Chinese Journal of Aeronautics, (2013),26(5): 1211–1217

CSAA

Chinese Society of Aeronautics and Astronautics
& Beihang University

Chinese Journal of Aeronautics

cja@buaa.edu.cn
www.sciencedirect.com

Moving object detection in aerial video based on spatiotemporal saliency

Shen Hao [a], Li Shuxiao [a], Zhu Chengfei [a,*], Chang Hongxing [a], Zhang Jinglan [b]

[a] *Institute of Automaton, Chinese Academy of Sciences, Beijing 100190, China*
[b] *Queensland University of Technology, Brisbane, Australia*

Received 17 September 2012; revised 17 December 2012; accepted 21 March 2013
Available online 1 August 2013

KEYWORDS

Aerial video;
Computer vision;
Object detection;
Saliency;
Unmanned aerial vehicles

Abstract In this paper, the problem of moving object detection in aerial video is addressed. While motion cues have been extensively exploited in the literature, how to use spatial information is still an open problem. To deal with this issue, we propose a novel hierarchical moving target detection method based on spatiotemporal saliency. Temporal saliency is used to get a coarse segmentation, and spatial saliency is extracted to obtain the object's appearance details in candidate motion regions. Finally, by combining temporal and spatial saliency information, we can get refined detection results. Additionally, in order to give a full description of the object distribution, spatial saliency is detected in both pixel and region levels based on local contrast. Experiments conducted on the VIVID dataset show that the proposed method is efficient and accurate.

1. Introduction

With the development of technology, unmanned aerial vehicles (UAVs) have played a vital role in modern wars and industries. Moving object detection in aerial video as the foundation of higher targets, such as tracking and object recognition, is essential for UAV intelligence. 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 video, such as changing background and low resolution. Therefore, much attention has been paid to moving object detection in aerial video.

Generally object detection methods can be categorized in three approaches, namely temporal-based, spatial-based, and combined approach. For moving object detection from a video, motion cue is the most reliable information, so the proposed moving object detection methods are mainly based on temporal information, such as background subtraction[1,2] frame difference,[3,4] and optical flow.[5,6] Additionally, Cao[7] proposed to use the multi-motion layer analysis in moving object detection for airborne platform. Yu[8] used the long-term motion pattern in moving vehicle detection in aerial video. However, as the lack of spatial distribution, the results for the methods based on motion cues are usually undesirable. On the other hand, the spatial-based object detection method is principally used in the domain of object detection in static images. With the development of biological vision, many

[*] Corresponding author. Tel.: +86 10 62550985 21.
E-mail address: chengfei.zhu@ia.ac.cn (C. Zhu).

Peer review under responsibility of Editorial Committee of CJA.

ELSEVIER Production and hosting by Elsevier

1000-9361 © 2013 Production and hosting by Elsevier Ltd. on behalf of CSAA & BUAA.
http://dx.doi.org/10.1016/j.cja.2013.07.038

researchers have shifted their attentions to saliency detection, and plenty of saliency-based object detection methods have been designed. Initially, saliency detection is mainly based on low-level features, e.g., edges, colors, and textures. Recently, many new measures have been adopted in this literature, such as region contrast,[9] patch rarities,[10] and difference in frequency domain.[11,12] In addition, Wang[13] used visual saliency in aerial video summarization. Besides, in order to give a further description for moving objects, some researchers have also tried to combine temporal and spatial information in moving object detection.[14–17] Yin[14] used a 3D Markov random field (MRF) to predict each pixel's motion likelihood and the message was passed in a 6-connected spatiotemporal neighborhood. As every pixel needs to be predicted by MRF, the computational cost is huge. Liu[15] introduced saliency in moving object detection. They developed an efficient information theoretic-based procedure for constructing an information saliency map (ISM), which was calculated from spatiotemporal volumes.

