Deep learning - based Hand Gesture Recognition System to control VLC media player

Diya
191IT215
diya.191it215@nitk.edu.in

Lakshmi Raj
191IT225
lakshmisraj.191it225@nitk.edu.in

Sanjeevi Meghana
. 191IT144
sanjeevimeghana.191it144@nitk.edu.in

Patchipulusu Sindhu
191IT137
sindhu.191it137@nitk.edu.in

Abstract:

Hand gestures are a form of nonverbal communication that can be used in several fieldssuch as communication between deaf-mute people, robot control, human–computer interaction (HCI), home automation and medical applications. For the Recognition of hand gestures, model can be fed with static as well as dynamic datasets but In this paper, we propose a Hand Gesture Recognition System with which a video can be controlled using static hand gesture, captured with the help of webcam.This paper mainly focuses on the Recognition of Static Hand Gesture and connecting it with an action in the VLC player like play,pause,mute,unmuteetc.Therefore, we need large-scale real-world datasets to build a responsive,accurate system, that can recognize complex gestures with subtle differences between them.Multiple frames sampled from the video sequence are given as an input to the classification network.To extract the representative frames, wepropose to use deep learning framework, that is the convolutional neural network.The creation of the model includes the deep learning concepts along with open-cv, keras and tensorflow modules.
A significant and improvedaverage recognition rate of 97.10% is achieved through inherentfeature learning capability of CNN.

Introduction:

Hand gestures are an aspect of body language that can be conveyed through the center of the palm, the finger position and the shape constructed by the hand. Hand gestures can be classified into static and dynamic. As the name implies, the static gesture refers to the stable shape of the hand, whereas the dynamic gesture comprises a series of hand movements such as waving. Hand gestures offer an inspiring field of research because they can facilitate communication and provide a natural means of interaction that can be used across a variety of applications like surveillance, robotics, gaming technologies, virtual reality, augmented reality, sign language, medical environment, human-computer interaction etc.In this paper we present an application which uses different computer vision techniques for recognizing hand gestures for controlling the VLC media player. The aim and objectives of this

application is to use a natural device free interface, which recognizes the hand gestures as commands. With the massive influx and advancement of technologies, a computer system has become a very powerful machine which has been designed to make the human beings' tasks easier. Due to which the HCI (human computer interaction)has become an important part of our lives. In current world many facilities are available for providing input to any application some needs physical touch and some without using physical touch (speech, hand gesture etc.).But not many applications are available which are controlled using current and smart facility of providing input which is by hand gesture .By this method user can handle application from distance without using keyboard and mouse. This application provides a novel human computer interface by which a user can control media player (VLC) using hand gestures.

A major challenge evolves is the complexity and robustness linked with the analysis and evaluation for recognition of gestures.Algorithms have been developed based on computer vision methods to detect hands using different types of cameras. The algorithms attempt to segment and detect hand features such as skin color, appearance, motion, skeleton, depth, 3D model, deep learning detection and more. Previously, hand gesture recognition was achieved with wearable sensors attached directly to the hand with gloves. But in this we are using the webcam to detect the hand gesture.The idea of the project is to create a model to recognize Hand Gesture and accordingly connect these gestures with the respective actions in the vlc player. This project is implemented using deep learning framework i.e CNN. Below is the order in which the project functions:

- **The application uses a webcam which is used for image acquisition and hand is put in the specified region of interest.**
- **Then we extract the contour and segment the hand region. Finally, we recognize the gesture.**
- **Recognized gesture is connected to the respective action in the VLC player like play, pause etc.**

Rest of the paper organization is as follows: the following section provides a literature review of hand gesture recognition systems. Section 3 covers the problem statement. Section 4 shows the methodology designed for the application. Analysis and application results are highlighted in section 5. Conclusion in section 6 with the future work is discussed.

Literature Survey:

The power of deep learning tool for the recognition of hand postures from raw images (RGB) has been utilized in [4]. The proposed CNN architecture eliminates the need of detection and segmentation of hands from the captured images, thus reducing the computational burden faced during hand posture recognition with classical approaches. The performance of the proposed method has been evaluated on two publicly available datasets through five fold cross validation.

