Moving Object Recognition for Airport Ground Surveillance Network

Zhizhuo Zhang, Xiang Zhang(✉), Donghang Chen, and Haifei Yu

Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou 324000, Zhejiang, China
{zhangzhizhuo,chendh,yuhaifei}@std.uestc.edu.cn,
uestchero@uestc.edu.cn

Abstract. In this paper we first introduce an airport ground surveillance network, which is composed of data acquisition terminal based on multiple cameras, data transmission based on high-speed optical fiber, and processing terminal including some airport intelligent applications, e.g. intrusion warning and conflict prediction. Next we present a moving object recognition algorithm named AMORnet which is the basis of the intelligent applications in this surveillance network. Unlike the traditional object detection which cannot distinguish static and moving objects and moving object detection requiring accurate silhouette segmentation, the AMORnet only locate moving object and much faster than the time-consuming segmentation. To achieve this purpose, firstly we estimate the scene background through a motion estimation network, compared to the commonly used temporal histogram based approach, our background estimation method can better cope with infrequent aircraft movements in airports. Secondly, we use feature pyramids to perform regression and classification at multiple levels of feature abstractions. In this way, only moving objects are correctly recognized. Finally, experiments are conducted on an airport ground surveillance benchmark to verify the effectiveness of the proposed AMORnet.

Keywords: Airport ground surveillance · Moving object recognition

1 Introduction

In recent years, the number of passengers carried by civil aviation in the world has continued to increase, the airport structure has become increasingly complex. The difficulty of visual monitoring of ground moving objects in the field by administrators has gradually increased. The safety hazards of manual command and management have become more obvious. To satisfy the need for automated scene surveillance, the airport surveillance network that cover the entire airport field area are used in modern airport.

This work was supported by the Project of Quzhou Municipal Government (2020D011), and National Science Foundation of China (U1733111, U19A2052).

C. T. Calafate et al. (Eds.): MONAMI 2021, LNICST 418, pp. 335–343, 2022.
https://doi.org/10.1007/978-3-030-94763-7_25

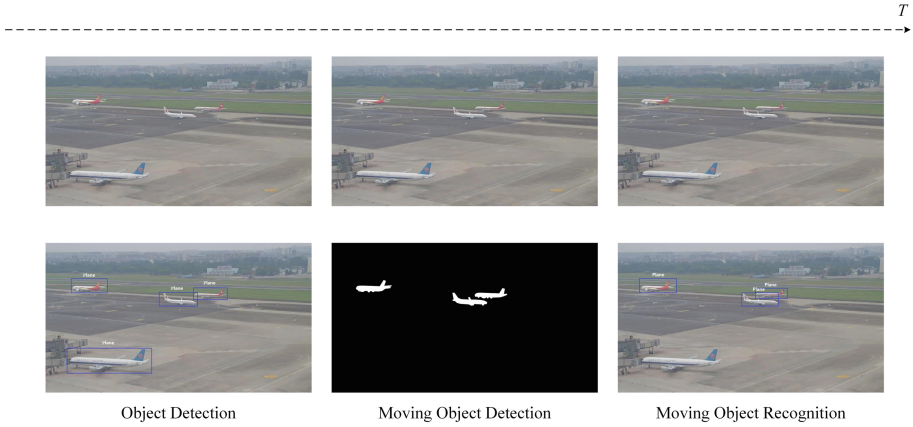

Fig. 1. Difference between the three tasks: object detection, moving object detection, moving object recognition.

The Airport ground surveillance network has access to a large amount of real-time video information, on which we can develop intelligent applications such as intrusion warning, conflict prediction, etc. Moving object detection and object detection are the foundation of many video based intelligent applications, but are subject to some limitations when applied in airport ground surveillance networks. Moving object detection aims to present and label foreground objects that undergo spatial position changes in the image sequence or video. Moving object detection algorithm use few input video frames to estimate an initially clean background image, and then the pixel-wise segmentation is carried out between estimated background and input video frames. Traditional approaches are unsupervised and have poor performance because of lighting and shadows in the airport ground [1–4]. The supervised moving object detection algorithm s based on convolutional neural networks (CNNs) [5–8] has good accuracy in AGVS benchmark [9], but the algorithms cannot distinguish the class of the target and are time-consuming. In addition, the objective of object detection is to find objects with different geometries as well as to assign an accurate label to each detected object. Object detection algorithms can be divided into two approaches: two-stage [10–12] and one-stage [13–15], and it is often believed that the former works slower, but the detection accuracy is higher. However, existing object detection algorithms cannot distinguish static and moving objects, which are not intuitive enough in the airport ground surveillance network.

