Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding

Junwei Han [*], Peicheng Zhou, Dingwen Zhang, Gong Cheng, Lei Guo, Zhenbao Liu, Shuhui Bu, Jun Wu

Department of Control and Information, School of Automation, Northwestern Polytechnical University, 127 Youyi Xilu, Xi'an 710072, PR China

ARTICLE INFO

Article history:
Received 26 June 2013
Received in revised form 11 November 2013
Accepted 26 December 2013
Available online 29 January 2014

Keywords:
Geospatial target detection
Visual saliency
Discriminative sparse coding

ABSTRACT

Automatic detection of geospatial targets in cluttered scenes is a profound challenge in the field of aerial and satellite image analysis. In this paper, we propose a novel practical framework enabling efficient and simultaneous detection of multi-class geospatial targets in remote sensing images (RSI) by the integration of visual saliency modeling and the discriminative learning of sparse coding. At first, a computational saliency prediction model is built via learning a direct mapping from a variety of visual features to a ground truth set of salient objects in geospatial images manually annotated by experts. The output of this model can predict a small set of target candidate areas. Afterwards, in contrast with typical models that are trained independently for each class of targets, we train a multi-class object detector that can simultaneously localize multiple targets from multiple classes by using discriminative sparse coding. The Fisher discrimination criterion is incorporated into the learning of a dictionary, which leads to a set of discriminative sparse coding coefficients having small within-class scatter and big between-class scatter. Multi-class classification can be therefore achieved by the reconstruction error and discriminative coding coefficients. Finally, the trained multi-class object detector is applied to those target candidate areas instead of the entire image in order to classify them into various categories of target, which can significantly reduce the cost of traditional exhaustive search. Comprehensive evaluations on a satellite RSI database and comparisons with a number of state-of-the-art approaches demonstrate the effectiveness and efficiency of the proposed work.

1. Introduction

The advance of remote sensing technology has led to the explosive growth of geospatial images in quantity and quality. Especially, the present high-resolution geospatial images contain rich visual information to describe the surface appearance of the earth. It has necessitated study into automatic analysis and understanding of visual content to acquire useful information. A fundamental aspect of this research is target detection, which can be specifically defined as the problem of "Does an image window at a certain location contain a given target?"

Target detection in geospatial images has been extensively studied in recent years. Many approaches cast target detection as a classification problem. A set of features that can characterize the targets is extracted first, and then classification is performed using the extracted features and predefined classifiers. The establishment of the classifiers can be based on template matching (Akçay and Aksoy, 2008; Li et al., 2010b, 2012; Sirmaçek and Ünsalan, 2009, 2010) or direct application of a variety of sophisticated machine learning methods such as Support Vector Machines (SVM) (Bi et al., 2012; Cheng et al., 2013; Li et al., 2010a; Li and Itti, 2011; Sun et al., 2012; Tao et al., 2011), Kernel-Based Density Estimation (Sirmaçek and Ünsalan, 2011), Gaussian Mixture Models (Bhagavathy and Manjunath, 2006), Hough Forests (Lei et al., 2012), Latent Dirichlet Allocation models (Sun et al., 2010b), and Support Tensor Machines (Zhang et al., 2011). According to the feature types used in the models, geospatial target detectors can be briefly categorized into two broad groups: frequency/transformation domain feature-based and spatial domain feature-based. Li et al. (2010b) adopted the ridgelet transform to detect straight road edges. Tello et al. (2005) proposed to detect ships in synthetic aperture radar (SAR) images using the discrete wavelet transform. Bhagavathy and Manjunath (2006) developed a system to detect the presence of geospatial objects such as harbors, golf courses, and parking lots by means of detecting spatially recurrent patterns that are characterized by the Gabor filters. Sirmaçek and Ünsalan (2010) and Li et al. (2010a) also leveraged Gabor filters to obtain local feature points to detect the urban areas and buildings. Zhang et al. (2011) incorporated the spatial relationship of neighboring pixels into the textural feature modeled by Gabor functions to detect targets. Another group of approaches extracted features in the

[*] Corresponding author. Tel./fax: +86 29 88431318.
E-mail address: junweihan2010@gmail.com (J. Han).

0924-2716/$ - see front matter © 2014 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS) Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.isprsjprs.2013.12.011

spatial domain because spatial analysis is adequate to describe the structure and layout of geospatial targets. Some other approaches extracted spatial features based on segmented regions. For example, Akçay and Aksoy (2008) proposed a new method to detect geospatial objects using multiple hierarchical segmentations by combining spectral information with structure information. The works of Bi et al. (2012), Li et al. (2012) and Sun et al. (2010b) utilized the spatial contextual information of segmented regions to achieve promising detection results. Recently, local invariant features with the powerful ability of characterizing target structural features have been successfully applied to geospatial object detection. To be specific, Tao et al. (2011) described airports by a set of scale-invariant feature transform (SIFT) keypoints. Xu et al. (2010) applied the visual Bag-of-Words (BoW) model to describe and classify geospatial objects. Sun et al. (2012) built the Spatial Bag-of-Words model by incorporating the geometric information of the targets into the clustering of SIFT points to represent targets. The detection was achieved based on the SVM training.

Although the topic of geospatial target detection has been deeply investigated, less concern has been given to two critical problems. The first problem is the multi-class target detection. Generally, a geospatial image contains multiple classes of interesting targets instead of only a single one. Currently, most of target detection methods still utilize individual class detectors, for example, a detector only for airplanes, another only for ships, etc. The multi-class target detection means to directly, simply, and subsequently apply those independent detectors. However, this approach may not be successful due to the following two reasons. First, the features extracted in many existing individual detectors are customized for the particular type of targets, which are not able to generate generally good results across different categories of targets. This group of approaches is also not scalable towards large numbers of target classes. More importantly, these traditional approaches typically train a binary classifier for each class of targets independently, which loses the discriminative information hidden in various classes of targets. Therefore, it is of great significance to develop a unified framework that can automatically identify multiple classes of targets simultaneously.

The current remote sensing technology makes more and more large scene geospatial images readily available. The second problem needed to be solved urgently is to rapidly localize targets in these images using limited computational resources. Many conventional methods (Lei et al., 2012; Li and Itti, 2011; Sun et al., 2012; Zhang et al., 2011) scan over the entire image to detect targets and require equal computational time to process each location. This exhaustive search is time-consuming. Some approaches adopt segmented regions rather than individual pixels as the basic unit for target detection, which may reduce the computational cost to some extent. Nevertheless, in this case, the performance of target detection significantly depends on the segmentation results. Image segmentation is still a challenging problem in the computer vision field and it is difficult to find a segmentation algorithm that can deal with all different targets with various appearances effectively in complicated geospatial images.

Human visual attention is a cognitive procedure which can rapidly select a small set of highly informative or visually salient objects from a scene for further detailed processing such as object recognition while removing redundant information. Inspired by the human visual attention mechanism, this paper presents a practical framework for efficient and simultaneous detection of multi-class target in geospatial images by the integration of a computational attention model and discriminative learning of sparse coding. As illustrated in Fig. 1, the proposed framework consists of two major components: a learning-based saliency prediction model and a discriminative sparse coding-based multi-class target detection model. In the former model, a set of visual features which are believed to direct attention is calculated first. Subsequently, the optimal combination of various visual features is learned via a direct mapping from these features to an expert-labeled ground truth set of salient objects, which results in a computational model. The computational model is then combined with a simple image segmentation approach to predict a small set of target candidate areas (or regions of interest, ROI) for a given geospatial image. To the best of our knowledge, this ground truth dataset is among the earliest datasets to be used to investigate visual saliency modeling in the field of RSI analysis. In the second component, we train the multi-class object detector by incorporating the Fisher discrimination criterion into the dictionary learning of sparse coding. This yields a set of coding coefficients by minimizing their within-class distribution but maximizing their between-class distribution. In the target detection procedure, the trained multi-class classifier is integrated with a multi-scale scanning window mechanism to classify each location in target candidate areas, instead of the entire image, into various categories of targets.

The rest of the paper is organized as follows. Section 2 describes the learning based saliency prediction model. Section 3 presents the multi-class target detection based on the discriminative learning of sparse coding. Section 4 reports evaluation results. Finally, conclusions are drawn in Section 5.

2. Learning-based saliency prediction model

The guidance of visual attention can facilitate humans to detect and recognize meaningful targets rapidly. Computational saliency models aimed at quantitatively predicting the importance of each location in the image has been extensively studied. These models (Achanta et al., 2008, 2009; Borji, 2012; Han et al., 2006, 2011; Harel et al., 2006; Hou and Zhang, 2007; Itti et al., 1998; Judd et al., 2009; Zhang et al., 2008) generally assume that locations in the visual field that are distinctive from their contextual background are more likely to be interesting objects and applied contrast features to modulate the distinctiveness.

The saliency map can be applied to predict locations of the potential candidate targets because targets are generally distinctive from the contextual background. However, most existing saliency detection approaches (Achanta et al., 2008, 2009; Borji, 2012; Han et al., 2006, 2011; Harel et al., 2006; Hou and Zhang, 2007; Itti et al., 1998; Judd et al., 2009; Zhang et al., 2008) were developed for natural images instead of RSIs. In real optical satellite scenes, many geospatial objects such as airports, rivers, and bridges, are also conspicuous from their surroundings, which indicates that computational visual attention models can be used to predict candidate targets. Lately, a few works (Bi et al., 2012; Li and Itti, 2011; Sun et al., 2010a) which applied the attention features or models to RSI analysis have been published. For instance, Li and Itti (2011) combined saliency features and gist features for classifying small image chips. Bi et al. (2012) and Sun et al. (2010a) directly applied existing saliency models generated for natural images to detect ships (Bi et al., 2012) and salient regions (Sun et al., 2010a) from satellite images, respectively. These previous works have demonstrated the feasibility of using visual saliency models for RSI analysis.

