 JAI, vol.1, no.1, pp.1-8, 2019

Intelligent Mobile Drone System Based on Real-Time Object Detection

Chuanlong Li[1, 2], Xingming Sun[1, 2, *] and Junhao Cai[3, *]

Abstract: Drone also known as unmanned aerial vehicle (UAV) has drawn lots of attention in recent years. Quadcopter as one of the most popular drones has great potential in both industrial and academic fields. Quadcopter drones are capable of taking off vertically and flying towards any direction. Traditional researches of drones mainly focus on their mechanical structures and movement control. The aircraft movement is usually controlled by a remote controller manually or the trajectory is pre-programmed with specific algorithms. Consumer drones typically use mobile device together with remote controllers to realize flight control and video transmission. Implementing different functions on mobile devices can result in different behaviors of drones indirectly. With the development of deep learning in computer vision field, commercial drones equipped with camera can be much more intelligent and even realize autonomous flight. In the past, running deep learning based algorithms on mobile devices is highly computational intensive and time consuming. This paper utilizes a novel real-time object detection method and deploys the deep learning model on the modern mobile device to realize autonomous object detection and object tracking of drones.

Keywords: Drone, UAV, CNN, object detection, mobile application, Ios.

1 Introduction

Unmanned aerial vehicle (UAV) is an aircraft without the need of an actual onboard pilot. UAV can be used in various scenarios including military and civil use [Li and Sun (2018); Quinn, Nyhan, Navarro et al. (2018)]. For instance, drones can be used to patrol certain area of interest and execute boarder reconnaissance in military operation. Also, many areas utilize the flexibility of the drone and the camera footage acquired by it to realize 3D reconstruction and autonomous flight [Kim, Liu, Lee et al. (2019)]. However, traditional methods of realizing autonomous flight or object detection usually use simultaneous localization and mapping (SLAM) based algorithms which require lots of sensors. Also, the image-based objection detection or avoidance methods typically require huge

[1] School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, China.

[2] Jiangsu Engineering Centre of Network Monitoring, Nanjing, 210044, China.

[3] School of Politics, Economics and International Relations, University of Reading, Berkshire, RG6 6BG, United Kingdom.

[*] Corresponding Authors: Xingming Sun. Email: sunnudt@163.com.
Junhao Cai. Email: junhao.cai@student.reading.ac.uk.

computational power and complicated algorithms which make it hard to realize. Recent advance in deep learning has proven that convolutional neural network (CNN) is very useful in computer vision related tasks. Many researchers and companies take advantage of the exceptional power of CNN and put the computer vision field to a new era [Nguyen, Arsalan, Kooet al. (2018); Xiang, Zhai, Lv and El Saddik (2018); Rivas, Chamoso, González-Briones et al. (2018); He, Gkioxari, Dollár et al. (2017); Li, Jiang and Cheslyar (2018); Deniz, Vallez, Espinosa-Aranda et al. (2017)]. Object detection and object segmentation are among these successful applications [Girshick, Donahue, Darrell et al. (2014); Girshick, (2015); Ren, He, Girshick et al. (2017); Liu, Anguelov, Erhan et al. (2016)]. Object detection focuses on locating the independent object and the corresponding object class. This is harder than normal image classification tasks for it also has to determine the location of the object and puts a suitable bounding box around the object. The bounding box should contain the detected object with high confidence that the object belongs to certain class. This process has highly computational complexity which makes it hard to use in real-world applications. Furthermore, object detection on mobile video stream requires real-time features while maintaining high detection accuracy. You Only Look Once (YOLO) as the representative of the one-shot object detection method is specialized for high detection speed while still reserves relatively high accuracy which makes it very suitable for real-time applications [Redmon, Divvala, Girshick et al. (2016)]. This paper takes advantage of the flexibility of the modern mobile devices and deep learning based real-time object detection method to realize a mobile system which can automatically detect certain objects from the video stream of drone.

