Fast Object Detection for Quadcopter Drone using Deep Learning

Widodo Budiharto[1], Alexander A S Gunawan[2], Jarot S. Suroso[3], 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]Mathematics Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
[3]Information Systems Management Department, Binus Graduate Program, Bina Nusantara University, Jakarta, Indonesia 11480
e-mail: wbudiharto@binus.edu

***Abstract*—The paper presents our research progress in the development of object detection using deep learning based on drone camera. The grand purpose of our research is to deliver important medical aids for patients in emergency situations. The case can be simplified into delivery of an item from start to the goal position. We will exploit the drone technology for transporting items efficiently. In sending process, our drone must detect the object target, where the items will be delivered. Therefore, we need object detection module that can detect what is in video stream and where the object is by using GPS as well. To implement the module, we use combination of MobileNet and the Single Shot Detector (SSD) framework for fast and efficient deep learning-based method to object detection. The ability of deep learning to detect and localize specific objects is studied by conducting experiments using drone camera and, as comparison, using stereo camera Minoru.**

Keywords-deep learning; object detection; delivery problem; drone; MobileNet; SSD

I. INTRODUCTION

Deep learning is a fast-growing domain of machine learning, mainly for solving problems in 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. Deep learning has ability to learn multiple levels of representations that correspond to hierarchies of concept abstraction [1]. One of the implementation of deep learning are object localization and detection based on video stream. Object localization and detection are crucial in computer vision. Recent advances in object detection are mainly using deep learning such as region-based convolutional neural networks (R-CNNs). From previous work, we used conventional machine learning approach like fast algorithm for object detection using SIFT (Scale Invariant Features Transform) as key point detector [2] and FLANN (Fast Library for Approximate Nearest Neighbor) based matcher [3]. Previously, we made experiments using stereo vision, but unfortunately our past approach was not too accurate and very slow comparing the recent development which using deep learning.

Technically, deep learning is based on the backpropagation algorithm, which is a method for training the weights in neural network. Backpropagation network has been known for its ability to learn from data a3nd improving itself during training process, but its performance depends on the initial values. If backpropagation network algorithm is combined with genetic algorithm, we can achieve higher accuracy by defining the best initial values for the network's architecture [4]. We will use deep learning for object localization and detection in our research.

We will use the drone technology for transporting items efficiently. While in sending process, our drone must localize and detect the object target. Therefore, object detection module is developed based on camera drone. In this paper, we exploit a fast deep-learning framework for object localization and detection, that is: Mobilenet and Single Shot Detector (SSD) based on camera of Parrot quadcopter drone. The quadcopter drone which used in the experiments is shown in fig. 1. We compare the drone results of object localization and detection experiments to stereo camera Minoru. In here, we also make experiments using popular deep learning architectures, that is: GoogleLeNet, ResNet and VGGNet. The initial progress will be used in our future research in delivering medical aids for patients in emergency situations by using drone.

Figure 1. Parrot AR. Drone that used in the experiment [5].

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.

978-1-5386-6350-9/18/$31.00 ©2018 IEEE

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 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. 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].

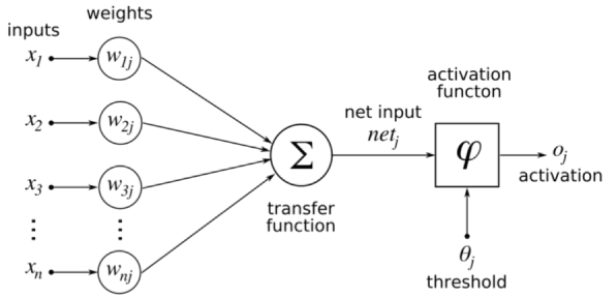

Figure 2. Basic concept of perceptron.

Deep learning belongs to the family of ANN algorithms, and in most cases, the two terms can be used interchangeably. 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 architecture in deep learning, called as LeNet [6] as shown in fig. 3. The CNN takes a small square and starts applying it over the image, this square is often referred to as a window. This key component is named as the convolutional layer, which can capture the structure of 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.

Recently, deep learning has been significantly developed and improved in computer vision, particularly in object recognition and classification. Referred as the black box approach, 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 [7] 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 [8], Krizhevsky, Sutskever & Hinton [9], Qi et al [10], Simonyan & Zisserman [11], and He et al [12] 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 [12] 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).

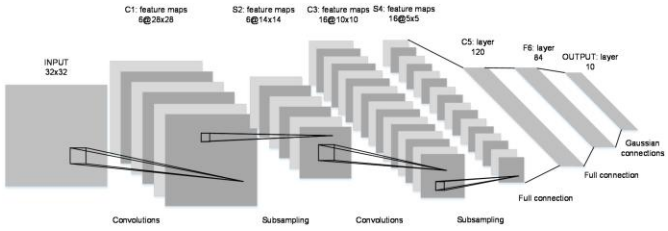

Figure 3. Deep Learning model based on CNN [4].

