Automated Hand Gesture Recognition using a Deep Convolutional Neural Network model

Ishika Dhall
Department of Computer Science & Engineering
Amity University, Uttar Pradesh
India
ishikadhall11@gmail.com

Shubham Vashisth
Department of Computer Science & Engineering
Amity University, Uttar Pradesh
India
shubham.vashisth.delhi@gmail.com

Garima Aggarwal
Department of Computer Science & Engineering
Amity University, Uttar Pradesh
India
gmehta@amity.edu

***Abstract*—The tremendous growth in the domain of deep learning has helped in achieving breakthroughs in computer vision applications especially after convolutional neural networks coming into the picture. The unique architecture of CNNs allows it to extract relevant information from the input images without any hand-tuning. Today, with such powerful models we have quite a flexibility build technology that may ameliorate human life. One such technique can be used for detecting and understanding various human gestures as it would make the human-machine communication effective. This could make the conventional input devices like touchscreens, mouse pad, and keyboards redundant. Also, it is considered as a highly secure tech compared to other devices. In this paper, hand gesture technology along with Convolutional Neural Networks has been discovered followed by the construction of a deep convolutional neural network to build a hand gesture recognition application.**

***Keywords*—Convolutional Neural Networks, Hand Gesture Recognition system, Feature Map, Deep neural network**

I. INTRODUCTION

A gesture is a body movement that conveys a noteworthy implication. Gesture recognition is a computer science technology that helps a user in interacting with their digital devices using simple and natural body gestures. Gesture recognition technology can be beneficial at many places like automated home appliances, hand signal interpretation [1], automobiles, etc. Hand gesture recognition is a part of gesture recognition that is based on recognizing the movements of hands meant to be delivered, for example: showing a forefinger could denote the number "1" or a thumbs up could be an indication of agreement.

Deep learning is a fragment of a wide-ranging family of Artificial Intelligence. It essentially puts a light on the concept of a multi-layer perceptron learning. A Convolutional Neural Network commonly known as a Comp Net is a neural network class used in deep learning which is most applied to images and videos for their analysis. A CNN is a technique, or a machine learning model that can be applied to images to make them interpretable by machines. It can be implemented in other data analysis and classification problems as well. It is a type of artificial neural network which has a specialty of being able to deduce or distinguish patterns and understanding them. It is different than other deep learning models as it has an extra set of hidden layers called the convolutional layers along with the standard hidden layers. It can have one or more than one convolutional layer followed by the fully connected layers. The system will be learning features from each gesture and then further classify it. The entire notion of making a machine learn and making it smart is based on the abundance of data or information.

Our data fuels the machines and is used to make the machine learn to make predictions. The main aim of this paper is to train an algorithm which enables it to classify images of various hand gestures and signs like thumbs up, bolted fist, finger count, etc. Since the analysis of visual imaginings is being used, the class used to perform deep learning will be Convolutional Neural Network with Keras and TensorFlow as it is the standardized version of a multilayer perceptron.

This research provides the reader with a profound knowledge of a deep convolutional neural network. Also, this paper uses the data captured using the OpenCV library which will contribute to improving the accuracy score of the existing hand gesture recognition techniques.

In this research, a real-time anti-encroaching hand gesture recognition and hand tracking mechanism has been proposed which will improve the human-computer interactions and bring ease for the ones who rely on gestures for their day-to- day communication. It can be a significant communication tool for deafened people and people with ASD or autism spectrum disorder. It can be of great assistance for SOS signaling.

II. LITERATURE REVIEW

A technique of hand gesture recognition on a video game-based application has been proposed in [1].

978-1-7281-2791-0/20/$31.00 ©2020 IEEE

A new algorithm has been discussed in [1] to recognize and track hand gestures for better interaction with a video game. It is consisting of four hand gestures and four-hand direction classes to fulfil requirements that could have been extended to make it more powerful. It uses segmentation and tracking [2]. The proposed algorithm was performed on 40 samples and the accuracy turned out to be quite impressive.

Use of a convolutional neural network to reduce the feature extraction process and parameters being used has been discussed in [2]. The hand gesture recognition is performed using a convolutional neural network but the one used in our paper shows a deep convolutional neural network implementation. Results shown in [3] are very impressive when a training set of 50% of the database is used. Max Pooling Convolutional Neural Network to advance Human- robot interactions using color segmentation, edge blurring with morphological digital image processing and then experimenting with mobile robots using ARM 11 533 MHz [3].

