https://doi.org/10.2352/ISSN.2470-1173.2020.8.IMAWM-084

A New Training Model for Object Detection in Aerial Images

Geng YANG [a], Yu Geng [a]*, Qin LI [a], Jane YOU [b], and Mingpeng Cai [c];
[a] School of Software Engineering, Shenzhen Institute of Information Technology, Shenzhen, China
[b] Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China
[c] Shenzhen Shangda Xinzhi Information Technology Co., Ltd., China
*gengy@sziit.edu.cn

Abstract

This paper presents a new training model for orientation invariant object detection in aerial images by extending a deep learning based RetinaNet which is a single-stage detector based on feature pyramid networks and focal loss for dense object detection. Unlike R3Det which applies feature refinement to handle rotating objects, we proposed further improvement to cope with the densely arranged and class imbalance problems in aerial imaging, on three aspects: 1) All training images are traversed in each iteration instead of only one image in each iteration in order to cover all possibilities; 2) The learning rate is reduced if losses are not reduced; and 3) The learning rate is reduced if losses are not changed. The proposed method was calibrated and validated by comprehensive for performance evaluation and benchmarking. The experiment results demonstrate the significant improvement in comparison with R3Dec approach on the same data set. In addition to the well-known public data set DOTA for benchmarking, a new data set is also established by considering the balance between the training set and testing set. The map of losses which dropped down smoothly without jitter and overfitting also illustrates the advantages of the proposed newmodel.

Introduction

Analysis of aerial images is very useful in applications of intelligent transportation system, land and resources management, national security protection and so on. Aerial images were frequented collected from satellites or unmanned aerial vehicles. Valuable information such as different sizes of objects can be extracted from aerial images using image processing algorithm. Various objects including cars, routes, bridges, planes, ships, buildings and so on may exist in aerial images with different sizes and quantities. In addition, many aerial images were static images that were distinguished from dynamic videos. Therefore, it is hard to utilize traditional moving object detection algorithms, for example, the background subtraction method and the optical flow method.

Moving objects can be obtained by subtracting the latest frame from the background image, which was acquired by a weighted summation of the previous background and this frame[1] or was generated by using the Gaussian mixture model [2, 3]. Besides, moving objects can be detected in the information of a motion that was related to moving pixels in an image[4]. Four kinds of optical flow methods were frequently used in previous studies: the differential based method, the region-based matching method, the energy based method and the phase based method [5]. However, static objects can be found by using feature based methods [6,7,8], which using SVM or Adboost to classify a potential object based on its features regardless it was moving or was static.

Recently, there are many emerging object detection methods based on deep learning models, for instance, Faster-RCNN[9], SSD[10] and yolo[11]. In these methods, potential object proposals can be selected or generated. Nevertheless, ideas of Region Proposal Network, convolution neural network and classification networks are used. Not only moving objects but also static objects can be extracted from video sequences or static images. In the study field of object defections in aerial images, Faster-RCNN, SSD and YOLOv2 were analyzed on a large-scale data set named DOTA in previous studies[12]. Results were good when these methods were qualified and validated on DOTA. However, these methods with their initial training ways were hard to get good results in a new and different data set.

In this paper, an improved training model based on R^3Det[13] is proposed for object detection in aerial Images. It is trained and tested by a new data set provided by the Remote Sensing Image Sparse Representation and Intelligent Analysis Competition[14]. The new data set will be introduced in details in Section 2. In Section 3, results of the proposed method are analyzed and then compared to those of using the traditional training method provided by the re-implemented R^3Det based on open source codes[15]. The result can be improved significantly by using the proposed method compared to the traditional method. Finally, conclusions are given and future studies are indicated in the last section.

Methodology

The original R^3Det model was based on the RetinaNet [16]. The RetinaNet was an one-stage object detector consist of ResNet, feature pyramid net, class subnet and box subnet. The reason to select the R^3Det model is that it is robust to large aspect ratio, densely arranged and category unbalance problems. These problems frequently occurred in aerial images. By adopting this model, these problems can be temporarily ignored so that improvement of the training mode can be focused in this study. However, codes of R^3Det[13] are not yet open source, open source codes of RetinaNet [15] are used in this paper.

There are 18 types of objects in aerial images of the new data set. It is different from DOTA. Comparisons of types between these two data sets are shown in Table 1. The new data set is separated into a training set and a testing set evenly in types.

Table 1. Comparisons of types between DOTA and the new data set

Classifications	DOTA	The new data set
Soccer ball field	√	√
helicopter	√	√
Swimming pool	√	√
Roundabout	√	√
Large vehicle	√	√
Small vehicle	√	√
bridge	√	√
harbor	√	√
Ground track field	√	√
Basketball court	√	√
Tennis court	√	√
Baseball diamond	√	√
Storage tank	√	√
ship	√	√
plane	√	√
airport	X	√
Container crane	X	√

In this paper, the tensorflow framework is implemented. Some parameters such as learning rate, image size, batch size and so will be set in the training progress. Before calculating the accuracy, intersection over union (IoU) is used to define whether a detected object is valid. If IoU is larger than 0.5, the object is valid, otherwise it is invalid. The accuracy will be measured based on valid detected objects. In order to measure the accuracy, Mean Average Precision (MAP) is used to evaluate the performance of the proposed method.

