ISPRS Journal of Photogrammetry and Remote Sensing 99 (2015) 30–44

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

A generic discriminative part-based model for geospatial object detection in optical remote sensing images

Wanceng Zhang [a,b,c], Xian Sun [b,c], Hongqi Wang [b,c], Kun Fu [b,c,*]

[a] University of Chinese Academy of Sciences, Beijing, China
[b] Institute of Electronics, Chinese Academy of Sciences, Beijing, China
[c] Key Laboratory of Technology in Geo-spatial Information Processing and Application System, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 14 April 2014
Received in revised form 14 August 2014
Accepted 29 October 2014
Available online 20 November 2014

Keywords:
Geospatial object detection
Part-based model
Rotation invariance
Deformation feature

ABSTRACT

Detecting geospatial objects with complex structure has been explored for years and it is still a challenging task in high resolution optical remote sensing images (RSI) interpretation. In this paper, we mainly focus on the problem of rotation variance in detecting geospatial objects and propose a generic discriminative part-based model (GDPBM) to build a practical object detection framework. In our model, a geospatial object with arbitrary orientation is divided into several parts and represented via three terms: the appearance features, the spatial deformation features and the rotation deformation features. The appearance features characterize the local patch appearance of the object and parts, and we propose a new kind of rotation invariant feature to represent the appearance using the local intensity gradients. The spatial deformation features capture the geometric deformation of parts by representing the relative displacements among parts. The rotation deformation features define the pose variances of the parts relative to the objects based on their dominant orientations. In generating the two deformation features, we introduce the statistic methods to encode the features in the category level. Concatenating the three terms of the features, a classifier based on the support vector machine is learned to detect geospatial objects. In the experiments, two datasets in optical RSI are used to evaluate the performance of our model and the results demonstrate its robustness and effectiveness.

1. Introduction

With the development of airborne and spaceborne sensors, an increasing amount of data is being produced in an ever broader range. Meanwhile, the imagery from the commercial satellite with high resolution better than half a meter has become operational, such as Worldview-1 with the resolution 0.44 m panchromatic (Blaschke, 2010). In this way, a great deal of spatial patterns and hidden information can be achieved to analyze the Earth's surface, and object detection as one of the fundamental applications in RSI can provide more detail information. So the task of building an efficient and practical framework to describe and detect objects in high resolution optical RSI is necessary, and many works have been proposed until now. However, it is still a big challenge to detect geospatial objects with complex inner structure, and the abundant spatial information has not been explored well to conquer the disturbing background.

In high resolution optical RSI, kinds of low-level features, such as the texture, shape and structure information, can be achieved, which have drawn the attention of the research community in computer vision for years (e.g., Hinz and Stilla, 2006; Grabner et al., 2008; Yang and Newsam, 2013). Many practical methods have been researched and applied in object detection with different focuses on high resolution RSI. For example, Grabner et al. (2008) proposed an on-line version of boosting to detect cars from aerial images, employing the local features. Sun et al. (2010) proposed the learned taxonomic semantics to combine object detection with segmentation in Latent Dirichlet Allocation models. Lei et al. (2012) introduced Hough forest to learn the explicit codebook of objects that enhanced with color space to localize objects in RIS. Dealing with object detection problem, researchers also would like to treat it as a classification problem and many successful classifiers have been applied, such as the boosting (Grabner et al., 2008), the extended template matching (Akcay and Aksoy, 2008; Sirmaçek and Ünsalan, 2009, 2010), the *k*-nearest neighbor

* Corresponding author at: Institute of Electronics, Chinese Academy of Sciences, Beijing, China.
E-mail address: kunfuiecas@gmail.com (K. Fu).

http://dx.doi.org/10.1016/j.isprsjprs.2014.10.007
0924-2716/

(Cheng et al., 2013b), and the support vector machines (SVM) (Mountrakis et al., 2011; Bi et al., 2012; Sun et al., 2012). Among the classifiers, the SVM has also been successfully used in other applications in RSI, such as changing detection (Nemmour and Chibani, 2006), and classification (Walker, 2004; Bruzzone and Carlin, 2006; Mallinis et al., 2008; Xu et al., 2010), which inspire us to study SVM with our detection model.

