TRACKING-BASED MOVING OBJECT DETECTION

Hao Shen[1], *Shuxiao Li*[1], *Jinglan Zhang*[2], *Hongxing Chang*[1]

[1]Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]Queensland University of Technology, Brisbane, Australia

ABSTRACT

We present a novel approach for multi-object detection in aerial videos based on tracking. The proposed method mainly involves three steps. Firstly, the spatial-temporal saliency is employed to detect moving objects. Secondly, the detected objects are tracked by mean shift in the subsequent frames. Finally, the saliency results are fused with the weight map generated by tracking to get refined detection results, and in turn the modified detection results are used to update the tracking models. The proposed algorithm is evaluated on VIVID aerial videos, and the results show that our approach can reliably detect moving objects even in challenging situations. Meanwhile, the proposed method can process videos in real time, without the effect of time delay.

Index Terms—moving object detection, aerial video, tracking, real-time video processing, detection-by-tracking

1. INTRODUCTION

Moving object detection is one of the most challenging problems in computer vision, especially for the aerial videos, which are captured by the Unmanned Aerial Vehicles (UAV). In recent years, the UAV is a growing field with military and civilian applications, including surveillance, rescue, and reconnaissance. In contrast to applications with fixed cameras, such as traffic monitoring and building surveillance, aerial surveillance has the advantages of higher mobility and larger surveillance scope. Meanwhile, more challenges are involved in aerial videos, such as smaller object size, motion blur and camera motion. Therefore, much attention has been paid to moving object detection in aerial videos.

Much work has been done on moving object detection in aerial videos. The existing methods can be roughly classified into two categories. The first category is based on image processing techniques, and the second one is based on machine learning approaches. For the first category, Yin[1] used the forward-backward motion history image, which is obtained by accumulating the image differences, to detect moving objects. Reilly[2] proposed a method for human detection in aerial videos based on a number of geometric constraints that are obtained from the metadata. For the second category, the researchers tried to learn the patterns of object appearance by using various kinds of features. For example, Benedek[3] have learned a three-layer Markov Random Field (MRF) model to detect the moving regions. Rosenbaum[4] used gentle AdaBoost with Haar-like features to generate a confidence image. Cheng[5] has tried to segment the aerial surveillance video using a Mixture of Experts.

Although the detection methods have been shown to be able to locate moving objects even in complex scenes, the false positives have remained frequent and the detection results are sensitive to occlusions and registration errors. On the other hand, the tracking methods have good ability to find a particular object in image sequences but they accumulate error during runtime(drift). Therefore, to achieve a more reliable performance, some researchers have tried to combine the detection and tracking results. Ali[6] proposed to use motion and appearance contexts to achieve persistent tracking of objects in aerial imagery. Koppen[7] used the detection results to build a graph structure, and tracking is reformulated as a heuristic search for optimal paths. Li [8] used the observation result to resample the particle filters. In [9], the authors proposed a framework for tracking multiple targets, where the candidate detection results for individual frames are used to recover trajectories of targets. However, As pointed out by Garcia-Martin et al [10], most of the methods that combine detection and tracking are designed mainly with the aim of improving tracking results (tracking-by-detection), and the improvements introduced in the detection task are a byproduct of the tracking task (detection-by-tracking). Few research has tried to improve explicitly the detection result using the tracking history. In [10], the authors used the tracking information to extrapolate the intermittent people detection results and remove the detector errors. Kalal[11] proposed a novel tracking-learning-detection(TLD) method, that explicitly decomposed the long-term tracking task into tracking, learning and detection. The learning estimates the detector's errors and updates it to avoid errors.

The main goal of this research is to develop a robust algorithm to detect small moving objects in aerial videos. To overcome the shortcomings that exist in the pure detection methods, we introduce a novel tracking-based moving object detection method. The main contributions of this paper include:

(1)The targets are detected in a hierarchical manner. First, the candidate motion regions are segmented using motion history and trajectory prediction; then the exact target location is detected based on spatial-temporal saliency.

(2)The tracking information is explored in both the prediction stage and the result fusion stage. In the prediction stage, the previous tracking results are used to predict the candidate object location; in the result fusion stage, the tracking results are combined with the detection results to get the optimized detection results.

(3)The detected results are classified as temporary and formal targets respectively to avoid the environmental noise, meanwhile, the static model and the dynamic model are used to give a full representation of the targets.

2. TRACKING-BASED MOVING OBJECT DETECTION

The overall flowchart of the proposed tracking-based detection algorithm is presented in Fig.1. The method mainly consists of two parts: detection and tracking. By combining the detection and tracking results we can get the optimized detection results.