Inspired by Judd et al. (2009) and Borji (2012), this paper proposes a supervised learning-based saliency model for prediction of target candidate areas in RSIs. This model consists of four components. At first, a ground truth database that contains salient targets in each RSI is manually constructed. To the best of our knowledge, our ground truth is among the earliest benchmarks for exploring visual saliency for RSIs. Second, a set of visual features, which have already been demonstrated to be powerful by the best existing saliency models, is calculated. Third, a supervised learning model by training classifiers directly from human-labeled ground truth data is yielded to form a saliency map. Finally, an

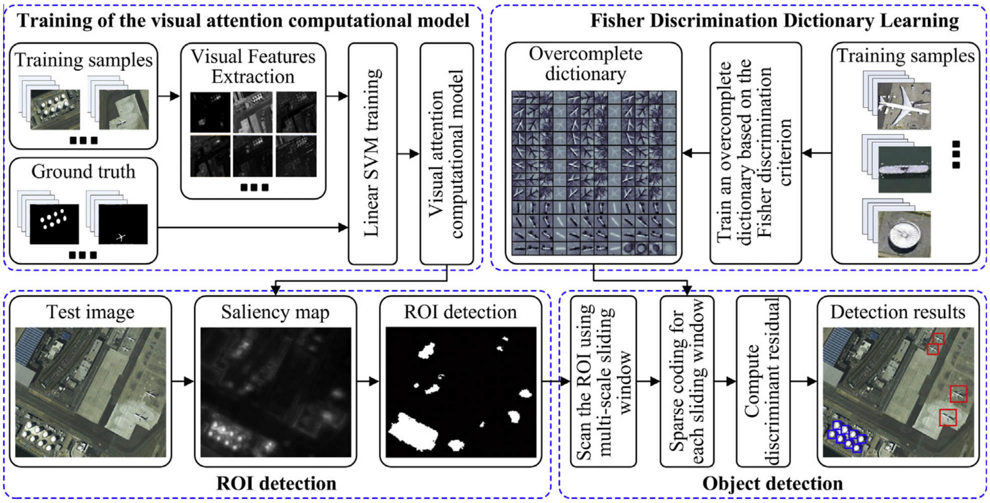

Training of the visual attention computational model
Training samples
Ground truth
Visual Features Extraction
Linear SVM training
Visual attention computational model
Fisher Discrimination Dictionary Learning
Overcomplete dictionary
Train an overcomplete dictionary based on the Fisher discrimination criterion
Training samples
Test image
Saliency map
ROI detection
ROI detection
Scan the ROI using multi-scale sliding window
Sparse coding for each sliding window
Compute discriminant residual
Detection results
Object detection

Fig. 1. The architecture of the proposed multi-class geospatial target detection framework.

adaptive algorithm is used to binarize the saliency map and predict the potential target locations.

2.1. Ground truth construction

150 satellite RSIs with the size of 1000 × 800 are collected from the publicly available Google Earth service. The spatial resolution of these images ranges from 0.5 m/pixel to 2 m/pixel. Three human subjects, who are experts in RSI processing are asked to view each image and specify salient targets, which are meaningful to represent this image. A voting strategy is taken by multiple subjects to reduce the labeling inconsistency. After the salient targets in each image are decided, the experts manually draw an accurate contour for each target. The ground truth database of 150 images is thus generated. Fig. 2 shows a number of samples from this database.

2.2. Visual features

For each RSI, the low- and mid-level features are extracted for every pixel to train the computational saliency model. These features have already been demonstrated by previous works (Achanta et al., 2008, 2009; Borji, 2012; Han et al., 2006, 2011; Harel et al., 2006; Hou and Zhang, 2007; Itti et al., 1998; Judd et al., 2009; Zhang et al., 2008) to correlate with visual saliency or be biologically plausible.

2.2.1. Low-level features

Given a pixel $I_i(R_i, G_i, B_i)$, the following low-level features are included in our feature vector $\mathbf{f}_i$. Here, R_i, G_i, B_i are the values of the red, green, and blue colors.

(1) Local contrast of intensity, orientation, and color, IC_i, OC_i, CC_i : these features were originally developed by Itti et al. (1998) and played a critical role in modeling saliency. At first, one intensity and four color channels are generated based on R_i, G_i, B_i. Four orientation channels in $\{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ are calculated using Gabor filters upon the intensity of image. Afterwards, for each of these feature channels, nine spatial scales from scale 0 to scale 8 are produced using dyadic Gaussian pyramids, which progressively subsample the original image. Then, the center-surround operator is calculated by the difference between fine and coarse scales. In total, 6, 12, and 24 center-surround difference maps are obtained for intensity, color, and orientation channels, respectively. These maps are normalized and linearly combined into 3 overall contrasts of intensity, color, and orientation. In summary, these three features are derived from the rarity of a location with respect to its local neighborhoods, which indicate the local contrast.

(2) Color values in red, green, and blue color channels, R_i, G_i, B_i, as suggested by Judd et al. (2009).

(3) Global color contrast by measuring the rarity of a pixel's color with respect to the entire image. By following Judd et al. (2009), five global color contrast features are formed based on the probability of each color estimated using color histograms.

2.2.2. Mid-level features

The outputs of five state-of-the-art saliency models, GBVS (Graph-based visual saliency) (Harel et al., 2006), SDS (Salient region detection and segmentation) (Achanta et al., 2008), WSCR (Bottom-up saliency based on weighted sparse coding residual) (Han et al., 2011), SR (Saliency detection: A spectral residual approach) (Hou and Zhang, 2007), and FT (Frequency-tuned salient

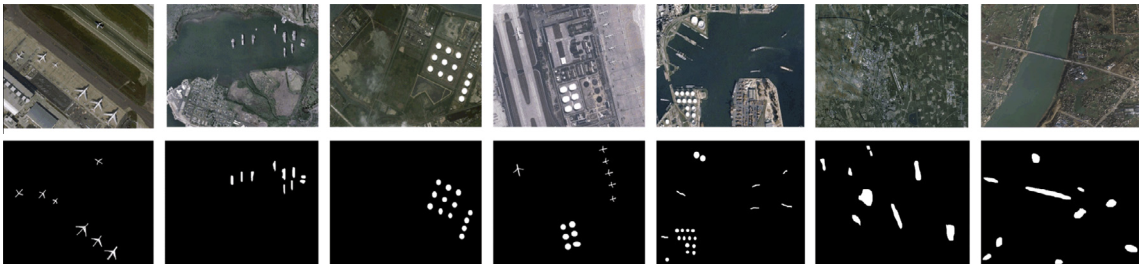


Fig. 2. A number of samples (the top row) with their corresponding ground truth data (the bottom row). In the ground truth data, the pixels in white form the expert labeled salient targets.

region detection) (Achanta et al., 2009) are utilized as the mid-level features in our work. These models are selected for several reasons, for example, they explain and characterize saliency using various mechanisms, they are fast, and they have achieved outstanding performance in predicting saliency in natural images.

Overall, the feature vector, $\mathbf{f}_i$, of the pixel I_i, can be represented as $\mathbf{f}_i = (IC_i, OC_i, CC_i, R_i, G_i, B_i, P_i^1, P_i^2, P_i^3, P_i^4, P_i^5, GB_i, SR_i, SD_i, FT_i, WS_i)$. Here, GB_i is derived from the GBVS model (Harel et al., 2006), SD_i is derived from the SDS model (Achanta et al., 2008), WS_i is derived from the WSCR model (Han et al., 2011), SR_i is derived from the SR model (Hou and Zhang, 2007), and FT_i is derived from the FT model (Achanta et al., 2009). An image example and its corresponding feature maps are shown in Fig. 3.

2.3. Learning model

In contrast to previous saliency models that fuse various visual features using a set of fixed weights, we adopt a probabilistic model via direct learning from human labeled ground truth. Essentially, the learning is able to optimize the feature weights and maximize the correlation between the features and salient targets in the ground truth.

Given a pixel I_i located at (x_i, y_i) with the feature vector $\mathbf{f}_i$, we assume the binary random variables, T, L, F, indicate whether or not a pixel belongs to the salient target, the location of a pixel, and the visual features of a pixel, respectively. Here, $i = 1, 2, \ldots, N$ and N is the total number of pixels in the image. The variable T can be described as:

$$T = \begin{cases} 1 & \text{if } I_i \text{ belongs to the salient target} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Inspired by Zhang et al. (2008), the saliency of I_i can be defined as a posterior probability as follows:

$$S_i = \Pr(T|F = \mathbf{f}_i, L = (x_i, y_i)) \tag{2}$$

According to Bayes' rule:

$$S_i = \Pr(T|F = \mathbf{f}_i, L = (x_i, y_i)) = \frac{\Pr(F = \mathbf{f}_i, L = (x_i, y_i)|T)\Pr(T)}{\Pr(F = \mathbf{f}_i, L = (x_i, y_i))} \tag{3}$$

It is reasonable to assume F and L are independent and conditionally independent given T. Therefore, Eq. (3) can be rewritten as:

$$\begin{aligned} S_i &= \frac{\Pr(F = \mathbf{f}_i, L = (x_i, y_i)|T)\Pr(T)}{\Pr(F = \mathbf{f}_i, L = (x_i, y_i))} = \frac{\Pr(F = \mathbf{f}_i|T)\Pr(L = (x_i, y_i)|T)\Pr(T)}{\Pr(F = \mathbf{f}_i)\Pr(L = (x_i, y_i))} \\ &= \Pr(T|F = \mathbf{f}_i)\frac{\Pr(T|L = (x_i, y_i))}{\Pr(T)} \propto \Pr(T|F = \mathbf{f}_i)\Pr(T|L = (x_i, y_i)) \end{aligned} \tag{4}$$

In traditional saliency models for natural images, the location prior, $\Pr(T|L = (x_i, y_i))$, is often assumed to be biased towards the image center because photographers take pictures with a strong tendency to put interesting objects close to the center. However, there is basically no center bias in RSIs. We therefore can assume that there is no prior information on the location of targets and the term of $\Pr(T|L = (x_i, y_i))$ can be ignored in Eq. (4). The saliency of the pixel I_i can be estimated by the posterior probability as:

$$S_i \propto \Pr(T|F = \mathbf{f}_i) \tag{5}$$

In our model, we approximate $\Pr(T|F = \mathbf{f}_i)$ via learning a classifier from the manually labeled ground truth described in Section 2.1. Due to its outstanding performance, we adopted a liblinear SVM to train the classifier.