They manage to get an accuracy score of around 96%. The vocabulary in the proposed project was up to 11 classes. It could have been used in a milieu of human-swarm interaction. Driver's hand gesture recognition via a 3D convolutional neural net was followed [4]. It engages spatial data augmentation techniques and pre-processing techniques for better results.

They achieved a score of 77% which could have been improved by constructing a deeper neural network. The challenges of a 3D CNN to perform classification and detection on the given dataset was addressed in [5] which also introduces a multi-modal dynamic challenging dataset and achieved an accuracy of 83.8%.

III. PRELIMINARIES

A convolutional neural network is a deep learning neural network class which is most applied to images and videos for its analysis. It is a kind of artificial neural network using machine learning algorithms for a unit and perceptron for supervised problems.

A CNN is basically a technique, or a machine learning model applied to images to make them interpretable by machines. It can have one or more than one convolutional layer followed by the fully connected layers. All types of cognitive tasks are performed using CNNs like Natural Language processing, image processing, etc. The concept of machine learning is not a contemporary thing, the first Artificial Intelligence-based program which came into play with a learned version of a game in which an Artificial Intelligence program was built that understood natural language finally.

In the given Fig 1, the image given as the input is used to find the primitive features like horizontal and vertical lines.

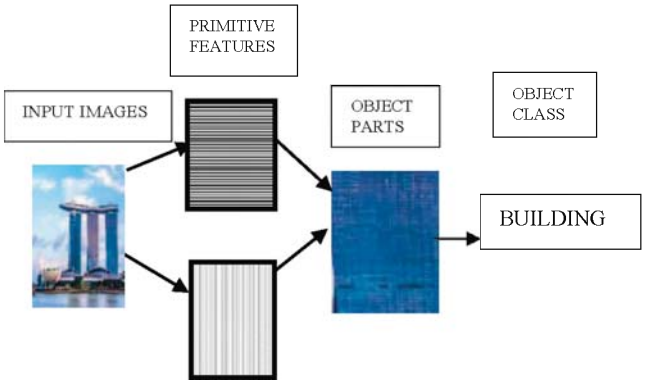


Fig 1. Working of a CNN Model

After the primitive feature extraction phase, the next stage determines the part of the given object using the features extracted. Objects parts are then used to interpret the class of the object.

Interfaces that cannot be touched not only improves the driver's focus and prevent possible mishaps but also makes the devices much more user-friendly due to which implementation of such technologies in several control systems is preferred.

IV. METHODOLOGY

Fig 2 shows a convolutional Neural Network model and various layers involved.

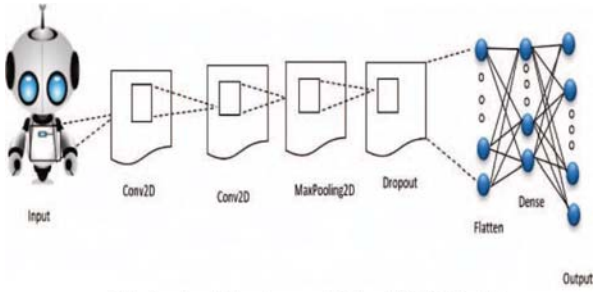


Fig 2. Architecture of the CNN Model

CNN is a type of neural network that is empowered with certain specific layers, such as:

1. Input layer
2. Hidden layer:

2.1. A Convolutional Layer
2.2. A Max Pooling Layer
2.3. A Fully Connected Layer
3. Output layer

A convolutional layer is the first Hidden Layer of convolutional deep learning and its key purpose is to detect and extract features from the image that was the input to our model like the edges and vertices of an image. Let's take an example, assume that we have an input image to be detected, then the hidden convolutional layer will help the system in finding the edges of that image. After that, we define a filter matrix also called a 'Kernel' which is further used for finding a new image with all the edges. Slide this filter for certain strides all over the image to find a new image with just the edges. Apply the dot product over the pixels of the image to find the image with all the edges. Edge image is useful for the initial steps and layers of CNN, in fact, it is the very first set of primitive feature sets for the working in a hierarchy of all the features.

The convolutional process is about detecting the edges, small patterns, orientations present in an image and it is present as a mathematical function. Convert the image into a matrix of binary integer values, i.e., 0s and 1s where appoint the value 0 to all the places where the image is black and appoint 1 to all the white parts of the image. Commonly while making this matrix, a value between 0 and 255 is used where all these numbers represent different shades of grey in case of a grayscale image. A three-channel image is used in case of input images which are not in grayscale format but are colored. Randomly select the type of filter and arbitrarily initialize the value of matrix elements.

