Computer Vision and Image Understanding 135 (2015) 31–48

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Combine color and shape in real-time detection of texture-less objects ☆

Xiaoming Peng *

School of Automation Engineering, University of Electronic Science and Technology of China, Qingshuihe Campus, No. 2006, Xiyuan Ave, West Hi-Tech Zone, 611731, Chengdu, Sichuan, China
School of Computer Science and Software Engineering, The University of Western Australia, M002, 35 Stirling Highway, Crawley, WA 6009, Australia

ARTICLE INFO

Article history:
Received 27 July 2014
Accepted 21 February 2015
Available online 7 March 2015

Keywords:
Real-time texture-less object detection
The Dominant Orientation Templates (DOT) method
Color name
Speed-up strategy

ABSTRACT

Object instance detection is a fundamental problem in computer vision and has many applications. Compared with the problem of detecting a texture-rich object, the detection of a texture-less object is more involved because it is usually based on matching the shape of the object with the shape primitives extracted from an image, which is not as discriminative as matching appearance-based local features, such as the SIFT features. The Dominant Orientation Templates (DOT) method proposed by Hinterstoisser et al. is a state-of-the-art method for the detection of texture-less objects and can work in real time. However, it may well generate false detections in a cluttered background. In this paper, we propose a new method which has three contributions. Firstly, it augments the DOT method with a type of illumination insensitive color information. Since color is complementary to shape, the proposed method significantly outperforms the original DOT method in the detection of texture-less object in cluttered scenes. Secondly, we come up with a systematic way based on logistic regression to combine the color and shape matching scores in the proposed method. Finally, we propose a speed-up strategy to work with the proposed method so that it runs even faster than the original DOT method. Extensive experimental results are presented in this paper to compare the proposed method directly with the original DOT method and the LINE-2D method, and indirectly with another two state-of-the-art methods.

1. Introduction

Object instance detection is a fundamental problem in computer vision, and many applications require it as a necessary step. For instance, image-based tracking usually needs an object detection step in order to initialize the tracking. However, it is not a trivial problem because the object to be detected can undergo viewpoint, scale, and illumination changes, and sometimes can even be subject to partial occlusions. Generally, methods in this area are divided into two types. The first type of methods aims to identifying an instance of a *specific* object, while the other type of methods means to recognize an instance of an object that belongs to a *generic* object category, e.g., buildings, cars, or daily used items. In a recent tutorial, Gauman and Leibe [1] discuss both of these two types of methods. Of the methods in the first type, the solution to the detection of a texture-rich object has become quite standard and very successful [2,3], with its pipeline summarized as follows: (1) Construct a sparse model of features for an object using local features (e.g., SIFT features) extracted from multiple views of the object and the Structure from Motion (SfM) technique. (2) Match the features of the constructed model with those extracted from an input image, establishing the 3D–2D correspondences between the matched ones. After that, the pose of the object can be estimated [2] and wrong matches eliminated. This paper is focused on the detection of a specific object that has very few textures on its surface. Compared with the previous problem of detecting a texture-rich object, this problem is more involved because the detection is usually based on matching the shape of the object with the shape primitives extracted from an image, which is not as discriminative as matching appearance-based local features, such as the SIFT features. Hsiao and Hebert [4] compare invariant and non-invariant approaches for shape-based object instance detection. "Invariant" methods create a unified object representation across different viewpoints by explicitly modeling the structural relationships of high level shape primitives (e.g., curves and lines). In contrast, "non-invariant" methods use view-based templates and capture viewpoint variations by sampling the view space and matching each template independently. They conclude that non-invariant approaches are well-suited for specific object recognition and meanwhile can also be computationally efficient.

☆ This paper has been recommended for acceptance by Kyoung Mu Lee.

* Address: School of Automation Engineering, University of Electronic Science and Technology of China, Qingshuihe Campus, No. 2006, Xiyuan Ave, West Hi-Tech Zone, 611731, Chengdu, Sichuan, China.
E-mail addresses: pengxm@uestc.edu.cn, xiaoming.peng@uwa.edu.au

http://dx.doi.org/10.1016/j.cviu.2015.02.010
1077-3142/

In this paper, we propose a new method for the detection of texture-less objects in an image by extending the Dominant Orientation Templates (DOT) method proposed by Hinterstoisser et al. [5], which is one of the state-of-the-art methods for the detection of texture-less objects and can work in real time. The DOT method represents the sparse shape of an object at a given view using the dominant orientations of the gradients of the view's image. However, one issue with the DOT method is that it may well generate false detections in a cluttered background, due to the fact that many shapes partially similar to an instance of the object can exist in a cluttered background. Intuitively, the performance of shape-based object detection can be enhanced by combining with complementary information. Inspired by the very recent success of fusing color and shape information in visual object recognition [6,7], in this paper we propose a new method that incorporates color information into the DOT method. The new method significantly enhances the robustness of texture-less object detection in a cluttered background, but does not compromise the real-time speed of the original method. In the rest of this section, we will discuss the related work and argue the contributions of the paper.

1.1. Related work

In this subsection, we will mainly discuss the related work in the area of specific texture-less object detection. Also, we will touch upon the pertinent work in color-based object recognition.

1.1.1. Texture-less object detection

Since the object to be detected has very few textures on its surface, appearance-based methods are not likely to work well. Let's take one powerful appearance-based object detection method, the Implicit Shape Model (ISM) [8], as an example. Assume that one has a codebook of entries containing local appearances of a specific object observed under various viewpoints, represented by image patches extracted around interest points in the various views of the object. At the detection stage, image patches are also extracted around interest points in an input image, and compared against the entries of the codebook. The matched codebook entries then cast probabilistic Hough votes, which lead to object location hypotheses that can later be refined. If the object were rich in texture, one could expect that many of the hypotheses are generated by the matches within the "interior" of the object, which abounds with interest points. However, when the object is largely texture-less, the interest points will mostly be located along the boundaries of the object. As a result, the image patches will inevitably contain contents not belonging to the object and thus easily lead to false matches. For this reason, most methods for the detection of text-less objects are shape-based.

One type of shape-based methods relies on matching contours or edge segments. Among them, some methods [9–11] are meant to detect objects of a generic category. Ferrari et al. [9] propose a contour structure called *k*AS (*k* Adjacent Segments), which consists of *k* roughly straight contour segments adjacent to each other. They use the *k*AS histograms of the training samples of an object category to train a Support Vector Machine (SVM) classifier, which in turn is used to detect new instances of the object category. Although the *k*AS structures are invariant to scale and translation changes, the accuracy of matching them is heavily dependent on the consistent continuity of the extracted contours across images. To overcome the difficulty of one-to-one matching when the contours of the input image are fragmented, Srinivasan et al. [10] allow multiple contours of the input image to match to one contour of the object model (many-to-one matching). Since in both methods a detector only works for a particular view of the object category, if they were to be used to detect a specific object that can appear under arbitrary viewpoints, one would need to train a different detector for *each* different view of the object. Given that each detector is relatively slow (in [9], the *k*AS method requires one second to process an input image on a standard workstation with a C++ implementation; in [10], a trained detector further works in conjunction with a separate linear programming optimization procedure to perform the many-to-one matching), it is not likely for both methods to work in real time in the specific object detection case. Payet and Todorovic [11] use Bag of Boundaries (BoBs) to jointly match the contours of all the object templates to the contours of the input image, and at the same time estimate the pose of the object in the input image. Also, their method is not ready to be implemented in real time (requiring several minutes in processing a single image with a MATLAB® implementation).

By contrast, some researchers devise real-time contour-based methods particularly for the detection of texture-less specific objects [12–17]. Damen et al. [12,13] propose to match contours using constellations of edgelets, where an edgelet is a short straight segment represented by its center point and orientation. A descriptor is constructed for each constellation of edgelets to encode the relative orientations and positions of the edgelets in the constellation. The scalability of their method is achieved by using hash tables to store configurations of edgelets shared across different objects. At the detection stage, the descriptors of the various templates of an object are compared with the descriptors of the input image along a fixed-path (a path is the order of how the edgelets in a constellation are connected) to generate candidate detection results. Because the direct matching of the constellations of edgelets has limited accuracy, the candidate detection results need to be verified by establishing a homography from the matched template to the input image using the matched constellations of edgelets, followed by an Iterative Closest Edgelet (ICE) refinement which registers the template to the image. For this reason, when there are more clutters present in the input image, the speed of this method will drastically drop and meanwhile will result in significantly increased false matches. Tombari et al. [14] come up with Bunch Of Lines Descriptor (BOLD) features to describe and detect line segments in the pipeline of SIFT-feature description and detection [3]. A BOLD feature describes the relative position and orientation of a straight line segment with respect to its multiple adjacent neighbors. Obviously, their method also heavily relies on the accuracy and continuity of line segment extraction. To match edge segments at different scales, Lee and Soatto [15] construct pyramids for the template images and the input image, and extract edge segments at the various levels of the pyramids. The edge segments are then aligned to some "canonical" reference frame so that they are locally invariant to scale and rotation changes. Afterwards, the features of the edge segments can be represented and compared in a way similar to the DOT method. However, while their method works very well for the detection of planar objects, it is not likely to work equally well for the detection of 3D objects, because the transformations between the different views of a 3D object cannot be sufficiently approximated using the similarity transformation as assumed in their method. Seo et al. [16] increase the robustness of locating a texture-less object in a cluttered background with the aid of a 3D model of the object. In their method, the object location problem reduces to an optimization problem of finding the correspondences between the projections of the 3D model's contours and the edges extracted from the input image. However, they have not addressed the issue of how to initialize the camera pose which is required to start the optimization. In light of this, their method is better to be used for tracking where the camera pose can be easily initialized using the tracking results of the previous frame, rather than be used for detection where the camera pose is mostly unknown. In addition, constructing a 3D model of a texture-less object almost