Multimed Tools Appl (2017) 76:14781–14798
DOI 10.1007/s11042-016-4025-7

Small target detection combining regional stability and saliency in a color image

Jing Lou[1] · **Wei Zhu**[1] · **Huan Wang**[1] · **Mingwu Ren**[1]

Received: 6 July 2016 / Revised: 25 September 2016 / Accepted: 28 September 2016 /
Published online: 9 November 2016

Abstract In this paper, we will address the issue of detecting small target in a color image from the perspectives of both stability and saliency. First, we consider small target detection as a stable region extraction problem. Several stability criteria are applied to generate a stability map, which involves a set of locally stable regions derived from sequential boolean maps. Second, considering the local contrast of a small target and its surroundings, we obtain a saliency map by comparing the color vector of each pixel with its Gaussian blurred version. Finally, both the stability and saliency maps are integrated in a pixel-wise multiplication manner for removing false alarms. In addition, we introduce a set of integration models by combining several existing stability and saliency methods, and use them to indicate the validity of the proposed framework. Experimental results show that our model adapts to target size variations and performs favorably in terms of precision, recall and F-measure on three challenging datasets.

Keywords Small target detection · Stable region · Visual saliency · Color image

1 Introduction

Small target detection plays an important role in many computer vision tasks, including early warning system, remote sensing and visual tracking. Different from conventional object

✉ Mingwu Ren
mingwuren@163.com

Jing Lou
jinglou@gmail.com

Wei Zhu
zw_njust@163.com

Huan Wang
wanghuanphd@njust.edu.cn

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094 Jiangsu Province, People's Republic of China

recognition and detection in natural scenes, small targets usually lack high-level appearance features and semantic patterns due to the long imaging distance. In many cases, small targets are immersed in sea clutter and cloud clutter. Several factors such as sensor noise, size variations and artificial interference make the problem more challenging.

Over the past decades, numerous detection models aiming at small infrared targets have been proposed, whose strategies can be divided into three main categories: target enhancement [26], background suppression [4], and figure-ground segregation [7, 13, 17]. With the advantage of simultaneously enhancing target signal and suppressing background clutter, the third category of methods usually exploits different contrast techniques to model the problem. In computer vision, the contrast mechanisms, including local center-surround difference and global rarity, are closely related to human visual perception and widely used in bottom-up saliency detection models. Since visual saliency which stemmed from psychological science has attracted more attention [1–3, 9, 12, 16, 22, 24, 25, 27], some saliency map based methods are also proposed in recent years [14, 15, 20].

However, we focus on small target detection in a color image rather than infrared image. For one thing, compared with sequential detection, single frame detection is more suitable for fast changing backgrounds and inconsistent targets [10]. For another, although the visible spectrum is a rather narrow portion of the whole electromagnetic spectrum, the visual band is the most familiar in all human activities [11] and often used in conjunction with infrared imaging in many multispectral vision tasks [6]. Studies of small target detection in the visual band may extend the relevant techniques to more general cases, and offer some complementary solutions to the existing infrared target detection methods.

Compared with infrared small targets, the small targets obtained in the visual band usually have smaller intensity values. But if we are shown two examples of the small targets obtained in the visual and infrared bands individually, we will find somewhat similar features between them. Intuitively, both of them are connected foreground components, which have some desirable properties as follows: 1) spot-like shape, 2) small entropy, 3) nearly uniform intensity, and 4) local center-surround contrast with their neighbors. These intrinsic properties imply two important pieces of information, i.e., an ideal small target is locally stable, as well as locally salient. Thus our research mainly focuses on two issues: “how to measure stability” and “how to detect saliency”.

Two classical works are closely correlated with the above mentioned issues. Relevance to stability is a new type of affinely invariant region namely “Maximally Stable Extremal Region” (MSER) introduced in [18]. The MSER is a local part of the gray-scale input where its binarization is stable over a range of thresholds. However, due to the absence of optimal thresholds, there may exist multiple stable thresholds for some certain parts of input image. Besides false alarms, these redundant patterns need to be removed in the post-processing stage.

For the second issue related to saliency, Achanta *et al.* [3] propose a Frequency-Tuned (FT) method which considers globally rare color features and tends to emphasize large salient objects. In a color image, however, the color and luminance properties of small targets are not always globally rare, thereby preventing them from distinctly popping out. In order to remove the noisy results caused by such confusion, we need more sophisticated techniques to refine the saliency value of each target region. Furthermore, the variations of the target size make it even trickier to choose the appropriate thresholds for target segmentation.

Mainly inspired by [3, 18], we propose a novel model in this paper for extracting locally stable and salient regions from a color image. This model will be called “RSS” in the following sections. By exploiting four regional structure metrics, a stability extractor is designed to

produce a stability map consisting of a set of small target proposals. Meanwhile, a saliency detector is presented to generate the saliency map by suppressing the low frequencies for the likely small targets pop-out. All of the small target proposals then compete for entire saliency amongst themselves by a simple integration of the above two maps, so that only the relatively salient regions are retained. We will demonstrate the validity and adaptability of this integrated model, which yields the performance improvement with higher precision and recall.

The main contributions of this work include:

1. We propose an algorithm to detect stable regions for an individual image, which generates more accurate target proposals.
2. An improved local contrast mechanism is proposed for pixel-level saliency detection, which simultaneously highlights small salient regions and suppresses uniform patterns.
3. We present a simple integration technique by combining stability and saliency maps instead of just using a single one in isolation, which achieves more favorable results.
4. We provide a benchmark database containing three datasets (totally 1,093 images) for small target detection. For each image, the database provides the pixel-wise ground truth. We also provide our MATLAB code for evaluating detection results on this database.

This paper is organized as follows. In Section 2, we present the small target detection model combining regional stability and saliency. In Section 3, we discuss experimental details, evaluation measures and results. Conclusions and possible extensions are presented in Section 4.

2 The Proposed Approach

2.1 General Framework

As shown in Fig. 1, our RSS model contains two separate parts: a stability extractor that proposes stable regions, and a saliency detector that gives each candidate region a saliency score. It is closely related to the intrinsic properties of small targets, explaining human visual observation strategies.

