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 has 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:mohammadburan@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.*
- *Md. Buran Basha, S. Ravi Teja, K. Pavan Kumar, M.V.S.D. Anudeep.*

not increase the recognition level by increasing the sample size. Several areas of application may benefit from recognition of pose of a human or the gestures he/she performs , such as recognition of sign language , gaming, medical applications involving assessment of human 's condition and even navigation in virtual reality environments. There are different approaches and techniques involving some type of device, either "worn" by the subject (e.g.,acclerometers, gyroscopes,etc.) or tracking the movement of the subject(e.g.,Cameras). For finding the depth information a typical RGB Camera is used in the few years. The user need to stand in front of camera without wearing any kind of sensor. Several parts of body are represented in terms of 3D space. For training of pose or typical features are extracted and to recognise the gesture. The acknowledgment technique dependent on Convolutional neural systems can examine the reality changes of motions, yet the acknowledgment speed of the strategy isn't acceptable. With the quick advancement of AI also, profound learning in PC vision, strategies dependent on AI and profound learning have pulled in more and more specialists' consideration. Among them, the profound neural system has the qualities of a neighbourhood association, weight sharing, programmed include extraction, and so on., which brings new thoughts of signal acknowledgment. Accordingly, based on the multifaceted nature of signal changes, we propose a motion acknowledgment technique dependent on profound Convolutional Neural Network (CNN). We utilize the camera of the PC to gather the information straightforwardly. In any case, the nature of the inspected information is clearly influenced by the enlightenment. So we play out the light recognition right off the bat. The reason for this progression is to get top-notch tests. We just need to alter the camera edge, light force or on the other hand different strategies to accomplish this progression. Next, we use DCGAN to create new pictures to take care of the issue of over fitting. At long last, 5/6 of the information is utilized for preparing and the remaining 1/6 is utilized for testing. We plan two system structures to understand the articulation acknowledgment capacity, computation and content yield. It, for the most part, changes the profundity of the system and the number of parameters as indicated by the task. In this paper, we present a motion acknowledgment approach that spotlights

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, I, 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 t 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 perusing 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

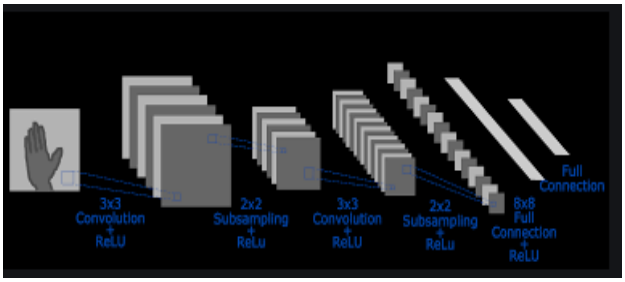
models, including the four most renowned ones: ZFNet VGGNet GoogleNet and ResNet They accomplished a higher characterization exactness. As far as structure, the number of layers of CNN expanded. The number of layers of the ILSVRC 2015 boss ResNet was multiple times further than AlexNet and multiple times 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. Oliveir 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%accuarcy. 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 tiled image patterns and tiled binary pattern within a semantic segmentation-based deep learning framework, deconvolution neural network. The advantage of using novel tiled image patterns is that if the image contains multiple non overlapping blocks and represent entire video sequence within a single tiled 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.



Fig(1): *Block diagram of model.*

The input to model is an image to extract features. For feature extraction convolution neural network is one of the excellent algorithm for extracting features.

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.

