A Novel Target-Objected Visual Saliency Detection Model in Optical Satellite Images

Xiaoguang Cui, Yanqing Wang, and Yuan Tian
Institute of Automation, Chinese Academy of Sciences, Beijing, China
Email: {xiaoguang.cui, yanqing,wang, yuan.tian}@ia.ac.cn

***Abstract*—A target-oriented visual saliency detection model for optical satellite images is proposed in this paper. This model simulates the structure of the human vision system and provides a feasible way to integrate top-down and bottom-up mechanism in visual saliency detection. Firstly, low-level visual features are extracted to generate a low-level visual saliency map. After that, an attention shift and selection process is conducted on the low-level saliency map to find the current attention region. Lastly, the original version of hierarchical temporal memory (HTM) model is optimized to calculate the target probability of the attention region. The probability is then fed back to the low-level saliency map in order to obtain the final target-oriented high-level saliency map. The experiment for detecting harbor targets was performed on the real optical satellite images. Experimental results demonstrate that, compared with the purely bottom-up saliency model and the VOCUS top-down saliency model, our model significantly improves the detection accuracy.**

***Index Terms*—Visual Salience; Target-Oriented; Hierarchical Temporal Memory**

I. Introduction

With the development of remote sensing technology, optical satellite images have been widely used for target detection, such as harbors and airports. In recent years, high spatial resolution satellite images provide more details for shape, texture and context [1]. However, data explosion for high resolution remote sensing images, brings more difficulties and challenges on fast image processing. Visual saliency detection aims at quickly identifying the most significant region of interest in images by means of imitating the mechanism of the human vision system (HVS). In this way, significant regions of interest can be processed with priority by the limited computing resource, thus substantially improving the efficiency of image processing [2]-[3].

There are two models for HVS information processing, namely, bottom-up data driven model and top-down task driven model. Bottom-up model often acts as the unconscious visual processing in early vision and is mainly driven by low-level cues such as color, intensity and oriented filter responses. Currently, many bottom-up saliency models have been proposed for computing bottom-up saliency maps, by which we can predict human fixations effectively. Several bottom-up models are based on the well known biologist saliency model by Itti et al [4]. In this model, an image is decomposed into low-level feature maps across several spatial scales, and then a master saliency map is formed by linearly or non-linearly normalizing and combining these maps. Different from the biological saliency models, some bottom-up models are based on mathematical methods. For instance, Graph-based Visual Saliency (GBVS) [5] formed a bottom-up saliency map based on graph computations; Hou and Zhang [6] proposed a Spectral Residual Model (SRM) by extracting the spectral residual of an image in spectral domain; Pulsed Cosine Transform (PCT) based model [7] extended the pulsed principal component analysis to a pulsed cosine transform to generate spatial and motional saliency.

Although the bottom-up saliency models are shown to be effective for highlighting the informative regions of images, they are not reliable in target-oriented computer vision tasks. When apply bottom-up saliency models in optical satellite images, due to the lack of top-down prior knowledge and highly cluttered backgrounds, these models usually respond to numerous unrelated low-level visual stimuli and miss the objects of interest. In contrast, top-down saliency models learn from training samples to generate probability maps for localizing the objects of interest, and thus produce more meaningful results than bottom-up saliency models. A well-known top-down visual saliency model is Visual Object detection with a CompUtational attention system (VOCUS) [8], which takes the rate between an object and its background as the weight of feature maps. The performance of VOCUS is influenced by object background. Although it performs well in nature images, it does not work reliably in the complicated optical satellite images. Recently, several top-down methods have been proposed based on learning mappings from image features to eye fixations using machine learning techniques. Zhao and Koch [9]-[10] combined saliency channels by optimal weights learned from eye-tracking dataset. Peters and Itti [11], Kienzle et al. [12] and Judd et.al. [13] learned saliency using scene gist, image patches, and a vector of features at each pixel, respectively.

It is established that top-down models achieve higher accuracy than bottom-up models. However, bottom-up models often take much lower computational complexity due to only taking into account of low-level visual stimuli. In this case, an integrated method of combining

doi:10.4304/jmm.9.2.302-309

bottom-up and top-down driven mechanisms is needed to get benefits from both types of mechanisms.

How to effectively integrate bottom-up and top down driven mechanisms is still an unsolved problem for the visual saliency detection. According to the mechanism of HVS, this paper proposes a target-oriented visual saliency detection model, which is based on the integration of both the two driven mechanisms. The proposed model consists of three parts, namely pre-attention phase module, attention phase module and post-attention module. Firstly, a low-level saliency map is quickly generated by the pre-attention phase module to highlight the regions with low-level visual stimuli. Then the attention phase conducts an attention shift and selection process in the low-level saliency map to find the current attention region. After obtaining the attention region, a target probability of the region evaluated by the post-attention module is fed back to the low-level saliency map to generate a high-level saliency map where the suspected target regions are emphasized meanwhile the background interference regions are suppressed. The main contributions of this paper are:

A new method is presented for combining top-down and bottom-up mechanisms, i.e. revising the low-level saliency map with target probability evaluation so that the attention regions containing suspected targets are enhanced, meanwhile inhibiting the non-target regions.

