Target Evaluation of Remote Sensing Image Based on Scene Context Guidance

Wenjuan Li[1(✉)], Shunan Shang[2], and Ling Tong[1]

[1] Beijing Institute of Spacecraft System Engineering, Beijing 10094, China
wjolee@l26.com
[2] Institute of Telecommunication Satellite, CAST, Beijing 10094, China

Abstract. The correlation between scenes and targets in remote sensing images can provide useful and important information and guidance for satellite to achieve onboard targets evaluation in order to find valuable targets to image. The relationship between the target and the scene, as well as the spatial location association it contains, determines what the system should "focus on" and "what areas to focus on" in different scenarios. Referring to the guiding role of context information in the visual system, this paper studies how to identify potential targets through the scene context information, and a saliency model based on the task context information to achieve the target evaluation under different scenarios is proposed. At the end of the paper, a simulation experiment is given. It can be seen from the experiment that through scene context guidance, different parameters can be loaded in different scenarios to realize the evaluation and discrimination of different targets.

Keywords: Context · Saliency model · Target detection

1 Preface

In the case of complex targets and backgrounds, the human visual system can still quickly identify and classify a large number of targets, and has very good adaptability to the illumination, attitude, texture, deformation and occlusion of the target imaging. In addition, the human brain has powerful learning and reasoning ability to identify targets which have never been seen by observing a set of targets. This is because when humans recognize objects in the real world, other surrounding objects and specific environments provide a rich contextual connection to the visual system. Targets appear in consistent or common scenarios, making detection and recognition tasks more accurate and faster than appearing in uncoordinated scenarios. The broad connection between the target and its environment is called context information. Studies in the human visual system [1–8], cognitive neuroscience [9–13], and computer vision [14–16] have shown that context information plays an important role in the target classification of the human brain. In fact, the human brain using context information during feature analysis. Inspired by the recognition goals of human cognitive system, many scholars imitate the human visual system to improve the performance of computer recognition systems. Some new methods consider the context information of the target,

M. Jia et al. (Eds.): WiSATS 2019, LNICST 280, pp. 425–436, 2019.
https://doi.org/10.1007/978-3-030-19153-5_44

and introduce the scene information and the mutual constraints between the targets into the target classification task to improve the performance and eliminating the uncertainty.

2 Context Feature

2.1 Definition

Cognitive psychology believes that perception and human knowledge and experience are inseparable. The impact of experience knowledge on vision is multifaceted, and the most striking one is the role of context information. The human visual process can be viewed as being guided and planned under context information. Thus, context information is an important reminder when humans are performing target detection and recognition. Context information can indicate what is worth noting and what is negligible, which greatly reduces the processing burden of the visual system and reduces processing time. Biederman et al. have pointed out that when humans detect targets, the prompts that violate the context information not only increase the processing time, but also make them more error prone. The MRI results also confirmed the use of context information when human brain is detecting and identifying man-made targets. If the human body is the evolution of the visual upper and lower information, then context information is an effective way to understand the visual world. Thus, it can be inferred that context information is also beneficial to machine cognitive systems.

In practical applications, the scene configuration in which the target is located or the structure inside the target is often highly structured. The context information of the target can be defined as information contained in the scene or the entire scene information for detecting and identifying the target. In a general sense, the context is about the surrounding environment in which the objects are located. In a natural image, there is a strong specific relationship between the target and the scene, and the context is to describe it. Generally, context information can be divided into the following three categories.

(1) Local context. The image to be detected contains many local regions, each of them have a relationship with their surroundings. The information describing the relationship is called a local context feature. It includes the neighborhood context of the local region and the geometric relationships between the local regions.
(2) Target context. The target to be detected has a certain relationship with the surrounding targets, including whether these targets appear, and the location and scale relationship between them. This information is called the target context.
(3) Scene context. The targets are in a certain scene, and the scene in which the targets are located is very helpful for the detection and recognition of the target. This scene information is called the scene context.

2.2 Detection and Recognition Methods Based on Context Features

Context in computer vision can be defined as all information related to the target but not the apparent description of the target itself. To some extent, this reveals to some extent

the "connotation" of the target. Computer vision uses context information from three different levels, namely: local context feature layer, target context target layer, and scene context scene layer. The local context includes a location relationship between the context based on the neighborhood and the local region of the geometric context.