During the training stage, in each training image we randomly select 400 positively labeled pixels from the salient targets of the ground truth map and 500 negatively labeled pixels from the rest of the ground truth map. All visual features are normalized to $[0,1]$. For a testing example, I_i, rather than adopting the predicted labels $(\mathrm{sgn}(\mathbf{w}^{\mathrm{T}}\mathbf{f}_i + b)$, where $\mathbf{w}, b$ are parameters in the SVM), the value of $\mathbf{w}^{\mathrm{T}}\mathbf{f}_i + b$ is used to approximate $\Pr(T|F = \mathbf{f}_i)$, which results in a continuous saliency map.

2.4. Segmentation of target candidate areas

The continuous saliency map obtained in the above subsection can quantitatively predict the likelihood that each pixel belongs to the potential salient targets. We subsequently integrate the saliency map and the image segmentation technique to segment target candidate areas from RSIs. This paper follows an elegant algorithm proposed by Achanta et al. (2008, 2009). Firstly, the mean-shift segmentation in Lab color space is applied to segment the image into small regions. Secondly, those regions whose average saliency is greater than a threshold are retained as the potential target areas. The threshold G, is adaptively determined based on the image saliency (Achanta et al., 2009):

$$G = \frac{1.6}{N}\sum_{i=1}^{N} S_i \tag{6}$$

where S_i is the saliency value of a pixel and N is the total number of pixels in an image, respectively. In this work, our target detection does not largely rely on the salient region segmentation. Therefore, the simple and fast segmentation approach can be used in our work.

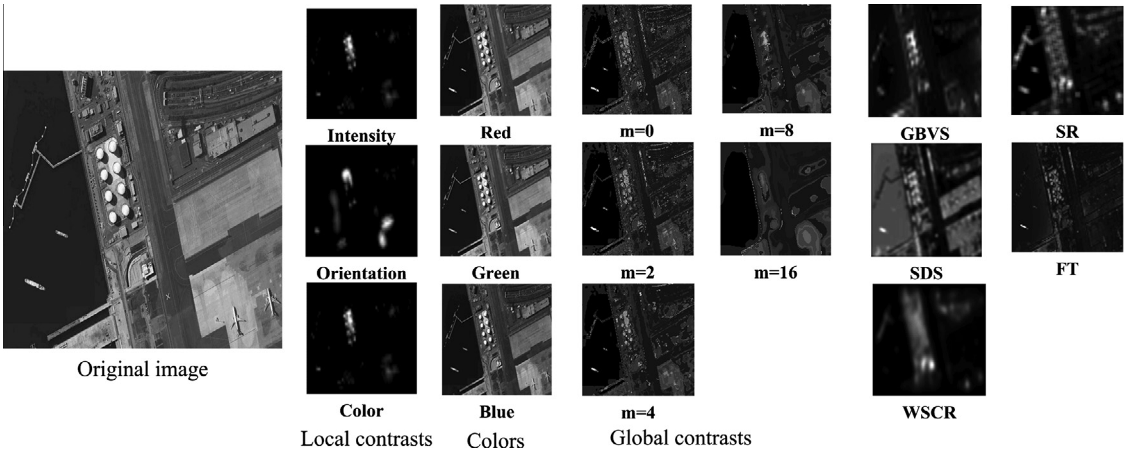


Fig. 3. An example of a RSI and its corresponding feature maps. Low-level features are shown from the second column to the fifth column (m indicates the scale of the used median filter). Mid-level features are shown in the last two columns.

3. Multi-class geospatial target detection

3.1. Problem formulation

The multi-class target detection in an image can be formulated as the issue of determining if the multi-scale windows at each location of the image contain one of a number of predefined classes of objects. The multi-scale window mechanism is adopted here to deal with the problem of target scale variation. We suppose a window be represented as $W^{\sigma_j}_{(x_i,y_i)}$ and $\mathbf{g}^{\sigma_j}_{(x_i,y_i)}$ be its corresponding feature vector. Here, $(x_i, y_i), i = 1, 2, \ldots, N$ and $\sigma_j, j = 1, 2, \ldots, M$ indicate the center and scale of the window, respectively. Assume we have C object models $\boldsymbol{\Omega} = \{\varphi_p, p = 1, \ldots, C\}$, and each of them corresponds to an object class. Suppose $o_i \in \{0, 1, \ldots, C\}$ be the object label at the location (x_i, y_i) (the 0 label indicates the background). o_i can be achieved via maximizing the following probability:

$$(\hat{p},\hat{j}) = \arg\max_{j,p} \Pr(\Omega = \{\phi_p : p = 1,2,\ldots,C\} | \{\mathbf{g}^{\sigma_j}_{(x_i,y_i)} : j = 1,2,\ldots,M\})$$
$$o_i = \hat{p} \tag{7}$$

The scale of the detected object at (x_i, y_i) is $W^{\sigma_{\hat{j}}}_{(x_i,y_i)}$.

3.2. Discriminative classifier using sparse coding

One key issue is to design the classifier for multi-class object detection. In this paper, we adopt the sparse representation-based classification (SRC) algorithm. We select the SRC algorithm for the following two reasons. First, it is one of the state-of-the-art approaches for classification and has been successfully applied to tasks such as face recognition (Mairal et al., 2008; Wright et al., 2009; Yang et al., 2011), digit and texture classification (Mairal et al., 2008; Yang et al., 2011), and image denoising. Recently, a few efforts have been devoted to apply SRC to target detection in RSI. For example, Chen et al. (2011) proposed a method for target detection in hyperspectral imagery based on sparse representation, which exploits the sparsity constraint, reconstruction error and the neighboring pixels that have similar spectral attributes. Sun et al. (2012) integrated sparse coding and BoW model to detect airports in RSIs. In this paper, we attempt to explore the feasibility of leveraging SRC for multi-class target detection in RSIs. Second, many traditional single-class target detectors still rely on the capability of their extracted features. Normally, the features used in one detector work effectively only for a specific target and may not be powerful enough to tackle other types of targets. This problem limits the usage of these detectors to multi-class target detection. Nevertheless, as demonstrated in Wright et al. (2009), the theory of compressed sensing implies that the performance of the SRC algorithm does not depend on powerful features. Moreover, SRC algorithm is robust to occlusion (Wright et al., 2009). These properties make the SRC appropriate for the task of multi-class geospatial target detection.

The basic assumption of SRC algorithm is that a high-dimensional image can be described by a few representative atoms of a learned dictionary in a low-dimensional manifold. The classical SRC algorithm can be used to solve the multi-class classification problem as follows. Given a set of training samples $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_p$, -$\ldots, \mathbf{X}_C]$, where $\mathbf{X}_p$ is the subset of the training samples from class p, sparse representations of these data can be addressed by solving the following minimization problem:

$$\min_{\mathbf{D},\boldsymbol{\alpha}} ||\mathbf{X} - \mathbf{D}\alpha||_2^2 + \eta||\boldsymbol{\alpha}||_1 \tag{8}$$

Here, $\mathbf{D} = [\mathbf{D}_1, \ldots, \mathbf{D}_p, \ldots, \mathbf{D}_C]$ is an over-complete dictionary to be learned, where $\mathbf{D}_p$ is a sub-dictionary associated with the pth class, and $\boldsymbol{\alpha}$ is the coding coefficients. In Eq. (8), $||\mathbf{X}$ - $\mathbf{D}\boldsymbol{\alpha}||_2$ adopts the ℓ_2 norm to measure the reconstruction error, $||\boldsymbol{\alpha}||_1$ is the sparsity penalty based on ℓ_1 norm, and η is a constant. To classify a testing data $\hat{\mathbf{x}}$, we firstly sparsely code it by Wright et al. (2009)

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} ||\hat{\mathbf{x}} - \mathbf{D}\alpha||_2^2 + \eta||\boldsymbol{\alpha}||_1 \tag{9}$$

Afterwards, the residuals (or the reconstruction errors) are computed by

$$e_p(\hat{\mathbf{x}}) = ||\hat{\mathbf{x}} - \mathbf{D}_p\hat{\boldsymbol{\alpha}}_p||_2 \tag{10}$$

where $\hat{\boldsymbol{\alpha}}_p$ is the coefficient vector associated with the pth class. Finally, $\hat{\mathbf{x}}$ is classified to the class with $\arg\min e_p(\hat{\mathbf{x}})$.

For the multi-class classification, the discriminative information hidden in various categories of targets is potentially useful, which is ignored by the original SRC algorithm. In this paper, we basically follow Mairal et al. (2008) and Yang et al. (2011) and develop a multi-class classifier by incorporating discriminative information into the dictionary learning of SRC. The objective function in Eq. (8) can be improved as (Yang et al., 2011):

$$\min_{\mathbf{D},\boldsymbol{\alpha}} \{DRE(\mathbf{D}, \mathbf{X}, \boldsymbol{\alpha}) + \lambda_1 CS(\boldsymbol{\alpha}) + \lambda_2 CD(\boldsymbol{\alpha})\} \tag{11}$$

where $DRE(\mathbf{D}, \mathbf{X}, \boldsymbol{\alpha})$ is discriminative reconstruction error, $CS(\boldsymbol{\alpha})$ is coefficient sparsity, $CD(\boldsymbol{\alpha})$ is coefficient discrimination, and λ_1 and λ_2 are constants.