An effective method for focus shift and attention region selection is proposed to focus on the suspected target regions rapidly and accurately.

The original HTM model is improved in several respects including the input layer, the spatial module and the temporal module, leading to a robust estimation of the target probability.

This paper is structured as follows: Section II describes the framework of the proposed model. The details of the three parts i.e. pre-attention phase module, attention phase module and post-attention module are presented in Section III, IV and V, respectively. Experimental results are shown in Section VI. Finally, we give the concluding remarks in Section VII.

II. Framework of the Proposed Model

A new model is presented to simulate HVS attention mechanism, and composed of three functional modules, namely, pre-attention phase module, attention phase module and post-attention phase module, as shown in Fig. 1. The pre-attention phase is a bottom-up data driven process. It is employed to extract the lower features to form the low-level saliency map. According to principles of *winner takes all*, *adjacent proximity* and *inhibition of return* [4], the attention phase module carries out the focus of attention shift on the low-level saliency map and proposes a self-adaptive region growing method to rationally select the attentions regions. The post-attention phase is a top-down data driven process, and its major function is to apply the HTM model [14]-[15] to evaluate the target probability of the selected attention regions. The probability is then multiplied with the corresponding attention region on the low-level saliency map, thus a high-level saliency map which is more meaningful to locate objects of interest is generated.

III. Pre-Attention Phase

In this phase, we first extract several low-level visual features to give rise to feature maps, and then we compute saliency map for each feature map using the PCT-based attention model. Finally, saliency maps are integrated to generate the low-level saliency map. The block diagram of the pre-attention phase is shown in Fig. 2.

A. Feature Extraction

If a region in the image is salient, it should contain at least one distinctive feature different from its neighborhood. Therefore, visual features of the image should be extracted first. For this, we extract three traditional low-level visual features, i.e. color, intensity and orientation.

1) Color and intensity: HSI color space describes a color from the aspect of hue, saturation and intensity, more consistent with human visual features than RGB color space. Hence, we transfer the original image from RGB to HIS in order to obtain the color feature map H, S and the intensity feature map I :

$$\begin{aligned} H &= \frac{1}{360}\left[\begin{array}{l} 90 - \arctan\left(\dfrac{2R-G-B}{\sqrt{3}(G-B)}\right) + \\ \{0, G \ge B; 180, G < B\} \end{array} \right] \\ S &= 1 - \left[\frac{\min(R,G,B)}{I}\right] \\ I &= \frac{R+G+B}{3} \end{aligned} \tag{1}$$

2) Orientation: Artificial targets in optical satellite images generally possess obvious geometrical characteristics. Therefore, orientation feature is crucial to identify the artificial targets. Here we adopt Gabor filters $(\theta_{\mathrm{k}} = 0^o, 45^o, 90^o, 135^o)$ to extract the orientation feature. The kernel function of a 2-D Gabor wavelet is defined as:

$$\begin{aligned} \Psi_{\theta_k}(z) &= \frac{\left\|v_{\theta_k}\right\|^2}{\sigma^2} \cdot e^{\frac{\left\|v_{\theta_k}\right\|^2 \|z\|^2}{2\sigma^2}} \cdot \left(e^{iv_{\theta_k}} - e^{-\frac{\sigma^2}{2}}\right) \\ v_{\theta_k} &= \left(\cos\theta_k, \sin\theta_k\right) \end{aligned} \tag{2}$$

where $\mathrm{z} = (x, y)$ denotes the pixel position, and the parameter σ determines the ration between the width of Gaussian window and the length of wave vector. We set $\sigma = 7\pi/4$ in the experiment. Four orientation feature maps can be obtained by convoluting the intensity feature map I with Ψ_{θ_k} :

