INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

IJARIIT

ISSN: 2454-132X

Impact factor: 4.295

(Volume 4, Issue 1)

Available online at www.ijariit.com

Hand Detection and Gesture Recognition in Real-Time Using Haar-Classification and Convolutional Neural Networks

Kartik Wadehra
kwadehra@gmail.com
Maharaja Agrasen Institute of Technology, Rohini, Delhi

Mukul Kathpalia
mukulkathpalia@gmail.com
Maharaja Agrasen Institute of Technology, Rohini, Delhi

Vasudha Bahl
vasudha.bahl@gmail.com
Maharaja Agrasen Institute of Technology, Rohini, Delhi

ABSTRACT

Over the past few years, with the development of hardware and software, Human-Computer Interaction (HCI) has evolved from the use of keyboard and mouse to a more gesture-based approach to make the computer function according to your own will. Gesture Recognition has been used successfully in various applications like gaming (Xbox Kinect), recognizing sign languages and many others. The use case for the gesture recognition problem is very vast and therefore is being worked on continuously.

Keywords: *Gesture Recognition, Haar Feature-based classification, Convolutional Neural Networks, Hand Detection.*

1. INTRODUCTION

The gesture is a symbol of physical behavior or emotional expression. It includes body gesture and hand gesture. It falls into two categories: static gesture and dynamic gesture. For the former, the posture of the body or the gesture of the hand denotes a sign. For the latter, the movement of the body or the hand conveys some messages.

Human-Computer Interaction (HCI) can be more natural with gesture recognition. Movement of a mouse with hand movements, using fingers to select right clicks or left clicks et cetera can make HCI more effective. In the recent times, many innovations have been made in the field of Gesture Recognition and Human-Computer Interaction [1] [2] [3]. These innovations have led to various enlightening and important results in this field.

Gestures can be distinguished into two different categories – static and dynamic. A static gesture is a particular hand configuration and pose, represented by a single image. A dynamic gesture is a moving gesture, represented by a sequence of images.

We focus on the recognition of static gestures. Although, the gestures are static the movement of hands as an error is considered and therefore the gesture is recognized after a certain interval or can be set to be recognized at the click of a key.

With the recent development of deep learning, a few methods have been developed based on Convolutional Neural Networks (ConvNets) [4], [5], [6], [7], [8], [15] and Recurrent Neural Networks (RNNs) [9], [10], [11], [12]. However, proposed method considers a video as a sequence of still images with some form of progressive difference, or as a continuous system of images or image features.

2. SYSTEM DESIGN FOR HAND DETECTION AND GESTURE RECOGNITION

The system will take a two-step approach:

- Detection of a hand in a live video feed using a trained Haar Classifier.
- Recognition of the Gesture made by the hand using a trained Convolutional Neural Network.

