Target Tracking Method in Aerial Video Based on Saliency Fusion

Jie Han,
ICIE Institute, School of Aerospace Science and Technology, Xidian University,
Xi'an,China
e-mail:hanjie289@163.com

Wei Sun * ,
ICIE Institute, School of Aerospace Science and Technology, Xidian University,
Xi'an,China.
*corresponding author:wsun@xidian.edu.cn

Baolong Guo,
ICIE Institute, School of Aerospace Science and Technology, Xidian University,
Xi'an,China
e-mail:blguo@xidian.edu.cn

Abstract—For the problem that when doing target tracking under dynamic background, the object extraction is susceptible to interference. Based on the static and dynamic saliency maps of video sequence, this paper presents a method of feature fusion. By calculating the weight for each of the significant area of the extracted static and dynamic saliency, based on human visual attention mechanism, the weighted comprehensive significant figure is get, so as to determine the status of the target. Several experiments are done by the mathod put forward in this paper and the experimental results show that the algorithm is able to solve the problem of susceptible to interference in target tracking, while ensuring robustness and meeting the real-time and accuracy needed. And the algorithm in this paper not only improves the real-time performance and robustness on the premise of guaranteeing the tracking accu racy, but also gives attribution to the field of target tracking.

Keywords-salience;aerial video;target tracking; dynamic background ;biological vision

I. INTRODUCTION

As an important research direction of the field of computer vision, target tracking attracts the interest of researchers. In most cases, the background is dynamic, which brings more difficult for moving object extraction [1]. In contrast, human and animal eyes can accurately locate target in a complex environment. Therefore, the principle of biological vision gives important scientific ways for intelligent processing of video information [2]. Simulating biological visual system for target tracking has broad application prospects.

To solve those problems, this paper proposes a method for saliency fusion based on biological vision, to achieve aerial target tracking. First, extract static saliency map based on the phase spectrum. Second, extract dynamic saliency map using ORB (oriented FAST and rotated BRIEF) [3]. And then fuse the two maps, resulting in a comprehensive saliency map, and ultimately complete tracking aerial targets. Thus, the rapidity, accuracy, and robustness of the system can be optimized.

II. STATIC SALIENCY DETECTION

Using the visual attention mechanism based on phase spectrum can quickly and accurately detect the target in the image. Compared with the particle filter and the Itti model, it has lower computational complexity, and can effectively separate target and background.

$$E\left(\frac{(\bar{f},\bar{x})}{\|\bar{f}\|\cdot\|\bar{x}\|}\right)\geq 0.5,\ for\ |\Omega_b|<\frac{N}{6} \quad (1)$$

$$\bar{f}=IDCT(Sign(DCT(f))) \quad (2)$$

$$\bar{x}=IDCT(sign(DCT(x))) \quad (3)$$

Where $E\left(\frac{(\bar{f},\bar{x})}{\|\bar{f}\|\cdot\|\bar{x}\|}\right)$ represents expectation, Ω_b is the cardinal number of background image in DCT domain。

f consists of independent Gaussian distribution points, on the DCT inverse transform, 79.8% [4] of the energy belongs to the foreground image.

$$E\left(\frac{\alpha}{\sqrt{\alpha^2+\beta^2}}\right)\geq\sqrt{\frac{2}{\pi}}\approx 0.798 \quad (4)$$

Where $\alpha=\sqrt{\sum_{i\in T_f}\bar{f}_i^2}$, $\beta=\sqrt{\sum_{i\notin T_f}\bar{f}_i^2}$, T_f is non-zero collection of the foreground image. Thus, the image saliency map can be efficiently obtained using the DCT inverse transformation[5].The Gauss template of an aerial

image can be selected as 5×5， δ =2； Its Static saliency is shown in Fig .1.

(a)Original image (b)Static salience map

Figure 1. Static salience map

III. DYNAMIC SALIENCY DETECTION

The speed of ORB operator is 10 times that of SURF, and ORB makes up the imperfection of no-direction of FAST and no-rotational invariance of BRIEF. So in this paper, the ORB is used to feature [6].

