Region of Interest Detection Based on Visual Attention and Threshold Segmentation in High Spatial Resolution Remote Sensing Images

Libao Zhang[1*] and Hao Li[2]
[1] College of Information Science and Technology, Beijing Normal University
Beijing, P. R. China
[e-mail: libaozhang@bnu.edu.cn]
[2] College of Information Science and Technology, Beijing Normal University
Beijing, P. R. China
[e-mail: hower708@gmail.com]
*Corresponding author: Libao Zhang

Received April 19, 2013; revised June 30, 2013; accepted July 20, 2013; published August 30, 2013

Abstract

The continuous increase of the spatial resolution of remote sensing images brings great challenge to image analysis and processing. Traditional prior knowledge-based region detection and target recognition algorithms for processing high resolution remote sensing images generally employ a global searching solution, which results in prohibitive computational complexity. In this paper, a more efficient region of interest (ROI) detection algorithm based on visual attention and threshold segmentation (VA-TS) is proposed, wherein a visual attention mechanism is used to eliminate image segmentation and feature detection to the entire image. The input image is subsampled to decrease the amount of data and the discrete moment transform (DMT) feature is extracted to provide a finer description of the edges. The feature maps are combined with weights according to the amount of the "strong points" and the "salient points". A threshold segmentation strategy is employed to obtain more accurate region of interest shape information with the very low computational complexity. Experimental statistics have shown that the proposed algorithm is computational efficient and provide more visually accurate detection results. The calculation time is only about 0.7% of the traditional Itti's model.

Keywords: Image processing, region of interest, visual attention, threshold segmentation

This research was supported National Natural Science Foundation of China (61071103), and Fundamental Research Funds for the Central Universities (2012LYB50). We express our thanks to these editors and reviewers who checked our manuscript.

http://dx.doi.org/10.3837/tiis.2013.08.006

1. Introduction

Remote sensing technology, which is the most important tool to collect geographic data, has been applied to various fields. In recent years, with the launch of several advanced remote sensing satellites, such as IKONOS and GeoEye-1, the spatial resolution of remote sensing images has been increased greatly, meaning that the information in a remote sensing image has been much more than a low spatial resolution remote sensing image, which brings great challenge to the analysis and processing of high spatial remote sensing images [1, 2].

Generally speaking, the so-called high spatial resolution remote sensing images are the ones with resolution of meter-level. Comparing with the traditional low spatial resolution remote sensing images, the high spatial resolution images have the following properties:

1) The amount of data in an image is extraordinary large. A high spatial resolution remote sensing image contains complicated spatial information, clear details and well-defined geography objects.
2) The intensity, structure, shape and texture information in a high spatial resolution remote sensing image is abundant and clear. The background information becomes much more complex.

All these properties allow people to observe the earth in detail at a smaller scale. At the same time, a more efficient information processing technology for high spatial resolution remote sensing images is required.

Among various applications of remote sensing images, target detection is one of the most popular ones. Traditional approaches to detecting targets in remote sensing images are commonly using statistical pattern recognition methods based on features such as spectrum and texture. These methods can be divided into two categories: the supervised and the unsupervised [3, 4]. The supervised method uses classifiers such as a neural network [5] and a support vector machine (SVM) [6] which are trained by a prior knowledge library. The unsupervised method uses the clustering algorithms like C-means [7] and ISODATA [8]. However, with the increase of spatial resolution, the geography objects cannot be detected using single classifier. In recent years, new methods based on the expert knowledge have also been proposed and put into practice [9]. Most of these methods need a prior knowledge library which is difficult to build and has a great influence on the detection result. Moreover, global searching is an essential part of these methods, which is both time-consuming and storage expensive [2, 10, 11].

Nowadays, the study of human visual system (HVS) has become an important trend. HVS has an ability to suppress some information in a scene and draw the attention rapidly toward the unusual [12]. This mechanism is called visual attention. The content that draws human beings' attention has a characteristic called visual saliency to stand out from the surroundings. A point that draws people's attention is called a focus of attention (FOA) and the region centered by FOA is called a region of interest. In recent years, several models have been proposed to simulate visual attention mechanism. Inspired by the feature integration theory, Itti et al. have defined a system in which multi-scale features including intensity, color and orientation are extracted to generate the saliency map through center-surround difference [13, 14]. Besides Itti's model there are valuable achievements obtained by other researchers from other perspectives. Achanta et al. proposed a completely computational model computing the saliency map in frequency domain and this model can produce saliency map with the original size [15]. Harel et al. proposed a model similar to Itti's but using feature vectors to create "activation maps" to build the saliency map, which is both biologically based and computational [16]. Besides all these models, valuable achievements were also made on quality assessment of visual saliency models and feature selection for visual saliency computation [17, 18].

In the context that the target to be detected usually stands out from its surroundings, it is possible to bring visual attention mechanism to target detection in high spatial resolution remote sensing images, eliminating the prior knowledge library and global searching. Generally speaking, the regions that may contain the valuable targets take only small proportion of the whole image. Region of interest detection can greatly reduce the data needed for further processing, which is of great value for the real-time processing.

In this paper, a model for region of interest detection based on visual attention and threshold segmentation (VA-TS) in high spatial resolution remote sensing images is proposed (see **Fig. 1**). The original image is firstly pretreated to decrease the amount of data. The intensity, color, orientation and the DMT features are extracted to compute the saliency map. Furthermore, a new feature competition strategy is employed to attach different weights to different feature maps according to the size and saliency of the salient regions. The detected ROIs are described using threshold segmentation which improves the accuracy and the speed of detection. Experiments have shown that the proposed model is time-saving and the detection result is satisfying.

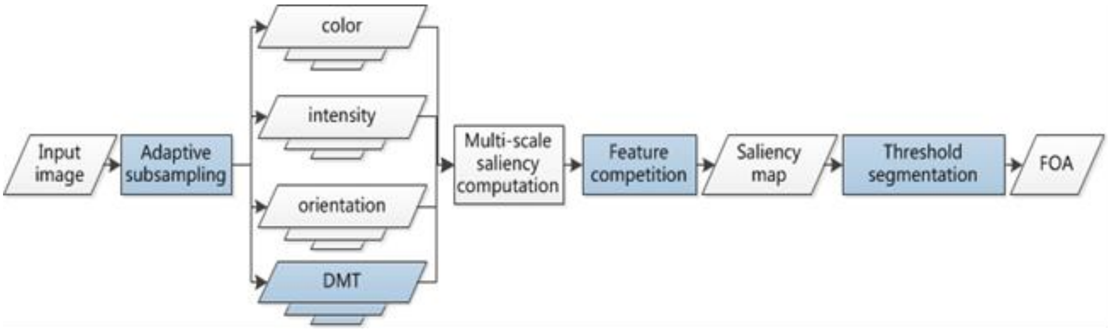


Fig. 1. Framework of the VA-TS

2. Traditional Visual Attention Models and Their Limitations

In the traditional visual attention models, saliency map computation is the key part. The methods to compute a saliency map can be classified as biologically based, purely computational, or a combination of the two. In this section, we will introduce the typical method of each category, including Itti's model, Achanta's model and Harel's model.

