# Hand Posture Detection and Classification using You Only Look Once (YOLO v2) Object Detector

Upadrasta Tanmaie[1], Ch.Srinivasa Rao[2]

1.PG scholar, JNTUK-UCEV, Vizianagaram, 535003, India

2. Professor, JNTUK-UCEV, Vizianagaram, 535003,India

**Abstract— Hand gestures can be used to play an important role for establishing Human-Computer Interaction interface (HCI) in modern techniques. Direct use of hand as input device is an attractive method for providing natural HCI. Human gestures can substitute the use of mouse and keyboard as inputs to computer. Gesture commands can be used to control computers and other intelligent machines. The proposed work aims to implement a hand posture detection method based on You Only Look Once (YOLO v2) approach for detecting and classifying static hand postures. YOLO v2 object detection uses a single stage object detection network and is faster than other two stage deep learning object detectors like Faster R-CNN. For training and testing YOLO v2, NUS Hand Pose Dataset II is used. This dataset consists of 2000 hand images and 750 hand images with noise. Results are evaluated by calculating the average accuracy and average testing time per image. The proposed system achieved a good performance for static hand postures.**

**Keywords—Deep Learning, NUS Hand Pose Dataset II, Feature Extraction, Anchor boxes, K-means clustering**

## I. INTRODUCTION

Communication can be broadly categorized into verbal and non-verbal communication. Gesture is a form of non-verbal communication in which movement of hand, face or other parts of body communicate a particular message. Hand is one of the most commonly used body parts to make gestures for establishing communication [1]. Hand Gesture Recognition is a well-researched topic in machine learning, deep learning and image processing. Hand Gestures can be static or dynamic hand gestures. Static hand gestures are hand postures which deals with still images and dynamic hand gestures deals with image sequences. In the proposed work, the focus is on detecting and classifying 10 different static hand postures from the NUS Hand posture-II dataset.

There are several researches works on hand posture recognition, the oldest approach for hand gesture recognition uses hand gloves with sensors or wires, markers (eg LED) or any other devices for recognition [1] [2]. These approaches give a good accuracy only when the lighting conditions are stable otherwise the classification of hand is very difficult problem. After deep learning techniques were proposed, CNN became more popular technique for object detection and classification task which replaces the traditional methods. In [3], PavloMolchanov et al. used 3D convolutional neural networks for recognizing dynamic hand gesture from a video sequence. In this the hand localization and recognition were done using multi-resolution sliding window approach. In this approach each region in image is selected and evaluated for hand gesture. Their method achieved a classification rate of 77.5% on the VIVA challenge dataset. CNN approaches need a lot of regions to predict accurately, hence it has high computation time.

Nowadays Region based convolutional neural networks [4] becoming trending in detection works. RCNN[5], Fast RCNN[6] and Faster RCNN [7],[8] were built for object localization. Recently You Only Look Once(YOLO) [9], [10], [11] also introduced for localization of area of interest. Girshick*et al.* [12] proposes region-based method for object localization because the area of interest is in various scales and ratios. In [12], author used selective search [13] algorithm to generate region of proposals extract around large number of regions from each and every image. It takes more time to evaluate the process because each and every region should send to the CNN (Convolutional Neural Network) separately. Same author [6] proposes Fast RCNN to solve the difficulties of RCNN like each and every image is given as input only one time to the CNN and generated the feature maps by using selective-search algorithm.Ren *et al.* did small modification to the extension of Fast R-CNN to reduce the time taken to detect that is called Faster R-CNN [7]. In which procedure is same as Fast R-CNN but in place of using selective-search method, author used RPN is used to extract the proposals. Soe and Naing [2] used Faster R-CNN method along with caffe frame work to localize the hand posture in clean background. In which author used boosting approach to localize the hand posture. Pisharady*et al.* [14] proposed the segmentation method to detect the hand posture and achieved 93 % accuracy using NUS dataset. Redmon *et al.* [11] proposes You Only Look Once(YOLO) to localize area of interest. It employs a completely different approach. The full image is applied to a single neural net. This network splits and anticipates bounding boxes as well as probabilities. Those bounding boxes were assessed by the probabilities estimated. deng*et al.* [15] proposed rotation estimation by using CNN for localizing hand region and Yuan Li *et al.* [16] designed deep attention network for hand gesture localization.