The ORB method of feature extraction adopts FAST-9[7] operator with direction and BRIEF operator to detect and describe feature points. That allows dynamic detection of targets a better robustness over scale of target and rotating light. The implementation process is as follows:

A. Extraction of Feature Points

As shown in Fig .2, construct a corner detector with the given n, and compare the 16 pixels of the arc of radius 3 to determine whether the pixel is a corner.

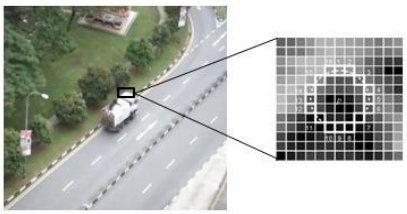


Figure 2. Corner detection

To avoid that some of the extracted feature points are local minima, non-maximum suppression with score function V is adopted to remove adjacent corner points of V value. For point P extracted by Segment-test in the point collection of M, its value of scoring function is V_p.In the $w \times w$ neighborhood of point P, $q \in M$ .And only when $V_p \geq V_q$, P is a feature point.

To add direction information to the FAST corner operator [8], ORB uses gray centroid, which represents direction with the offset vector between gray and the center of a corner, for:

$$\theta = a\tan 2(m_{01}, m_{10}) \quad (5)$$

$$C = (\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}})$$

Where the heart is .The result of corner extraction in the video frame of Fig .1 with oFAST [9] is shown in Fig .3.

Figure 3. Feagure points

B. Feature Describing and Matching with rBRIEF

In the position of (x_i, y_i) , a matrix is defined:

$$M = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \end{pmatrix} \quad (6)$$

For n arbitrary sets of feature collection of binary criteria, build the correct version of M:

$$M_\theta = R_\theta \cdot M \quad g_n(p,\theta) = f_{nd}(p) \mid (x_i, y_i) \in M_\theta \quad (7)$$

Where θ is the direction of neighbourhood, and R_θ is the corresponding rotation matrix. As a random parameter estimation algorithm, RANSAC can rule out outside interference to estimate global optimal parameters. For video frames, their feature points searching, point pairs matching and purification are as follows:

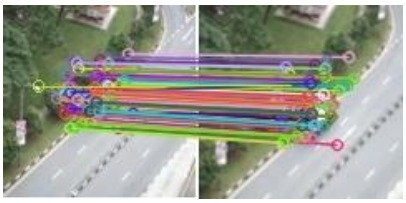


Figure 4. Feature matching

IV. FUSION OF IMAGE SALIENCY

A. Principle of Human Vision

The human visual system is composed of eyes, LGN and visual cortex and information in the brain delivers according to a certain path [10]. The brain processes complex visual information hierarchically. For the observer, not all external information is equally important, so the brain responds to some important information, which is an important basis for target tracking here.

B. *Salience Fusion Based on Visual Principle*

Based on selective attention mechanism in visual searching, a reasonable feature fusion strategy is required for static and dynamic full-time estimating and feature fusion to generate the final saliency map.

1) According to the principle of information in events, N represents the number of salient regions, and λ *the ratio of salient area to total area, whose weight is* w_a *:*

$$w_\lambda = -\frac{1}{N}\log_2(\lambda) \tag{8}$$

$$\lambda = \frac{\sum_{i=1}^{n} m_i}{M} \tag{9}$$

$I(x)$ is the information included in event x, M is the total area of image, and m_i is area i in salience map k.

2) The more concentrated salient regions saliency map is, the larger significant contribution the map provides. Therefore, in saliency map k, the corresponding weight of average distance between the significant region is:

$$w_d = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{N}\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{10}$$

(x_j, y_j) is the heart of the salient region.

3) The center position of the image is important central visual zone. So, the more salient region near the center, the greater contribution of the considered saliency map is.

$$w_l = \frac{1}{N}\sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2} \tag{11}$$

Where $(\bar{x}, \bar{y})$ is the center position of the image.