Based on moving object detection and object detection, we further implement moving object recognition. The difference between object detection, moving object detection and moving object recognition is given in Fig. 1. Compared to the first two methods, the moving object recognition localization and classification of moving objects which are more suitable for the development of intelligent applications for airport ground surveillance networks. The rest of the paper is organized as follows: the airport ground surveillance network is briefly

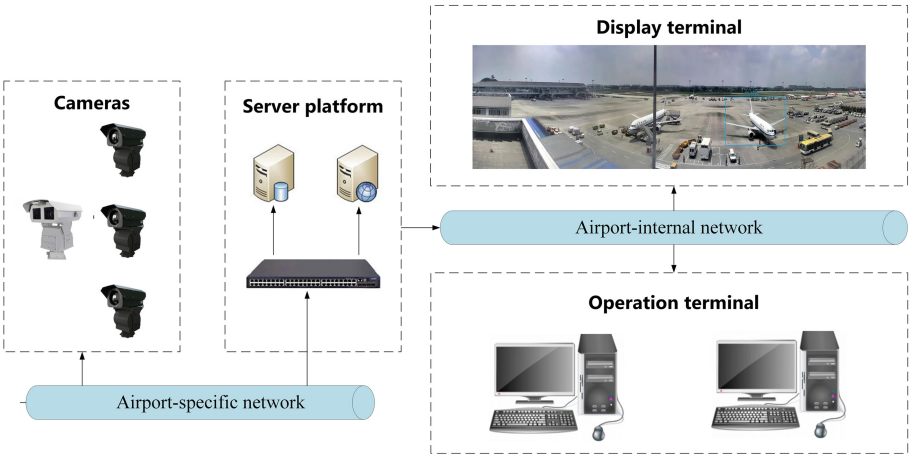

Fig. 2. Structure of airport ground surveillance network.

described in Sect. 2. Section 3 presents the details of the proposed moving object recognition algorithm. Experimental results are given in Sect. 4, followed by the conclusion in Sect. 5.

2 The Airport Ground Surveillance Network

The airport ground surveillance network is shown in Fig. 2. It consists of four parts: video surveillance front-end, server platform, display terminal, and operation terminal. The video surveillance front-end consists of multiple fixed cameras and pan-tilt cameras, which is because the airport field is very wide and needs to overlap multiple cameras field of view to realize the video surveillance of the whole airport, the format of the captured video is transmitted to the server platform in the form of H264 through the airport-specific network, the server platform will stitch and store the video according to the geometric relationship of the cameras, and transmit it to the display terminal and the operation terminal through the airport-internal network. The display terminal can display the real-time airport scenes processed by the operator terminal.

The proposed moving object recognition algorithm can be applied at the operation terminal, as shown in Fig. 2, where the display terminal shows the result of the algorithm proposed in this paper, the aircraft on the right side of the screen has just entered the ground without stopping, while the aircraft on the left side has stayed for some time, and the algorithm is able to detect the moving object separately. Based on the moving object detection algorithm, we can also develop higher-level applications. For example, we automatically send conflict prediction to staff when objects are too close to each other; we can also train classes of objects that may have trespassing, such as private cars, drones, etc., and automatic alarm when the target is in view. We also present the initial implementation of conflict prediction in the next session d.

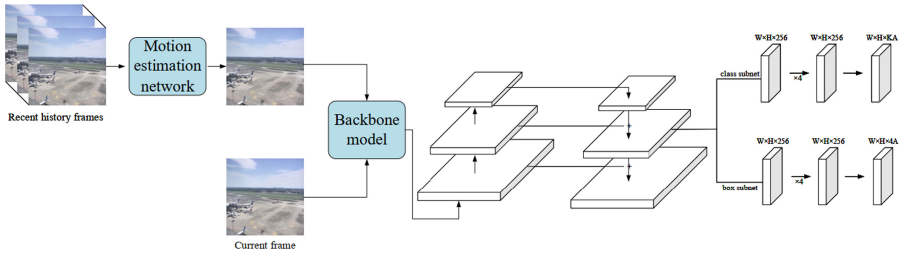

Fig. 3. Schematic illustration of the proposed AMORnet.

3 Moving Object Recognition

Our proposed method is called AMORnet (Airport moving object recognition). We propose a motion estimation network based on the moving object detection method, and a regression and classification block based on the object detection method, which together form AMORnet. The whole AMORnet is shown in Fig. 3, and we will discuss the functions of each module in the following subsections.