Part I Visual input is provided in the form of color image (Fig. 1a), and is converted to the gray-scale intensity image (Fig. 1b) due to no color variations inside a small target. Considering inner

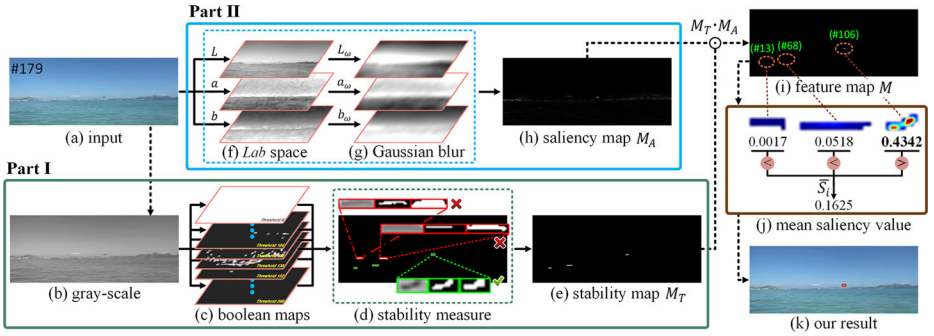


Fig. 1 Framework of the proposed RSS model

smoothness and intensity homogeneity, we exploit five structure descriptors to construct all potential target proposals (Fig. 1c). Hierarchical region analysis is further applied to filter out the erroneous results (Fig. 1d), and generate the final stability map M_T (Fig. 1e).

Part II We transform the *RGB* input (Fig. 1a) into the *Lab* color space (Fig. 1f) and apply a Gaussian blurred kernel on three color channels (Fig. 1g) to extract saliency information. Our aim is to judge the important parts of the entire scene automatically. As a result, we produce a full resolution saliency map M_A (Fig. 1h) which represents the conspicuity at each location and guides the proposals selection.

Combination The two maps M_T and M_A are fed, in a simple pixel-wise multiplication $M_T \cdot M_A$ manner, into a master feature map M (Fig. 1i), which topographically codes for local stability and conspicuity over the entire scene. Then, the average saliency of each target proposal is computed, such that only the proposals whose saliency values exceed the average value of all proposals (Fig. 1j) are retained. Figure 1k shows the detected target by using the thresholding value, with 0.1625, the average value of three candidate regions.

2.2 Regional Stability

In a subimage containing one small target, it is reasonable to conclude that the target image whose pixels tend to distribute uniformly, will have an appearance of high homogeneity and will not exhibit a large variety of gray tones. Similarly, the rest of the subimage which could be considered the background, also has a small change of intensities. In addition, from the rudimentary aspect of human vision system, humans can focus attention on target region of interest and detect such target, depending mainly on its contrast to the nearby area. Since there are two dominant modes with sufficient contrast in a subimage window, a typical way to extract the small target from its surrounds is to choose an optimal threshold for separating these two clumps. In this case, Otsu's method [19] is optimum, in the sense that it computes a threshold maximizing the between-class variance.

However, it is important to note that the key challenges are how to subdivide the input image and how to obtain the meaningful subimage windows we need automatically. In fact, the aforementioned subimage window is essentially a center-surround block. There exist numerous multi-size blocks having similar pattern in the entire image area, but most of them are false alarms. In practical scenarios, there are seldom adaptive constraints to help the detector filter out these false alarms, especially in some applications where accuracy is an important factor. Furthermore, for a specific small target, there also exist multiple subimages with various window sizes that result in different thresholds obtained by using Otsu's method. Although any of these thresholds sometimes seems to perform quite well, we are still faced with an intractable problem of how to determine which one is more suitable. Considering the variability of small targets involved, the choice of window size will have a material impact on the performance of the detection system.

In the proposed model, we tackle the issue of the optimal size of a subimage window as finding the smallest rectangle containing the candidate target region, which should satisfy the stability constraints based on certain structure attributes. Consider the gray-scale subimage shown in Fig. 2b whose values are between 0 and 255, both the small target and the background are nearly uniform. Figure 2c shows a set of binary masks by inversing and

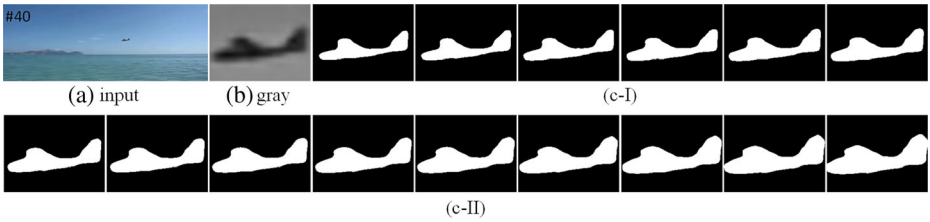


Fig. 2 Demonstration of the stability of a small target

thresholding Fig. 2b from 64 to 120 with a step size of 4. Note that although the white region corresponding to the local intensity minima grows gradually, the local binarization is virtually unchanged over a limited range of thresholds. So we conclude from these observations that a small target is a stable region in its local surrounds.

Based on the conclusion mentioned above, a two-stage detection algorithm, denoted as "RSt", is designed for computing five structure descriptors of each connected component and measuring the stability in each cluster. Here, we introduce four comparisons of attribute values to describe the similarity between two labeled regions u and v as follows, and all necessary definitions are given in Table 1.

- area variation:

$$\mathbf{D_r}(u,v) = ||u|-|v||. \tag{1}$$

- center distance:

$$\mathbf{D_c}(u,v) = \|c_u - c_v\|^2. \tag{2}$$

  where $\|\cdot\|$ is the ℓ_2-norm.
- fill rate difference:

$$\mathbf{D_f}(u,v) = \frac{\max(f_u, f_v)}{\min(f_u, f_v)}. \tag{3}$$

- aspect ratio difference:

$$\mathbf{D_a}(u,v) = \frac{\max(a_u, a_v)}{\min(a_u, a_v)}. \tag{4}$$

Stage I We first segment and complement the gray-scale image G using a set of sequential thresholds in the range 0 to 255 with a step size of δ. The output is a pool R of regions as

Table 1 Definitions used for measuring region similarity

boolean map I is a mapping $I: Q \subset \mathbb{Z}^2 \rightarrow \{0, 1\}$.
region $r \subset Q$, each r is a connected foreground component, i.e. $\forall (x, y) \in r : I(x, y) = 1$ and the Euler number of r equals to 1. $|r| = \sum_r I(x, y)$ indicates the number of pixels in r.
centroid c is the center of mass of r, c is a vector $[c_x, c_y]$ where $c_x = (\sum_r xI(x, y))/|r|$ and $c_y = (\sum_r yI(x, y))/|r|$.
bounding box b denotes the smallest rectangle containing the region r. b is defined by a vector $[b_x, b_y, b_w, b_h]$, where the upper left corner of b is in the form $[b_x, b_y]$, and b_w, b_h specify the width and height of b respectively.
fill rate f, defined as $|r|/(b_w \times b_h)$.
aspect ratio a is the width-to-height ratio of the region r, and is computed as b_w/b_h.

overlapping figure-ground segments, together with five structural descriptors as listed in Table 1. Suppose the pool R consists of m regions $\{r_1, \ldots, r_m\}$, we then use a clustering technique based on the spatial relationships between regions to partition them into several groups, by taking account of the constraint that the Euclidean distance of regional centroids must not be greater than Δ_c (see Alg. 1). It is motivated by the observation that each proper small target has a set of similar segmentation results with adjacent coordinates of the centroids over a limited range of sequential thresholds. The goal of clustering is to produce an over-complete coverage of potential targets that belong to the same cluster, which could be further used for measuring regional stability. In our experiments, we set:

$$\Delta_c = \frac{\min^2\left(b_{r_i w}, b_{r_j w}\right) + \min^2\left(b_{r_i h}, b_{r_j h}\right)}{4} \tag{5}$$

where the subscripts of b indicate the width w and height h of the regions r_i and r_j respectively.

