_HERSHEY_SIMPLEX,
                1, (0,0,255), 2, cv2.LINE_AA)
# Show the final output
cv2.imshow("Output", frame)
if cv2.waitKey(1) == ord('q'):
    break
```

II. CONCLUSION AND FUTURESCOPE

The best method for segmenting images using the histogram ROI technique is this one, but it has a very important drawback: it cannot be used in dynamic scenarios like live videos or videos. On the plus side, though, we can now classify differentgestures based on thresholds that are used inresearch papers to maximize variations in various hand gesture features.

However, there is a drawback: although it will slow down the process and work inmost applications, we can use computer vision to locate the hand and then place the histogram there to address dynamic situations. This problem can be solved with the aid of high-performance computation.

The Goal of this project is to develop a system in which hand gestures can be translated into text.

Future scope:

This system can also be used in platforms like “YouTube”, “Netflix” etc., videos where there is currently no feature for auto- text generation on the basic of gestures and sign languages.

• Even in video conference we can embed our system of transformation for better communication.

• Can also be used in places like smart devices to control them via gestures rather than voice (for dumb people)

•In other words, this proposed “Hand sign and gesture recognition system” can be used for both public welfare and commercial use.

REFERENCE :-

1. Dynamic Hand Gesture Based Sign Word Recognition Using Convolutional Neural Network with Feature Fusion
2. https://ieeexplore.ieee.org/document/8777621
3. A Static Hand Gesture Based Sign Language Recognition System using Convolutional Neural Networks
4. Hand Gesture Feature Extraction Using Deep Convolutional Neural Network for Recognizing American Sign Language
5. https://ieeexplore.ieee.org/document/8858563
6. https://ieeexplore.ieee.org/document/9057853
7. https://paperswithcode.com/paper/fast-and-robustdynamic- hand-gesture
8. https://paperswithcode.com/paper/real-time- handgesture-detection-an
9. https://paperswithcode.com/paper/deep-learning- forhand-gesture-recognition-on