2 Related work

Convolutional neural network (CNN) is a revolution technology in deep learning field, it typically consists of one input layer and one output layer and also includes multiple hidden layers between the input and output layer. The convolutional layer applies a convolution kernel over the input tensor and the convolutional operation is essentially a cross-correlation mathematical operation which can reduce huge amount of network parameters while achieve the same goal as the traditional fully connected network. Krizhevsky et al. [Krizhevsky, Sutskever and Hinton (2012)] first introduced deep convolutional neural network to classify the ImageNet dataset and outclassed the traditional classification methods. A large number of novel applications using CNN sprang up after their successful work [Zhou, Liang, Li et al. (2018)]. CNN is mostly popular used in image classification, image object detection, image object segmentation, image style transfer and so on in computer vision field.

Objection detection is a computer vision task which recognizes the class and the position of the object inside a given image. The detected object is annotated by a rectangle bounding box with certain class identifier. Recent studies have utilized deep learning to realize objection detection. Region-based convolutional neural network (R-CNN) for object detection is the representative of this kind of deep learning methods. It separates the object classification and bounding box prediction by using the region of interest method to detect object [Girshick, Donahue, Darrell et al. (2014)]. Fast R-CNN improves the speed of R-CNN by introducing Region of Interest (ROI) pooling which can share the convolutional layers [Girshick (2015)]. Faster R-CNN further improves Fast R-CNN by introducing

Region Proposal Network and becomes the top framework in several benchmarks [Ren, He, Girshick et al. (2017)]. YOLO stands for another object detection party and differs from the above algorithms. YOLO uses a single convolutional network which predicts the bounding boxes and the class probabilities for the detected boxes simultaneously [Redmon, Divvala, Girshick et al. (2016)]. It works by dividing the input image into several grids and predicts several bounding boxes within each grid. The network outputs the class probability, the confidence score and bounding box offset values for each predicted bounding box. The confidence score shows how likely the detected area contains object and the class probability determines what object is inside the bounding box area.

Quadcopter drones equipped with high resolution camera is a very suitable platform for real-time computer vision task. Štěpán et al. [Štěpán, Krajník, Petrlík et al. (2019)] present computer vision modules of a multi-unmanned aerial vehicle system which runs completely on board in real time. They split the system into two separate tasks. The first one is responsible for finding, tracking, and landing on a human-driven car while the second one focuses on finding small colored objects in a wide area. Furthermore, deep learning models can be deployed on the onboard system or mobile control station to realize autonomous control of the drones. Rivas et al. [Rivas, Chamoso, González-Briones et al. (2018)] use multirotor drones to detect cattle in real-time and employ the artificial intelligence techniques for the analysis of information captured by drones. They use a convolutional neural network to realize cattle detection and implement the system on a PC software. Xiang et al. [Xiang, Zhai, Lv et al. (2018)] use drones for vehicle counting in traffic monitoring. The detector they proposed can handle two situations including static background and moving background. Nguyen et al. [Nguyen, Arsalan, Koo et al. (2018)] propose a method which helps the autonomous landing of drones. They utilize the visible-light-camera to develop a novel remote-marker-based tracking algorithm which can enable the drones to land in heterogeneous areas without GPS signal. They also use convolutional neural network to extract image features automatically and outperform the state-of-the-art objector trackers.

3 System design

The object detection method used in this system is based on YOLO network. YOLO works by using a fixed grid of detectors instead of region proposal-based detectors. It consists of two networks where the first one serves as an image feature extractor and the second one is in charge of the real object detection. The first part is typically trained on the ImageNet dataset using Visual Geometry Group Net (VGGNet), InceptionNet or Residual Neural Network (ResNet) to obtain the best image feature representation. The object detection network predicts the class probabilities, the confidence score of whether or not a certain location contains an object and the bounding box coordinates. The main purpose of using grid cell is to limit the detector to find objects in certain regions so as to improve the accuracy and detection speed. Anchor box is used during the training process to help the detector understand the most common object shapes. The network is mainly comprised of these following basic convolutional blocks: $Conv \rightarrow ReLu \rightarrow MaxPool$. And the last couple of layers only contain the $Conv \rightarrow ReLu$ blocks to generate the final bounding box grid cells. The grid cells are further processed to obtain the final object detection information.