Because aerial video has the property of changing background and small objects, moving object detection is still an open problem that needs to be addressed further. As the camera is moving, it is not easy to build a reliable background. In addition, the computing resource available on a UAV platform is often limited, so the optical flow is not a suitable choice. Thus, most of the object detection methods are based on frame difference. Although motion information is very important for moving object detection, there are still several drawbacks:

(1) The detected object may be larger than its real size.
(2) There may be holes in detection results.
(3) When an object is moving slowly, its motion is unreliable.

Besides, most of the saliency detection methods are based on static images, which focus on application of image classification or recognition, so they are not suitable for moving object detection.

For the combined methods, there are also some aspects that need to be modified for moving object detection in aerial video. Firstly, most of the existing methods[15–17] are mainly aimed at applications with fixed cameras, so they are not easy to be adopted in aerial video. Secondly, calculation of pixel saliency in a whole image is time-consuming. Finally, most of the integrated spatial information is only extracted in the pixel level, so higher-level object descriptions, such as region, are missed.

In short, there are clearly three major challenges for moving object detection in aerial video: changing background, small objects, and real-time processing demand. To tackle with these problems, we propose a novel spatiotemporal saliency detection method, inspired by biological vision. Temporal and spatial saliency is adopted in moving object detection as employed by previous researchers. However, instead of calculating spatial and temporal saliency separately[15,16] we developed a hierarchical detection method. Temporal saliency is used to get a coarse segmentation, and spatial saliency is adopted to get the object's appearance details in candidate motion regions. Finally, we get refined detection results by fusing temporal and spatial saliency information. Our contributions can be summarized as follows:

(1) A novel framework for moving object detection in aerial video that combines both temporal and spatial saliency.
(2) A hierarchical saliency detection manner that can greatly reduce time cost for spatial saliency calculation.
(3) A novel spatial saliency representation method, in which spatial saliency is extracted in both pixel and region levels to give a full description of the object distribution.

2. Proposed detection algorithm

In aerial video such as the ones shown in Fig. 1, objects are usually very small; they normally show more saliency in the local region than in global image. Thus we only explore spatial saliency in candidate local regions which are obtained through temporal saliency detection. The final detection results are achieved by fusing spatial and temporal saliency information. Fig. 2 shows the flow chart of the proposed moving object detection algorithm.

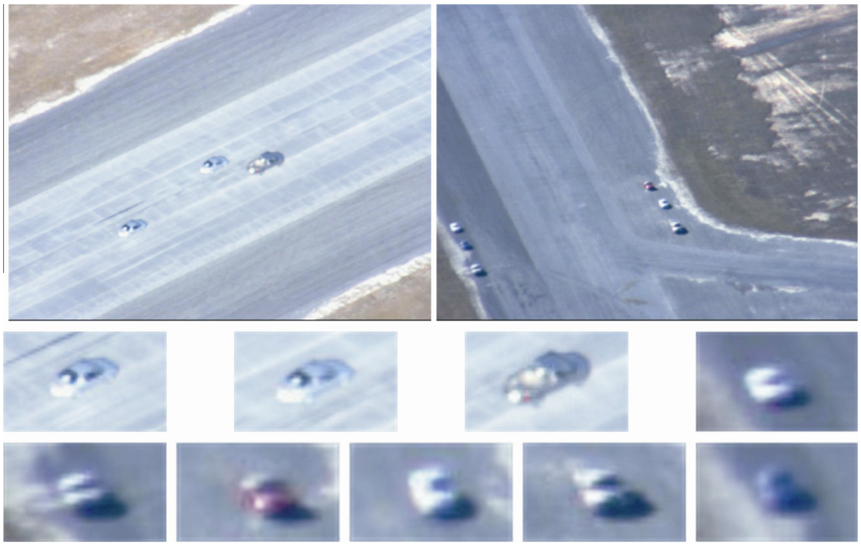


Fig. 1 Sample images. Upper: original images in the VIVID dataset. Middle and bottom: segmented objects in local regions.