Itti et al. proposed a model to simulate human's bottom-up visual attention mechanism in the year 1998 [13]. In this model, the early visual features are extracted from the input image, including intensity, color and orientation. For each feature image, nine spatial scales are created using dynamic Gaussian pyramid. A linear "center-surround difference" operation, denoted as "Θ", including interpolating the courser to the finer scale and point-by-point subtraction, is used between different levels of the pyramid to compute multi-scale visual saliency and generate several feature maps. All these feature maps are globally prompted using the normalization operation $N(\cdot)$ and are added using the across-scale combination operation "$\oplus$" including reduction each map to scale four and point-by-point addition to generate the conspicuity maps. The final saliency map is computed as the addition of these three conspicuity maps (see **Fig. 2**).

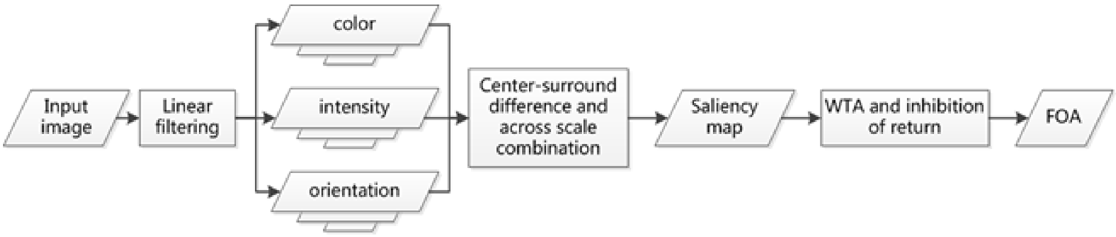

Input image
Linear filtering
color
intensity
orientation
Center-surround difference and across scale combination
Saliency map
WTA and inhibition of return
FOA

Fig. 1. Itti's visual attention model

Itti's model accords with HVS very well and it utilizes various visual features to generate the saliency map. Nevertheless, Itti's model has some drawbacks. The complicated convolution and float computation lead to a high computational complexity. The low-resolution saliency map makes it difficult to produce an accurate detection result.

In 2009, Achanta introduce a method for salient region detection that outputs full resolution saliency maps with well-defined boundaries of salient objects (see **Fig. 3**). These boundaries are preserved by retaining substantially more frequency content from the original image than other existing techniques. This method exploits features of color and luminance, is simple to implement, and is computationally efficient. Compared with Itti's model, this method is purely computational and discards the biology mechanism of HVS. In Achanta's model, the original image is firstly transformed to CIELAB color space and each pixel location is a $[L,a,b]^T$ vector. The saliency value of each pixel location is computed as the Euclidean distance between the mean image feature vector and the corresponding image pixel vector value in the Gaussian blurred version (using a 5×5 separable binomial kernel) of the original image. An image-adaptive threshold is set to binarize the saliency map. The adaptive threshold (*Ta*) value is determined as two times the mean saliency of a given image.

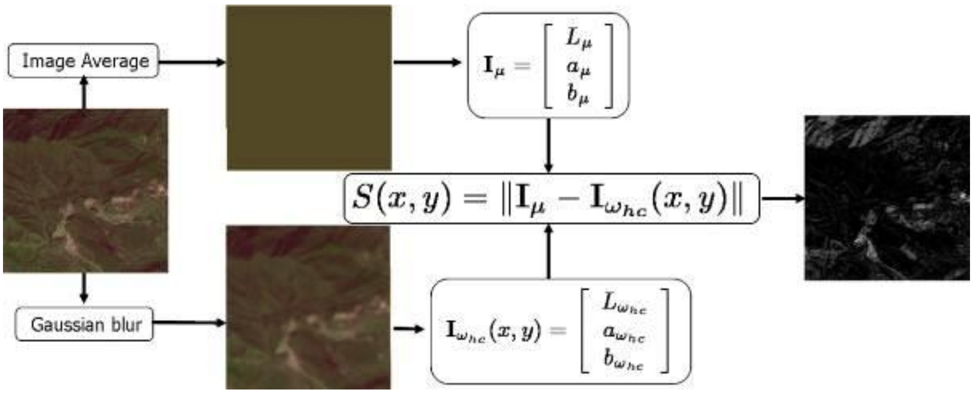


Image Average
Gaussian blur
$\mathbf{I}_\mu = \begin{bmatrix} L_\mu \\ a_\mu \\ b_\mu \end{bmatrix}$
$S(x,y) = \|\mathbf{I}_\mu - \mathbf{I}_{\omega_{hc}}(x,y)\|$
$\mathbf{I}_{\omega_{hc}}(x,y) = \begin{bmatrix} L_{\omega_{hc}} \\ a_{\omega_{hc}} \\ b_{\omega_{hc}} \end{bmatrix}$

Fig. 2. Achanta's visual attention model

Achanta's model is easy to implement, fast, and provides full resolution saliency maps. However, this method involves only intensity and color features, eliminating some valuable characteristics of HVS. Sometimes this method produces detection result that contains some unexpected regions.

Harel’s approach computes the graph-based visual saliency, which is both biological and computational. This method extracts image features just like Itti’s approach to produce feature vectors. Several “activation maps” are created using these feature vectors with Markov chains. A dissimilarity function is defined between each pair of nodes in the chain and the weight between those is proportional to the dissimilarity function. The weights are normalized to 1 and treated as the transition probability. All these activation maps are normalized and combined to produce the saliency map.

Harel’s method produces low resolution saliency map which is 1/64 of the original image size, so it has badly defined salient object borders just like Itti’s method.

Once a saliency map is obtained, the next step will be the extraction of ROIs. Generally speaking, there are three kinds of methods to do this.

In Itti’s method, the search for the region of interest is established on the saliency map under the “winner-take-all” network [19, 20] and “inhibition of return” mechanisms [21]. The global maximum is selected, and the disk region around it with a fixed radius is regarded as the region of interest. The region of interest previously detected is inhibited before the new search procedure. It is apparent that the disk-shape detection results are not accurate. To overcome this drawback, Ma et al. proposed the fuzzy growing method to perform an accurate detection of ROIs [22]. Since we just care about the detection results rather than the shift of ROIs, it is not necessary to perform an iterating search. Threshold segmentation is now a popular way to extract the most salient objects. But this method may encounter difficulty when the gray-scale of the salient objects is not evenly distributed.

3. Region of interest Detection Model Based on VA-TS in High Spatial Resolution Remote Sensing Images

Since we are aiming at building up a rapid and effective detection model, these limitations mentioned in the previous section should be taken into consideration seriously. Finally we found some methods to resolve these problems and proposed the region of interest detection model based on visual attention and threshold segmentation (VA-TS) in high spatial resolution remote sensing image. In this section, we will introduce VA-TS model in detail.