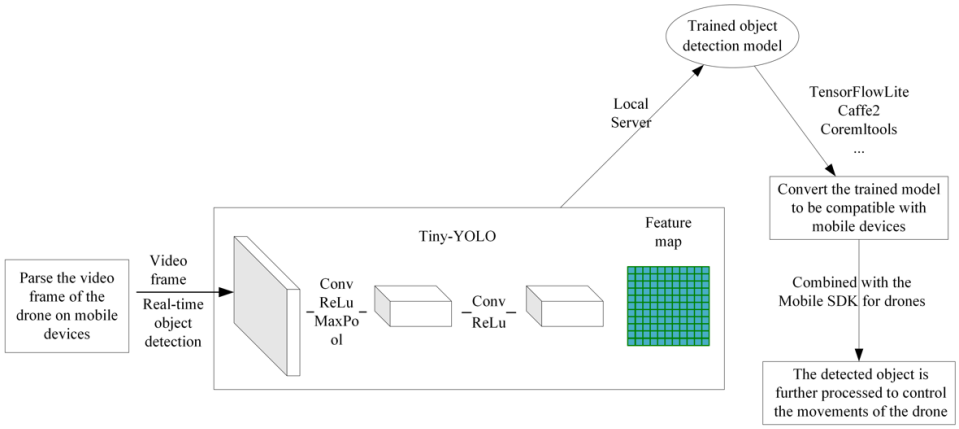

Figure 1: System framework

As Fig. 1 shows, the mobile system works by parsing the video frame of the drone camera first. And in order to achieve low video latency as the system processing the frame images, the object detection method is required to have a high detection speed. Tiny-YOLO is recommended in this circumstance which can achieve high detection speed while ensure a relative high detection accuracy. The object detection model is typically trained on a local server using a specific dataset such as (Common Objects in Context) COCO dataset. The trained object detection model is then converted to another format in order to be compatible with mobile devices. The conversion tools include TensorFlowLite, Caffe2, CoreMLTools, etc. The converted mobile version model is then deployed on the modern mobile devices while combining the mobile SDKs for drones. In this case, all the object detection process happens directly on the mobile device which can reduce the consumption of the battery of drones. The system can further process the detected objects to control the movements of the drone.

4 Implementation details

The proposed system is implemented on a modern iOS device to further demonstrate the feasibility. The object detection model is converted to CoreML format in order to deploy the model on the iOS device. CoreML is introduced by Apple company to enable better machine learning features on mobile devices. The mobile system receives the live video stream and displays it on the screen while the trained model is running at the background to detect known objects. The video stream is decoded into iOS CVPixelBuffer frame in order to apply the object detection algorithm. The bounding box rendering position is further calculated according to the anchor boxes and screen resolution. After detecting the object, user can choose whether to follow certain moving object or not. The detailed implementation architecture is shown as Fig. 1.

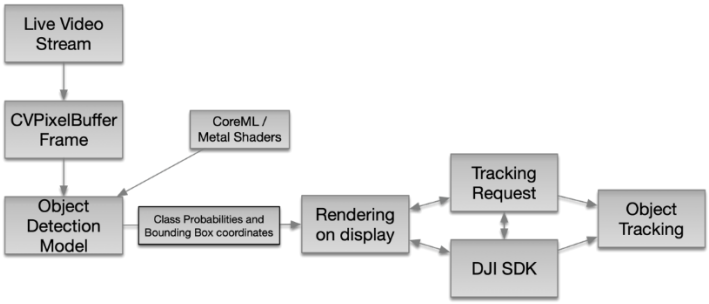

Figure 2: Detailed implementation architecture

Fig. 2 shows the object detection example result and Fig. 3 shows the target tracking result. Note that the full YOLO network is relatively big considering the computational power of the mobile phone. When the system detects certain object, the previous detected bounding box can be slightly lagging behind sometimes as shown in Fig. 2.