The values of matrix elements are updated by optimum values as it goes through the training phase. Take dot product of pixels with every matrix value, now slide the kernel window over the image and find a new matrix called convolved image. Perform the element-wise multiplication of the kernel matrix and the image matrix. Try to apply multiple kernels in one image. Multiple kernels being applied to one image will result in multiple convolved matrices.

Leveraging different kernels assistances, it helps to find divergent patterns present in the image like curves, edges, etc. At the end of this layer, we receive a feature map which is the output of the process of convolution. Rectified linear unit (ReLu) [13] which is a non- linear activation function unit for providing a mapping between the response variables and inputs. Swap all negative numbers with a '0' and all positive numbers stay the same while using ReLu function (tf.nn.relu(x)).

For a pooling layer, down sample the output images for relu function to perform dimensionality reduction for the activated neurons. Use the Max Pooling layer to perform dimensionality reduction. Max Pooling finds the extreme values in input and simplifies the inputs. It diminishes the number of parameters within the model and converts lower-level data into a piece of higher-level information. In Fig 3, process of extracting a feature map is shown.

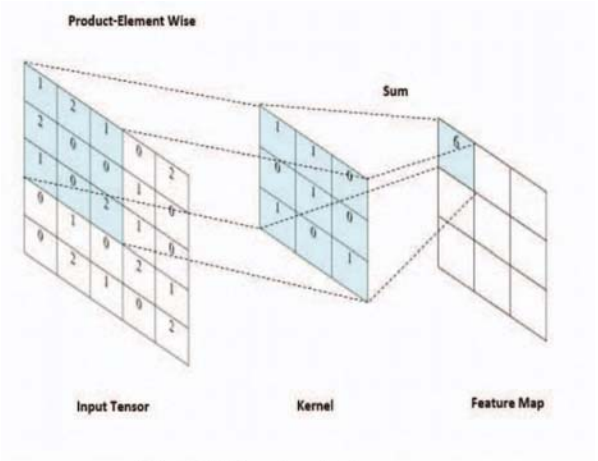


Fig 3. Extracting feature maps

For example, A 2x2 matrix results into a single pixel data by choosing the maximum value of the matrix. Strides dictates the sliding behavior of a max-pooling process [6]. It prevents overlapping, for example, if strides have a value two then the window will move 2 pixels every time. A fully connected layer takes the high-level images from the previous layer's filtered output and converts it into a vector. Each layer of the previous matrix will be first converted into a single (flatten) dimensional vector then each vector is fully connected to the next layer that is connected through a weight matrix [7]. SOFTMAX is an activation function that helps to find the class of each digit and generate probability for outputs.

The advantages of CNN include being a strong and computationally fast model, more efficient in terms of memory and it shows accurate results on images.

The disadvantages of CNN are that it is computationally expensive, quite complex, requires GPU and a large dataset along with the problem of Overfitting. It can also result in an imbalance in class. One of the parts of collected dataset is presented in Fig 4. It shows the images of class showing '1'.

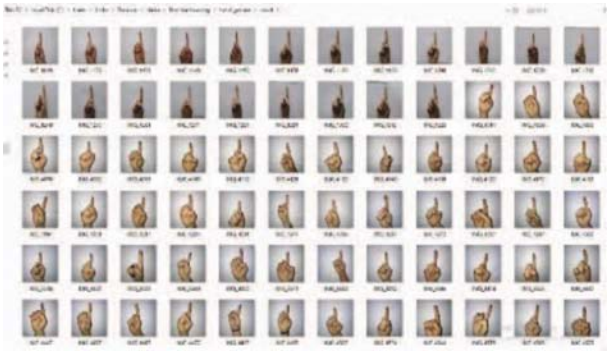


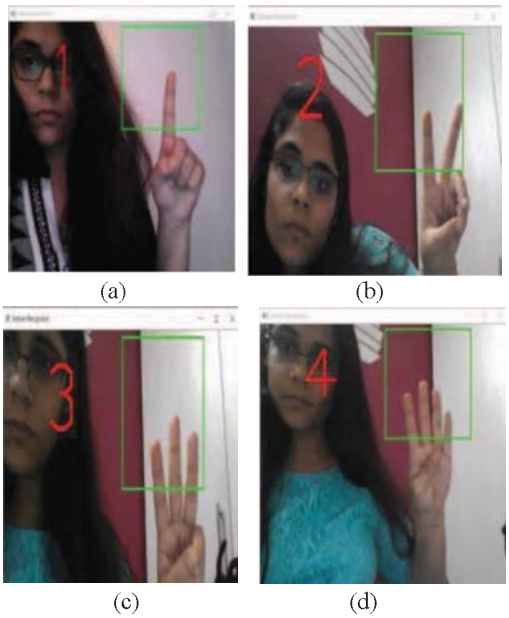
Fig. 4. Dataset of one of the labels

All the necessary libraries like Keras, TensorFlow, etc. were imported and data set was gathered using OpenCV library followed by an implementation of data augmentation technique which is an approach allows experts to expressively rise the variety of data available for model training, without collecting much data.