3.1 Motion Estimation Network

We use the method of moving object detection to realize the estimation of motion. And the key to this is how to get a clean and accurate background image. Recently, temporal histogram based approach for background estimation which is proposed [16]. At each pixel location, temporal histogram is obtained using Eqs. (1) and (2).

$$Hist_{(m,n)}(l) = \sum_{t=1}^{N} f(I(m,n,t),l); l \in [0,255] \tag{1}$$

$$f(x,y) = \begin{cases} 1 & x = y \\ 0 & else \end{cases} \tag{2}$$

From estimated pixel-level temporal histogram, accurate background pixel intensity at particular location (m, n) is obtained using Eq. (3).

$$BG_n(m,n) = \arg\max_{l}(Hist_{(m,n)}(l)); l \in [0,255] \tag{3}$$

where, argmax(.) is maximum value of the histogram bin index. The estimated background results of pixel-level temporal histogram based approach are illustrated in Fig. 5(b). However, Due to the slow speed of the moving objects in the airport scene and the frequent motion and stillness of the objects. Temporal histogram based approach can't yield good results. On this basis, we propose motion estimation network based on Convolutional Neural Network, the network structure is shown in Fig. 4. The background is learned through a sequence

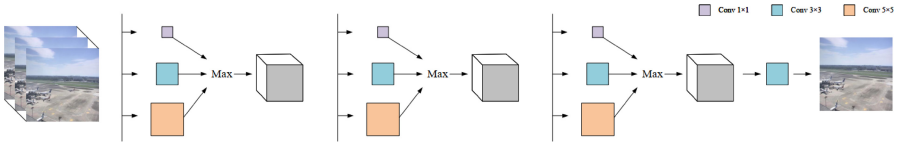

Fig. 4. Structure of motion estimation network, where each convolutional layer is followed by Relu as the activation function

of multi-scale receptive feature blocks using recent temporal history. Each stage captures the maximum response from multiple receptive fields of size 1×1, 3×3 and 5×5. This allows us to obtain background statistical information while still ensuring the adaptability of the network to different changing scenarios. The visualization results of the proposed motion estimation network are shown in Fig. 5.

Fig. 5. Comparison of proposed motion estimation network with temporal histogram based approach: (a) Current frame (b) Temporal histogram based approach (c) Motion estimation network

3.2 Regression and Classification Blocks

Regression and classification blocks are composed of a backbone network and two task-specific subnetworks, as shown in Fig. 3. The backbone is responsible for computing a convolutional feature map over the entire input images. We adopt the Feature Pyramid Network (FPN) from [16] as the backbone network.

FPN augments a standard convolutional network with a top-down pathway and lateral connections, the established multi-scale feature pyramids can be used to detect objects of different scales. At each level of pyramidal feature map, we take an input feature map with 256 channels and set 9 anchors, two task-specific subnetworks are used for regression and classification respectively, the layers follow similar configurations as in [15].

3.3 Conflict Prediction

We set when there are two bounding box between the IOU threshold of 0.7 or more will automatically send a conflict prediction to the relevant staff, the diagrammatic sketch is shown in Fig. 6, 16th. But due to the filming perspective, there may be a crisscross on the video between aircraft that are far away from each other, and the implemented algorithm is not completely accurate, there are often false alarms, so to achieve accurate conflict prediction also need to add auxiliary information of airports such as ADS-B.

4 Experimental Results

4.1 Experiment Settings

AMORnet takes two tensors of shape $608 \times 608 \times N$ (past temporal history) and $608 \times 608 \times 3$ (current frame) as input and returns the spatial coordinates of the moving object with class label. The regression uses smooth L1 loss and the classification use focal loss, the sum of which constitutes the training loss. The loss gradients are backpropagated through motion estimation network as well. Training is performed with batch size = 1 over Nvidia RTX 2060 GPU. We use adam optimizer with initial learning rate set to $1 \times 10^{-}5$.

4.2 Dataset

Due to the lack of available benchmark datasets with labeled bounding boxes for airport moving object recognition, based on the AGVS benchmark [9], we manually annotated 20 of the 25 segments. The first 10 videos are used for training and the last 10 videos are used for testing.