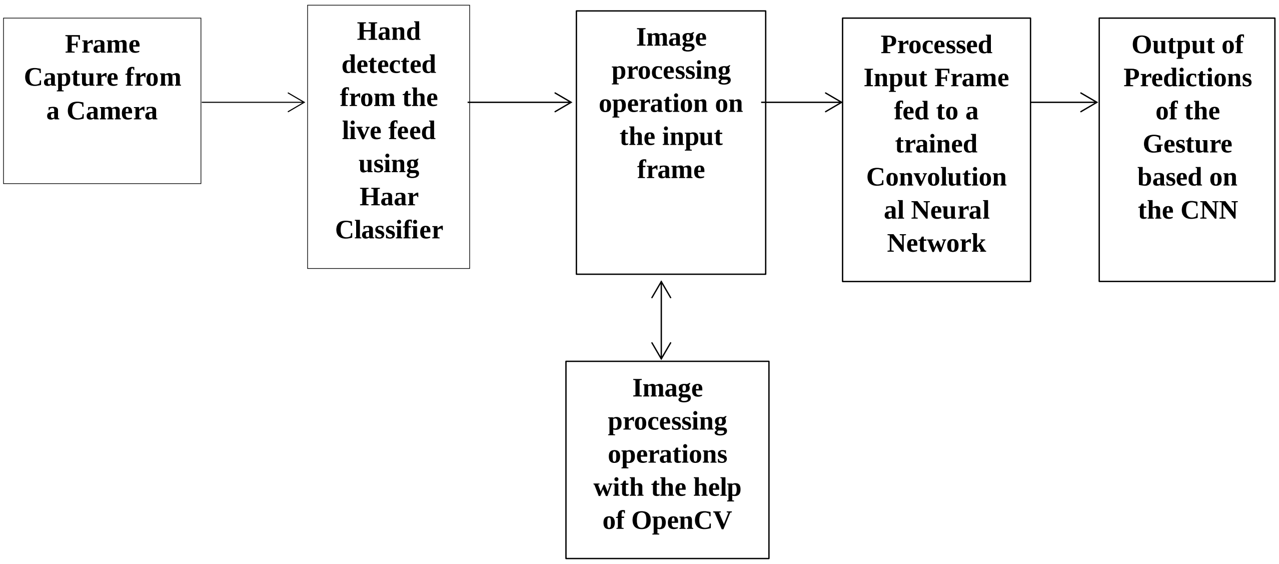

Frame Capture from a Camera
Hand detected from the live feed using Haar Classifier
Image processing operation on the input frame
Processed Input Frame fed to a trained Convolutional Neural Network
Output of Predictions of the Gesture based on the CNN
Image processing operations with the help of OpenCV

Fig. 1. Block Diagram of the System

3. DETECTION OF HAND IN A LIVE FEED USING HAAR CLASSIFICATION

Using the webcam of a laptop or an external camera, the input is taken. The input is taken frame by frame and each frame is compared with the previous to check if there has been a change in the frame. If there has been a change, then the process starts over else if the frame difference has been less than 80% then the gesture recognition starts.

Haar Classification

Haar features based cascade classifiers [13] is an object detection algorithm which can be used to detect specific features in an image, for example, a specific expression of emotion within a face or a specific object like a person on a sidewalk within traffic. In this process, the algorithm is fed with a lot of positive images (the images contain very specifically the features that want to be identified) and negative images (these images contain anything other than the specific feature to identify, for example, backgrounds like walls, wallpapers et cetera)done using Adaboost learning process and Integral images [13].

Haar features based classification is used to detect the hand out of the surroundings in the live video feed. Haar classifier is generated by training the classifier with over 2000 real-world positive images of hands and also some processed images for later use in training the Convolutional Neural Network.

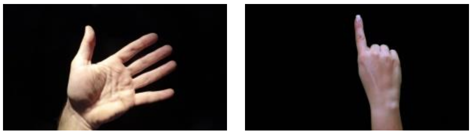


Fig 2: Examples of Processed Images with Black Backgrounds

The processed images were used so because when training the classifier it would be easier to mark the location of the required feature in the script written for training the Haar feature-based classifier.

The training of the classifier was done till 15 stages and took near 3 hours to complete. An XML file was generated to use it in the system.

4. FRAME CAPTURING AND IMAGE PROCESSING OPERATIONS

A frame is captured from the video feed from the webcam. RGB trackbars can be provided to adjust the skin color values according to the lighting and background. Firstly, before sending the frame for further processing, it is compared with the previously stored frame to check if the hand gesture position has been changed or not. If the difference in the current and previous frame is greater than 80%, then the frame is read again otherwise it is sent for further processing.

The image is converted to YCrCb format. The YCrCb color space is derived from the RGB color space and has three components.

- Y – Luminance part obtained after gamma correction of an RGB image.
- Cr = R – Y (deviation of the red part from Luminance).
- Cb = B – Y (deviation of the blue part from Luminance).

As YCrCb format takes advantage of the fact that an effective representation of a scene can be achieved by separating the luminance and chrominance components. It also uses lesser bits for chrominance than luminance using Color sub-sampling. This color space also helps in getting rid of some redundant information.

Gaussian Blurring is done to the input frame for removal of noise, edge smoothing et cetera which is considered high-frequency components, therefore edges are blurred a bit in the operation.

Then the skin region of the frame is detected using the trackbar input introduced earlier using it for lower and higher threshold and the values in the image lying between the thresholds would be considered as the skin region.

Contour detection in done on the frame and the largest contour is extracted and its background is removed. This contour is then made into a new image with a black background (corresponding to the training images used for our convolutional neural network).

Then a bounding rectangle is placed around the detected hand region.

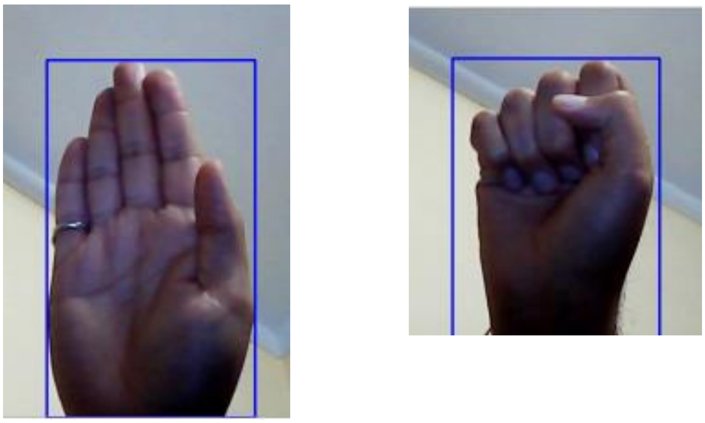


Fig 3: A Blue Bounding Rectangle around the Detected Hand Region

Contour detection in done on the frame and the largest contour is extracted and its background is removed. This contour is then made into a new image with a black background (corresponding to the training images used for our convolutional neural network).