3.1 Pretreatment to the Input Image

Once we obtain a remote sensing image, which usually has a format of tiff and contains multi-spectrum information, it should be transformed to the format which is suitable to be watched by people in order to apply visual attention mechanism. Since the color feature is needed in the following processing, the original image should be transformed to a chromatic image. Here we choose three spectrums and assign red, green and blue respectively to generate a RGB image.

In order to decrease the data to be processed, an adaptive subsampling strategy is used in our model. This strategy subsamples the original image using the Gaussian pyramid. The original image is filtered and then subsampled to a specific level p of the pyramid according to the original resolution.

$$p = \lfloor \log_2 M - \log_2 N \rfloor \tag{1}$$

In Eq. (1), p is the level of subsampling, M is the shorter border of the original image and N is set to 512 to balance the time complexity and the detection results. We have $M \geq 2N$. It is easy to see that the size of the subsampled image, I_s, is $1/2^{2p}$ of the input image. The following feature extraction is then done on I_s.

It is inevitable that the subsampled image will lose some information, while it is not as certain that the reduction of information will lead to a bad detection result. On one hand, some tiny ROIs may be completely lost during the subsampling, but small ROIs are often with low importance and can be suppressed. On the other hand, the reduction of information will make the detection result center on regions with largest saliency value and size.

Fig. 4 shows the ability to simplify the computation of the adaptive subsampling. In this figure, p stands for the subsampling level and t stands for the processing time. In this experiment, the adaptive subsampling is used to process the input image (with a resolution of 2048×2048 pixels) of Itti's model. The processing time spent on the original image is nearly 600 seconds while the processing time spent on the 2-level subsampled image is only 5.6 seconds. When the subsampling level increases to 3 and above, meaning that the size of the subsampled image is smaller than 512×512, the overall processing time has little change. That is why we choose to subsample the original image to the size which is nearest and larger than 512×512 pixels. Adaptive subsampling strategy is quite helpful in decreasing the computational complexity.

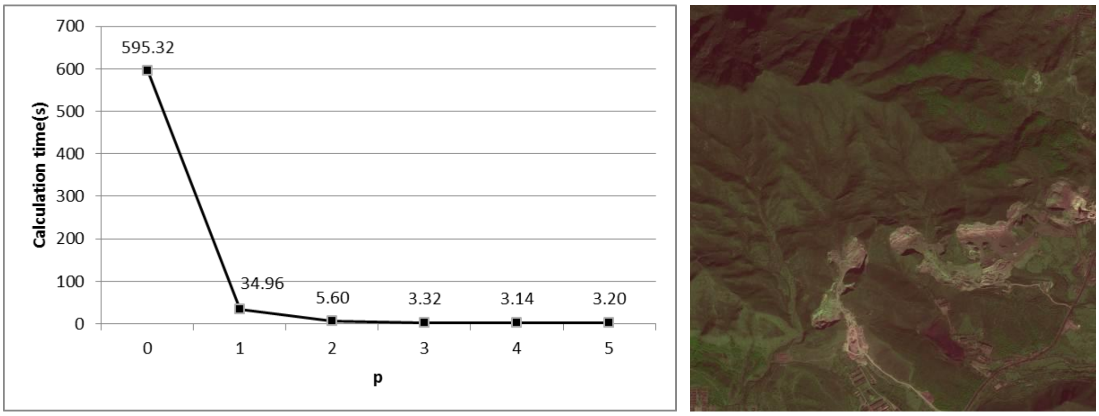


Fig. 3. Calculation time based on different subsampling levels

3.2 Feature Extraction

The subsampled image I_s acts as the input of the following feature extraction, which is similar to Itti's model.

Let r, g and b be the red, green and blue channels of the input image. The intensity image I is obtained as:

$$I = (r + b + g)/3 \tag{2}$$

The corresponding Gaussian pyramid is obtained as $I(\sigma)$ where $\sigma \in [0...6]$. Then the center-surround difference operation is applied with $c \in \{1,2,3\}$ and $s = c + \delta$, $\delta \in \{2,3\}$:

$$I(c,s)=\left|I(c)-I(s)\right| \tag{3}$$

As for the color feature, the *r*, *g* and *b* channels are first normalized by *I* to eliminate the influence of intensity, and then four broadly-tuned color channels are generated:

$$R=r-(g+b)/2 \tag{4}$$

$$G=g-(r+b)/2 \tag{5}$$

$$B=b-(r+g)/2 \tag{6}$$

$$Y=(r+g)/2-\left|r-g\right|/2-b \tag{7}$$

R, *G*, *B* and *Y* refer to red, green, blue and yellow respectively (negative values are set to zero). Four Gaussian pyramids are generated for these color channels, namely $R(\sigma)$, $G(\sigma)$, $B(\sigma)$ and $Y(\sigma)$. The center-surround difference is computed according to a so-called "color double-opponent" system [23]:

$$RG(c,s)=\left|(R(c)-G(c))\Theta(G(s)-R(s))\right| \tag{8}$$

$$BY(c,s)=\left|(B(c)-Y(c))\Theta(Y(s)-B(s))\right| \tag{9}$$

The orientation feature is obtained by using the Gabor pyramids [24], where $\sigma\in[0...8]$ referring to the scale and $\theta\in\{0^o,45^o,90^o,135^o\}$ representing the preferred orientation. Orientation feature maps are computed as follows:

$$O(c,s,\theta)=\left|O(c,\theta)\Theta O(c,s)\right| \tag{10}$$

Now that we are aiming at the processing of high spatial resolution remote sensing images, we should consider the unique characteristics of them. High spatial resolution remote sensing images contain abundant texture information which is useful in region of interest detection, because the ROIs usually have different texture characteristics from the background region. The discrete moment transform (DMT) is a kind of structural feature, which reflects the intensity distribution of a small region [25] and is widely used in image analysis and computer vision. The DMT is computed as follows, where *i* and *j* are the coordinates of the corresponding pixel, and $p,q=0,1,2\cdots$.

$$DMT_{p,q}(i,j)=\sum_{r=-k}^{k}\sum_{s=-k}^{k}I(i-r,j-s)(i-r)^p(j-s)^q \tag{11}$$

In VA-TS model, we choose $k=1$ and $(p,q)=(1,0),(0,1),(1,1)$ to obtain 3 DMT feature images $D(p,q)$. The corresponding Gaussian pyramid is generated as $D(p,q,\sigma)$. The feature maps are computed:

$$D(p,q,c,s)=\left|D(p,q,c)\Theta D(p,q,s)\right| \tag{12}$$

3.3 Saliency Map Computation

In Itti's model, a normalization operation is applied to each of the maps to be fused which consists of:

1) normalizing the values in the map to a fixed range $[0\cdots M]$, in order to eliminate modality-dependent amplitude differences;
2) finding the location of the map's global maximum M and computing the average $\overline{m}$ of all its other local maxima;
3) Globally multiplying the map by $(M-\overline{m})^2$.

This operation measures how different the most salient region is from the average. When the difference is large, the map will be strongly prompted. What's more, this operation is computationally simple. The normalization will strongly enhance the map with only one location which is much more conspicuous than others. However, when there are several conspicuity locations distributed sparsely in the map, the operation will suppress the map, which is an unexpected result [26].