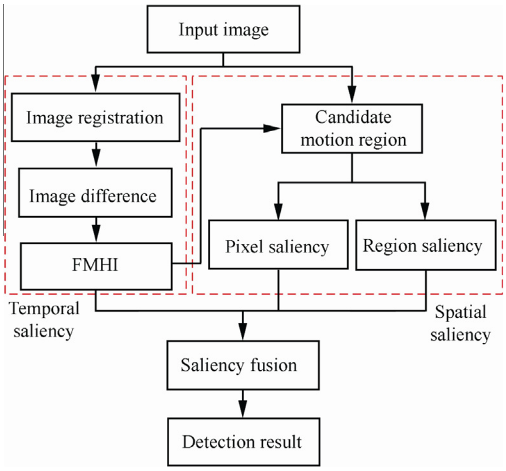


Fig. 2 Flow chart of the proposed algorithm.

2.1. Motion saliency detection

For aerial video, motion cues are salient and reliable in the global image. In this paper, motion information is used in both the candidate generation process and the saliency fusion stage. Considering the effect of time delay, the forward motion history image (FMHI), which is calculated from previous frames, is used as the temporal saliency information.

Firstly, the image registration between adjacent frames is implemented based on a point feature matching method that uses matched features to estimate the affine transformation between frames via the random sample consensus (RANSAC) approach.[18] Then, according to the estimated transformation between adjacent frames, we can get the image difference between them. After that, by fusing the previous FMHI and the current difference image, we can obtain the current FMHI, which is calculated as follows:

$$M_k(x,y) = \begin{cases} \max(0, M_{k-1}(x,y) + D_k(x,y) - d), & D_k(x,y) < T \\ 255, & D_k(x,y) \geqslant T \end{cases} \tag{1}$$

where $M_k(x,y)$ is the FMHI value of the kth frame at the location of (x,y), D_k the image difference between frames k and $k-1$, T the segment threshold, and d the decay term. The parameters are set as $T = 35$ and $d = 25$ by trial and error.

Finally, the FMHI is segmented as binary, and the connected component labeling algorithm[19] is used to get the candidate motion regions (MR).

2.2. Pixel saliency detection

As human visual system is sensitive to contrast in scenes and an object is compact in spatial distribution, we propose a modified histogram-based contrast method to define spatial saliency at the pixel level. Specifically, color contrast is weighted by its spatial distribution. As the segmented motion regions in aerial video are usually too small to calculate the saliency value, the original motion region is enlarged with a certain factor. The pixel saliency value at the location of (x,y) in the image is defined as

$$S_{\mathrm{P}}(x,y) = w_{\mathrm{RC}}(x,y)S_{\mathrm{HC}}(x,y) \tag{2}$$

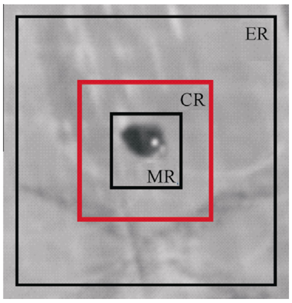


Fig. 3 CR and ER illustration.

where w_{RC} is the pixel distribution contrast between the center region (CR) and the extended region (ER), S_{HC} is the histogram-based color contrast. The CR is obtained by enlarging the MR with a factor of 2, and the ER is obtained by enlarging the MR with a factor of 4, as illustrated in Fig. 3. The largest rectangle is the ER, the middle one is the CR, and the smallest one is the MR which is obtained by temporal saliency.

Because the moving object is located mainly in the CR, the ER contains more background components other than the moving object. Therefore, the distribution contrast is utilized to suppress the background noise. The value of w_{RC} is formulated as follows:

$$w_{\mathrm{RC}}(x,y) = H_{\max}^{\mathrm{diff}} - H^{\mathrm{diff}}(x,y) \tag{3}$$

where $H^{\mathrm{diff}}(x,y)$ is the normalized histogram difference between the CR and the ER at the color value of $I_k(x,y)$, and $H_{\max}^{\mathrm{diff}}(x,y)$ is the maximum color difference.

