A Saliency Based Human Detection Framework for Infrared Thermal Images

Xinbo Wang[1,2,3], Dahai Yu[1,2,3(✉)], Jianfeng Han[1,2,3], and Guoshan Zhang[1,2,3]

[1] Tianjin University, Tianjin, China
wxb.tju@gmail.com, yudahai_dublin@hotmail.com
[2] Tianjin ShenQi Technology Co., Ltd., Tianjin, China
[3] Tianjin University of Commerce, Tianjin, China

Abstract. In this paper, a novel saliency framework for crowd detection in infrared thermal images is proposed. In order to obtain the optimal classifier from a large amount of data, the process of training consists of the following four steps: (a) a saliency contrast algorithm is employed to detect the regions of interest; (b) standard HOG features of the selected interest areas are extracted to represent the human object; (c) the extracted features, which are prepared for training, are optimized based on a visual attention map; (d) a support vector machine (SVM) algorithm is applied to compute the classifier. Finally, we can detect the human precisely after high-saliency areas of an image are input into the classifier. In order to evaluate our algorithm, we constructed an infrared thermal image database collected by a real-time inspection system. The experimental results demonstrated that our method can outperform the previous state-of-the art methods for human detection in infrared thermal images, and the visual attentional techniques can effectively represent prior knowledge for features optimization in a practicable system.

Keywords: Human detection · Saliency algorithm
Visual attention map · Infrared thermal images · Vision application

1 Introduction

With the development of imaging technology, infrared thermal (IRT) imaging systems have been widely used in many fields, such as local surveillance, fire detection and human inspection. In contrast to color imaging, IRT imaging is based on the temperature of objects, i.e. IRT imaging can actually represent the radiation range of infrared energy emitted, transmitted and reflected by any object. Therefore, surveillance integrated with IRT imaging can show a visual picture even in dark areas or blurred views, which makes it easier to detect the target, and this is why infrared target detection has received attention from many researchers in recent years. However, as a key technique, traditional target detection schemes have their limitations. For example, the threshold segmentation used in surveillance is sensitive to image content and methods based on background modelling cannot find stationary targets.

J. Yang et al. (Eds.): CCCV 2017, Part I, CCIS 771, pp. 282–294, 2017.
https://doi.org/10.1007/978-981-10-7299-4_23

Enlightened by the fact that the object to be detected in an image is usually correlated with its salience, many researchers combine a salience model with object detection. Ren et al. [1] introduced a new saliency model based on a region-based solution and adopted it in encoding features for object recognition. Jung and Kim [2] proposed a novel unified spectral-domain approach to build a saliency map, and applied it in object segmentation. According to their experiments, object detection fused with a saliency model can achieve a higher precision. Therefore, we intend to utilize a saliency model in infrared object detection. In fact, there have been many works in this sphere so far. In view of the different visual characteristics between small targets and background clutter, Qi et al. [3] presented a new directional saliency-based method, which is competitive with the state-of-the-art methods, especially for images with various typical complex backgrounds. Another effective approach for small maritime object detection is described in [4] in 2011, it used local minimum patterns (LMP) to obtain the saliency map and segment the potential regions. Then a fast clustering algorithm was utilized for localizing objects from segmented regions. In the same year, Han et al. [5] improved the local contrast scheme for calculating the saliency map and then obtained the target region precisely by adopting a threshold operation along with a traversal mechanism.

From this short review, we notice that most of the focus in infrared object detection is on small targets. However, it is sometimes difficult to locate the targets accurately. The reason can be attributed to two causes: (1) small targets that occupy only a few pixels are short of distinct features and are easily contaminated by unknown noise; (2) the contrast between targets and background is low, thus making them difficult to distinguish. In this paper, we start from another assumption and propose a scheme for the detection of certain targets which is more applicable and operable than small target detection in surveillance. Here, the detection target is a human, and a saliency model integrated with visual attention[1] technique is embedded in the scheme.

It is well known that, saliency detection can be formulated into two categories: bottom-up and top-down. The principle of the bottom-up methods is based on the center-surround mechanism, which is in line with the performance of the human vision system. Due to this reason, the definition of the saliency model in a bottom-up method is varied and it is hard to compare the features between point and point, area and area in the image. In the model calculation, we extract the underlying characteristics of the image first, such as brightness, color, direction, etc. Then the contrast of these characteristics is calculated. Finally, the saliency map is obtained, which can show the degree of significance of each pixel in the image.

In 1998, Itti et al. built a representative saliency detection model that implements RGB color channels, intensity and orientation as a standard benchmark for comparison [6]. Then, Harel and Koch improved Ittis model by adopting a Markov chain, and named it graph-based visual saliency (GBVS) [7]. Later,

[1] The term visual attention refers to the human observer who concentrates to a specific area of the visual scene, and processing is performed in a serial fashion.

Hou and Zhang developed a spectral residual model whose computation time is competitive [8]. Achanta et al. used a frequency-based approach with low-level features in Lab color space to measure the saliency of each pixel [9]. In addition, Bi et al. (2010) combined visual attention guidance and a local descriptors representation for generic object detection [10]. In general, bottom-up methods are fast, simple and automatic, but they are not suitable for detecting homogeneous and quite large objects.

In contrast, the definition of saliency in a top-down model is usually related to the specific object and detection task. For example, the saliency map always depends on the shape, location and size of the objects to be detected. In other words, the identifiable high-level information is often treated as an indicator to measure the saliency of the image in a top-down model. In the state-of-the-art, the basis of top-down saliency models is first to learn the prior knowledge related to the goal or task, and then adopt the prior knowledge to guide the process of visual attention. Therefore, the top-down model contains two stages: learning and decision-making. In the learning stage, the main task is to extract the underlying characteristics of the image and learn the differences between the target and background in order to obtain a probabilistic statistical model of the training set. In the decision-making stage, the model extracts the underlying characteristics of the image in the testing set, and then utilizes the statistical model to guide the influence of the underlying characteristics for visual attention.

There have been three design principles for the top-down saliency model so far.

(1) The saliency model based on simple parameter estimation. Itti [11] mentioned a multi-scale feature integration method for specific target detection in his doctoral thesis. Frintrop et al. [12] developed a VOCUS system for extracting the saliency associated with a specific goal. In addition, Lee and Lee [13] established the relationship between specific knowledge and the goal, and then utilized the model to detect faces.
(2) The saliency model based on production models [14–20]. Its core is first to build a model according to the conditions of density and prior probability, and then get a posterior probability distribution. In 2006, Navalpakkam and Itti [14] defined the signal-to-noise ratio (SNR) as the ratio between degrees of significance of the target and background and then computed linear parameters for the fusion of the underlying features by maximizing the SNR so that the saliency of target objects is higher than that of the background. Analogously, Oliva et al. [15] and Torralba [16] defined the saliency as finding the probability of a series of characteristics in an image. Gao et al. [17] defined the question as a one-to-many classification, and in his view, the significant features are those features that can best separate targets from background faultlessly.
(3) The saliency model based on discriminant models. Peters and Itti [21] learned what they call the gist features of people in a particular task and then used these to predict the gaze area in the next frame. In addition, Judd et al. [22] computed the location of the focus for human eyes, and

Borji [23] introduced a new method that learns the weight coefficients of the feature fusion by classifier instead of by parameter estimation. Recently, Yang and Yang [24] also developed a top-down saliency model, which is based on a Conditional Random Field CRF with latent variables, to detect various objects.

Actually, top-down methods perform better in object detection compared with bottom-up models, but they are time-consuming and limited due to the lack of high-level information in some images. Therefore in this paper, we introduce a bottom-up model and top-down model simultaneously considering they have different characteristics for accurate human detection.

Firstly, we automatically segment the regions-of-interest (ROI) that may contain the target, through a bottom-up saliency algorithm. Then, we use histograms of oriented gradients (HOGs) as a feature extraction technique for the detection of humans. In order to select proper features and reduce redundant features without losing important local information, we introduce a novel method to optimize standard HOG features by using a visual attention map. The visual attention map represents the distribution effects of human fixation points that are collected by eye tracker equipment. In this regard the visual attention map plays a role as prior knowledge, so we can treat the feature optimization as a top-down method. Finally, Support Vector Machines (SVM) is applied for recognizing humans on the basis of the above results. The block diagram is depicted in Fig. 1.

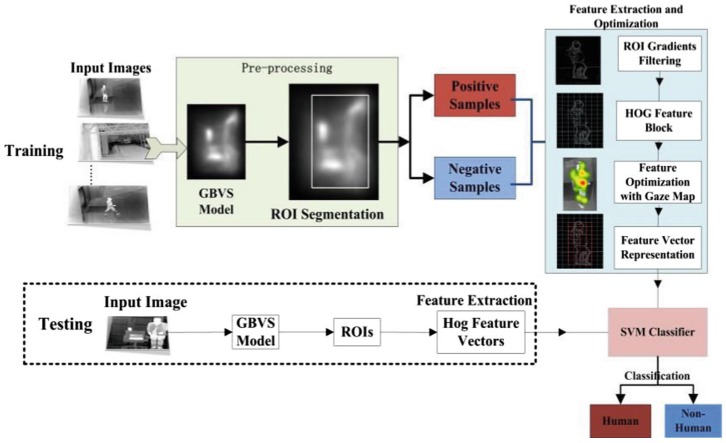

Fig. 1. Proposed framework.

The main contributions associated with this work are: (1) we practically analyse and develop a visual attention technique, which can be seen as a priori knowledge of the human subject, for feature optimization in a real-time object detection system; (2) we propose a new framework that combines a saliency

model and a HOG based feature process method to detect a known object rather than unknown small objects in infrared thermal images.

The rest of this paper is organized as follows. In Sect. 2, the proposed detection model is described in detail. Section 3 demonstrates the effectiveness of our algorithm through a series of experiments. Finally, concluding remarks and discussions on the future work are drawn in Sect. 4.

2 The Detection Framework

The detection framework in this paper involves two main processes: training and testing. The training process includes three key steps: (1) pre-processing, (2) feature extraction and optimization, (3) classifier construction, while the testing includes (1) pre-processing, (2) feature extraction and (3) classification.

2.1 Pre-processing

Pre-processing, which appears in both image training and testing has the effect of a preliminary screening. It is inspired by the essence of IRT imaging. As we know, the intensity value of infrared images reflects the temperature of the objects in the scene, as shown in Fig. 2(a). It is very important to emphasize that the greater the intensity the higher the temperature. Therefore, the object that we detect usually is located in regions where the background has a strong contrast with the objects, i.e. regions where the temperature (intensity value) of a human is higher relative to the temperature (intensity value) of background in most cases. Considering the observers often pay attention to those significant regions, here we can implement a bottom-up algorithm to choose some potential regions of interest (ROI) for the convenience of later processing and the reduction of computation.

As a matter of fact, there have been many developments in bottom-up algorithms. Generally, depending on the principle of each algorithm, they can be divided into four categories: (1) algorithms based on feature integration [5,25–28]; (2) algorithms based on image segmentation [29]; (3) algorithms based on frequency domain analysis [7,30]; (4) algorithms based on information theory [6,31,32]. Some representative algorithms have been de-scribed in the Introduction so we wont give a detailed description here. In a word, the motivation of ROI segmentation is to find some local regions with strong saliency. For this purpose, we choose the GBVS algorithm, to acquire the outline of regions with high saliency.

As a kind of classic bottom-up algorithm, the basic process is similar to the Itti algorithm. The first step is to extract features at multiple spatial scales. For example, we can compute different level of Gaussian features based on luminance, R, G, B, Y, orientations, etc. Here, given the characteristics of the IRT image, we extract only two kinds of features viz. luminance and orientation. Then we consolidate the feature maps using Markov theory, and obtain a saliency map. Finally we normalize the unique saliency map. The result of Fig. 2(a) processed by GBVS is shown in Fig. 2(b).

Fig. 2. An example of an Infrared Thermal Image (a) and its saliency map of (b).

2.2 Feature Extraction and Optimization

Feature extraction is an essential module in object recognition. Many types of features can be extracted from an image. In this paper, we mainly adopt the Histogram of Oriented Gradient (HOG) descriptor to represent IRT images. HOG was proposed by Dalal et al. in 2005 [33], and its core is that the appearance and shape of local targets within an image can be represented by the direction density distribution of its gradients or edges. In detail, first the image is divided into small regions referred to as cells. Then for each pixel in each cell, we construct a histogram of gradients or edges. Finally, we combine these histograms to constitute a feature descriptor. In this paper, because the potential regions (ROIs) have been selected previously, in this stage we compute the HOG features only for the cells in each ROI.

Even though we have introduced ROI segmentation before feature extraction in order to reduce computation compared with the traditional algorithm, we still have to face the problem that the training is usually time-consuming. Thus we integrate a feature optimization into feature extraction to select proper features with important local information and reduce the computation workload for training further. The specific process is shown in Fig. 3.

It is worth to notice that the proposed algorithm adds a selection of the obtained descriptor blocks based on a gaze distribution map which can represent the objective visual attention of observers. We generate the gaze distribution map with the help of eye tracker equipment. The eye tracker can record the fixation points of observers when they watch the image. If the observer paid more attention to one area of the image, there will be more fixation points on this area. Based on this result, we represent the density of fixation points using different intensity values of RGB on the image to demonstrate the gaze distribution. The result of gaze distribution is referred to as a visual attention map. In detail, some examples of a gaze distribution map are shown in Figs. 4 and 5. The colored areas (red/yellow/green) in the map represent the attention distribution of single human subjects and multi human subjects respectively, and the stronger the intensity value, the more important the degree of visual attention. When we choose the important descriptor blocks, we first set a threshold, and then count the number of pixels located in the corresponding attention area of the

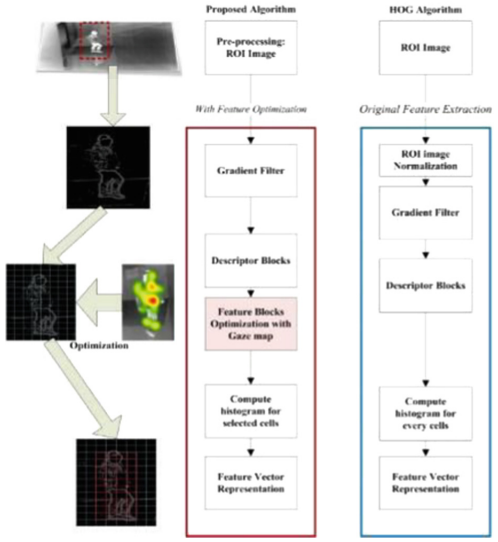

Fig. 3. The Process of feature extraction and optimization.

gaze distribution map for each cell. And only the blocks whose statistic is higher than a certain threshold are selected (the threshold is 40 for $[8 \times 8]$ cell in this paper). Finally, the selected blocks are weighted and normalized to get the HOG feature vectors for training.

2.3 SVM Classification

SVMs are a supervised learning mechanism, and its core is to find the optimal hyper-plane in a linear separable space which is mapped from high-dimensional inseparable space. Because SVMs are a very mature technique in object recognition, here we only describe the training data instead of elaborating its process at length. In this paper, the training data contains two sets. The positive set in which each image includes a human is selected manually from ROI images. And the negative set without a human is selected randomly from ROI images. In addition, the sizes of those sets are equal.

3 The Experiment

In order to demonstrate the effectiveness of our algorithm for human object detection in infrared thermal images, we collected 200 IRT images from website (60) and real-time inspection system (140) using a 640 by 380 resolution thermography camera. The number and size of human objects in the scene are varied. Specifically, we grouped all the images into two sets: training set and test set, with size 80 and 120 respectively. Further, 12 observers watched the image

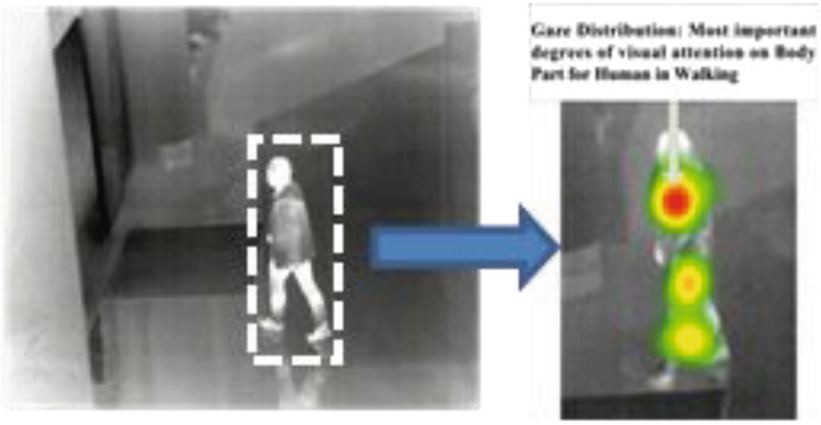

Fig. 4. A heat map of gaze distribution on single human object. (Color figure online)

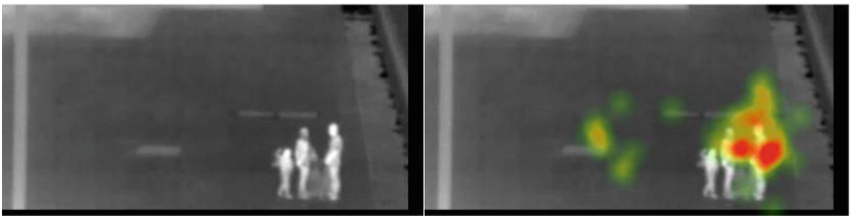

Fig. 5. A heat map of gaze distribution on multiple human objects. (Color figure online)

in a sequence of 80 training images one after another. Each image is shown on the screen about 3 s. During this processing, we collect the fixation results of each image using eye tracker equipment and its built-in software.

The experimental results for some example images are shown in Fig. 6. In the top group of images, there is only one hot object (person), while the rest con-tain many hot objects in different capture angles. Here we compare our method with four other algorithms, which are state-of-art saliency detection methods Itti and Koch [26], Yang and Yang [24], the traditional recognition methods [33], and an algorithm which is similar to ours apart from the absence of feature optimization. Obviously, the traditional method is not precise enough even for the single object detection. The results of our method and the other three methods are almost the same. Neverthe-less, it is worth to notice that our algorithm is superior to the other four algo-rithms whose results overlap or shows false positives. To sum up, our method performs best in the object detection.

To illustrate the validity of our algorithm at a deeper level, we computed the precision and error rate. In this regard, we manually selected the best region for each human for each testing image. Then we clustered all of the pixels into two groups: the correct pixels (the pixels that locate in the best region) and the error pixels (the pixels that locate in a region other than the best region), finally we

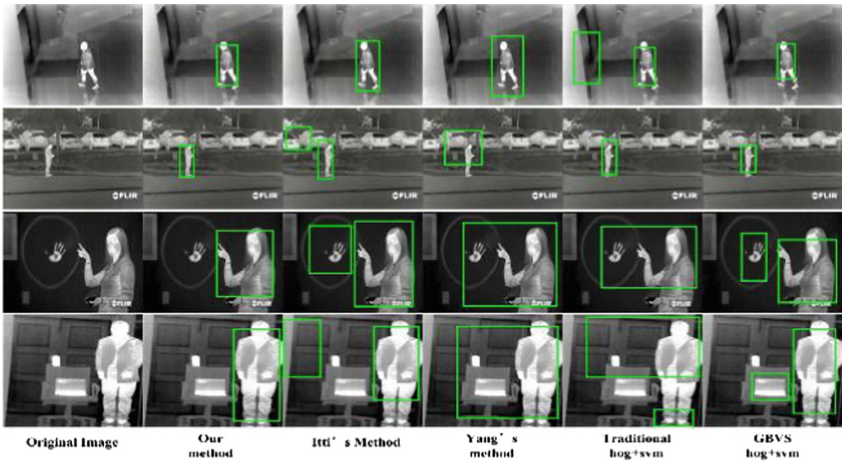

Fig. 6. Human detection comparison for single object.

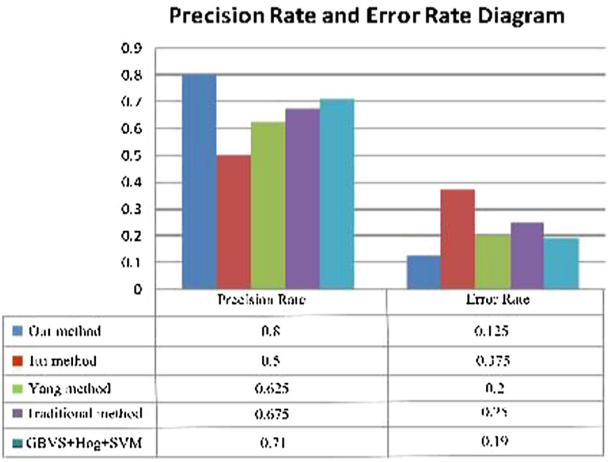

Fig. 7. The precision and error rate of different methods.

can cal-culate the precision rate and error rate respectively according to Eqs. (1) and (2). Figure 7 shows the statistical data. It is obvious that the precision rate of our algorithm is highest, and the order of other algorithms from high to low is the fourth algorithm (GBVS + HOG + SVM), the traditional method, Yangs method and Ittis method. Meanwhile, the order of the error rate is Ittis method, the tradi-tional method, Yangs method, the fourth kind of algorithm and our algorithm, which is the reverse of the order with the precision rate.

$$\text{precision rate} = \frac{\text{No. of correct pixels}}{\text{Total number of pixels in detection region}} . \qquad (1)$$

$$\text{error rate} = \frac{\text{No. of error pixels}}{\text{Total number of pixels in detection region}} . \quad (2)$$

In addition to this testing, we present more detection results using our algorithm under various conditions in IRT images (See Figs. 8 and 9). We can no-tice that all the people can be detected accurately. Finally, the capture sensor camera used was FLIR FC-S series. The eye tracker equipment used was Tobii X2 series. Our implementation tool is Matlab on a PC with Intel i5 CPU, 3G memory and Windows XP system. The average testing speed is 1.76 s for every 40 images.

Fig. 8. Detection results of proposed method based on website collected images.

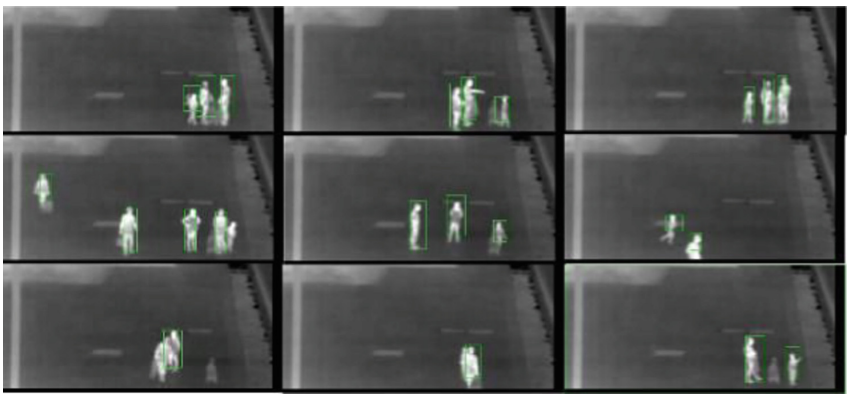

Fig. 9. Detection results of proposed method based on real-time images.

4 Conclusion

In this paper, we have proposed an efficient framework to solve the known object detection problem for infrared thermal images, which mainly relies on a saliency model and HOG feature optimization process. Evaluations on a number of real images captured from thermal cameras and comparisons with several state-of-the-art algorithms have demonstrated that the proposed method is effective and more accu-rate than other three other methods.

A few limitations of our method can be concluded as follows. First, there is some overlapping in detection results when the number of human objects is large. In our future work, we will examine how to adaptively classify overlapping objects to have better results. Second, the computational performance of the proposed algorithm is inadequate e for higher resolutions in real-time applications. Another task in the future is to improve the real-time performance using multi-resolution methods.

In order to extend the application fields of the proposed algorithm, we will put emphasis on extending the utilization of our algorithm in human body detection, animal detection, etc.

References

1. Ren, Z., Gao, S., Chia, L.T., Tsang, I.W.H.: Region-based saliency detection and its application in object recognition. IEEE Trans. Circ. Syst. Video Technol. **24**(5), 769–779 (2014)
2. Jung, C., Kim, C.: A unified spectral-domain approach for saliency detection and its application to automatic object segmentation. IEEE Trans. Image Process. **21**(3), 1272–1283 (2012)
3. Qi, S., Ma, J., Tao, C., Yang, C., Tian, J.: A robust directional saliency-based method for infrared small-target detection under various complex backgrounds. IEEE Geosci. Remote Sens. Lett. **10**(3), 495–499 (2013)
4. Qi, B., Wu, T., Dai, B., He, H.: Fast detection of small infrared objects in maritime scenes using local minimum patterns. In: 2011 18th IEEE International Conference on Image Processing (ICIP), pp. 3553–3556 (2011)
5. Han, J., Ma, Y., Zhou, B., Fan, F., Liang, K., Fang, Y.: A robust infrared small target detection algorithm based on human visual system. IEEE Geosci. Remote Sens. Lett. **11**(12), 2168–2172 (2014)
6. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
7. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems, pp. 545–552 (2007)
8. Hou, X., Zhang, L.: Saliency detection: a spectral residual approach. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2007, pp. 1–8 (2007)
9. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2009, pp. 1597–1604 (2009)

10. Bi, F., Bian, M., Liu, F., Gao, L.: A local descriptor based model with visual attention guidance for generic object detection. In: 2010 3rd International Congress on Image and Signal Processing (CISP), vol. 4, pp. 1599–1604 (2010)
11. Itti, L.: Models of bottom-up and top-down visual attention. Ph.D. thesis, California Institute of Technology (2000)
12. Frintrop, S., Backer, G., Rome, E.: Goal-directed search with a top-down modulated computational attention system. In: Kropatsch, W.G., Sablatnig, R., Hanbury, A. (eds.) DAGM 2005. LNCS, vol. 3663, pp. 117–124. Springer, Heidelberg (2005). https://doi.org/10.1007/11550518_15
13. Lee, Y.B., Lee, S.: Robust face detection based on knowledge-directed specification of bottom-up saliency. Etri J. **33**(4), 600–610 (2011)
14. Navalpakkam, V., Itti, L.: An integrated model of top-down and bottom-up attention for optimizing detection speed. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2049–2056 (2006)
15. Oliva, A., Torralba, A., Castelhano, M.S., Henderson, J.M.: Top-down control of visual attention in object detection. In: Proceedings of 2003 International Conference on Image Processing. ICIP 2003, vol. 1, pp. I–253 (2003)
16. Torralba, A.: Modeling global scene factors in attention. JOSA A **20**(7), 1407–1418 (2003)
17. Gao, D., Han, S., Vasconcelos, N.: Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **31**(6), 989–1005 (2009)
18. Torralba, A., Oliva, A., Castelhano, M.S., Henderson, J.M.: Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol. Rev. **113**(4), 766 (2006)
19. Borji, A., Sihite, D.N., Itti, L.: Probabilistic learning of task-specific visual attention. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 470–477 (2012)
20. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: Sun: a bayesian framework for saliency using natural statistics. J. Vis. **8**(7), 32 (2008)
21. Peters, R.J., Itti, L.: Beyond bottom-up: incorporating task-dependent influences into a computational model of spatial attention. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2007, pp. 1–8 (2007)
22. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 2106–2113 (2009)
23. Borji, A.: Boosting bottom-up and top-down visual features for saliency estimation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 438–445 (2012)
24. Yang, J., Yang, M.H.: Top-down visual saliency via joint CRF and dictionary learning. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2296–2303 (2012)
25. Le Meur, O., Le Callet, P., Barba, D., Thoreau, D.: A coherent computational approach to model bottom-up visual attention. IEEE Trans. Pattern Anal. Mach. Intell. **28**(5), 802–817 (2006)
26. Itti, L., Koch, C.: Computational modelling of visual attention. Nature Rev. Neurosci. **2**(3), 194 (2001)
27. Wai, W.K., Tsotsos, J.K.: Directing attention to onset and offset of image events for eye-head movement control. In: Proceedings of the 12th IAPR International Conference on Pattern Recognition, Conference A: Computer Vision and Image Processing, vol. 1, pp. 274–279 (1994)

28. Milanese, R., Bost, J.M., Pun, T.: A bottom-up attention system for active vision (1992)
29. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.M.: Global contrast based salient region detection. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 569–582 (2015)
30. Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Trans. Image Process. **19**(1), 185–198 (2010)
31. Bruce, N., Tsotsos, J.: Saliency based on information maximization. In: Advances in Neural Information Processing Systems, pp. 155–162 (2006)
32. Hou, X., Zhang, L.: Dynamic visual attention: searching for coding length increments. In: Advances in Neural Information Processing Systems, pp. 681–688 (2009)
33. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2005, vol. 1, pp. 886–893 (2005)
34. Alpher, A.: Frobnication. J. Foo **12**(1), 234–778 (2002)