Algorithm 1 Procedure for clustering regions

Input: set R of regions
Output: set $\mathbf{S}$ of clusters
$\mathbf{s}_1 = \{r_1\}$, $\mathbf{S} = \{\mathbf{s}_1\}$ ▷ initialization
for each region $r_j \in R$ **do**
  if $\forall \mathbf{s}_{exist} \in \mathbf{S}, \exists r_i \in \mathbf{s}_{exist} : \mathbf{D_c}(r_i, r_j) \leq \Delta_c$ **then**
    $\mathbf{s}_{exist} = \mathbf{s}_{exist} \cup \{r_j\}$ ▷ existing cluster
  else
    $\mathbf{s}_{new} = \{r_j\}$, $\mathbf{S} = \mathbf{S} \cup \{\mathbf{s}_{new}\}$ ▷ new cluster
  end if
end for

Suppose the spatial clustering algorithm produces n clusters, i.e., $\mathbf{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ where $\mathbf{s}_k$ denotes the k^{th} cluster. Due to the closely adjacent locations, all regions in each cluster could be viewed as multiple representations for a certain target over the corresponding range of thresholds. In fact, such regions are of interest since they share some homologous properties with the notion of stability, we can readily evaluate these properties and obtain the optimal subimage for this target.

Stage II In the second stage as shown in Alg. 2, regional comparisons of fill rate and aspect ratio are first invoked for seeking the maximally stable target region $\mathbf{r}$ in each cluster $\mathbf{s}_k$. Here, we directly consider the global minima of two functions $\mathbf{D_f}$ and $\mathbf{D_a}$ for measurement, by which we mean that the optimal pair of regions is more likely to leave the binary representation unchanged and hence reflects the stability associated with a particular target. After two optimal pairs of regions (r_i, r_j) and (r_m, r_n) are found, we select the largest region as the choice of $\mathbf{r}$ for the purpose of accurate target representation. By using the bounding box of $\mathbf{r}$, the optimal subimage $G_{\mathbf{r}}$ for $\mathbf{s}_k$ can be extracted from the input gray-scale image G.

As discussed at the beginning of this subsection, we then compute the Otsu's threshold ϑ and use it to segment the target region $\mathbf{r}^{\vartheta}$ from the subimage $G_{\mathbf{r}}$. It should be noted that the segmentation result may have multiple binary regions, our **SEGMENT** function only returns one region which has the longest boundary, because we consider that any subimage $G_{\mathbf{r}}$

contains only a unique small target, that is, all regions in each cluster $\mathbf{s}_k$ are homologous as mentioned before.

The next step is to investigate whether $\mathbf{r}^{\vartheta}$ is subject to the constraint with regard to regional area. In order to ensure that each resulting region is stable over at least a threshold interval δ, we proceed to segment $G_{\mathbf{r}}$ using two specified thresholds ϑ_- and ϑ_+ in the interval spanned by δ, and obtain two output regions in the same manner as $\mathbf{r}^{\vartheta}$. Finally, $\mathbf{r}^{\vartheta}$ is rejected if the area variation $\mathbf{D}_{\mathbf{r}}$ exceeds Φ_r, otherwise it is added to the set R_T of stable regions.

Algorithm 2 Procedure for generating the set of stable regions

Input: gray-scale image G, and set $\mathbf{S}$

Output: set R_T of stable regions

1: $R_T = \varnothing$

2: **for** each cluster $\mathbf{s}_k \in \mathbf{S}$ **do**

3: $\mathbf{r} = \arg \max_{\mathbf{r} \in \{r_i, r_j, r_m, r_n\}} |\mathbf{r}|$

$$where \begin{cases} (r_i, r_j) = \arg \min_{r_i, r_j \in \mathbf{s}_k} \mathbf{D}_{\mathbf{f}}(r_i, r_j) \\ (r_m, r_n) = \arg \min_{r_m, r_n \in \mathbf{s}_k} \mathbf{D}_{\mathbf{a}}(r_m, r_n) \end{cases}$$

4: $G_{\mathbf{r}} = \mathbf{EXTRACT}(G, \mathbf{r})$

5: $\vartheta = \mathbf{OTSU}(G_{\mathbf{r}})$

6: $\mathbf{r}^{\vartheta} = \mathbf{SEGMENT}(G_{\mathbf{r}}, \vartheta)$

7: $\mathbf{r}^{\vartheta_-} = \mathbf{SEGMENT}(G_{\mathbf{r}}, \vartheta_-)$ ▷ $\vartheta_- = \vartheta - \delta/2$

8: $\mathbf{r}^{\vartheta_+} = \mathbf{SEGMENT}(G_{\mathbf{r}}, \vartheta_+)$ ▷ $\vartheta_+ = \vartheta + \delta/2$

9: **if** $\mathbf{D}_{\mathbf{r}}(\mathbf{r}^{\vartheta_-}, \mathbf{r}^{\vartheta_+}) \le \Phi_r$ **then**

10: $R_T = R_T \cup \{\mathbf{r}^{\vartheta}\}$

11: **end if**

12: **end for**

Obviously, the performance of the stable region extraction is typically associated with the parameter Φ_r, which is important and is case-sensitive with respect to the size of the small target. Society of Photo-Optical Instrumentation Engineers defines a small target to have a total spatial extent of less than 80 (9×9) pixels [7]. This classification includes: point source targets, small extended targets, and clusters of point source targets and small extended targets [23]. However, considering the variations of the target size, we express Φ_r as a certain percentage Δ_r of the relationship ϕ_r between the regional areas of $\mathbf{r}^{\vartheta_-}$ and $\mathbf{r}^{\vartheta_+}$ (cf. Eq. (6)). In the implementation of the proposed stability-based detection algorithm, we set the size of the small target $t_s = 100$ (10×10) to ensure that our algorithm adapts to the changes of the target size in a wide range. In Section 3.2, we will discuss the influence of the two parameters t_s and Δ_r.