$$O_k(z) = I(z) \otimes \Psi_{\theta_k}(z) \tag{3}$$

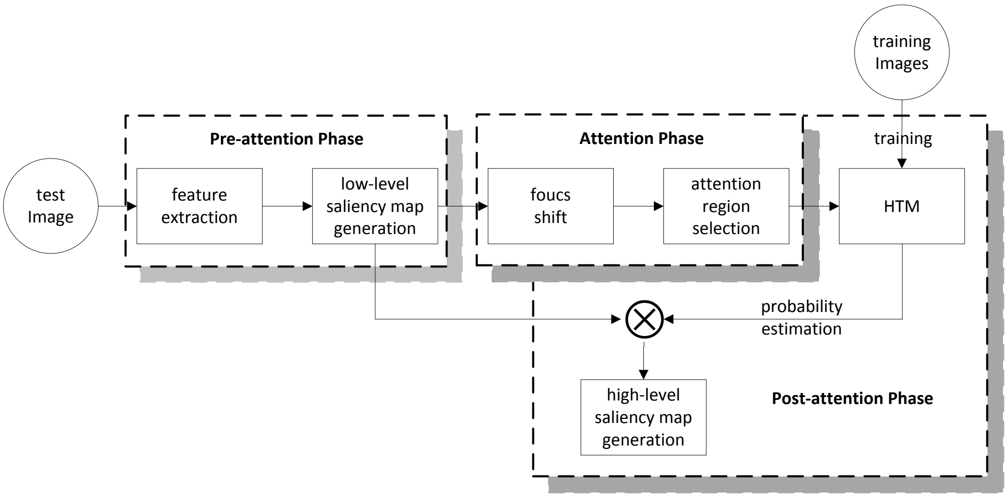


Figure 1. The framework of the proposed model

B. The Generation of the Low-Level Saliency Map

Recently, many effective approaches for saliency detection have been proposed. Here we employed PCT-based attention model because of its good performance in saliency detection and fast speed in computation [7]. According to the PCT model, the feature saliency map S_F of a given feature map F can be calculated as:

$$\begin{aligned} P &= sign(C(F)) \\ A &= abs(C^{-1}(P)) \\ S_F &= G * A^2 \end{aligned} \tag{4}$$

where $C(\cdot)$ is the 2-D discrete cosine transform and $C^{-1}(\cdot)$ is its inverse transform. G is a 2-D low-pass filter. We apply linear weighted method to integrate the feature maps. Due to the lack of priori information, the weight of each feature map is set to $1/N$ (N is the number of feature maps, here $N = 7$) and the low-level saliency map S_{low} can be obtained as:

$$S_{low} = \frac{1}{N}\left[S_H + S_S + S_I + \sum_{k=1,2,3,4} S_{O_k} \right] \tag{5}$$

IV. Attention Phase

Attention phase provides a set of attention regions so that the significant area of interest can be processed with priority in the post-attention phase. This phase includes two parts, namely, the focus of attention shift and the attention region selection.

A. Focus of Attention Shift

According to principles of *winner takes all*, *adjacent proximity* and *inhibition of return*, an un-attended pixel, of the highest salience and closest to the last focus of attention on the low-level saliency map, is chosen as the next focus of attention, which is based on the following formula:

$$\begin{aligned} \left(px^{t+1}, py^{t+1}\right) &= \arg\max_{x,y} \begin{pmatrix} S_{low}(x,y) \times D(x,y) \times \\ B(x,y) \end{pmatrix} \\ D(x,y) &= \left((x - px^t)^2 + (y - py^t)^2\right)^{-1/2} \\ B(x,y) &= \begin{cases} 0 & (x,y) \, has\, been\ \ focused \\ 1 & otherwise \end{cases} \end{aligned} \tag{6}$$

where （px^t, py^t） is the location of the current focus of attention, （px^{t+1}, py^{t+1}） is the location of the next focus of attention, $D(\cdot)$ serves as the *adjacent proximity*, i.e. areas close to the current focus of attention will be noticed with priority, $B(\cdot)$ serves as the *inhibition of return*, i.e. the noticed areas will not participate in the focus shift.

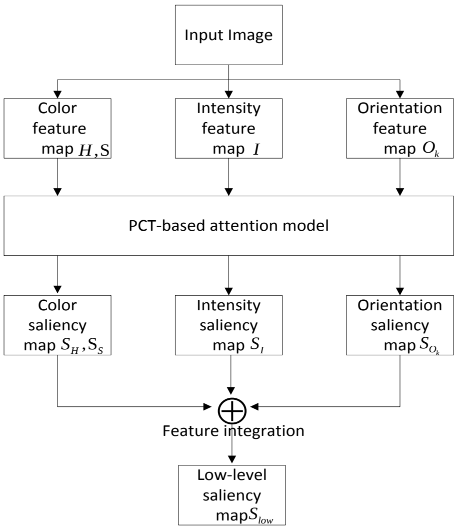

Input Image
Color feature map H,S
Intensity feature map I
Orientation feature map O_k
PCT-based attention model
Color saliency map S_H,S_S
Intensity saliency map S_I
Orientation saliency map S_{O_k}
Feature integration
Low-level saliency mapS_{low}

Figure 2. Block diagram of the pre-attention phase

B. Attention Region Selection

Different from the attention region selection with fixed size in Itti's model [4], the attention region in this research is identified by a self-adaptive region growing method: taking the focus of attention as seed point, the region growing is conducted by computing the saliency difference between the current growing area and its surrounding areas according to a given step-size sequence. Once the difference tends to be decreasing, the growth will be terminated. Finally, the minimum area-enclosing rectangle of the growing area is deemed to the attention region. Here we define R_i as the growing area obtained in each growth, n_i as the number of pixels in R_i, A_i as the saliency difference between R_i and its surrounding area. Given a step-size sequence $N_i(i \in [0,T])$, where T denotes the maximum times of growing, the algorithm for the self-adaptive region growing is as Algorithm 1.