Over the past decades, there are generally two main problems to face in the field of object detection in either ground-shot images or high resolution optical RSI: the descriptors to choose and the model to fuse the descriptors. But unlike the objects with upright direction in the ground-shot images due to the gravity, the objects in RSI are usually in arbitrary orientations, since the images are taken from the vertical viewpoint. The problem of rotation variance has to be considered when choosing the descriptor and model in geospatial object detection. Some researchers address the two problems together to select the feature and the model that are common to all directions, but they do not utilize the explicit structure information of objects directly (Sirmaçek and Ünsalan, 2011; Li et al., 2012; Sun et al., 2012; Swearingen and Cheriyadat 2012; Han et al., 2013). For example, Sun et al. (2012) proposed a spatial sparse coding Bag-of-Word (BoW) model for target detection, which applied the scale-invariant feature transform (SIFT) descriptor (Lowe, 2004) and the corner detector (He and Yung, 2004) to generate the dictionary of objects. Li et al. (2012) proposed a contour-based spatial model to detect objects with multiple segmentations. These approaches paid more attention to the common property of images, but they lost the specific property of objects and they are easily disturbed by the complex background. Another group takes full advantage of the spatial information of objects to incorporate the descriptors with models, such as the geometric or inner structure information (e.g., Inglada, 2007; Cheng et al., 2013a; Zhang et al., 2014). These models cover the possible orientations of the objects to achieve rotation invariance. For example, Cheng et al. (2013a) introduced a discriminatively trained mixture model (Felzenszwalb et al., 2010) to detect geospatial objects. The mixture model for an object included a number of independent part-based models, and each part-based model was trained in a certain range of orientations based on the histogram of oriented gradients (HOG) (Dalal and Triggs, 2005) features. This type of strategy augments the original data to train the model or the descriptor with different orientations, and there is not a single common model to cover the rotation invariance. With the increasing resolution of RSI, the geospatial or structure information of objects has drawn more attention for the object description, however, the problem of rotation invariance has not been handled well to combine with the structure information to detect objects.

In this paper, we make a further study on the object detection in high resolution optical RSI, focusing on the problem of rotation variance. The method here depicts the inner structure information of objects in a single generic model, and it is adapt to the object with arbitrary orientation. Our model is extended from the discriminatively trained part based models (DTPBM) (Felzenszwalb et al., 2008, 2010), which have been successfully used to detect objects on difficult benchmarks such as the PASCAL dataset (Everingham et al., 2007). However, the DTPBM are rotation variant, since the HOG features they use are only adapt to the objects with small rotation, and their spatial deformation feature would be correct when the object is *upright*. The above shortcomings make the DTPBM unable to be applied directly in geospatial object detection.

Among the existing studies, the most related work is our rotation invariant part-based model (RIPBM) (Zhang et al., 2014), in which an object contained several models in different orientations. But the part-based model we proposed here is based on different image features and deformation features, and the model describes the geospatial object in a more general way by three terms: the appearance features, the spatial deformation features and the rotation deformation features. The contributions comparing with our previous work are as follows. Firstly, the previous work used the rotation invariant HOG (RIHOG) descriptor to represent the appearance of the object and parts, where each patch would have several features in different dominant orientations. But here we propose a new rotation invariant appearance feature to capture the more detail information of objects and parts instead. Each patch only has a unique discriminative feature. Secondly, for the deformation features generation, the deformation features in RIPBM were calculated from each sample independently, and their rotation deformation was dependent on the appearance feature. Here, we introduce the statistic methods to generate the two deformation features. The statistic methods focus on the common property of the same category from all the training samples, and they make the rotation deformation features independent, which allow the better description of intra-class variability. Besides, in labeling the results, the search strategy here considers the objects with different sizes, which makes our model more robust than RIPBM. Integrated with a multi-scale sliding window mechanism, a more flexible model is explored to detect objects in high resolution optical RSI.

The remainder of the paper is organized as follows. Section 2 gives the overview of our object detection framework based on GDPBM, followed by Section 3 which describes the details of our model in representing geospatial objects and the searching method in the detection step. Section 4 shows the experiments of our model. The conclusions are given in Section 5.

2. Framework overview

Fig. 1 illustrates the procedure of geospatial object detection framework based on our model. There are two phases in application: model training and object detection.

In the first phase, our model is trained using the supervised data. Each positive sample is manually labeled with bounding boxes for the object and parts. Firstly, the appearance features are encoded on the labeled regions based on our new rotation invariant feature (see Section 3.1.1). Before the deformation features generation, we count the distributions of the spatial positions among parts and their dominant orientations from the positive samples as in the center of Fig. 1. The distributions are obtained in rotation invariant manners and they are utilized to generate the spatial and rotation deformation features (see Sections 3.1.2 and 3.1.3). The spatial deformation features address the geometric deformation of parts relative to their "anchor" positions, and the rotation deformation features define the "pose" variances of parts relative to their "ideal" directions in the object. Then, the appearance features, spatial deformation features and rotation deformation features are concatenated to train a whole SVM classifier for our GDPBM. Meanwhile, the SVM classifiers for each part are trained independently based on the appearance features.

In the detection phase, given an image, firstly the appearance SVM classifiers are utilized to find the latent objects and their parts to build the hypothesis space. To reduce the scale of the hypothesis space, the distributions of the spatial and rotation statistics from the training phase are applied to filter the candidates by thresholds in probabilities, and they are used to encode the spatial and rotation deformation features for the left ones. Then, we concatenate the above three kinds of features, and give the last responses by the whole SVM classifier at each hypothesis. Finally, a clustering method is introduced to fuse the detection results at each position with different sizes, and the labels of the objects are given in the image (see Section 3.2).