$$\Phi_r = \Delta_r \phi_r, \tag{6}$$

where

$$\phi_r = \begin{cases} \max\left(\left|\mathbf{r}^{\vartheta_-}\right|, \left|\mathbf{r}^{\vartheta_+}\right|\right), & \text{if } \min\left(\left|\mathbf{r}^{\vartheta_-}\right|, \left|\mathbf{r}^{\vartheta_+}\right|\right) \ge t_s \\ t_s, & \text{otherwise} \end{cases}.$$

Figure 3 shows the detection results by applying our stability-based algorithm (RSt) to several video frames from three benchmark datasets. The stability map M_T is a binary image of the same size as the input. The on pixels correspond to all of the stable regions in R_T, and all other pixels are off. In the first row of Fig. 3, we see that our algorithm successfully achieves the initial objective of detecting the targets of different sizes.

The principal drawback of RSt is that there exist several other target regions which are the false alarms. This is illustrated in Fig. 4, in which other two small regions stemming from the clusters #93 and #96 are respectively obtained. Although such regions also satisfy our stability criteria, they actually are erroneous results which will lead to the performance deterioration in terms of the detection accuracy. Intuitively, with reference to the three subimages in the lower left corner of Fig. 4c, humans prefer to focus attention on the first subimage due to high contrast stimulus. As Fig. 4e shows, the average saliency values of three regions are 0.281, 0, and 0.001 respectively. Thus the saliency of each stable region can be employed to remove the latter two regions from the stability map M_T. In the next subsection, an improved saliency map generation method by exploiting local contrast mechanism is put forward.

2.3 Regional Saliency

Based on the principle of center-surround contrast, we directly define pixel-level saliency by comparing the *Lab* color vector of each pixel with its Gaussian blurred version. This simple process corresponds to the local difference of the center and surrounding color distributions. In order to suppress the uniform patterns and detect small salient regions, a large filter scale should be chosen for the Gaussian kernel. Consequently, we blur each channel of the *Lab* color space by a Gaussian low-pass filter ω of size $3\sigma \times 3\sigma$ with the standard deviation $\sigma = \min(W, H)/\sigma_s$, where W and H indicate the width and height of the input image, and the parameter σ_s controls the strength of weighting. Unlike some elaborate filter methods, here we roughly set the scale parameter of the Gaussian kernel instead of finding the optimal scale or exploiting multi-scale fusion, because we focus only on the entire saliency of each candidate target rather than strong details. Finally, the saliency map M_A can be formulated as Eq. (7) and be further normalized to the range [0, 1].

$$M_A = \|(L, a, b) - (L_\omega, a_\omega, b_\omega)\|^2. \tag{7}$$

where L_ω, a_ω and b_ω are the Gaussian blurred versions of L, a and b respectively.

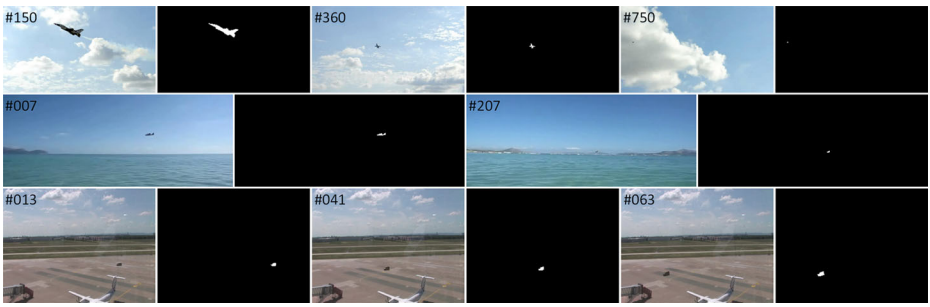


Fig. 3 Visual results of the proposed RSt algorithm

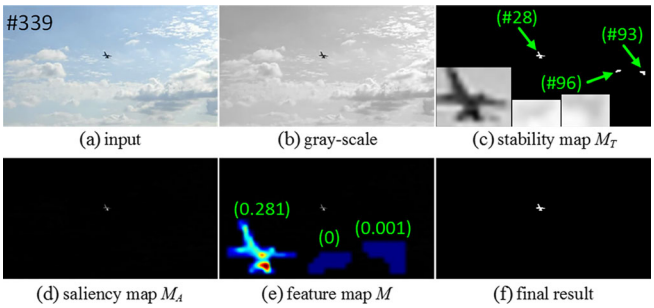


Fig. 4 Principal drawback of the proposed RSt algorithm

Although most saliency methods are not specifically designed for detecting small targets, some of them are suitable for this task. Li *et al.* think the SR method [12] can work well in detecting small salient regions where the center-surround contrast is very strong [16]. Considering the salient regions of different sizes, they design the HFT method [16] for detecting both large and small salient regions, which achieves the state-of-the-art performance.

In Fig. 5, we show the saliency maps produced by applying ten saliency methods to an input frame from the first dataset. The proposed saliency method, denoted as "RSa", correctly detects the small blob of interest in the color image against other nine methods. An obvious difference with them is the nature of our method which highlights the small salient region and simultaneously suppresses the rest of the image. In Section 3.2, we will discuss the influence of the parameter σ_s.

Although the proposed saliency method does a better job, there still exist two severe drawbacks. This is illustrated with another example in Fig. 6. The first is that, the resulting saliency map is a gray-scale intensity image, so we need to perform the thresholding operation to get a binary segmentation for the task of target detection. Unfortunately, this also raises the

