Fast Object Detection for Quadcopter Drone using Deep Learning

Widodo Budiharto[1], Alexander Agung Santoso Gunawan[1], Jarot S. Suroso[2] and Andry Chowanda[1],
Aurello Patrik[1] and Gaudi Utama[1]
[1]Computer Science Department, School of Computer Science, Bina Nusantara University,
Jakarta, Indonesia 11480
[2]Information Systems Management Department, BINUS Graduate Program, Master of Information Systems Management,
Bina Nusantara University,
Jakarta, Indonesia 11480
Email: jsembodo@binus.edu

***Abstract*—This paper presents our progress in the development of object detection using deep learning for quadcopter drone. Nowadays, drone technology is important for delivering items with minimum time needed. The purpose of this drone is to delivers important aids for patients in emergency situations or items from start to goal position. Object detection can not only tell us what is in an image but also where the object is as well. we use combination of MobileNet and the Single Shot Detector (SSD) framework for fast, efficient deep learning-based method to object detection. The ability for deep learning to detect and localize specific objects and the experimental results using built in camera and stereo camera Minoru are presented.**

Keywords-component; deep learning, drone, CNN, neural network, object detection

I. INTRODUCTION

Deep learning is a fast-growing domain of machine learning and computer vision. It is a class of machine learning algorithms that use a cascade of many layers of nonlinear processing. It also part of the broader machine learning field of learning representations of data facilitating end-to-end optimization. It has the ability to learn multiple levels of representations that correspond to hierarchies of concept abstraction [1]. One of the implementation of deep learning are object detection and recognition that can be used by drone for specific purpose.

Object detection and recognition are crucial in computer vision. Recent advances in object detection are from the success of region proposal methods and region-based convolutional neural networks (R-CNNs). From previous work, we proposed a framework, fast algorithm for object detection using SIFT (Scale Invariant Features Transform) key point detector and FLANN (Fast Library for Approximate Nearest Neighbor) based matcher. Stereo vision is used in order the system knows the position (pose estimation) of the object [2][3], unfortunately it was very slow and not based on deep learning concept.

Backpropagation network has been known for its accuracy because it allows itself to learn and improving itself thus it can achieve higher accuracy. Backpropagation network algorithm was combined with genetic algorithm to achieve both accuracy and training swiftness for recognizing alphabets. Genetic algorithm is used to define the best initial values for the network's architecture and synapses' weight thus within a shorter period of time, the network could achieve the best accuracy [4]. In this paper, we propose a fast object detection using Mobilenet SSD for drone based on quadcopter drone from Parrot. This result will be used by drone for specific purpose in the future. We use OpenCV and popular network architectures compatible with OpenCV 3.3 include GoogleLeNet, ResNet and VGGNet. The quadcopter drone is shown in fig. 1:

Figure 1. Parrot AR. Drone that used in the experiment [10]

II. LITERATURE REVIEW

A. Deep learning

Deep learning is an area of machine learning that emerged from the intersection of Artificial Neural Networks (ANNs), artificial intelligence, graphical modeling, optimization, pattern recognition and signal processing. ANNs are a class of machine learning algorithms that learn from data and specialize in pattern recognition, inspired by the structure and function of the brain. The basic building block is a neuron. A neuron takes a weighted sum of inputs and calculates an activating function. The basic of neural network called perceptron is shown in fig. 2.

Neural networks are usually composed of several layers of interconnected neurons. In the first layer, called the input layer, each neuron corresponds to an input feature (in our case, a pixel). The second layer neurons' inputs are the first layer neurons, third layer neurons' inputs are the second layer neuron. Training a neural network means selecting the best weights for all of the neuron connections. The weights are learned using an algorithm called backpropagation. Layers that have been used in deep learning include hidden layers of an artificial neural network and sets of propositional formulas [10]. They may also include latent variables organized layer-wise in deep generative models such as the nodes in Deep Belief Networks and Deep Boltzmann Machines [6].

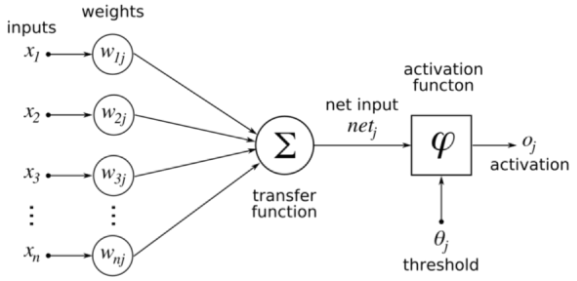


Figure 2. Basic concept of perceptron

Deep learning belongs to the family of ANN algorithms, and in most cases, the two terms can be used interchangeably. SSDs developed by Google, and the algorithm is more straightforward than Faster R-CNNs.When building object detection networks, we normally use an existing network architecture such as VGG or ResNet. We can use MobileNets [12], we call these networks "MobileNets" because they are designed for resource constrained devices such as your smartphone. If we combine both the MobileNet architecture and the Single Shot Detector (SSD) framework, we arrive at a fast, efficient deep learning-based method to object detection. A major contribution of SSD is using default boxes of different scales on different output layers.