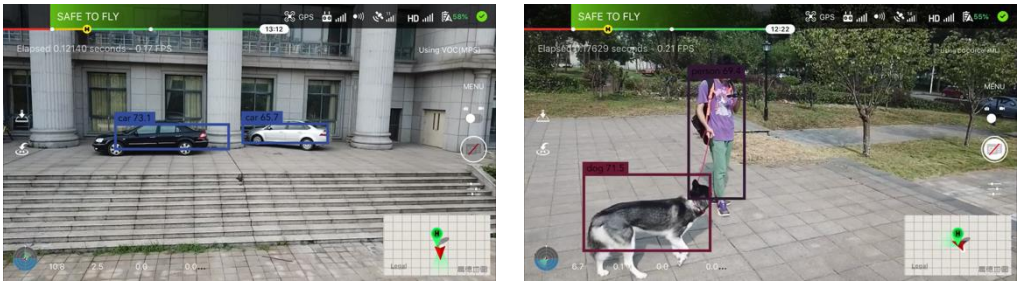

Figure 3: Drone object detection using Full-YOLO model

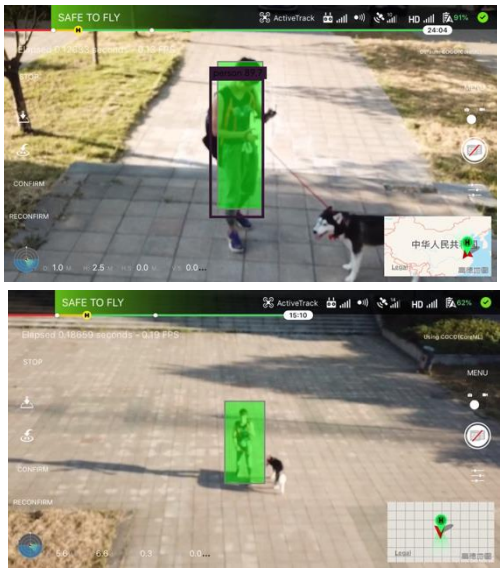

Figure 4: Object tracking

Fast-YOLO (Tiny-YOLO) is a refined object detection model which resembles Full-YOLO network but specializes for real-time applications. The entire Tiny-YOLO architecture has fewer $Conv \rightarrow ReLu \rightarrow MaxPool$ blocks but similar architecture as the Full-YOLO. It shows much lesser bounding box latency than the full model one but also results in lower detection accuracy. As shown in Fig. 4, the object position can be detected with high deviation or even some objects are not detected at all.

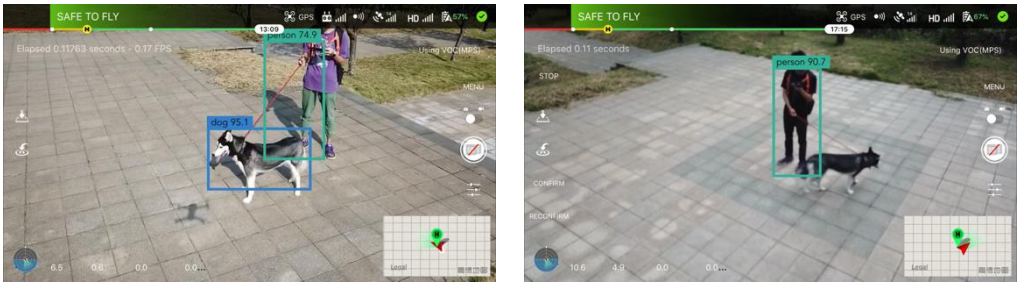

Figure 5: Tiny-YOLO detection

The testing mobile device for this experiment is an Apple iPhone 7 plus which could also be one of the potential reasons the algorithm runs slow. According to the official announcement of the Apple company, the newest device should run the same model with a 10x times faster speed. The drone used in the experiment is a DJI Mavic Pro and the mobile system is written in swift language using iOS SDK and DJI Mobile SDK. The object detection network is first trained on a local computer using COCO dataset [Lin, Maire, Belongie et al. (2014)]. The learning rate is set to 0.001 and the network is trained using one Nvidia GTX1080 GPU for 10000 steps. All the training images are reformatted to 416x416x3 shape with basic data augmentation technique.