The discriminative reconstruction error aims to fulfill three objectives simultaneously. First, the dictionary $\mathbf{D}$ can represent the training samples $\mathbf{X}$ well. Second, the sub-dictionary $\mathbf{D}_p$ can well represent the samples $\mathbf{X}_p$ from the same class. Third, the sub-dictionary $\mathbf{D}_p$ is not able to well represent the samples $\mathbf{X}_q$, $q \neq p$ from the different classes. Overall, $DRE(\mathbf{D}, \mathbf{X}, \boldsymbol{\alpha})$ is defined as:

$$DRE(\mathbf{D}, \mathbf{X}, \boldsymbol{\alpha}) = ||\mathbf{X} - \mathbf{D}\alpha||_F^2 + \sum_{p=1}^{C} ||\mathbf{X}_p - \mathbf{D}_p\boldsymbol{\alpha}_p^p||_F^2$$
$$+ \sum_{p=1}^{C} \sum_{q=1, q\neq p}^{C} ||\mathbf{D}_q\boldsymbol{\alpha}_p^q||_F^2 \tag{12}$$

where $\boldsymbol{\alpha}_p^q$ is the coding coefficient of $\mathbf{X}_p$ over $\mathbf{D}_q$, and $||\bullet||_F$ denotes the Frobenius norm.

The second term, $CS(\boldsymbol{\alpha})$, in Eq. (11) is defined by following the classical SRC algorithm as $||\boldsymbol{\alpha}||_1$ (Wright et al., 2009). The coefficient discrimination, $CD(\boldsymbol{\alpha})$, is based on the assumption the coding coefficients are discriminative across various classes. Following Yang et al. (2011), we adopt the Fisher discrimination criterion to define $CD(\boldsymbol{\alpha})$:

$$CD(\boldsymbol{\alpha}) = \mathrm{tr}(SW(\boldsymbol{\alpha})) - \mathrm{tr}(SB(\boldsymbol{\alpha})) + \lambda_3||\boldsymbol{\alpha}||_F^2$$
$$SW(\boldsymbol{\alpha}) = \sum_{p=1}^{C} \sum_{\boldsymbol{\alpha}_{p,k} \in \boldsymbol{\alpha}_p} (\boldsymbol{\alpha}_{p,k} - \bar{\boldsymbol{\alpha}}_p)(\boldsymbol{\alpha}_{p,k} - \bar{\boldsymbol{\alpha}}_p)^{\mathbf{T}} \tag{13}$$
$$SB(\boldsymbol{\alpha}) = \sum_{p=1}^{C} n_p(\bar{\boldsymbol{\alpha}}_p - \bar{\boldsymbol{\alpha}})(\bar{\boldsymbol{\alpha}}_p - \bar{\boldsymbol{\alpha}})^{\mathrm{T}}$$

where $\mathrm{tr}(\bullet)$ is the trace operator, $\bar{\boldsymbol{\alpha}}_p$ and $\bar{\boldsymbol{\alpha}}$ are the mean vector of $\boldsymbol{\alpha}_p$ and $\boldsymbol{\alpha}$ respectively, n_p is the number of samples in the pth class, and $||\boldsymbol{\alpha}||_F^2$ is an elastic term to make $CD(\boldsymbol{\alpha})$ convex. The minimization of $CD(\boldsymbol{\alpha})$ can basically lead to minimizing the within-class scatter of coding coefficients $SW(\boldsymbol{\alpha})$, and maximizing the between-class scatter of coding coefficients $SB(\boldsymbol{\alpha})$, simultaneously.

Similar to Mairal et al. (2008) and Yang et al. (2011), we optimize the objective function in Eq. (11) by iterating between updating $\mathbf{D}$ while fixing $\boldsymbol{\alpha}$ and updating $\boldsymbol{\alpha}$ while fixing $\mathbf{D}$. In brief, the following four steps are performed for the optimization. First, all the atoms of each $\mathbf{D}_p$ are initialized as random vectors having unit ℓ_2 norm. Second, we fix $\mathbf{D}$ and compute the sparse coding coefficients $\boldsymbol{\alpha}_p, p = 1, \ldots, C$, class by class, by optimizing the objective function:

$$\min_{\boldsymbol{\alpha}_p}\{DRE(\mathbf{D},\mathbf{X}_p,\boldsymbol{\alpha}_p)+\lambda_1 CS(\boldsymbol{\alpha}_p)$$
$$+\lambda_2\left(\|\boldsymbol{\alpha}_p-\overline{\mathbf{A}}_{\mathbf{p}}\|_F^2-\sum_{k=1}^{C}\|\overline{\mathbf{A}}_{\mathbf{k}}-\overline{\mathbf{A}}\|_F^2+\lambda_3\|\boldsymbol{\alpha}_p\|_F^2\right)\Bigg\} \tag{14}$$

where $\overline{\mathbf{A}}_{\mathbf{k}}$ and $\overline{\mathbf{A}}$ denote the mean vector matrices of class k, $k = 1, \ldots, C$, and all classes, respectively. They can be obtained by taking n_k mean vectors $\bar{\boldsymbol{\alpha}}_k$ and $\bar{\boldsymbol{\alpha}}$ as their column vectors. Third, we fix $\boldsymbol{\alpha}$ and update each $\mathbf{D}_p$ by optimizing the objective function:

$$\min_{\mathbf{D}_p}\left\{\|\mathbf{X}-\mathbf{D}_p\boldsymbol{\alpha}^p-\sum_{q=1,q\neq p}^{C}\mathbf{D}_q\boldsymbol{\alpha}^q\|_F^2+\|\mathbf{X}_p-\mathbf{D}_p\boldsymbol{\alpha}_p^p\|_F^2+\sum_{q=1,q\neq p}^{C}\|\mathbf{D}_p\boldsymbol{\alpha}_q^p\|_F^2\right\} \tag{15}$$

where $\boldsymbol{\alpha}^p$ is the coding coefficients of $\mathbf{X}$ over $\mathbf{D}_p$. Fourth, the second and third steps are repeated until the convergence or the maximum number of iterations is achieved.

3.3. Target detector training

We build the geospatial target detector based on the multi-class classifier described in Section 3.2. In principle, we can build a detector to detect a large number of various categories of geospatial targets. In this paper, our experiment focuses on the evaluation of the detector using four different categories of targets: airplanes, ships, storage tanks, and baseball diamonds. These targets are generally important targets in satellite images. In geospatial target detection, target orientation variation and scale variation are two nontrivial issues. However, the discriminative sparse coding-based classifier actually does not take into account these two issues in theory. We therefore deal with them in the classifier training and target detection stages. In the training stage, training images are downloaded from Google Earth. In these images, a human expert manually labeled targets using the bounding box. The labeled targets are utilized as the samples for training the detector. To deal with the problem of the target orientation variation, our idea is to collect training samples with various orientations. For each manually labeled target sample, we firstly rotate it in the step of 10° within a range, which results in another a number of training samples. Specifically, for the sample of airplane and baseball diamond targets, the range for rotation is between 0° and 360°. For the sample of ship targets, the range for rotation is between 0° and 180° because their shape has bilaterally symmetry. For the sample of storage tank targets, it is unnecessary to perform the rotation because their shape is a circle. In this way, we can collect 36, 18, 1, 36 training samples for each manually labeled sample of airplane, ship, storage tank and baseball diamond, respectively. Afterwards, we firstly smooth all training samples using a Gaussian low pass filter and down-sample them from their original scale to a uniform scale. In our current implementation, the uniform scale is empirically set to 19 × 19 that will be described in Section 4.2. Fig. 4 shows a number of training examples.

The pixel gray values of each training sample of size 19 × 19 are then normalized to have unit ℓ_2 norm and concatenated to form a 361-D features. Feature vectors of training samples from the same class are stacked as columns of the matrix $\mathbf{X}$. The discriminative dictionary is learned via the optimization of Eq. (11), which eventually forms the multi-class classifier. Fig. 5 shows the learned dictionary for classifying airplanes, ships, storage tanks, and baseball diamonds.

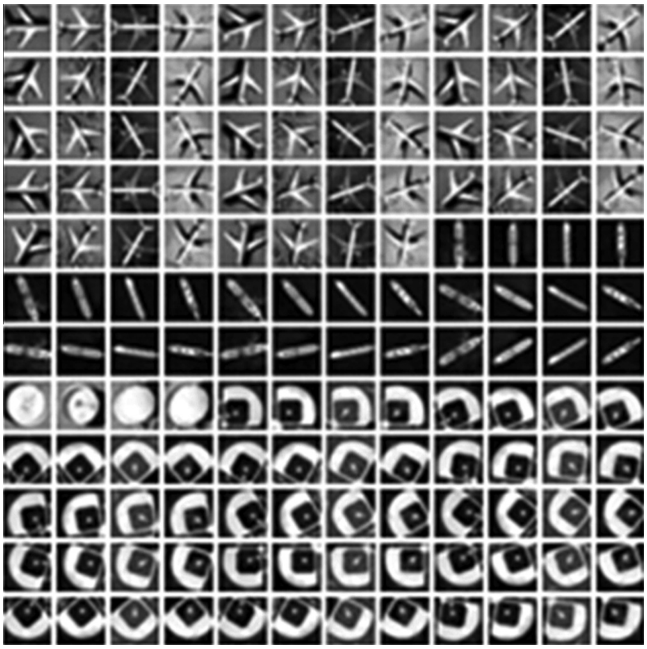


Fig. 4. A number of training samples.

3.4. Target detection

Given a test image, we firstly apply the learned saliency model described in Section 2 to predict candidate locations that potentially contain targets. Afterwards, we adopt multi-scale sliding windows to scan the image over candidate locations and determine if these windows contain those predefined classes of targets. The multi-scale windows can deal with the problem of target scale variation. By following the problem formulation described in Section 3.1, we assume the multi-scale windows at the location (x_i, y_i) be represented as $W_{(x_i,y_i)}^{\sigma_j}$. Similar to the training stage, we smooth all $W_{(x_i,y_i)}^{\sigma_j}$ using a Gaussian low pass filter and down-sample them from their original scale to the uniform scale of 19 × 19. The unit ℓ_2 norm normalized pixel values in each down-sampled window form a 361-D feature vector $\mathbf{g}_{(x_i,y_i)}^{\sigma_j}$. Afterwards, $\mathbf{g}_{(x_i,y_i)}^{\sigma_j}$ is encoded by the trained dictionary $\mathbf{D}$ in terms of:

$$\hat{\boldsymbol{\alpha}}=\arg\min_{\boldsymbol{\alpha}}||\mathbf{g}_{(x_i,y_i)}^{\sigma_j}-\mathbf{D}\alpha||_2^2+\eta||\boldsymbol{\alpha}||_1 \tag{16}$$

Then, the classification metric (Yang et al., 2011) with respect to each target class p is finally defined based on the reconstruction error and the distance between the encoding coefficients and the coefficient mean of the pth class:

$$e_p(\mathbf{g}_{(x_i,y_i)}^{\sigma_j})=||\mathbf{g}_{(x_i,y_i)}^{\sigma_j}-\mathbf{D}_p\hat{\boldsymbol{\alpha}}_p||_2^2+\lambda_4||\hat{\boldsymbol{\alpha}}-\bar{\boldsymbol{\alpha}}_p||_2^2 \tag{17}$$

where λ_4 is the trade off weight.

