Hand Poses Detection Using Covolutional Neural Network

Md. Buran Basha, Vijayawada

Abstract: Hand gesture recognition is the method by which an individual involving only hands understands specific forms of shape and movement. There are many applications where it is possible to apply hand gesture recognition to enhance control, usability, interaction and training. Communication through hand gesture has been shown successful results for humans and active research continues to replicate the same performance in computer vision systems. Interaction between humans and computers can be significantly enhanced by developments in systems capable of recognizing different hand movements. In this paper we have considered leapGestRecog data set which consists of 1000 images of 10 different members and each of them constituting 9 different labels of hand i.e palm, fist, thumb, index, ok, c, down, palm moved, fist moved. We are using convolutional neural networks which provide a very good result when dealing with images. This hand gesture recognition can be widely in case of physically impaired people and video games which involve gestures to move or play. Now a days by showing some gestures we can open some applications in mobile phone. The main goal of hand poses detection is to detect the gesture and able to control it.

Keywords: Convolutional Neural Network, Human computer interaction, Gesture recognition, Deep learning, Hand gesture.

1 INTRODUCTION:
The way human-computer interaction has also been dramatically changed in recent years with the growing development of science and technology. There have also been various new types of human-computer interaction methods of communication in the field of vision. The mouse and keyboard’s collaborative mode has become a touch screen and voice. The more effective form of interaction, however, is to allow the computer to understand the language of human body. Gestures and poses are one of the fundamental means of communication between humans, while they can also play a crucial role in interaction between humans and computers, as they transfer some kind of meaning. The research area of pose and gesture recognition is aimed at recognising such expressions, typically involving some posture and or movement of entire body’s hands, arms, head, or even skeletal joints. The meaning may differ in some cases, based on facial expression. The characteristics of the gesture are first extracted when performing gesture recognition, and the gesture recognition is performed according to characteristics extracted. There are various gesture recognition methods existing. In traditional hand gesture recognition features are extracted using segmentation techniques with thresholding and drawback of this some of the features are getting lost and now-a-days various emerging technologies came into existence like neural networks which could capture complete features present in an image. Neural networks have ability to classify and identify. But the overfitting problem arises when number of neural network layers are shallow.

- Md. Buran Basha, Asst.Prof. Dept. of Ece, KLEF, Vijayawada
  Mail id:mohammadaburan@kluniversity.in
- S.Ravi Teja Pursuing Btech in Dept. Of Ece KL University 160041027 Mail Id: raviteja.singamsetty115@gmail.com.
- K. Pavan Kumar Pursuing Btech in Dept. Of Ece KL University 160040339 Mail Id: Kakarlapavankumar899@gmail.com.
- M.V.S.D. Anudeep Pursuing Btech in Dept. Of Ece KL University 160040501 Mail Id: anudeepmvsd143@gmail.com.
close by signals. We propose a novel profound learning engineering that uses a Convolutional Neural Network (CNN). All the more explicitly, we utilize the camera images so as to recognize and follow the subject's skeletal joints in the spatial coordinates. We at that point select a subset of these points, i.e., all that is included at any of the signals of our informational collection. At that point, we make a counterfeit picture dependent on these 3D facilitates. We apply the Discrete Fourier Transform on these pictures and utilize the subsequent ones to prepare the CNN. We contrast our methodology and past work, where a lot of hand-made measurable highlights on joint directions had been utilized. At long last, we exhibit that it is conceivable to productively perceive hand signals without the need of a component extraction step. The assessment happens utilizing another dataset of 10 hand signals of 10 different people by changing the centre of camera at various angles. In this paper we have 1000 images of 10 different people of 10 different signs. The main contribution of this paper based on CNN a gesture recognition method is proposed and we had evaluated with some data sets and attained high accuracy when comparing with others. A specific gesture by using recognition model can provide actual meaning of gesture. The problem arises as most number of samples in data set are getting trained and the model lose the ability to predict if any new input is given and this is over fitting problem and we can overcome this by using some additional algorithms. The labels included as hand poses in this paper are palm, l, fist, fist moved, thumb, index, ok, palm moved, c, down.