V. RESULT

The experimental results of this paper show that the model proposed in this paper can distinguish among several dominant and low-level features for the input images and can classify various hand gestures [8] with greater accuracy and a negligible model loss of 0.0504.

Fig 5(a) illustrates the hand gestures showing the value "1", 5(b) represents "2", 5(c) represents 3", 5(d) represents "4" and 5(e) represents "5" being captured and detected successfully.



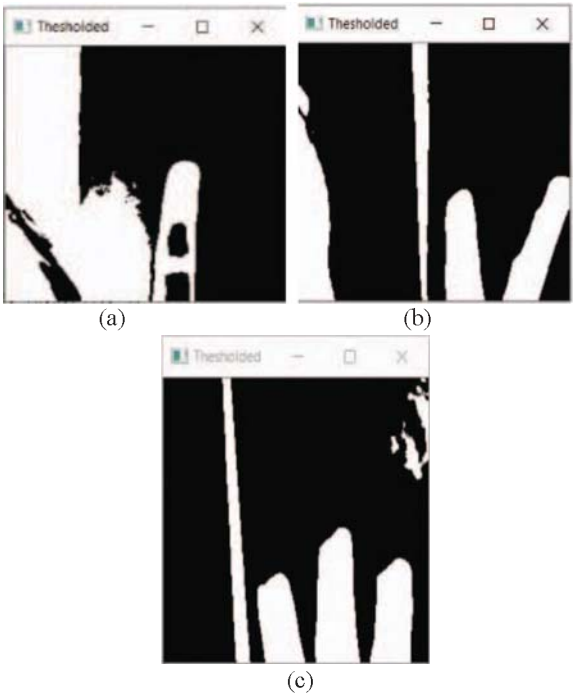
(a) (b)

(c) (d)

(e)

Fig 5. Hand gesture detection

Fig 6(a), 6(b), 6(c), 6(d) and 6(e) represents the pre-processed threshold images of "1", "2", "3", "4" and "5" respectively.



(a) (b)

(c)

(d) (e)

Fig 6. Threshold images

Before actual detection, the images to be trained are segmented. Afterwards, features [9] are extracted and reduced (finding the significant features and removing the unnecessary details that may create noise or reduce accuracy). Fig 7 presents a graph of results received from the presented model.

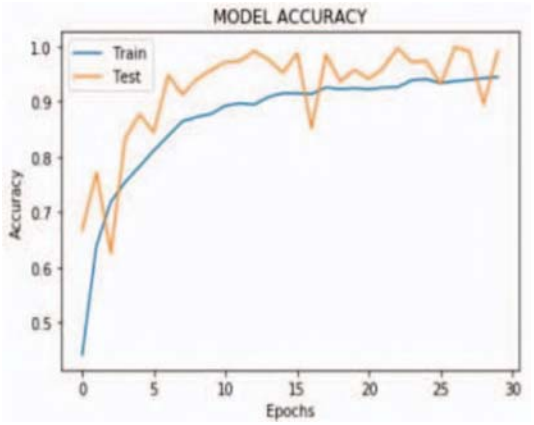


Fig 7. Result accuracy graph of the model

The model was able to achieve a testing accuracy of 99.13% which was much higher than the existing models. Gaining an accuracy as high as that was a tough job but it can be achieved by tuning various hyperparameters [10] and by performing proper data pre-processing and augmentation techniques.

Data pre-processing and data augmentation values:

Rescaling = 1./255
Rotation range = 20
Width shift range = 0.2
Height shift range = 0.2
Horizontal flip = True

The techniques like rescaling, cropping, padding, and horizontal flipping are very common for training dense neural networks. Data pre-processing [11] is also an important step which is mostly used in cases where we need to reduce the amount of data. It reduces the number of attributes, number of attribute values and the number of tuples. The most optimal values upon testing the accuracy of model [12] were as mentioned in the results above.

Here, Table 1 represents the values of various parameters that were tuned. The model was tested and the validation score of the model is calculated. The resultant Deep Convolution Network model comprises of seven hidden layers and tuned Hyperparameters [14] for our application.

Table 1. Optimal hyperparameters for the CNN

Parameters	Value
Hidden Layers	7
Dropout layer	1
Number of filters per Convolutional layer	32,64
Nodes in Dense layer	128
Batch size	32
Activation Function	ReLU
Optimizer	Adam
Epochs	30
Kernel Size	(3, 3)
Size of Pool	(2, 2)
Strides	(2, 2)

VI. CONCLUSION AND FUTURE WORK

This paper discusses and offers a state-of-the-art deep Multi-layer Convolutional Neural Network for performing hand gesture recognition in Human-Robot Interaction systems. It is an efficient model to be used on image data when tuned properly and with proper image pre-processing [15]. While detecting the images on live-videos or static images, the images and labels that were fed and trained in the model are used to compare the output. Its palpable ability to determine the invariant problem of recognizing gestures despite all the noise and complications is undefeatable. Also, it has been posed with real hand gesture images and despite hefty numbers and boundless structural overlapping of noises [16], the program worked fine along with providing a decent accuracy notch. This work can further be extended by adding more functionalities to the model and by making a greater number of classes.

Above mentioned explications put it in persuasive distinction to propose the mechanisms of recognition [17] which only work on meek images having fewer varying objects to be distinguished [18]. Therefore, this work is afoot in a door. Its remaining complications are meant to be resolved progressively.

REFERENCES

[1] Pigou, Lionel, et al. "Sign language recognition using convolutional neural networks." *European Conference on Computer Vision*. Springer, Cham, pp. 572-578, 2014.

[2] Li, Gongfa, et al. "Hand gesture recognition based on convolution neural network." Cluster Computing pp.2719-2729, 2017.
[3] Nagi, Jawad, et al. "Max-pooling convolutional neural networks for vision-based hand gesture recognition." 2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA). IEEE, pp. 342-347, 2011.
[4] Molchanov, Pavlo, et al. "Hand gesture recognition with 3D convolutional neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 1-7, 2015.
[5] Molchanov, Pavlo, et al. "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4207-4215, 2016.
[6] Tolias, Giorgos, Ronan Sicre, and Hervé Jégou. "Particular object retrieval with integral max-pooling of CNN activations." *arXiv preprint arXiv:1511.05879,* 2015.
[7] Ren, Zhou, et al. "Robust part-based hand gesture recognition using kinect sensor." IEEE transactions on multimedia 15.5, pp.1110-1120, 2013.
[8] Nishihara, H. Keith, et al. "Hand-gesture recognition method." U.S. Patent No. 9,696,808. 4 Jul, 2017.
[9] Chaudhary, Anita, and Sonit Sukhraj Singh. "Lung cancer detection on CT images by using image processing." *2012 International Conference on Computing Sciences*. IEEE, pp. 142-146, 2012.
[10] Binh, Nguyen Dang, Enokida Shuichi, and Toshiaki Ejima. "Real-time hand tracking and gesture recognition system." *Proc. GVIP* pp.19-21, 2005.
[11] Murthy, G. R. S., and R. S. Jadon. "A review ofvision based hand gestures recognition." *International Journal of Information Technology and Knowledge Management* 2.2, pp.405-410, 2009.
[12] Manresa, Cristina, et al. "Hand tracking and gesture recognition for human-computer interaction." ELCVIA Electronic Letters on Computer Vision and Image Analysis 5.3, pp.96-104, 2005.
[13] Agarap, Abien Fred. "Deep learning using rectified linear units (relu)." *arXiv preprint arXiv:1803.08375* 2018.
[14] Wang, Binghui, and Neil Zhenqiang Gong. "Stealing hyperparameters in machine learning." *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, pp. 36-52, 2018.
[15] Poostchi, Mahdieh, et al. "Image analysis and machine learning for detecting malaria." *Translational Research*194, pp.36-55, 2018.
[16] Santhanam, T., and S. Radhika. "A Novel Approach to Classify Noises in Images Using Artificial Neural Network1." 2010.
[17] Simonyan, Karen, and Andrew Zisserman. "Verydeep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* 2014.
[18] Kim, Youngwook, and Brian Toomajian. "Hand gesture recognition using micro-Doppler signatures with convolutional neural network." *IEEE Access* 4,pp.7125-7130 2016.