The saliency of a pixel is defined as its color contrast to all other pixels in the CR. The saliency is calculated by using color statistics in the CR, which can be simplified by histogram manipulations as proposed by Cheng.[9] S_{HC} is calculated in the CR as follows:

$$S_{\mathrm{HC}}(x,y) = S_{\mathrm{H}}(u) = \sum_{v=0}^{N} f_v |u-v| \tag{4}$$

where u is the color value of $I_k(x,y)$, $S_{\mathrm{H}}(u)$ the saliency value of color u, f_v the probability of pixel color v in the CR, and N the color depth value, which is set as 256 in this paper.

2.3. Region saliency detection

Because segmented regions are compact and informative, we also extract the saliency over segmented regions to provide further information for object detection. Since the moving objects in aerial video are very small in the whole image, the region saliency is detected in the local region which is obtained from temporal saliency. The region saliency is defined as the distinctiveness of a patch from other patches in the local region. The widely used graph-based algorithm[20] is adopted to partition the local region into different patches. A color histogram is built to represent the information of each patch. By integrating the effects of dissimilarity and spatial distribution, the saliency for the patch R_i is defined as follows:

$$S_{\mathrm{R}}(i) = f_{\mathrm{spatial}}(i)\sum_{j=1}^{M} f_{\mathrm{area}}(j) D_{\mathrm{C}}(i,j) \tag{5}$$

where M is the number of segmented patches in the local region, $f_{\text{area}}(j)$ the area weight of patch R_i, and $D_C(i,j)$ the color contrast between patches i and j. Here the area weight is used to emphasize color contrast to bigger patches. The color contrast is defined as the histogram distance between two patches. f_{spatial} is the spatial weight coefficient, which is used to increase the effects of center patches and decrease the effects of patches that are near the local region boundary. The spatial weight coefficient is composed of the centroid offset coefficient $f^{\text{C}}_{\text{spatial}}$ and the boundary offset coefficient $f^{\text{B}}_{\text{spatial}}$. In order to emphasize the center patches, the spatial weight coefficient is formulated as follows:

$$f_{\text{spatial}}(i) = f^{\text{B}}_{\text{spatial}}(i) f^{\text{B}}_{\text{spatial}}(i) + \left(1 - f^{\text{C}}_{\text{spatial}}(i) f^{\text{C}}_{\text{spatial}}(i)\right) \tag{6}$$

where $f^{\text{C}}_{\text{spatial}}$ is used to measure the centroid distance between patch i and the entire local region. A smaller distance corresponds to a larger f_{spatial} value. Suppose that (x_c,y_c) is the centroid of the entire local region, W and H are the width and height of the region, and $(x_c(i),y_c(i))$ is the centroid of the current patch, so the centroid offset coefficient can be represented as

$$f^{\text{C}}_{\text{spatial}}(i) = \text{MAX}\left\{\frac{|x_c(i) - x_c|}{W/2}, \frac{|y_c(i) - y_c|}{H/2}\right\} \tag{7}$$

$f^{\text{B}}_{\text{spatial}}$ is used to measure the boundary distance between the patch and the entire local region. A smaller distance corresponds to a smaller f_{spatial} value. Suppose that B is the minimum boundary rectangle of the current patch, and B_l, B_r, B_t, B_b represent the left, right, top, bottom boundary respectively, so the boundary offset coefficient can be formulated as

$$f^{\text{B}}_{\text{spatial}}(i) = \text{MIN}\left\{\frac{\text{MIN}(B_l, W - B_r)}{W}, \frac{\text{MIN}(B_t, H - B_b)}{H}\right\} \tag{8}$$

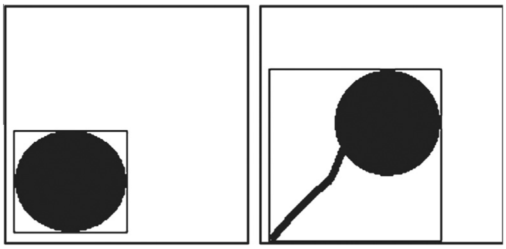


Fig. 4 Example of two situations that can be distinguished by the centroid offset coefficient.