Deep neural networks (DNN) categorized as unsupervised, supervised, and hybrid. The unsupervised learning does not use any task specific supervision information in the learning process. It generates meaningful samples by sampling from the networks. Convolutional Neural network (CNN) is a fundamental in deep learning as shown in fig. 3, it is learning filters and applying them to the images. The CNN takes a small square and starts applying it over the image, this square is often referred to as a 'window'. The key component is the convolutional layer. These layers exploit the structure of the data - an image. A convolutional layer connects each output to only a few close inputs, as shown in the illustration above. Intuitively, this means the layer will learn local features. The pooling layer then combines nearby inputs as shown below that has deeper architecture because we able to increase the number of hidden layers, or the number of units, called LeNet [4].

Several papers have proposed ways of using deep networks for predicting object bounding boxes [9]. Some of deep learning-based object detection are Faster R-CNNs [8] and Single Shot Detectors (SSDs) [7]. Faster R-CNNs, the technique is hard to implement, and challenging to train. Even with the "faster" implementation R-CNNs, the algorithm can be quite slow, on the order of 7 FPS. Faster R-CNN, is composed of two modules. The first module is a deep fully convolutional network that proposes regions, and the second module is the Fast R-CNN detector [8].

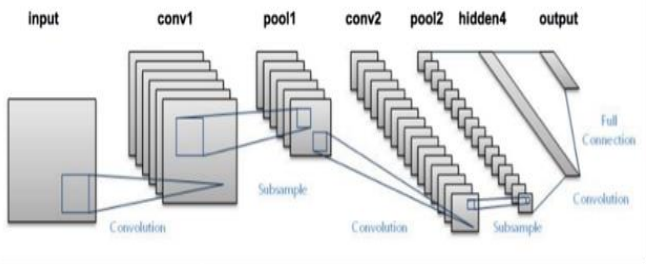


Figure 3. Deep Learning model based on CNN [5]

Deep learning has been significantly developed and improved in the area of computer vision, particularly in object recognition and classification. Referred as the "*the work of the devil*", deep learning methods provide significant improvement to object recognition and classification problem. Deep learning allows the learning architecture to learn the important features that identify the object from a ton of images. Zhou et al [19] implements deep learning for scene recognition problem, the researcher reached the accuracy of 94.42 ± 0.76 % with more than 7 million datasets. Socher et al [18], Krizhevsky, Sutskever & Hinton [17], Qi et al [15], Simonyan & Zisserman [16], and He et al [14] also proposed deep learning method to solve object recognition and classification problem. They trained millions of datasets to recognize hundreds of object classes. Most of them reached more than 80% of accuracy in both 3D and RGB descriptors with their proposed deep learning algorithm. Moreover, He et al [14] claims that the level of accuracy of a system with deep learning, in recognizing objects in ImageNet Classification, is surpassing the human capability. Currently, technology allows us to build and deploy a system with deep learning features in a mobile device (e.g. Tensorflow in Android and CoreML in iOS).

B. GPS

This research used the principle of Global Positioning System (GPS), which is radio navigation system and positioning by using satellite. This system is designed to provide the position and speed of three dimensions and information about time continuously. GPS consists of three

major segments: a segment of space consisting of GPS satellites, control segment segments consisting of monitoring stations and satellite controllers, and user segments consisting of users GPS includes GPS signal and data processing equipment. Figure 4. gives an illustration of how data from GPS is used to control the hexacopter.

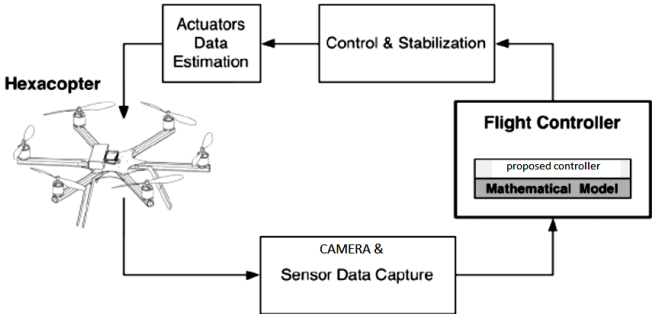


Figure 4. Control of hexacopter with GPS

The GPS receiver will collect signals transmitted from the 24 satellite system. This system is called NAVSTAR. At any time, five GPS receivers are on the surface of the earth. Positioning can be done by trilateration To determine the exact position required three distance values between the GPS signal receiver and GPS satellites. The three distances form a circle on each satellite with the third point of the circle which defines the position of the GPS receiver (Figure 5).