[5] analyzed the effect of data augmentation in deep learning framework that involves convolutional neural networks .Hand gestures can be recognized successfully but some extension is still possible.More gestures can be added to the list of recognition. It was assumed that the background should

be less complex. Therefore, recognition of gestures in complex backgrounds can be another extension. Recognition of gestures made with both hands is not possible by this system.

The proposed system in [1] used webcam in order to track handThen used skin color technique and morphology to remove the background. In addition, kernel correlation filters (KCF) used to track ROI. Then the resulted image enters into a deep convolutional neural network (CNN) where the CNN model is used to compare performance of two modified from Alex Net and VGG Net. The recognition rate for training data and testing data, respectively 99.90% and 95.61% .

Table 1

SUMMARY OF LITERATURE SURVEY

Authors	Methodology	Merits	Limitations	Dataset Used
[1]	skin color detection and morphology & background subtraction	Recognition of hand gestures in case of complex backgrounds.	Recognition rate for training dataset is 99.90% whereas for test dataset its only 95.61%.	4800 image collected for train and 300 for test
[2]	No segment stage, Image direct fed to CNN after resizing	Hand gesture recognition system works in real time and gives a result with simple backgrounds as 97.1%	Accuracy of recognition of hand gestures decreases to 86% in case of complex backgrounds.	Mantecón et al.* dataset for direct testing
[3]	skin color -Y–Cb–Cr color space & Gaussian Mixture model	Recognizing human gestures with various hand positions and orientations and light conditions.	Average recognition rate is 95.96% which could further be increased.	image information collected by Kinect
[4]	automatic recognition of hand postures using convolutional neural networks with deep parallel architectures.	Accuracy in terms of recognition of hand gestures in case of even complex backgrounds is too high.(99.6%)	This is applicable only for static hand gestures.	NUS hand posture dataset and American fingerspelling A dataset)
[5]	skin color detection & background subtraction	The loss percentage of model is less.	Recognition of gestures made with both hands is not possible by this system .	8000 images and tested on 1600 images

Problem Statement:

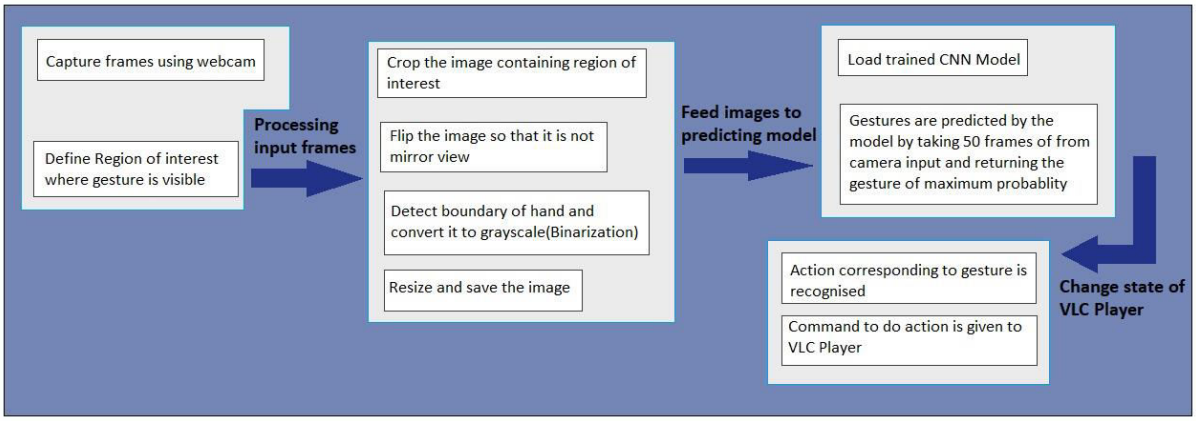
Controlling VLC Media player using a vision based hand gestureinterface for Human-Computer Interaction(HCI) solved using CNN.

Objectives:

- Detecting the hand of a person placed in a rectangular box i.e Region of Interest(ROI).
- Building a deep learning model for theRecognition of Hand Gestures.
- Evaluating the model by selecting the required action from the benchmark dataset.
- Controlling VLC media player using hand gestures.