All the region-based methods are two stage networks whereas YOLO is a single stage network which is faster and smaller compared to two stage networks can detect and classify hand postures at once.In this paper YOLO v2 network is used which can detect hand postures in human noise backgrounds with different lighting conditions while training with hand images without any noisy background.

## II. PROPOSED ALGORITHM

YOLO network can see the entire image and uses the features from the image to draw the bounding box. It can predict all the bounding boxes for all the classes present in an image simultaneously. This means our network reasons globally about entire image and all the objects in the image. YOLO gives a high-speed performance with reasonable accuracy. YOLO v2 is used in our model for detection and classification of hand postures. The typical architecture of our model is shown in figure 1.

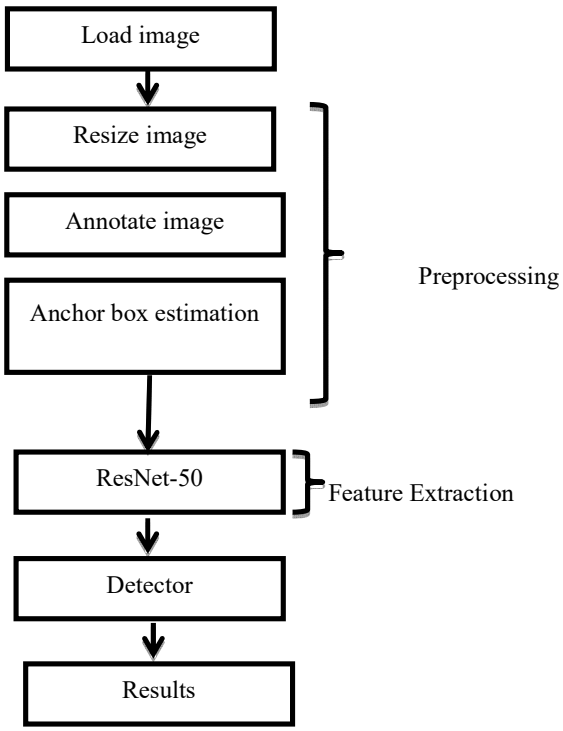


Figure. 1. Architecture of proposed model

Our model architecture is divided into three stages as Preprocessing, Feature Extraction, and Detector as shown in the figure. 1.

*a) Preprocessing:*

Initially all the images are loaded from NUS dataset [18] which are used for training. All the images in NUS datasetare of different sizes, we need to resize the testing images to [224 224 3] as we are using ResNet-50 for feature extraction which can only accept images of size [224 224 3]. For annotating the images to get ground truth images Image Labeler app was used.

Anchor boxes are estimated by using k-mean clustering algorithm[18]. Anchor boxes are a collection of predefined bounding boxes of a specific height and width and are important measurements of deep learning detectors of objects like Faster RCNN and YOLOv2. Anchor box quality affects the prediction speed as well as the accuracy of properly predicting exact position of target object in the frame. Instead of manually predicting the number of anchor boxes and the height and width dimensions as in case of Faster R-CNN, Redom et al. proposed a method for dimensional clustering called k-means clustering where k denote number of anchor boxes. The height and width of the k clustering center boxes are the anchor box dimensions. In our model we have considered k=9, that means we have taken 9 anchor boxes.

*b) Feature Extraction:*

ResNet-50 is used for feature extraction.The main innovation behind using of ResNet is the skip connection [19]. As we know, without adjustments, deep networks often suffer from vanishing gradients, i.e., as the model back propagates; the gradient gets smaller and smaller. Tiny gradients can make learning intractable.