5 Citations

This paper applies the object detection algorithm using convolutional neural network to mobile drone applications. The live video transmitted from the main camera is shown on the mobile device, and the video stream is decoded to separated frames in order to send to the YOLO object detection network to obtain the object bounding box and class prediction. It also implements the feature that user can track certain detected moving objects. This system can further enable drones to realize autonomous flight control, autonomous object tracking and obstacle avoidance. Also, by implementing the instance segmentation method such as Mask R-CNN, the mobile drone system will be able to detect the instance contour and track objects more accurately. Furthermore, there are some privacy and security issues that need to be considered in the future.

Acknowledgement: This work is supported by the National Key R&D Program of China under grant 2018YFB1003205; by the National Natural Science Foundation of China under grant U1836208, U1536206, U1836110, 61602253, 61672294; by the Jiangsu Basic Research Programs-Natural Science Foundation under grant numbers BK20181407; by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD)

fund; by the Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET) fund, China.

References

Deniz, O.; Vallez, N.; Espinosa-Aranda, J.; Rico-Saavedra, J.; Parra-Patino, J. et al. (2017): Eyes of things. *Sensors*, vol. 17, no. 5, pp. 1173.

Girshick, R. (2015): Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440-1448.

Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. (2014): Rich feature hierarchies for accurate object detection and semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580-587.

He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. (2017): Mask R-CNN. *IEEE International Conference on Computer Vision*, pp. 2980-2988.

Kim, D.; Liu, M.; Lee, S.; Kamat, V. R. (2019): Remote proximity monitoring between mobile construction resources using camera-mounted UAVs. *Automation in Construction*, vol. 99, pp. 168-182.

Krizhevsky, A.; Sutskever, I.; Hinton, G. E. (2012): Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, pp. 1097-1105.

Li, C.; Jiang, Y.; Cheslyar, M. (2018): Embedding image through generated intermediate medium using deep convolutional generative adversarial network. *Computers, Materials & Continua*, vol. 56, no. 2, pp. 313-324.

Li, C.; Sun, X. (2018): A novel meteorological sensor data acquisition approach based on unmanned aerial vehicle. *International Journal of Sensor Networks*, vol. 28, no. 2, pp. 80-88.

Lin, T.; Maire, M.; Belongie, S.; Hays, J.; Perona, P. et al. (2014): Microsoft COCO: common objects in context. *European Conference on Computer Vision*, vol. 8693, pp. 740-755.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S. et al. (2016): SSD: single shot multiBox detector. *European Conference on Computer Vision*, pp. 21-37.

Nguyen, P. H.; Arsalan, M.; Koo, J. H.; Naqvi, R. A.; Truong, N. Q. et al. (2018): LightDenseYOLO: a fast and accurate marker tracker for autonomous UAV landing by visible light camera sensor on drone. *Sensors*, vol. 18, no. 6, pp. 1703.

Quinn, J. A.; Nyhan, M. M.; Navarro, C.; Coluccia, D.; Bromley, L. et al. (2018): Humanitarian applications of machine learning with remote-sensing data: review and case study in refugee settlement mapping. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2128.

Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. (2016): You only look once: unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779-788.

Ren, S.; He, K.; Girshick, R.; Sun, J. (2017): Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 6, pp. 1137-1149.

Rivas, A.; Chamoso, P.; González-Briones, A.; Corchado, J. (2018): Detection of cattle using drones and convolutional neural networks. *Sensors*, vol. 18, no. 7, pp. 2048.

Štěpán, P.; Krajník, T.; Petrlík, M.; Saska, M. (2019): Vision techniques for on-board detection, following, and mapping of moving targets. *Journal of Field Robotics*, vol. 36, no. 1, pp. 252-269.

Xiang, X.; Zhai, M.; Lv, N.; El Saddik, A. (2018): Vehicle counting based on vehicle detection and tracking from aerial videos. *Sensors*, vol. 18, no. 8, pp. 2560.

Zhou, S.; Liang, W.; Li, J.; Kim, J. (2018): Improved VGG model for road traffic sign recognition. *Computers, Materials & Continua*, vol. 57, no. 1, pp. 11-24.