**Gesture Recognition**

In the present world, augmented simulation has continuously shown up in individuals’ day by day life, and it is without a doubt the standard of human-PC association later on. Nonetheless, at the contribution of human-PC association, there is no bound together way. With the novel focal points of signals, it will turn into the standard of future associations. At present, signal acknowledgment is predominantly isolated into two kinds: contact and non-contact. The contact association technique predominantly procures three-dimensional data of signals by methods for hardware, for example, gloves, however, the way of utilizing peripherals to a great extent confines the adaptability of human-PC connection. The non-contact sort of cooperation is essentially a visual-based strategy, which takes out the requirement for the administrator to wear any peripherals, also, the collaboration is progressively common and agreeable. Early motion acknowledgment depended on information gloves. In 1983, Grime et al. first utilized gloves with hub markers. They utilized the palm skeleton to perceive motions and complete basic signal acknowledgment. During the 1990s, with the bit of leeway of exact situating of peripherals, numerous incredible frameworks showed up at home and abroad, utilized information gloves to accomplish the acknowledgment of 46 explicit signals; The finger checking technique supplanted the information gloves and finished the acknowledgment of a few explicit motions, it accomplished great outcomes. In numerous human-PC collaborations, dynamic motions were regularly required, subsequently advancing the advancement of dynamic signals, utilized the data entropy calculation to portion the hand from the foundation picture and effectively applied it to the video information stream through the parallel processing calculation, and distinguished the removed objective picture with a precision pace of 95%, however, there were fewer signal classes that could be perceived. During this period, signal acknowledgment, for the most part, expected to be performed by methods for peripherals. In this manner, the use of signal cooperation was incredibly restricted. In 2010, Microsoft discharged a profundity sensor “Kinect” for sensory games, which could quantify the separation between the human body and the gadget, and could follow the developments of the human body. From that point forward, many signal acknowledgment calculations and frameworks have been founded on Kinect. Simultaneously, numerous electronic data organizations had additionally joined the point of motion collaboration and accomplished great outcomes. Juan et al. utilized face acknowledgment, discourse acknowledgment, and motion acknowledgment to apply it to ES8000 arrangement TVs for purusing pages, TV remote control also, different capacities. Around the same time, Microsoft utilized the Doppler impact, worked in speakers and mouthpieces to accomplish target situating and signal acknowledgment and created the signal cooperation device “Sound Wave”; Richard et al. presented the signal acknowledgment device “Handpose” in view of profundity data to follow the development of the turn in genuine time. Sungho and Wonyong likewise attempted to perceive dynamic signals. At this stage, some signal calculations and gadgets had arrived at the prerequisites of viable applications. Notwithstanding, such items calculations still had extraordinary issues, and there were numerous confinements in the application procedure. There was as yet a hole between the distinguishing proof and use of uncovered hands.

**Neural Network**

Convolution Neural Network is a typical profound learning design roused by organic common visual acknowledgment instruments. In 1959, Hubel and Wiesel found that creature visual cortical cells were answerable for recognizing optical signals. During the 1990s, distributed a paper that built up the cutting edge structure of CNN and later improved it. They planned a multi-layer fake neural system called LeNet-5 to order written by hand numbers. Like other neural systems, LeNet-5 could likewise be prepared by utilizing back propagation calculations. LeNet-5 had accomplished satisfying outcomes. In any case, due to the absence of capacity to process huge scale preparing information, LeNet-5 didn’t perform well on complex issues. In this way, the convolution neural system once fell into a low tide. With the advancement of GPU quickening agents and enormous information, the quantity of CNN layers has been developed, and the acknowledgment exactness has been significantly improved, so it has gotten a great deal of consideration and research. Since 2006, scientists have structured numerous approaches to defeat the trouble of profound convolution neural system preparing. Among them, AlexNet was one of the most popular. AlexNet utilized a great CNN structure to accomplish leap forward execution in picture acknowledgment. The general structure of AlexNet was like that of LeNet-5, however with more layers. After the accomplishment of AlexNet, specialists further planned a great deal of better characterization
models, including the four most renowned ones: ZFNet VGGNet GoogleNet and ResNet. They accomplished a higher characterization exactness. As far as structure, the number of layers of CNN expanded. The number of layers of the ILSVRC 2015 boss ResNet was multiple times further than AlexNet and multiple times further than VGGNet. By expanding the profundity, the system can utilize nonlinearity to determine the estimated structure of the goal work, in this manner further better portraying the highlights and accomplishing better order results. In view of real outcomes, they seem to deliver better examples (all the more sharp and clear pictures) than others. DCGAN was an expansion of GAN that brought a convolution neural system into a generative model for solo preparing, utilizing the ground-breaking highlight extraction abilities of the convolution arrange to improve the learning of the produced model. These days, different neural systems develop in an unending stream and are applied to a wide scope of fields, applied it to picture acknowledgment, applied it to data stowing away, applied it to common language handling.