Algorithm 1 Self-adaption region growing

```
Input: $N_i(i \in [0,T])$, $R_0 = \{f\}$, where $f$ is the present focus of attention; $n_0 = 1$; $i = 1$.
Iteration:
while not reach the maximum growing time do
    Initialize $R_i$ and $n_i$: $n_i = n_{i-1}$; $R_i = R_{i-1}$.
    while do
        produce a new growing point $p$ : $p = \arg\max_{p_j} S(p_j)$, where $p_j \in A$, $A$ is the adjacent pixel set of $R_i$, $S(p_j)$ is the saliency of $p_j$.
        update $R_i$ and : $R_i = \{R_i, p\}$; $n_i = n_{i-1} + 1$.
    end while
    Calulate :
    $A_{i-1} = \sum_{p_j \in R_i} S(p_j) \big/ N_i - \sum_{p_j \in R_{i-1}} S(p_j) \big/ N_{i-1}$ when $A_{i-1}$ tends to decrease, the growth is terminated:
    if then
        the growth is terminated.
    else
        $i = i + 1$; growth continues.
    end if
end while
Output:
the minimum area-enclosing rectangle of .
```

V. Post-Attention Phase

In the post-attention phase, we optimize the original version of the HTM model [14] to estimate the target probability of attention regions. The probability is then fed back to the low-level saliency map, and finally the target-oriented high-level saliency map is generated.

A. The Optimization of HTM

HTM model is the newest layering network model that imitates the structure of the new human neocortex [14]. HTM model takes time and space factors which depict samples into account in order to tackle with ambiguous rule of inference, presenting strong generalization ability. Thus, it has been gradually highlighted in the field of pattern recognition [16]-[19].

Different from most HTM-based applications [15]-[18] which apply the pixel's grayscale as the input layer of HTM, in this research, the low-level visual features extracted in the pre-attention phase are taken as the input layer for the purpose of improving the precision of the model. Fig. 3 shows the structure of our HTM model, where the notes in the second layer conduct the learning and reasoning of the low-level visual features, meanwhile, the notes above the third layer conduct the learning and reasoning of the spatial position relationships. Notes in different layers use the same mechanism to conduct the learning and reasoning process, and they have the same node structure which is formed by a spatial module and a temporal module.

1) Spatial module: The main function of spatial module is to choose the quantization centers of the input samples, that is, to select a few representative samples in the sample space. These centers should be carefully selected to ensure that the spatial module will be able to learn a finite quantization space from an infinite sample space. It is assumed that the learned quantization space in the spatial module of a node is $Q = [q_1, q_2, ..., q_n]$, where q_i is quantization center and N is the number of the existing centers. All the Euclidean distances d between these centers are calculated and their sum S is considered as a distance metric of the quantization space:

$$S = \sum_{i}^{N} \sum_{j}^{N} d(q_i, q_j) \tag{7}$$

when a new input sample q_c appears in the node, we first add q_c to Q, and the distance increment inc caused by q_c can be calculated as follows:

$$inc = \sum_{i}^{N} d(q_i, q_c) \tag{8}$$

The change rate of the distance increment inc/S is then examined against a given threshold η. If $inc/S > \eta$, q_c is retained in Q otherwise, q_c is removed from Q. This algorithm ensures that input samples which contain substantial information will be considered as new quantization centers, whereas those which do not contain representative information will be discarded.

The learning of the spatial module is stopped when the added quantization centers are sufficient to describe the sample space. In practice, the learning is completed when the rate of adding new centers falls bellow a predefined threshold.

2) Temporal module: The temporal module proposed in [14] is suitable in applications where the input samples have obvious time proximity such as video images. However, the input images for training the HTM model rarely share any amount of time correlation in our research. Therefore, instead of the time adjacency matrix proposed in [14], we exploit a correlation coefficient matrix C to describe the time correlation between different samples. We adopt Pearson's coefficient as the

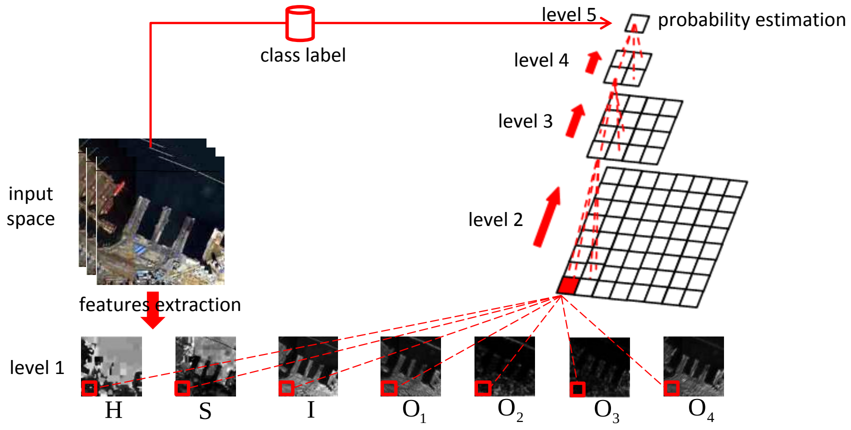


Figure 3. The proposed HTM network structure

measure of correlation. The $N \times N$ correlation matrix, which contains the Pearson's correlation coefficients between all pairs of centers, is calculated as follows:

$$C(q_i, q_j) = \frac{E[(q_i - \mu_{q_i})(q_j - \mu_{q_j})]}{\sigma_{q_i}\sigma_{q_j}} \quad (9)$$

where E is the expected value operator, μ_q and σ_q denotes the mean and the standard deviation of the respective quantization center, respectively. The larger the absolute value of correlation is the stronger the association between the two centers.