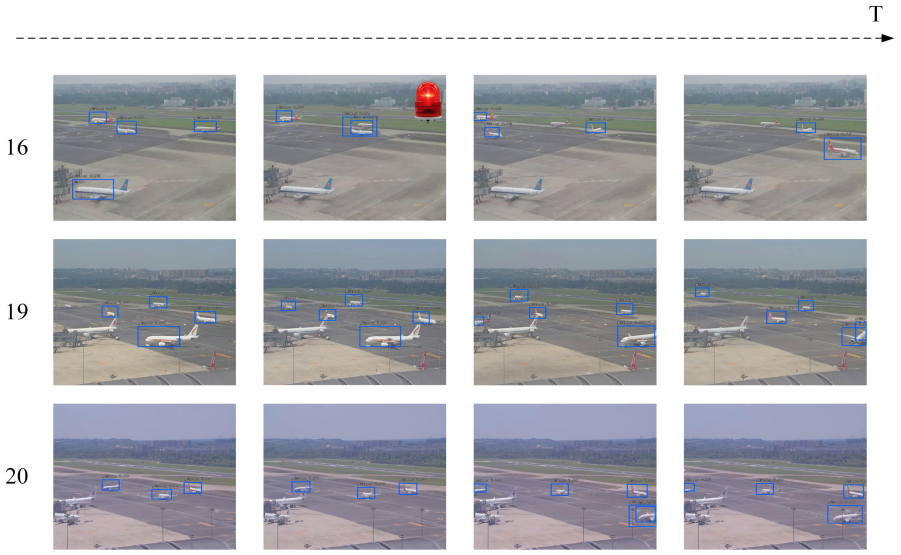

Fig. 6. Qualitative results of our method for 16th,19th and 20th from AGVS benchmark.

4.3 Evaluation Metrics

Since the experimental results are similar to those of objection detection, we use mAP, which is commonly used for object detection, as the evaluation metrics. The mAP is used to measure the average of the Precision at different Recall. A predicted object instance is predicted to be True Positives if the predicted object instance has at least 50% crosslinking with the corresponding ground truth object instance.

Table 1. Algorithm performance under different input past temporal history frames

Depth\mAP	14	16	19	20	Overall
20	56.5	73.5	67.9	82.4	70.0
30	59.2	76.2	**73.7**	82.2	72.8
50	**61.2**	76.3	71.4	**83.2**	**73.0**
70	51.0	**76.9**	69.9	82.9	70.1

4.4 Qualitative Analysis

The detection results of our method on the AGVS dataset are shown in Table 1, which we can conclude that the best result is when the number of historical

Fig. 7. Comparison of proposed method with Faster RCNN: (a) Current frame (b) Faster RCNN (c) AMORnet

frames input by Motion estimation network is 50. Among them, the 16th video has the process of moving the aircraft from motion to static, the 19th video has occlusion between the aircrafts, and the 20th video the aircraft is incomplete in the image. The algorithm in these three video detection results are shown in Fig. 6, and the comparison result with the Faster RCNN [12] detection results is shown in Fig. 7, from the figure it can be seen that our method can not only distinguish between stationary and moving objects in all these scenes, but also fully detect occluded or defective objects.

5 Conclusion

An airport ground surveillance network was introduced in this paper. It consists of four parts: video surveillance front-end, server platform, display terminal, and operation terminal. The proposed moving object recognition algorithm based on object detection and moving object detection can be applied to the airport ground surveillance network. The proposed method AMORnet first obtains a relatively accurate background through the motion estimation network, and then implements the moving object detection through the classification and

regression network. Experimental results on the AGVS benchmark demonstrated the effectiveness of the proposed method.

References

1. Stauffer, C., Grimson, E.: Adaptive background mixture models for real-time tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, October 1999
2. Kim, K., Chalidabhongse, T., Harwood, D., Davis, L.: Background modeling and subtraction by codebook construction. In: IEEE International Conference on Image Processing, October 2004
3. Zivkovic, Z.: Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recogn. Lett. **27**(7), 773–780 (2006)
4. Barnich, O., Droogenbroeck, M.V.: ViBe: a powerful random technique to estimate the background in video sequences. In: International Conference on Acoustics, Speech, and Signal Processing, April 2009
5. Lim, L., Keles, H.: Foreground segmentation using a triplet convolutional neural network for multiscale feature encoding. Pattern Recogn. Lett. **112**, 256–262 (2018)
6. Liao, J., Guo, G., Yan, Y., Wang, H.: Multiscale cascaded scene-specific convolutional neural networks for background subtraction. In: Hong, R., Cheng, W.-H., Yamasaki, T., Wang, M., Ngo, C.-W. (eds.) PCM 2018. LNCS, vol. 11164, pp. 524–533. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00776-8_48
7. Tezcan, M.O., Ishwar, P., Konrad, J.: BSUV-Net: a fully-convolutional neural network for background subtraction of unseen videos. In: IEEE Winter Conference on Applications of Computer Vision, March 2020
8. Bakkay, M., Rashwan, H., Salmane, H., Khoudour, L., Puig, D., Ruichek, Y.: BSCGAN: deep background subtraction with conditional generative adversarial networks. In: IEEE International Conference on Image Processing, October 2018
9. www.agvs-caac.com
10. Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. Int. J. Comput. Vis. **4**(2), 154–171 (2013)
11. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, June 2014
12. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017)
13. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, June 2016
14. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
15. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. **42**(2), 318–327 (2017)
16. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, July 2017