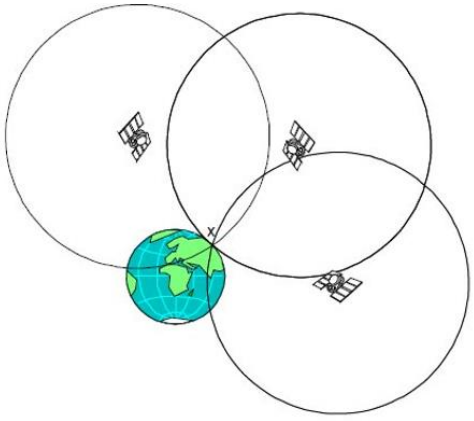


Figure 5. Trilateration Using Three Satellites

III. Proposed Method

A. Architecture of the Drone

Parrot AR Drone as the heart of the quadcopter drone from Parrot.com, includes built-in camera, Processor ARM Cortex A8 1 GHz 32-bit processor with DSP video 800 MHz, Linux 2.6.32, DDR2 1 GB RAM at 200 MHz, accelerometer, High-speed USB 2.0 for extensions, Wi-Fi b g n and 3 axles gyroscope and support Python programming as shown in fig. 6.

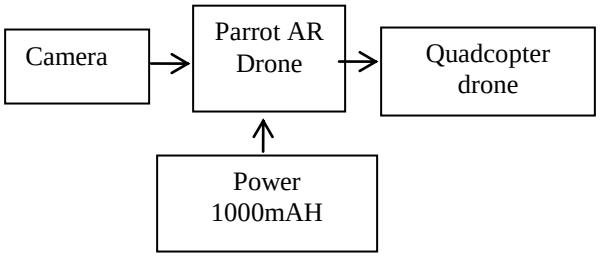


Figure 6. Block diagram of quadcopter drone

B. Object detection

The family of popular object detectors in the deep learning are SSD that implements different activation maps (multiple-scales) for prediction of classes and bounding boxes and YOLO that use a single activation map for prediction of classes and bounding boxes. Using multiple scales helps to achieve a higher mAP *(mean average precision)* by being able to detect objects with different sizes on the image better. SSD only needs an input image and ground truth boxes for each object during training. The MobileNet SSD was first trained on the COCO dataset and was then fine-tuned on PASCAL VOC reaching 72.7% mAP (mean average precision). We can therefore detect 20 objects in images (+1 for the background class), including airplanes, bicycles, birds, boats, bottles, buses, cars, cats, chairs, cows, dining tables, dogs, horses, motorbikes, people, potted plants, sheep, sofas, trains, and TV monitors. First, we train the training images, aster that we got the model and will be used in testing [7].

The SSD training objective is derived from the MultiBox objective, but is extended to handle multiple object categories and SSD model adds several feature layers to the end of a base network, which predict the offsets to default boxes of different scales and aspect ratios and their associated confidences, the architecture of SSD shown in fig. 7:

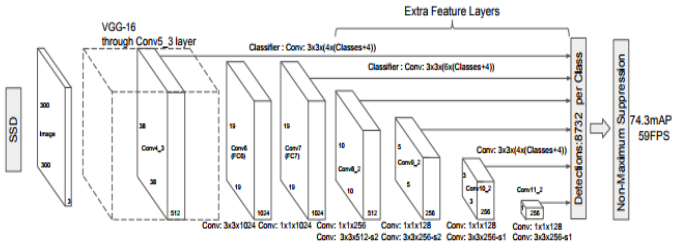


Figure 7. SSD proposed by Liu et al [7]

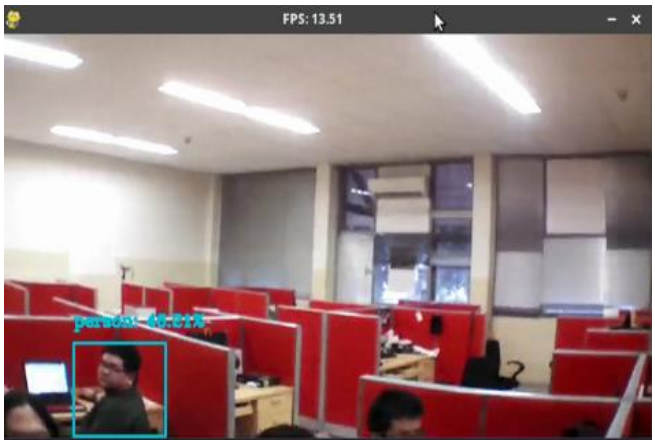
C. Algorithm

First, the program will initialize the list of class labels MobileNet SSD was trained to detect, then generate a set of bounding box colors for each class then we need to load our

model. Then load the input image and construct an input blob for the image by resizing to a fixed 300x300 pixels and then normalizing it. The program also checks the confidence (i.e., probability) associated with each detection [13]