*c) Detector:*

YOLO v2 object detector is used in our model. It analyzes entire image at once, being a very fast at processing time. It divides the input image into g× g grids. If the center of the object falls into the grid, that grid contains an object in it.Each grid

has L bounding boxes and confidence score along with the class. Confidence score determines how confident the box contains an object in it. During training, each box of YOLO v2 calculatesconfidence score as Pr(object) ∗IOU, where IOU is Intersection over union which is the overlap between ground truth and predicted box. Its value is between 0 and 1.The resultant detected image contains bounding box along with the confidence score and the class of the hand detected.

## III. EXPERIMENT AND RESULT

### *A. Dataset*

The proposed work uses NUS Hand Pose Dataset II [17] which is an open source dataset available for academic research purpose. This dataset is a 10-class dataset which consists of 2000 hand images (40 subjects having 5 images per class per subject) with an image size of $160 \times 120$ and 750 hand images with human noise (15 subjects having 5 images per class per subject) with an image size of $320 \times 240$ and 2000 background images. All the hand postures were taken in and around National University of Singapore (NUS). This dataset was selected as it contains hand postures of different sized taken in different complex backgrounds and lighting conditions. In Figure. 2, some of the sample hand postures without human noise and in Figure. 3, some of the sample hand postures with human noise are shown respectively.

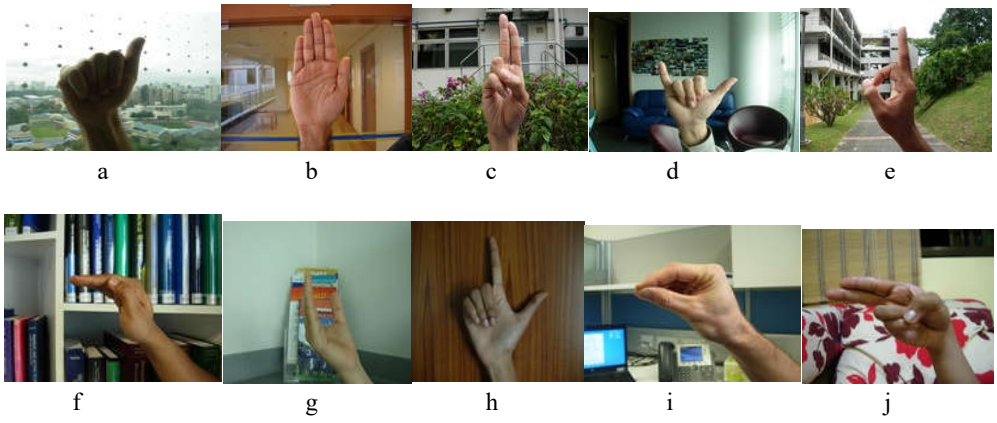


Figure. 2. Sample hand postures taken from NUS hand posture dataset-II showing all classes from 1 to 10 (a to j).

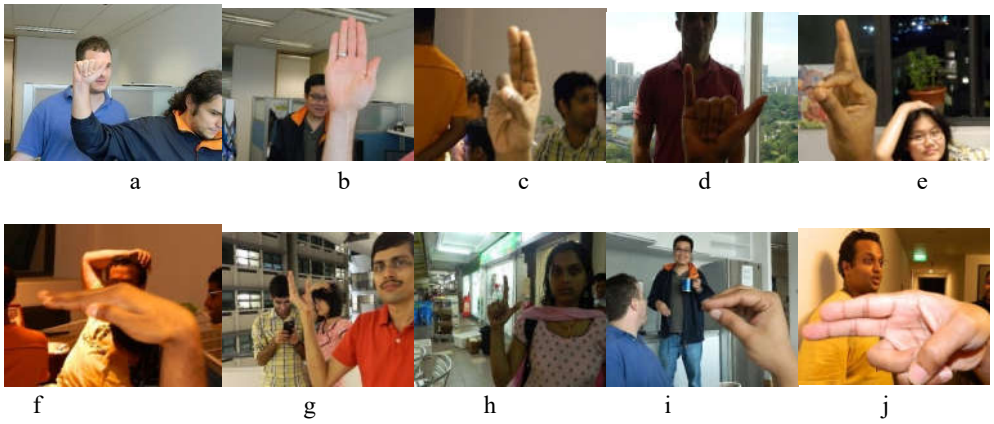


Figure. 3. Sample hand postures with human noise taken from NUS hand posture dataset-II showing all classes from 1 to 10 (a to j).