METHODOLOGY:

ARCHITECTURE DIAGRAM :



A) TRAINING AND TESTING THE MODEL:

The dataset is divided into an appropriate ratio for training and testing respectively. A sequential model is created and fed with datasets for training the model. The model is then tested on the remaining dataset

1. EXTRACTING THE DATASET

The datasets can be downloaded from the below link - https://www.kaggle.com/gti-upm/leapgestrecog
The dataset images are processed before creating the model.The model uses images of size (128,128), so the images are resized. The resized images are then converted from RGB to grayscale. OpenCv was used for processing the images. Then after resizing, the dataset is then split into training and test dataset in the ratio 7:3. To increase the accuracy, dataset images are suffled before feeding it to the network. The model is then trained on this prepared dataset.

2) CREATING AND TRAINING THE MODEL

A sequential model is created using Deep Convolutional Neural Network.

Passing the model through a bunch of filters namely **conv2D** and **MaxPooling2D** is a part of creating the model**.**

The Keras Conv2D parameter filter determines the number of kernels to convolve with the input volume. Each of these operations produces a 2D activation mapwhereas, Maxpooling2D is then used to reduce the spatial dimensions of the output volume. Dropout is a technique where randomly selected neurons are ignored during training. They are “dropped-out” randomly. This means that their contribution to the activation of downstream neurons is temporarily removed on the forward pass and any weight updates are not applied to the neuron on the backward pass.

Model configuration is done for training purposes, this uses **adam** optimizer.We used **loss** of 'categorical_crossentropy' type.Using 'categorical_crossentropy' type the model will output a *probability*that an example belongs to any of the categories.

Now,the model can be trained by passing the training datasetsand for validation, test datasets are passed.

3) TESTING THE MODEL

The loss and accuracy on test data can be calculated with the help of test datasets. We make predictions on the test data and compare the predicted value with the actual label value, ideally both should be the same.

We used a function to plot images and labels based on predictions and their actual values for validation purposes and then we created a **Confusion Matrix** for evaluation.It is used to measure the performance of a classification model.

B) CONTINUOUS GESTURE RECOGNITION

Initially average weight is set and the reference from the webcam is taken.

Region of interest (ROI) is set by giving all the four coordinates.Whatever is visible in ROI is our point of interest. As per our project, hand gestures are to be placed in that region.

The camera starts recording after entering ‘s’ on the keyboard. Once the recording starts, the continuous frames are extracted from the video being recorded. Resizing and flipping of frame is done in order to get the actual view of frame.

Now, ROI frame is extracted from the captured image and this frame is converted from RGB to grayscale.

Once the threshold is reached, the background is obtained so that running average model gets calibrated. Here in our case the threshold is set to a fixed number. After achievingthreshold, the prediction starts. Fixed number of frames for same gesture are fed to the trained model andfor every frame of input the probability for each gesture is found out and the gesture that has maximum probability is considered as the final predicted gesture.

CONTROLLING VLC PLAYER

We then use the gesture predicted from the model to control respective action of the VLC player like play, pause and so on.
After creating an object to control the state of the VLC player, we give in the media file to be played

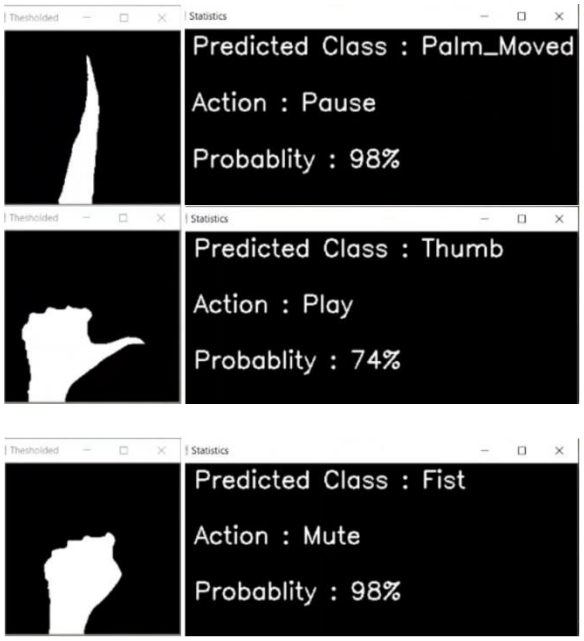
The actions controlled by gestures are -