IV. Experimental Result

We use OpenCV3.3 and Python 2.7 for the experiment. We use some training and testing images and the result shown in fig. 8-9. For streaming video, the average is about 14.50 FPS and using Stereo camera Minoru only 6 FPS as shown in figure 3. It is superior comparing with YOLO. The Parrot AR Drone programmed using Parrot AR Drone client library for Python and Ubuntu 16 in order to be able to fly from start to goal position. The drone's property image contains always the latest image from the camera. The drone's property navdata contains always the latest navdata.



(a)

(b)

Figure 8. SSD Object detector could detect object from the drone (a) and successfully detect all of object (b)

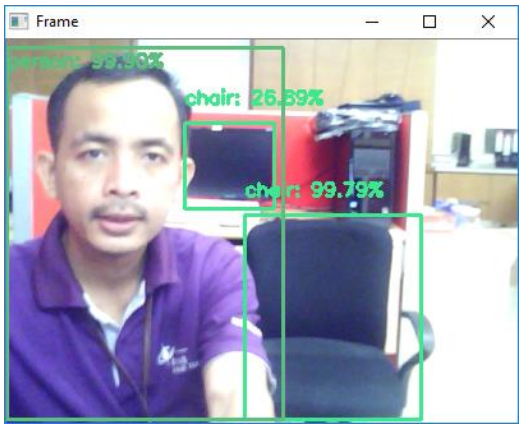


Figure 9. SSD Object detector in 6FPS using Stereo camera Minoru

V. Conclusion

This paper presents the implementation of deep learning technology and MobileNet SSD Detector for object detection that can be used in quadcopter drone. Our method using MobileNet SSD Detector can be used as object detector with high-accuracy detection with average about 14.50 FPS and using stereo camera Minoru only 6 FPS. The resulting system is interactive and engaging and we able to control the Parrot AR Drone easily with low specification in hardware. Moreover, the Parrot AR Drone also is able to correctly recognize the common objects such as person, desk, or chair with high accuracy.

For the future work, the Parrot AR Drone will be deployed to an outdoor environment along with the improved features for object recognition. As for the object recognition, the drone will be equipped with higher accuracy to recognize the objects as well as more object to be recognized. Moreover, this research also aims to create an autonomous drone where the object/scene recognition/detection helps the drone to make a decision in where or how to move.

Acknowledgment

We say thank you very much for Bina Nusantara University for supporting this research based on BINUS PUB grant 2017.

References

[1] Y. Bengio, A. Courville, P. Vincent, "Representation Learning: A Review and New Perspectives". IEEE Transactions on Pattern Analysis and Machine Intelligence. 35 (8): 1798–1828., 2013, doi:10.1561/2200000006.

[2] W., Budiharto, "Robust vision-based detection and grasping object for manipulator using SIFT keypoint detector, International Conference on Advanced Mechatronic Systems (ICAMechS 2014), Japan, pp. 448-452, 2014.

[3] H. Yeremia, N.A. Yuwono, P. Raymond, W. Budiharto, "Genetic algorithm and neural network for optical character recognition", Journal of Computer Science, pp. 1435-1442, 2013.

[4] Deep Learning, accessed at https://towardsdatascience.com/deep-learning-2-f81ebe632d5c

[5] https://blog.dataiku.com/deep-learning-with-dss

[6] Y. Bengio, "Learning Deep Architectures for AI" (PDF). Foundations and Trends in Machine Learning. 2 (1): 1–127, 2009, doi:10.1561/2200000006.

[7] W. Liu, et al, "SSD: Single Shot MultiBox Detector", Computer Vision and Pattern Recognition, 2015.

[8] Girshcik et al, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". NIPS, 2015.

[9] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in Neural Information Processing Systems (NIPS), 2013.

[10] Parrot AR Drone, accessed at https://www.parrot.com/global/drones

[11] www.opencv.com

[12] A. G. Howard et al, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications" accessed at https://arxiv.org/abs/1704.04861

[13] https://www.pyimagesearch.com/2017/09/11/object-detection-with-deep-learning-and-opencv/

[14] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision (pp. 1026-1034).

[15] Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. Proc. Computer Vision and Pattern Recognition (CVPR), IEEE, 1(2), 4.

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

[17] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[18] Socher, R., Huval, B., Bath, B., Manning, C. D., & Ng, A. Y. (2012). Convolutional-recursive deep learning for 3d object classification. In Advances in Neural Information Processing Systems (pp. 656-664).

[19] Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. In Advances in neural information processing systems (pp. 487-495).