In this paper, the key improvement is the modification of training mode in the R^3Det model. The proposed training method in the training progress is designed as following,

(i) Total number of training images are used instead of only one image in each training iteration.

(ii) The training rate will be reduced by T1 (a numerical value can be set) if the loss is not reduced.

(iii) The training rate will be divided by T2 (a numerical value can be set) if the loss it not changed.

Experiment and Result Analysis

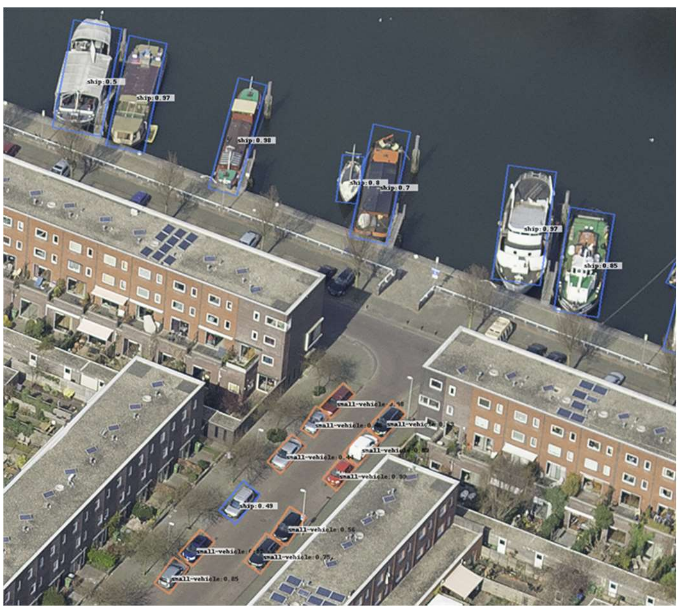
In the experiment, at the first, the R^3Det model is implemented according to its paper and open source codes of RetinaNet. Second, the same parameters and settings of the R^3Det model in its paper are adopted. Third, the original R^3Det model on the new data set is trained and tested to get a baseline result. Finally, the new training mode for the R^3Det model on the same data set is calibrated and then validated to get a result for comparisons.

In the new data set, there are 800 different images for training, while there are 283 images for testing. All these images were annotated by human beings as a ground truth. By applying the same parameters in the paper of the R^3Det model with its training method, the MAP is about 23.1%. By applying the same parameters of the R^3Det model with the proposed training method (T1=T2=10), the MAP is about 57.3%. It is a good improvement for the MAP target if using the proposed training method.

As Fig 1. and Fig. 2 show, very small objects and dense objects can be detected by applying the proposed training mode based on the R^3Det model.

Fig 1. Ships in the sea

(a)



(b)

Figure 2. Dense ships and small vehicles (a) the original size (b) the zoom in size

There are three types of loss in the model: the classification loss, the object detection loss and the total loss that is the summation of the classification loss and the object detection loss. As Fig. 3 depicts, these losses were going down nice-looking when the number of iterations was increasing. As mentioned before, all training images were traversed in each iteration. It can be found that losses were decreased significantly when the number of iterations was increased to 2 and 24. It was noted that after the 24th iteration, a better optimization result was obtained. Therefore, the improved training mode was working and a good result was acquired.

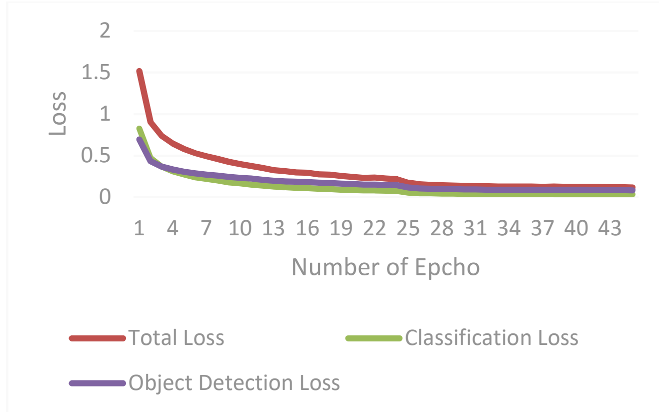


Figure 3. Changes of Losses in the training progress

Conclusion and Future Work

In this paper, an improved training mode based on the R^{3}Det model is proposed. The proposed method was calibrated and validated by utilizing a new aerial image data set. A significant result was obtained in the experiment. The reason to include all 800 images in each iteration is to traverse all possibilities in the training images. Losses might be influenced by these possibilities. If only one image is traversed in each iteration based on the original training mode, losses will be hard to reduce commendably. The reason to dynamically reduce the learning rate is to avoid over-fitting and to obtain a better optimal result. When losses are not changed, the optimal result may be skipped two continuous iterations. As expected, the improved training mode works well. However, there are many analyses not yet carried out. In the future, more analyses will be promoted such as analyses of results of each classification type, analyses of influences of image resolutions, analyses of results if applying different parameters.