Several papers have proposed ways of using deep networks for predicting object bounding boxes [13]. Some of deep learning-based object detection are Faster R-CNNs [14] and Single Shot Detectors (SSDs) [15]. Faster R-CNNs 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. Even with the faster implementation R-CNNs, the algorithm can be quite slow, on the order of 7 frame per second (FPS). On the other hand, SSD has much better accuracy and it can achieve 58 FPS on a Nvidia Titan X which outperforming Faster R-CNN model. A major contribution of SSD is using default boxes of different scales on different output layers. Furthermore, there is MobileNets architecture [16]. It is called as MobileNets because it is designed for resource constrained devices such as smartphone. If we combine both the MobileNet architecture and the Single Shot Detector (SSD) framework, we can get at a fast, efficient deep learning-based method to object detection.

B. GPS

For recording the location of object target, this research used the 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. Our drone is equipped with GPS system and it can be used to record geolocation and to improve stability. When flying at high altitude, our drone becomes increasingly difficult to see and stability is therefore essential for controlling it. The GPS receiver calculates drone's position and helps it to remain stable against the wind and then improve drone stability.

III. PROPOSED METHOD

A. *Architecture of the Drone*

Parrot AR Drone [5], as the heart of our drone system, has 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 bgn and 3 axles gyroscope and support Python programming as shown in fig. 4.

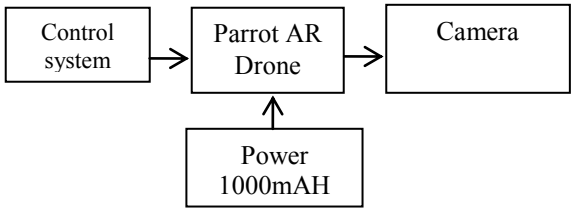

Figure 4. Block diagram of quadcopter drone.

B. *Object Localization and Detection*

The family of popular object detectors in the deep learning are Single Shot Detector (SSD) that use a single activation map for prediction of classes and bounding boxes and Faster R-CNN that implements different activation maps (multiple-scales) 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, after that we got the model and will be used in testing.

The SSD training objective is derived from the MultiBox objective, which 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. 5.

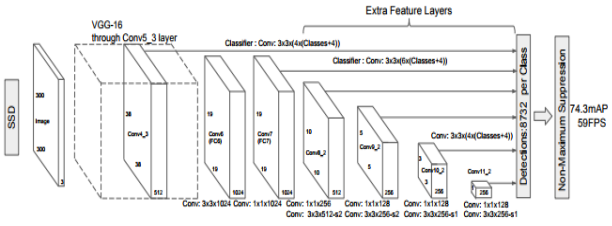

Figure 5. SSD proposed by Liu et al [15].

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.

IV. EXPERIMENTAL RESULT

The Parrot AR Drone programmed using Parrot AR Drone client library based on Python and Ubuntu 16, in order to fly from start to goal position. The drone's property image contains always the latest image from the camera.

To experiment with SSD algorithm, we use OpenCV 3.3 [17] and Python 2.7. The experiment results are shown in fig. 6 for camera drone and Fig. 7 for stereo camera Minoru. The average of SSD processing speed of camera drone is about 14 FPS and processing speed of Stereo camera Minoru is only 6 FPS. In our experiments, SSD algorithm shows to be superior comparing with Faster R-CNN algorithm.

Figure 6. SSD Object detector could detect object from the flying drone

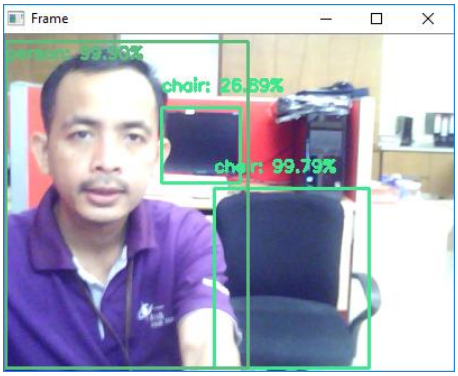

Figure 7. SSD Object detector using Stereo camera Minoru.

V. CONCLUSION

This paper presents the implementation of deep learning technology and MobileNet SSD Detector for object localization and detection that can be fitted in quadcopter drone. Our method using MobileNet SSD Detector can be used as object detector with high-accuracy detection with average about 14 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 can correctly detect the common objects such as person, desk, or chair with high accuracy.

Furthermore, the purpose of the research is to develop an autonomous drone where the object and scene recognition helps the drone to decide in where or how to move. For the future work, the Parrot AR Drone will be deployed to an outdoor environment along with the improved features mainly for object recognition. As for the object localization and detection, the drone will be equipped with object recognition module to decide where the drone will reach the correct target.

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

[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] 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).

[8] 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).

[9] 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).

[10] 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.

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

[12] 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).

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

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

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

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

[17] www.opencv.com