A new feature competition strategy is proposed in VA-TS model. This strategy is based on one obvious fact: the salient regions in the feature maps or in the conspicuity maps should stand out strongly from their surroundings. Points with intensity above a pre-set threshold T are selected as "salient points," representing the ROIs. The associated maps are weighted according to the difference between the average intensity of the salient points and the average intensity of the whole map. The weighting computation includes the following steps:

1) Normalize all the maps to a fixed range $[0\cdots 1]$, and interpolate the map to level 1 of the Gaussian pyramid. The motivation for interpolating all maps to level 1 is to retain as much useful information as possible.
2) Compute the thresholds. Otsu's method [27] is applied to each map to compute a series of thresholds, whose average is the salient point threshold *T*.
3) Use *T* to find all the salient points in one map and then compute the average intensity $\overline{m}$ of all the salient points.
4) Compute the average intensity $\overline{M}$ of the whole map.
5) Compute the normalized weight w_{n_i} for each map as

$$w_i=[\overline{M}_i-\overline{m}_i]^2 \tag{13}$$

$$w_{n_i}=\frac{w_i-(w_i)_{\min}}{(w)_{\max}-(w_i)_{\min}} \tag{14}$$

The final conspicuity maps are computed as follows:

$$\overline{D} = \sum_{(p,q)=(1,0),(0,1),(1,1)} w_{n_{p,q}} \times \left(\bigoplus_{c=1}^{3} \bigoplus_{s=c+2}^{c+3} w_{n_{c,s}} \times D(c,s,p,q) \right) \tag{15}$$

$$\bar{I} = \bigoplus_{c=1}^{3} \bigoplus_{s=c+2}^{c+3} w_{n_{c,s}} \times I(c,s) \tag{16}$$

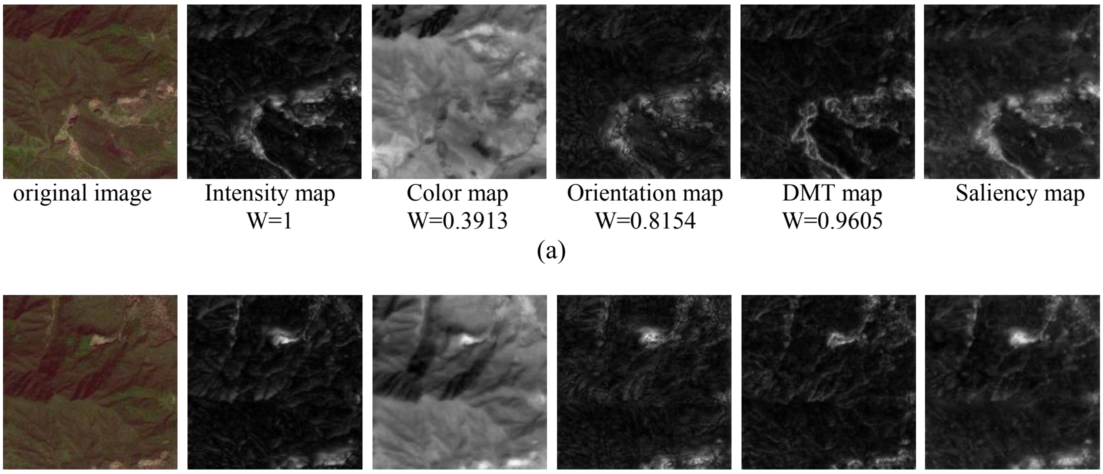
$$\overline{C} = \bigoplus_{c=1}^{3} \bigoplus_{s=c+2}^{c+3} w_{n_{c,s}} \times \left(w_{n_{RG}} \times RG(c,s) + w_{n_{BY}} \times BY(c,s) \right) \tag{17}$$

$$\overline{O} = \sum_{\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}} w_{n_\theta} \times \left(\bigoplus_{c=1}^{3} \bigoplus_{s=c+2}^{c+3} w_{n_{c,s}} \times O(c,s,\theta) \right) \tag{18}$$

The "$\oplus$" consists of interpolating the map to level 1 and point-by-point addition. The saliency map S is computed as:

$$S = w_{n_{\bar{I}}} \times \bar{I} + w_{n_{\overline{C}}} \times \overline{C} + w_{n_{\overline{O}}} \times \overline{O} + w_{n_{\overline{D}}} \times \overline{D} \tag{19}$$

Fig. 5 shows example results of feature maps and the final saliency map. As we can see, in both cases the color map is almost useless thus is attached the smallest weight. On the other hand, we can notice the DMT map strongly prompts the edges in the image.



original image | Intensity map W=1 | Color map W=0.3913 | Orientation map W=0.8154 | DMT map W=0.9605 | Saliency map

(a)

original image | Intensity map W=1 | Color map W=0.4106 | Orientation map W=0.8022 | DMT map W=0.8098 | Saliency map

(b)

Fig. 4. Examples of conspicuity Maps and final saliency maps

3.4 Description of Region of Interest

Just as described in section 2, the iterating region of interest description method in Itti's model and the fuzzy growing method are time consuming. Since we are aiming at building up a rapid model, the threshold segmentation method is used in VA-TS model.

The size of the saliency map is 1/2 of the subsampled size. We have to interpolate the saliency map to the original size to obtain the final detection result.

The threshold is set using Otsu's method. The algorithm assumes that the image contains two classes of pixel or bi-modal histogram, and then calculates the optimum threshold separating those two classes so that their intra-class variance is minimal.

In Otsu's method we exhaustively search for the threshold that minimized the intra-class variance, defined as a weighted sum of variances of the two classes:

$$\sigma_w^2(t) = w_1(t)\sigma_1^2(t) + w_2(t)\sigma_2^2(t) \tag{20}$$

Weights ω_i are the probabilities of the two classes separated by a threshold t and σ_i^2 are variances of these classes.

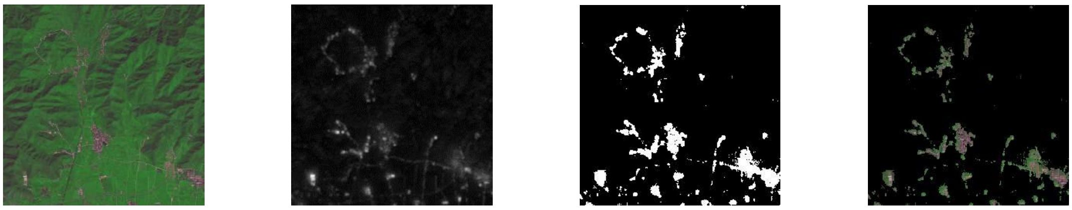
Otsu shows that minimizing the intra-class variance is the same as maximizing inter-class variance:

$$\sigma_h^2(t) = \sigma^2 - \sigma_\omega^2(t) = \omega_1(t)\omega_2(t)[\mu_1(t) - \mu_1(t)]^2 \tag{21}$$

Which is expressed in term of class probabilities ω_i and class means μ_i .

These more details about Otsu's method can be found from [27].

