e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

# RECOGNITION OF DYNAMIC HAND GESTURES USING 3D CONVOLUTIONAL NETWORK

**Ankit Anurag *1, Dr Amita Goel*2, Nidhi Sengar*3, Vasudha Bahl*4**

*1Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, India.

*2Professor, Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, India.

*3Assistant Professor, Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, India.

*4Assistant Professor, Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, India.

## ABSTRACT

Hand gesture recognition are a major part of human-computer interaction (HCI). Given the progress in usage of augmented reality technology and increased integration of technology in everyday life, a better understanding of human gestures and actions have gained greater importance. Although there are many devices available which can be outfitted on an individual's hand to assist hand gesture recognition, such technologies are not economically feasible. Additionally, for increased convenience, it is necessary to avoid dependence on such devices. This work seeks to understand and implement hand gesture recognition models for better performance. A model to detect and recognise both static and dynamic gestures is designed.

**Keywords:** Human Computer Interaction, Hand Gesture Detection, Gesture Classification, Convolutional Neural Network, Dynamic Gestures.

## I. INTRODUCTION

Hand gesture recognition finds application in various scenarios. These include video games, device manipulation and augmented reality. Gestures are also more intuitive than pressing buttons or typing in data. There is no requirement for complex instructions to operate a device which is based on gesture recognition. Furthermore, gestures are an integral part of human-human interactions, and hence play a greater role in advent of robotics. Hand gesture recognition has various facets - spatial structure, temporal structure and contextual structure. Integration of hand gesture recognition in HCI carries great importance. Advancements in machine learning and deep learning have greatly helped researchers and organisations build more sophisticated and efficient gesture recognition models.

Figure - 1 shows the categorisation of hand gesture recognition techniques in two prominent groups - 3D model based and appearance based.

3D model based techniques are further categorised as volumetric - analysis of interactions of the subject's fingers with an object, geometric - analysis of articulation points in the hand and wrist, and skeletal - analysis of motion and shape according to skeletal structure.

Appearance based recognition techniques are also further categorised as colour-based - analysis using pixel values in image, silhouette geometry - analysis using background subtraction and hand geometry, deformable gabarit - analysis by using two dimensional templates to extract features, and motion-based - analysis using a sequence of frames.

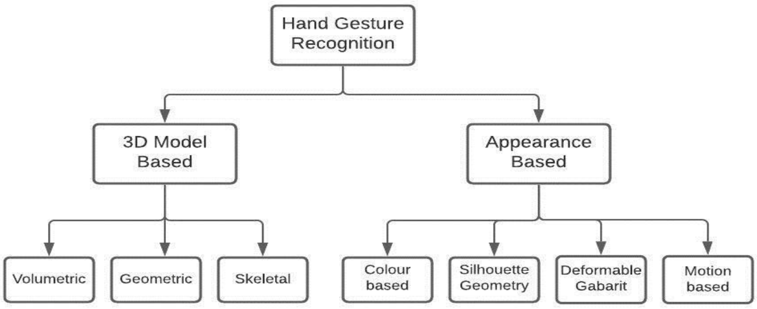


**Fig. 1** - Hand Gesture Recognition Techniques

The problem consists of identifying gestures with unknown backgrounds and noise, and to classify them accordingly. An object detection and object tracking module are used to identify and track the movement of hand. A 3D Convolutional Neural Network (CNN) is the used to identify the gestures. Most works on the topic are based on identification of hand gestures using sensors and other devices attached to the hand, or especially built to detect and recognise gestures, which does not assist in increasing the feasibility of applying this technology.

## II. LITERATURE REVIEW

Advancements in deep learning and neural network technology has facilitated greater work and research in the field of gesture recognition. Various algorithms and technologies have been developed to increase efficiency and accuracy of models.

Okan Kopuklu et al.[1] proposed a model that worked in two phases - a light weight gesture detection model to identify if a recognisable gesture is being performed and a heavy weight gesture recognition model to identify the gesture itself. EgoGesture and NVIDIA Dynamic Hand Gesture datasets were used with the ResNeXt-101 model.

Guillaume Devineau et al.[2] developed a novel dynamic hand gesture recognition algorithm based on deep learning. Model performance was analysed on gesture classification tasks. Parallel convolution layers process sequences of skeletal joints in a CNN model. The model achieved an accuracy of 91.28 percent, and 84.35 percent on a 14-gesture dataset and 28-gesture dataset respectively.

Amirhossein Dadashzadeh et al.[3] developed a two stage CNN architecture which first performed semantic segmentation and then identified the gestures based on the segments. It is called HGR-Net and is a combination of fully convolutional residual network and atrous spatial pyramid pooling. This algorithm operated at a rate of 23ms per frame.

Yuxiao Chen et al.[4] proposed a Dynamic Graph-based Spatial Temporal Attention method, where a fully-connected graph was constructed based on a hand skeleton and node features and edges were learned via self-attention mechanism performed in spatial and temporal domains.