For each location (x_i, y_i) in the image, we can use $e_p(\mathbf{g}_{(x_i,y_i)}^{\sigma_j})$ to approximate $\Pr(\Omega=\{\phi_p:p=1,2,\ldots,C\}|\{\mathbf{g}_{(x_i,y_i)}^{\sigma_j}:\ j=1,2,\ldots,M\})$ in practice. Therefore, we calculate:

$$(\hat{p},\hat{j})=\arg\min_{p,j}\{e_p(\mathbf{g}_{(x_i,y_i)}^{\sigma_j}):p=1,2,\ldots,C;j=1,2,\ldots,M\} \tag{18}$$

If $e_{\hat{p}}(\mathbf{g}_{(x_i,y_i)}^{\sigma_{\hat{j}}})>\tau$ (τ is a predefined threshold), we decide that the location (x_i, y_i) is the background. Otherwise, we detect the $\hat{p}$th class of target at (x_i, y_i) with the scale of $W_{(x_i,y_i)}^{\sigma_{\hat{j}}}$. Fig. 6 shows a number of detection examples. As can be seen, the coefficient vectors of four targets are sparse including a small number of nonzero entries whereas the coefficient vector of background is relatively dense including many nonzero entries. These illustrations imply that the sparsity of coefficients is a discriminative attribute for the classification.

In practice, a target may be successfully detected by a couple of windows with different scales or slightly different locations. A

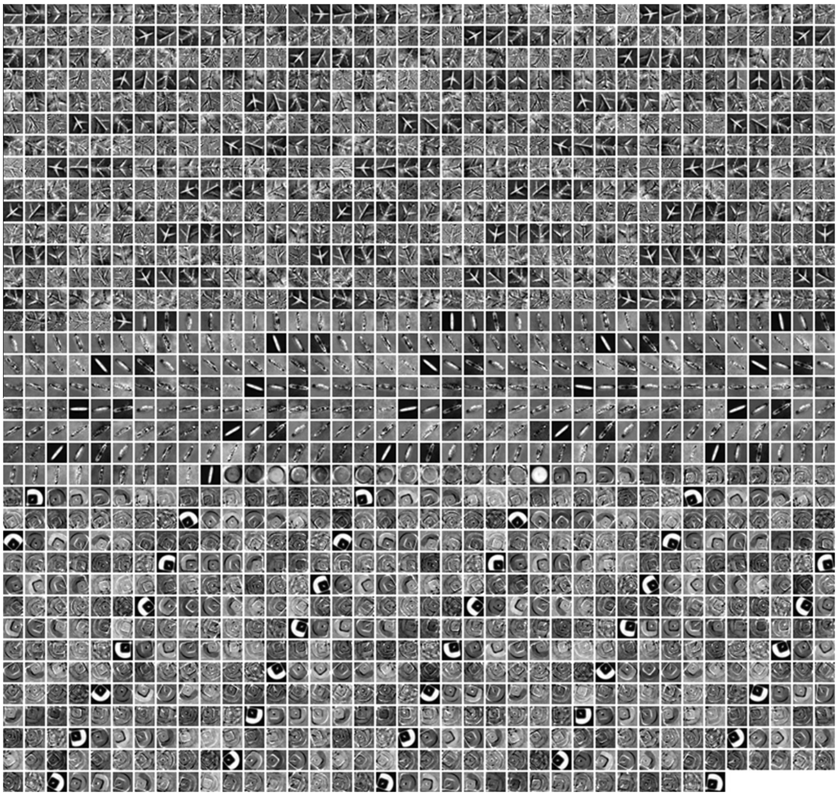


Fig. 5. The learned discriminative dictionary for the targets of airplanes, ships, and storage tanks, and baseball diamonds.

post-processing procedure called non-minimum suppression is used to reduce false positives. Specifically, we sort all scan windows in ascending order based on their corresponding $e_{\tilde{p}}(\mathbf{g}^{\sigma_j}_{(x_i,y_i)})$ and greedily select the windows with the lowest $e_{\tilde{p}}(\mathbf{g}^{\sigma_j}_{(x_i,y_i)})$. The windows which are at least 50% covered by any previously selected windows are eliminated.

4. Experimental results

4.1. Experimental setup

In theory, the proposed multi-class target detection framework can detect arbitrary number of classes of geospatial targets. However, in our experiments, we used the task of detection of four various types of targets to evaluate the performance of the proposed work. These four classes of targets are airplanes, ships, storage tanks, and baseball diamonds. We collected 400 high resolution RSIs from Google Earth for our evaluations. The resolution of these images is 1000 by 800 pixels and a majority of these images contain at least one type of targets we intend to detect. The spatial resolution of these images ranges from 0.5 m/pixel to 2 m/pixel. We separated these images into three independent sets: the set containing 150 images for testing saliency models, the set containing 43 images for training the multi-class target detector, and the set containing 207 images for testing the performance of the target detector. From those 43 training images, we labeled 21 airplane samples, 21 ship samples, 21 storage tank samples, and 21 baseball diamond samples. To deal with the orientation variance problem, we rotated each sample as described in Section 3.3 and finally we collected 756 airplanes, 378 ships, 21 storage tanks, and 756 baseball diamonds as the training samples in total. The testing set contains 756 airplane targets, 295 ship targets, 1876 storage tank targets, and 106 baseball diamond targets. The sizes of the targets in the testing images vary from 30 × 30 to 150 × 150 pixels approximately. In our developed framework, there are some critical parameters including η, λ_1, λ_2, λ_3, λ_4, which have important effect on the performance of the proposed approach. We trained a set of different classifiers by using different values of λ_1, λ_2 and λ_3, and then we evaluated how the trained classifiers affected the detection performance by varying the values of η and λ_4. According to our experiment, these parameters influenced the detection performance moderately and the best performance was achieved when $\eta = 0.15$, $\lambda_1 = 0.005$, $\lambda_2 = 0.05$, $\lambda_3 = 1$ and $\lambda_4 = 0.5$. Consequently, we empirically used these optimal parameters in our all subsequent evaluations.

4.2. Parameter analysis

In the proposed framework, the dictionary plays an important role in deciding the overall detection performance. There are two critical parameters that influence the dictionary construction and determine the size of the dictionary. One is the uniform scale described in Section 3.3 and the other is the number of training samples. We conducted an experiment to analyze how detection performance is affected by the values of these two parameters. 30 RSIs randomly selected from our testing database were used in this experiment. We varied the values of these two parameters and adopted the average precision (AP) to quantitatively evaluate the target detection performance. AP is a standard metric for object detector evaluation, which calculates the area under the precision-recall curve (PRC). The details on PRC will be provided in Section 4.4. The evaluation results are listed in Table 1. As can be seen, these two parameters influence the detection performance moderately and the best performance was achieved when uniform scale is set to 19 × 19 and the number of training samples is set to