(1) Local context

At present, the mainstream target detection and recognition method is based on local features. It is simple, easy to calculate, insensitive to affine transformation, and also has certain resistance to occlusion, illumination and intra-class changes. The Bag of Features (BoF) model is a very hot model in recent years, and the target detection and recognition effect is very good. However, the model does not consider the positional relationship between local features, which is very important, because the target is a whole, and the local regions which constitute the target are not unrelated, but organized according to certain rules. Recently, some researchers have proposed a method to add rough spatial positional relations to a local feature model, such as embedding spatial information into the BoF model. This method needs to balance discriminative ability and generalization ability, which needs to be considered from two aspects of feature quantization and spatial feature extraction.

In addition, the local context also includes the neighborhood context of the local area. The neighborhood context information is mainly used in image labeling. When labeling an area, the information of the surrounding area is considered, and the labeling of the area is constrained to improve the accuracy of labeling. Even in unsupervised cases, tags based on neighborhood contexts also have good results. The models describing this neighborhood context constraint are mainly random fields, including Markov Random Field (MRF), Discriminative Random Fields (DRF), multi-scale conditional random fields (mCRF) [17].

(2) Target context

The target context refers to the co-occurrence relationship and location relationship between the target and other objects in the scene. There is no doubt that other objects in the scene can be very helpful in detecting and identifying the target object. There are many target recognition methods based on target context in computer vision. However, based on the target context method, a condition is also required, that is, the identification of other objects is relatively accurate. If you use some very unreliable information to infer, the results will not be very good. In order to solve this problem, the commonly used target context-based detection and recognition method is implemented by iteration, using the most reliable target recognition result to infer other targets, and then repeating the process until convergence. For example, Finks proposes a method to detect local features of targets and other targets using cascaded target detection structures [18]. All the interdependencies are calculated in an iterative manner. Torralba et al. also proposed a similar framework that using boosting and graph networks to learn the possibility of co-occurrence and related locations of targets [19]. In addition, the method must have multi-target information in the training. If the object to be identified is the only target in the database, then the target context information cannot be learned, and the target context cannot be used [20].

(3) Scene context
In the research of computer vision, the most commonly used method of scene classification is target based method. There are some fixed and obvious signs in some scenes. The object-based scene classification method identifies the scene by recognizing these marks. But these methods also include some intermediate steps, such as segmentation, feature organization and target recognition, which are also the key problems in computer vision. On the other hand, the human vision system performs very well in scene classification, far surpassing the computer vision system. Human beings start from the whole in scene recognition, because of this, there are some recent research methods analyzing the whole scene to achieve the classification tasks and the results are very good. There are two main methods based on scene context features: the first is to add rough context relations to the model based on local features. For example, Lazebnik et al. put forward a Beyond Bag of Features (BBoF) model, which uses spatial pyramid structure to extract the spatial relationship of local features. And then it is added to the BoF model, referred to as the "BBoFs" model. The second is based on the spatial distribution of multi-scale and multi-directional filters, which is called "gist" model. Extracting more robust and semantically explicit scene context features is very important for scene classification. But since the first model only aims at local features and the second model is based on filter features, it is necessary to extend the two models so that they can be applied to both features. Moreover, the combination of multiple features can make the scene representation more robust. Recent studies have shown that low dimensional statistics of low-level features can be used as a description of the context of the scene. As it contains certain semantic information, which can be used to predict the target in the scene, including the probability, location and scales of the target etc. Many traditional methods are based on local features, but high-level semantic descriptions are more stable, so the semantic description of adding context can fill the gap between low-level features and high-level semantics.
In this paper the context information is added to the saliency model trying to imitate the prior knowledge guidance in human cognitive system, which will be used to load different algorithms to realize onboard target evaluation in difference satellite scenarios such as the sea, deserts, etc.

3 Scene Context Guidance Based Target Evaluation

3.1 Scene-Target Context

The context relationship between the targets in the scene of the remote sensing image includes the relative position, scale size and panoramic spatial relationship of the target and the surrounding environment. A thorough understanding of these relationships can increase or decrease the probability of different targets appearing in different scenarios and can be used to guide the system in selecting the appropriate target evaluation method. For example, the possibility of an airport around a city increases; the possibility of ships and islands in the ocean is high.

For Earth observation, there is a mapping relationship between the flight orbit of the remote sensing satellite and the geographical location of its imaging area: the observation field can be determined according to the width of the imaging field, flight orbit and flight time. Figure 1 shows the relationship between the strip number, line number and administrative division, latitude and longitude of the Landsat satellite using the global reference system WRS2 in China. From this mapping relationship, scene semantic information of the current observation area of the satellite can be obtained, such as mountains, cities, oceans, ports, and so on.