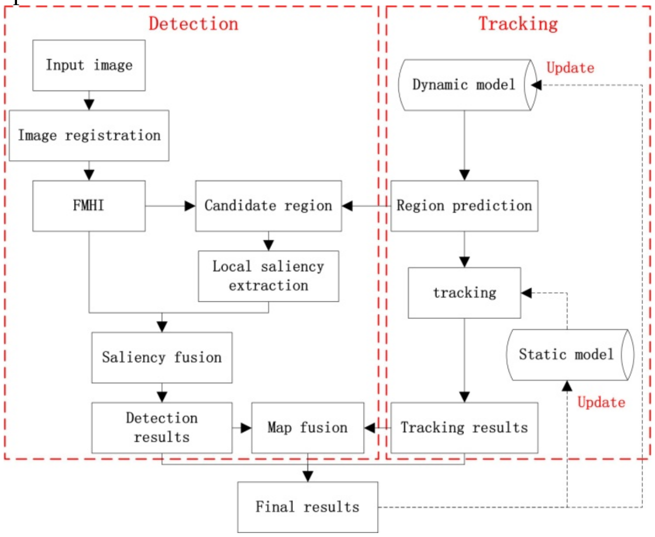

Fig.1.Overall flow chart of the proposed tracking-based detection method.

2.1. Detection

For aerial videos, the camera platform is moving, so the image registration is implemented based on the feature matching method before motion detection to compensate the camera motion. Then, the Motion History Image(MHI)[1] is used to get the candidate motion regions. Besides, in order to avoid the effect of time delay that caused by backward MHI, we only used the Forward MHI (FMHI). As the FMHI only accumulates the information in the past, which results in lacking of future prediction, especially when the target is moving fast, the result of FMHI is usually unreliable. In this paper we involve the tracking trajectory information to complement this. To make more robust prediction for irregular motions as well as regular motions we try to combine several directions and magnitudes of the previous trajectories to predict the target position in the current frame. Let $\mathbf{x}_t^i$ denotes the position of *i*th predicted result at time *t*, X_{t-1} denotes the position at time *t-1*, and v^i denotes the velocity at time *t-i*. The prediction can be formulated as follows.

$$\mathbf{x}_t^i = \mathbf{x}_{t-1} + \mathbf{v}^i * T \qquad (1)$$

where *T* denotes the interval of time step, here we set T=1. Typical segmentation results are shown in Fig.2. From the results we can see that, when the targets move slowly (the left image), both FMHI and the prediction results works well, however, when the targets move fast (the right image), the results obtained by FMHI only cover parts of the targets, but the prediction results can still work well. So when the tracking trajectory is available, the candidate object regions are obtained by fusing the results of trajectory prediction and FMHI. Otherwise, the FMHI results are used directly as the candidate regions.

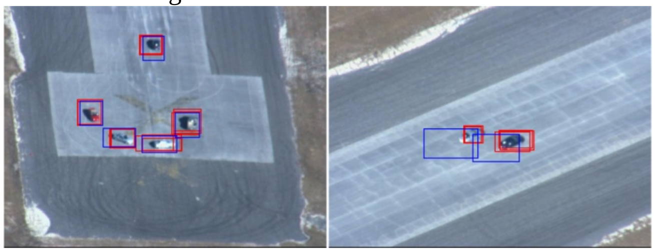

Fig. 2. Segmentation results of FMHI and prediction. The FMHI results are marked by rectangles with blue color, and the results obtained by prediction are marked by rectangles with red color.

After the coarse segmentation, the spatial saliency is extracted to get the object's appearance details in candidate motion regions. Here various spatial saliency methods can be used to detect the object's appearance details. For the sake of simplicity, the pixel-level saliency[12] and region based contrast[13] are used to detect the object in the candidate motion region. Finally, by linear combining the FMHI and spatial saliency, we can get the detection results shown in Fig.3.

Considering the noise that comes from the environment, the detection results are classified as temporary and formal targets respectively. The reliable detection results that appear many times are considered as formal targets. The objects that appear less than a certain number of times will be recognized as temporary targets. When the number of appearances reaches the predefined threshhold, the temporary targets will be changed to formal targets. In our system, only the formal targets are tracked in order to reduce the time cost and avoid the noise effect.

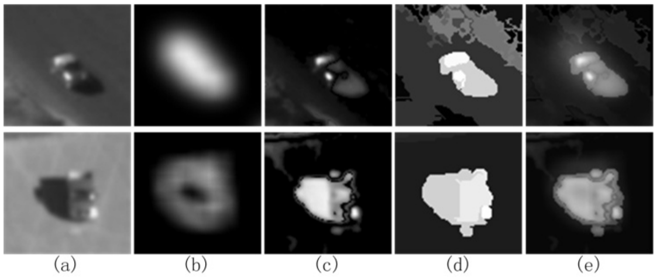

Fig. 3. The results of the detection procedure. (a)the original candidate local image;(b)result of FMHI; (c)result of pixel-level saliency; (d)result of region based contrast; (e)fusion result

2.2. Tracking

As the detection algorithms estimate the object location in every frame independently, the false alarms and missing detections are unavoidable. In this paper we try to deal with this by combining tracking information with detection. Here tracking serves two purposes: the first is to provide target trajectories for prediction (as illustrated in section 2.1); the second is to refine the detection results.