A temporal grouping procedure is then utilized to separate the quantization space Q into highly correlated coherent subgroups. The major advantage of replacing the time adjacency matrix with the correlation coefficient matrix is that it enables the grouping procedure to be irrelevant with the temporal sequence of sample images, so as to improve the precision of the model.

In [14], a computationally efficient greedy algorithm is introduced to the temporal grouping procedure. The algorithm is briefly described as follows:

Select the quantization center with the greatest connectivity.

Find the M quantization centers with greatest connectivity to the selected quantization center, and create a new group for the M centers.

Repeat step 1 and step 2 until all quantization centers have been assigned.

The greedy algorithm requires the groups to be disjoint, i.e., no quantization center can be part of more than one group. However, in real applications, rarely groups can be clearly identified. Some quantization centers usually lie near the boundaries of two of more groups. As a result, The greedy algorithm can lead to ambiguity because the quantization centers are forced to be member of only one group. To overcome shortcomings of the greedy algorithm, here we propose a fuzzy grouping algorithm that allows quantization centers to be member of different groups according to the correlation.

We define a $n_q \times n_g$ matrix PQG (n_q and n_g is the numbers of quantization centers and groups, respectively), in which element $PQG[i, j] = p(q_i \mid g_j)$ denotes the conditional probability of quantization centers q_i given the group g_j. $PQG[i, j]$ can be obtained as follows:

$$PQG[i, j] = \sum\nolimits_{q_k \in g_j} \left(C(q_k, q_j) \frac{p(q_k)}{\sum\nolimits_{q_l \in g_j} p(q_l)} \right) \quad (10)$$

where $p(\cdot)$ is the prior probability of quantization centers. $PQG[i, j]$ shows the relative probability of occurrence of coincidence q_i in the context of group g_j, by which we design the fuzzy grouping algorithm, as described bellow: We first use the greedy algorithm to generate a initial grouping solution; then the groups with less than a given threshold n_t centers are removed because they often bring limited generalization; the quantization centers grouped by the greedy algorithm are expected to be the most representative for the group, however, other centers not belonging to the group could have high correlation to centers in the group, we allow a center q_i to be added to a group g_j if $PQG[i, j]$ is high. The fuzzy grouping algorithm is shown in Algorithm 2.

B. The Generation of High-Level Saliency Map

The low-level saliency map predicts interesting locations merely based on bottom-up mechanism. By means of introducing top-down mechanism to obtain more meaningful results, simultaneously inspired by [14], we multiply the probability (estimated by the HTM model) with the according attention region on the low-level saliency map to generate a high-level saliency map. By this way, the suspected target regions are emphasized in the high-level saliency map meanwhile the background interference regions are suppressed. Assuming R^t is the present attention region, P^t is the estimated probability of R^t , Let $S^0{}_{high} = S_{low}$, the current high-level saliency map $S^t{}_{high}$ can be obtained as follows:

$$S^t{}_{high}(x,y) = \begin{cases} S^{t-1}{}_{high}(x,y) * P^t & \text{if } (x,y) \in R^t \\ S^{t-1}{}_{high}(x,y) & \text{otherwise} \end{cases} \tag{11}$$

where $S^{t-1}{}_{high}$ is the corresponding high-level saliency map of the last attention region.

```
Algorithm 2 The fuzzy grouping algorithm
1. Create initial groups using the greedy algorithm.
2. Remove groups with less than $n_t$ (a given threshold) quantization centers.
3. Compute the matrix $PQG$ , each element $PQG[i,j]$ is calculated according to equation(10).
4. for each $q_i$ do
   for each $g_j$ do
      if $PQG[i,j] < \varepsilon$ (we set $\varepsilon = 0.8$ in the experiment) then
         $g_j = g_j \cup q_i$
      end if
   end for
end for
```

VI. Experiment and Discussion

To verify the effectiveness of our model, the experiment for detecting harbor targets is performed on the real optical satellite images. There are 50 images used in the experiment, all from Google Earth. Each image contains 1 to 5 harbor targets. A total of 187 targets are involved in the experiment, and 30 are chosen as the training samples of HTM model. Related parameters in the experiment are set as follows:

The step-size sequence is set according to the size range of targets as:

$$N = \{1, 10\times10, 15\times15, 20\times20, 25\times25, 30\times30, 35\times35, 40\times40, 45\times45, 50\times50\}$$

The threshold value of inc/S is set to *0.08* according to experiences, the learning of the spatial module is completed when the rate of adding new centers falls below *0.2*, i.e. for every *10* new input vectors, when less than *2* new centers are added, the learning procedure should be stopped.