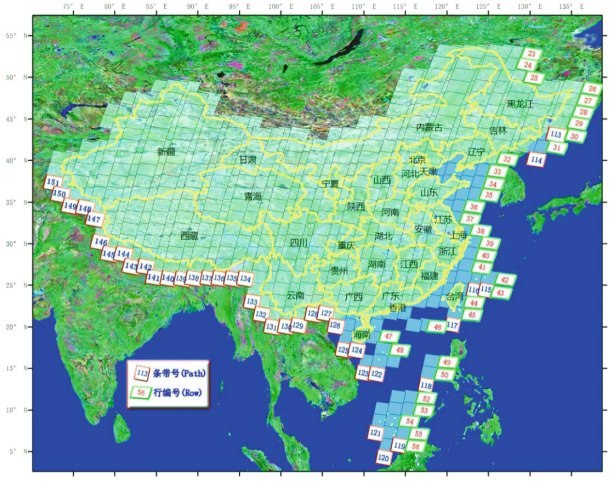

Fig. 1. The WRS2 of China for Landsat imaging [figure is obtained from internet]

For the realization of the onboard target evaluation, the scene semantic information not only contains the scene feature, but also the information of the potential valuable targets. The scene feature gives a characterization of the scene, and the potential valuable targets information can determine the observation targets, thereby guiding the system to "where/what to pay attention to" in the scene. The scene semantic information also reveals the relationship between the target and the scene. For example, the airport is usually built on the suburbs of the city, so it can be inferred that the probability of the aircraft appearing in that scene is high. It can be seen that the analysis of the scene semantics not only provides the general information and features of the scene, but also determines the probability of potential targets appearing in the scene.

3.2 Model

In order to realize the onboard target evaluation, it is necessary to analyze the scene semantics, based on the parsed information, and select appropriate parameters of the payload to determine the salient features of the potential targets and other corresponding parameters, such as the size range, length-width ratio of the target etc. According to the analyzed scene information and potential targets, the appropriate

saliency detection method is selected to realize the targets evaluation, and the valuable targets and their regions are determined, which can guide the satellite to select image and provide basis for further detailed observation of other satellites. In this paper, a framework for target evaluation based on scene information is proposed. In order to verify the validity of the model, several simple application scenarios are designed. By inputting scene semantic information, the model can independently select the corresponding algorithm to complete the target evaluation and detection of the scenario. The more detailed scene division, the more precise the constraint relationship between the target and the scene, and the more accurate the result of the target evaluation and detection can be.

According to the types of terrain, it can be simply divided into three categories: sea area, sea-land junction and lands, of which the lands can also be subdivided. From the perspective of remote sensing images target evaluation application can be divided into: man-made targets detection, natural disasters and scientific phenomena detection, and hot spot change detection. Considering the scene and the application point of view, the scene and the targets or events that may need to be observed are divided as follows, and are listed in Table 1.

(1) Sea: ships, islands, air planes.
(2) Sea-land junction: vessels, nuclear power plants, airports.
(3) According to characteristics of land features, lands can be roughly divided into: mountains, deserts, plains, fields, cities, woodlands, rivers, lakes, and glaciers. And the targets and events that may need to be observed in these different surface types are divided as follows:
    (a) Desert and Gobi: oasis, man-made buildings, winds power stations;
    (b) Mountain: man-made buildings, volcanoes, wind power stations;
    (c) Plain land: man-made buildings;
    (d) Rivers: dams and bridges, ships, floods;
    (e) Fields: man-made buildings;
    (f) Cities: special buildings or designated buildings, airports;
    (g) Woodlands: man-made buildings, fires.