From the above analysis, the weight of saliency map k w_k and the comprehensive map is obtained as follows:

$$w_k' = w_{\lambda k} + w_{dk} + w_{lk} \tag{12}$$

$$w_k = \frac{\frac{1}{w_k'}}{\sum_{k=1}^{l}\frac{1}{w_k'}} \tag{13}$$

C. *Fusion of Static and Dynamic Saliency*

From the above, the fused comprehensive saliency map [11] is:

$$S = \sum_{k=1}^{l} w_k * S_k \tag{14}$$

Detect the static and dynamic salience of the video frame in Fig .1, and then fuse the detected saliency [12], the final fusion and tracking results are shown in Fig .5:

(a) Comprehensive saliency map (b) Final tracking result

Figure 5. Fusion map and final tracking result

V. ALGORITHM IMPLEMENTATION PROCESS

Based on the above idea, Fig .6 shows the algorithm implementation process:

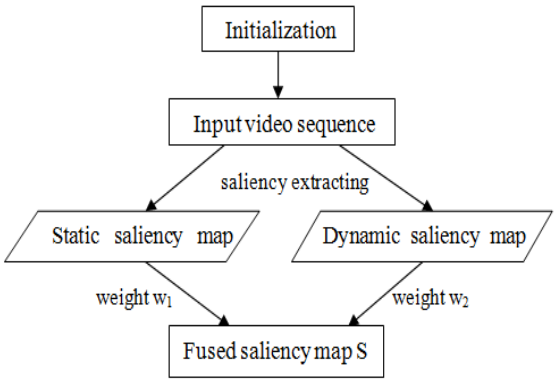


Figure 6. Algorithm block diagram in this paper

Steps of the algorithm are as follows：

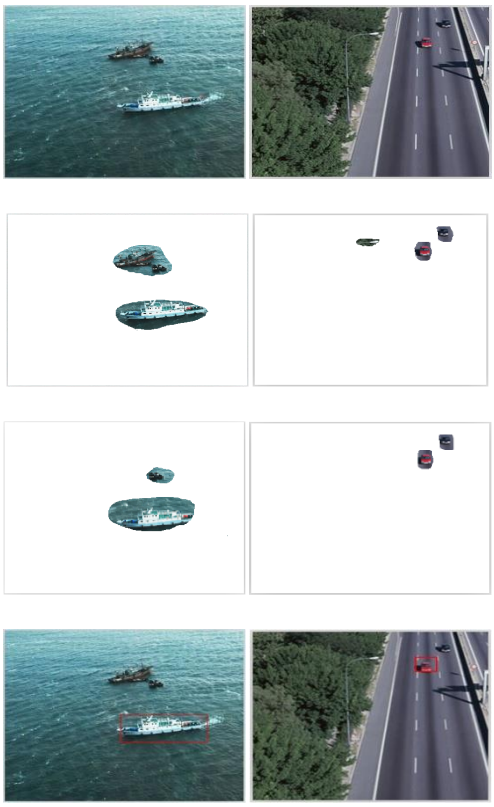
step1: Obtain a static saliency map S1 of the video sequence using DCT inverse transform;

step2: Use oFAST and rBRIEF to extract and match feature points, and pure the matches with RANSAC;

step3: Extract the dynamic saliency map S2 of a video sequence;

step4: Fuse according to $S = \sum_{k=1}^{l} w_k * S_k$ [13] to get integrated saliency map S.

Fig .7 shows the result of two aerial video sequences, i t can be seen that the target can be accurately located.



(a) Boat video sequence (b) car video sequence

Figure 7. Experimental results of boat, car video sequence

VI. EXPERIMENT AND ANALYSIS

A. *Rapidity Analysis*

For three aerial video sequences, as shown in tab.1, detect the feature points and match them using SIFT and ORB(the method in this paper)[14].The consumed time shows the speed of ORB detecting feature points is much faster than SIFT.

B. *Accuracy Analysis*

Fig .8 shows the tracking results of the 20th ,60th ,and 100th fame from the same video sequence, it can be seen that when there is another car interference in 60th fame[15], particle filter algorithm began tracking the other non-target car while tracking by detector can better track the target vehicle.