It should be noted that here we employ two kinds of spatial coefficients to give a full illustration of the patch spatial property. For situations illustrated in Fig. 4, $f^{\text{B}}_{\text{spatial}}$ is nearly the same while $f^{\text{C}}_{\text{spatial}}$ can distinguish them clearly. In Fig. 4, the left image denotes a patch that is near boundary, the right image denotes a patch that is in the center but affected by noise. For patches that have different areas but the same centroid location, the boundary offset coefficient can work well.

2.4. Saliency fusion

We define the final saliency value of a pixel as a weighted linear combination of the detected saliency values. The fusing method can be formulated as

$$S(x,y) = w_1 M_k(x,y) + w_2 S_P(x,y) + w_3 S_R(i) \tag{9}$$

where i is the patch to which the pixel located at (x,y) belongs. w_i the weight value, that is obtained via offline training. Here the weight values are set as $w_1 = 0.31$, $w_2 = 0.42$, $w_3 = 0.27$.

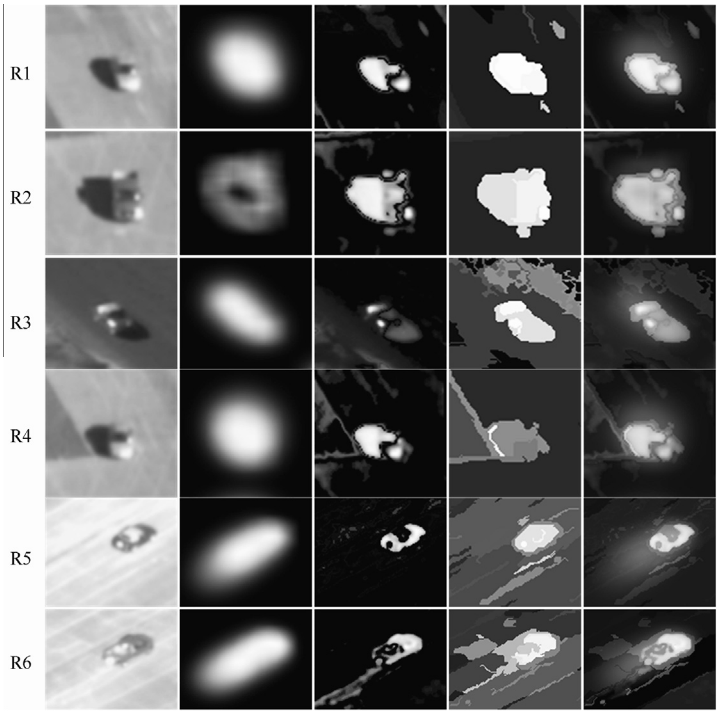


Fig. 5 Saliency detection results.

Some detected saliency results are shown in Fig. 5. The first column is the original local region that contains the moving object. The second to fourth columns are the temporal, pixel, and region saliency results, respectively, and the fifth column is the saliency fusion results.

From the results in Fig. 5, we can see that three kinds of saliency complement each other. The temporal saliency can provide candidate moving object regions for spatial saliency detection, as illustrated in the previous section. The spatial saliency can deal with the hollow and streaking effects in the temporal saliency that are caused by too slow or fast motions, as shown in the results of Rows 2, 5, and 6. Furthermore, the results of Row 3 to Row 6 illustrate that the pixel saliency and the region saliency can also complement each other.