Table 1. The example of the potential targets in different scenes

Scene		Potential valuable targets
Sea-land junction	Sea	Ship, island, air plane
	Land	Port, vessels, nuclear power plants, airport
Lands	Desert, Gobi	Oasis, man-made buildings, winds power stations, roads
	Mountain	Man-made buildings, volcanoes, wind power stations
	Plain land	Man-made buildings, airports
	Rivers	Dams, bridges, ships
	Fields	Man-made buildings, waters
	Cities	Special buildings or designated buildings, airports, waters
	Woodlands	Man-made buildings, waters

Applying saliency detection features combined with the image data and its potential targets, this paper proposes a saliency model for targets evaluation to achieve valuable targets detection guided by scene semantic information. Firstly, in the process of top-down attention mechanism, the potential targets are derived from the scene context information of the scene semantic analysis, such as searching for ships, man-made buildings, etc. in the remote sensing image. Meanwhile, in the process of bottom-up visual simulation, the appropriate scale and method are selected according to the analyzed scene information, and the feature maps of the image are calculated according to the selected method, and the saliency detection of the valuable targets are realized according to the saliency features of the scene at last. In this process, the scene context information introduces "attention" into the region of interest driven by the task, and combines the top-down information with the data-driven saliency map to achieve the valuable target detection without human involved. The model is shown in Fig. 2.

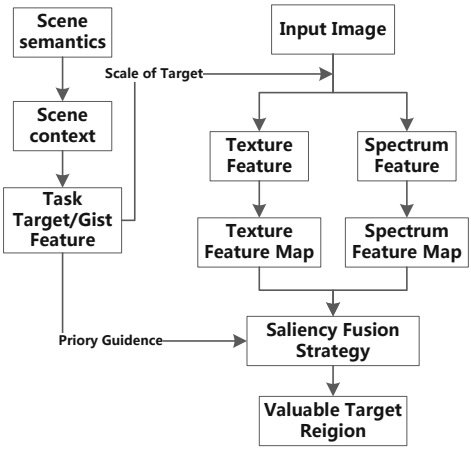

Fig. 2. The schematic diagram of the target evaluation model

As can be seen from Fig. 2, the scene context information provides task target information and gist features that guide the extraction of the scene, which provide prior knowledge as attention guidance information. Under the guidance of the target information, it is first necessary to extract the features of the image, and provide the basis for distinguishing the significant target and the background in different scenarios. The extracted features are mainly as follows:

(1) Texture features. As a form of perception representation of an object in the human visual system, the image texture features refer to the gray-scale or color change of the pixel points in the image and the regular pattern of variation distribution. The texture features reflect the surface roughness, certain regularity and directionality of the object. Different objects have different texture features. It is one of the causes of human visual difference and an important basis for distinguishing

different objects. Texture is also important information of remote sensing images. It not only reflects the brightness statistics of the image, but also the relationship between the structural features and the spatial arrangement of the object itself. The texture features are composed of local texture information and global texture information, wherein the local texture feature is represented by the gray level or color distribution of the pixel and its surrounding spatial neighborhood; and the global texture feature is repeated by different degrees of the local texture. Therefore, image texture features can be used to characterize different scenes contained in the image, such as deserts, grasslands and so on. Their extraction methods are divided into various types such as gray level co-occurrence matrix, LBP, etc. And the extracted texture features reflect the characteristics of objects or features from different angles, and the feature maps $S_{texture}$ are obtained.

(2) Spectrum features. Fourier transform is a typical spectral feature extraction method. The spectrum of different frequency bands corresponds to different features in the image. Therefore, the identification between target and background can also be realized by spectrum analysis. Similar to the extraction of texture features, the different scale spectrum features of the image are extracted according to the spectral distribution quantization method, and the spectrum feature map $S_{spetral}$ of the image is calculated.

(3) Fractal features. Fractal features are one of the important features that distinguish man-made targets from natural backgrounds. A method for calculating multi-scale fractal dimension features is proposed in [21]. The difference of multi-scale fractal dimension between natural background and man-made is obvious, which can distinguish natural background and man-made objects, and beneficial to the detection of man-made objects.

Thus, three feature maps $S_{texture}$, S_{frac}, and $S_{spetral}$ are obtained. Finally, the saliency map of the image is calculated according to the task target information and Eq. 1, and the salient object is the high-value target in the scene. Where, the exponents α_1, α_2 and α_3 are weighting factors, and satisfy $\alpha_1 + \alpha_2 + \alpha_3 = 1$.

$$SM = S_{texture}^{\alpha_1} \cdot S_{frac}^{\alpha_2} \cdot S_{spetral}^{\alpha_3} \tag{1}$$

Table 1 lists potential targets in different scenarios of remote sensing images. As mentioned above, different targets constitute a scene, and a scene contains various targets. It is assumed that scene A contains target C. And if scene X is determined to be the scene A, the scene-target context infers that the probability of occurrence of target C is relatively high. The probability of the same target appears in different scenes varies. For example, the probability of a road appearing in a city or a desert is different, and can be expressed as p_1 and p_2 respectively. Therefore, the probability can be used to represent the relationship between the target and the scene. Set S to be the scene, T is the target, $P(T = T_j | S = S_i)$ indicating the probability of occurrence of the target T_j under the scene S_i. The value of the probability determines the value of the weighting factor, so that the most effective target evaluation method in the scene can be selected according to the scene semantics. And the scene context information also provides the gist feature of background for the algorithm as a reference, and the size of the target to

select the appropriate scale for calculation. Take the sea scenario for example, the multi-scale fractal dimension method has a large weighting factor due to the high probability of the ship appearing, and then the multi-scale fractal dimension method is selected to realize the ship detection. In the mountain background, the detection of man-made buildings can also be achieved by multi-scale fractal dimension. However, due to the different dimensions of the building and the ship target and the different background gist features, it is necessary to change the settings of the parameters such as the scale, weighting factors to realize the detection of man-made buildings in the mountain background. Thus, the corresponding effective methods and parameters are selected according to different scenarios and potential targets to realize the target evaluation under the scenario.

4 Experiment

In order to illustrate the target evaluation in different scenarios under the framework of this model, this paper builds an experimental environment using MATLAB, in which the image to be processed and the scene type of the image can be selected, as shown in Fig. 3. Each type of scene corresponds to the known and stetted context information of the scene, including: size range of potential valuable targets, the extracted features, gist features of the scene, the probability of the potential targets. And the probability of the potential valuable targets determines the weighting factors, which are used to select the effective features of the targets in the scene. As the scene classification is refined, the information of the scene will be more and more detailed, and the parameters for guiding, such as weighting factors, will be more and more precise, and the result of the target evaluation will become more and more accurate.

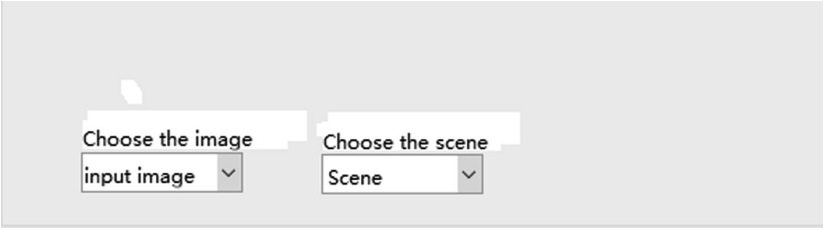

Fig. 3. The Menu of image and scene selection

By analyzing the semantic information of the regional scene, and comparing the context information between the scene and the target, it is possible to select appropriate features and target evaluation method for the region. Further, the scene estimation of the future observation area may be performed in advance according to the flight orbit of the satellite, and an appropriate method is selected according to the method to achieve the target evaluation. Defining the scene semantics of the image into a format as $\{Scene, Target, Scale, Feature, Summary\ Information\}$. The 'target' represents the potential targets in the scene; some features of the target are mostly presented as local

saliency, and the 'scale' is based on potential targets to determine the size of the target area in order to highlight the targets, as well as determining the scale range of different size targets in a scene to achieve multi-target multi-scale target evaluation; 'features' is to select the effective features of the target in the scene; 'gist information' is a feature representing the scene, as a reference of the potential targets saliency. The parsing of the scene semantic information and the pseudo code selected by the target evaluation algorithm are shown in Fig. 4.

```
IF Scene = S1 THEN Target = {T1, T2}; end
IF Scene = S1 and Target = T1
THEN
    scale = wi;
    feature = F{f1, f2, f3};
    gist = {fgist1, fgist2, fgist3};
    P(S1|T1) = p1
    IF P(S1|T1)= p1 THEN weight = {α1, α2, α3}; end
end
IF Scene = S1 and Target = T2
THEN
    scale = wj;
    feature = F{f1, f3};
    gist = {fgist1, fgist3};
    P(S1|T2)=p2
    IF P(S1|T2)=p2 THEN weight = {β1, β2, β3}; end
end
IF Scene = S1and Target = {T1, T2}
THEN
    Method = Model1;
end
```

Fig. 4. Pseudo code of the algorithm

From the platform built by MATLAB, semantic representation of the scene can be selected, such as sea and Gobi. And each scene semantic corresponds to scene context information contains features, setting parameters of targets and the scene, which can guide the selection of appropriate methods to complete the target evaluation under the scene. Figure 5 shows the result of the target evaluation of the image according to the selected scene. The simulation experiment selects the scene of the image by manual selection. While onboard the semantic description of the scene needs to be obtained according to the mapping relationship between the actual observed field and the geographical location. Each scene corresponds to an algorithm configuration file, and the corresponding parameters of the algorithm in the configuration file can be improved by experiments to make the system continuously self-learning and correcting.