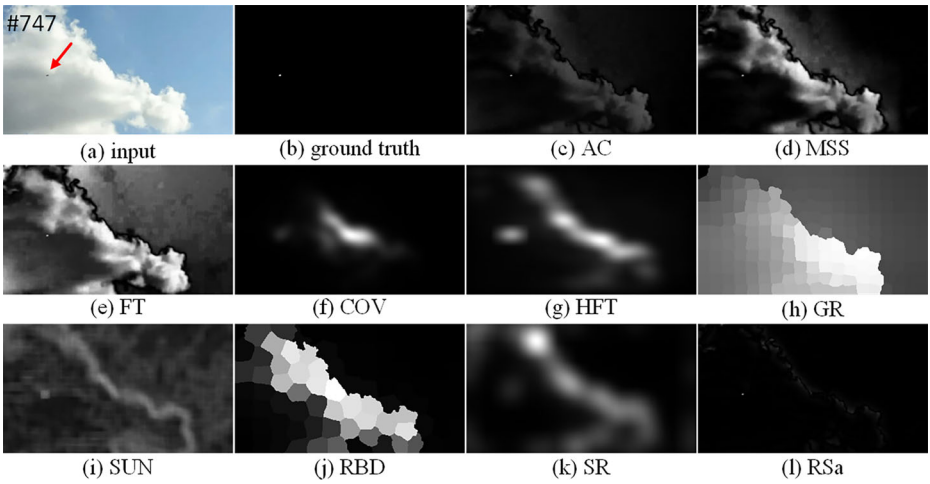


Fig. 5 Visual comparison of saliency maps. (**a**) input image, (**b**) ground truth, and saliency maps produced by using (**c**) Achanta *et al.* [2], (**d**) Achanta and Süsstrunk [1], (**e**) Achanta *et al.* [3], (**f**) E. Erdem and A. Erdem [9], (**g**) Li *et al.* [16], (**h**) Yang *et al.* [22], (**i**) Zhang *et al.* [24], (**j**) Zhu *et al.* [27], (**k**) Hou and Zhang [12], and (**l**) our RSa method

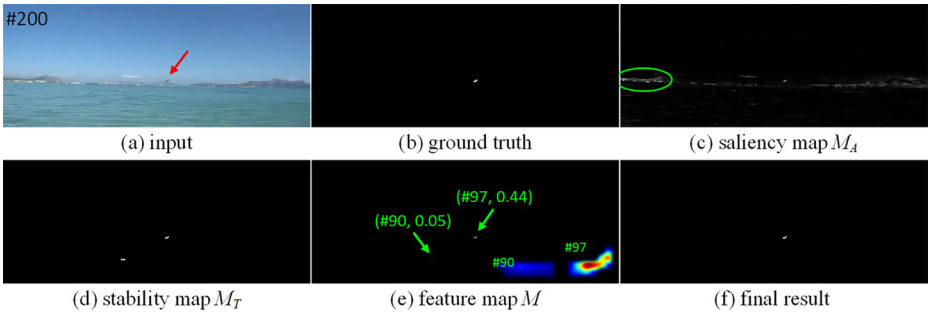


Fig. 6 Two severe drawbacks of the proposed RSa method

question of how to choose one or more suitable thresholding values automatically as mentioned before. The second drawback stems from the detection mechanism of our computational model. We evaluate the saliency of a local image region by using the contrast with respect to its surrounding area, so the regions with strong center-surround difference will be assigned nearly the same or higher saliency values. As Fig. 6c illustrates, the labeled elliptical area contains several small salient regions. Obviously, all these regions are the false alarms and are also difficult to be removed from the saliency map effectively.

As Fig. 6d shows, our solution to these problems is to invoke the stability map. Note that there are no detected stable regions in the same position as labeled in Fig. 6c, that is to say, such salient regions do not satisfy our stability criteria. In addition, there is no need to consider the thresholding of the saliency map, because we can exploit the entire saliency of each stable region (cf. Fig. 6e) to automatically determine whether it should be retained.

3 Experiments

3.1 Experimental Setup

We present empirical evaluation and analysis of the proposed RSS model on three benchmark datasets of different scenes with manually labeled ground truth. Some statistics and features of these datasets are summarized in Table 2. Since both stability and saliency maps are computed and combined to detect small targets in our model, the evaluation measures are divided into three parts: 1) stability-based detection methods, 2) saliency methods, and 3) integration models via combining the former two parts. In all experiments, we only detect the targets

Table 2 Description of three evaluation datasets

No.	Background	Images	Features
1	Sky	805	Frames #001~#752: Single target; Frames #753~#805: No target
2	Sea-Sky	208	Single target
3[a]	Ground	80	Single target

[a] Dataset 3 is available at http://people.ee.ethz.ch/~dragonr/943/ [8], in which each image is shrunk to 20 % of the original image size for the purpose of demonstration in our experiments

which do not connect to the image boundary, and the target sizes are between 4 and $0.2 \times W \times H$ where W and H indicate the width and height of the input image.

In the first part, we evaluate our stability-based detection method (RSt) in comparison with MSER [18]. For the MSER method, we directly extract the co-variant regions utilizing the VL_MSER function of the VLFeat open source library [21]. The main parameters are discussed in Section 3.2, and the statistical metrics and results are presented in Section 3.3.

In the second part, we compare the proposed saliency method (RSa) with nine state-of-the-art methods including AC [2], MSS [1], FT [3], COV [9], HFT [16], GR [22], SUN [24], RBD [27], and SR [12]. For these baseline methods, we run the C++ and MATLAB codes provided by Borji *et al.* [5] to generate the saliency maps. The evaluation measures will be discussed in Section 3.3.

In the third part, we combine the stability-based and saliency-based methods via pixel-wise multiplication as our RSS model, and evaluate these integration models in terms of precision, recall, and F-measure in Section 3.3. The video demos of our model on three benchmark datasets can be found in the project webpage: http://www.loujing.com/rss-small-target. We demonstrate in the first video that our model works stably when the target size varies within a large range. It should be noted that although each resulting demo is shown in the form of video, all video files are created from the sequential frames in which each frame is obtained by applying the proposed RSS model to the individual image.

3.2 Parameter Analysis

Three main parameters of the VL_MSER function are sample step δ, minimum diversity d_m of region, and maximum variation v_m (absolute stability score) of regions. For the proposed RSS model, four parameters are involved in its implementation: target size t_s, sample step δ, threshold Δ_r of area variation, and weight σ_s of standard deviation of the Gaussian low-pass filter. Figures 7 and 8 show the influence of these parameters on the $\bar{F}$ scores on each dataset. The black baseline, denoted as "Average", is simply the average $\bar{F}$ scores of all datasets. In the next subsection, we will give the definition of $\bar{F}$.