Honghai Liu et al.[5] proposed an image segmentation method to improve the rate of gesture recognition and studied popular methods like Graph Cut and Random Walker. Interactive image segmentation using geodesic star convexity, are studied. The Gaussian Mixture Model was used for image modelling and continuous execution of the Expectation Maximum algorithm was used to learn its parameters.

Quentin De Smedt et al.[6] developed a new skeleton-based method where the geometric shape of hand was utilised to extract an effective descriptor from joints. The descriptor was encoded by a Fisher Vector representation and a linear SVM classifier was used for gesture recognition. This was evaluated on a dataset having 14 gestures performed by 20 participants.

## III. METHODOLOGY

### Dataset

For our work, we used images from the 20BN-Jester, which is a large collection of video clips labelled to show humans performing pre-defined gestures in front of a laptop camera or webcam.

The dataset comprises of 1,48,092 total clips which are divided into three categories of training set with 1,18,562 clips, validation set with 14,787 clips and test set with 14,743 clips. There are a total of 27 labelled gestures in dataset.

For our use, we selected five gestures from the set.

1. Swiping Left
2. Swiping Right
3. Thumb Up
4. Pulling Hand In
5. No Gesture

### Architecture

Before providing a description of the architecture and various components in use, we first attempt to understand a deep convolutional neural network.

Convolutional Neural Network, or more commonly known as CNN, is one of the most popular and useful techniques in computer vision problems. It traditionally comprises of a convolutional layer followed by a pooling layer. Once an image is processed by these layers, it is flattened to the form a 1D vector and used to train a neural network. Convolutional layers make use of filters which identify and extract information from images like edges. Many such filters are used in every layer. This is followed by a pooling layer, which can use strategies like max pooling and min pooling to actuate the important information from the obtained processed images. While there is loss of information in this layer, it works to reduce the computational load on the system. The final neural network layer classifies the image into various classes, acting as a traditional neural network system.

3D CNN are designed to process a sequence of frames and are utilised to extract and identify information from video clips. In case of dynamic gestures, where the information is contained in a clip instead of an image, a 3D CNN is better utilised.

Any image or clip originally contains noise and non-uniformity. This prevents proper analysis and recognition of information from the data. Furthermore, any neural network expects uniformity in data in terms of size and dimensions. For preprocessing, the image was resized and changed to grayscale. The pixel values were then standardised by subtracting the mean to avoid excess variation in data.

We actualised our model as a convolutional neural network comprising of two convolutional layers and two pooling layers. After flattening, there are three dense layers to process the vector, separated by two dropout layers. The rectified linear unit function has been used as the activation function. We have used 'adam' as the optimiser in compilation. The loss function in use was categorical cross entropy.

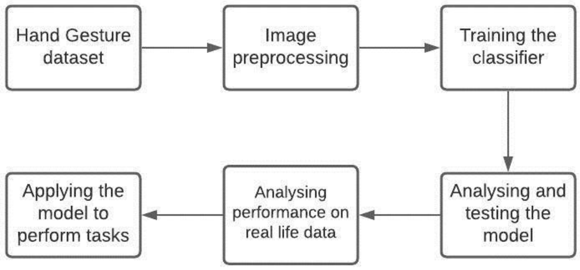

Hand Gesture dataset
Image preprocessing
Training the classifier
Analysing and testing the model
Analysing performance on real life data
Applying the model to perform tasks

**Fig. 2** - Block diagram of proposed methodology

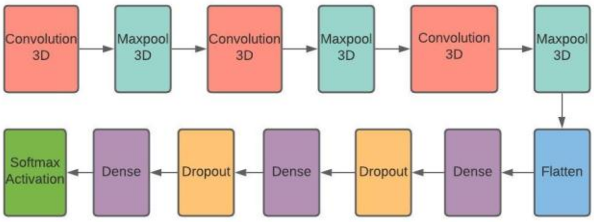


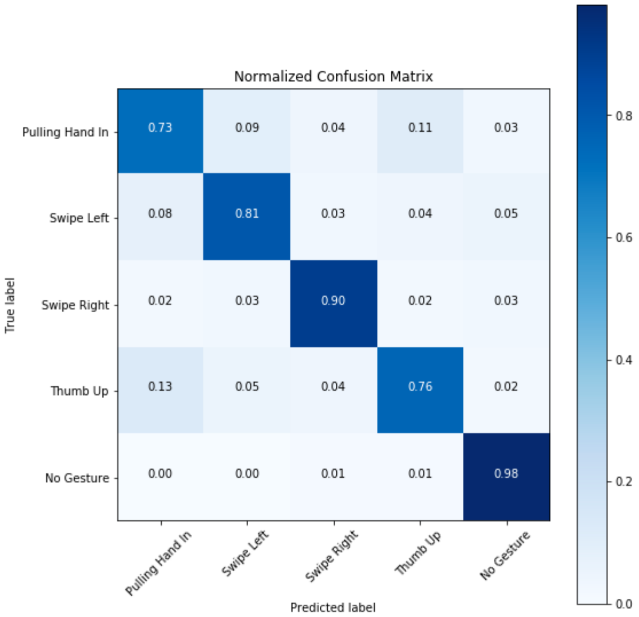
**Fig. 3** - Model Architecture

## IV. MODELING AND ANALYSIS

The model was trained on the 20BN-Jester dataset to recognise five gestures in real time image capture. The 'OpenCV 'python library was utilised for capturing and processing the image. Tensorflow.Keras api was utilised for developing and training the model. The 'scikit-learn 'and 'matplotlib.python 'libraries were used for analysing the performance of the model.

