VHR Object Detection Based on Structural Feature Extraction and Query Expansion

Xiao Bai, Huigang Zhang, and Jun Zhou *Senior Member, IEEE*

Abstract

Object detection is an important task in very high resolution remote sensing image analysis. Traditional detection approaches are often not sufficiently robust in dealing with the variations of targets, and sometimes suffer from limited training samples. In this paper, we tackle these two problems by proposing a novel method for object detection based on structural feature description and query expansion. The feature description combines both local and global information of objects. After initial feature extraction from a query image and representative samples, these descriptors are updated through an augmentation process to better describe the object of interest. The object detection step is implemented using a ranking support vector machine (SVM), which converts the detection task to a ranking query task. The ranking SVM is firstly trained on a small subset of training data with samples automatically ranked based on similarities to the query image. Then a novel query expansion method is introduced to update the initial object model by active learning with human inputs on ranking of image pairs. Once the query expansion process is completed, which is determined by measuring entropy changes, the model is then applied to the whole target dataset in which objects in different classes shall be detected. We evaluate the proposed method on high resolution satellite images, and demonstrate its clear advantages over several other object detection methods.

Index Terms

Object detection, VHR remote sensing images, feature augmentation, query expansion, active learning.

X. Bai and H. Zhang are with the School of Computer Science and Engineering, Beihang University, Beijing 100191, China. (e-mail: baixiao.buaa@googlemail.com.)

J. Zhou is with School of Information and Communication Technology, Griffith University, Nathan, QLD 4111, Australia.

I. INTRODUCTION

With the development of imaging technology, many airborne or satellite sensors have provided very high resolution (VHR) imagery. These images have been used in a wide range of remote sensing applications, including agriculture, land planning, disaster control, environmental management, defense and surveillance [1], [2], [3], [4]. Of particular interest in these applications is the detection of natural or man-made targets, which is facilitated by abundant spatial information and details of objects in the VHR imagery. This is, however, a non-trivial task due to the intra-class variations of object classes. Therefore, developing robust VHR object detection methods is very challenging.

Object detection tasks often start from grouping pixels into segments as object candidates. For example, Chaudhuri *et al.* grouped candidate bridge pixels into possible bridge segments based on their connectivity and geometric properties, and then verified these segments according to their directions and connectivity with road segments [5]. Akcay *et al.* also used segmentation algorithm to help detection, where the segments were extracted by applying connected components analysis to the pixels selected according to their morphological profiles [6]. Liu *et al.* proposed a method to automatically detect vehicles from QuickBird images [2]. This method used morphological filters in various sizes to remove road lines while retaining vehicle shapes. Then neighboring pixels in similar intensity were grouped into objects whose intensity histograms were generated. Finally, the vehicle detection task was implemented by discriminating vehicle histograms with those from roads and background. Morphological methods have been used in other reported work when spatial context has to be considered. They assign pixels to connected components of objects so that shape information can be recovered to detect objects such as roof, buildings, roads, and vegetation [7], [8], [9], [10].

Due to the large size of VHR image, an image can also be split into small patches to facilitate the detection task. Various image classification method can be employed to judge whether these patches contain the object of interest. From the image patches, spatial relationship of different parts of objects can be extracted. This leads to structural feature extraction methods. Jin *et al.* combined structural, contextual, and spectral information to detect buildings in urban areas [11]. Wang *et al.* presented a hierarchical connection graph (HCG) algorithm for roof extraction from aerial imagery [12]. Sirmacek *et al.* combined scale-invariant feature transform (SIFT) [13] with

graph theory to detect urban areas and buildings in grayscale Ikonos images [14]. Gabor filters have also been used to extract spatial characteristics of buildings, such as edges and corners, in different orientations [15]. In [16], intermediate level inputs were used to study whether or not structural information can complement or replace spectral information for road detection. Li *et al.* [17] proposed a texture preceded algorithm for VHR image segmentation, in which texture was clustered and then combined with distance, spectral, and shape features to measure the similarity between subregions. Then graph models were used to merge regions into land cover types. Belongie *et al.* proposed a widely used descriptor, the shape context [18], which is a globally discriminative feature that captures not only the local points but also the distribution of the remaining points relative to them. In [19], local self-similarity (LSS) feature was applied to capture the internal geometric layouts of regions in an image for object and action detection.

Whilst various features and their combinations can be very effective for object description, feature augmentation is a step that further improves the performance of object detection. This is based on the rationale that objects of the same category share a set of common features with each other, but not those of different categories and background. Therefore, these shared features can be used to improve category specific description. Feature augmentation is normally achieved by enhancing the useful and discriminative features, e.g., by assigning higher weights to them, or/and weakening the influence of noisy features, e.g., by assigning lower weights to or even eliminating some of them. Efforts in this direction include co-occurrence analysis [20], [21], and feature selection using random forest [22], graph-based method [23], or multiple kernel learning [24].

During the feature augmentation and object classification process, large number of training samples are usually needed to produce a good model. However, collecting training samples is often difficult and can be very time consuming. To address this problem, incremental learning is a viable solution. Amongst various incremental learning models, active learning [25] updates a model by interaction with users. It requires only few training samples at the beginning, and can gradually improve itself by obtaining manual labels for data it has chosen.

Active learning methods have shown to be very useful and practical in remote sensing image analysis [26], [27]. Most of them have been found effective for classification tasks, but have rarely been used for detection which is the problem addressed in this paper. Ferecatu *et al.* [27] put forward an active learning selection criterion to update a ranking model by user feedback at

every iteration. They used most ambiguous and orthogonal (MAO) criterion to make decisions on which image should be chosen as feedbacks. In the work of Demir *et al.* [28], the feedback selection criterion was a combination of uncertainty and diversity. The uncertainty criterion was used to improve the correctness of sample classification, while the diversity criterion allowed a set of unlabeled samples with higher diversity be selected. The follow-up work in [29] was developed using a Bayesian approach in order to select aligned unlabeled pixels in two images with most uncertainty. Moreover, Tuia *et al.* [30] proposed a simple but effective method that used active learning to solve the data set shift problems when a classifier trained on one portion of the image dataset was applied to the rest of the dataset.

In developing active learning methods, it is desirable that a recognition model has high accuracy while being capable of ranking unseen samples so that those with high ranks can be selected for manual labelling. Among various recognition methods, support vector machines (SVMs) have shown superior performance and have been widely used for VHR images analysis [31], [32], [33], [34]. Nevertheless, common SVMs can not provide accurate ranking information. To address this problem, Joachims proposed the ranking SVM algorithm [35]. It allows pairwise ranking of two data points in a feature space by user inputs or distance measurement. This property makes it soon be used for image retrieval [36] and attribute studies [37], [38].

In this paper, we tackle the structural feature extraction and query expansion problem. The proposed method first represents a query image using a structured image descriptor. Then a feature augmentation step is employed to update image descriptors using a small number of representative samples from which a ranking SVM is also learnt. This initial model is enriched with the additional information obtained from the query expansion, which allows active learning from human inputs on ranking of image pairs in the training set. This step refines the ranking SVM classification model, which is then applied to all target image patches. By doing so, the learned model can be updated in an incremental way, thus, does not need large amount of representative samples. The proposed method has been tested on VHR images captured by commercial satellites such as QuickBird and GeoEye-1.

The main contributions of this paper are threefold. Firstly, we present a novel image descriptor that combines SIFT feature with spatial information of objects, which is an effective representation to describe objects in VHR images. Secondly, we propose a feature augmentation process which augments common features in the query object and improves the query performance.

Thirdly, a novel query expansion strategy is proposed to update the query model. This strategy uses the state-of-the-art ranking SVM and active learning methods, and is a practical solution for many remote sensing applications that require human-in-the-loop processing.

The remaining part of this paper is organized as follows. In Section II, we give an overview of the proposed object detection method. Section III introduces the feature extraction and augmentation methods. Section IV describes the object detection method, including query expansion process, ranking SVM and active learning. Experimental results on several datasets are presented and analyzed in Section V. Finally, we conclude this paper with discussions in Section VI.

II. Method Overview

In this section, we give an overview of the proposed method. A flow chart of key steps to generate the detection model is shown in Fig. 1. Given a query image of an object class, the objective of this work is to detect instances of the same class in a dataset of VHR satellite images. Because these images are normally very large and may contain one or multiple objects, a straightforward way of processing is to segment the images into small patches and then predict whether or not each patch contains the query object.

As a preprocessing step of the proposed approach, sliding windows in different sizes and aspect ratios, which are determined by the size of the query image, are used to scan through the images in the dataset and divide them into many patches. Some patches are randomly selected to form a training set, while the rest patches become the target set for object detection. The training patches, whose labels are unknown, will be used in the following feature augmentation and classifier learning steps.

The first key step of the proposed method is to convert images into vectors via feature extraction so that they can be used to train the detection model. Here, we propose a novel structured feature extraction method. It extracts SIFT features from an image patch. Then an image descriptor is constructed as a weighted frequency histogram of these SIFT descriptors, where the weights are determined by the spatial distribution of SIFT keypoints with respect to the image center. This descriptor contains rich structural and statistical information of an image.

We then introduce a feature augmentation step to refine the image descriptors. The descriptors of the query image and representative samples are updated using selected similar representative

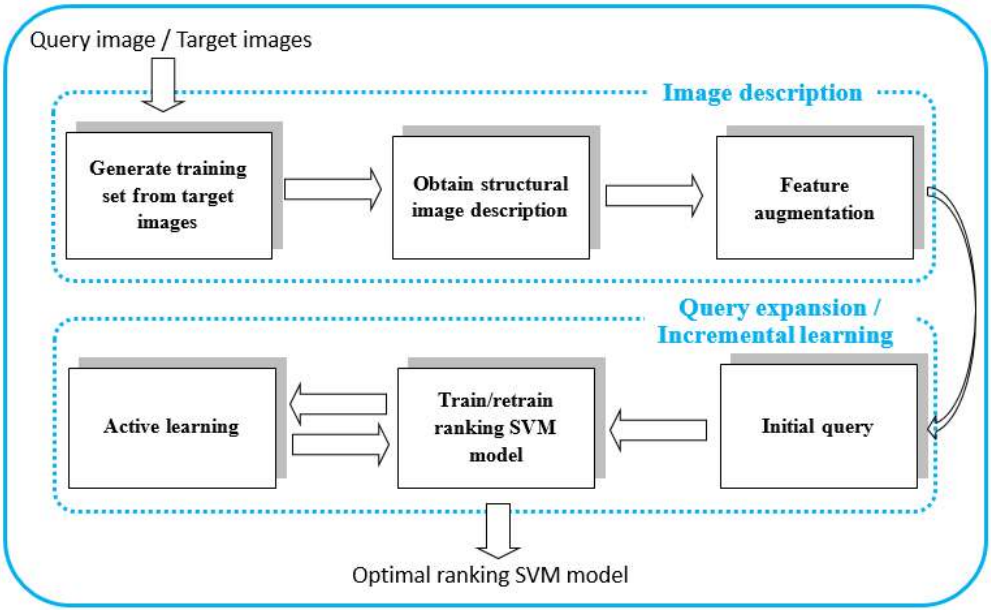


Fig. 1. A flow chart of key steps to generate the proposed object detection model.

samples. The goal is to augment the weight of those discriminative foreground features. By highlighting these features, we can get a more robust image descriptor.

Having obtained the augmented description, we calculate and rank the similarities between the representative samples and the query image. The ranking results are used to train an initial ranking SVM classifier. An incremental learning process then gradually refines this initial model. This process firstly applies the learned ranking SVM classifier to all training image patches and ranks them according to the uncertainty of classification measured by an entropy function. Then top ranked image samples are presented to users for pairwise comparison. Users provide input on the ranking of similarities between the samples and the query image, so that the ranking SVM can be updated. This incremental process iterates until a stopping condition is met. Details on image description, feature augmentation, and incremental learning process can be found in Section III-A, III-B, and IV, respectively.

Finally, the target image patches are classified by the refined ranking SVM model for object detection. If the ranked score is larger than a threshold, an image patch is considered as containing the query target, and vice versa.

III. Image Descriptor and Feature Augmentation

When applied to the VHR image analysis, traditional feature extraction methods are characterized by two main problems. Firstly, most feature extraction operations are performed at pixel level such as textures, or are region-based such as SIFT and MSER [39]. These features are incapable of describing the global information of objects and characterizing the spatial relationships between object keypoints or parts. Secondly, when extracted features are statistical in nature, for example, constructed by the popular bag-of-words model [40], large amount of training samples are required in order to generate accurate feature distribution. This, however, is usually difficult to achieve in VHR image analysis due to the lack of labelled data.

To tackle these two problems, we propose a novel feature extraction and augmentation method. The feature extraction method generates a descriptor for each image patch, which captures not only the local properties of an object but also the spatial connections between keypoints where local features are extracted. The development of this structural descriptor is based on the rationale that the spatial distribution of features can contribute to object description, and that the features closer to the geometric center of the object shall make higher contributions. Therefore, this feature extraction method provides comprehensive information about object candidates in images. Once structural image descriptors are generated, a feature augmentation process is used to enhance the descriptors of both query image and representative samples so that they become more distinctive.

A. Feature extraction and image description

The proposed feature extraction method consists of several steps. The first step is extracting keypoints from an image patch. To do so, the scale-invariant feature transform (SIFT) method is used. This method generates keypoints and corresponding descriptors that are invariant to object scale and rotation. Using K-means method [41], a codebook can be constructed by clustering the SIFT descriptors and treating the cluster centers as codewords. Then each SIFT descriptor is assigned to a single cluster center via nearest neighbor search. In this way, an image patch can be represented by a histogram of the codewords.

Once the SIFT feature is extracted and the codebook is generated, in the second step, we construct log-polar coordinates at the center of the image patch. This step forms five concentric circles as shown in Fig. 2(a). The radiuses of these five circles are set to 0.125, 0.25, 0.5, 1, and 2 times the mean distance from key points to the image patch center, respectively. Therefore, these

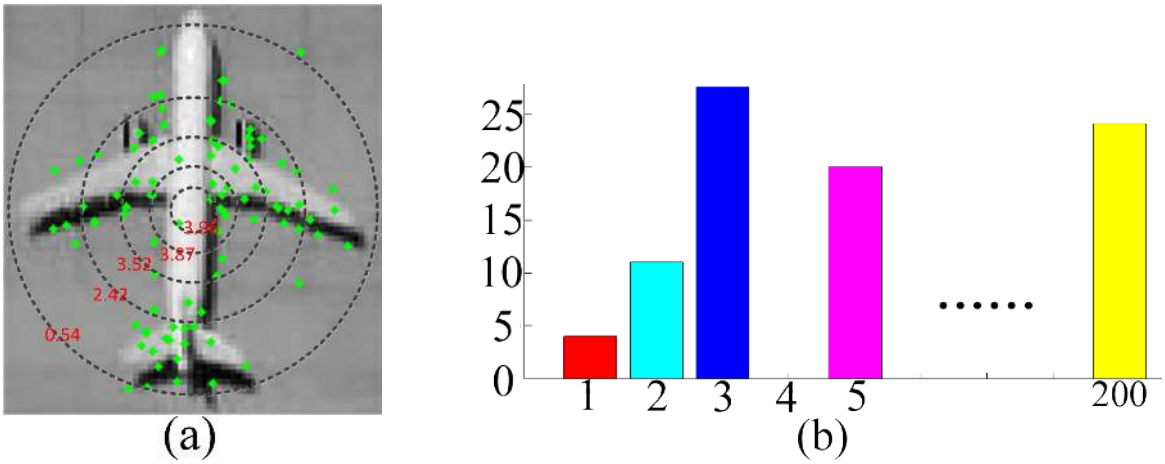


Fig. 2. Construction and representation of the proposed image descriptor. In (a), an image patch is described using circles, and the features in each circle are assigned with different weights. In (b), the image patch is converted to a K dimensional histogram, with each dimension represents a feature codeword. ($K = 200$ in this example)

circles divide the image patch into five regions around its center. Note that such concentric setting is only related to the relevant locations of the keypoints to the image patch center, therefore, it can be applied to images in various sizes. If a keypoint q falls into the region between the $(j-1)$-th and the j-th circle, we define $q \in region(j)$. Note that some outer points are not within any circle, and the outer circles may also go beyond the boundaries of the image patch. Because the sliding window segmentation method ensures at least one window may contain the target in most cases, the above two conditions do not influence the effectiveness of the proposed descriptor.

Different weights α_j are assigned to features in each region using a Gaussian function

$$\alpha_j = \frac{10}{\sqrt{2\pi}} \exp(-\frac{r_j^2}{2}) \tag{1}$$

where r is the ratio of circle radius over the mean distance of keypoints. Therefore, the weights are set to 3.96, 3.87, 3.52, 2.42 and 0.54 from inside to outside, respectively, as marked in Fig. 2(a). Note that larger weights are given to features closer to the image center, because these features are more likely to be relevant to the target object other than the background.

In the third step, all SIFT features around the image center are mapped to the circular description, so that the image patch is represented as a weighted frequency histogram $f(i) = [f(i)_1, f(i)_2, ..., f(i)_K]^T$, where i is the index of the image and K is the total number of

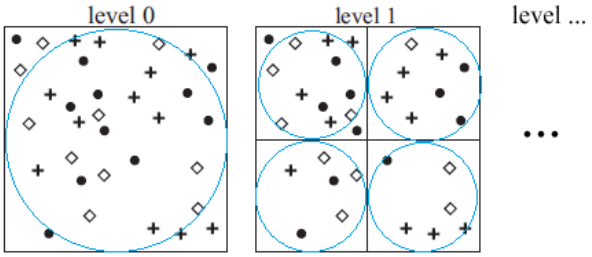


Fig. 3. Example of a spatial pyramid descriptor construction. It partitions the image into increasingly finer sub-regions. Then histograms of local features in each region are computed and concatenated to form a new bag-of-words image representation.

codewords. Each entry f_k is computed as

$$f(i)_k = \sum_{j=1}^{5} (\alpha_j \times \#\{q : q \in codebook(k) \quad and \quad q \in region(j)\}) \tag{2}$$

where k indices the codeword, and α_j is the weight for the j-th circle as defined in Equation (1). An example of such feature construction is shown in Fig. 2(b).

Furthermore, we extend the constructed descriptor following a spatial pyramid method as proposed by Lazebnik *et al.* [42]. This method partitions an image into increasingly finer sub-regions. Then histograms of local features in each region are computed and concatenated to form an effective and computationally efficient extension of an unordered bag-of-words image representation, which characterizes the coarse to fine spatial relationship of local features. In our work, we construct a sequence of circles not only at the original level but also at split subregions, as shown in Fig. 3.

We only calculate the circular representations in level 0 and level 1 for efficiency purpose. Therefore, features can be extracted from 5 regions, and each of them is treated independently. Suppose that the length of the histogram for each region is K, this allows a feature vector of dimension $5 * K$ be constructed, which forms the descriptor of the image patch. As done in [42], each histogram is multiplied by a weight before concatenation. The weight associated with level l is set to $\frac{1}{2^{L-l}}$, where L is the maximum level. Thus the weights of the two levels are set to 1 and 0.5, respectively.

The proposed descriptor is invariant to scale changes because it is built using relative distance to the center of an image patch. When only level 0 is used, this descriptor is also rotation invariant because it is based on statistics of SIFT features in each circle. However, when more

than one levels are used, the descriptor is not rotation invariant anymore, just like the widely used spatial pyramid method [42]. This problem is alleviated by symmetric nature of many object classes in the remote sensing image. The structured feature also shows great advantages in characterize the spatial relationships of different parts of objects, and in turn effectively helps the object detection task.

In the proposed feature descriptor, the SIFT features are extracted from small neighborhood of the keypoints. They describe the local information of image regions. On the other hand, the proposed circular and hierarchical feature representation combines these local features through their spatial information. This gives a global description of the image patches and the objects contained in these patches. Therefore, this proposed descriptor combines local SIFT features with their spatial relationship. It contains more structural information of a target image than the traditional bag-of-words method. Moreover, this descriptor is also statistical, which makes the following classification tasks easier than those methods that only use spatial information.

B. Feature augmentation

After the proposed image description step, we introduce a feature augmentation strategy to improve the descriptors of the query image and representative samples. The goal is to augment the weights of those discriminative foreground features. This is done by an unsupervised processing step that identifies matching features extracted from the unlabeled training images.

Our work is motivated by Turcot *et al.* [20], based on the fact that features extracted from the background likely exist in only one single image, while useful foreground features are likely to be found in more than one images that contain the same class of objects. We also use the concept of image adjacency, in which two images that have high similarity are considered to be adjacent. Here we construct a graph $G = (V, E)$ to represent the relationships between image patches. In this graph, each vertex $v \in V$ represents an image patch, and an edge $e_{ij} \in E$ connects two vertices which are verified to be matching.

The feature augmentation starts from finding the closest images among the representative samples to the query image and extracting common features. The reciprocal of Euclidean distance is used to measure the similarities between the image descriptors. For the query image and each representative sample $I(i)$, the similarities between $I(i)$ and all the other images in the training set are calculated. Then a ranked list is obtained according to the degree of similarities.

Those samples with high ranks (larger than a similarity degree S) are considered as images that contain the same class of objects as $I(i)$. These samples can be used for feature augmentation if their SIFT feature points are found to be geometrically consistent with each other via feature verification.

The feature verification is done as follows. Firstly, the initial point-to-point feature matching from two images is achieved by matching SIFT descriptors. Then geometrical checking is performed using RANSAC method [43] in order to select those feature points that are consistent in spatial relationship. If two images have more than M spatially matching feature points, they are considered as neighbors in the graph, and the corresponding matching features are used to augment the description of image $I(i)$. Detailed analysis on how parameters S and M affect the overall performance of our method is given in Section V.

The new description of an image $I(i)$ is then updated as

$$f(i)_k = f(i)_k + \sum_{j,\{(i,j),(j,i)\}\in E} f(j)'_k \tag{3}$$

where $f(i)_k$ is the number of occurrences of codeword k in image $I(i)$, $f(j)'_k$ is the number of occurrences of codeword k in the recalculated description of image $I(j)$, $k = 1, \cdots, K$.

In this way, entries of the original feature descriptor are augmented. It shall be mentioned here that rather than augmenting the vector with all codewords from images, it is preferable to use only those spatially verified matching features for augmentation. This is the main difference from the method in [20], where all visual words from neighboring images are used for feature augmentation.

A summary of the proposed feature augmentation method is given in Algorithm 1.

IV. Classification of object candidates

As mentioned before, given a query image of a target object class, our objective is to retrieve instances in the same class from the VHR satellite images. Rather than merely using the query image to model the object, we use an incremental learning process to get more information of object class and gradually refine the object model. Here we adopt a query expansion strategy to gain additional knowledge from human input. The proposed method consists of three main steps, i.e., getting the initial query results, ranking results by SVM, and updating the prediction model by active learning.

Algorithm 1 Feature augmentation steps

Input: Image descriptors, similarity threshold S, validation threshold M

Output: Image graph $G = (V, E)$, augmented image descriptors

For each image $I(i)$ in the training dataset **do**

  Add $I(i)$ to G as a vertex

  Query the training dataset using image $I(i)$

  $R \leftarrow$ list of the ranking results above S

  for each image $I(j) \in R$ **do**

    Calculate the pairwise validated features between $I(i)$ and $I(j)$

    if the number of validated features is larger than M **then**

      Add an edge between image $I(i)$ and $I(j)$

      Save all the validated features in $I(j)$ as $f(j)'_k$ (Sec. III-B)

    end

  end

  Update the descriptor of $I(i)$ using Equation (3)

end

A. Initial query

Note that in Section II, a set of representative samples have been randomly sampled from image patches in the VHR dataset. These representative samples are unranked and their labels are unknown. The goal of the initial query step is ranking these representative samples, so that they can be used to train a ranking SVM. The ranking is done by calculating the Euclidean distances between the structural descriptors of the query image and all representative samples. Then the representative samples are sorted based on their similarities.

A list of images in the initial detection results for a query sample are given in Fig. 4. It can be observed that these results may not be correct due to the variation of images. In addition, the Euclidean distance measure has ignored the inherent relationships among intra-class images. Therefore, we explore query expansion to improve the detection model.

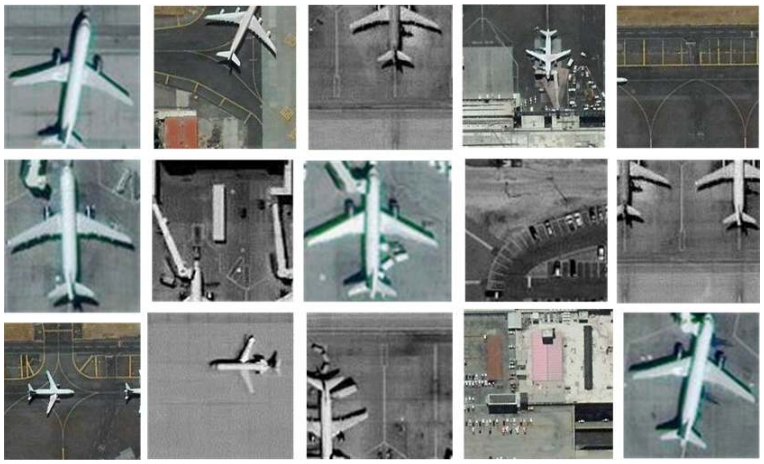


Fig. 4. Top 15 images patches from the initial query.

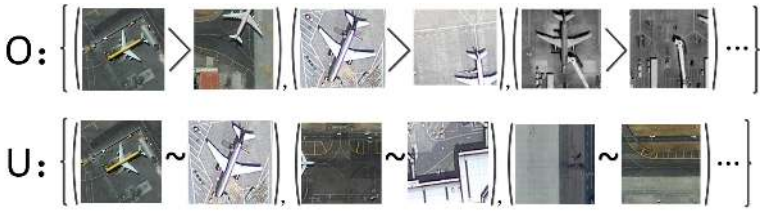


Fig. 5. Examples of the ordered (**O**) and unordered (**U**) set of image pairs.

B. Ranking SVM for object detection

The query expansion method uses a ranking SVM method so as to improve the accuracy of object detection. In this method, a training set consists of some ordered pairs of images $\mathbf{O} = \{(x_i, x_j)\}$ and some unordered pairs $\mathbf{U} = \{(x_i, x_j)\}$. If $(x_i, x_j) \in \mathbf{O}$, then image x_i is preferable to x_j, denoted as $x_i \succ x_j$. Alternatively, $(x_i, x_j) \in \mathbf{U}$ suggests that image x_i and x_j have similar ranking scores, and is denoted as $x_i \sim x_j$. Fig. 5 gives some examples on the pairwise relationships between image patches in $\mathbf{O}$ and $\mathbf{U}$ respectively.

Note that the initial query results have given abundant ranking information for $\mathbf{O}$. We randomly select image pairs (x_i, x_j) in the ranking list where x_i has a higher ranking score than x_j, and put these pairs in $\mathbf{O}$. As for $\mathbf{U}$, the initial ranking results do not give such information, so it is defined as an empty set in the first round. While in the following active learning step, $\mathbf{O}$ and $\mathbf{U}$ are updated iteratively as discussed in Section IV-C.

In order to learn the ranking function $f : \mathbb{R}^n \to \mathbb{R}$, following constraints have been used

$$\forall x_i \succ x_j : f(x_i) > f(x_j), \qquad \forall x_i \sim x_j : f(x_i) = f(x_j) \tag{4}$$

This function formulation follows [37], where $f(x_i) = w^T x_i$. Solving w is an NP hard problem. Similar to the solution to the regular SVMs, we can solve this problem by introducing slack variables ξ_{ij} and η_{ij}. In addition, L2 regularization is imposed on w to maximize the margin between the closest projections, which leads to the following optimization target

$$\begin{aligned} min\ & \frac{1}{2}\|w^T\|_2^2 + \gamma(\sum \xi_{ij}^2 + \sum \eta_{ij}^2) \\ s.t.\ \ & w^T(x_i - x_j) \geq 1 - \xi_{ij}; \forall (i,j) \in \mathbf{O} \\ & |w^T(x_i - x_j)| \leq \eta_{ij}; \forall (i,j) \in \mathbf{U} \\ & \xi_{ij} \geq 0, \quad \eta_{ij} \geq 0 \end{aligned} \tag{5}$$

where γ is the trade-off constant between maximizing the margin and satisfying the pairwise image constraints. The above optimization problem can be solved using Newton's method [44].

In fact, a regular SVM can also be used for this object detection tasks if we transform the classification results into probabilities [45]. However, there are several advantages in using ranking SVM as the detection option. First of all, the ranking SVM introduces the ranking relativity concept. By comparing the pairwise images one by one, we can obtain more abundant learning knowledge for our model. Secondly, the ranking property is more suitable for this object detection task, which retrieves more likely images before the less likely ones. Thirdly, and most important, the margin definition of ranking SVMs is more precise than that of regular SVMs. This has been supported by theoretical analysis [37], and by experiments to be introduced in later sections.

We also show the advantage using an example in Fig. 6. The margin of a ranking SVM is the distance between the closest projections within all representative samples. Nevertheless, in a regular SVM classifier, the margin is the distance between the nearest binary-labeled examples. So the margin in a regular SVM is rougher, which may lead to wrong ranking results. As we can see, the ranking SVM correctly ranked samples 4 and 5, but the regular SVM did it wrong.

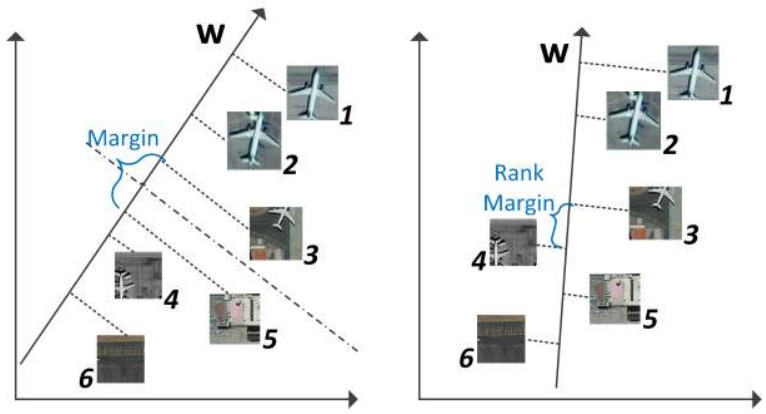


Fig. 6. Illustration of the distinction between learning a regular SVM (left) and a ranking SVM (right). The regular SVM separates two classes, while the ranking SVM preserves a desired ordering of samples.

C. Active learning process

The learning step from previous sections generates an initial ranking SVM model from the training images. To further improve the accuracy of this detection model, we propose a novel active learning based ranking SVM approach. This approach adopts a similar iterative framework as the traditional active learning methods [26]. The differences lie in two aspects. First, our model is based on ranking SVM rather than traditional binary or multi-class classification models. Second, the user intervention is given in the form of ranking of representative samples, rather than giving labels to the novel samples. An entropy function is used to determine which samples are selected for human input, which also forms the basis of a stopping criterion.

The initial ranking SVM is applied to the training set and get a ranking score s for each training image patch. These scores are used to calculate $P(y_i|x)$, which is the probability of assigning candidate x into class y_i. To do so, an exponential function is used, such that

$$P(y_i|x) = \frac{1}{1+\exp(-s)} \tag{6}$$

Because the object detection can be seen as a two-class problem, so $P(y_1|x) = 1 - P(y_2|x)$.

The active learning process then selects those samples that have high uncertainty for human input and model updating, where the uncertainty is evaluated by an entropy function

$$\tilde{x} = -\sum_i P(y_i|x) \log P(y_i|x) \tag{7}$$

where $\tilde{x}$ is the entropy of sample x. Based on the information theory, high uncertainty often suggests more novel information that can be used to refine the prediction model [46]. Therefore, those samples that are the most uncertain, i.e., with high $\tilde{x}$ value, are used to update the ranking model via an active learning strategy. For each iteration, we choose the top 20% highest entropy samples from the training set to update the model. Note that this iterative update process uses the same training set, rather than adding new samples to the training set gradually.

The ranking SVM model is updated using the chosen samples. This requires pairwise ranking of samples as the input to the training process. To do so, these samples are divided into pairs randomly. For each pair, the human user gives input on which sample is closer to the the query object. If one sample is more likely to be positive than the other, they are assigned to the ordered set $\mathbf{O}$. Otherwise, if these two images appears to have similar affinity to the query, they are added to the unordered set $\mathbf{U}$. Thus, if 20 samples are used, we only need to make 10 inputs. Then $\mathbf{O}$ and $\mathbf{U}$ are used to train a new ranking SVM model for incremental learning. Note that some comparison results have already been available from previous iterations, we keep these results in $\mathbf{O}$ and $\mathbf{U}$, and only add the newly ordered pairs to these two sets.

The proposed query expansion task follows an iterative process. Once the ranking SVM is re-trained, it is used to predict the rankings of representative samples again, so as to start a new active learning iteration. This process terminates until a stopping criterion is satisfied. Normally the active learning process terminates when a classifier yields the highest performance on a validation set [47]. In the proposed method, we use the changes of entropy values between two iterations as the stopping criterion. The initial ranking SVM model may have high uncertainty, which leads to high total data entropy at the first iteration. The entropy then decreases as the model is gradually refined in the following iterations. An example of such process is shown in Fig. 7. We define the stopping criterion as the condition that the entropy change between two adjacent iterations is less than 10%. In other words, if the total entropy value of all samples is more than 90% of that of the last iteration, the iteration is stopped. Let x_i be a training image, and $\tilde{x}_i$ be the entropy of x_i, this criterion can be formally defined as follows

$$\frac{\sum_1^N \tilde{x}_i^{t+1}}{\sum_1^N \tilde{x}_i^t} < 0.9 \tag{8}$$

where t indices the iteration number and N is the total number of samples.

After the ranking SVM model has been improved by the active learning method, it is used to

classify all image patches in the target set. If the ranked score of an image patch is larger than a threshold DT, we consider an object in the same class as the query object has been detected, and vice versa.

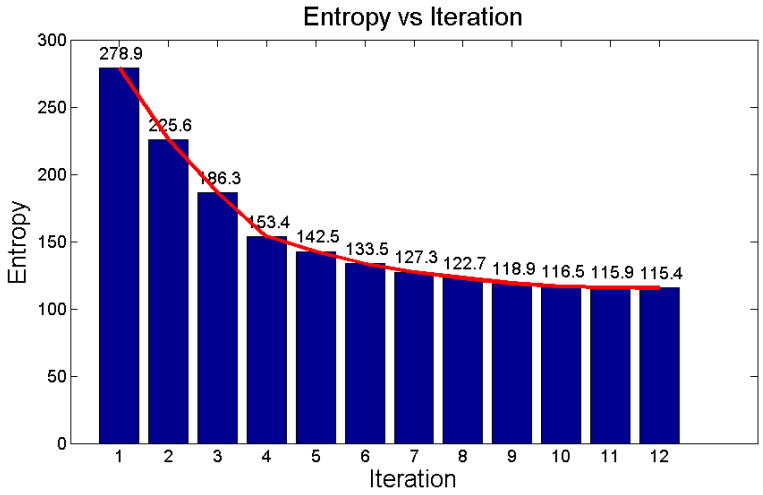


Fig. 7. An example of how entropy changes in each iteration. This example uses an airplane as the query.

In summary, the novelty of the proposed active learning framework lies in that it is not necessary to decide which image sample contains an object in the same class as the query. This property is particularly useful for object detections tasks, in which determining whether an image patch contains an object or not may be difficult because it contains only part of an object. In these cases, our method focuses the attention on which sample is closer to the query target than another and takes full advantage of the ranking SVM model.

V. Experimental results

In this section, we validate the effectiveness of several important steps of the proposed method. Firstly, the proposed image descriptor and the feature augmentation process are compared with other features in terms of classification performance. Secondly, we evaluate the proposed query expansion method and its improvement over the initial query result. Finally, we analyzed the effectiveness of the ranking SVM and active learning methods. We also compare the options of active learning and random learning. All experiments were implemented on a desktop computer with an Intel Core i3 Duo 2.93-GHz processor.

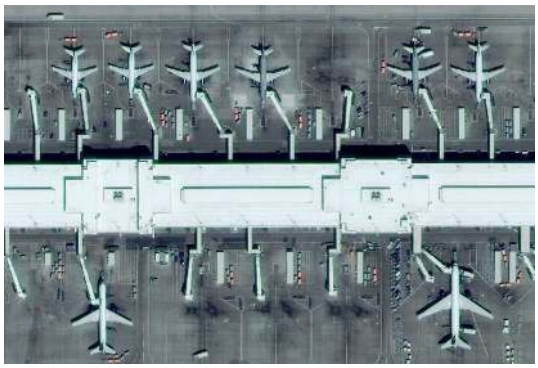
The proposed method has been used to detect objects from images captured by QuickBird and GeoEye-1 satellites. These images were downloaded from *GodEyes* website[1]. All the QuickBird and GeoEye-1 images are in the size of 3000×3000 pixels, but different in ground resolution (0.6m and 0.5m, respectively). We have used 200 images in the experiments. Each image contains one or more classes of the six classes of objects, i.e., airplanes, ships, houses, stadiums, bridges and vegetated land. The numbers of instances, i.e., objects, and query images in each class are shown in Table I. Objects in these images vary in size, shape, orientation and background. Fig. 8 shows two airplane examples cropped from two images in the target dataset. The query images are selected from the downloaded images. They are image patches with the object occupying the majority of the image. Some sample query images are displayed in Fig. 9.

TABLE I

NUMBER OF INSTANCES AND QUERY IMAGES IN EACH CLASS.

Target objects	airplanes	ships	houses	stadiums	bridges	vegetated land	Total
Number of instances	375	256	381	93	47	208	1360
Number of query images	15	10	15	10	10	10	70

(a) QuickBird image sample



(b) Geoeye-1 image sample

Fig. 8. Examples of airplane images.

As mentioned previously in Section II, images in the target dataset are segmented into small patches using sliding window method. We have assigned the sizes of the sliding windows to be

[1]http://www.godeyes.cn/

Fig. 9. Query image examples in six classes which include airplanes, ships, houses, stadiums, bridges, and vegetated land.

0.5, 1, 1.5 and 2 times the size of the query image, respectively. The number of sliding windows is expanded with various aspect ratios by stretching the height of the windows to 0.6, 1 and 1.5 times of their original sizes. These patches may or may not contain objects in the same class as the query, and sometimes, may contain multiple or parts of target objects. It should be mentioned here that there is no optimum sliding window size for a particular object class due to the variation of objects. Rather, it is expected that the proposed system will automatically pick the one with the best image descriptor at each location.

In all experiments, 5% image patches were randomly selected from the target dataset to form the representative samples for feature augmentation and classifier learning, and the rest were used for object detection. An exception was in the active learning evaluation experiment, in which we varied the number of representative samples from 1% to 10% of the target dataset to show the influence of training set size. The training patches were converted to feature descriptors using the proposed feature extraction method. To construct the initial feature descriptor, we have empirically set the total number of codewords K to be 200. Then the descriptors were updated in the feature augmentation step. The iterative classifier learning stage followed an active learning process, which used human inputs to update the ranking SVM model. In this step, we set the trade-off parameter γ in Equation 5 to 0.5.

After training, the learned ranking SVM classifier was used to classify the target set for object detection. For sliding windows in different scales and ratios, there may be significant overlap on detected target. For example, several sliding windows may contain the same object or object parts. Some windows may contain several objects. To solve these problems, a non-maxima suppression step is introduced to retain only the sliding window candidate with the highest ranked score amongst the overlapping ones.

A. Analysis on image descriptor and feature augmentation

In this section, we compare the proposed image descriptor against related baseline methods with or without feature augmentation step. The baseline feature descriptors include SIFT and LSS (local self-similarity [19]). LSS is a structural feature which describes the similarity between a small region and its neighboring regions in an image. It offers a simple unified way to describe the internal relations in the image. We used the bag-of-words method (BOW) [40] and used the ranking SVM to perform the detection task. BOW is the most popular coding scheme for image classification. The idea is grouping feature descriptors (in our experiments, SIFT and LSS) into clusters, and then calculating the frequency that features in an image be assigned to each cluster. This process converts an image into a vectorized form which can be used to train a classifier for image classification. The methods were tested on all six object classes. For quantitative evaluation, we manually labelled the target image patches.

The universal mean average precision (MAP) was adopted to evaluate the object detection performance. Given a set of queries, MAP is the mean of the average precision scores for each query, which is computed as

$$MAP = \frac{1}{Q}\sum_{q=1}^{Q} AP(q) \tag{9}$$

where Q is the number of queries. $AP(q)$ is the average precision of the q-th query, which is defined as

$$AP = \frac{\sum_{k=1}^{n}(P(k) \times r(k))}{|R(q)|} \tag{10}$$

where $|R(q)|$ is the number of images relevant to the query q. k is the rank in the sequence of retrieved images, n is the total number of retrieved images. $P(k)$ is the precision at cut-off k in the list, and $r(k)$ is an indicator function whose value is 1 if the image at rank k is relevant, and is 0 otherwise [48].

TABLE II

OBJECT DETECTION PERFORMANCE WITH DIFFERENT FEATURES AND THE EFFECT OF FEATURE AUGMENTATION.

Target objects	SIFT+BOW		LSS+BOW		OURS	
	regular	augmented	regular	augmented	regular	augmented
airplanes	77.1%	82.5%	77.4%	81.9%	88.2%	91.6%
ships	76.0%	82.2%	75.5%	81.7%	85.4%	92.0%
houses	74.8%	79.4%	73.2%	76.7%	83.6%	87.8%
stadiums	73.1%	78.1%	75.5%	78.3%	84.0%	90.2%
bridges	74.5%	74.5%	70.2%	72.3%	85.1%	87.2%
vegetated land	79.8%	81.7%	76.9%	79.8%	85.1%	86.5%
Overall	76.3%	80.5%	75.2%	79.2%	85.8%	89.9%

Based on the above criterion, the experimental results are shown in the "regular" column of Table II. To make a fair comparison, all three methods were tested on the same process except that the image descriptors were different. This experiment was done via the proposed query expansion process but without feature augmentation. From Table II, it can be seen that the classification accuracy of the proposed image descriptor significantly exceeds those of the traditional bag-of-words representations on all six target classes. This is due to the contribution of the circular and spatial pyramid description, which provides more precise information on spatial distribution of features than simple statistical description of images.

We also tested the proposed feature augmentation process on all three descriptors. The results are shown in the "augmented" column of Table II. This step has delivered an average of 4% improvement over the baseline methods. This includes more than 5% improvement on the *ships* and the *stadiums* classes over all descriptors considered. The reason is that these two classes are influenced by more background noise. After the feature augmentation process, most noisy features are ignored, leading to a remarkable accuracy improvement. The LSS descriptor has been improved the least among the three descriptors. This is because LSS feature can partly reduce some background noise.

In the above feature augmentation process, several free parameters have to be set to proper values. These include the similarity threshold S and the validation threshold M defined in Section III-B, and the detection threshold (DT) defined in Section IV-C. To show how their values affect the overall performance of the proposed method, we performed experiments on all

the six categories and evaluate the detection accuracy. When evaluating one parameter, the other two were fixed. The results are shown in Fig. 10. We chose the best values, and set S, M, and DT to 1, 10, and 0.8, respectively.

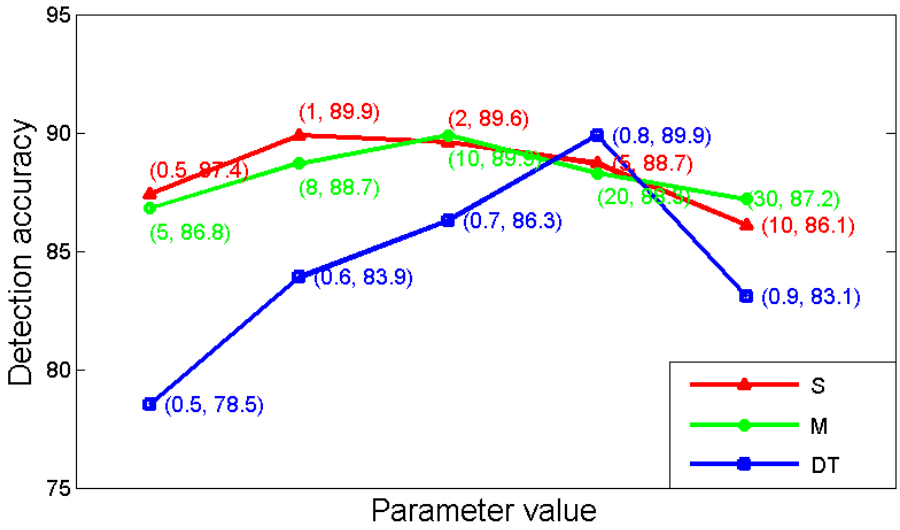


Fig. 10. Influence of key parameters to the detection accuracy. These parameters are similarity threshold (S), validation threshold (M), and detection threshold (DT). The number pairs give the parameter values and their corresponding detection accuracies.

B. Query expansion evaluation

In this section, we evaluate the proposed query expansion method by comparing it with three other baseline retrieval methods. Baseline 1 is the basic detection method without query expansion, which is also our initial retrieval result. Baseline 2 is a regular query expansion method. The query image representation is replaced by the mean image vectors of the top 20% of the initial retrieval result. Baseline 3 is the proposed method but with regular SVM instead of ranking SVM.

For baseline 3, we treated the detection task as a binary classification problem. So long as a target image patch contains an object in the same class as the query, even if multiple objects in different classes may present, it is considered as positive. Therefore, an image patch may be considered as positive for multiple classes. For the binary SVM model, we adopted the C-SVC in LIBSVM [49] with an RBF kernel. The parameters for the SVM were obtained by

TABLE III

COMPARISON OF DIFFERENT TARGET DETECTION METHODS. (MAP)

Area	Baseline 1	Baseline 2	Baseline 3	Ours
airplanes	68.7%	78.9%	89.1%	92.3%
ships	65.2%	75.3%	87.5%	92.4%
houses	58.0%	76.4%	83.4%	90.5%
stadiums	71.9%	74.7%	87.0%	88.7%
bridges	80.1%	83.3%	87.2%	90.7%
vegetated land	68.1%	76.6%	84.8%	87.5%
Overall	68.6%	77.5%	86.5%	90.4%

cross-validation on a subset of manually labeled representative samples. For example, the kernel parameter Γ and cost parameter C were set to 2 and 0.5, respectively, for the airplane model. For the active learning process, user provided labels to the top 20% representative samples with the highest entropy as calculated in equation (7), given the probabilistic output from the SVM model.

Let the total number of objects be (NP), the correct number of detections be the true positives (TP), and the number of spurious detections of the same object be false positives (FP), then precision-recall curve can be plotted. Precision is the number of retrieved positive images out of the total number of images retrieved, i.e., $Precision = TP/(TP + FP)$, while recall is the number of retrieved positive images out of total number of positives in the target dataset, i.e., $Recall = TP/NP$. The area under the precision-recall curve is related to how well the methods under comparison have performed.

The experiments followed the same feature extraction and augmentation steps for all four methods being compared. The proposed image descriptor was used for feature extraction. The testing results are given in Table III. The precision-recall curves are shown in Fig. 11. The experimental results show that the proposed active learning based ranking SVM method has achieved the best results on all six classes of objects. The performance is followed by baseline 3, baseline 2, and baseline 1, in a descending order, respectively. Baseline 1 method achieved the lowest accuracy because its simple model cannot provide enough information to represent an object category accurately, which proves the effectiveness of query expansion methods.

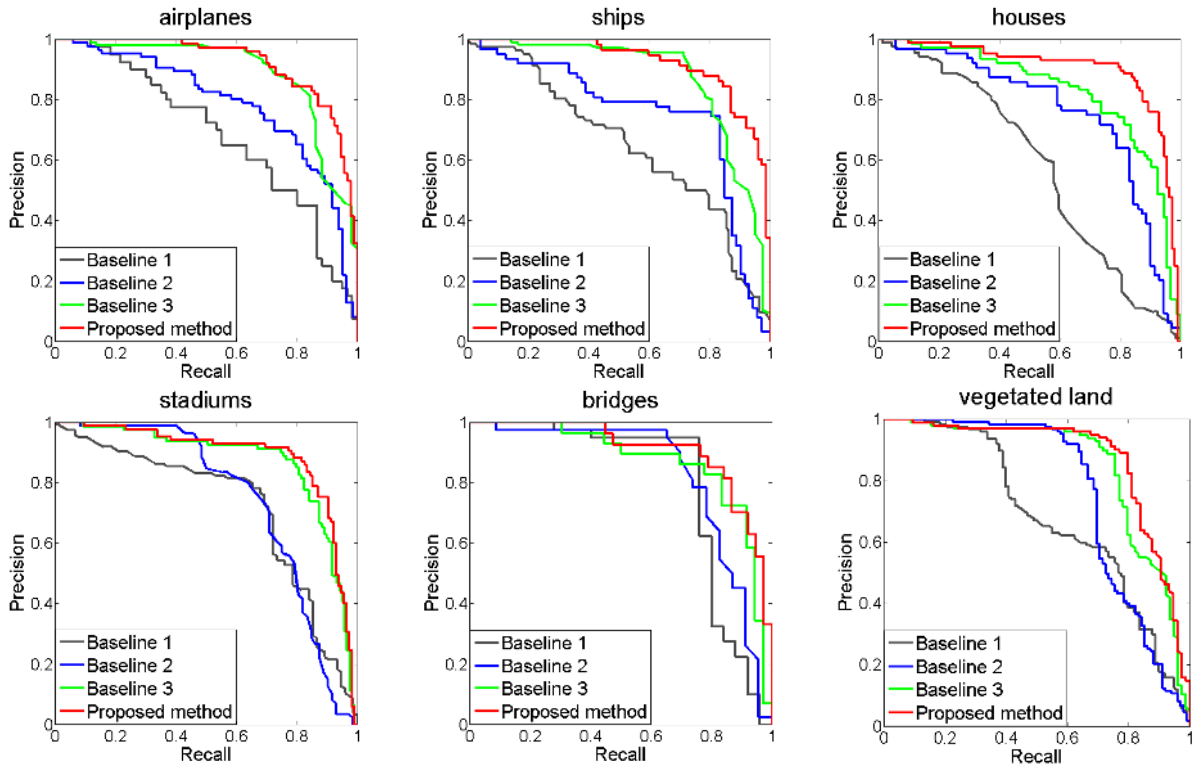


Fig. 11. Precision-recall curves for different detection methods of six object classes.

Furthermore, active learning has significantly refined the prediction model, which leads to improvement on the object detection performance. The fact that both regular SVM and ranking SVM have achieved high accuracy has verified the effectiveness of the proposed query expansion with active learning for object detection.

We also compared the active learning method with a random learning strategy. This strategy randomly selected representative samples from the image patches in the training set rather than using the highest entropy ones. Furthermore, we evaluated the influence of the number of representative samples in these two settings. The number of representative samples varied from 1% to 10% of the total target patches. The experiments were done on all six categories, with the overall accuracy presented in Table IV.

It can be observed from this table that when the number of representative sample increases, the detection accuracy is improved. This is natural because more representative samples provide richer information on the data distribution, allowing better classifiers to be built. The proposed active learning process has outperformed the random learning method under all training number

TABLE IV

OBJECT DETECTION PERFORMANCE (MAP) WITH DIFFERENT TRAINING STRATEGY.

Number of Representative Samples	1%	3%	5%	7%	10%
active learning	80.8%	86.2%	89.6%	90.8%	91.2%
random learning	74.1%	79.5%	85.7%	86.8%	87.4%

conditions, especially when the number of representative samples is small. This is because active learning process can find more informative samples than the random learning method by using the entropy theory.

We show some object detection results in Fig. 12. These image patches were cut out from the original images (3000×3000 pixels in size) in the target set. The green rectangles show the correct detection results, while the red ones represent wrong detections or missed detections. This figure shows that the proposed method can achieve very good object detection results even in some difficult conditions, i.e., objects with different orientation, or blurred object boundaries by background noise.

It should be mentioned here that it is difficult to evaluate the cost of user intervention in the proposed active learning process due to the possible different numbers of iterations for each query, the randomness of training set generation, and the uncertainty on time that different users take to make judgment on the ranking of image pairs. Based on our experience, 20% of the samples in training set is a sound amount of data to allow training of a good ranking SVM model, while not taking too long for the user intervention.

VI. CONCLUSION

In this paper, we tackle the problem of target detection in VHR images. Unlike previous researches on object detection that usually adopt binary classification solution, we propose to learn ranking models, and employ image pairs with preference relationships to update the model. By using an active learning based ranking SVM method, we have learned a refined model for final detection. In addition, we also propose a novel image descriptor which considers not only local features but also their spatial relationships. This method gives more accurate information for object description. Moreover, a feature augmentation process is proposed to improve the image descriptor and make features more distinctive. The method has achieved very good performance

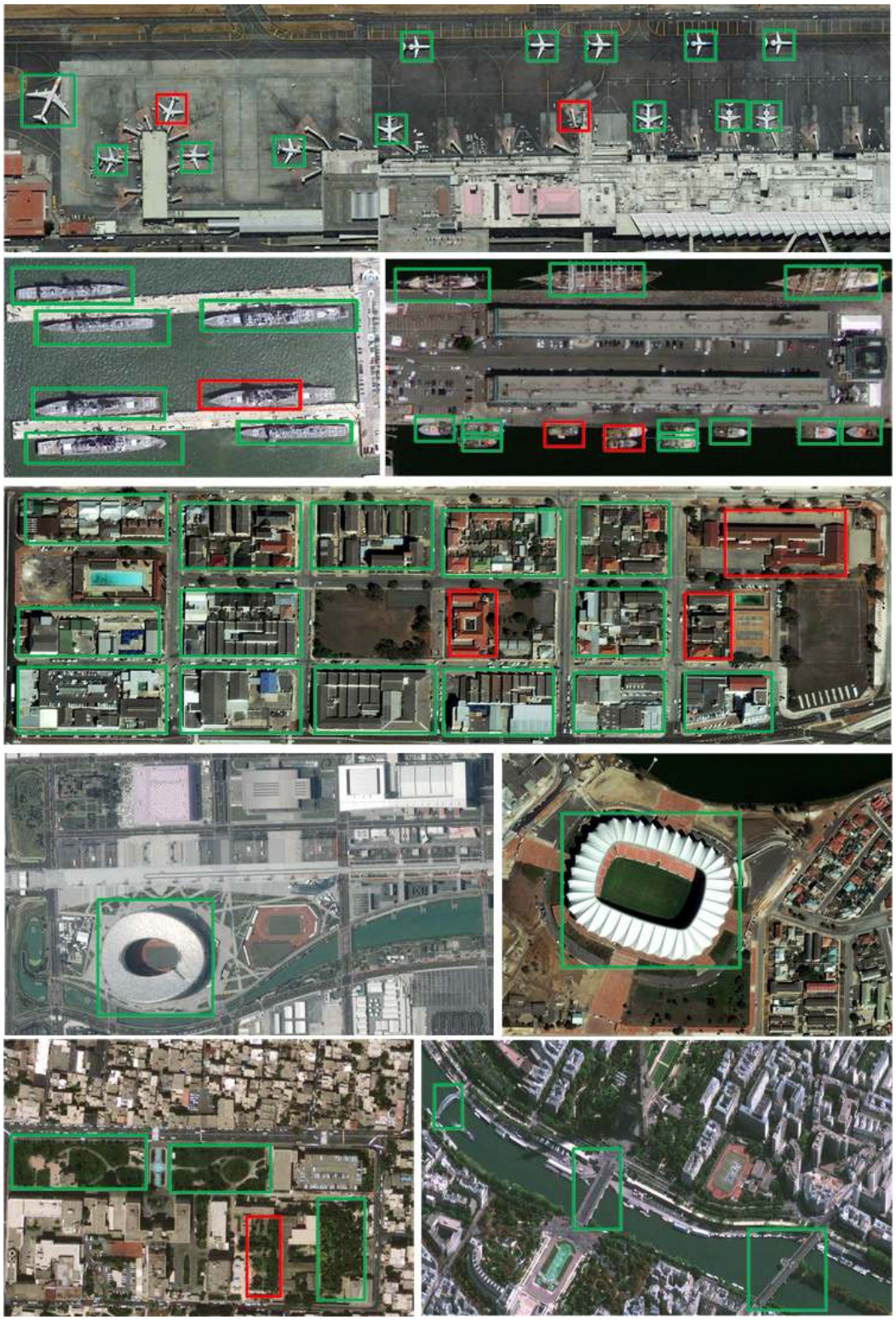


Fig. 12. Detection examples of six object classes. They are airplanes, ships, houses, stadiums, bridges, and vegetated land, respectively.

on VHR satellite images, and has significantly outperformed several baseline methods. For future study, we would like to explore spectral information and combine multispectral image data with VHR images for more accurate object detection.

ACKNOWLEDGMENT

This work was supported in part by NSFC under Grant 61105002 and Australian Research Councils DECRA Projects funding scheme (project ID DE120102948).

REFERENCES

[1] C. Unsalan, "Measuring land development in urban regions using graph theoretical and conditional statistical features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 12-1, pp. 3989–3999, 2007.

[2] W. Liu, F. Yamazaki, and T. T. Vu, "Automated vehicle extraction and speed determination from QuickBird satellite images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 4, no. 1, pp. 75–82, 2011.

[3] I. T. K. Aksoy, S.; Yalniz, "Automatic detection and segmentation of orchards using very high resolution imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 8, pp. 3117–3131, 2012.

[4] D. Brunner, G. Lemoine, and L. Bruzzone, "Earthquake damage assessment of buildings using VHR optical and SAR imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 5, pp. 2403 – 2420, 2010.

[5] D. Chaudhuri and A. Samal, "An automatic bridge detection technique for multispectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 9, pp. 2720–2727, 2008.

[6] H. G. Akcay and S. Aksoy, "Automatic detection of geospatial objects using multiple hierarchical segmentations," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 7, pp. 2097–2111, 2008.

[7] L. Gueguen, M. Pesaresi, A. Gerhardinger, and P. Soille, "Characterizing and counting roofless buildings in very high resolution optical images," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, no. 1, pp. 114–118, 2012.

[8] K. Stankov and D. He, "Building detection in very high spatial resolution multispectral images using the hit-or-miss transform," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 1, pp. 86–90, 2013.

[9] J. Leitloff, S. Hinz, and U. Stilla, "Vehicle detection in very high resolution satellite images of city areas," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 7, pp. 2795–2806, 2010.

[10] C. Zhu, H. Zhou, R. Wang, and J. Guo, "A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 9, pp. 3446–3456, 2010.

[11] X. Jin and C. Davis, "Automated building extraction from high-resolution satellite imagery in urban areas using structural, contextual, and spectral information," *EURASIP Journal on Applied Signal Processing*, pp. 2196–2206, 2005.

[12] Q. Wang, Z. Jiang, J. Yang, D. Zhao, and Z. Shi, "A hierarchical connection graph algorithm for gable-roof detection in aerial image," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 1, pp. 177–181, 2011.

[13] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[14] B. Sirmacek and C. Unsalan, "Urban-area and building detection using SIFT keypoints and graph theory," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 4, pp. 1156–1167, 2009.

[15] ——, "Urban area detection using local feature points and spatial voting," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 1, pp. 146–150, 2010.

[16] G. Mountrakis and L. Luo, "Enhancing and replacing spectral information with intermediate structural inputs: A case study on impervious surface detection," *Remote Sensing of Environment*, vol. 115, no. 5, pp. 1162–1170, 2011.

[17] N. Li, H. Huo, and T. Fang, "A novel texture-preceded segmentation algorithm for high-resolution imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 7, pp. 2818–2828, 2010.

[18] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[19] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[20] P. Turcot and D. Lowe, "Better matching with fewer features: The selection of useful features in large database recognition problems," in *12th IEEE International Conference on Computer Vision Workshops*. IEEE, 2009, pp. 2109–2116.

[21] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2911–2918.

[22] D. Duro, S. Franklin, and M. Dubé, "Multi-scale object-based image analysis and feature selection of multi-sensor earth observation imagery using random forests," *International Journal of Remote Sensing*, vol. 33, no. 14, pp. 4502–4526, 2012.

[23] X. Chen, T. Fang, H. Huo, and D. Li, "Graph-based feature selection for object-oriented classification in VHR airborne imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 1, pp. 353–365, 2011.

[24] D. Tuia, G. Camps-Valls, G. Matasci, and M. Kanevski, "Learning relevant image features with multiple-kernel classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 10, pp. 3780 – 3791, 2010.

[25] B. Settles, "Active learning literature survey," *University of Wisconsin, Madison*, 2010.

[26] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 606–617, 2011.

[27] M. Ferecatu and N. Boujemaa, "Interactive remote-sensing image retrieval using active relevance feedback," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 4, pp. 818–826, 2007.

[28] B. Demir, C. Persello, and L. Bruzzone, "Batch-mode active-learning methods for the interactive classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 3, pp. 1014–1031, 2011.

[29] B. Demir, F. Bovolo, and L. Bruzzone, "Detection of land-cover transitions in multitemporal remote sensing images with active-learning-based compound classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 5, pp. 1930–1941, 2012.

[30] D. Tuia, E. Pasolli, and W. Emery, "Using active learning to adapt remote sensing image classifiers," *Remote Sensing of Environment*, vol. 115, no. 9, pp. 2232–2242, 2011.

[31] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 3, pp. 247–259, 2011.

[32] B. Waske, S. Van Der Linden, J. Benediktsson, A. Rabe, and P. Hostert, "Sensitivity of support vector machines to random feature selection in classification of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 7, pp. 2880–2889, 2010.

[33] N. Segata, E. Pasolli, F. Melgani, and E. Blanzieri, "Local SVM approaches for fast and accurate classification of remote-sensing images," *International Journal of Remote Sensing*, vol. 33, no. 19, pp. 6186–6201, 2012.

[34] A. Salberg and R. Jenssen, "Land-cover classification of partly missing data using support vector machines," *International Journal of Remote Sensing*, vol. 33, no. 14, pp. 4471–4481, 2012.

[35] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 133–142.

[36] Y. Hu, M. Li, and N. Yu, "Multiple-instance ranking: Learning to rank images for image retrieval," in *IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[37] D. Parikh and K. Grauman, "Relative attributes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 503–510.

[38] S. Ma, S. Sclaroff, and N. Ikizler-Cinbis, "Unsupervised learning of discriminative relative visual attributes," in *the 12th ECCV Workshop on Parts and Attributes*. Springer, 2012, pp. 1–10.

[39] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.

[40] F. Li, R. Fergus, and A. Torralba, "Recognizing and learning object categories," in *IEEE International Conference on Computer Vision, Short Course*, 2005.

[41] G. Hamerly and C. Elkan, "Alternatives to the k-means algorithm that find better clusterings," in *Proceedings of the eleventh International Conference on Information and Knowledge Management*, vol. 4, no. 09, 2002, pp. 600–607.

[42] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 2169–2178.

[43] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[44] O. Chapelle, "Training a support vector machine in the primal," *Neural Computation*, vol. 19, no. 5, pp. 1155–1178, 2007.

[45] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[46] L. Li and L. Fei-Fei, "Optimol: automatic online picture collection via incremental model learning," *International journal of computer vision*, vol. 88, no. 2, pp. 147–168, 2010.

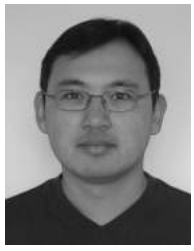
[47] F. Olsson and K. Tomanek, "An intrinsic stopping criterion for committee-based active learning," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, 2009, pp. 138–146.

[48] M. Zhu, "Recall, precision and average precision," *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo*, 2004.

[49] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Xiao Bai received the B.Eng. degree in computer science from Beihang University of China, Beijing, China, in 2001, and the Ph.D. degree from the University of York, York, U.K., in 2006. He was a Research Officer (Fellow, Scientist) in the Computer Science Department, University of Bath, until 2008. He is currently an Associate Professor in the School of Computer Science and Engineering, Beihang University. He has published more than forty papers in journals and refereed conferences. His current research interests include pattern recognition, image processing and remote sensing image analysis. He has been awarded New Century Excellent Talents in University in 2012.

Huigang Zhang received the Bachelors degree in mathematics from Hebei University of Technology, Tianjin, China, in 2010, and the M.Tech. degree in computer science from Beihang University, Beijing, China, in 2013. His current research interests include structural pattern recognition, statistical machine learning, and image processing.



Jun Zhou (M'06, SM'12) received the B.S. degree in computer science and the B.E. degree in international business from Nanjing University of Science and Technology, China, in 1996 and 1998, respectively. He received the M.S. degree in computer science from Concordia University, Canada, in 2002, and the Ph.D. degree in computing science from University of Alberta, Canada, in 2006.

He joined the School of Information of Communication Technology at Griffith University, Nathan, Australia, as a Lecturer in June 2012. Previously, he had been a Research Fellow in the Research School of Computer Science in the Australian National University, Acton, Australia, and a Researcher in the Canberra Research Laboratory, NICTA, Canberra, Australia. His research interests include pattern recognition, computer vision, and machine learning with human in the loop, with their applications to spectral imaging and environmental informatics.