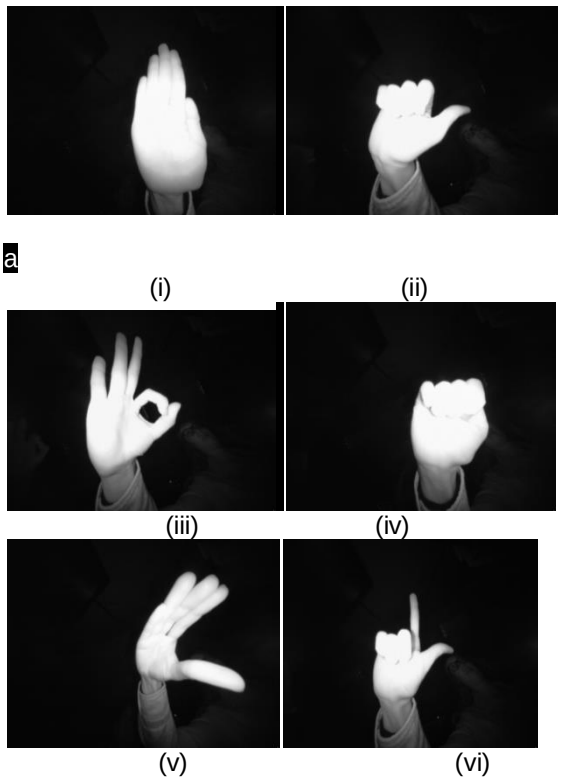
6 PARAMETERS FOR CNN:

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 8, 126, 126)	224
activation_1 (Activation)	(None, 8, 126, 126)	0
conv2d_2 (Conv2D)	(None, 8, 124, 124)	584
max_pooling2d_1 (MaxPooling2	(None, 8, 62, 62)	0
conv2d_3 (Conv2D)	(None, 16, 60, 60)	1168
conv2d_4 (Conv2D)	(None, 16, 58, 58)	2320
max_pooling2d_2 (MaxPooling2	(None, 16, 29, 29)	0
conv2d_5 (Conv2D)	(None, 32, 27, 27)	4640
conv2d_6 (Conv2D)	(None, 32, 25, 25)	9248
max_pooling2d_3 (MaxPooling2	(None, 32, 12, 12)	0
conv2d_7 (Conv2D)	(None, 64, 10, 10)	18496
conv2d_8 (Conv2D)	(None, 64, 8, 8)	36928
conv2d_9 (Conv2D)	(None, 64, 6, 6)	36928
conv2d_10 (Conv2D)	(None, 64, 4, 4)	36928
max_pooling2d_4 (MaxPooling2	(None, 64, 2, 2)	0
flatten_1 (Flatten)	(None, 256)	0
dense_1 (Dense)	(None, 32)	8224
dropout_1 (Dropout)	(None, 32)	0
dense_2 (Dense)	(None, 32)	1056
dropout_2 (Dropout)	(None, 32)	0
dense_3 (Dense)	(None, 10)	330

Total params: 157,074
Trainable params: 157,074
Non-trainable params: 0

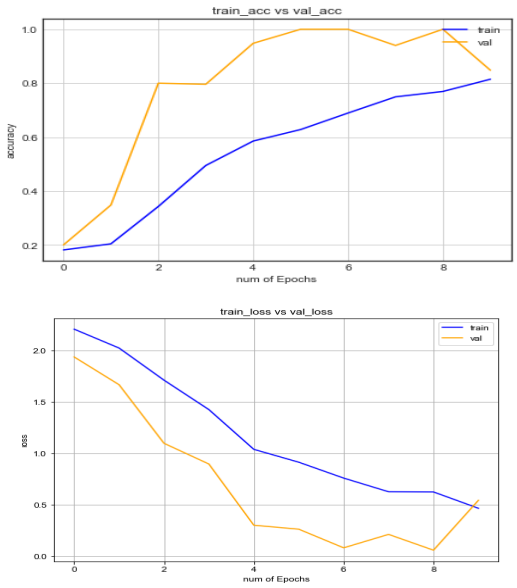
Fig(2): *Layer Information and parameters of CNN*