The focus of attention transition is stopped when the transition times reach 20.

A. Accuracy Evaluation of the Optimized HTM

The original version of HTM [14] was implemented for benchmarking against the optimized HTM. Both versions used a 5-level network structure with the input images of size 128 by 128 pixels. Firstly, the efficiency of the original HTM and the optimized HTM were examined. Then the input layer, spatial module and temporal module of the original HTM was replaced individually by the optimized version, and the resulting efficiency was examined. The results are shown in TABLE I.

Obviously, the optimized HTM shows much better performances than the original HTM and both the improvement in the input layer, spatial and temporal module results in higher accuracy than the original version.

The efficiency of the HTM could be further increased with the utilization of a stronger classifier in the top layer [15]. Therefore, we applied Support Vector Machine (SVM) to estimate the probability in the top layer to get higher accuracy results. To further verify the effectiveness of the optimized HTM, a single SVM classifier with a dimensionality reduction process via Principal Component Analysis (PCA) was used as a reference. TABLE II shows the detection accuracy of the original HTM+SVM, the optimized HTM+SVM and SVM+PCA. Obviously, by using a stronger classifier in the top layer, both the original HTM and the optimized HTM achieve higher accuracy than SVM+PCA.

TABLE I. Detection Accuracy of the original HTM and the optimized HTM

	Detection rate of test set (%)	Detection rate of train set (%)
Original HTM	72.51	81.63
Original HTM with feature maps	77.42	85.17
Original HTM with the proposed spatial module	75.12	83.42
Original HTM with the proposed temporal module	79.74	87.94
Optimized HTM	81.34	89.28

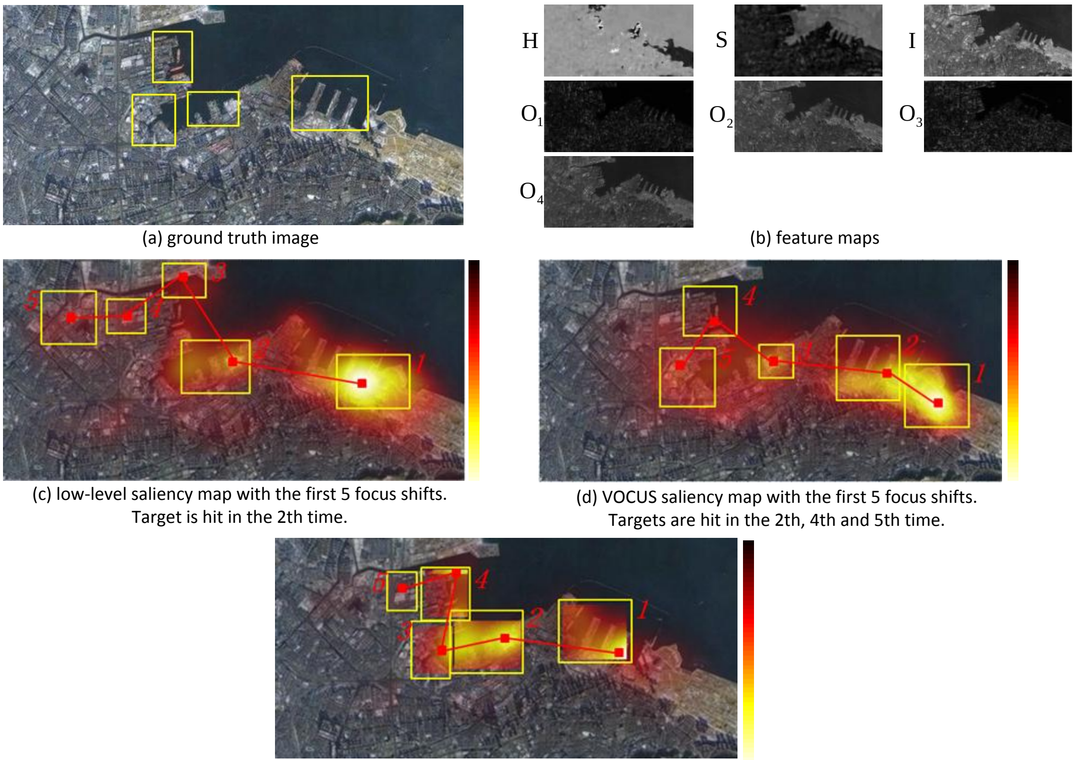
TABLE II. Detection accuracy of original HTM+SVM, the optimized HTM+SVM and SVM+PCA

	Detection rate of test set (%)	Detection rate of train set (%)
Original HTM+SVM	76.73	84.67
Original HTM +SVM	85.81	92.48
SVM+PCA	71.57	82.79

B. Saliency Detection Performance

Three methods are compared for accuracy evaluation, including the low-level saliency map with the bottom-up mechanism only, VOCUS, and the proposed model. Fig. 4 shows an experiment result and it can be seen that: 1) the location of most harbors is significant on the low-level saliency map. However, the most significant regions are not harbors but other ground objects. 2) focus of attention is shifted according to the order of the declining intensity of significance. Moreover, the selection of attention regions shows self-adaption (see Fig. 5 for an example), which is more consistent with the HVS mechanism compared with the option of fixed size. 3) In post-attention phase, the suspected target attention regions on the low-level saliency map are enhanced while the non-target regions are inhibited. 4) our model performs better than VOCUS for it is more efficient to hit target regions.