2.2.1. Tracker introduction

To achieve a reliable tracking result, we used the tracking method based on foreground and background ratio[14]. The discriminative tracking features are selected online from linear combinations of RGB space based on variance ratio, then the selected features are used to compute the weight image, and the mean-shift process is adopted to get the tracking result.

Each detected object is represented by a dynamic model *D* and a static model *S*. For the dynamic model *D*, we adopt the velocity and position to describe the object each time. The dynamic model is mainly used for position prediction. The static model *S* is defined as the selected features which are obtained by tracking. The static model is used as the tracking template, as well as a measure to associate the results of detection and tracking.

2.2.2. Birth/death move

The birth/death move consists of adding/removing a tracker from the current trackers. For birth move, once a detected object is recognized as a formal target, a new tracker will be created, and initialized with the object appearance. For death move, a tracker will be terminated if it experiences the no-association situation (including the case of tracking failure) for a long time span, or the tracking result is beyond the boundaries of the image.

2.3. Result association

Different from traditional association methods[8][9] that only consider the trajectory information, we also explore the tracking confidence level as an external evidence to modify the detection results. The detection and tracking results consist of two parts: spatial description and weight image(confidence level). Here the spatial description is denoted as $\boldsymbol{S}=(x,y,s)$, in which (x,y) is the position, s is the size of the target.

First of all, the detection and tracking results are associated based on the proximity of spatial description. A tracking result is associated with a detection result only if both the spatial distance and size difference are smaller than the thresholds. We assume that one tracking result can only be associated with one detection result, so only the most alike pair is selected as a successful association.

After setting up the association between the detection and tracking results, the associated tracking results are used to modify the detection results, and in turn the modified detection results are used to update the tracking target model. For those trackers that correctly associate with detection, the detection results are considered as the correct results directly. For those trackers that are located closely with the detection results but the difference of size is larger than some threshold, the detection response and tracking confidence are combined to generate a new weight map, and then the mean-shift algorithm is adopted to get the location of the modified detection results with the size of the tracking result. On the contrary, for the trackers that neither have associated with detection, nor have near detection result, the tracking results are directly used as the detection result to make up the missed detection. After the modification, the modified detection results are in turn used to update the target models, which can help get rid of the accumulated error during runtime for normal tracking.

3. EXPERIMENTS

To validate the effectiveness of the proposed tracking-based moving object detection algorithm, we test it on the public VIVID dataset. VIVID is a public aerial video dataset, with the challenge of small moving objects, low contrast and occlusion.

Fig.4 shows the visual comparison results where the results of MHI[1] are drawn with green rectangles and the proposed tracking-based detection results are drawn with red rectangles. We can see that the proposed tracking-based method outperforms the motion based method in the situation of occlusion and environment noise.

Also, the precision and recall measures are used to evaluate the performance of the proposed tracking-base detection method comprehensively. Precision corresponds to the fractions of detected target pixels that are true positive, while recall indicates the ratio of correctly detected target pixels to the number of actual target pixels. The precision-recall curve is shown in Fig.5.

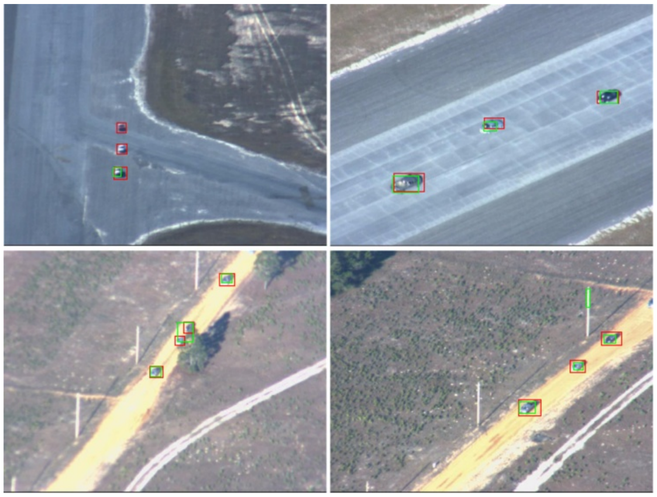

Fig. 4. Experimental results

In the experiments, the final result maps that are obtained by the proposed tracking-based (TD), the pure saliency-based (SD), and the MHI-based[1] (MD) detection methods, are binarized using various thresholds. The values of precision and recall are computed vis-a-vis ground truth data that are labeled manually at pixel level.