Once the threshold is obtained, the saliency can be turned into a binary image and the final detection result is obtained by multiplying the binary image with the original image (**Fig. 6**).



(a) original image (b) saliency map (c) binary mask (d) detection result

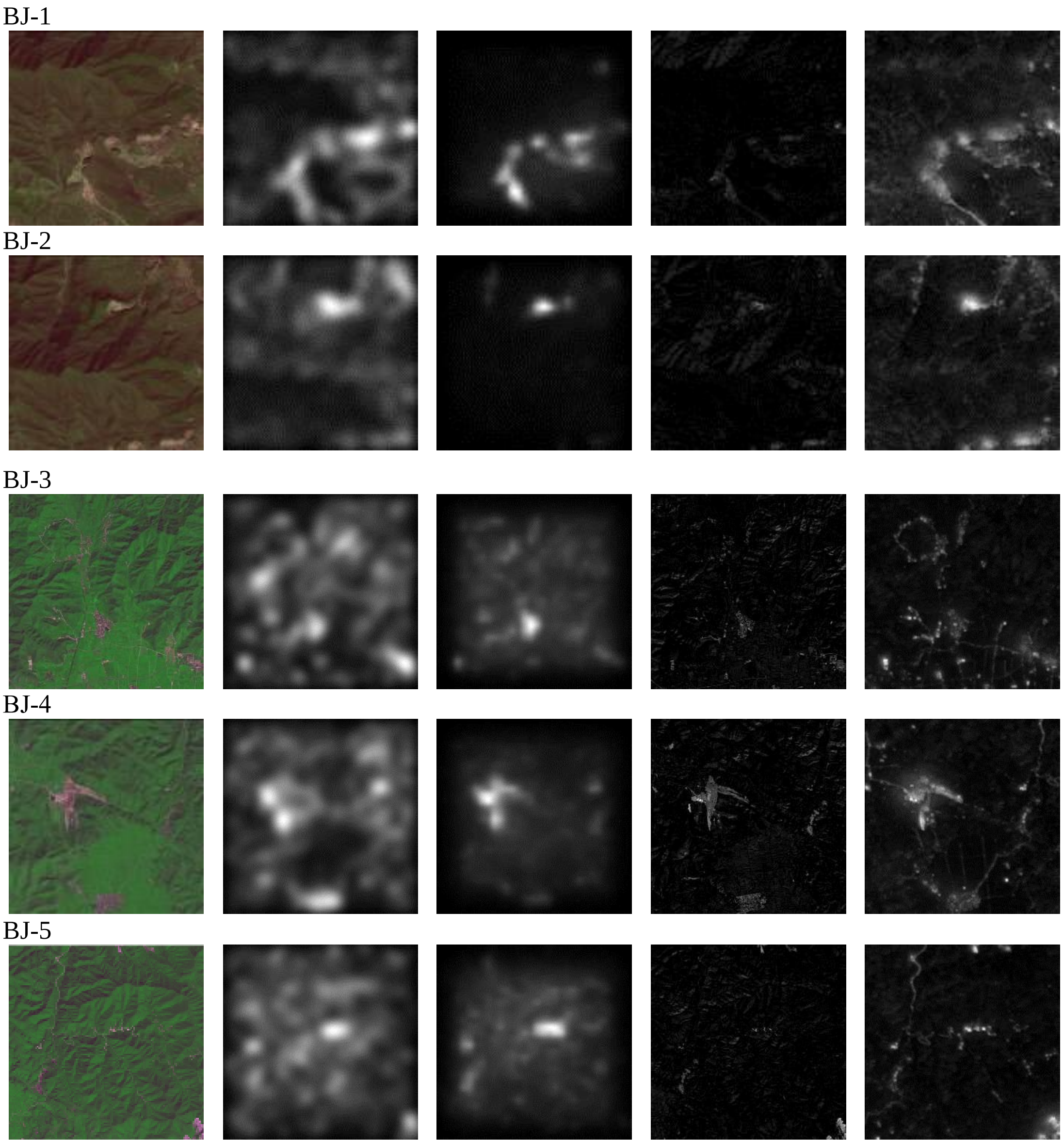
Fig. 5. Example of obtaining the final detection result

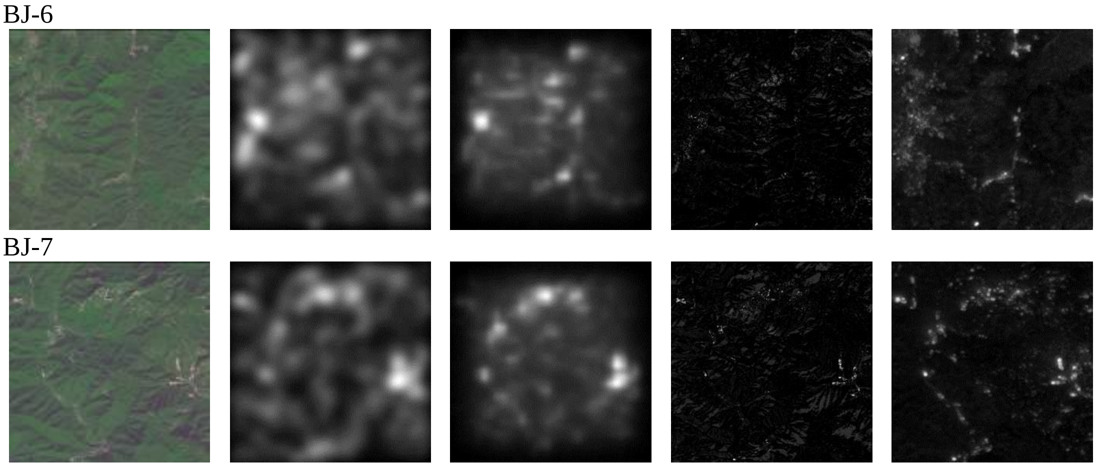
4. Experimental Results and Analysis

To evaluate the performance of VA-TS model, several experiments were established on selected high spatial resolution remote sensing images. These images were taken from a large image depicting Beijing area taken by the SPOT5 satellite. Three bands were taken from the four bands image to make an approximation to the RGB color space. All the selected images have the resolution of 2048×2048 pixels and marked as BJ-1 to BJ-7. The experiments were established on a computer with an AMD Phenom(TM) II X6 1055T Processor 2.8GHz and a 4G memory.

4.1 Performance Comparison

Fig. 7 shows the comparison of saliency maps across Itti's model, Harel's model, Achanta's model and the VA-TS model. The saliency maps from Itti's model and Harel's model have low resolutions and cannot make an accurate description of ROIs. At the same time, saliency maps from Itti's model include much back ground areas. Achanta's model produces saliency maps of the same size with the original images, so we can see much more details and well-defined borders. But sometimes Achanta's model assigned great saliency values to pixels in background areas (BJ-2 and BJ-7). Saliency maps from the VA-TS model have a resolution of 256×256, making it possible to produce a much more accurate detection than Itt's and Harel's model. The most salient areas are strongly prompted in the saliency maps and can be separated from the surroundings easily. What's more, thin edges are better illustrated than Achanta's model (BJ-3, BJ-4 and BJ-5).