（a） （b）

(a)Tracking result of particle filter for frames 20＃,60＃,110＃ (tracking directly)

(b)Tracking result of algorithm in this paper for frames 20＃,60＃,110＃（tracking by detector）

Figure 8. Compare of tracking results

TABLE I. CONTRASTION TABLE OF FEATURE EXTRACTION ALGORITHMS

Algorithmin	*Video sequence*	*Number of frames*	*Number of feature points*	*Detection of feature points (ms)*	*Total time (s)*	*Average time (ms)*
SIFT	S1	360	131	487	73.4	203
	S 2		212	541	99.5	276
	S 3		143	495	92.3	256
ORB（in this paper）	S 1		243	16	13.0	36
	S2		480	78	18.5	51
	S 3		507	81	14.7	41

VII. CONLUSION

This paper puts forward a new method for target tracking in aerial video based on feature points extraction and saliency map fusion. First foreground image is obtained by DCT inverse transform phase spectrum collection of nonzero, resulting in static saliency map, and then get a dynamic map based on ORB feature point matching. Then fuse them to get the comprehensive

saliency map according to the fusion rule. And finally locate and track. The experimental results show that compared to SIFT, the algorithm in this paper improves the real-time performance and robustness [16] on the premise of guaranteeing the tracking accuracy.

ACKNOWLEDGMENT

This work was supported by Fundamental Research Funds for the Central Universities under Grant JB141307, National Nature. Science Foundation of China (NSFC) under Grants 61201290, and other Grants 51205301, 61201089, 61305041, 61305040, the China Scholarship Council (CSC) and the National Institutes of Health Grants No.R01CA165225 of the United States.

REFERENCES

[1] Wei Sun, Objects detecting and tracking with a new particle filter; CECNet2012, Yichang,pp.3340 -3343,2012.

[2] Yilmaz A, Javed O, Shah M. "Object tracking": Acm computing surveys (CSUR), 38(4), pp. 13. 2006.

[3] Wei Sun,Chen Long.Binocular Vision-based Position Determination Algorithm and System. CDCIEM 2012,Vol.1, pp.170-173,2012.

[4] Baolong Guo, Wei Sun. "Introduction to Digital Image Processing System Engineering". People Post Press, pp.65-80, 2012.

[5] Wei Sun, Xu Zhang, Yunyi Yan. An Objects Detecting and Tracking Method based on MSPF and SVM. International Journal of Soft Computing and Software Engineering [JSCSE], Vol.2, No.2,pp. 9-16, 2012.

[6] TANG Ti, WEI Ming, KE Jia. Multi-object tracking algorithm based on data association matrices, Computer Engineering, 36(23), pp.158-161.,2010.

[7] P. Viola, M. Jones. "Rapid Object Detection Using a Boosted Cascade of Simple". IEEE Computer Vision and Pattern Recognition,3(2), pp. 511-518,2008.

[8] Nummiaro K, Koller-Meier E, Van Gool L. "An adaptive color-based particle filter". Image and vision computing , 21(1), pp. 99-110, 2008.

[9] Wei Sun, Guo Bao-long. Multiple Objects Tracking with Particle Filter and Mean Shift Clustering. ISDA-2008, Vol.2, pp.482- 486, 2008.

[10] Gordon N J, Salmond D J, Smith AFM. "Novel approach to nonlinear/non-Gaussian Bayesian state estimation"//IEE Proceedings F (Radar and Signal Processing). IET Digital Library, 11(2), pp.37-43. 2003.

[11] Evans C. "Notes on the opensurf library". University of Bristol, Tech. Rep.,January, pp.21-33, 2009.

[12] WANG Xiao,SUN Zhang, LIN Jing. "Energy-efficient adaptive sensor scheduling for target tracking".Control Theory, 8,pp.8-92,2010.

[13] XU Jin, LI Jinxing. "State estimation with quantised sensor information in wireless sensor networks".IET Signal Process,5, pp.1-26,2011.

[14] SHOU Tiande. "Processing mechanism of visual information". Shanghai Press, pp. 58-80,1997.

[15] L. Itti, P. F. Baldi. "Bayesian Surprise Attracts Human Attention". Vision Research, 49(10),pp. 1295-1306.2009

[16] Long Chen, Baolong Guo,Wei Sun .A new multiple-objects tracking method with particle filter. IAS 2009, Vol.1, pp.281-284, 2009.