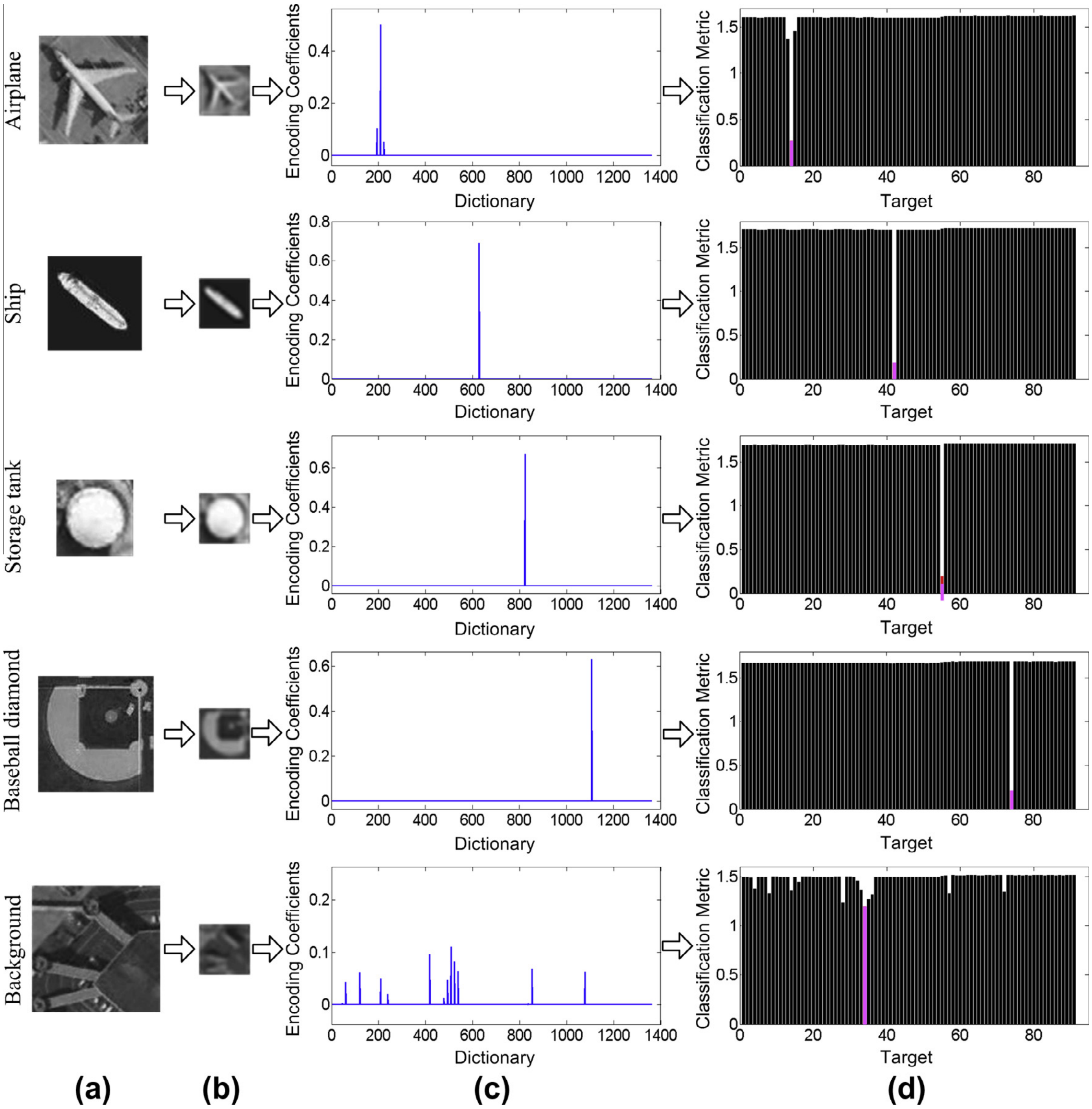


Fig. 6. A number of detection examples. (a) Testing examples; (b) down-sampled examples; (c) encoding coefficients on the learned dictionary; (d) classification metrics with respect to each target class.

1365 (540 airplane samples, 270 ship samples, 15 storage tank samples, and 540 baseball diamond samples). Consequently, we empirically set uniform scale as 19 × 19 and the number of training samples as 1365 in dictionary construction, which results in a dictionary with the size of 361 × 1365. We also used this dictionary size in our all following evaluations.

4.3. Evaluation of saliency prediction model

In our work, we proposed a saliency prediction model to yield the saliency map for each RSI based on the assumption that targets are generally distinctive from the contextual background. Afterwards, the segmentation was performed on the saliency map to estimate the potential target locations. To learn and evaluate our saliency prediction model, we collected 150 high resolution satellite images from Google Earth. As we described in 2.1, the ground truth dataset was manually built. 130 images were randomly selected as the training data to learn the model and the remaining 20 images were adopted as the test data to evaluate the model performance. We also compared the proposed model with two state-of-the-art models called FT (Achanta et al., 2009) and WSCR (Han et al., 2011). Fig. 7 shows a number of saliency maps and their corresponding segmentation results by using the three different models. As can be seen, the segmentation results based on our saliency prediction model basically can precisely provide the informative regions and candidate locations that contain targets. Although the two other methods can also generate quite good

Table 1
The effect of dictionary size on detection performance (the underlined values denote the most appropriate parameters and the best result).

Uniform scale	The number of training samples	AP
11 × 11	910	0.6754
11 × 11	1365	0.6474
11 × 11	1820	0.6433
15 × 15	910	0.7031
15 × 15	1365	0.6527
15 × 15	1820	0.7173
19 × 19	910	0.6837
<u>19 × 19</u>	<u>1365</u>	<u>0.7680</u>
19 × 19	1820	0.7321

segmentation results which contain those targets, they obtain much more false positives compared with our method. It implies that the subsequent target detection on these results has to be performed over larger areas, which will take much time and thus significantly limit the efficiency.

We also constructed two quantitative evaluations to test the proposed saliency prediction model. The first evaluation is to test the performance of the saliency map. We repeated the above evaluation (130 randomly selected images as training data and other 20 images as testing data) five times and calculated the average true positive rates (TPR) and false positive rates (FPR). Similar to most previous works on visual saliency models, we adopted the Receiver Operating Characteristic (ROC) curve and the Area Under the ROC Curve (AUC) as the metrics. Given a saliency map with saliency values in the range [0, 255], we threshold the saliency map at a threshold *Th* within [0, 255] to obtain a binary mask for the salient object. To quantitatively compare the performance of various saliency maps, the threshold *Th* is varied from 0 to 255 and the FPR and the TPR are computed at each value of *Th*, which results in the ROC. Here, the FPR and TPR were computed by comparing each pixel in the binarized saliency map against the ground truth. The FPR measures the fraction of negative examples that are falsely classified as positive examples whereas the TPR measures the fraction of positive examples that are correctly classified. The bigger AUC value indicates the better performance of the saliency map. Fig. 8 shows the comparison ROCs and their corresponding AUCs denoted in parentheses. As can been seen from Fig. 8, the proposed saliency model is better than other two state-of-the-art models (Achanta et al., 2009; Han et al., 2011).

The purpose of our saliency prediction model is to provide the target candidates for the subsequent target detector. In our model, we perform the segmentation on the saliency map and the segmented areas are considered as the target candidate areas. Our second experiment is to evaluate the performance of predicted results. We expect that the segmented target candidate areas are as small as possible while they contain as many as possible true targets. As a result, we use two metrics to measure the performance. One metric is called potential recall, which measures the percentage that targets to be detected are contained in the segmented candidate areas. A target is regarded as "contained" when its center falls in the segmented candidate areas. The other metric is called prediction area rate, which measures the percentage of the predicted target candidate areas over the entire image size. The smaller prediction area rate generally leads to the higher efficiency of the subsequent detection. Table 2 lists the recalls and the prediction area rates obtained in terms of different methods. As can be seen from the comparison results, the proposed prediction model can achieve the highest recall and the smallest prediction area rate. Although the model of Han et al. (2011) also can provide quite good results in predicting target candidates, its segmented area is much bigger than that generated by our model, thus its detection takes much longer time than ours.

4.4. Evaluation of target detection

In the detection stage, because the scales of targets may be different in RSIs, we adopted the multi-scale window mechanism. A number of multi-scale windows were slid over the predicted target areas. Our current implementation set the window scales from 30 × 30 pixels to 150 × 150 pixels with the interval of 10 pixels for the window side. The sliding step-size was set to be 10% of the window side length similar to the work (Sun et al., 2012). 207 satellite images containing 3033 targets were used to test the performance of the detectors. Fig. 9 shows a number of detected results by using the proposed approach ($\tau = 0.5031$). As can be seen, the proposed multi-class detector can locate the targets of various classes effectively although these targets may have different orientations and sizes.

We also constructed two quantitative evaluations to test the performance of the proposed work. We manually labeled the targets using bounding boxes in all testing images as the ground truth. For each category of targets, let *NP*, *TP*, and *FP* denote the total number of target objects in the ground truth, the number of true positives, and the number of false positives in all testing images, respectively. A detection is marked as a true positive when its corresponding window can cover more than 50% of a ground

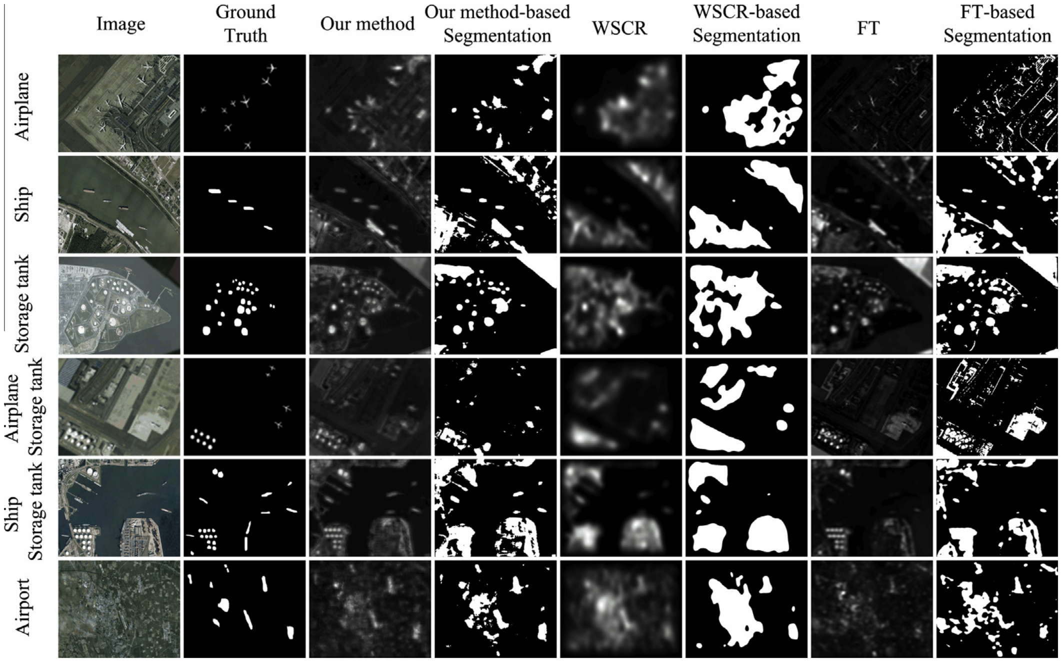


Fig. 7. A number of saliency maps and their corresponding segmentation results by using the proposed model, WSCR model and FT model. (The words shown in the leftmost column indicate the meaningful geospatial targets contained in the corresponding RSIs.)

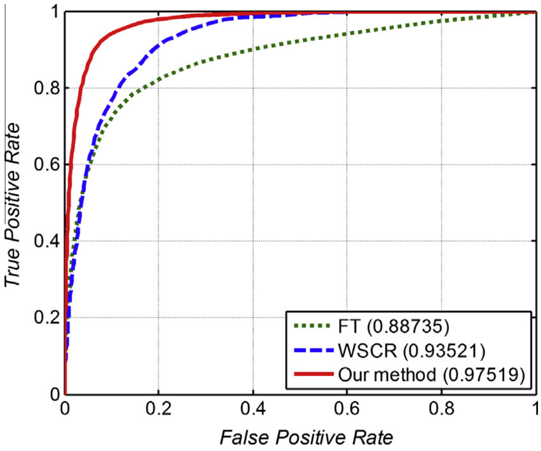


Fig. 8. ROC curves of various saliency maps using different methods and their corresponding AUC values denoted in parentheses.