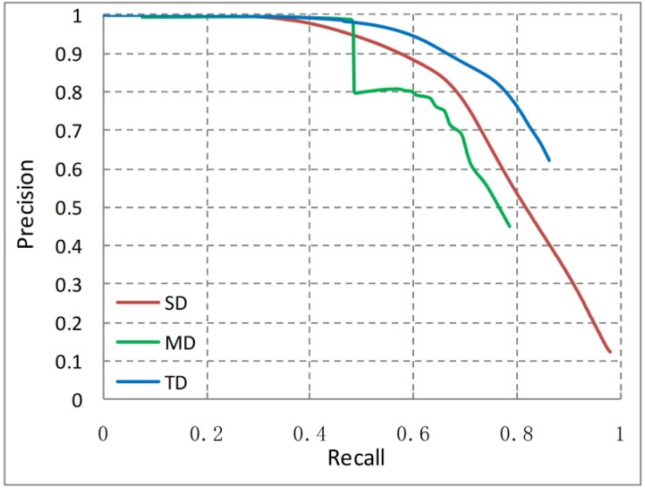

Fig. 5. Precision-Recall Curve(PRC) for naive thresholding of result maps in VIVID dataset egtest01(1800 frames).

From the experimental results we can see that, the proposed method performs more robustly than the motion-based method and the pure saliency-based detection method.

The algorithm is implemented with C++ on a personal computer with Pentium dual-core 2.5GHz CPU and 2G RAM. For a video with a resolution of 640*480, the time cost of our algorithm is about 100ms per frame, which is suitable for near-real-time moving target detection applications.

4. CONCLUSIONS

In this paper, we propose a novel tracking-based object detection framework for aerial videos. The moving target is only detected in the candidate object region, and the tracking information is used in both the prediction stage and the result optimization stage. The experimental results demonstrate that the proposed method can detect moving objects in aerial videos with high efficiency and accuracy.

5. REFERENCES

[1] Z. Yin and R. Collins, “Moving object localization in thermal imagery by forward-backward MHI,” in *CVPR Workshop*. IEEE, 2006.

[2] V. Reilly, B. Solmaz, and M. Shah, “Geometric constraints for human detecion in aerial imagery,” in *ECCV*, 2010.

[3] C. Benedek, T. Sziranyi, Z. Kato, and J. Zerubia, “Detection of object motion regions in aerial image pairs with a multilayer markovian model,” *TIP*, vol. 18, no.10, pp. 2303–2315, 2009.

[4] D. Rosenbaum, F. Kurz J. Leitloff, O. Meynberg, and T. Reize, “Real-time image processing for road traffic data extraction from aerial images,” in *in Proc. ISPRS Commission VII Symp.*, 2010.

[5] H. Cheng and D. Butler, “Segmentation of aerial surveillance video using a mixture of experts,” in *Proceedings of the Digital Imaging Computing: Techniques and Applications*, 2005.

[6] S. Ali, V. Reilly, and M. Shah, “Motion and appearance contexts for tracking and re-acquiring targets in aerial videos,” in *CVPR*, 2007.

[7] W.P. Koppen and M. Worring, “Backtrcking: retrospective multi-target tracking,” *CVIU*, vol. 116, pp. 967–980, 2012.

[8] Y. Li, H. Ai, T. Yamashita, S. Lao, and M. Kawade, “Tracking in low frame rate video: a cascade particle filter with discriminative observers of different life spans,” *TPAMI*, vol. 30, no. 10, pp. 1728–1740, 2008.

[9] Q. Yu and G. Medioni, “Multiple-target tracking by spatiotemporal monte carlo markov chain data association,”*TPAMI*, vol. 31, no. 12, pp. 2196–2210, 2009.

[10] A. Garcia-Martin and J. M. Martinez, “On collaborative people detetion and tracking in complex scenarios,” *Image and Vision Computing*, vol. 20, pp. 345–354, 2012.

[11] Z. Kalal, K. Mikolajczyk, and J. Matas, “Tracking-learning-detection,” *TPAMI*, vol. 34, no. 7, pp. 1409–1422, 2012.

[12] Y. Zhai and M. Shah, “Visual attention deteciton in video sequences using spatiotemporal cues,” in *ACM Multimedia*, 2006.

[13] M. Cheng, G. Zhang, N. J. Mitra, X. Huang, and S. Hu, “Global contrast based salienct region detection,” in *CVPR*, 2011.

[14] R. T. Collins, Y. Liu, and M. Leordeanu, “Online selection of discriminative tracking features,” *TPAMI*, vol.27, no. 10, pp. 1631–1643, 2005.