Fig. 6 shows the performance curve of the three methods. The proposed model presents higher detection precision than the other two methods, and can hit more than 75% targets under 25% saliency ratio.



(a) ground truth image

(b) feature maps

(c) low-level saliency map with the first 5 focus shifts. Target is hit in the 2th time.

(d) VOCUS saliency map with the first 5 focus shifts. Targets are hit in the 2th, 4th and 5th time.

(e) high-level saliency map with the first 5 focus shifts. All Targets are hit in the first 4 shifts. The probability of the 5 attention regions, in sequences, is 0.77, 0.86, 0.73, 0.69, 0.21.

Figure 4. Experiment results of low-level saliency map, VOCUS and high-level saliency map.

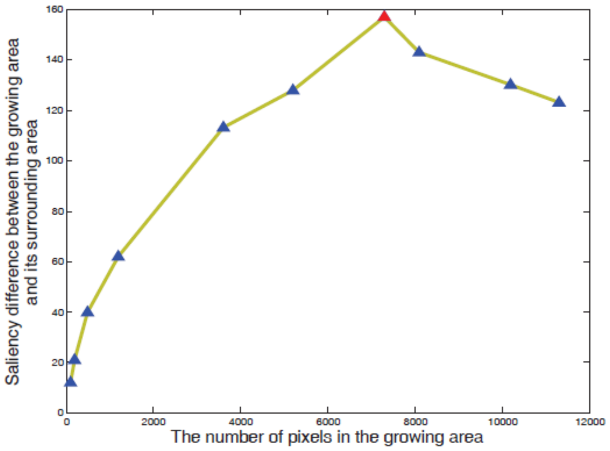


Figure 5. The self-adaption region growing of the first focus in Fig. 4(c). The growth is terminated in the downward inflection point (marked as a red triangle in the figure).

In order to further assess the precision of our model, we introduce three definitions: 1) hit number: the rank of the focus that hits the target in order of saliency; 2) average hit number: the arithmetic mean of the hit numbers of all targets 3) detection rate: the ratio between the hit target number in the precious 10 focus shifts and the total target number. The accuracy analysis of the three approaches is expressed in TABLE III and Fig. 7.

It can be seen from the experiment results that due to the introduction of top-down mechanism, VOCUS and our method are better than the low-level saliency map with bottom-up mechanism only. At the same time, our approach is excellent to VOCUS. This is mainly because the top-down procedure of VOCUS only takes the weight of lower feature into consideration while that of our approach applies HTM model comprehensively took account of the lower features and spatial location relationship, possessing more effective target orientation.

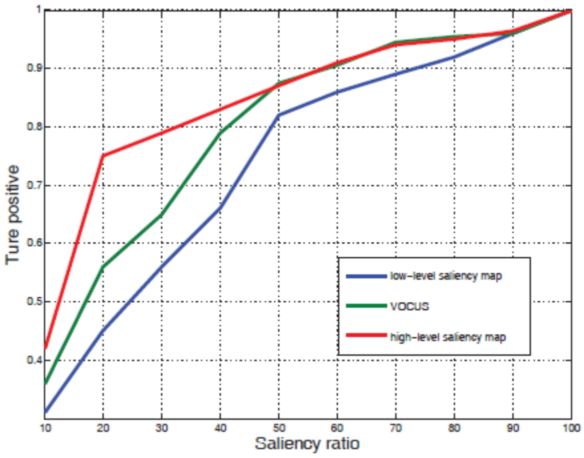


Figure 6. The performance curve of low-level saliency map, VOCUS and high-level saliency map. Saliency ratio is the ratio between the size of saliency area and of the total image.

TABLE III. AVERAGE HIT NUMBER AND DETECTION RATE OF THE THREE METHODS.