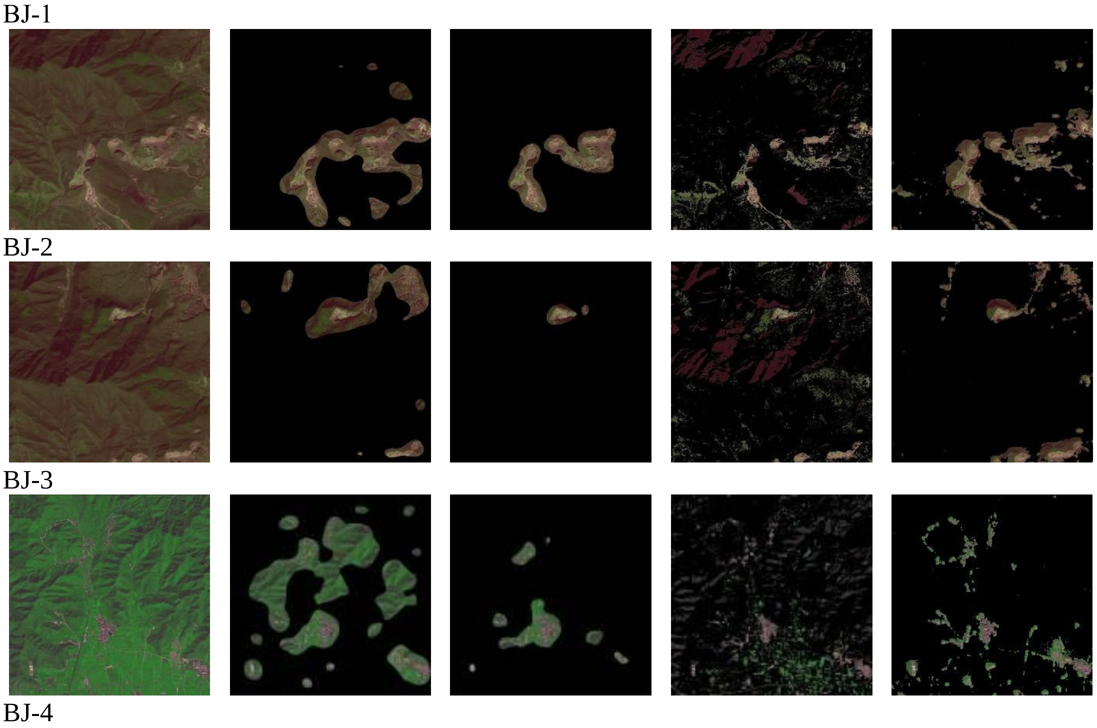


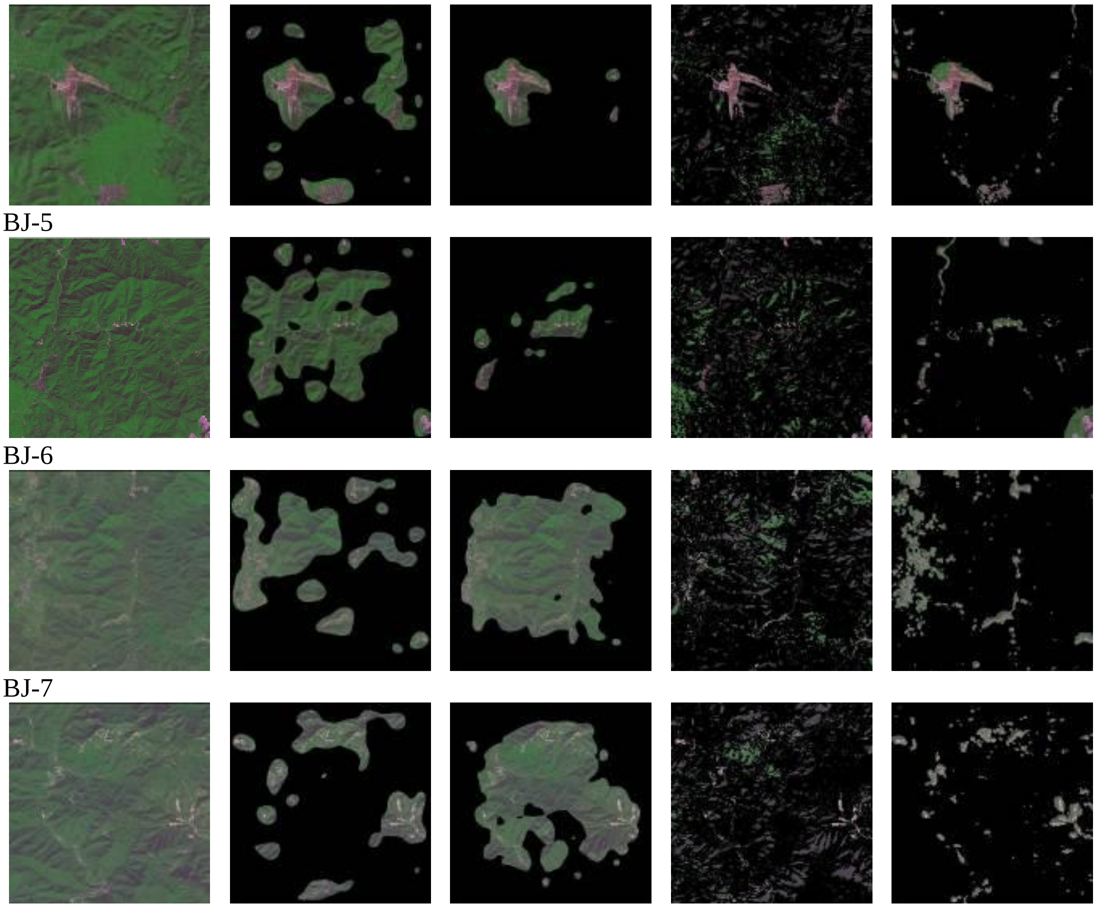


(a) Original image (b) Itti’s method (c) Harel’s method (d) Achanta’s method (e) VA-TS model

Fig. 6. Comparisons of saliency maps

Fig. 8 shows the final detection results of these four models. It is clears that detection results from Itti’s and Harel’s model fail to make an accurate description of the salient regions. Itti’s model can find most of the visually salient regions though with much back ground information while Harel’s model sometimes fails to detect some salient regions (BJ-1, BJ-2, BJ-3 and BJ-4). Achanta’s model has the ability to make an accurate description of salient areas but it has a significant drawback. The salient regions are difficult to figure out because they are hidden in so many detected regions. Finally, the VA-TS model produces much more accurate detection than Itt’s and Harel’s model. Though the accuracy of the detected regions is lower than Achanta’s model, the detection result involves few completely useless regions.





(a) Original image (b) Itti's method (c) Harel's method (d) Achanta's method (e) VA-TS model

Fig. 7. Comparisons of final detection results

4.2 Calculation Time Comparison

Fig. 9 shows the average saliency map calculation time for 20 randomly selected remote sensing images in terms of the resolution of the image in pixels. Different resolutions were obtained by resizing the initial image (2048×2048). One can see that for lower resolution images (with size of 1024×1024 pixels) Achanta's model (red bars) outperforms the others. For resolution from 2048×2048 pixels up to 4096×4096 pixels, Harel's approach is the quickest while the calculation time of Achata's model increases fast and finally became the most time-consuming method. The VA-TS model is the slowest for resolution 1024×1024 pixels and 2048×2048 pixels, but the calculation time increases slowly. For the resolution of 4096×4096, the VA-TS model is just a little slower than Harel's approach and much faster than Achanta's model. It is clear that the VA-TS model has an advantage efficient for high spatial resolution remote sensing images.

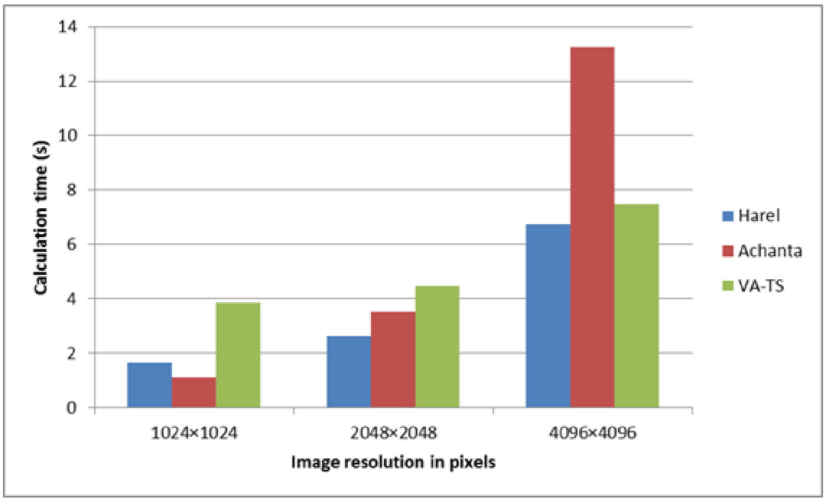


Fig. 8. Comparison of average saliency map calculation time

Table 1 shows the calculation time of region growing, WTA and inhibition of return, and threshold segmentation for different remote sensing images. RG means region growing. WTA & IR means winner-take-all and inhibition of return. TS mean threshold segmentation. Threshold segmentation is the most time-saving method to describe these ROIs and followed by WTA and inhibition of return. Threshold segmentation maintains a stable performance for different images. Region growing is the most time-consuming method and the calculation time varies according to different images.

Table 1. Comparisons of region of interest description calculation time (s)

Images	RG	WTA & IR	TS
BJ-1	16.7403	0.2913	0.0024
BJ-2	12.8532	0.2911	0.0024
BJ-3	12.7796	0.2937	0.0024
BJ-4	9.1253	0.2102	0.0024
BJ-5	3.9229	0.2020	0.0023
BJ-6	16.7688	0.3471	0.0024
BJ-7	13.5894	0.2989	0.0024

Table 2 shows the comparison of calculation time across Itti’s model and the VA-TS model. It can be seen that the VA-TS model is much more time-saving than Itti’s model. The average calculation time of the VA-TS model is only about 5 seconds and the model maintains a stable performance with different images. On the contrary, it takes Itti’s model hundreds of seconds to do the same thing, and the calculation time varies significantly with different images.

Table 2. Comparisons of overall calculation time across Itti’s model and the VA-TS model (s)

Images	Itti’s model	VA-TS model
BJ-1	781.7331	5.1553
BJ-2	912.1392	5.1373
BJ-3	624.6888	5.1609
BJ-4	614.6462	5.1640
BJ-5	582.7580	5.1779
BJ-6	721.3192	5.1186
BJ-7	707.0790	5.1398

In addition, we use the Receiver Operator Characteristic (ROC) curve to compare the performance across the four models quantitatively in **Fig. 10**. 20 randomly selected image fragments with dimension of 2048×2048 are used as the image database. For each image, a manually segmented map is generated as the background truth. The ROC curves are generated by classifying the locations in a saliency map into salient regions and non-salient regions with varying quantization thresholds. It can be seen that our model has the best performance across all the four models.

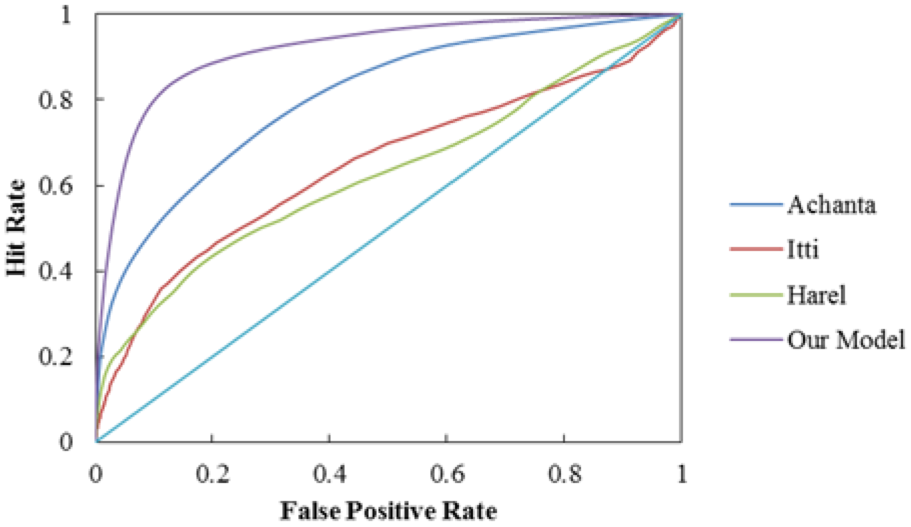


Fig. 9. ROC curves of VA-TS and other three existing models

5. Conclusion

In this paper, a novel model for rapid region of interest detection based on visual attention and threshold segmentation in high spatial resolution remote sensing images, namely VA-TS model, is proposed.

High spatial resolution remote sensing images contain large amount of data, which brings difficulty to fast processing. To tackle this issue, in the pre-process step, the input image is adaptively subsampled using Gaussian Pyramids to a smaller dimension to decrease the data for further processing. To utilize texture information, the discrete moment transform (DMT) is used in the feature extraction step. In computing the saliency map, a feature competition strategy based on the number of “salient points” and “strong points” is proposed. Threshold segmentation is applied to the saliency map generate a binarized mask to present the detected region of interest, which is much more accurate and brings little redundancy. The experimental results have shown that the VA-TS model produces accurate detection results and the calculation time increases slowly when the resolution of a images increases, meaning that the VA-TS model is more efficient when applied to high spatial resolution remote sensing images. The calculation time is less than 1% of Itti’s model.

Generally speaking, the VA-TS model can resolve to some extent the problem of computation efficiency in high spatial resolution remote sensing image processing. The detection results are visually and statistically satisfying. Though the overall computational complexity is a little higher than Harel's methods, it is much faster than Achanta's method when the resolution increases and show great potential in processing extraordinarily high spatial resolution remote sensing images. Nevertheless, further research is still needed to obtain a better performance including an improved way of computing the saliency maps and a better way of resolving the blurred edges.

References

[1] D. Dai, and W. Yang, "Satellite image classification via two-layer sparse coding with biased image representation," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 1, pp. 173-176, 2011. Article (CrossRef Link)

[2] A. Sedaghat, M. Mokhtarzade, and H. Ebadi, "Uniform robust scale-invariant feature matching for optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 11, pp. 4516-4527, 2011. Article (CrossRef Link)

[3] Z.-g. Liu, J. Dezert, G. Mercier et al., "Dynamic Evidential Reasoning for Change Detection in Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 5, pp. 1955-1967, 2012. Article (CrossRef Link)

[4] Y. Lina, Z. Guifeng, and W. Zhaocong, "A Scale-Synthesis Method for High Spatial Resolution Remote Sensing Image Segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 10, pp. 4062-4070, 2012. Article (CrossRef Link)

[5] P. Liang, T. Yang, "The Neural Network Classification of Remote Sensing Image Supplemented by Texture Characteristic," *Geomantic & Spatial Information Technology,* vol. 31, no. 4, pp. 66-67, 2008. Article (CrossRef Link)

[6] X. Zheng, W. Chen, and B. Cui, "Multi-Gradient Surface Electromyography (SEMG) Movement Feature Recognition Based on Wavelet Packet Analysis and Support Vector Machine (SVM)," in *Proc. of 2011 5th International Conference on Bioinformatics and Biomedical Engineering*, Wuhan, China, pp. 1-4, May, 2011. Article (CrossRef Link)

[7] J. Fan, H. Min, and W. Jun. "Single Point Iterative Weighted Fuzzy C-means Clustering Algorithm for Remote Sensing Threshold segmentation," *Pattern Recognition,* vol. 42, no. 11, pp. 2527-2540, 2009. Article (CrossRef Link)

[8] O. Rozenstein, and A. Karnieli, "Comparison of methods for land-use classification incorporating remote sensing and GIS inputs," *Applied Geography*, vol. 31, no. 2, pp. 533-544, 2011. Article (CrossRef Link)

[9] T. Lei, S. Wan, and T. Chou, "The comparison of PCA and discrete rough set for feature extraction of remote sensing image classification–A case study on rice classification, Taiwan," *Computational Geosciences*, vol. 12, no. 1, pp. 1-14, 2008. Article (CrossRef Link)

[10] Zhang Guomin. Researches on Object Detection in Remote Sensing Image with Complicated Scenes, PHD Thesis, National University of Defense Technology, Changsha, 2010. Article (CrossRef Link)

[11] Sun Ning. Research on Target Recognition Methods for Building Detection in High Spatial Resolution Remote Sensing Images, PHD Thesis, Zhejiang University, Hangzhou, 2010. Article (CrossRef Link)

[12] J. M. Wolfe, and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?," *Nature Reviews Neuroscience*, vol. 5, no. 6, pp. 495-501, 2004. Article (CrossRef Link)

[13] 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, no. 11, pp. 1254-1259, 1998. Article (CrossRef Link)

[14] L. Itti, and C. Koch, "Computational modelling of visual attention," *Nature reviews neuroscience,* vol. 2, no. 3, pp. 194-203, 2001. Article (CrossRef Link)

[15] Achanta R, Hemami S, Estrada F, and Susstrunk S, "Frequency-tuned Salient Region Detection," in *Proc. of* IEEE *International Conference on Computer Vision and Pattern Recognition (CVPR)*,

pp. 1597-1604, June 20-15, 2009. Article (CrossRef Link)

[16] Harel J, Koch C, Perona P, "Graph-based Visual Saliency," *Advances in Neural Information Processing Systems,* vol. 19, pp. 545-552, 2009. Article (CrossRef Link)

[17] Milind S. Gide, Lina J. Karam, "Comparative Evaluation of Visual Saliency Models for Quality Assessment Task," in *Proc. of International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, Jan, 2012. Article (CrossRef Link)

[18] Pal R, Mitra P, Mukhopadhyay J, "Suitable features for visual saliency computation in monochrome images," in *Proc. of 4th International Congress on Image and Signal Processing* (CISP), pp. 1457-1460, Oct 15-17, 2011. Article (CrossRef Link)

[19] J. K. Tsotsos, S. M. Culhane, W. Y. Kei Wai et al., "Modeling visual attention via selective tuning," *Artificial intelligence*, vol. 78, no. 1, pp. 507-545, 1995. Article (CrossRef Link)

[20] Koch C, Ullman S, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," *Human Neurobiology*, vol. 4, pp. 219–227, 1985. Article (CrossRef Link)

[21] M. I. Posner, and Y. Cohen, "Components of visual orienting," *Attention and performance X: Control of language processes,* vol. 32, pp. 531-556, 1984. Article (CrossRef Link)

[22] Ma Y. F, Zhang H. J, "Contrast-Based Image Attention Analysis by Using Fuzzy Growing, Proceedings of the Eleventh," in *Proc. of ACM International Conference on Multimedia*, pp. 374–381, Jan, 2003. Article (CrossRef Link)

[23] S. Engel, X. Zhang, and B. Wandell, "Colour tuning in human visual cortex measured with functional magnetic resonance imaging," *Nature*, vol. 388, no. 6637, pp. 68-71, 1997. Article (CrossRef Link)

[24] Greenspan H, Belongie S, Goodman R, Perona P, Rakshit S, and Anderson C. H, "Overcomplete Steerable Pyramid Filters and Rotation Invariance," *IEEE Computer Vision and Pattern Recognition*, pp. 222-228, 1994. Article (CrossRef Link)

[25] S. O. Belkasim, M. Shridhar, and M. Ahmadi, "Pattern recognition with moment invariants: a comparative study and new results," *Pattern recognition,* vol. 24, no. 12, pp. 1117-1138, 1991. Article (CrossRef Link)

[26] L. Itti, and C. Koch, "A comparison of feature combination strategies for saliency-based visual attention systems," *SPIE human vision and electronic imaging IV (HVEI'99),* vol. 3644, pp. 373-382, 1999. Article (CrossRef Link)

[27] N. Otsu, "A Threshold Selection Method from Gray-Level Histogram," *IEEE Trans on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62-66, 1979. Article (CrossRef Link)

Libao Zhang is an associate professor in College of Information Science and Technology, Beijing Normal University, China. He received his B.S. degree from Jilin University, China, in 1999, his M.S. degree from College of Communication Engineering, Jilin University, China, in 2002, and his Ph.D. form College of Communication Engineering, Jilin University, China, in 2005. He has taken charge of two National Science Foundations of China. Now, his interested research fields include image compression, object recognition, and wavelet transform.

Hao Li received the B.S. degree in automation from Beijing Normal University, Beijing, China, in 2010. He is currently pursuing the M.S. degree at Beijing Normal University, Beijing, China. His primary research interest lies in modeling biologically-plausible computational visual attention and object detection/recognition.