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

A R T I C L E   I N F O

Article history:
Received 23 April 2014
Received in revised form 14 October 2014
Accepted 14 October 2014
Available online 9 November 2014

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

A B S T R A C T

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
0924-2716/

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; Vǎduva 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.

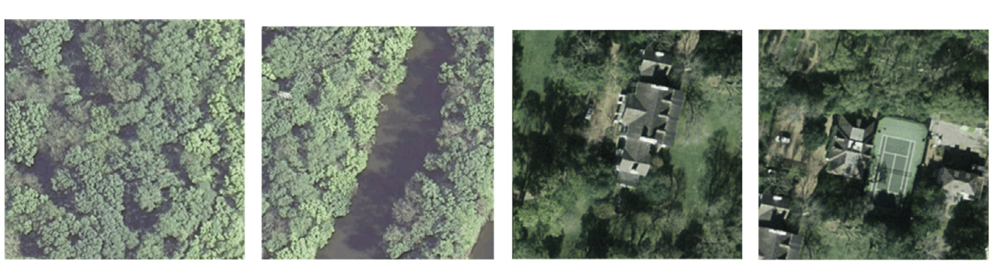

Fig. 1. Four images from a publicly available 21-class LULC data set (Yang and Newsam, 2010, 2011). (a) Forest. (b) River. (c) Sparse residential. (d) Tennis court.