First, the sample step size δ has a direct impact on the detection results. For MSER, the $\bar{F}$ scores start to drop significantly when δ is greater than 20, and for RSS, this threshold is 16. Second, the parameters d_m, v_m, t_s, and σ_s have similar influence on all test datasets. Overall, the $\bar{F}$ scores show a clear upward trend as the parameter values increase, and then start to drop slightly when the parameter values exceed certain thresholds. For the saliency detection module, our RSa method is not sensitive to the parameter σ_s on the datasets 1 and 2, but on the third dataset, the $\bar{F}$ scores reach the peak when σ_s is equal to 16. Accordingly, we set $\delta =$

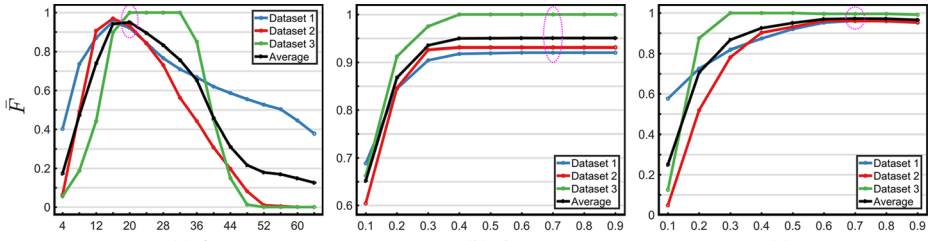


Fig. 7 Parameter analysis of MSER

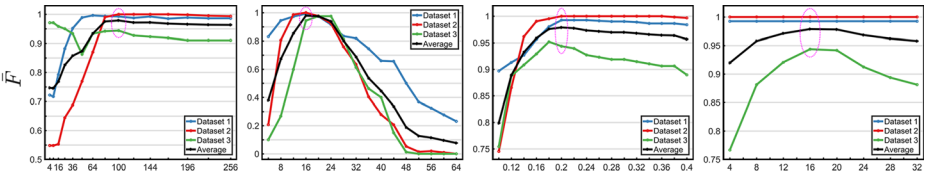


Fig. 8 Parameter analysis of RSS

20, $d_m = 0.7$, and $v_m = 0.7$ in the implementation of MSER. And for RSS, we set $t_s = 100$, $\delta = 16$, $\Delta_r = 0.2$, and $\sigma_s = 16$. Moreover, our RSS model only detects the regions with local intensity minima, so the parameters *DarkOnBright* and *BrightOnDark* of the VL_MSER function are set to 1 and 0 respectively.

3.3 Results

1) *Evaluation of Stability-Based Detection Methods:* We evaluate our RSt method in comparison with MSER using the evaluation metric of detection rate (DR), which is defined as the number of target regions overlaps. For each individual frame, *Precision* (*P*) is the ratio of the number of successfully detected targets to the number of all detected targets, and *Recall* (*R*) is the ratio of the number of successfully detected targets to the number of ground truth targets. To combine precision and recall, a standard *F-measure* (*F*) is defined as Eq. (8). In our experiments, a small target is considered as being detected if the following criterion is satisfied. That is, the intersection of a detected target and the corresponding ground truth is no less than 50 %. Then, the average values of precision, recall and F-measure, denoted by $\overline{P}$, $\overline{R}$ and $\overline{F}$ respectively, are obtained over the whole dataset.

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{8}$$

As Table 3 shows, our RSt method achieves the highest recall scores on all datasets. This means that the MSER method has the problem of missing targets. However, the RSt method has lower precision scores especially on the third dataset, thereby decreasing the values of F-measure. This is mainly caused by the existence of the false alarms. In Section 2.2, we have thrown up this issue and introduced the saliency method to tackle it. Furthermore, our RSt method does not detect the foreground regions which connect to the image boundary, because we think this kind of regions is neither complete nor connective. This mechanism results in that RSt misses the only one target in the #752

Table 3 Statistical comparison of stability-based detection methods (%)

Method	Dataset 1			Dataset 2			Dataset 3			Average		
	$\overline{P}$	$\overline{R}$	$\overline{F}$	$\overline{P}$	$\overline{R}$	$\overline{F}$	$\overline{P}$	$\overline{R}$	$\overline{F}$	$\overline{P}$	$\overline{R}$	$\overline{F}$
MSER [18]	95.3	98.0	96.2	95.4	97.1	96.0	99.4	100	99.6	96.7	98.4	97.2
RSt	89.9	99.3	92.4	90.6	100	93.5	51.0	100	64.5	77.1	99.8	83.4

frame of the first dataset. For MSER, we also invoke this detection mechanism to make a fair comparison.

2) *Evaluation of Saliency Methods:* One of the most widely used metrics for saliency method evaluation is the *Precision-Recall* metric. For a saliency map, we can convert it to a binary mask B using a fixed threshold which varies from 0 to 255, and compute *Precision* (P) and *Recall* (R) by comparing B with its ground truth A on each threshold:

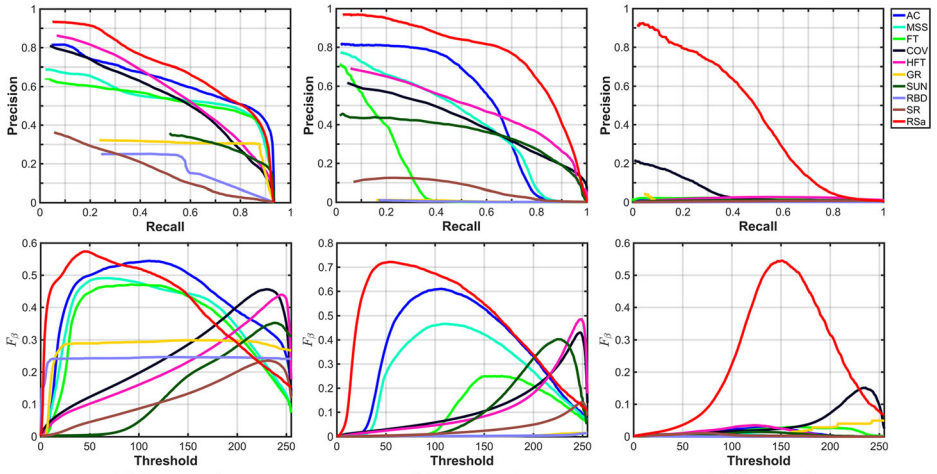
$$Precision = \frac{|B \cap A|}{|B|}, \quad Recall = \frac{|B \cap A|}{|A|}. \tag{9}$$

We also report the *F-measure* (F) metric which jointly considers precision and recall with a non-negative weight β:

$$F_\beta = \frac{\left(1+\beta^2\right) \times Precision \times Recall}{\beta^2 \times Precision + Recall}. \tag{10}$$

where we also set $\beta^2 = 0.3$ to emphasize the precision as suggested in [3].

The precision-recall and F-measure curves on three datasets are plotted in Fig. 9, and the average scores of precision, recall and F-measure (i.e., $\overline{P}$, $\overline{R}$ and $\overline{F}$) are reported in Table 4. Overall, our RSa method has the highest $\overline{R}$ scores on all the datasets, and outperforms all other methods with large margins. On the datasets 2 and 3, we also obtain the highest $\overline{F}$ scores. Besides, the AC method achieves close performance and performs slightly better than our RSa method in terms of $\overline{F}$ on the first dataset. With regard to the precision metric, the RSa method has no advantage due to our saliency detection mechanism. We emphasize the detection of small salient regions, which leads to the weak responses to the slightly bigger salient regions. In Sections 2.2 and 2.3, it has been pointed out that we employ RSa to assist the stability-based method to remove the false alarms, and we only focus on the roughly entire saliency of the candidate regions rather than strong details.