3. Experimental results

To validate the efficacy of the proposed saliency-based moving object detection algorithm, we test it on the public VIVID dataset, and the results are compared with a motion-based

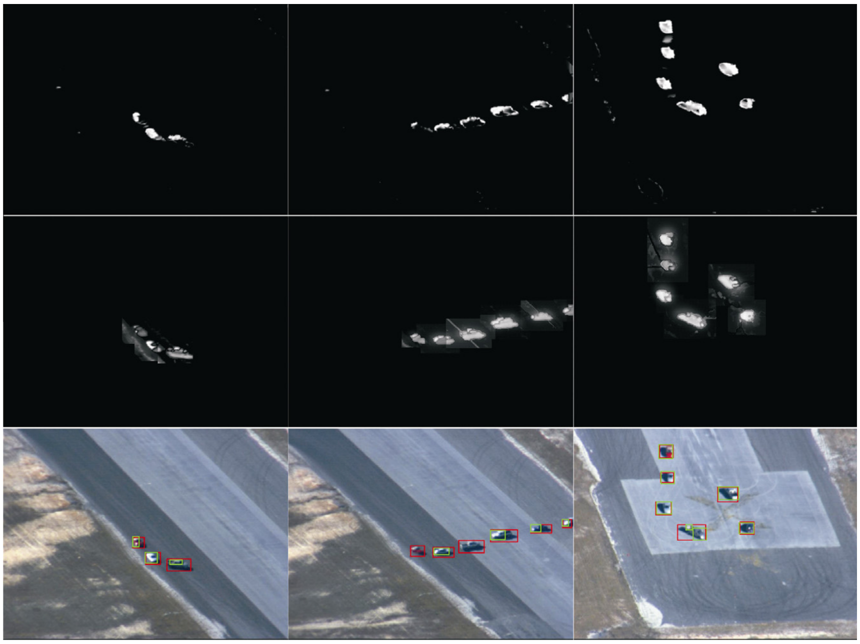


Fig. 6 Comparison results.

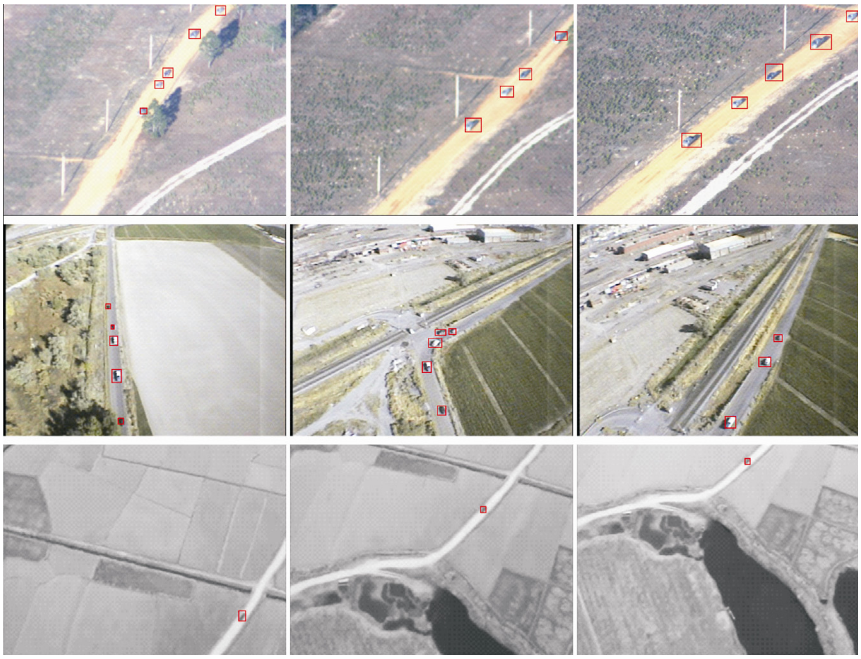
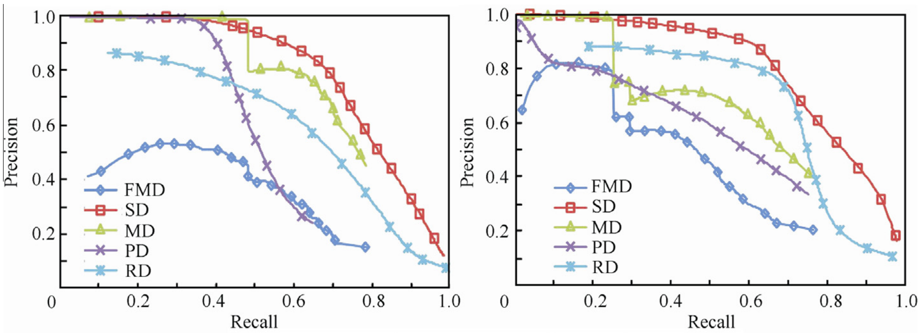


Fig. 7 More experimental results.