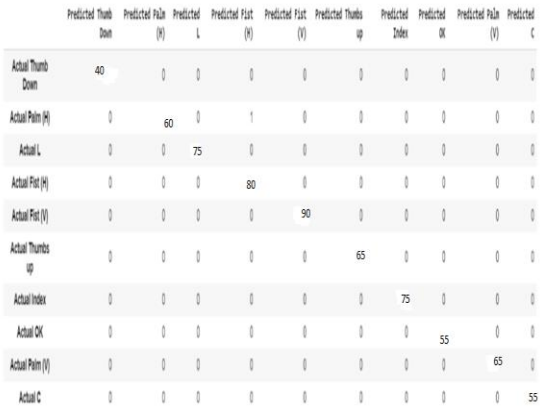
7 TRAINING DATASET:



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

Obtained Results:





	Predicted Thumb Down	Predicted Palm (H)	Predicted L	Predicted Fist (H)	Predicted Fist (V)	Predicted Thumbs up	Predicted Index	Predicted OK	Predicted Palm (V)	Predicted C
Actual Thumb Down	40	0	0	0	0	0	0	0	0	0
Actual Palm (H)	0	60	0	1	0	0	0	0	0	0
Actual L	0	0	75	0	0	0	0	0	0	0
Actual Fist (H)	0	0	0	80	0	0	0	0	0	0
Actual Fist (V)	0	0	0	0	90	0	0	0	0	0
Actual Thumbs up	0	0	0	0	0	65	0	0	0	0
Actual Index	0	0	0	0	0	0	75	0	0	0
Actual OK	0	0	0	0	0	0	0	55	0	0
Actual Palm (V)	0	0	0	0	0	0	0	0	65	0
Actual C	0	0	0	0	0	0	0	0	0	55

Fig(4): *Confusion Matrix*

8 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.

9 REFERENCES:

[1] WEI FANG et.al., proposed “Gesture recognition based on CNN & DCGAN for calculation and text output” IEEE Transactions on 10.1109/ACCESS 2019.2901930.

[2] CHRISTAIN WOLF et.al., proposed “Multi-Scale Deep Learning for Gesture Detection and Localization” ECCV Workshop pp 474-490 on 19 March 2015.

[3] M.Stecher et.al., proposed “Tracking down the illutiveness of gesture interaction in the truck domain” 6th International Conference on AHFE 3(2015) 3176-3183.

[4] Jawad Negi et.al., proposed “Max pooling convolutional neural networks foe vision-based hand gesture recognition” International Conference on Signal and Image Processing Applications 2011.

[5] Vijay John et.al., proposed “Deep Learning-Based Fast Hand Gesture Recognition Using Representative Frames” International Conference on digital image computing: Techniques and Applications 2016.

[6] Carl A. Pickering et.al., proposed “A reasearch study of hand gesture recognition technologies and applications for huamn vehicle interaction” 3rd Institute of Engineering and Technology Conference on Automative Electronics 2007.

[7] Soeb Hussain et.al.,proposed”Hand Gesture Recognition using deep learning” International SoC Design Conference 2017.

[8] Adnan Khashman et.al., proposed “Deep Learning in vision-based static hand gesture” Neural Computing and Applications Volume 28 pp 3944-3951 2016.

[9] S.Marcel et.al., proposed “Hand Gesture Recognition using input-output hidden markov models” Fourth IEEE International Conference on Automatic Face and Gesture Recognition PR00580 2000.

[10] Hatice Gunes et.al., proposed “Face and Body Gesture Recognition for a Vision-Based Multimodal Analyzer Computer Research Group PO 123 2007.

[11] Simonyan, K., Zisserman, A.: Two-Stream Convolutional Networks for Action Recognition in Videos. In: arXiv preprint arXiv:1406.2199v1 (2014)

[12] Bilal, S., Akmeliawati, R., El Salami, M.J., Shafie, A.A.: Vision-based hand posture detection and recognition for sign languagea study. In: 2011 4th International Conference on Mechatronics (ICOM), pp. 1–6. IEEE (2011)

[13] Molchanov, P., Gupta, S., Kim, K., Kautz, J.: Hand gesture recognition with 3d convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–7 (2015)

[14] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).

[15]) Rafiqul Zaman Khan et.al., proposed “Comparative Study of Hand Gesture Recognition System” ACSIT DOI:10.5121/csit.2012.2320 2012.