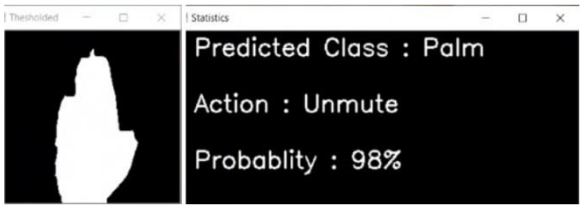
1. Play
2. Pause
3. Mute
4. Unmute
5. Volume up
6. Volume down
7. Forward
8. Rewind

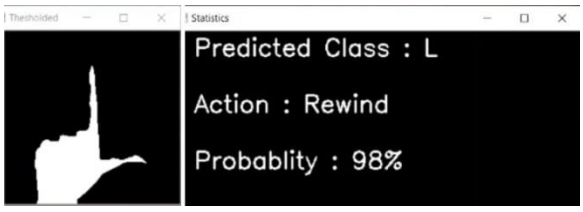
.

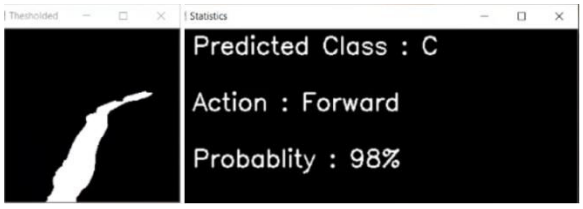
RESULTS:

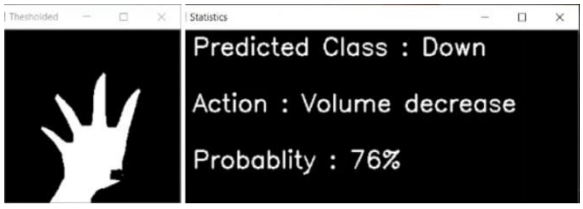
The following are the snapshots which contain results of the recognition of different hand gestures along with the threshold part and the respective actions performed by gestures.

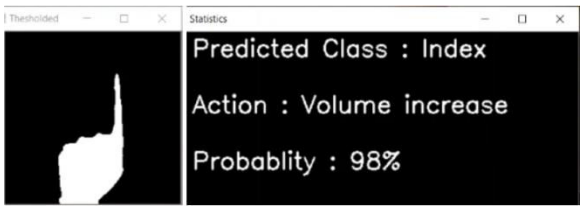




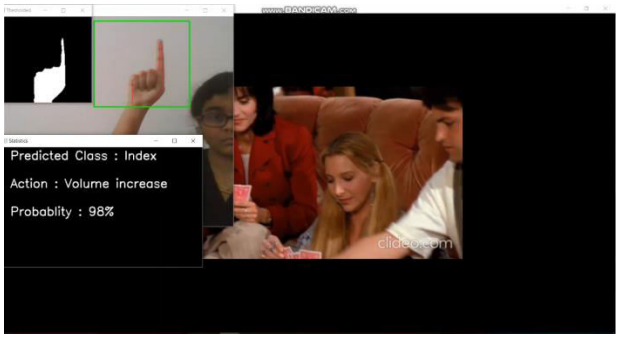




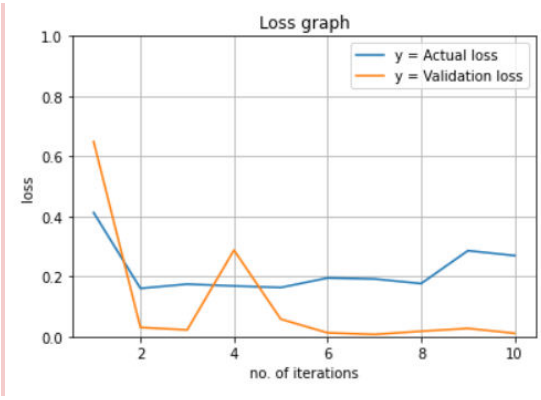


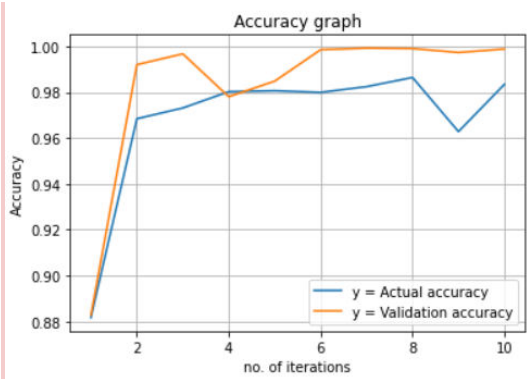


Below snapshot depicts the working of VLC player. In this, the hand is put in the bounding box and the action performed by gesture is reflected through the video playing in background.



ANALYSIS:





Above are the graphs representing accuracy and loss for hand gesture recognition system for 10 epochs. The overall accuracy is high (i.e 99.88%) which represents the effectiveness of the model proposed.

Confusion matrix displayed below presents the results of recognition for different hand gestures. The values of confusion matrix represent that the recognition rate for some hand gestures is 100%, while for some it is near to 98%.

	Predicted Palm	Predicted L	Predicted Fist	Predicted Thumb	Predicted Index	Predicted Palm Moved	Predicted C	Predicted Down
Actual Palm	626	0	0	0	0	0	0	0
Actual L	0	587	0	0	0	0	1	0
Actual Fist	0	0	617	0	0	0	0	0
Actual Thumb	0	0	0	617	0	0	0	0
Actual Index	0	0	0	0	596	0	0	0
Actual Palm Moved	0	0	0	0	0	580	0	0
Actual C	0	0	0	0	0	0	625	0
Actual Down	0	0	0	3	0	0	0	549

Likert analysis:

A Likert analysis was conducted through google form which included the video that showed the working of our project and a group of 18 people were asked the different questionnaires regarding the project and their responses(rated out of 5) are clubbed and recorded in the table below:

	5	**4**	**3**	**2**	**1**
Effectiveness of the actions performed	11	6	1	0	0
Time taken for a action to be performed	8	5	5	0	0
Comfortness while using hand gestures	14	3	1	0	0

Conclusion:

In this paper, a deep learning hand gesture recognition algorithm is proposed for controlling the functions of vlc media player like mute, unmute, play, pause, volume-increase, volume-decrease, moving-forward and moving-backward .Based on the results presented in the previous section, we can conclude that our algorithm successfully classifies different hand gestures images with an accuracy of 97.10% based on a Deep Learning model.

The convolutional neural network is utilized to perform the hand gesture recognition. The classification accuracy and computational efficiency of the convolutional neural network is improved by extracting 50 representative frames from the video sequence. The recognized hand gesture is connected with the actions in the VLC player to control the video. Our proposed algorithm is evaluated on the Hand Gesture Recognition Dataset available on kaggle. The accuracy of our model is directly influenced by a few aspects of our problem. The gestures presented are reasonably distinct, the images are clear and without background. Future work will look at the variety of complex backgrounds and recognition of gestures made with both hands.

Contribution:

Although we all worked together but the main contribution of individuals in creating the project are:

Patchipulusu Sindhu:

Main contribution: In finding the datasets and extraction of hand from the video captured through webcam.

Diya:

Main contribution: Creating the model and training dataset.

Sanjeevi Meghana:

Main contribution: Continuous Hand Gesture Prediction

Lakshmi S Raj:

Main contribution: Connecting the hand gestures to the VLC media player.

References:

[1] .Chung, H.; Chung, Y.; Tsai, W. An efficient hand gesture recognition system based on deep CNN. In Proceedings of the 2019 IEEE International Conference on Industrial Technology (ICIT), Melbourne, Australia, 13–15 February 2019; pp. 853–858.

[2]. Bao, P.; Maqueda, A.I.; del-Blanco, C.R.; García, N. Tiny hand gesture recognition without localization via a deep convolutional network. IEEE Trans. Consum. Electron. 2017, 63, 251–257. [CrossRef].

[3]. Lin, H.-I.; Hsu, M.-H.; Chen, W.-K. Human hand gesture recognition using a convolution neural network. In Proceedings of the 2014 IEEE International Conference on Automation Science and Engineering (CASE), Taipei, Taiwan, 18–22 August 2014; pp. 1038–1043.

[4]. Islam, Md & Hossain, Mohammad & Islam, Raihan & Andersson, Karl. (2019). Static Hand Gesture Recognition using Convolutional Neural Network with Data Augmentation. 10.1109/ICIEV.2019.8858563.

[5]https://doi.org/10.1016/j.procs.2020.04.255

Timeline: