Visual Attention Model with a Novel Learning Strategy and Its Application to Target Detection from SAR Images

Fei Gao[1], Xiangshang Xue[1], Jun Wang[1], Jinping Sun[1], Amir Hussain[2], Erfu Yang[3]

Abstract.—*The* selective visual attention mechanism in human visual system helps human to act efficiently when dealing with massive visual information. Over the last two decades, biologically inspired attention model has drawn lots of research attention and many models have been proposed. However, the top-down cues in human brain are still not fully understood, which makes top-down models not biologically plausible. This paper proposes an attention model containing both the bottom-up stage and top-down stage for the target detection from SAR (Synthetic Aperture Radar) images. The bottom-up stage is based on the biologically-inspired Itti model and is modified by taking fully into account the characteristic of SAR images. The top-down stage contains a novel learning strategy to make the full use of prior information. It is an extension of the bottom-up process and more biologically plausible. The experiments in this research aim to detect vehicles in different scenes to validate the proposed model by comparing with the well-known CFAR(constant false alarm rate) algorithm.

Keywords: Visual attention model, object detection, learning strategy, synthetic aperture radar (SAR) images.

1 Introduction

Human visual system possesses the astonishing ability to perceive the inputs from visual scenes. Whatever a visual scene is simple or complicated, humans can efficiently pick the most interesting part (weather it is free viewing or under the condition of a specific task), which is far beyond the development of the field of computer vision. Research has shown there are massive visual data (10^8-10^9 bits) entering the eyes every second [1].Without the help of any effective mechanism, the real-time processing seems impossible. Luckily, there exists a localization ability called visual attention or

This work was supported by the National Natural Science Foundation of China (61071139; 61471019; 61171122; 61501011), the Aeronautical Science Foundation of China (20142051022), the Pre-research Project(9140A07040515HK01009), the National Natural Science Foundation of China (NNSFC) under the RSE-NNSFC Joint Project (2012-2014) (61211130210) with Beihang University, and the RSE-NNSFC Joint Project (2012-2014) (61211130309) with Anhui University.

1. School of Electronic and Information Engineering, Beihang University, Beijing 100191, China.
2. Cognitive Signal-Image and Control Processing Research Laboratory, School of Natural Sciences, University of Stirling, Stirling FK9 4LA, UK.
3. Space Mechatronic Systems Technology Laboratory, Department of Design, Manufacture and Engineering Management, University of Strathclyde, Glasgow G1 1XJ, UK.

selective attention which enables human to act effectively and precisely in complex environment. When dealing with a complex visual scene, humans tend to turn their attention to one or few more salient objects or areas, while ignoring those which are not salient enough. Talking about whether an object or a region is salient, we need to consider it from a biological perception. In the retina, the photoreceptor cells are connected with ganglion cells which have unique receptive fields. The circular-shaped receptive field has a center area and a surrounding area. On the receptive field there exist on-center cells which are activated by stimuli, and off-center cells inhibited to stimuli. In terms of different stimuli, the receptive fields have different functions. Some are sensitive to intensity-contrast of the light, some are sensitive to color-contrast, while some are sensitive to motion, and et.al. Thus a flat region is not able to activate the receptive field. On the contrary, a region or an object with bright colors or anything else that are different from its surrounding can cause the sensitivity. As a result, attention or focus is led to the object which is assumed salient.

As the human visual system has so much potentials, researchers began to model it into a mathematically computational system. Almost all the attention models can be dated back to the feature integration theory (FIT) proposed by Treisman and Gelade in 1980 [2].This theory claims the visual input is first decomposed into a set of topographic feature maps and then feed in a bottom-up manner into a master map which depicts the local conspicuity of a visual scene. Koch and Ullman [3]then proposed a purely bottom-up model to combine the features and introduced the concept of saliency map [1]. But until 1998, the first fully implementation based on [3] was formally proposed by Itti and Koch [4]. The Itti model is believed to be biologically inspired because it imitates the early stages of human visual system [4].Since then, this field has drawn lots of attention and various models have emerged. However, most of the existing visual attention models follow a basic framework of Itti's to generate a topological saliency map. Indeed, attention is not merely caused by the visual scene's conspicuity, but other factors like knowledge, expectations, rewards and current goals also play important roles in the visual search, which is considered as a task-dependent top-down process [1]. Based on this phenomena, a lot of models combining both bottom-up cues and top-down cues have been proposed in recent years with specific application like car detection, face and pedestrian recognition and et al.

Top-down cues.

Neurobiological and psychophysical evidences have shown that the top-down mechanisms exist in the human brain for visual processing [5, 6]. Although the top-down attention is essential and inevitable, the computational models for top-down attention are fewer than the bottom-up ones, because how prior knowledge influence attention is still not fully understood. The existing top-down models can be classified into two categories. One is related to combine the low-level features in a top-down manner. The revised guided search structure (GS2 model) is believed to be the earliest computational model proposed by Wolfe in 1996 [6, 7]. Itti and Koch presented four strategies for combining the bottom-up cues in their original work: (1) Simple summation after scaling to a fixed dynamic range; (2) linear combination with weights learned; (3) nonlinear combination; (4) local nonlinear iterative competition between salient locations [8]. The second one is validated as the best one but needs a supervised additive training. By maximizing the signal-to-noise ratio of the target versus the background, Navalpakkam

and Itti [9] derived an optimal integration of bottom-up cues when detecting targets. Frintrop [10] proposed the VOCUS model with a top-down extension which includes a learning mode and a search mode. Armmanfard et al. [11] proposed a feature fusion technique which applied a weighted feature summation block whose weights are optimized by the genetic algorithm, instead of both across scale combination and normalization and linear combination block. In [12], Han et al. proposed a saliency map generated from the weighted features where the rough sets are used to assign the weights for every feature. But, the problem in the aforementioned work is that they don't make the full use of prior information, which matters a lot in the human brain. The other one involves the representation of the top-down cues using tools like conditional random field (CRF), fuzzy theory and et al. Tsotsos et al. proposed a hierarchical system and a new winner-takes-all (WTA) updating rule to match the current related knowledge [13]. In [14] there's a top-down model considering the visual memory which adopts a fuzzy adaptive resonance theory neural network with the learning function. Borji et al. [15] used evolutionary algorithms to search some parameters inside the basic saliency model as the top-down priors. Ban et al. [16] proposed a growing fuzzy topology adaptive resonance theory (GFTART) with two roles: one is to form the bottom-up features of arbitrary objects, and the other is to generate the top-down bias. Yang et al. [17] proposed a top-down saliency model that jointly learns a conditional random field (CRF) and a visual dictionary. The model has a three-layered structure from the bottom to the top: CRF, sparse coding, and a visual dictionary. Obviously, those methods are more mathematical rather than biologically plausible.

Application of visual attention.

It is said that the visual information is interpreted in a need-manner in the brain to serve the task demands [18]. A lot of attention models can be put into one category of computer vision. Usually, these models are applied to detect or recognize targets like faces, cars, pedestrians and so on [1, 19-23]in the context of real-life visual scenes. But we want to address the application in the area of remote sensing in this work. As the remote sensing images, for example, SAR images are quite different from the optical images due to the completely different mechanisms of imaging, it is not proper to directly apply the attention model to the SAR images. As a result, there is few research on understanding the remote sensing images using attention model.

In our paper, we propose a visual attention model specifically for the application of the vehicle targets detection from SAR images by integrating a bottom-up stage and a top-down stage. The bottom-up stage follows the procedure of the Itti model but with some simplification and modification in some aspects. The top-down stage also generates a saliency map similar to that in the bottom-up stage. During the top-down stage, two weighting parameters are learned from the training set to instruct how the feature maps are combined. A training set is used for two reasons: one is to get the best weighting coefficients for two conspicuity maps; the other one is to get targets' average length or size used as thresholds. The global saliency maps is then generated through the linear combination of bottom-up and top-down saliency maps. Finally, the detection result is acquired from the global saliency through binarization and thresholding processes.

2 Proposed Method

The proposed method has both the bottom-up and top-down computational process to mimic the human visual system. Our proposed method is based on the saliency map, which means that the computational process is restricted to generate saliency maps including a bottom-up saliency map, a top-down saliency map and a combined one. The framework of the proposed method is depicted in Fig. 1. The input is first processed by the bottom-up stage to generate a BU saliency map. Then after learning the optimal weights in the top-down stage, a TD saliency map is generated. At the decision stage, two saliency maps are combined into a single global map, and prior information are used as thresholds. The bottom-up saliency map is a modified version of Itti's and specifically tuned for SAR images. As for the top-down saliency map, it is generated from the intermediate e bottom-up stage by applying a learning strategy. Then, the global map is computed from the two former saliency maps.

2.1 Bottom-up attention

The bottom-up process in this paper is based on the well-known Itti model with some modifications in consideration of the characteristics of SAR images. For SAR images, there's no color information thus color channel in the Itti model can be ignored. Intensity and orientation channels are consistent with those in the Itti model, but possess some specific modifications. As can be seen from the flowchart, the bottom-up saliency consists of a feature extraction stage, a weighting operation and a saliency map generating process. The detailed bottom-up saliency is described down below.

Feature extraction.

Intensity channel.

For an intensity SAR image I , a five-scaled Gaussian pyramid is first created by applying a Gaussian filter and sub-sampling. The Gaussian image pyramid with five scales s_0 - s_4 is further transformed into the feature maps by applying a center-surround operation. Unlike Itti's 9 scales, the image pyramid in our method has only 5 scales with almost the same function.

In Itti' model, the center-surround operation is implemented as the difference between the fine and coarse scales: the center is represented as scales $c \in \{2,3,4\}$, while the corresponding surround is at the scales $s = c + \delta$, where $\delta \in \{3,4\}$. The across-scale difference is obtained by interpolating the coarse scale to the finer one and then point by point subtraction. However, we find that the operation has another alternative. The detailed implantations are presented down below.

Step 1: Create five-scaled Gaussian pyramid I_σ , where $\sigma \in [0..4]$ is the scale. The Gaussian low-pass filter is:

$$G(x, y, o) = \frac{1}{2\pi o^2} \exp(-\frac{x^2 + y^2}{2o^2}) \tag{1}$$

where x, y denote the coordinate of an arbitrary pixel and o=3 .

Step 2: Represent the surround of an arbitrary pixel.

For each pixel $I^{"}_{\sigma=c}(x,y)$ in the center, its surround is represented as:

$$I^{"}_{\sigma=c|\delta}(x,y)\frac{\sum\left(I_{s_i}(x-\delta,y-\delta)+...I_{s_i}(x+\delta,y+\delta)\right)}{\delta^2} \tag{2}$$

where $c\in\{2,3,4\}$ is the center, $\delta\in\{4,8\}$ is the length of the side of neighborhood.

Step 3: Apply the center-surround operation and yield 6 feature maps $I^{"}_{c,\delta}$.

The feature maps are defined as:

$$I^{"}_{c,\delta}=I^{"}_{\sigma=c}-I^{"}_{\sigma=c|\delta} \tag{3}$$

Through these steps, 6 intensity maps are acquired and wait to be further processed.

Orientation channel.

In SAR images, the targets for example vehicles are usually small, therefore the orientation information seems to matter only a little but still be indispensable. We accept the operation in [10] where only 5 scales are needed. The absence of the center-surround operation in orientation extraction is because the oriented center-surround difference is already determined implicitly by the Gabor filter [10], and it can also prevent the images from getting blurred due to the center-surround mechanism. The oriented Gabor pyramid $O^{"}_{\sigma,\theta}(x,y)$ is acquired by:

$$H(x,y,\sigma,\theta)=\frac{1}{\sigma^2}\exp(-\pi\frac{x^2+y^2}{\sigma^2})\{\exp[i2\pi(x\cos\theta+y\sin\theta)]-\exp(-\frac{\pi^2}{2})\} \tag{4}$$

where $\sigma\in\{2,3,4\}$ is the scale and $\theta\in\{0^\circ,45^\circ,90^\circ,135^\circ\}$ is the angle.

Eventually, 12 raw orientation feature maps are yielded.

Weighting and normalization.

After acquiring all the 6 intensity and 12 orientation feature maps, the feature maps should be normalized to the same scale and fused to form two conspicuity maps. In human visual systems, the fusion mechanism is quite complicated which is not clearly figured out even in a bottom up way let alone the high-level neural activity. Different features contribute differently to perceptual saliency [24] and the relevant feature fusion or weighting approaches are influenced by tasks, goals expectations et al. If the feature maps are combined in a purely straightforward way, they contribute equally [10]. To prevent this effect, we have to determine the most important maps and raise their influence. Therefore, a brand new weighting function is designed in this paper and somehow tested faithful. The weighting function is defined as:

$$W(X)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(M-\bar{m})^2}{2\sigma^2}}\cdot\frac{\bar{m}}{\bar{r}}\cdot X \tag{5}$$

where M is the global maxima within a feature map, $\bar{m}$ is the expectation of local maxima, σ is the standard deviation of the feature map, $\bar{r}$ is the expectation of the rest of the feature map when taking out the local maxima.

$$I'=\bigoplus_{c,\delta}W\left(I^{"}_{c,\delta}\right) \tag{6}$$

$$O^{'} = \bigoplus_{\sigma,\theta} W\left(O^{''}_{\sigma,\theta}\right) \tag{7}$$

where $\oplus$ indicates the point by point addition.

Bottom up saliency map.

After acquiring the conspicuity map for each channel, the bottom-up saliency map is then computed by fusing the conspicuity maps together.

$$S_{BU} = W\left(I^{'}\right) + W\left(O^{'}\right) \tag{8}$$

Actually, saliency map is a topographic map which indicates the saliency or conspicuity of an area within the map.

2.2 Top-down attention

In SAR image with vehicle targets, even for human, it is very difficult to determine whether an object is a vehicle or not. Due to the mechanism of SAR imaging, the vehicles in SAR images are completely different with those in optical images, let alone the low-resolution of SAR images compared to optical images. But if observers are offered to watch some targets in advance, it then becomes very easy for observers to recognize it in a SAR scene. Apart from the vehicle's low-level characteristic like intensity and orientation which raise the attention even before we know it is a target, the information like size, outline, texture play an important role in human's understanding process. So the use of this prior information provides a promising way to help detect vehicle targets.

In our proposed method, the top-down process is also based on the saliency map, but needs a learning process first.

Learning strategy.

In the previous bottom-up saliency, two conspicuity maps are weighted and fused to generate the saliency map. But how can we know the weights computed from (5) are the perfect weights to generate the most accurate saliency map, what if there're other weights that outperform the former ones? The learning process is designed to make sure that the perfect weights are selected. Therefore, we need the following learning strategy.

The general process of the learning strategy is depicted in the middle of Fig. 1. A set of image slices with targets therein, is needed. For slice X_i , (i is the number of slices), two corresponding conspicuity maps I_i and O_i are computed with the aforementioned bottom-up stage. Instead of using the weighting function $W(\square)$ to form the saliency map, we obtain the most accurate weights by benchmarking the saliency maps generated from different weights. The F-measure is adopted to benchmark the most salient one. Below is the detailed steps.

Step 1: For each slice, compute the bottom-up weights of conspicuity maps: $w_{I\_i} = W\left(I_i^{'}\right), w_{O\_i} = W\left(O_i^{'}\right)$;

Step 2: Determine the intervals of top-down weights $\left[w_{I\_\min}, w_{I\_\max}\right]$ and $\left[w_{O\_\min}, w_{O\_\max}\right]$. The interval is defined as:

$$w_{I\_\min} = \min(w_{I\_i}) - \sigma_I \tag{9}$$

$$w_{I\_\max} = \max(w_{I\_i}) + \sigma_I \tag{10}$$

$$w_{O\_\min} = \min(w_{O\_i}) - \sigma_O \tag{11}$$

$$w_{O\_\max} = \max(w_{O\_i}) + \sigma_O \tag{12}$$

where σ_I and σ_O are the standard deviation of w_{I_i} and w_{O_i}, respectively.

Step 3: Select 10 weights from every interval at a regular distance and compute 100 saliency maps for each target slice.

Step 4: Benchmark the 100 saliency maps and find the best one with its corresponding weights w'_{I_i} and w'_{O_i}. Here we use the Precision (P), Recall (R) and F-measure (F) as the benchmarks, defined as follows:

$$P = \sum(S \otimes A) / \sum(S) \tag{13}$$

$$\mathrm{R} = \sum(S \otimes A) / \sum(A) \tag{14}$$

$$F = \frac{\left(\alpha^2 + 1\right) P * R}{\alpha^2 \left(P + R\right)} \tag{15}$$

where S is the saliency map, A is the segmentation map. Operator $\otimes$ is the point by point multiplication.

Step 5: The final weights are the means of the two set of weights.

$$w_I = mean(w'_{I\_i}) \tag{16}$$

$$w_O = mean(w'_{O\_i}) \tag{17}$$

Top-down saliency map.

After learning the weights, the top-down saliency map is generated from the two bottom-up conspicuity maps and the top-down weights.

$$S_{TD} = w_I \cdot I' + w_O \cdot O' \tag{18}$$

It should be noticed that, here we only compute the weights for the conspicuity maps. Actually, this approach is also suited to compute the weight for each raw feature map, but takes a lot of computing resource apparently.

2.3 Global saliency map

The global saliency map is then generated from the combination of the bottom-up and top-down maps. Parameter t determines how much the top-down process contribute to the global saliency map.

$$S = S_{BU} + t * S_{TD} \tag{19}$$

2.4 Decision

Apart from the top-down weights learned from the target slices, the size and length of a specific type of vehicle also play important roles. In our model, the size represented by the number of pixels a vehicle possesses and the length used as two thresholds for the final decision stage. But first, we need to transform the saliency map to a binary map. Here we use the Ohtsu [25] method to create the threshold.

$$S_{bw}(x,y)=\begin{cases}1 & S(x,y)>threshold\\ 0 & S(x,y)<threshold\end{cases} \tag{20}$$

For an arbitrary region in the binary map, it is determined whether it's a target or not by the size and length of the target.

$$R_i=\begin{cases}1 & size\in[a,b]\,\&\,length\in[c,d]\\ 0 & size\notin[a,b]\,|\,length\notin[c,d]\end{cases} \tag{21}$$

where R_i is the suspicious areas, 1 for target, 0 for not. The confidence intervals [a, b], [c, d] are computed from the segmentation maps in the learning stage.

3 Experiment

In this section, the experiments on both the proposed method and the constant false alarm rate (CFAR) which is well an acknowledged method for SAR image detection in the literature are carried out. The result of the proposed method with only the bottom-up process is also presented to demonstrate the effectiveness of our top-down strategy.

We picked up a heavy cluttered image from the spotlight SAR images of ground vehicles in the moving and stationary target acquisition and recognition (MSTAR) database with the size of 1478×1784 pixels. The images is added with 20 vehicles targets. One image has little distracters, whereas the other has much more. The image with the added targets is depicted in Fig. 2.

There are plenty of benchmarks to quantitatively evaluate the effectiveness of a detection algorithm, among them the probability of detection (P_d) and probability of false alarm (P_f) are often used, thereby they are included in the experiments. P_d and P_f are defined as:

$$P_d=\frac{\text{Number of detected targets}}{\text{Total number of targets}} \tag{22}$$

$$P_f=\frac{\text{Number of false alarm}}{\text{Number of detected units}} \tag{23}$$

Besides, the Precision, Recall and F-measure are fundamental measures in statistics, therefore they are also included in our experiments. In our cases, R has the same definition as P_d. P and F-measure are defined as:

$$P=\frac{\text{Number of detected targets}}{\text{Number of detected units}} \tag{24}$$

$$F=\frac{(\alpha^2+1)P*R}{\alpha^2(P+R)} \tag{25}$$

Because of the learning strategy in the proposed method, a training set is needed. We selected 100 vehicle target slices from the MSTAR database as the training set. The training set is also used to determine the confidence interval mentioned in (21), calculated as [35.15, 46.40] and [420.30, 484.34] using (26). The interval [a, b] and [c, d] are defined as:

$$\begin{aligned} a &= \overline{\mu_s} - \sigma_s,\ b = \overline{\mu_s} + \sigma_s \\ c &= \overline{\mu_l} - \sigma_l,\ d = \overline{\mu_l} + \sigma_l \end{aligned} \tag{26}$$

where $\overline{\mu_s}$ and $\overline{\mu_l}$ are the expectations of the size and length of each training target, σ_s and σ_l are the relevant standard deviations.

It is noticed from Fig. 2 that the vehicles in this image are distinct from the surrounding and thus possess strong conspicuity. The saliency maps are shown in Fig. 3 and the detection results are shown in Fig. 4. The green rectangles mark the detected targets, while the red and white ones mark the false alarms and the undetected targets respectively. There are 2 targets undetected and 2 false alarms generated for the proposed method, whereas the CFAR has 4 targets undetected and generated 5 false alarms. As for the bottom-up way, the result seems unacceptable with 3 targets undetected and generated astonishing 11 false alarms.

Table Ⅱ shows the quantitative evaluation for the two methods. The proposed method outperformed the CFAR by 10% detection rate higher and 13.81% false alarm rate lower. And, the results demonstrated the effectiveness of the top-down strategy.

4 Conclusion

In this paper a visual attention based approach has been presented for the underlying application in detecting targets from remote sensing images. The proposed method contains a bottom-up stage, which is a modified version of the Itti model, and specifically tuned for SAR images, as well as a top-down stage. The top-down process contains a novel learning strategy, but it is a once-for-all job, because once the weights are learned it can be adapted to most of the scenes. The novelty of the method lies in the following three aspects. First, the brand new weighting function makes multi-target popping out possible. Second, the learning strategy selects the optimal weights from the training set. Last but not least, target's prior information like size and length are used as thresholds in the decision stage. Experiment results have demonstrated that the proposed method possesses greater ability in detecting vehicle targets in comparison with the CFAR. In addition, the results from the only bottom-up way were presented, which were far inferior to that of the complete method, which further validated the effectiveness of the top-down strategy.

Though the proposed method was applied to the vehicle detection from SAR images, it can be adapted to other areas. For example, it can be potentially applied to detect other targets or applied to the optical images on fixation prediction. Our future work will explore these potentials by applying our method to other research fields.

5 REFERENCES

1. A. Borji and L. Itti, "State-of-the-Art in Visual Attention Modeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, pp. 185-207, 2013.
2. A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychol*, vol. 12, pp. 97-136, 1980.
3. C. Koch and S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," *Human Neurobiology*, vol. 4, pp. 219-27, 1985.
4. L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1254-1259, 1998.
5. J. B. Hopfinger, M. H. Buonocore, and G. R. Mangun, "Hopfinger JB, Buonocore MH, Mangun GR. The neural mechanisms of top-down attentional control. Nat Neurosci 3: 284-291," *Nature Neuroscience*, vol. 3, pp. 284-91, 2000.
6. L. Zhang and W. Lin, "Computational Models for Top-down Visual Attention," in *Selective Visual Attention:Computational Models and Applications*, ed: Wiley-IEEE Press, 2013, pp. 167-205.
7. J. M. Wolfe, "Guided Search 2.0 A revised model of visual search," *Psychonomic Bulletin & Review*, vol. 1, pp. 202-238, 1994.
8. L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *Redele Revista Electrónica De Didáctica Ele*, vol. 10, pp. 161-169, 2001.
9. V. Navalpakkam and L. Itti, "An Integrated Model of Top-Down and Bottom-Up Attention for Optimizing Detection Speed," in Cvpr, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, pp. 2049-2056.
10. S. Frintrop, VOCUS: A visual attention system for object detection and goal-directed search vol. 3899: Springer, 2006.
11. Z. Armanfard, H. Bahmani, and A. M. Nasrabadi, "A novel feature fusion technique in Saliency-Based Visual Attention," in *Advances in Computational Tools for Engineering Applications, 2009. ACTEA '09. International Conference on*, 2009, pp. 230-233.
12. B. Han, L. Tcheang, V. Walsh, and X. Gao, "A Novel Feature Combination Methods for Saliency-Based Visual Attention," in *2009 Fifth International Conference on Natural Computation*, 2009, pp. 18-22.
13. J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial Intelligence*, vol. 78, pp. 507-545, 1995.
14. B. Kim, S. W. Ban, and M. Lee, "Growing fuzzy topology adaptive resonance theory models with a push–pull learning algorithm," *Neurocomputing*, vol. 74, pp. 646-655, 2011.
15. A. Borji, M. N. Ahmadabadi, and B. N. Araabi, "Cost-sensitive learning of top-down modulation for attentional control," *Machine Vision & Applications*, vol. 22, pp. 61-76, 2011.
16. S. W. Ban, B. Kim, and M. Lee, "Top-down visual selective attention model combined with bottom-up saliency map for incremental object perception," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, 2010, pp. 1-8.
17. J. Yang and M. H. Yang, "Top-Down Visual Saliency via Joint CRF and Dictionary Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, pp. 1-1, 2016.
18. J. Triesch, D. H. Ballard, M. M. Hayhoe, and B. T. Sullivan, "What you see is what you need," *IEEE Comput Soc*, vol. 3, p. 102, 2003.
19. J. Najemnik and W. S. Geisler, "Optimal eye movement strategies in visual search," *American Journal of Ophthalmology*, vol. 434, pp. 387-91, 2005.

20. P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *Pattern Analysis & Machine Intelligence IEEE Transactions on*, vol. 32, pp. 1627-45, 2010.
21. P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, vol. 57, pp. 137-154, 2004.
22. M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, pp. 1019-25, 2010.
23. L. Itti and C. Koch, "Target detection using saliency-based attention," 1999.
24. L. Itti and C. Koch, "Computational modelling of visual attention," *Nature reviews neuroscience*, vol. 2, pp. 194-203, 2001.
25. N. Ohtsu, "A Threshold Selection Method from Gray-Level Histograms," *Systems Man & Cybernetics IEEE Transactions on*, vol. 9, pp. 62-66, 1979.

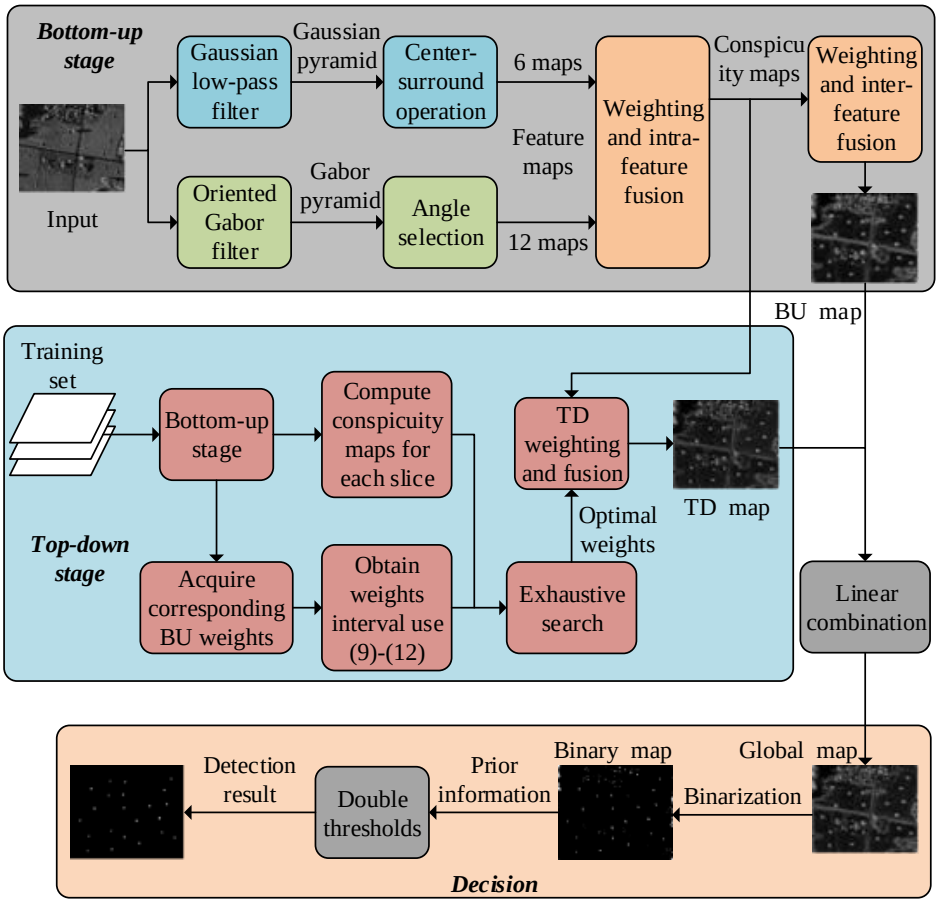


Fig. 1. Framework of the proposed method.

Table 1. QUANTITATIVE MEASURES OBTAINED BY CFAR AND THE PROPOSED METHOD FOR SCENE 2.

Methods	P_d	P_f	P	R	$F_{\alpha=1}$

CFAR	80.00%	23.81%	76.19%	80.00%	78.05%
Bottom-up way	85.00%	39.29%	60.71%	85.00%	70.83%
Proposed method	90.00%	10.00%	90.00%	90.00%	90.00%

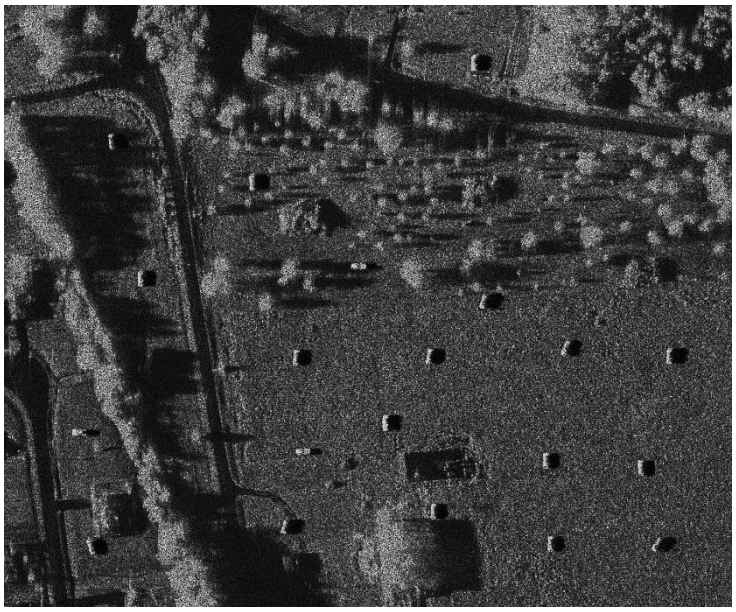
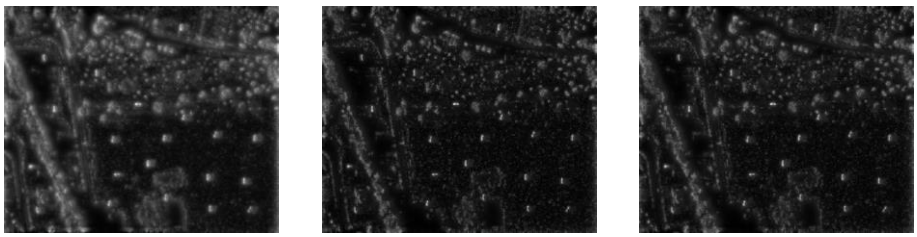
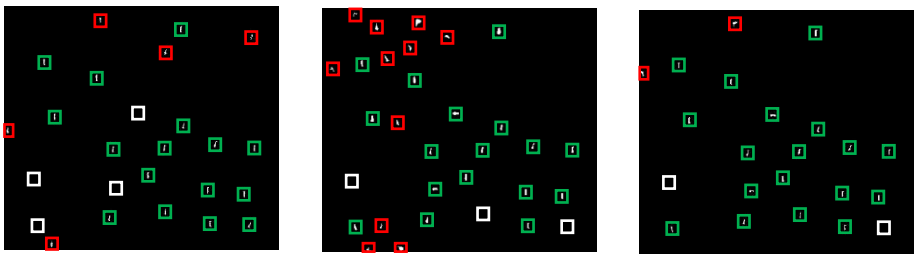


Fig. 2. Scene 2 with 20 vehicle targets inside.



(a) (b) (c)

Fig. 3. Saliency maps of scene 2. (a) BU; (b) TD; (c) Global



(a) (b) (c)

Fig. 4. Detection results of the CFAR and the proposed method of scene 2. (a) Result of the CFAR; (b) Result of the only bottom-up way; (c) Result of the proposed method.