## V. RESULTS AND DISCUSSION

The model was trained over 200 epochs with a batch size of 32, with 80-20 split for training and test data respectively. 1500 clips of each of the five gestures were taken, with a total of 7500 images being used for training the model. The model displayed a peak accuracy of 83.73 % on validation data after 200 epochs.



| True label \ Predicted label | Pulling Hand In | Swipe Left | Swipe Right | Thumb Up | No Gesture |
|---|---|---|---|---|---|
| Pulling Hand In | 0.73 | 0.09 | 0.04 | 0.11 | 0.03 |
| Swipe Left | 0.08 | 0.81 | 0.03 | 0.04 | 0.05 |
| Swipe Right | 0.02 | 0.03 | 0.90 | 0.02 | 0.03 |
| Thumb Up | 0.13 | 0.05 | 0.04 | 0.76 | 0.02 |
| No Gesture | 0.00 | 0.00 | 0.01 | 0.01 | 0.98 |

**Fig. 4** - Confusion matrix

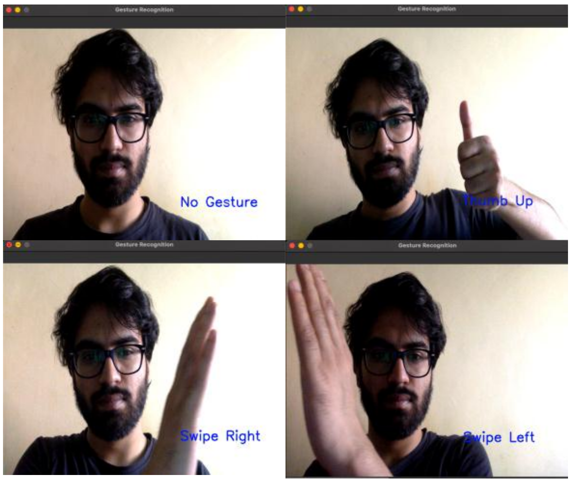


**Fig. 5** - Model performance with live feed

## V. CONCLUSION

A 3D CNN to reliably identify gestures was successfully implemented for the relatively smaller number of gestures. Increasing the classes will also require a more complex architecture. Furthermore, hyperparameter tuning can be performed to better improve the model performance.

The overall accuracy of the model was obtained to be 83.7%. As can be noted from the confusion matrix, the “pulling hand in” gesture is mostly responsible for lower accuracy. This can be contributed to the lack of definition in the gesture itself. A different gesture might result in better model accuracy.

Another point of note is that a very generic image processing model is used. Utilisation of hand detection and segmentation techniques can further be done to improve reliability and accuracy of the model. Image processing is always the most crucial part in computer vision problems and can greatly enhance model performance.

## ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude towards my mentors Dr. Amita Goel, Ms. Nidhi Sengar and Ms. Vasudha Bahl for their guidance and constant supervision during the development of this work, as well as their support in providing the necessary information regarding the project

## VI. REFERENCES

[1] Okan Köpüklü, Ahmet Gunduz, Neslihan Kose, Gerhard Rigoll : Real-time Hand Gesture Detection and Classification Using Convolutional Neural Networks, Jan 2019

[2] Guillaume Devineau, Wang Xi, Jie Yang, Fabien Moutarde : Deep Learning for Hand Gesture Recognition on Skeletal Data, 2018

[3] Amirhossein Dadashzadeh, Alireza Tavakoli Targhi, Maryam Tahmasbi, Majid Mirmehdi : HGR-Net: A Fusion Network for Hand Gesture Segmentation and Recognition, Jun 2018

[4] Yuxiao Chen, Long Zhao, Xi Peng, Jianbo Yuan, Dimitris N. Metaxas : Construct Dynamic Graphs for Hand Gesture Recognition via Spatial Temporal Attention, Jul 2019

[5] Honghai Liu, Hui Yu, Zhaojie Ju, Heng Tang, Guozhang Jiang, Jianyi Kong, Ying Sun, Gongfa Li, Disi Chen : An Interactive Image Segmentation Method in Hand Gesture Recognition, 2017

[6] Quentin De Smedt, Hazem Wannous, Jean-Philippe Vandeborre : Skeleton-based Dynamic hand gesture recognition, 2016