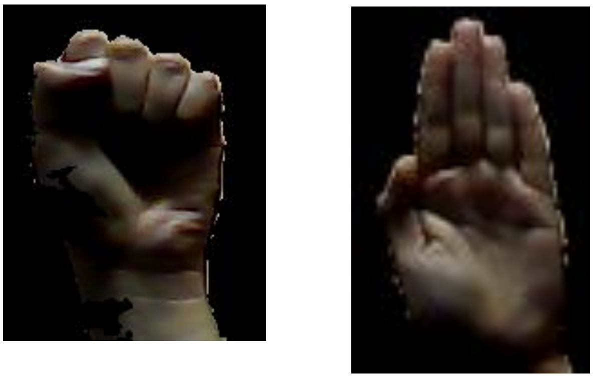


Fig 4: Example of Images Sent for Prediction

Additional tasks include finding the convexity defects and doing further improvements to fix them.

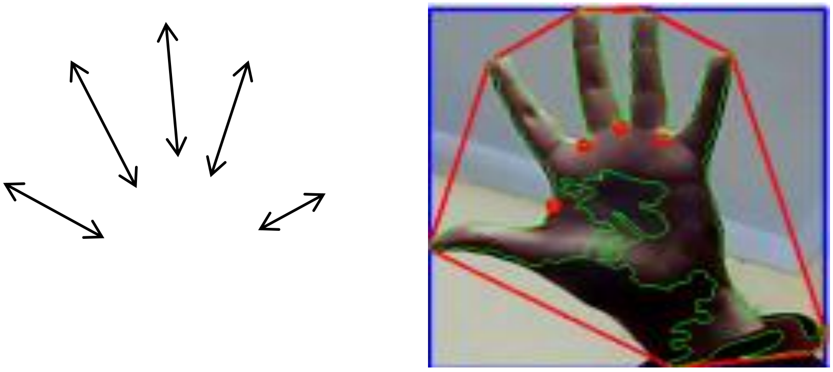


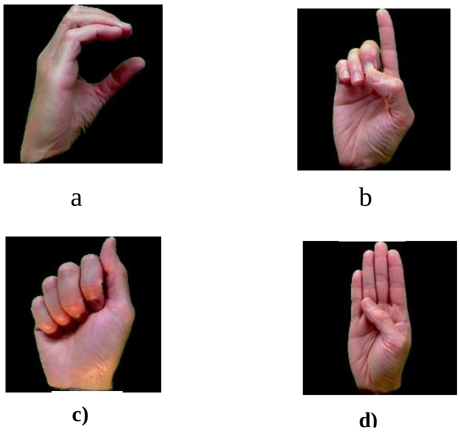
Fig 5: Red Represents the Convex Hull and Black Arrows Represent the Convexity Defects

5. CONVOLUTIONAL NEURAL NETWORK AND ITS ARCHITECTURE

For the prediction process for this system, a Convolutional Neural Network [14] is used. Convolutional Neural Networks takes as input an image, define a weight matrix and convolution is applied to the input to extract the required specific features. The information about the spatial arrangement is not lost during this process. Gesture Recognition using Convolutional Neural Networks have been a part of much research before [16] [18] and we aim to add further to it.

Dataset used

The dataset used consists of images of hand gestures used in American Sign Language. Each of the hand gesture has been extracted from its original image and pasted on a black background.



a b

c) d)

Fig 6: Examples of Images in the Dataset:
a) A in ASL b) B in ASL c) C in ASL d) D in ASL