2 LITERATURE SURVEY:
Wei Fang et.al., proposed based on CNN and DCGAN for calculation and text output. Normally for gesture we use traditional method like feature extraction but some of the features are getting lost and for this purpose CNN with DCGAN are used to express gesture recognition, calculation and text output and the achieved results. It is less susceptible to illumination and background interface and achieve an efficient real-time recognition effect. Oliveira Berneir et.al., presents a new method of hand gesture recognition based on input-output hidden markov models and diverse aspects of gestures are discussed in this process where gestures are extracted from sequence of video images by tracking skin. The algorithm used in this paper is IOHMM which deal with dynamic aspects of gestures. The drawback with this algorithm is that it uses current observation only and not a temporal windows fixed with priori. The main objective of this paper is to understand two types of gestures: deictic and symbolic. Jawad Nagi et.al., proposed max pooling convolution neural networks for vision based hand gesture recognition. This project mainly focuses on sign language recognition an human robot interaction. Using the deep convolution neural networks that incorporates convolution and max pooling to supervise feature training and assign human gestures to mobile robots. The contour of the hand is retrieved by segmentation of colour, the smoothed by processing of morphological image, eliminating noisy edges. In this paper they considered six gesture classes and obtained 96%accuracy. Vijay John et.al., proposed deep learning based fast hand gesture recognition using representative frames. The algorithm used is long term recurrent convolution neural network. The input is video sequence where multiple frames are sampled and from these some representative frames are selected to perform classification. To extract the selected frames novel tilted image patterns and tilted binary pattern within a semantic segmentation-based deep learning framework, deconvolution neural network. The advantage of using novel tilted image patterns is that if the image contains multiple non overlapping blocks and represent entire video sequence within a single tilted image. The binary patterns represent the output generated from video sequence using dictionary learning and sparse modelling framework and a comparative analysis is performed with baseline algorithms.

3 METHODOLOGY:
In this paper we are using convolution neural networks in which it consists of three main stages first is convolution layer second is max pooling layer and third is fully connected layer and each of them have their importance.

4 CNN:
Profound learning approaches have been playing a key job in the field of AI by and large. They are of PC vision and are among those that have profited the most. A few profound structures have been proposed during the last scarcely any years. In any case, the Convolution Neural Networks (CNN) still remain the predominant profound engineering in PC vision. A CNN takes after a customary neural system (NN), yet it varies since it will probably get familiar with a lot of convolution channels. In any case, preparing happens as with each and every other NN; a forward proliferation of information and a regressive proliferation of mistakes do happen to refresh loads. The key part of a CNN are the convolution layers. They are framed by gathering neurons in a rectangular lattice. The preparation procedure intends to become familiar with the parameters of the convolution. Pooling layers are normally set after a solitary or a lot of sequential or parallel convolution layers and take little rectangular squares from the convolution layer and subsample them to create a solitary yield from each square. At long last, thick layers (otherwise called "completely associated" layers) perform arrangement utilizing the highlights that have been separated by the convolution layers and afterward have been sub sampled by the pooling layers.