### *B. Data preparation*

The NUS Hand Posture Dataset-II [17] is divided into training data and testing data for validating the results as shown in Table 1.

TABLE I. NUSHANDPOSEDATASET-II DATAPREPARATIONFORRESULTSEVALUATION

| **Data** | **Consists** |
|---|---|
| Training data | 1800 hand posture color images (180 images per class) |
| Testing data | 200 hand posture color images (20 images per class) and 750 hand posture color images with human noise as background (75 images per class). |

*C. Test results*

For testing, 200 hand postures which were not used for training and 750 hand posture images with human noise in different complex backgrounds are taken from NUS Hand Posture Dataset-II [17]. The test results of hand postures without and with human noise are shown in Figure. 4 and 5 respectively. The detected hand posture is shown with bounding box along with the class detected and the confidence score. The performance evaluation is done by calculating the accuracy of each class along with the average accuracy of all the classes and also the testing time per image is calculated. The evaluated results are tabulated in Table 2. The accuracy for each class and the average accuracy of all classes are calculated by (1) and (2) respectively [2].

$$\text{Accuracy} = (\text{True Positive} + \text{True Negative}) \div \text{Total number of test images} \quad (1)$$

$$\text{Average Accuracy} = \text{Sum of all classes accuracies} \div \text{ total number of classes} \quad (2)$$

From the results we can see that the accuracy of few hand postures decreases as there are few hand postures which were wrongly detected which results in false positives. As there are only 180 images used per class for training which is very less data for the network to detect and classify a hand posture correctly. For getting better accuracy we need to train the network with more data. The average testing time is found to be 2.075 sec per image.

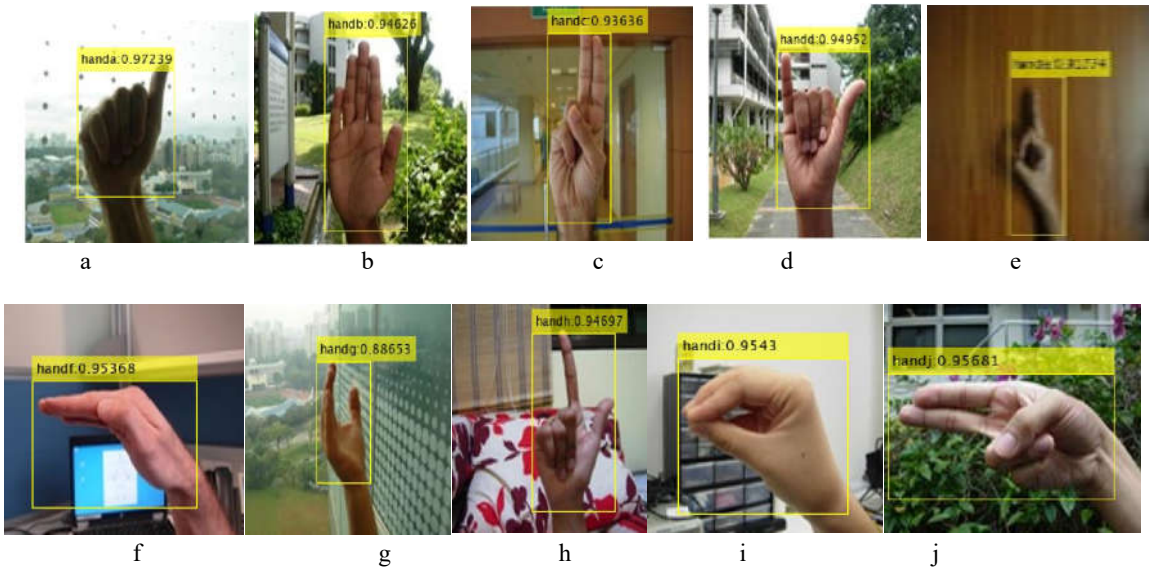


Figure. 4 Sample test results of hand postures taken from NUS hand posture dataset-II showing all classes from 1 to 10 (a to j) along with confidence score.