Acknowledgement

The authors would like to thank the support from The Hong Kong Polytechnic University, Zhujiang Scholar Scheme with Shenzhen Institute of Information Technology and the Shenzhen Fundamental Research fund under Grant (No. JCYJ20170306100015508).

References

[1] M. Piccardi, "Background subtraction techniques: a review", IEEE International Conference on Systems, Man and Cybernetics,vol. 4, pp. 3099-3104,2004.

[2] P. Kaewtrakulpong and R. Bowden, "Video-based surveillance systems", Springer US, pp. 135-144, 2002.

[3] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction", Proceedings of the 17th IEEE International Conference on Pattern Recognition, 2004, vol. 2, pp. 28-31, 2004.

[4] Z. Sun, G. Bebis and R. Miller, "On-road vehicle detection: a review ", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 5, pp. 694-711, 2006.

[5] J.L. Barron, D.J. Fleet, S.S. Beauchemin and T.A. Burkitt, "Performance of optical flow techniques", Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol.12, no.1, pp. 236-242, 1992.

[6] Puja and Er. Rajput, "Feature extraction in face recognition using SVM-LBP detection technique", International Journal of Innovative Research in Computer and Communication Engineering, vol.4, issue 10, pp.18890-1898, 2016.

[7] X. Wen, L. Shao, W. Fang and Y. Xue, "Efficient feature selection and classification for vehicle detection", IEEE Transactions on Circuits & Systems for Video Technology, vol. 25, no.3, pp. 508-517, 2015.

[8] P. Negri, X. Clady and L. Prevost, "Benchmarking Haar and histograms of oriented gradients features applied to vehicle detection", International Conference on Icinco, pp.359-364, 2007.

[9] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks", IEEE Transactions on Pattern Analysis & Machine Intelligence, vol.39, no.6, pp. 1137-1149, 2017.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu and A. C. Berg, "SSD: single shot multibox detector", arXiv:1512.02325, 2015.

[11] J. Redmon, A. Farhadiar, "YOLOv3: An Incremental Improvement", arXiv:1804.02767,2018.

[12] G. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, L. Zhang, "DOTA: A Large-scale Dataset for Object Detection in Aerial Images", arXiv:1711.10398, 2018.

[13] X. Yang, Q. Liu, J. Yan, A. Li, "R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object", arXiv:1908.05612, 2019.

[14] The Remote Sensing Image Sparse Representation and Intelligent Analysis Competition : http://rscup.bjxintong.com.cn

[15] Open sources code of RetinaNet: https://github.com/DetectionTeamUCAS/RetinaNet_Tensorflow_Rotation

[16] T. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, “Focal Loss for Dense Object Detection”, arXiv:1708.02002, 2017.

[17] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, andS. Belongie. Feature pyramid networks for object detection.In CVPR, 2017.

Author Biography

Geng Yang is currently an assistant professor at the School of Software, Shenzhen Institute of Information Technology, Shenzhen, China. Mr. Yang obtained his Doctor of Engineering degree from the Hong Kong Polytechnic University in 2018, MSc. in Electronic & Information Engineering (EIE, Multimedia Signal Processing and Communications) with Distinction from PolyU in 2010 and BEng in Telecommunications from Xidian University, China in 2009. His research interests include pattern recognition, intelligent transportation system, intelligent fitness system, etc.

Yu Geng received his B.Eng. degree in optical engineering from the Yanshan University, Qinhuangdao, China, in 2006, his master degree in optical engineering from Zhejiang University, Hangzhou, China, in 2008, and his Ph.D. degree in electrical and computer engineering from the Hong Kong University of Science and Technology, Hong Kong, China, in 2015. Now, he is an assistant professor at the School of Software, Shenzhen Institute of Information Technology, Shenzhen, China. His research interests include semiconductor devices simulation and fabrication, image processing, and pattern recognition.

Jane You is currently a professor in the Department of Computing at the Hong Kong Polytechnic University. Prof. You obtained her BEng. in Electronic Engineering from Xi'an Jiaotong University in 1986 and Ph.D in Computer Science from La Trobe University, Australia in 1992. Prof. Jane You has worked extensively in the fields of image processing, medical imaging, computer-aided detection/diagnosis, pattern recognition. So far, she has more than 260 research papers published. Prof. You is also an associate editor of Pattern Recognition and other journals.

Mingpeng Cai received his Diploma degrees in computer science from the Shenzhen Institute of Information Technology in 2017. Since Feb. 2019, he has been working at the Shenzhen Shangda Xinzhi Information Technology Co., Ltd., China. His research interests include image processing and pattern recognition, etc.

Qin Li received his B.Eng. degree in computer science from the China University of Geoscience, Wuhan, China, in 2001, his M.Sc. degree (with distinction) in computing from the University of Northumbria at Newcastle, Newcastle,U.K.,in2003,and his Ph.D. degree in computing from the Hong Kong Polytechnic University, Hong Kong,in 2010. Now, he is an associate professor at the School of Software, Shenzhen Institute of Information Technology, Shenzhen, China. His research interests include medical image analysis, biometrics, image processing, and pattern recognition.