5 CONVOLUTION LAYER:
The first layer in convolution neural network is convolution layer and this can be used as filter and filter used is guassian filter where random samples are taken into consideration and this is used to extract features and also builds a relationship between pixels by considering small squares of input data. Two inputs are taken by this layer is that one of them is image matrix and the other is filter or kernel. By applying different filters we can perform operations such as edge detection, blur and sharpen the image. In this proposed method we use a stride of 2 where
it represents shifting of number of pixels over the given input matrix. As the stride is 2 we move the filters to 2 pixels at a time and so on and problem with this is some times filter do not perfectly match to the input image and then zeros must be added to fit to image and this is known as zero padding.

**Pooling Layer**

When the images are too large to reduce number of parameters we use this layer where down sampling takes place and removes the unwanted data and this can be done applying spatial pooling. This is broadly classified into max pooling layer, Sum pooling, Average pooling. In this paper we are using max pooling layer where it selects the largest element in given stride. As we have provided a stride value of 2 we consider 2x2 matrix and then select maximum element and replace it.

**Fully Connected Layer**

The extracted features are represented in form of vector and this layer mainly concentrate on high level features which correlate to particular class and particular weights so that we can compute the weights of previous layers.

### 6 PARAMETERS FOR CNN:

<table>
<thead>
<tr>
<th>Layer Information and parameters of CNN</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>conv2d1(Conv2D) (None, 8, 128, 128)</td>
<td>224</td>
</tr>
<tr>
<td>activation1(Activation) (None, 8, 128, 128)</td>
<td>0</td>
</tr>
<tr>
<td>conv2d2(Conv2D) (None, 8, 124, 124)</td>
<td>594</td>
</tr>
<tr>
<td>max_pooling2d1(MaxPooling2D (None, 8, 62, 62)</td>
<td>0</td>
</tr>
<tr>
<td>conv2d3(Conv2D) (None, 16, 60, 60)</td>
<td>1168</td>
</tr>
<tr>
<td>conv2d4(Conv2D) (None, 16, 58, 58)</td>
<td>2219</td>
</tr>
<tr>
<td>max_pooling2d2(MaxPooling2D (None, 16, 29, 29)</td>
<td>0</td>
</tr>
<tr>
<td>conv2d5(Conv2D) (None, 32, 27, 27)</td>
<td>4648</td>
</tr>
<tr>
<td>conv2d6(Conv2D) (None, 32, 25, 25)</td>
<td>9248</td>
</tr>
<tr>
<td>max_pooling2d3(MaxPooling2D (None, 32, 12, 12)</td>
<td>0</td>
</tr>
<tr>
<td>conv2d7(Conv2D) (None, 64, 10, 10)</td>
<td>18486</td>
</tr>
<tr>
<td>conv2d8(Conv2D) (None, 64, 8, 8)</td>
<td>30928</td>
</tr>
<tr>
<td>conv2d9(Conv2D) (None, 64, 6, 6)</td>
<td>36928</td>
</tr>
<tr>
<td>conv2d10(Conv2D) (None, 64, 4, 4)</td>
<td>36928</td>
</tr>
<tr>
<td>max_pooling2d4(MaxPooling2D (None, 64, 2, 2)</td>
<td>0</td>
</tr>
<tr>
<td>flatten1(Flatten) (None, 256)</td>
<td>0</td>
</tr>
<tr>
<td>dense1(Dense) (None, 32)</td>
<td>8224</td>
</tr>
<tr>
<td>dropout1(Dropout) (None, 32)</td>
<td>0</td>
</tr>
<tr>
<td>dense2(Dense) (None, 32)</td>
<td>1856</td>
</tr>
<tr>
<td>dropout2(Dropout) (None, 32)</td>
<td>0</td>
</tr>
<tr>
<td>dense3(Dense) (None, 10)</td>
<td>310</td>
</tr>
</tbody>
</table>

**Fig(2): Layer Information and parameters of CNN**

### 7 TRAINING DATASET:

**Fig(3): Hand Poses (i) palm (ii) Thumb (iii) Super (iv) Fist (v) Down (vi) One.**

Obtained Results:
CONCLUSION:
We propose an robust CNN model for efficient and effective hand poses detection and we have trained and tested using 1000 samples and the obtained the number of each label in confusion matrix with an 98% accuracy.

REFERENCES: