Multi-class geospatial object detection and geographic image classification based on collection of part detectors

Article *in* ISPRS Journal of Photogrammetry and Remote Sensing · November 2014

DOI: 10.1016/j.isprsjprs.2014.10.002

CITATIONS
80

READS
445

4 authors:

Gong Cheng
Northwestern Polytechnical University
41 PUBLICATIONS **939** CITATIONS
SEE PROFILE

Junwei Han
Northwestern Polytechnical University
193 PUBLICATIONS **2,639** CITATIONS
SEE PROFILE

Peicheng Zhou
University of Technology Sydney
14 PUBLICATIONS **382** CITATIONS
SEE PROFILE

Kaiming Li
Emory University
275 PUBLICATIONS **3,509** CITATIONS
SEE PROFILE

ISPRS Journal of Photogrammetry and Remote Sensing 98 (2014) 119–132

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Multi-class geospatial object detection and geographic image classification based on collection of part detectors

Gong Cheng, Junwei Han *, Peicheng Zhou, Lei Guo

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

ARTICLE INFO

Article history:
Received 23 April 2014
Received in revised form 14 October 2014
Accepted 14 October 2014

Keywords:
Geospatial object detection
Geographic image classification
Very-high-resolution (VHR)
Remote sensing images
Part-based model
Collection of part detectors (COPD)

ABSTRACT

The rapid development of remote sensing technology has facilitated us the acquisition of remote sensing images with higher and higher spatial resolution, but how to automatically understand the image contents is still a big challenge. In this paper, we develop a practical and rotation-invariant framework for multi-class geospatial object detection and geographic image classification based on collection of part detectors (COPD). The COPD is composed of a set of representative and discriminative part detectors, where each part detector is a linear support vector machine (SVM) classifier used for the detection of objects or recurring spatial patterns within a certain range of orientation. Specifically, when performing multi-class geospatial object detection, we learn a set of seed-based part detectors where each part detector corresponds to a particular viewpoint of an object class, so the collection of them provides a solution for rotation-invariant detection of multi-class objects. When performing geographic image classification, we utilize a large number of pre-trained part detectors to discovery distinctive visual parts from images and use them as attributes to represent the images. Comprehensive evaluations on two remote sensing image databases and comparisons with some state-of-the-art approaches demonstrate the effectiveness and superiority of the developed framework.

1. Introduction

In recent years, the rapid development of remote sensing technology has increasingly facilitated us the acquisition of remote sensing images with higher and higher spatial resolution, which gives researchers the new opportunity for advancing the interpretation of remote sensing images, especially with regard to automated analysis and understanding of the meanings and contents of remote sensing images. Geospatial object detection and scene-level geographic image classification, as two fundamental yet challenging research aspects of remote sensing analysis, have recently attracted considerable attention and have been extensively studied.

Automated object detection in remote sensing images is a core requirement for high-level scene understanding and semantic information extraction. A number of recent works have proposed various methods for different object detection tasks. For example, Bhagavathy and Manjunath (2006) developed a method to learn a Gaussian mixture model from training samples using texture motifs and then detected compound objects based on the learned model. Grabner et al. (2008) developed an online boosting algorithm for car detection from large-scale aerial images. Ünsalan and Sirmacek (2012) explored a road network detection system which consists of probabilistic road center detection, road shape extraction, and graph-theory-based road network formation. Aytekin et al. (2013) proposed a novel airport runway detection method by using an Adaboost learning algorithm employed on a large set of textural features. In addition, the detection methods for some other object classes such as ships (Bi et al., 2012; Corbane et al., 2010; Zhu et al., 2010), buildings (Aytekın et al., 2012; Sirmacek and Ünsalan, 2011), and landslide (Cheng et al., 2013a; Martha et al., 2011), have also been explored.

Although the topic of geospatial object detection has been deeply investigated, most of the current object detection methods are still dominated by the detection of a single object class and fewer concerns have been given to scalable multi-class object detection. Furthermore, the features extracted in most existing individual detectors are customized for the particular type of objects, preventing them from scaling up to deal with the simultaneous detection of a large number of object classes. Generally, a large-scale remote sensing image always contains multiple object classes

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

http://dx.doi.org/10.1016/j.isprsjprs.2014.10.002

instead of only a single one, so it is a very important issue to develop a scalable multi-class object detection method for scene understanding and semantic information extraction where many object classes need to be identified.

Scene-level geographic image classification also plays an important role for diverse applications of remote sensing images analysis, such as land-use/land-cover (LULC) image classification (Xu et al., 2010; Yang and Newsam, 2010, 2011), semantic interpretations of images (Aksoy et al., 2005; Văduva et al., 2013), geographic image retrieval (Schroder et al., 2000; Shyu et al., 2007; Yang and Newsam, 2013), and forest type mapping (Kim et al., 2009). In recent years, the bag-of-features (BoF) model (Csurka et al., 2004; Li and Perona, 2005) has been among the most successful models for scene-level image categorization tasks. This group of methods represents an image as a collection of unordered local features, quantizes them into discrete visual words, and then computes a compact histogram representation for image classification. Nevertheless, since the BoF method disregards all information about the spatial layout of the features, it is incapable of capturing the shape information or locating an object. By overcoming this problem, one successful extension of the BoF model is spatial pyramid matching (SPM) (Lazebnik et al., 2006), which partitions the image into increasingly finer spatial sub-regions and computes histograms of local features from each sub-region. Although the resulted "spatial pyramid" is a computationally efficient extension of the unordered BoF representation and has shown very promising performance, it only characterizes the absolute location while ignores the relative spatial arrangement of the visual words in an image, which limits the descriptive ability of the image representation. Accordingly, Yang and Newsam proposed two novel image representation approaches termed spatial co-occurrence kernel (SCK) (Yang and Newsam, 2010) and spatial pyramid co-occurrence kernel (SPCK) (Yang and Newsam, 2011), respectively. The former method considered the relative spatial arrangement of the visual words while the latter one characterized both the absolute and relative spatial layout of an image. These two approaches have been shown to perform better on a challenging 21-class LULC data set (Yang and Newsam, 2010, 2011) than BoF method and SPM.

A common characteristic of those above-mentioned methods (Csurka et al., 2004; Lazebnik et al., 2006; Li and Perona, 2005; Yang and Newsam, 2010, 2011) is that nearly all of them are based on some kind of low-level image features, such as scale invariant feature transform (SIFT) (Lowe, 2004), color histogram, and texture. Although low-level image features have proven to be effective for some moderate visual recognition tasks, they may not be powerful for many challenging recognition tasks. For example, Fig. 1 shows four remote sensing images from a publicly available 21-class LULC data set (Yang and Newsam, 2010, 2011). A classification method based on texture statistics or color histogram would easily confuse all the four, especially the last two images as the same LULC class. Even if we use some contextual information such as spatial layout of the whole image, it is still difficult to differentiate the third "sparse residential" class from the fourth "tennis court" class. However, humans would classify the third and the fourth images as belonging to different LULC classes based on the discriminative visual parts (buildings and tennis court) and the high-level semantic concepts pertaining to the classes. This example and our visual experiences suggest that a straightforward way to recognize many complex real-world scenes would be discriminative visual parts-based method.

With the rapid advance of remote sensing technology, more and more high-resolution or very-high-resolution (VHR) remote sensing images have been providing us detailed spatial and textural information. Thanks to the higher spatial resolution, a greater range of objects and recurring spatial patterns can be observed than ever before, and even individual objects, such as cars, trees, and buildings, have become recognizable. This provides us new opportunity for further advancing the performance of automatic image interpretation by adopting object-guided image analysis scheme, and this can be easily achieved by training a great deal of discriminative visual parts detectors.

More recently, part model-based methods have achieved state-of-the-art results for object detection (Bourdev and Malik, 2009; Felzenszwalb et al., 2010; Malisiewicz et al., 2011) and image classification (Juneja et al., 2013; Li et al., 2013; Singh et al., 2012; Sun and Ponce, 2013) on natural scene (non-overhead) images, which represent an object category or an image by a number of important visual parts. Their success is largely owing to the introduction of the notion of "part detector", a discoverer of mid-level visual elements, or a linear support vector machine (SVM) classifier that can explicitly capture the locations, scales, and appearances of some discriminative visual parts. These distinctive visual parts can better complement or substitute low-level image features such as SIFT (Lowe, 2004). However, very different from natural scene images, in which objects are typically upright due to the Earth's gravity and the orientation variations across images are generally small, remote sensing images are taken overhead, in which geospatial objects usually have arbitrary orientations. Consequently, although part model-based methods have achieved impressive success on natural scene images, these methods cannot be directly used to detect objects and recurring spatial patterns from remote sensing images because they are difficult to effectively handle the problem of targets rotation variation.

Guided by this observation and motivated by the idea of using a large number of part detectors to explore a possible solution to address the rotation variation problem, in this paper, we develop an effective and rotation-invariant framework based on collection of part detectors (COPD) for multi-class geospatial object detection and geographic image classification. To be specific, the COPD is composed of a set of representative and discriminative part detectors, where each part detector is used for the detection of objects or recurring spatial patterns within a certain range of orientation. Here, we use the word "part" in its very general form—while smaller pieces of objects are parts, recurring visual patterns are parts, so are the whole objects in different viewpoints.

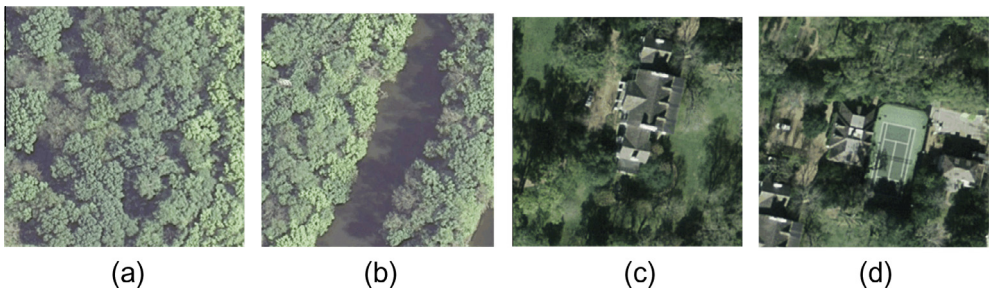

Fig. 1. Four images from a publicly available 21-class LULC data set (Yang and Newsam, 2010, 2011). (a) Forest. (b) River. (c) Sparse residential. (d) Tennis court.

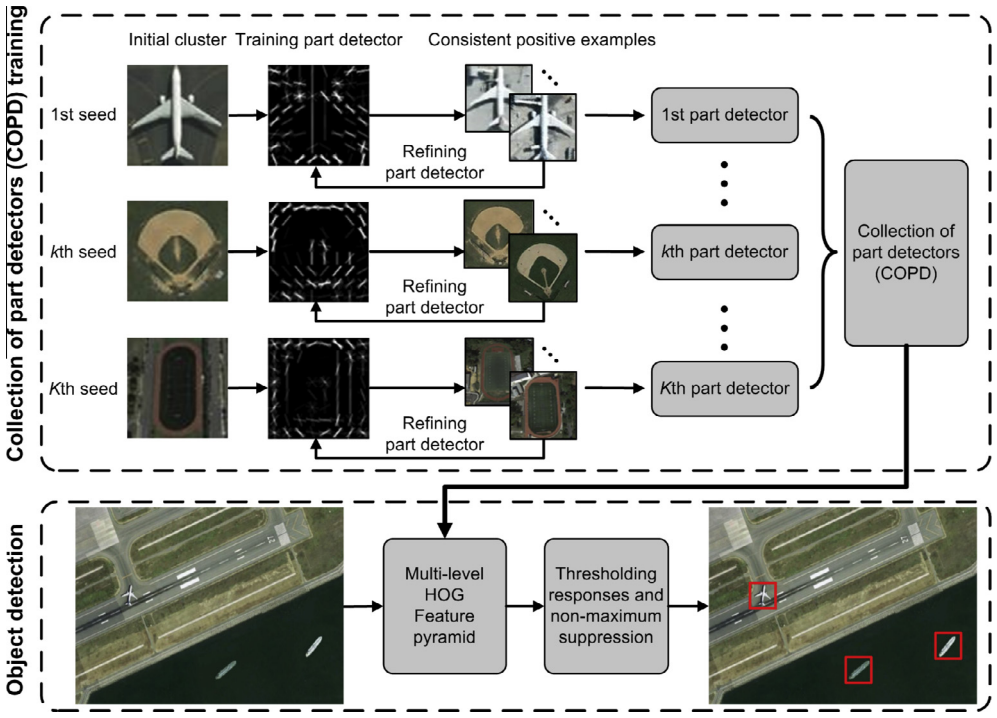

Fig. 2. Overview of the developed COPD-based multi-class geospatial object detection framework.

To sum up, the primary contribution of this paper is fourfold. First, by extending the notion of "part detector" to high-resolution remote sensing images analysis, we introduce a practical and rotation-invariant framework for multi-class geospatial object detection and geographic image classification based on collection of part detectors, where each part detector is used for the detection of objects or recurring spatial patterns within a certain range of orientation. Second, when training part detectors for multi-class object detection, we improve the traditionally used exemplar-SVM detector (Malisiewicz et al., 2011) training process by alternatively refining part detectors and incorporating consistent positive examples for each exemplar. The quantitative comparison results on first-of-its-kind 10-class objects data set, as shown in Fig. 7 and Table 3, demonstrate huge performance gain of our method compared with state-of-the-art approaches. Third, by taking advantage of the technology of mid-level visual elements discovery (Juneja et al., 2013; Li et al., 2013; Singh et al., 2012; Sun and Ponce, 2013), we achieve an effective image representation method for VHR geographic image classification by using discriminative visual parts as attributes, which provides a more informative description of an image. As shown in Fig. 12 and Table 5, superior and encouraging results are obtained on a publicly available 21-class LULC benchmark for image classification. To the best of our knowledge, this result is the best on this data set, which adequately shows the superiority and effectiveness of the developed framework. Fourth, a high-spatial-resolution remote sensing images data set containing 10-class objects (airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle) is constructed and will be made publicly available to other researchers.[1] We anticipate this data set will help other researchers to conduct further study or compare different algorithms.

The rest of the paper is organized as follows. Section 2 describes a COPD-based multi-class geospatial object detection framework and reports comparative experimental results on a high resolution remote sensing image data set. Section 3 details a COPD-based geographic image classification framework and gives comparative experimental results on a publicly available 21-class LULC benchmark. Finally, conclusions are drawn in Section 4.

2. COPD-based multi-class geospatial object detection

2.1. Framework overview

Fig. 2 gives an overview of the developed COPD-based multi-class geospatial object detection framework. It is mainly composed of two stages: COPD training and object detection. In the COPD training phase, we first pick a set of representative seeds to serve as initial clusters, where each seed corresponds to a particular viewpoint of an object class and each cluster corresponds to a part detector needed to be trained. Then, we train a set of part detectors using an iterative procedure (Bourdev and Malik, 2009; Cheng et al., 2013b; Felzenszwalb et al., 2010; Singh et al., 2012) that alternates between refining detectors and incorporating consistent positive examples for each seed from training images. Given K seeds, we can finally obtain a COPD that is composed of K seed-based part detectors. Since each part detector corresponds to a particular viewpoint of an object class, the collection of them could provide an effective solution for rotation-invariant and simultaneous detection of multi-class geospatial objects. In the object detection stage, given a new test image, we first run all detectors simultaneously on the input image, in Histograms of Oriented Gradients (HOG) (Dalal and Triggs, 2005) feature pyramid space, to obtain the response and potential object class for each sliding-window. Then, multi-class object detection is implemented by thresholding the responses and eliminating repeated detections via non-maximum suppression (Bourdev and Malik, 2009; Felzenszwalb et al., 2010).

2.2. Framework details

2.2.1. COPD training

When training a COPD for multi-class object detection, the input is composed of a "positive image dataset" P in which each

[1] http://pan.baidu.com/s/1c0w8h3q.

Table 1
The object sizes, object numbers from optimizing set and object numbers from testing set.

Object classes	Object sizes (pixels)	Object numbers from optimizing set	Object numbers from testing set
Airplane	33 × 33–129 × 129	146	561
Ship	40 × 40–128 × 128	62	214
Storage tank	34 × 34–103 × 103	63	326
Baseball diamond	49 × 49–179 × 179	83	246
Tennis court	45 × 45–127 × 127	121	317
Basketball court	52 × 52–179 × 179	32	96
Ground track field	192 × 192–418 × 418	36	102
Harbor	68 × 68–222 × 222	62	118
Bridge	98 × 98–363 × 363	36	81
Vehicle	42 × 42–91 × 91	77	155

image contains at least one target of interest and a "negative image dataset" N in which all images do not contain any targets of the given object classes. The COPD training is performed in terms of the following steps (Felzenszwalb et al., 2010; Juneja et al., 2013; Malisiewicz et al., 2011; Singh et al., 2012):

(1) Pick a set of seeds from P to serve as initial clusters (seeds generation will be described in Subsection 2.3.2), where each seed corresponds to a particular viewpoint of an object class. Given K seeds, we have K part detectors to be trained.
(2) Train a linear SVM classifier $\Gamma_k = (w_k, b_k)(k = 1, \ldots, K)$ for each cluster, in HOG (Dalal and Triggs, 2005) feature space, using image patches within the cluster as positive examples and all hard negative examples of N as negative examples. It is noted that when we train the SVM classifiers first time, this can be seen as a special situation of Malisiewicz et al. (2011) because each cluster contains one image patch only, i.e. the picked seed. For each cluster, learning the parameters w_k and b_k amounts to optimizing the following objective function:

$$(w_k, b_k)^* = \arg\min_{(w_k, b_k)} \left\{ \frac{1}{2}\|w_k\|^2 + \kappa \sum_{x^+ \in X_k^+} h\left(w_k^T \Phi(x^+) + b_k\right) + \kappa \sum_{x^- \in X_k^-} h\left(-w_k^T \Phi(x^-) - b_k\right) \right\} \quad (1)$$

where X_k^+ and X_k^- denote the sets of positive examples and negative examples of kth cluster. $\Phi(x^+)$ and $\Phi(x^-)$ denote the feature vectors of positive example x^+ and negative example x^- obtained by concatenating all the HOG feature vectors within the examples. $h(\tau) = \max(0, 1 - \tau)$ is the standard hinge loss function that allows us to use hard negative mining technique to cope with millions of negative examples (Bourdev and Malik, 2009; Felzenszwalb et al., 2010; Malisiewicz et al., 2011). κ is a constant and we set $\kappa = 0.1$ in our work.

(3) Run $\Gamma = \{\Gamma_k\}_{k=1}^K$ on P to obtain new clusters by selecting the top-m high-scoring patches for each part detector. In our work, we set $m = 5$ to keep each cluster having a high purity.
(4) Repeat the steps of (2) and (3) $L1$ iterations until the maximum measured by Average Precision (AP) (Everingham et al., 2010) is reached, thus we obtain an updated COPD $\Gamma = \{\Gamma_k\}_{k=1}^K$ with K part detectors. Here, it should be pointed out that the COPD is trained on data sets of P and N, while the parameter optimization of $L1$ is performed on another data set with pre-generated ground truth which are called "optimizing set" hereinafter. We will provide a detailed description of the data sets and report the effect of the parameter $L1$ in Subsection 2.3.4.

2.2.2. Object detection

Given a test image I, its HOG (Dalal and Triggs, 2005) feature pyramid $H(I)$ is first constructed. Then, for each sliding-window S, we run all detectors $\Gamma = \{\Gamma_k\}_{k=1}^K$ on $H(I)$ to obtain its response $R(S)$ and potential object class $O(S)$. $R(S)$ is defined as the maximum response of all part detectors:

$$R(S) = \max_{(w_k, b_k) \in \Gamma} (w_k^T \Phi(S) + b_k) \quad (2)$$

where $\Phi(S)$ denotes the feature vectors of sliding-window S by concatenating all the HOG feature vectors within it, $w_k^T \Phi(S) + b_k$ is the response of sliding-window S of detector Γ_k. $O(S)$ is defined by the class of part detector with the maximum response. Finally, multi-class object detection is implemented by thresholding the responses using a threshold ρ (the optimal threshold can be derived from the highest F1-measure), and each hypothesis is defined by a response, a potential object class and a bounding box. However, in practice, when we use the above described detection approach solely, a number of sliding-windows near each instance of an object are likely to be detected as the targets, which results in multiple overlapping detections for a single object. We therefore apply non-maximum suppression (Bourdev and Malik, 2009; Felzenszwalb et al., 2010) to eliminate repeated detections. In brief, the bounding boxes are sorted by their responses, and we greedily

Fig. 3. All 45 seeds used in our work for 10-class object detection.

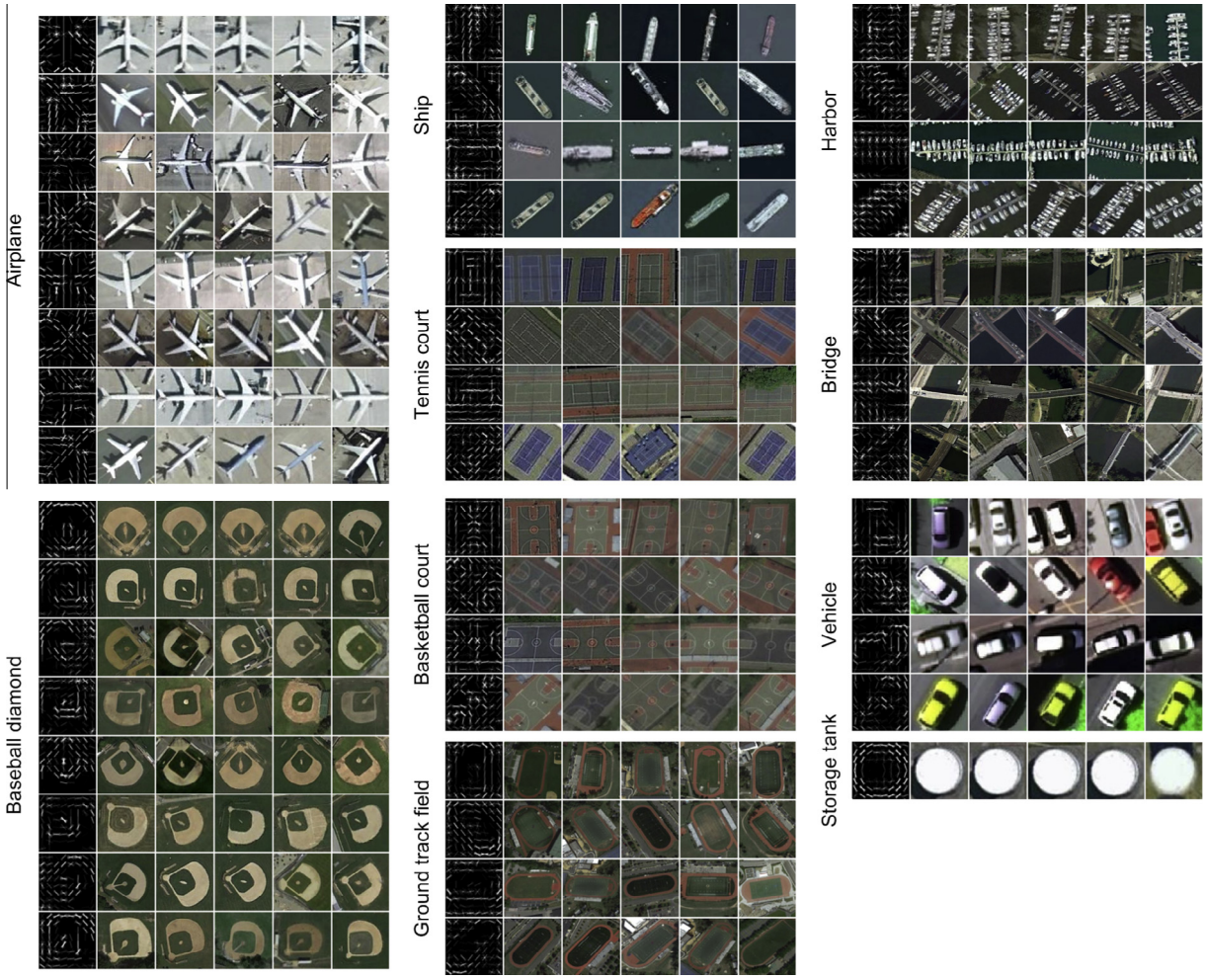

Fig. 4. The visualization of weight vectors of 45 detectors for airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle classes, respectively. Their top-5 high-scoring positives from the training data set are also shown subsequently. We have resized these positives to 60 × 60 pixels for visualization.

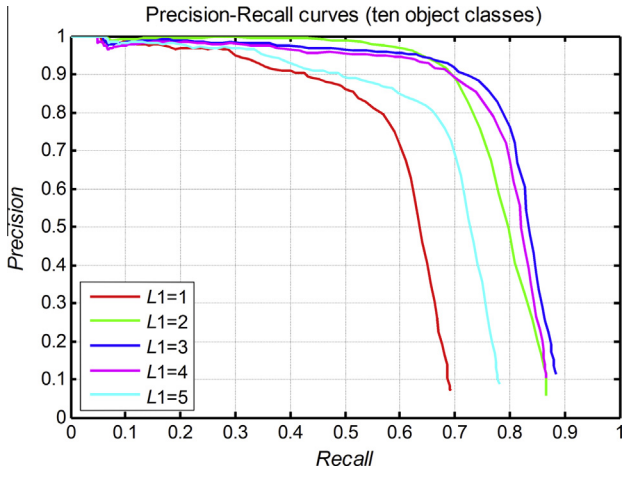

Fig. 5. Precision–Recall curves obtained by varying the values of $L1$.

Table 2
Performance comparisons of different $L1$ in terms of AP.

$L1$	1	**2**	3	4	5
AP	0.5929	0.7798	0.8044	0.7838	0.6798

select the highest scoring ones while removing those that are at least 50% covered by a previously selected bounding box.

2.3. Experiments

2.3.1. Data set description

In theory, the developed multi-class object detection framework can detect a large number of classes of geospatial objects. However, in our experiments, we used the task of detection of ten different types of objects to evaluate the performance of the developed framework. These ten classes of objects are airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle.

We collected 715 high-spatial-resolution color images from Google Earth and 85 very-high-spatial-resolution pansharpened color infrared (CIR) images from Vaihingen data set (Cramer, 2010) used for our evaluations, where the spatial resolution of Google Earth images ranges from 0.5 m to 2 m and the spatial resolution of CIR images is 0.08 m. The Vaihingen data was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) (Cramer, 2010): http://www.ifp.uni-stuttgart.de/dgpf/DKEPAllg.html. We divided these images into four independent datasets: a "negative image set" containing 150 images, a "positive image set" containing 150 images, an "optimizing set" containing 150 images, and a testing set containing 350