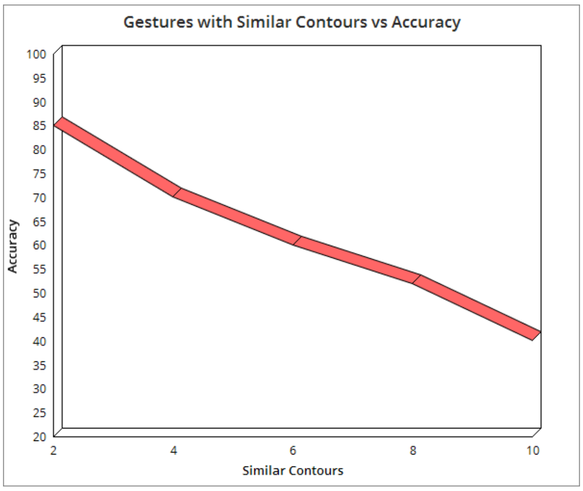
The architecture of the Convolutional Neural Network used is as follows:

- First two convolutional layers both have 32 filters each of size 3x3. Both of these layers are followed by an Activation layer using Rectified Linear Unit (ReLu) activation.
- The above system is then followed by a Max Pooling layer with downsizing factor of 2.
- The above combination of layers is then followed by a convolution layer with 64 filters each of size 3x3. Then a ReLu Activation layer and then a Max Pooling layer with a downsizing factor of 2.
- The above combination is followed by a Flatten layer, then a fully connected layer with the number of classes to be identified as an argument and a softmax Activation layer at the end.

The images used for training have been reduced to the size 100x100 and are then fed to the Convolutional Neural Network in the batches of 32 going for 100 epochs.

6. OBSERVATIONS AND FEW NOTES ON THESE OBSERVATIONS

When the network is trained for only two classes (A and B) the classification is accurate to 90%.
But as the similarity between the contours of the gestures increases, the accuracy decreases.



As the graph exhibits, when gestures that form similar contours increase, the accuracy of the system decreases continuously.

7. IMPROVEMENTS

- Better training of Haar feature-based classifier can lead to better hand detection.
- Model improvement and refinement of the CNN can lead to better results and feature extraction.

8. APPLICATIONS

We hope to further improve the accuracy of the system and use it in Sensor-based applications like a drive-thru ordering system in restaurants for the people that communicate through sign language.

9. REFERENCES

[1] Victor Adrian Prisacariu, Ian Reid:"3D hand tracking for human-computer interaction", Image and Vision Computing 30 (2012) 236–250.

[2] V. Prisacariu, I. Reid, and PWP3D: Real-time segmentation and tracking of 3D objects, Int. J. Comput. Vision (2011), doi: 10.1007/s11263-011-0514-3.

[3] Attila Licsa´ra, Tama´s Szira´nyi "User-adaptive hand gesture recognition system with interactive training" Image and Vision Computing 23 (2005) 1102–1114.

[4] P. Wang, W. Li, Z. Gao, C. Tang, J. Zhang, and P. O. Ogunbona, “Convents-based action recognition from depth map through virtual cameras and pseudocoloring,” in Proc. ACM international conference on Multimedia (ACM MM), 2015, pp. 1119–1122.

[5] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. Ogunbona, “Action recognition from depth maps using deep convolutional neural networks,” Human-Machine Systems, IEEE Transactions on, vol. 46, no. 4, pp. 498– 509, 2016.

[6] P. Wang, Z. Li, Y. Hou, and W. Li, “Action recognition based on joint trajectory maps using convolutional neural networks,” in Proc. ACM international conference on Multimedia (ACM MM), 2016. pp. 1–5.

[7] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, “Dynamic image networks for action recognition,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[8] P. Wang, W. Li, S. Liu, Y. Zhang, Z. Gao, and P. Ogunbona, “Large-scale continuous gesture recognition using convolutional neural networks,” in Proceedings of ICPRW, 2016.

[9] Y. Du, W. Wang, and L. Wang, “Hierarchical recurrent neural network for skeleton-based action recognition,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1110– 1118.

[10] V. Veeriah, N. Zhuang, and G.-J. Qi, “Differential recurrent neural networks for action recognition,” in Proc. IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4041–4049.

[11] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, “Concurrence feature learning for skeleton-based action recognition using regularized deep LSTM networks,” in The 30th AAAI Conference on Artificial Intelligence (AAAI), 2016.

[12] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, “NTU RGB+ D: A large-scale dataset for 3D human activity analysis,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[13] P. Viola and M. Jones, Rapid object detection using a boosted cascade of simple features, CVPR, 2001.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks.

[15] E. Ohn-Bar and M. Trivedi. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. IEEE Trans. on Intelligent Transportation Systems, 15(6):1–10, 2014.

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale visual recognition. In ICLR, 2015.

[17] P. Molchanov, S. Gupta, K. Kim, and K. Pulli. Multi-sensor system for driver’s hand-gesture recognition. In IEEE Automatic Face and Gesture Recognition, 2015.

[18] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition. In NIPS, 2014.