Machine Learning Capstone Project

Project: Real time person and vehicle detection in UAV systems
Jun Zhang
November 18th, 2017

Domain Background

In recent years, autonomous UAVs have been widely deployed for inspection, surveillance, search-and-rescue, traffic monitoring and infrastructure inspection[1-3]. As UAV application becomes widespread, a higher level of autonomy is required to ensure the safety and operational efficiency. To achieve this, the real-time visual object detector is necessary in UAV system. In this capstone project we will focus on person and vehicle detection in UAV system.

Problem Statement

Visual object detection is a classic problem in the computer vision world. However, it is even more challenging to perform those tasks from UAV images due to the top-down view, distortion due to UAV motion and real-time requirement with limited hardware resource. On the other hand, the objects interested in UAV system are different compared to images taken from cameras on ground. There are a few other work either focusing on a single object detection such as pedestrian detection [7] or vehicle detection [10]. Those two kinds of objects are particularly interesting in UAV applications. In this project, we only focus on person and vehicle detection at the same time.
Usually, a object detector will be evaluated with metric mean average precision (mAP)[6]. However, since there are only two objections, we will simply will evaluate precision and recall for each object as our detection accuracy. In addition, the UAV system typically is equipped with limited hardware, detection, detection accuracy will be not our solo target. We will only evaluate the speed, hardware consumption etc to have a more thorough evaluation.

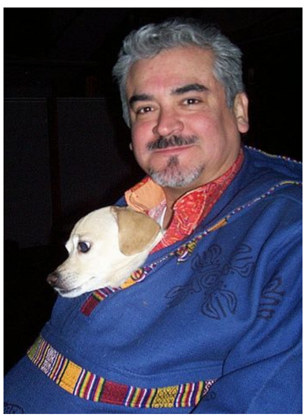
Datasets and Inputs

We are going to use two dataset Pascal Visual Object Classes (VOC)[6] and UAV123[11]. There are around 20000 images in VOC dataset, which are downloaded from flickr and most of images are captured by handheld device and images are typically large. We use this dataset as there are a lot of benchmark results from other literature on this dataset so we can easy to compare the detector we design with other know detector.

For VOC dataset, both 2007 and 2017 data will be used. The images of this dataset are downloaded from flickr and this dataset has 20 objects. For each object of one image, the label and the bounding box information (xmin, ymin, xmax, ymax), which are the coordinates of the top-left and bottom-right corners is provided.

In this capstone project, we are only interested in person and certain vehicles (‘bus’, ‘car’, ‘train’). Therefore, we have to do some preprocessing. We remove all the images which does not have ‘person’, ‘bus’, ‘car’, ‘train’ labels first, then we remap the ‘bus’, ‘car’, ‘train’ to ‘vehicle’ label.

The UAV123 is a very new dataset for visual object tracking released in 2016. This dataset contains labeled images from a UAV camera. Below are two images from the two dataset. The left side image is from VOC 2007 dataset and the right one is from UAV123 dataset. One obvious difference is the objects in VOC is significantly larger compared to objects on a UAV123 dataset.




This dataset has 123 annotated 1280x720 video sequences captured from a low-altitude aerial perspective. The video is stored as a sequence of jpeg images. It has tracking labels for various kind of single objects such as ‘bike’, ‘bird’, ‘boat’, ‘car’, ‘building’, ‘trunk’ etc and we only keep the video with ‘person’ and ‘car’ and change the label ‘car’ to ‘vehicle’. One problem with this dataset is not all the objects are labeled. To use this dataset for object detection purpose, we only keep the images which only have labeled object. In addition, images from a video clip typically have large amount of spatial redundancy. If use all the images, we might end up overfitting. To deal with this issue, we only randomly select around 30% of the selected images as training data and randomly select another 30% from it for validation and testing purpose.

Solution Statement

In this project, we are going to adapt YoLo into UAV systems and will propose a few necessary modifications to improve the accuracy and reduce the hardware resource consumption on a TX2 platform. We will use the original pre-trained YoLo-v2 on COCO dataset[8] as our benchmark and will evaluate the accuracy, speed and hardware consumption on TX2 platform.

To improve the the detection accuracy for UAV objects, we will borrow some idea from SSD method and will add another extra scale of feature from earlier ConvNet layer output for detection. Therefore, total three scale of features will be used for detection compared 2 scale features of original YoLov2.

To reduce the hardware resource consumption, we will not use the original Darknet19 as backbone feature extractor. Instead, we will use MobileNet[14] as the feature extraction layers. Also, we are going to design a shallow MobileNet so that the final feature map will not be too small. This could potentially increase the detection performance for small objects.

References

[1] M. Bhaskaranand and J. D. Gibson, “Low-complexity video encoding for uav reconnaissance and surveillance,” in Military Communications Conference (MILCOM). IEEE, 2011, pp. 1633–1638.

[2] T. Tomic, K. Schmid, P. Lutz, A. Domel, M. Kassecker, E. Mair,I. L. Grixa, F. Ruess, M. Suppa, and D. Burschka, “Toward a fully autonomous uav: Research platform for indoor and outdoor urban search and rescue,” IEEE robotics & automation magazine, vol. 19, no. 3, pp.46–56, 2012.

[3] Andriluka, M.; Schnitzspan, P.; Meyer, J.; Kohlbrecher, S.; Petersen, K.; Von Stryk, O.; Roth, S.; Schiele, B. Vision based victim detection from unmanned aerial vehicles. In Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Taipei, Taiwan, 18–22 October 2010;pp. 1740–1747.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. arXiv preprint arXiv:1506.02640, 2015.

[5] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. arXiv preprint arXiv:1612.08242, 2016

[6] Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C., Winn, J., and Zisserman, A. The Pascal visual object classes challenge: A retrospective. IJCV, 111(1):98–136, 2015

[7] T. Huster and N. C. Gale. Deep learning for pedestrian detection in aerial imagery. In MSS Passive Sensors, 2016.

[8] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B.Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. CoRR, abs/1405.0312, 2014

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E.Reed, C. Fu, and A. C. Berg. SSD: single shot multi-box detector. CoRR, abs/1512.02325, 2015.

[10] J Carlet, B Abayowa, Fast Vehicle Detection in Aerial Imagery. arXiv preprint arXiv:1709.08666, 2017 - arxiv.org

[11] M. Mueller, N. Smith, and B. Ghanem, “A benchmark and simulator for uav tracking,” in Proc. of the European Conference on Computer Vision (ECCV), 2016.

[12]S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.

[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan.Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9):1627–1645, 2010. 1, 4

[14] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T., Marco Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.