(a) PRC in egtest01 (1800 frames)

(b) PRC in egtest02 (1300 frames)

Fig. 8 Precision-recall curve (PRC) for naive thresholding of saliency maps in the VIVID dataset.

method.[3] Fig. 6 shows the results of visual comparison. The first row is the result map of Ref. 3 The second row is the saliency map of our method. The third row is the final segmentation, in which the results of Ref. 3 are drawn with green rectangles and the results of our method are drawn with red rectangles.

In order to test the robustness of the proposed method, more experiments are implemented in many other environments, and the results are shown in Fig. 7.

Similar to Refs. 9,12 precision and recall measures are used to evaluate the performance of the proposed method comprehensively. Precision corresponds to the fraction of salient pixels that are truely positive, while recall indicates the ratio of correctly detected salient pixels to the actual number of salient pixels.

In the test, the final saliency maps that are obtained by the proposed saliency-based (SD), FMHI-based (FMD), pixel saliency-based (PD), region saliency-based (RD), and MHI-based[3] (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 the pixel level. Here the egtest01 and egtest02 in VIVID are used to evaluate the algorithm performance. Fig. 8 shows that our method performs more robustly than the motion-based methods, and the saliency fusion result outperforms the individual saliency result.

The algorithm is implemented with C++ programming language on a personal computer with Pentium dual-core 2.5 GHz CPU and 2G RAM. For a video with a resolution of 640 × 480, the time cost of our algorithm is about 80 ms per frame, which is suitable for near-real-time moving target detection applications.

4. Conclusions

In this paper, we utilize spatiotemporal saliency in moving object detection. Temporal and spatial saliency is extracted in a hierarchical manner, and both pixel saliency and region saliency are extracted to give a full illustration for spatial distribution. The experimental results show that the proposed method can detect moving objects in aerial video with high efficiency and accuracy. Meanwhile, compared with an HMI-based method, our method does not have the effect of time delay.

However, as the detection algorithms estimate object locations in every frame independently, false alarms are unavoidable. We will deal with this by combining tracking information in our future study.

Acknowledgements

This study was co-supported by the National Natural Science Foundation of China (Nos. 61005028, 61175032, and 61101222). The authors are grateful to Yanhong Chen, Chenxu Wang, and Duoyu Gu for helping with the data labeling, and Yuqiang Qian in UMN for helping with the improvement of English writing quality. They would also like to thank the anonymous reviewers for their insightful comments and useful suggestions that help to improve the quality of this work.

References

1. Crivelli T, Bouthemy P, Cernuschi-Frías B, Yao J. Simultaneous motion detection and background reconstruction with a conditional mixed-state markov random field. *Int J Comput Vision* 2011;**94**(3):295–316.
2. Barnich O, Droogenbroeck MV. ViBe: a universal background subtraction algorithm for video sequences. *IEEE Trans Image Process* 2011;**20**(6):1709–24.
3. Yin Z, Collins R. Moving object localization in thermal imagery by forward-backward MHI. In: *IEEE conference on computer vision and pattern recognition, workshop*, Jun 17–22; 2006.
4. Benedek C, Szirányi T, Kato Z, Zerubia J. Detection of object motion regions in aerial image pairs with a multilayer Markovian model. *IEEE Trans Image Process* 2009;**18**(10):2303–15.
5. Medioni G, Cohen I. Event detection and analysis from video streams. *IEEE Trans Pattern Anal and Mach Intell* 2001;**23**(8):873–89.
6. Kim J, Ye G, Kim D. Moving object detection under free-moving camera. In: *17th IEEE international conference on image processing*. 2010 Sept 26–29, Hongkong, China. Piscataway: IEEE; 2010. p. 4669–72.
7. Cao X, Lan J, Yan P, Li X. Vehicle detection and tracking in airborne videos by multi-motion layer analysis. *Mach Vision Appl* 2012;**23**(5):921–35.
8. Yu Q, Medioni G. Motion pattern interpretation and detection for tracking moving vehicles in airborne video. In: *IEEE conference on computer vision and pattern recognition*, 2009 Jun 20–25, Miami, FL, USA. Piscataway: IEEE; 2009. p. 2671–78.