Table 2
Performance comparison of different methods in predicting target candidate areas.

	Our method	FT	WSCR
Potential recall	0.9977	0.8153	0.9935
Prediction area rate	0.1165	0.1692	0.1999

truth (Felzenszwalb et al., 2010; Santosh et al., 2010). By following most previous works (Akçay and Aksoy, 2008; Lei et al., 2012; Li and Itti, 2011; Tao et al., 2011), we used the PRC as the metric to measure the performance of the object detectors, where the recall and precision are defined as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{NP}}$$
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (19)$$

In each evaluation, the PRC was used as the metric, which is plotted using the recalls and precisions on all testing data under different threshold values of τ. In our experiment, the threshold τ can considerably affect the detection performance. To be specific, a high value of τ may yield a good *Recall* but a bad *Precision* and vice versa. The optimal τ is empirically determined to the value which leads to the highest *F*1 - *measure*. Here, *F*1 - *measure* is computed as: *F*1 - *measure* = (2 × *Recall* × *Precision*)/(*Recall* + *Precision*). In the first evaluation, we compared the proposed approach with three other state-of-the-art detection algorithms (Xu et al., 2010; Sun et al., 2012; Yang et al., 2009) using the same sets of training data and testing data. The first algorithm (Xu et al., 2010) is based on BoW feature description and SVM classifier, which is called BoWSVM in this paper, and the size of the codebook is set to 200 which obtained by K-means clustering. The second algorithm (Sun et al., 2012) is based on spatial sparse coding bag-of-words

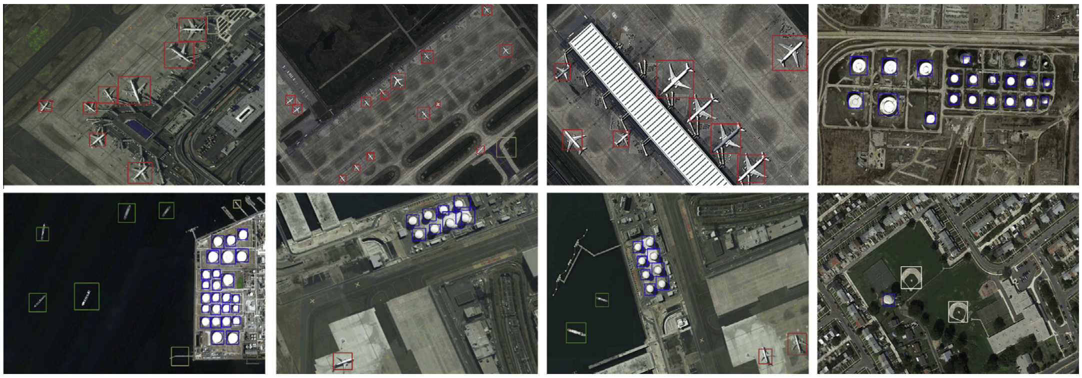


Fig. 9. A number of multiple target detection results by the proposed approach ($\tau = 0.5031$). The truly detected airplane targets, ship targets, storage tanks, and baseball diamond are labeled by red, green, blue, and white boxes, respectively. The false positive examples are labeled by yellow boxes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

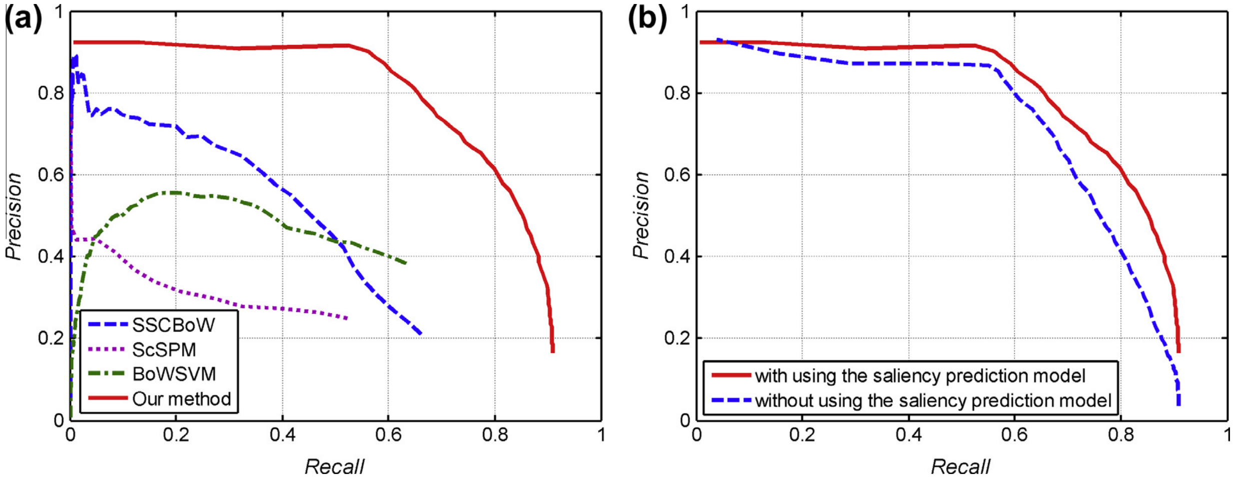
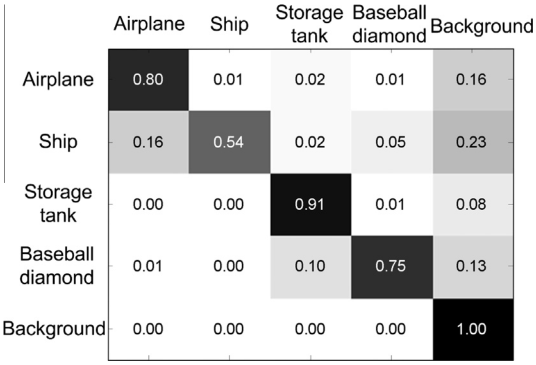


Fig. 10. Quantitative comparison results by using various approaches. (a) The PRCs of using different target detection approaches. (b) The PRCs of using and not using saliency prediction model.

Table 3
Performance comparisons of different approaches in terms of AP.

Approach	AP
Our method with using the saliency prediction model	0.7420
Our method without using the saliency prediction model	0.6481
BoWSVM (Xu et al., 2010)	0.3020
SSCBOW (Sun et al., 2012)	0.3822
ScSPM (Yang et al., 2009)	0.1692



	Airplane	Ship	Storage tank	Baseball diamond	Background
Airplane	0.80	0.01	0.02	0.01	0.16
Ship	0.16	0.54	0.02	0.05	0.23
Storage tank	0.00	0.00	0.91	0.01	0.08
Baseball diamond	0.01	0.00	0.10	0.75	0.13
Background	0.00	0.00	0.00	0.00	1.00

Fig. 11. The confusion matrix for evaluation of multi-class classification ($\tau = 0.5031$).

and SVM classifier, which is called SSCBOW in this paper, and the size of codebook is set to 400 which trained using sparse coding. The third algorithm (Yang et al., 2009) is based on the sparse coding of SIFT features, spatial max pooling, and SVM classifier, which is called ScSPM in this paper. For this algorithm, we use 3 scale levels which generate 21 segments and the codebook size is set to 200. For the task of multi-class classification, their SVM classifier is built using one-against-all strategy (Yang et al., 2009). Fig. 10(a) shows the PRCs of three different approaches. It is easy to see that the proposed algorithm can achieve the best performance for the purpose of multi-class target detection.

The second evaluation is to compare the performance with using and without using the saliency prediction model. The latter approach indicates that the detection is performed over the entire image instead of only over potential target areas predicted by the saliency model. Fig. 10(b) shows the comparison results. With the same recall value, the precision with using the saliency prediction model is higher than that without using the saliency prediction model, which means the proposed method yields a lower false alarm rate with the same true positives. With the same precision value, the recall with using the saliency prediction model is also higher than that without using the saliency prediction model, which means the proposed method can obtain more true positives with the same false alarm rate. The comparison results demonstrate that the saliency prediction model can significantly improve the performance in terms of PRC.

In addition to the PRC evaluation, we also adopted the AP to quantitatively evaluate the performance of different methods. Table 3 shows the APs of different approaches. As can be seen, the proposed method is substantially better than other methods in terms of AP. This result is consistent with the result in terms of PRC.

To evaluate the discrimination of multi-class classification, we provided the confusion matrix in Fig. 11. As can be seen from the results, our method can discriminate the inter-class difference among different category targets.

4.5. Evaluation of computational complexity

The proposed algorithm was implemented using MATLAB R2010b and run on a PC with Intel Pentium 2.13 GHz CPU and 2 GB memory. The offline discriminative dictionary training took a few hours. The online sparse encoding of each sliding window averagely took 0.0037 s. The average time consumed for saliency prediction was 10 s approximately. The selection of step-size for sliding essentially decides the running time of the target detection. There is a trade-off between the detection accuracy and cost when we vary the value of sliding step-size. A large value of sliding step-size leads to the faster detection speed and the lower detection accuracy. In contrast, we can obtain the better detection accuracy whereas the detection will take longer time when we use small value of sliding step-size. Fig. 12 presents the average running time for detecting multiple targets in 1000 × 800 satellite RSIs associated with different sliding step-size values and the use of saliency prediction model or not. Note that we utilized a fixed set of

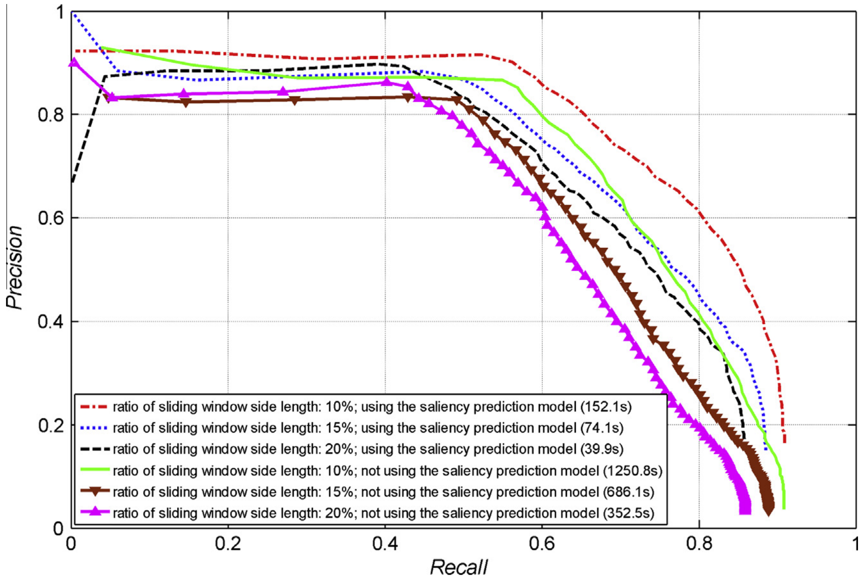


Fig. 12. PRC and running time comparisons of using different sliding step-size values and using or not using the saliency prediction model. Each method's running time is shown in the parentheses and s indicates seconds.

parameter values except for the sliding step-size for all test images, and the sliding step-size was set to be proportional to the sliding window side length. As can be seen, the sliding step-size greatly influences the detection speed whilst having a moderate impact on detection accuracy. The usage of the saliency model can substantially reduce the computational cost of detection and improve its efficiency.