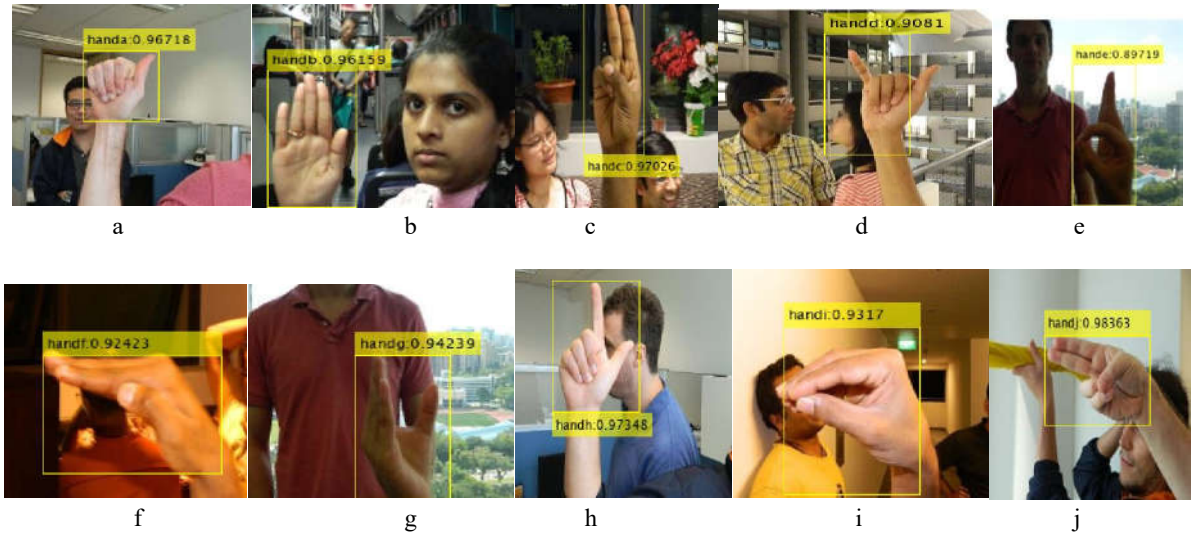


Figure. 4 Sample test results of hand postures taken from NUS hand posture dataset-II showing all classes from 1 to 10 (a to j) along with confidence score.

TABLE II. PERFORMANCEEVALUATIONOFTHEMODEL

| Posture Class | Hand Images from Dataset | Hand Images with Human Noise |
|---|---|---|
| Class a | 100% | 100% |
| Class b | 100% | 100% |
| Class c | 100% | 100% |
| Class d | 100% | 80% |
| Class e | 100% | 98.67% |
| Class f | 100% | 100% |
| Class g | 100% | 98.67% |
| Class h | 100% | 98.67% |
| Class I | 100% | 100% |
| Class J | 100% | 98.667% |
| **Average Accuracy** | **100%** | **97.46%** |

## IV. CONCLUSION

In this work, the main focus was done on hand posture detection and classification in complex backgrounds i.e human noise background and background with different lighting conditions. We trained images of hand postures without noise and observed the test dataof hand postures with noise in different light conditions. The work involved in designing a network by using YOLO v2 for fast detection and to get better accuracy compared to previous detection networks like faster R-CNN.

This work can be extended to control real time applications such as VLC media player, human robot interaction and other Virtual reality applications. The performance can be improved by using more training data and high-speed processor along with GPU. Same work can be executed by using latest techniques like YOLO version 3 or 4.

## REFERENCES


[1] Hernandez-Belmonte, U. H., and Ayala-Ramirez, V., 2016. “Real-time hand posture recognition for human-robot interaction tasks”. *Sensors,* 16(1), p. 36.

[2] Soe, Hsu Mon, and Tin Myint Naing. "Real-time hand pose recognition using faster region-based convolutional neural network." International Conference on Big Data Analysis and Deep Learning Applications. Springer, Singapore, 2018.

[3] 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 (2015)

[4] Simonyan, K., and Zisserman, A., 2014. “Very deep convolutional networks for large-scale image recognition”. *arXiv preprint arXiv:1409.1556*.

[5] Girshick, R., Donahue, J., Darrell, T., and Malik, J., 2014. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580–587.

[6] Girshick, R., 2015. “Fast r-cnn”. In Proceedings of the IEEE international conference on computer vision, pp. 1440–1448.

[7] Ren, S., He, K., Girshick, R., and Sun, J., 2015. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In Advances in neural information processing systems, pp. 91–99.

[8] Abbas, S. M., and Singh, S. N., 2018. “Region-based object detection and classification using faster r-cnn”. In 2018 4th International Conference on Computational Intelligence & Communication Technology (CICT), IEEE, pp. 1–6.

[9] Ni, Z., Chen, J., Sang, N., Gao, C., and Liu, L., 2018. “Light yolo for high-speed gesture recognition”. In 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, pp. 3099–3103.

[10] Tao, J., Wang, H., Zhang, X., Li, X., and Yang, H., 2017. “An object detection system based on yolo in traffic scene”. In 2017 6th International Conference on Computer Science and Network Technology (ICCSNT), IEEE, pp. 315–319.

[11] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A., 2016. “You only look once: Unified, real-time object detection”. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788.

[12] Girshick, R., Donahue, J., Darrell, T., and Malik, J., 2015. “Region-based convolutional networks for accurate object detection and segmentation”. *IEEE transactions on pattern analysis and machineintellsigence,* 38(1), pp. 142–158.

[13] Uijlings, J. R., Van De Sande, K. E., Gevers, T., and Smeulders, A. W., 2013. "Selective search for object recognition". *International journal of computer vision,* 104(2), pp. 154–171.
[14] Pisharady, P. K., Vadakkepat, P., and Loh, A. P., 2013. "Attention based detection and recognition of hand postures against complex backgrounds". *International Journal of Computer Vision,* 101(3), pp. 403–419.
[15] Deng, X., Zhang, Y., Yang, S., Tan, P., Chang, L., Yuan, Y., and Wang, H., 2017. "Joint hand detection and rotation estimation using cnn". *IEEE transactions on image processing,* 27(4), pp. 1888–1900.
[16] Li, Y., Wang, X., Liu, W., and Feng, B., 2018. "Deep attention network for joint hand gesture localization and recognition using static rgb-d images". *Information Sciences,* 441, pp. 66–78.
[17] Pramod Kumar, P., Vadakkepat, P., Poh, L.A.: The NUS hand posture datasets II. ScholarBank@NUS Repository, 11 June 2017
[18] Likas, A., Vlassis, N., and Verbeek, J. J., 2003. "The global k-means clustering algorithm". *Pattern recognition,* 36(2), pp. 451–461.
[19] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A., 2017. "Inception-v4, inception-resnet and the impact of residual connections on learning". In Thirty-first AAAI conference on artificial intelligence.
[20] Targ, S., Almeida, D., and Lyman, K., 2016. "Resnet in resnet: Generalizing residual architectures". *arXiv preprint arXiv:1603.08029*.

**Upadrasta Tanmaie**is pursuing her MTech. in Systems and Signal processing area from University College of Engineering, JNTUK, Vizianagaram, A.P, India. She pursuedherB.TechfromGITAMUniversity,Visakhapatnam,Andhra Pradesh,India.Herresearch interestsinclude one of the Deep Learning Techniques for Hand detection and classification.

**Srinivasa RaoChanamallu** is currently working as a Professor of ECE at JNTUK, University College of Engineering, Vizianagaram, AP, India. He obtained his Ph.D. in Digital Image Processing area from University College of Engineering, JNTUK, Kakinada, A.P, India. He has 26 years of teaching and research experience. He published 65 Research papers in reputed International Journals and Conferences. He is a Fellow of IETE and member of CSI. His research interests include Content based Image and video retrieval, Medical image processing, video water marking and Image forensics. He is also acting as the Reviewer for many reputed International journals.