9. Cheng M, Zhang G, Mitra N J, Huang X. Hu S. Global contrast based salient region detection. In: *IEEE conference on computer vision and pattern recognition*, 2011 Jun 20–25, Providence, RI. Piscataway: IEEE; 2011. p. 409–16.
10. Borji A, Itti L. Exploiting local and global patch rarities for saliency detection. In: *IEEE conference on computer vision and pattern recognition*; 2012.
11. Hou X, Zhang L. Saliency detection: a spectral residual approach. In: *IEEE conference on computer vision and pattern recognition*, 2007 Jun17–22, Minneapolis, MN, USA. Piscataway: IEEE; 2007. p. 1–8.
12. Achanta R, Hemami S, Estrada F, Süsstrunk S. Frequency-turned salient region detection. In: *IEEE conference on computer vision and pattern recognition*, 2009 Jun 20–25, Miami, FL, USA. Piscataway: IEEE; 2009. p. 1597–1604.
13. Wang J, Wang Y, Zhang Z. Visual saliency based aerial video summarization by online scene classification. In: *Proceedings of the sixth international conference on image and graphics*, 2011 Aug 12–15, Hefei, Anhui, China. Piscataway: IEEE; 2011. p. 777–82.
14. Yin Z, Collins R. Belief propagation in a 3D spatio-temporal MRF for moving object detection. In: *IEEE Conference on computer vision and pattern recognition*, 2007 Jun 17–22, Minneapolis, MN, USA. Piscataway: IEEE; 2007. p. 1–8.
15. Liu C, Yuen PC, Qiu G. Object motion detection using information theoretic spatio-temporal saliency. *Pattern Recog* 2009;**42**:2897–906.
16. Mahadevan V, Vasconcelos N. Spatiotemporal saliency in dynamic scenes. *IEEE Trans Pattern Anal Mach Intell* 2010;**32**(1):171–7.
17. Yang H, Tian J, Chu Y. Spatiotemporal smooth models for moving object detection. *IEEE Signal Process Lett* 2008;**15**:497–500.
18. Fischler MA, Bolles RC. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun ACM* 1981;**24**(6):381–95.
19. Suzuki K, Horiba I, Sugie N. Linear-time connected-component labeling based on sequential local operation. *Comput Vision Image Understanding* 2003;**89**:1–23.
20. Felzenszwalb P, Huttenlocher D. Efficient graph-based image segmentation. *Int J Comput Vision* 2004;**59**(2):167–81.

Shen Hao is a Ph.D. student in Institute of Automation, Chinese Academy of Sciences. He received his B.S. degree in mechanical engineering and automation from Hohai University in 2008. His main research interests are moving object detection and machine vision.

Zhu Chengfei is an assistant professor in Institute of Automation, Chinese Academy of Sciences. He received his Ph.D. degree from Chinese Academy of Sciences in 2010. His main research interests are computer vision, object detection and recognition.

Chang Hongxing is a professor with Institute of Automation, Chinese Academy of Sciences. He received his B.Sc. and M.Sc. degrees in mechanical engineering from Beijing University of Aeronautics and Astronautics, China, in 1986 and 1991, respectively. Presently, he is the director of Integrated Information System Research Center. His research interests include computer and machine vision, pattern recognition, and intelligent UAV systems.