5. Conclusions

In this paper, a practical framework for detecting multi-class geospatial targets from RSIs has been proposed. The unique contributions that can distinguish the proposed work from existing works are threefold. First, a supervised learning model was built based on a variety of saliency features and models, which enables it to predict candidate locations containing targets. Second, a multi-class classifier based on the discriminative dictionary learning of sparse coding representation was developed to detect multiple targets from different classes simultaneously and efficiently. The developed classifier leverages the discriminative information hidden in various categories of targets. Finally, a ground truth dataset using images from Google Earth was constructed by three experts for studying visual saliency modeling in RSIs, which is among the earliest benchmarks in this area to the best of our knowledge.

We plan to extend this work along the following directions. First, we will apply the developed target detector to detect a larger number of classes of targets. Second, the false positive rate of our current saliency prediction model is still slightly high, which reduces the efficiency of the detector. Our future work will incorporate more powerful features or improve the salient region segmentation to further decrease the false alarm rate. Third, in this work, our detector was based on multi-scale window scanning over predicted target candidate areas. We may adopt Hough voting on predicted target candidate areas, which may further improve the efficiency.

Acknowledgements

The authors appreciate the constructive suggestions from Dr. Alistair Sutherland. This work was partially supported by the National Science Foundation of China under Grant 61005018, 91120005, 61333017, 61103061, NPU-FFR-JC20120237, Program for New Century Excellent Talents in University under grant NCET-10-0079, and NPU-Z2013105.

References

Achanta, R., Estrada, F., Wils, P., Süsstrunk, S., 2008. Salient region detection and segmentation. In: Proceedings of the 6th International Conference on Computer Vision Systems (ICVS 2008), Santorini, Greece, 12–15 May, pp. 66–75.

Achanta, R., Hemami, S., Estrada, F., Susstrunk, S., 2009. Frequency-tuned salient region detection. In: Proceedings of the 2009 Computer Vision and Pattern Recognition, 2009 (CVPR 2009), Miami, Florida, USA, 20–25 June, pp. 1597–1604.

Akçay, H.G., Aksoy, S., 2008. Automatic detection of geospatial objects using multiple hierarchical segmentations. IEEE Trans. Geosci. Remote Sens. 46 (7), 2097–2111.

Bhagavathy, S., Manjunath, B.S., 2006. Modeling and detection of geospatial objects using texture motifs. IEEE Trans. Geosci. Remote Sens. 44 (12), 3706–3715.

Bi, F., Zhu, B., Gao, L., Bian, M., 2012. A visual search inspired computational model for ship detection in optical satellite images. IEEE Geosci. Remote Sens. Lett. 9 (4), 749–753.

Borji, A., 2012. Boosting bottom-up and top-down visual features for saliency estimation. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012), Providence, RI, USA, June 16–June 21, pp. 438–445.

Chen, Y., Nasrabadi, N.M., Tran, T.D., 2011. Sparse representation for target detection in hyperspectral imagery. IEEE J. Sel. Topics Signal Process. 5 (3), 629–640.

Cheng, G., Han, J., Guo, L., Qian, X., Zhou, P., Yao, X., Hu, X., 2013. Object detection in remote sensing imagery using a discriminatively trained mixture model. ISPRS J. Photogrammetry Remote Sens. 85, 32–43.

Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. 32 (9), 1627–1645.

Han, J., Ngan, K.N., Li, M., Zhang, H.J., 2006. Unsupervised extraction of visual attention objects in color images. IEEE Trans. Circuits Syst. Video Technol. 16 (1), 141–145.

Han, B., Zhu, H., Ding, Y., 2011. Bottom-up saliency based on weighted sparse coding residual. In: Proceedings of the 19th ACM International Conference on Multimedia (ACM 2011), Scottsdale, AZ, USA, November 28–December 01, pp. 1117–1120.

Harel, J., Koch, C., Perona, P., 2006. Graph-based visual saliency. In: Proceedings of the 20th Neural Information Processing Systems (NIPS 2006), Vancouver, British Columbia, Canada, 4–7 December, pp. 545–552.

Hou, X., Zhang, L., 2007, June. Saliency detection: a spectral residual approach. In: Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007), Minneapolis, Minnesota, USA, 18–23 June, pp. 1–8.

Itti, L., Koch, C., Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. 20 (11), 1254–1259.

Judd, T., Ehinger, K., Durand, F., Torralba, A., 2009. Learning to predict where humans look. In: Proceedings of the 12th IEEE International Conference on Computer Vision (ICCV 2009), Kyoto, Japan, September 27–October 4, pp. 2106–2113.

Lei, Z., Fang, T., Huo, H., Li, D., 2012. Rotation-invariant object detection of remotely sensed images based on texton forest and hough voting. IEEE Trans. Geosci. Remote Sens. 50 (4), 1206–1217.

Li, Z., Itti, L., 2011. Saliency and gist features for target detection in satellite images. IEEE Trans. Image Process. 20 (7), 2017–2029.

Li, P., Xu, H., Guo, J., 2010a. Urban building damage detection from very high resolution imagery using OCSVM and spatial features. Int. J. Remote Sens. 31 (13), 3393–3409.

Li, X., Zhang, S., Pan, X., Dale, P., Cropp, R., 2010b. Straight road edge detection from high-resolution remote sensing images based on the ridgelet transform with the revised parallel-beam Radon transform. Int. J. Remote Sens. 31 (19), 5041–5059.

Li, Y., Sun, X., Wang, H., Sun, H., Li, X., 2012. Automatic target detection in high-resolution remote sensing images using a contour-based spatial model. IEEE Geosci. Remote Sens. Lett. 9 (5), 886–890.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A., 2008. Supervised Dictionary Learning. In: Proceedings of the 22th Conference on Neural Information Processing Systems (NIPS 2008), Vancouver, Canada, 8–11 December, pp. 1033–1040.

Santosh, K.D., Larry, Z., Ashish, K., Simon, B., 2010. Detecting objects using unsupervised parts-based attributes. Carnegie Mellon University Robotics Institute Technical, Report, CMU-RI-TR-11-10.

Sirmaçek, B., Ünsalan, C., 2009. Urban-area and building detection using SIFT keypoints and graph theory. IEEE Trans. Geosci. Remote Sens. 47 (4), 1156–1167.

Sirmaçek, B., Ünsalan, C., 2010. Urban area detection using local feature points and spatial voting. IEEE Geosci. Remote Sens. Lett. 7 (1), 146–150.

Sirmaçek, B., Ünsalan, C., 2011. A probabilistic framework to detect buildings in aerial and satellite images. IEEE Trans. Geosci. Remote Sens. 49 (1), 211–221.

Sun, J., Wang, Y., Zhang, Z., Wang, Y., 2010a. Salient region detection in high resolution remote sensing images. In: Proceedings of the 19th IEEE Wireless and Optical Communications Conference (WOCC 2010), Shanghai, China, 14–15 May, pp. 1–4.

Sun, X., Wang, H., Fu, K., 2010b. Automatic detection of geospatial objects using taxonomic semantics. IEEE Geosci. Remote Sens. Lett. 7 (1), 23–27.

Sun, H., Sun, X., Wang, H., Li, Y., Li, X., 2012. Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model. IEEE Geosci. Remote Sens. Lett. 9 (1), 109–113.

Tao, C., Tan, Y., Cai, H., Tian, J., 2011. Airport detection from large IKONOS images using clustered SIFT keypoints and region information. IEEE Geosci. Remote Sens. Lett. 8 (1), 128–132.

Tello, M., López-Martínez, C., Mallorqui, J.J., 2005. A novel algorithm for ship detection in SAR imagery based on the wavelet transform. IEEE Geosci. Remote Sens. Lett. 2 (2), 201–205.

Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y., 2009. Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. 31 (2), 210–227.

Xu, S., Fang, T., Li, D., Wang, S., 2010. Object classification of aerial images with bag-of-visual words. IEEE Geosci. Remote Sens. Lett. 7 (2), 366–370.

Yang, J., Yu, K., Gong, Y., Huang, T., 2009. Linear spatial pyramid matching using sparse coding for image classification. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, Florida, USA, 20–25 June, pp. 1794–1801.

Yang, M., Zhang, L., Feng, X., Zhang, D., 2011. Fisher discrimination dictionary learning for sparse representation. In: Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV 2011), Barcelona, Spain, 6–13 November, pp. 543–550.

Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W., 2008. SUN: a Bayesian framework for saliency using natural statistics. J. Vision 8 (7), 1–20.

Zhang, L., Zhang, L., Tao, D., Huang, X., 2011. A multifeature tensor for remote-sensing target recognition. IEEE Geosci. Remote Sens. Lett. 8 (2), 374–378.