(a) Dataset 1 (b) Dataset 2 (c) Dataset 3

Fig. 9 Precision-Recall (*top*) and F-measure (*bottom*) curves of ten saliency methods on three datasets

Table 4 Statistical comparison of saliency methods (%)

Method	Dataset 1			Dataset 2			Dataset 3			Average		
	$\overline{P}$	$\overline{R}$	$\overline{F}$	$\overline{P}$	$\overline{R}$	$\overline{F}$	$\overline{P}$	$\overline{R}$	$\overline{F}$	$\overline{P}$	$\overline{R}$	$\overline{F}$
AC [2]	48.3	60.8	42.7	40.3	61.7	36.2	41.6	0.89	1.12	43.4	41.1	26.7
MSS [1]	44.2	53.3	36.7	38.9	50.8	27.8	27.1	0.25	0.33	36.7	34.8	21.6
FT [3]	45.7	50.0	34.5	35.4	28.0	10.6	54.3	1.23	1.47	45.1	26.4	15.5
COV [9]	77.4	26.8	24.9	93.0	11.2	12.2	47.2	4.18	3.99	72.5	14.1	13.7
HFT [16]	83.1	22.6	21.4	94.6	10.6	11.6	50.3	1.00	1.28	76.0	11.4	11.4
GR [22]	55.6	29.8	27.9	81.0	0.23	0.30	46.1	1.28	1.33	60.9	10.4	9.84
SUN [24]	87.1	12.5	14.3	85.6	12.4	12.0	53.3	0.40	0.52	75.3	8.46	8.92
RBD [27]	42.4	24.6	24.1	66.9	0.22	0.29	8.20	0.02	0.03	39.2	8.29	8.14
SR [12]	59.2	11.5	11.7	79.1	2.51	3.08	36.7	0.26	0.33	58.3	4.76	5.04
RSa	31.8	80.8	41.2	39.0	82.2	47.9	48.6	45.7	23.1	39.8	69.6	37.4

3) *Evaluation of Integration Models:* Based on the stability and saliency methods mentioned before, we finally introduce a set of integration models which is generated in a pixel-wise multiplication manner as our RSS model. For the purpose of demonstration, we select three saliency methods including AC, FT, and HFT, and combine them with MSER and RSt. The proposed RSa method is also employed with MSER for evaluation. For these integration models, we still use the DR metric to compute the average scores of precision, recall and F-measure as mentioned in Section 3.3, and the experimental setups follow in Sections 2.1 and 3.3.

Among all the models shown in Table 5, our RSS model which combines RSt and RSa is the highest performing model on the first two datasets. We also obtain the highest recall score on the third dataset, while the models integrating MSER have higher precision performance than ours. Overall, the evaluation results of these models mainly depend on the accuracy of the stability-based methods. From Tables 3, 4 and 5, we can see the improvements of the precision

Table 5 Statistical comparison of integration models (%)

Model	Dataset 1			Dataset 2			Dataset 3			Average		
	$\overline{P}$	$\overline{R}$	$\overline{F}$	$\overline{P}$	$\overline{R}$	$\overline{F}$	$\overline{P}$	$\overline{R}$	$\overline{F}$	$\overline{P}$	$\overline{R}$	$\overline{F}$
MSER+AC	98.0	98.0	98.0	97.1	97.1	97.1	100	100	100	98.4	98.4	98.4
MSER+FT	98.0	98.0	98.0	97.1	97.1	97.1	100	100	100	98.4	98.4	98.4
MSER+HFT	98.0	98.0	98.0	97.1	97.1	97.1	98.8	98.8	98.8	98.0	98.0	98.0
MSER+RSa	98.0	98.0	98.0	97.1	97.1	97.1	100	100	100	98.4	98.4	98.4
RSt+AC	99.3	99.3	99.3	100	100	100	75.0	91.3	80.4	91.4	96.8	93.2
RSt+FT	99.3	99.3	99.3	98.6	99.0	98.7	83.5	100	89.0	93.8	99.4	95.6
RSt+HFT	99.2	99.3	99.2	100	100	100	51.3	58.8	53.8	83.5	86.0	84.3
RSS (RSt+RSa)	99.3	99.3	99.3	100	100	100	91.9	100	94.4	97.0	99.8	97.9

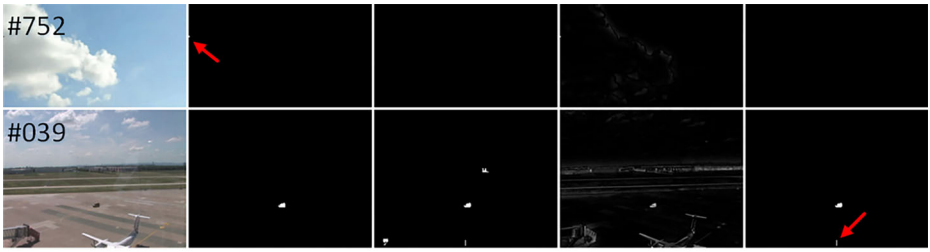


Fig. 10 Hard image cases of RSS in detection small targets

performances with the introduction of the saliency methods. This indicates the validity of the proposed RSS model, which integrates the stability-based and saliency-based detection techniques into a common framework, instead of just using a single one in isolation.

Although RSS has performed well in the experiments, it does fail in certain cases. Our RSS model does not achieve the evaluation values of 100 % on the datasets 1 and 3. Except for the only one target missed on the first dataset discussed in Section 3.3 (see the first row in Fig. 10), the main problem is the existence of the false alarms. RSS could not satisfactorily remove the erroneous stable regions with strong local-surround contrast for a "hard" image shown in the second row in Fig. 10.

As mentioned in Section 2.1, RSS only retains the candidate regions whose saliency values exceed the average of all stable regions. The issue of false alarms can be tackled by changing this constraint to only retain one region having the maximum saliency value. But, the modified version only works well on the image which contains a single small target. When multiple small targets appear in a scene, it does fail due to the problem of missing targets, which is often considered more serious than the issue of false alarms. Although each image in the test datasets contains at most one target, we still use the original constraint, considering the potential possibility for RSS to detect more than one target in some specific application scenarios.

4 Conclusion

Throughout this paper, we have tackled the problem of small target detection in the visual band. Considering the intrinsic properties of small targets, a novel model combining regional stability and saliency is designed to help in figure-ground segregation. To validate the effectiveness of the proposed framework, a set of integration models is introduced by exploiting several existing methods. Experimental results show the performance improvement of the proposed integration model in terms of precision and recall.

Although this work focuses on small target detection in a color image, the proposed model is also suitable for detecting certain target which has nearly uniform intensity or color, e.g., infrared small target, and featureless object with intermediate size. In the future, we plan to apply the proposed technique to detect infrared small target, or use it as a complementary solution to the existing infrared target detection methods. We will test the proposed model in more cluttered scenes and increase the performance speed of our detection system. In addition, due to the lack of complex visual features, we plan to invoke more top-down cues (e.g., spot shape and entropy) to solve the problem of false alarms.

Acknowledgments The authors thank all of the anonymous reviewers for their insights and suggestions, which were very helpful in improving this manuscript. They thank Haiyang Zhang for useful discussions. They also thank Mei Zhang and Huaiping Zhang for their kind proofreading of the manuscript. This work is supported by the National Natural Science Foundation of China under Grant 61231014.

References

1. Achanta R, Süsstrunk S (2010) Saliency detection using maximum symmetric surround. In: Proc. IEEE Int. Conf. Image Process., 2653–2656
2. Achanta R, Estrada F, Wils P, Süsstrunk S (2008) Salient region detection and segmentation. In: Proc. Int. Conf. Comput. Vis. Syst., 66–75
3. Achanta R, Hemami S, Estrada F, Süsstrunk S (2009) Frequency-tuned salient region detection. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 1597–1604
4. Bae T-W, Zhang F, Kweon I-S (2012) Edge directional 2D LMS filter for infrared small target detection. Infrared Phys Technol 55(1):137–145
5. Borji A, Cheng M-M, Jiang H, Li J (2015) Salient object detection: a benchmark. IEEE Trans Image Process 24(12):5706–5722
6. Chen H-Y, Leou J-J (2012) Multispectral and multiresolution image fusion using particle swarm optimization. Multimed Tools Appl 60(3):495–518
7. Chen CLP, Li H, Wei Y, Xia T, Tang YY (2014) A local contrast method for small infrared target detection. IEEE Trans Geosci Remote Sens 52(1):574–581
8. Dragon R, Ostermann J, Van Gool L (2013) Robust realtime motion-split-and-merge for motion segmentation. In Proc. Ger. Conf. Pattern Recognit., 425–434
9. Erdem E, Erdem A (2013) Visual saliency estimation by nonlinearly integrating features using region covariances. J Vis 13(4):1–20, 11
10. Gao C, Meng D, Yang Y, Wang Y, Zhou X, Hauptmann AG (2013) Infrared patch-image model for small target detection in a single image. IEEE Trans Image Process 22(12):4996–5009
11. Gonzalez RC, Woods RE (2002) Introduction. In: Digit. Image Process., 2nd ed. Prentice Hall. ch. 1: 12–13
12. Hou X, Zhang L (2007) Saliency detection: a spectral residual approach. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 1–8
13. Kim S, Lee J (2012) Scale invariant small target detection by optimizing signal-to-clutter ratio in heterogeneous background for infrared search and track. Pattern Recogn 45(1):393–406
14. Lee E, Gu E, Park K (2015) Effective small target enhancement and detection in infrared images using saliency map and image intensity. Opt Rev 22(4):659–668
15. Li W, Pan C, Liu L-X (2009) Saliency-based automatic target detection in forward looking infrared images. In: Proc. IEEE Int. Conf. Image Process., 957–960
16. Li J, Levine MD, An X, Xu X, He H (2013) Visual saliency based on scale-space analysis in the frequency domain. IEEE Trans Pattern Anal Mach Intell 35(4):996–1010
17. Li Y, Liang S, Bai B, Feng D (2014) Detecting and tracking dim small targets in infrared image sequences under complex backgrounds. Multimed Tools Appl 71(3):1179–1199
18. Matas J, Chum O, Urban M, Pajdla T (2004) Robust wide-baseline stereo from maximally stable extremal regions. Image Vis Comput 22(10):761–767
19. Otsu N (1979) A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern 9(1): 62–66
20. Qi S, Ma J, Tao C, Yang C, Tian J (2013) A robust directional saliency-based method for infrared small-target detection under various complex backgrounds. IEEE Geosci Remote Sens Lett 10(3): 495–499
21. Vedaldi A, Fulkerson B (2008) VLFeat: An open and portable library of computer vision algorithms, version 0.9.19. http://www.vlfeat.org
22. Yang C, Zhang L, Lu H (2013) Graph-regularized saliency detection with convex-hull-based center prior. IEEE Signal Process Lett 20(7):637–640

23. Zhang W, Cong M, Wang L (2003) Algorithms for optical weak small targets detection and tracking: Review. In: Proc. IEEE Int. Conf. Neural Networks Signal Process., 643–647
24. Zhang L, Tong MH, Marks TK, Shan H, Cottrell GW (2008) SUN: a Bayesian framework for saliency using natural statistics. J Vis 8(7):1–20, 32
25. Zhou C, Liu C (2015) An efficient segmentation method using saliency object detection. Multimed Tools Appl 74(15):5623–5634
26. Zhu B, Xin Y (2015) Effective and robust infrared small target detection with the fusion of polydirectional first order derivative images under facet model. Infrared Phys Technol 69:136–144
27. Zhu W, Liang S, Wei Y, Sun J (2014) Saliency optimization from robust background detection. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2814–2821

Jing Lou received the BE and ME degrees from Nanjing University of Science and Technology, Nanjing, Jiangsu, P.R. China, where he is currently working toward the PhD degree. His research interests include computer vision, image processing, and machine learning.

Wei Zhu received the BE degree in software engineering from Nanjing University of Science and Technology (NUST), Nanjing, Jiangsu, P.R. China. He is currently working toward the PhD degree in the School of Computer Science and Engineering, NUST. His research interests include image processing and deep learning.

Huan Wang received the PhD degree in pattern recognition and intelligent system from Nanjing University of Science and Technology (NUST), Nanjing, Jiangsu, P.R. China. He is currently a lecturer with the School of Computer Science and Engineering, NUST. His current research interests include pattern recognition, robot vision, image processing, and artificial intelligence.

Mingwu Ren received the PhD degree in pattern recognition and intelligent system from Nanjing University of Science and Technology (NUST), Nanjing, Jiangsu, P.R. China, in 2001. He is currently a Professor with the School of Computer Science and Engineering, NUST. His current research interests include computer vision, image processing, and pattern recognition.