Fig. 5. The processing result of an image with sea background

5 Conclusion

The scene context information in the remote sensing image provides useful and important information and guidance for the satellite to achieve onboard target evaluation. The guiding role of context information in the visual system is taken as the entry point, and the potential target in different scenes through the scene context information has been studied. Thus a novel integrated saliency model has been proposed, in which the analyzing the parameters corresponding to the scene context information and the target feature information has been introduced. In the model, the guiding system selects the appropriate method to achieve the target evaluation in different scenarios. In order to verify its validity, the experiment was carried out by using MATLAB to set up the experimental environment. Selecting the preset scene to call its configuration file, and read the information in it to realize the target evaluation under the certain scene. And the scene context guided valuable target detection is simulated and realized.

Acknowledgment. The project was supported is the independent research and development project in China Academy of Space Technology.

References

1. Auckland, M.E., et al.: Non-target objects can influence perceptual processes during object recognition. Psychon. Bull. Rev. **14**, 332–337 (2007)
2. Biederman, I., et al.: Scene perception: detecting and judging objects undergoing relational violations. Cognit. Psychol. **14**, 143–177 (1982)
3. Davenport, J.L., Potter, M.C.: Scene consistency in objects and background perception. Psychol. Sci. **15**, 559–564 (2004)
4. Friedman, A.: Framing pictures: the role of knowledge in automatized encoding and memory of gist. J. Exp. Psychol. Gen. **108**, 316–355 (1979)
5. Gordon, R.D.: Attentional allocation during the perception of scenes. J. Exp. Psychol. Hum. Percept. Perform. **30**, 760–777 (2004)

6. Henderson, J.M., et al.: Effects of semantic consistency on eye movements during scene viewing. J. Exp. Psychol. Hum. Percept. Perform. **25**, 210–228 (1999)
7. Palmer, S.E.: The effects of contextual scenes on the identification of objects. Mem. Cognit. **3**, 519–526 (1975)
8. Hollingworth, A., Henderson, J.M.: Does consistent scene context facilitate object detection. J. Exp. Psychol. Gen. **127**, 398–415 (1998)
9. Bar, M.: Visual objects in context. Nat. Rev. Neurosci. **5**, 617–629 (2004)
10. Bar, M., Aminoff, E.: Cortical analysis of visual context. Neuron **38**, 347–358 (2003)
11. Goh, J., et al.: Cortical areas involved in object, background, and object-background processing revealed with functional magnetic resonance adaptation. J. Neurosci. **24**, 10223–10228 (2004)
12. Aminoff, E., et al.: The parahippocampal cortex mediates spatial and non-spatial associations. Cereb. Cortex **27**, 1493–1503 (2007)
13. Gronau, N., Neta, M., Bar, M.: Integrated contextual representation for objects' identities and their locations. J. Cogn. Neurosci. **20**(3), 371–388 (2008)
14. Murray, N., Vanrell, M., Otazu, X., et al.: Saliency estimation using a non-parametric low-level vision model. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 433–440 (2011)
15. Torralba, A., et al.: Contextual guidance of attention in natural scenes: the role of global features on object search. Psychol. Rev. **113**, 766–786 (2006)
16. Hoiem, D., et al.: Putting objects in perspective. Proc. IEEE Comp. Vis. Pattern Recog. **2**, 2137–2144 (2006)
17. He, X., Zemel, R.S., Carreira-Perpiñán, M.Á.: Multiscale conditional random fields for image labeling. In: CVPR 2004, vol. 2, pp. II-695–II-702 (2004)
18. Viola, P., Jones, M.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
19. Torralba, A., Murphy, K.P., Freeman, W.T.: Contextual models for object detection using boosted random fields. In: Nips, pp. 1401–1408 (2004)
20. Wolf, L., Bileschi, S.: A critical view of context. Int. J. Comput. Vis. **69**(2), 251–261 (2006)
21. Li, W.J., Zhao, H.P., Guo, J., et al.: A multi-scale fractal dimension based onboard ship saliency detection algorithm. In: 2016 IEEE 13th International Conference on Signal Processing (ICSP), pp. 628–633 (2016)