	Low-level saliency map	VOCUS	The proposed model
Average hit number	11.67	8.46	3.75
Detection rate (%)	18.82	37.1	73.12

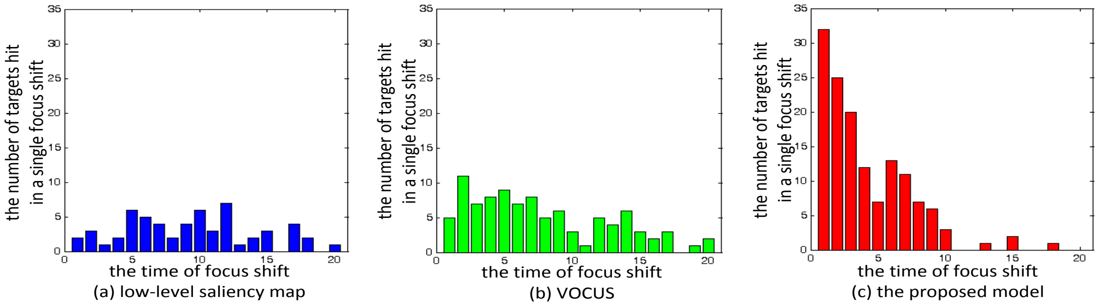


Figure 7. The number of targets hit in focus shifts. The total hit target number in the precious 10 focus shifts of the three methods is 35, 69, 136, respectively. It is obviously that our model can hit more targets in the first few focus shifts.

VII. Conclusion

In this paper we propose a novel target-oriented visual saliency detection model. Inspired by the structure of the human vision system, we build the model with three functional modules, i.e., pre-attention phase module, attention phase module and post-attention phase module. In the pre-attention phase module, a low-level bottom-up saliency map is generated to locate attention regions with low-level visual stimuli. In the attention phase module, we propose an effective method for focus shift and attention region selection to focus on the suspected target regions rapidly and accurately. In the post-attention phase, the original HTM is optimized in several respects including the input layer, the spatial module and the temporal module, leading to a robust probability estimation. Experimental results demonstrate that our model presents higher detection precision, compared with models of both low-level bottom-up saliency map and VOCUS model. It is proved that the proposed model provides a feasible way to integrate top-down and bottom-up mechanism in visual saliency detection.

Acknowledgment

This work was supported by the National Science Foundation of China No. 61203239, No. 61005067 and No. 61101222.

References

[1] M. Li, L. Xu, and M. Tang, "An extraction method for water body of remote sensing image based on oscillatory network," *Journal of multimedia*, vol. 6, no. 3, pp. 252–260, 2011.

[2] Q. Zhang, G. Gu, and H. Xiao, "Image segmentation based on visual attention mechanism," *Journal of multimedia*, vol. 4, no. 6, pp. 363–369, 2009.

[3] B. Yang, Z. Zhang, and X. Wang, "Visual important-driven interactive rendering of 3d geometry model over lossy wlan," *Journal of networks*, vol. 6, no. 11, pp. 1594–1601, 2011.

[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. 11, no. 20, pp. 1254–1259, 1998.

[5] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems*, 2007, pp. 542–552.

[6] X. Hou and L. Zhang, "Saliency detection: a spectral residual approach," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[7] Y. Yu, B.Wang, and L.Zhang, "Bottom-up attention: Pulsed pca transform and pulsed cosine transform," *Cognitive Neurodynamics*, vol. 5, no. 4, pp. 321-332, 2011.

[8] S. Frintrop, "Vocus: A visual attention system for object detection and goal-directed search," *Lecture Notes in Artificial Intelligence, Berlin Heidelberg*, 2006.

[9] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *Journal of Vision*, vol. 11, no. 3, pp. 1–15, 2011.

[10] ——, "Learning visual saliency," in *Information Sciences and Systems Conference*, 2011, pp. 1–6.

[11] R. Peters and L. Itti, "Beyond bottom-up: Incorporating task dependent influences into a computational model of spatial attention," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007, pp.1–8.

[12] B. Scholkopf, J. Platt, and T. Hofmann, "A nonparametric approach to bottom-up visual saliency," in *Advances in Neural Information Processing Systems*, 2007, pp. 689–696.

[13] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *International Conference on Computer Vision*, 2009, pp. 2106–2113.

[14] J. Hawkins and D. George, "Hierarchical temporal memory: Concepts, theory and terminology," *Whitepaper, Numenta Inc*, 2006.

[15] I. Kostavelis and A. Gasteratos, "On the optimization of hierarchical temporal memory," *Pattern Recognition Letters*, vol. 33, no. 5, pp. 670–676, 2012.

[16] A. Csap, P. Baranyi, and D. Tikk, "Object categorization using vfa-generated nodemaps and hierarchical temporal memories," in *IEEE International Conference on Computational Cybernetics*, 2007, pp. 257-262.

[17] W. Melis and M. Kameyama, "A study of the different uses of colour channels for traffic sign recognition on hierarchical temporal memory," in *Conference on Innovative Computing, Information and Control*, 2009, pp. 111–114.

[18] T. Kapuscinski, "Using hierarchical temporal memory for vision-based hand shape recognition under large variations in hands rotation," in Artificial Intelligence and Soft Computing, 2010, pp. 272–279.

[19] D. Rozado, F. B. Rodriguez, and P. Varona, "Extending the bioinspired hierarchical temporal memory paradigm for language recognition," *Neurocomputing*, vol. 79, pp. 75–86, 2012.