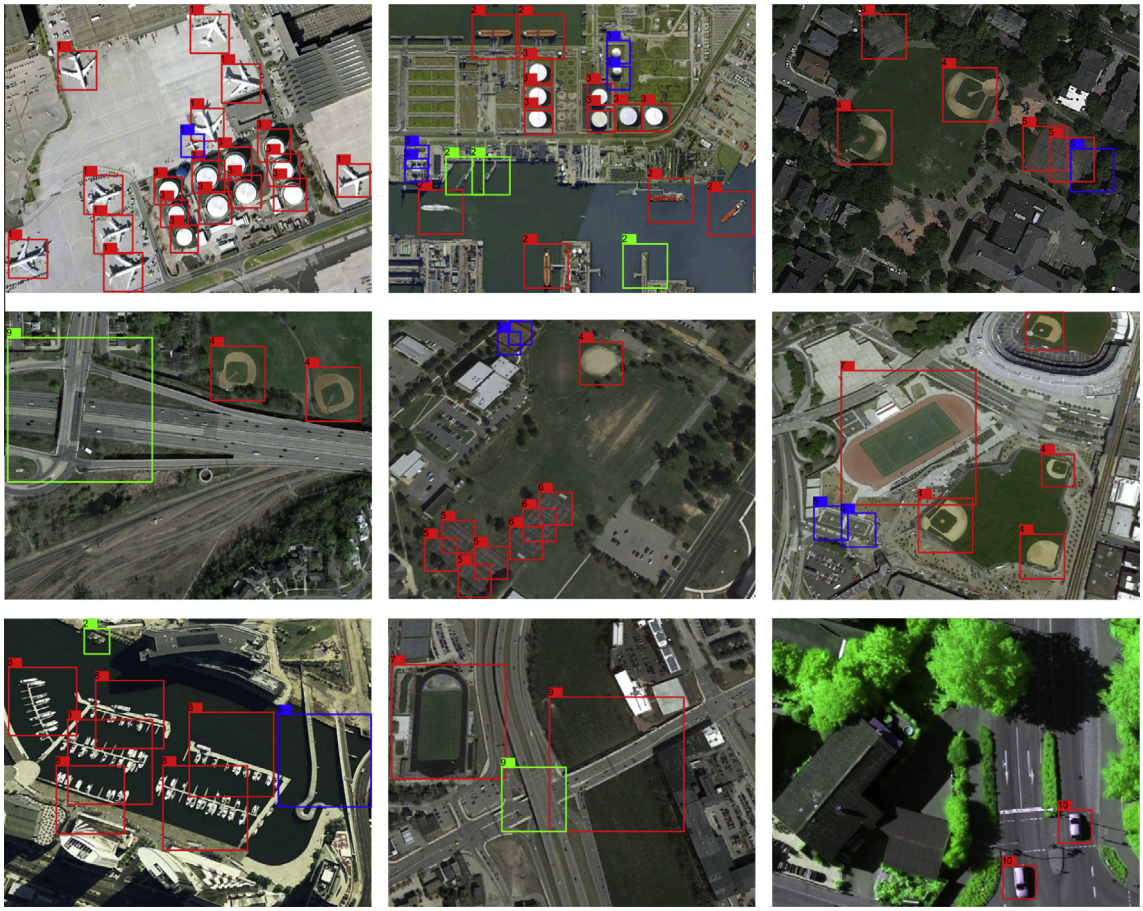

Fig. 6. A number of multi-class object detection results by using the developed framework.

images. All images from the first set do not contain any targets of the given object classes and each image from the last three sets contains at least one target to be detected. The "negative image set" and "positive image set" were used for the COPD training, the "optimizing set" was used for parameter optimization and the "testing set" was used for testing the performance of the developed framework. We labeled ground truths from the "optimizing set" and the "testing set", respectively. The detailed object sizes, object numbers from optimizing set, and object numbers from testing set of ten different object classes are listed in Table 1.

2.3.2. Seeds generation

Here, it should be pointed out that the seeds serving as initial clusters should be representative and have different orientations, which can be obtained by manually labeling a representative sample for each object class from the "positive image set", aligning them, and then rotating them with a certain angle. Specifically, for our 10-class object detection task, given ten manually labeled samples (each sample for each object class), we first align each of them to an unified orientation (e.g. approximately vertical in our implementation) and then rotate airplane and baseball diamond samples in the step of 45° from 0° to 360°, rotate samples of ship, tennis court, basketball court, ground track field, harbor, bridge, and vehicle in the step of 45° from 0° to 180° because their shapes are bilaterally symmetry, and perform no rotation for storage tank sample because its shape is circular. In this way, we can obtain eight seeds for each manually labeled sample of airplane and baseball diamond, four seeds for each manually labeled sample of ship, tennis court, basketball court, ground track field, harbor, bridge, and vehicle, and one seed for each manually labeled sample of storage tank. Fig. 3 illustrates the total 45 seeds used in our work for 10-class object detection.

Using the COPD training procedure as described in Subsection 2.2.1 and the 45 seeds as shown in Fig. 3, we trained a COPD consisting of 45 detectors for airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle classes on our training data set. Fig. 4 shows the visualization of weight vectors w_k of 45 detectors, where brighter "pixel" represents bigger weight and vice versa. Their corresponding top-5 high-scoring positives from the training data set are also shown subsequently.

2.3.3. Evaluation criterions

We consider a detection to be correct if its bounding box overlaps more than 50% with the ground truth bounding box, otherwise the detection is considered as a false positive. In addition, if several bounding boxes overlap with a same single ground truth bounding box, only one is considered as true positive and the others are considered as false positives. We adopted the standard Precision–Recall curve (PRC) (Buckland and Gey, 1994) and AP (Everingham et al., 2010) to quantitatively evaluate the performance of an object detection system. The *Precision* measures the fraction of detections that are true positives and the *Recall* measures the fraction of positives that are correctly identified. AP computes the average value of *Precision* over the interval from *Recall* = 0 to *Recall* = 1, i.e. the area under the PRC, so the higher the AP value is, the better the performance and vice versa. Let *TP*, *FP*, and *NP* denote the number of true positives, the number of false positives, and the number of total positives. The Precision and Recall can be formulated as:

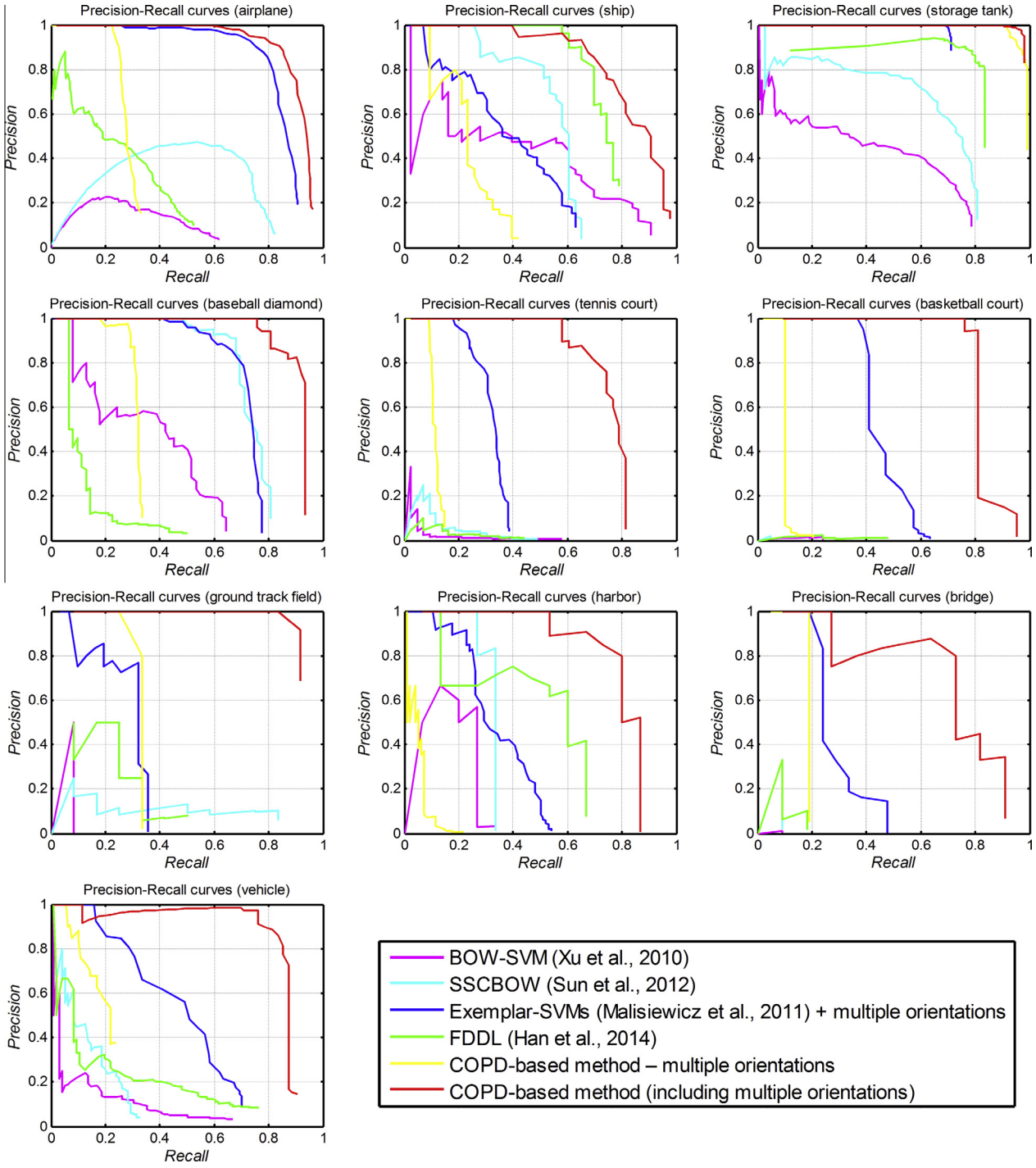

Fig. 7. Precision–Recall curves of the developed framework and some state-of-the-art approaches for airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle classes respectively.

$$Precision = TP/(TP + FP) \tag{3}$$

$$Recall = TP/NP \tag{4}$$

2.3.4. Experimental setting

In the implementation of multi-class object detection, to address the problem that the sizes of targets may be different in images, each image is represented by an 15-level HOG (Dalal and Triggs, 2005) feature pyramid and each octave contains five levels (i.e. for l th level, the sub-sampling factor is $2^{(l-1)/5}$). We follow the construction in Dalal and Triggs (2005) to extract the HOG feature for each pyramid level. Specifically, we partition the image at each pyramid level into non-overlapping cells of 6×6 pixels and use nine orientation bins to accumulate a one-dimensional histogram of gradient orientations over pixels in each cell. Then, each 2×2 neighbourhood of cells is grouped into one block (with a stride of one cell) and a robust normalization process based on 2-norm is run on each block to provide greater invariance to local illumination and spatial deformation, which finally forms a 36-dimensional HOG feature vector. Rather than using the 36-dimensional vector directly, in this work we project it onto a lower 31-dimensional

Table 3
Performance comparisons of six different methods in terms of AP values.

	BOW-SVM (Xu et al., 2010)	SSCBOW (Sun et al., 2012)	Exemplar-SVMs (Malisiewicz et al., 2011) + multiple orientations	FDDL (Han et al., 2014)	COPD-based method – multiple orientations	COPD-based method (including multiple orientations)
Airplane	0.0894	0.2848	0.8411	0.2310	*0.2807*	0.8911
Ship	0.3695	0.5212	0.3704	0.5218	*0.2100*	0.8173
Storage tank	0.3692	0.5961	0.7087	0.6503	*0.9782*	0.9732
Baseball diamond	0.3378	0.7159	0.7091	0.1058	*0.2995*	0.8938
Tennis court	0.0133	0.0283	0.3145	0.0147	*0.1100*	0.7327
Basketball court	0.0022	0.0024	0.4378	0.0056	*0.0851*	0.7341
Ground track field	0.0208	0.0926	0.2457	0.1089	*0.2417*	0.8299
Harbor	0.1354	0.2544	0.3307	0.4132	*0.0364*	0.7339
Bridge	0.0004	0.0152	0.2414	0.0225	*0.1429*	0.6286
Vehicle	0.0781	0.1148	0.4600	0.1766	*0.1669*	0.8330
Mean AP values	0.1416	0.2626	0.4659	0.2250	*0.2551*	0.8068

space as described by Felzenszwalb et al. (2010) and Singh et al. (2012). In addition, the size of each part detector is 8×8 blocks, i.e. an 8×8 HOG descriptors. Consequently, the sizes of image patches that each part detector can detect are 60×60, 69×69, 79×79, 91×91, 104×104, 120×120, 138×138, 158×158, 182×182, 209×209, 240×240, 276×276, 317×317, 364×364, and 418×418, respectively, which correspond to 15 different image scales.

In the developed framework, $L1$ used for COPD training is a critical parameter associated with object detection performance, so we constructed experiments on the "optimizing set" to evaluate how the performance is affected by it. Fig. 5 and Table 2 show the PRC and AP of all ten object classes, respectively, obtained by varying $L1$. As can be seen from them, $L1$ influences the detection result moderately. Specifically, the detection results measured by AP were improved in a certain range with the increase of $L1$, and then dropped off. When $L1 = 3$, the best performance can be achieved. Consequently, we empirically set $L1 = 3$ in our multi-class object detection evaluations.

2.3.5. Experimental results and comparisons

Using the pre-trained COPD, as illustrated in Fig. 4, we performed 10-class object detection on our testing dataset which contains 561 airplane targets, 214 ship targets, 326 storage tank targets, 246 baseball diamond targets, 317 tennis court targets, 96 basketball court targets, 102 ground track field targets, 118 harbor targets, 81 bridge targets, and 155 vehicle targets. Fig. 6 shows a number of multi-class object detection results by using the developed framework, in which the correctly detected targets, false alarms, and miss alarms are labeled by red, green, and blue boxes, respectively. The ten numbers of one to ten on the boxes denote the object classes of airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle. As can be seen from Fig. 6, although the objects of various classes have different orientations and sizes, the framework has successfully detected and located most of them.

In addition, to quantitatively evaluate the developed work, we compared it with four state-of-the-art object detection methods (Han et al., 2014; Malisiewicz et al., 2011; Sun et al., 2012; Xu et al., 2010). The method of Xu et al. (2010) is based on bag-of-words (BOW) feature and SVM classifier, which is called 'BOW-SVM' in this paper. The method of Sun et al. (2012) is based on spatial sparse coding bag-of-words and SVM classifier, which is called 'SSCBOW' in this paper. The method of Malisiewicz et al. (2011) is based on a set of exemplar-based SVMs, which is called 'Exemplar-SVMs' in this paper. The method of Han et al. (2014) is based on visual saliency modeling and Fisher discrimination dictionary learning, which is called 'FDDL' in this paper. For fair comparison, (1) We adopted the same training dataset and test dataset for various approaches, and used the same seeds as our method to train Exemplar-SVMs (Malisiewicz et al., 2011); (2) We implemented all comparison methods by adopting multi-scale scanning window scheme for each test image which was similar to our multi-level HOG feature pyramid. Following the work of Sun et al. (2012) and Xu et al. (2010), the vocabulary size was set to 400 and 450 for SSCBOW and BOW-SVM, respectively; (3) To address the rotation variation problem for traditionally used Exemplar-SVMs (Malisiewicz et al., 2011), we adopted the same scheme as our method to train a collection of Exemplar-SVMs by addressing multiple orientations for each object class. This comparison method is called 'Exemplar-SVMs+multiple orientations' in this paper. Moreover, to further demonstrate our COPD-based method is rotation-invariant for multi-class geospatial object detection, we also reported the results of not addressing multiple orientations while only using our improved exemplar-SVM detectors. This variant method is called 'COPD-based method – multiple orientations' in this paper.

Fig. 7 and Table 3 show the quantitative comparison results of six different methods, measured by PRC and AP values for each object class, respectively. As can be seen from Fig. 7 and Table 3, (1) Our COPD-based method outperforms all comparison approaches for all ten object classes in terms of AP and our COPD-based method also improves the second-best method by 73.17% in terms of mean AP; (2) For nine of these ten object classes (except for ship class), with the same *Recall*, the *Presision* of our COPD-based method is bigger than any of the comparison methods, which means that the false alarm rate (i.e. $1 - Presision$) of our method is lower with the same true positives; (3) For eight of these ten object classes (except for ship and vehicle classes),with the same *Presision*, the *Recall* of our COPD-based method is bigger that any of the comparison methods, which shows that our method can detect more actual targets with the same false alarm rate; (4) For seven of these ten object classes (except for ship, harbor and bridge classes), the false alarm rate of our COPD-based method is nearly zero before *Recall* is bigger than 0.6, which is highly competitive compared to the other four state-of-the-art approaches; (5) All four comparison methods are only effective for some specific object classes, whereas our method is adequate to all ten object classes. These adequately demonstrate that our framework is highly competitive compared to four state-of-the-art object detection methods.

The superior comparison results of our developed framework can be explained from the following five aspects: (1) The core idea of BOW-SVM method (Xu et al., 2010) is that it represents each image patch as a histogram of so-called visual words generated by k-means algorithm. This method is found robust against spatial variations but it ignores the spatial contextual relationships among the local features, so it is only effective to detect objects with simple shapes, such as ship targets, storage tank targets, and baseball

diamond targets; (2) Similar to BOW-SVM method, SSCBOW method (Sun et al., 2012) also represents each image patch as a histogram of visual words, but in which sparse coding is introduced to replace *K*-means algorithm for visual words encoding. This new spatial encoding strategy not only represents the relative position of the local features but also has the ability to encode the geometric information of an object, so SSCBOW method obtained better performance compared with BOW-SVM method. However, the detection results of these two methods depend largely on the extracted keypoints or local features such as SIFT descriptors. For those objects (e.g. tennis courts and basketball courts characterized by side lines, center lines, and goal lines) from which enough and discriminative keypoints are difficult to extract, the detection performances of these two methods are severely limited; (3) The method of 'Exemplar-SVMs (Malisiewicz et al., 2011) + multiple orientations' is based on training an individual linear SVM classifier for every selected exemplar in HOG feature space. Since each of these Exemplar-SVMs is defined by a single positive instance, each detector is quite specific to its exemplar and has poor generalization because of the appearances variation and deformation of objects; (4) In FDDL method, sparse representation based classification strategy is adopted to perform multi-class object detection, in which each image patch is described by a few representative atoms of a learned dictionary in a low-dimensional manifold. Unfortunately, as image patches need to be down-sampled to adapt the size of atoms, some critical and discriminative features of them (e.g. the straight lines and arc in tennis courts and basketball courts) are reduced and even removed. This fatal operation has significantly degenerated the detection accuracy; (5) Our developed framework performs object detection using a collection of part detectors derived from a set of representative seeds, where each detector corresponds to a particular viewpoint of an object class and is trained using an iterative procedure that alternatively refines part detectors and incorporates consistent positive examples for each seed from the training images. On the one hand, the incorporation of consistent positive examples could guarantee the learned detectors have good generalization and therefore can effectively handle object deformations and appearance variations compared to traditionally used Exemplar-SVMs (Malisiewicz et al., 2011). On the other hand, since each part detector corresponds to a particular viewpoint of an object class, the collection of them could provide an effective solution for rotation-invariant and simultaneous detection of multi-class geospatial objects. Consequently, our method can obtain promising results compared to the aforementioned four state-of-the-art approaches.

3. COPD-based geographic image classification

3.1. Framework overview

The flowchart of our COPD-based geographic image classification is illustrated in Fig. 8. It is mainly composed of two stages: COPD training and image classification. In the COPD training stage, given an image database, we first train class-specific part detectors for each image class based on class labels in a weakly supervised fashion. This can be achieved by sampling a large number of image patches from the positive training images, clustering them, and alternating between training discriminative part detectors and refining clusters (Bourdev and Malik, 2009; Felzenszwalb et al., 2010; Singh et al., 2012). Then, all part detectors of each image class are combined to generate a complete COPD. In the image classification stage, we first use the trained COPD to detect mid-level visual elements (i.e. discriminative image patches) from each image and represent the image as a feature vector of the responses of the top-*J* high-scoring image patches, which can provide more informative description of the image. Then, we train a linear one-vs-all SVM classifier for each image class by treating the images of the chosen class as positive instances and the rest images as negative instances. Finally, a test image is assigned to the label of the classifier with the highest response.

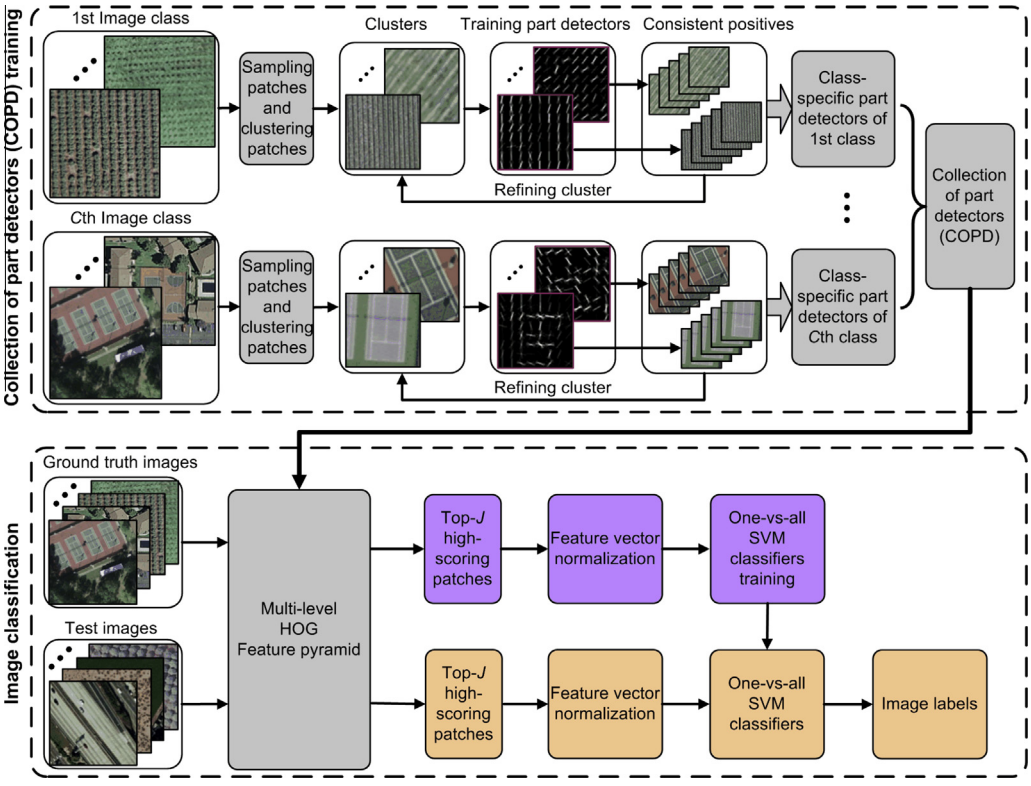

Fig. 8. Flowchart of the developed COPD-based geographic image classification framework.

3.2. Framework details

3.2.1. COPD training

Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_C\}$ denote the set of C image classes of an image database. We first learn class-specific part detectors $\Gamma^c = \{\Gamma_k^c\}_{k=1}^{K_c}$ for each image class $\omega_c (c = 1, \ldots, C)$, where K_c is the total number of part detectors and Γ^c should be representative for class ω_c and be discriminative against classes $(\Omega - \omega_c)$. Then, all part detectors of each class are combined to generate a complete COPD $\Gamma = \{\Gamma^c\}_{c=1}^{C}$. The training of Γ^c for class ω_c is performed in terms of the following steps (Felzenszwalb et al., 2010; Singh et al., 2012; Sun and Ponce, 2013):

(1) Construct "positive image dataset" P_c and "negative image dataset" N_c, where P_c is composed of the images of class ω_c and N_c is composed of the images of the classes $(\Omega - \omega_c)$.
(2) Randomly crop a large number of image patches from all images in P_c at different image scales, discard highly overlapping patches, perform standard k-means clustering over these image patches in HOG (Dalal and Triggs, 2005) feature space, and then retain sufficiently large clusters with size of 10 or more. The cluster number is adaptively set to be one tenth of the total number of sampled image patches in our work.
(3) Train a linear SVM classifier $\Gamma_k^c = (w_k^c, b_k^c)(k = 1, \ldots, K_c)$ for each cluster in HOG (Dalal and Triggs, 2005) feature space, using image patches within the cluster as positive examples and all hard negative examples of N_c as negative examples. Learning the parameters w_k^c and b_k^c amounts to optimizing the similar objective function as illustrated in Eq. (1).
(4) Run $\Gamma^c = \{\Gamma_k^c\}_{k=1}^{K_c}$ on P_c to form new clusters from the top-m high-scoring patches for each part detector. In our implementation, we set $m = 10$ to keep each cluster having a high purity.
(5) Repeat the steps of (3) and (4) $L2$ iterations until the maximum measured by image classification accuracy is reached, thus we can obtain a updated class-specific part detectors Γ^c for image class ω_c.

In our implementation, the aforementioned training procedure comes to a maximum after 4 iterations. We will report the effect of the parameter $L2$ in subsection 3.3. Using the above procedure, we trained five COPDs for all five held-out sets on a publicly available 21-class LULC data set (data set will be described in subsection 3.3.1). The total numbers of part detectors $K = \sum_{c=1}^{C} K_c$ on all five held-out sets are $K = \{3093, 3140, 3072, 3110, 2822\}$. Fig. 9 shows the visualization of two randomly selected part detectors for each image class from the first held-out set, and their corresponding top-5 high-scoring image patches. It is very interesting to see that the part detectors can capture more informative visual elements that seem very intuitive to us. For example, the part detectors for the "airplane" class capture the airplanes with different orientations and sizes; the ones for the "intersection" category capture the turnings and the zebra crossings. These discriminative detectors can therefore capture the essence of the scene in terms of these highly consistent and repeating patterns and hence provide a conceptually simple but surprisingly effective visual representation.

3.2.2. COPD-based image representation

Image representation plays a key role in scene-level geographic image classification. In this work, we present an effective image

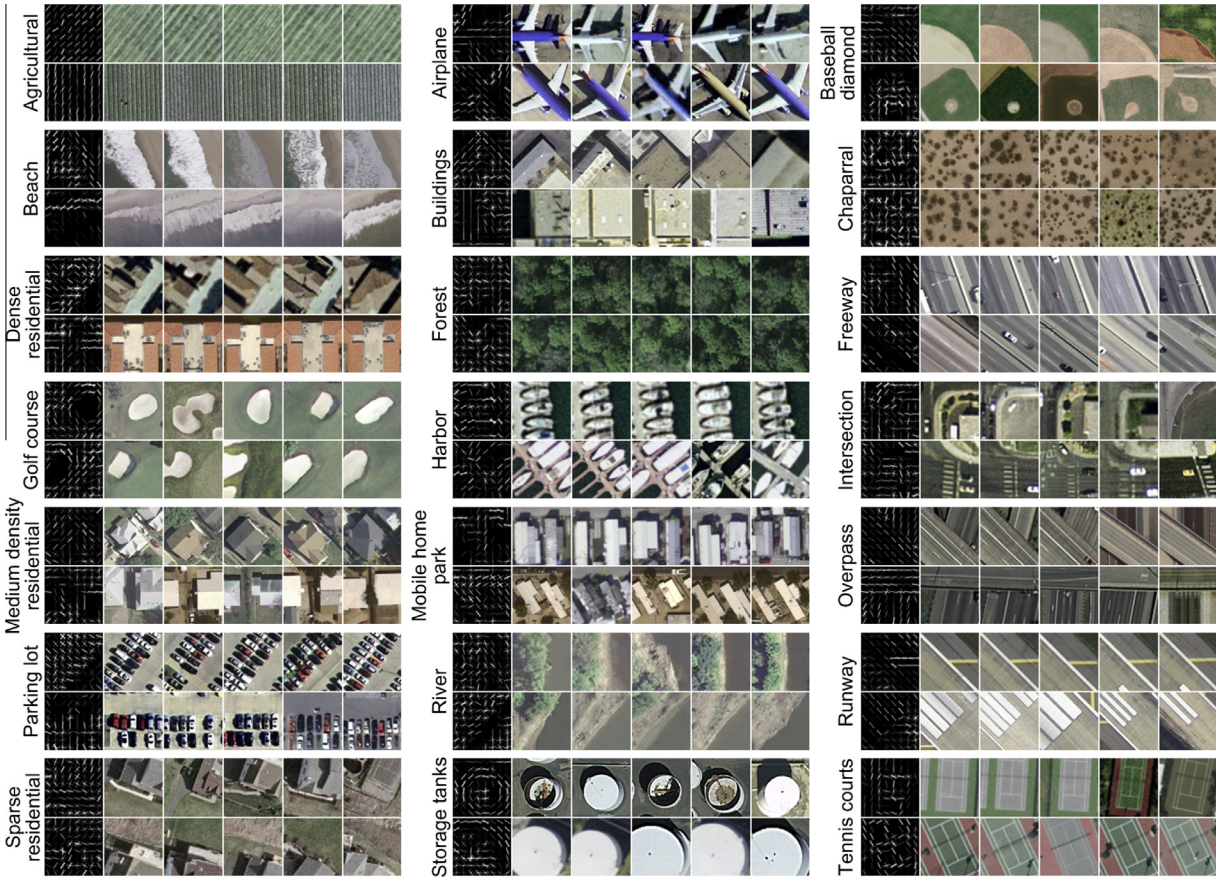

Fig. 9. The visualization of two randomly selected part detectors for each image class and their corresponding top-5 high-scoring image patches.

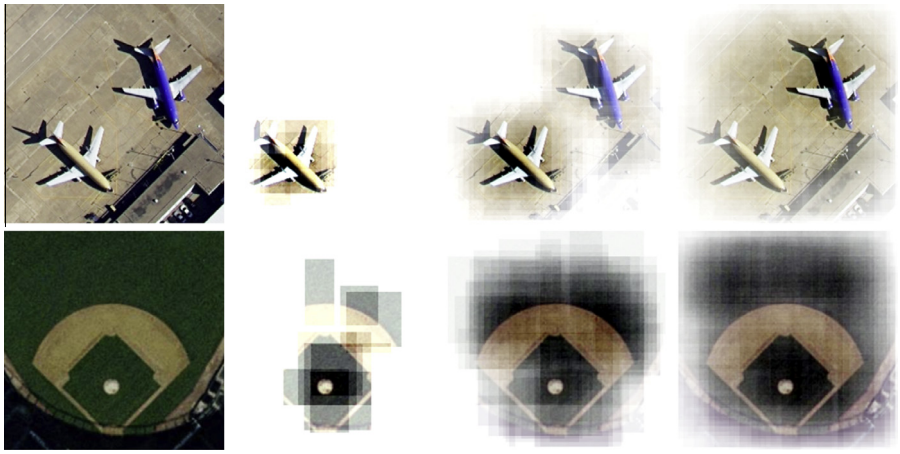

Fig. 10. Two original images (1st column) and their visualization images constructed by averaging their top-10 (2nd column), top-100 (3rd column), and top-500 (4th column) high-scoring patches.

representation method based on a collection of representative and discriminative part detectors that uses mid-level visual elements (i.e., discriminative image patches) as attributes for image representation. Its core is that through detecting discriminative image patches by using pre-trained part detectors, each image is represented as a feature vector of the responses of the top-J high-scoring image patches. For example, Fig. 10 shows two original images (1st column) and their visualization images constructed by averaging their top-10 (2nd column), top-100 (3rd column), and top-500 (4th column) high-scoring image patches, respectively. As can be seen from Fig. 10, a smaller value of J cannot obtain all image patches that are most related to the image class (e.g. the second column in Fig. 10). A bigger value of J can result in certain background except for the image patches that are most related to the image class (e.g. the fourth column in Fig. 10). Therefore, selecting an optimal parameter of J to capture the most discriminative essence of the scene is very important for high-level scene recognition tasks. We will report the detailed parameter optimization in Subsection 3.3.2.

Fig. 11. Some example images from the 21-class LULC data set.

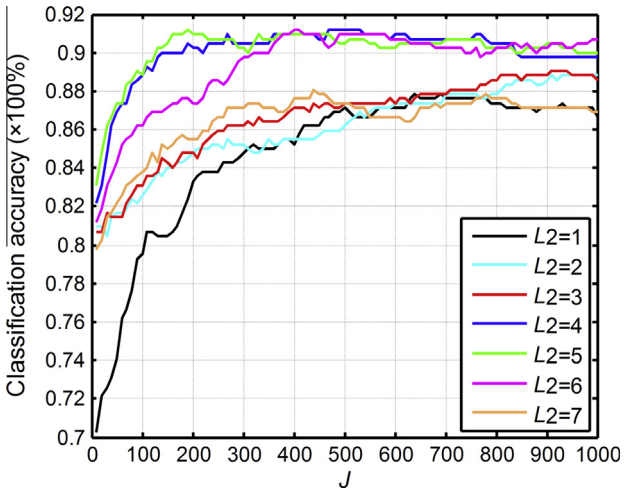

Fig. 12. Classification accuracy obtained by varying *L2* and *J*.

Specifically, given an input image I, we first run all part detectors $\Gamma = \{\Gamma^c\}_{c=1}^{C}$ on its HOG (Dalal and Triggs, 2005) feature pyramid $H(I)$ to compute the response and corresponding part detector label for each location by adopting the similar process used for object detection, as illustrated in Eq. (2). Next, the top-J high-scoring image patches (measured by their responses) and their part detector labels are obtained. Finally, the responses are normalized to [0, 1] and the input image is represented as a feature vector of $F(I)$ by accumulating all normalized responses to their corresponding part detectors. The dimension of $F(I)$ equals to the total number of part detectors in $\Gamma = \{\Gamma^c\}_{c=1}^{C}$.

3.2.3. Image classification

We use a simple one-vs-all scheme to perform image classification by constructing a set of binary SVM classifiers. Each one-vs-all SVM classifier is trained individually by treating the images of the chosen class as positive instances and the rest images as negatives instances. An unlabeled test image is assigned to label of the classifier with the highest response.

3.3. Experiments

3.3.1. LULC data set description

We comprehensively evaluate the performance of the explored COPD-based image classification method on a publicly available data set downloaded from http://vision.ucmerced.edu/datasets (Yang and Newsam, 2010, 2011). The data set comprises the following 21 LULC classes: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts. Each class consists of 100 images measuring 256 × 256 pixels, with a pixel resolution of 30 cm in the red-green-blue color space. Fig. 11 shows four samples of each class from this data set.

3.3.2. Experimental setting

In the implementation of image classification, when detecting discriminative patches from images, we need construct a multi-level HOG feature pyramid for each image in a similar way as object detection. The only difference between them is that the total number of feature pyramid level L is not limited to eight and it changes as the image size changes, i.e. $L = \lfloor 5\log_2 \min(rows, cols)/60 \rfloor + 1$, where *rows* and *cols* denote the image size in pixels in row and

Table 4
Classification accuracies over all 21 classes for five different held-out sets.

Held-out set number	1	2	3	4	5	Average
Classification accuracies (%)	90.95	90.24	93.33	90.48	91.67	91.33 ± 1.11

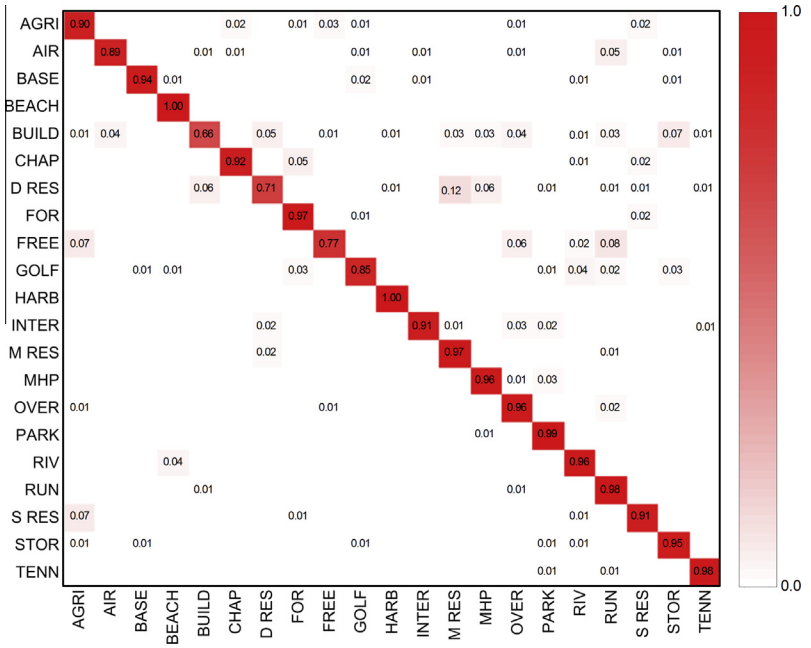

Fig. 13. Confusion matrix by averaging classification results over all five cross-validations.

Table 5
Average classification accuracies of 12 different methods.

Methods	BOVW	SPMK	SCK	BOVW + SCK	Color-RGB	Color-HLS
Average classification accuracies (%)	76.81[a]	75.29[a]	72.52[a]	77.71[a]	76.71[a]	81.19[a]
Methods	Color-Lab	Texture	SPCK	SPCK+	SPCK++	Our method
Average classification accuracies (%)	66.43[a]	76.91[a]	73.14[b]	76.05[b]	77.38[b]	91.33 ± 1.11

[a] The results are from Yang and Newsam (2010).
[b] The results are from Yang and Newsam (2011).

column, respectively. This results in the minimum size of image patch that each part detector could detect is 60 × 60 pixels, while the maximum could be as large as a full image.

We evaluate our framework using the same five-fold cross-validation methodology as the methods of Yang and Newsam (2010, 2011). To be specific, the images of each class are randomly split into five equal sets. The COPD and one-vs-all SVM classifier are then trained on four of the sets and evaluated on the held-out set. The classification accuracy is the fraction of the held-out images of 21 classes that are correctly labeled, and the average classification accuracy is the average over the five evaluations.

In the developed framework, $L2$ and J are two critical parameters associated with image classification result, so we constructed experiments on the first held-out set to optimize the parameters. Fig. 12 shows the classification accuracy when varying $L2$ and J. As can be seen, (1) Classification accuracy was improved with the increase of $L2$ and then dropped off; (2) Classification accuracy was improved with the increase of J and then stabilized in a certain range. Especially, when we fixed the value of $L2$ to be 4 and then changed J from 360 to 830 with a stride of 10, the classification accuracy only varies in the range of [0.9048, 0.9119]. Consequently, we empirically set $L2 = 4$ and $J = 360$ in our image classification evaluations. The iteration times for object detection ($L1 = 3$) is smaller than that for image classification ($L2 = 4$) can be easily explained: the seeds provided better initial clusters than that obtained by k-means clustering.

3.3.3. Experimental results and comparisons

Table 4 shows the classification accuracies averaged over all 21 classes for five different held-out sets. Fig. 13 presents the confusion matrix by averaging classification results over all five cross-validations, where the entry in row X and column Y denotes the rate of test images from class X that were classified as class Y. From the figure, we observed that for most of the categories (16/21) we have a classification rate higher than 90%, especially for "beach" class, the classification rate is 100%. In addition, the biggest confusion happens between "dense residential" and "medium density residential", due to their similar global structure and spatial layout.

Table 5 lists in the average classification accuracies of our method and some state-of-the-art methods. The results of Table 5 except for our method are from Yang and Newsam (2010, 2011). Comparison with state-of-the-art methods shows the huge performance gain (average gain of 15.86% on 11 different methods) resulting from our framework. To the best of our knowledge, this result is the best on this data set, which adequately shows the superiority and effectiveness of the developed framework.

4. Conclusions

In this paper, we developed a practical framework for multi-class geospatial object detection and geographic image classification based on collection of part detectors. Comprehensive evaluations on two remote sensing image databases and comparisons with a number of state-of-the-art approaches demonstrated that the developed framework is effective and superior. However, some problems still exist. One important consideration is computation cost. Since each part detector of COPD is independent, the computational resources increase linearly with the increasing of the number of part detectors, both in training and in the process of the parts detection.

The future work may include the following two issues. First, sharing part detectors, e.g. by using sparse coding scheme, to reduce the overall number of model parameters and improve the computational efficiency. Second, testing the developed framework in more vision applications, such as geographical image retrieval.

Acknowledgements

The Vaihingen data set was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) (Cramer, 2010): http://www.ifp.uni-stuttgart.de/dgpf/DKEPAllg.html. This work was partially supported by the National Science Foundation of China under Grant 91120005, 61473231, 61401357 and 61333017, Doctoral Fund of Ministry of Education of China under Grant 20136102110037, and China Postdoctoral Science Foundation under Grant 2014M552491. The authors would like to thank the anonymous reviewers for their constructive comments.

References

Aksoy, S., Koperski, K., Tusk, C., Marchisio, G., Tilton, J.C., 2005. Learning Bayesian classifiers for scene classification with a visual grammar. IEEE Trans. Geosci. Remote Sens. 43, 581–589.
Aytekın, Ö., Erener, A., Ulusoy, İ., Düzgün, Ş., 2012. Unsupervised building detection in complex urban environments from multispectral satellite imagery. Int. J. Remote Sens. 33, 2152–2177.
Aytekin, Ö., Zöngür, U., Halici, U., 2013. Texture-based airport runway detection. IEEE Geosci. Remote Sens. Lett. 10, 471–475.
Bhagavathy, S., Manjunath, B.S., 2006. Modeling and detection of geospatial objects using texture motifs. IEEE Trans. Geosci. Remote Sens. 44, 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, 749–753.
Bourdev, L., Malik, J., 2009. Poselets: Body part detectors trained using 3d human pose annotations. In: Proceedings of the Twelfth IEEE International Conference on Computer Vision (ICCV 2009). IEEE, Kyoto, Japan, pp. 1365–1372.
Buckland, M.K., Gey, F.C., 1994. The relationship between recall and precision. J. Am. Soc. Inform. Sci. 45, 12–19.
Cheng, G., Guo, L., Zhao, T., Han, J., Li, H., Fang, J., 2013a. Automatic landslide detection from remote-sensing imagery using a scene classification method based on BoVW and pLSA. Int. J. Remote Sens. 34, 45–59.
Cheng, G., Han, J., Guo, L., Qian, X., Zhou, P., Yao, X., Hu, X., 2013b. Object detection in remote sensing imagery using a discriminatively trained mixture model. ISPRS J. Photogramm. Remote Sens. 85, 32–43.
Corbane, C., Najman, L., Pecoul, E., Demagistri, L., Petit, M., 2010. A complete processing chain for ship detection using optical satellite imagery. Int. J. Remote Sens. 31, 5837–5854.
Cramer, M., 2010. The DGPF test on digital aerial camera evaluation – overview and test design. Photogramm. – Fernerkundung – Geoinform. 2, 73–82.
Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C., 2004. Visual categorization with bags of keypoints. In: Proceedings of ECCV International Workshop on Statistical Learning in Computer Vision, Prague, pp. 1–22.
Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005). IEEE, San Diego, CA, pp. 886–893.
Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. Int. J. Comput. Vision 88, 303–338.

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, 1627–1645.

Grabner, H., Nguyen, T.T., Gruber, B., Bischof, H., 2008. On-line boosting-based car detection from aerial images. ISPRS J. Photogramm. Remote Sens. 63, 382–396.

Han, J., Zhou, P., Zhang, D., Cheng, G., Guo, L., Liu, Z., Bu, S., Wu, J., 2014. Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding. ISPRS J. Photogramm. Remote Sens. 89, 37–48.

Juneja, M., Vedaldi, A., Jawahar, C., Zisserman, A., 2013. Blocks that shout: distinctive parts for scene classification. In: Proceedings of the 2013 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2013), Portland, OR, pp. 923–930.

Kim, M., Madden, M., Warner, T.A., 2009. Forest type mapping using object-specific texture measures from multispectral Ikonos imagery: segmentation quality and image classification issues. Photogramm. Eng. Remote Sens. 75, 819–829.

Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006). IEEE, New York, pp. 2169–2178.

Li, F.F., Perona, P., 2005. A bayesian hierarchical model for learning natural scene categories. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005). IEEE, San Diego, CA, pp. 524–531.

Li, Q., Wu, J., Tu, Z., 2013. Harvesting mid-level visual concepts from large-scale internet images. In: Proceedings of the 2013 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2013), Portland, OR, pp. 851–858.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60, 91–110.

Malisiewicz, T., Gupta, A., Efros, A.A., 2011. Ensemble of exemplar-svms for object detection and beyond. In: Proceedings of the thirteenth IEEE International Conference on Computer Vision (ICCV 2011). IEEE, Barcelona, Spain, pp. 89–96.

Martha, T.R., Kerle, N., van Westen, C.J., Jetten, V., Kumar, K.V., 2011. Segment optimization and data-driven thresholding for knowledge-based landslide detection by object-based image analysis. IEEE Trans. Geosci. Remote Sens. 49, 4928–4943.

Schroder, M., Rehrauer, H., Seidel, K., Datcu, M., 2000. Interactive learning and probabilistic retrieval in remote sensing image archives. IEEE Trans. Geosci. Remote Sens. 38, 2288–2298.

Shyu, C., Klaric, M., Scott, G.J., Barb, A.S., Davis, C.H., Palaniappan, K., 2007. GeoIRIS: Geospatial information retrieval and indexing system—content mining, semantics modeling, and complex queries. IEEE Trans. Geosci. Remote Sens. 45, 839–852.

Singh, S., Gupta, A., Efros, A.A., 2012. Unsupervised discovery of mid-level discriminative patches. In: Proceedings of the twelfth European Conference on Computer Vision (ECCV 2012). Springer, Firenze, Italy, pp. 73–86.

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

Sun, J., Ponce, J., 2013. Learning discriminative part detectors for image classification and cosegmentation. In: Proceedings of the fourteenth IEEE International Conference on Computer Vision (ICCV 2013), Sydney, Australia, pp. 3400–3407.

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, 109–113.

Ünsalan, C., Sirmacek, B., 2012. Road network detection using probabilistic and graph theoretical methods. IEEE Trans. Geosci. Remote Sens. 50, 4441–4453.

Văduva, C., Gavăt, I., Datcu, M., 2013. Latent dirichlet allocation for spatial analysis of satellite images. IEEE Trans. Geosci. Remote Sens. 51, 2770–2786.

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, 366–370.

Yang, Y., Newsam, S., 2010. Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the eighteenth SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, San Jose, California, pp. 270–279.

Yang, Y., Newsam, S., 2011. Spatial pyramid co-occurrence for image classification. In: Proceedings of the thirteenth IEEE International Conference on Computer Vision (ICCV 2011). IEEE, Barcelona, Spain, pp. 1465–1472.

Yang, Y., Newsam, S., 2013. Geographic image retrieval using local invariant features. IEEE Trans. Geosci. Remote Sens. 51, 818–832.

Zhu, C., Zhou, H., Wang, R., Guo, J., 2010. A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features. IEEE Trans. Geosci. Remote Sens. 48, 3446–3456.