Boosting masked dominant orientation templates for efficient object detection ☆

Reyes Rios-Cabrera [a,b,*], Tinne Tuytelaars [b]

[a] CINVESTAV, Robotics and Advanced Manufacturing, Av. Industria Metalurgica 1062, Ramos Arizpe 25900, Mexico
[b] KU Leuven, ESAT-PSI-VISICS, iMinds, Kasteelpark Arenberg 10, Leuven B-3001, Belgium

ABSTRACT

In this paper we present a novel template-based approach for fast object detection. In particular we investigate the use of Dominant Orientation Templates (DOT), a binary template representation introduced by Hinterstoisser et al., as a means for fast detection of objects even if textureless. During training, we learn a binary mask for each template that allows to remove background clutter while at the same time including relevant context information. These mask templates then serve as weak classifiers in an Adaboost framework.

We demonstrate our method on detection of shape-oriented object classes as well as multiview vehicle detection. We obtain a fast yet highly accurate method for category level detection that compares favorably to other more complicated yet much slower approaches. We further show how to efficiently transfer meta-data using the top most similar activated templates.

Finally, we propose an optimization scheme for detection of specific objects using our proposed masks trained by the SVM, resulting in an increment of up to 17% in performance of the DOT method, without sacrificing testing speed and it is able to run the training on real time.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

To date, the main application of template matching in image processing has been the detection in an image of an *a priori* known object instance from an *a priori* known viewpoint. To incorporate viewpoint variations (or non-rigid deformation), usually multiple templates are stored in memory and matched to the image one by one. The maximum response is selected and if this is greater than a threshold value, then it is counted as a detection. Fig. 1 illustrates this procedure.

Thanks to its speed, template matching has traditionally been a popular method in manufacturing environments, where occlusions and clutter can mostly be avoided and lighting and pose variations can be carefully controlled, although it has also been applied in more challenging settings such as, e.g., pedestrian detection [1]. Another advantage of template matching is that it is also possible to transfer offline annotated/created meta data to the object, once it is detected.

In other contexts, objects are mostly recognized based on local features (e.g. SIFT [2]). This usually works well with changing lighting conditions, partial occlusion and, to some extent, changes in the viewpoint. However, this kind of approach usually fails on texture-less objects such as mugs or bottles that are mostly determined by their projected contours or shapes. Another method, that seems to work well in both situations, is the histogram of oriented gradients (HOG) representation [3], combined with an SVM. However, its computation time is high. Here we investigate a simpler and more efficient alternative.

Hinterstoisser et al. proposed a template representation, coined *Dominant Orientation Templates* or DOT [4]. DOT is inspired by the Histograms of Oriented Gradients [3] descriptor, but designed to be fast. They show impressive results, recognizing 3D non-textured objects over cluttered background in realtime, using several templates per object. At the core of their method is a binary representation of the image and template, based on dominant gradient orientations. Evaluating the template for a given location in a new image is highly accurate and requires only simple bit-wise operations, making it very fast. By jittering the training image when building the template, they ensure invariance to small translations. This not only increases the robustness, but also allows skipping many locations while parsing an image. However, there are also some shortcomings. One of them is that the method cannot model negative data during training, which during detection, results in a relatively high number of false positives. Another limitation is that it cannot handle objects at the category level,
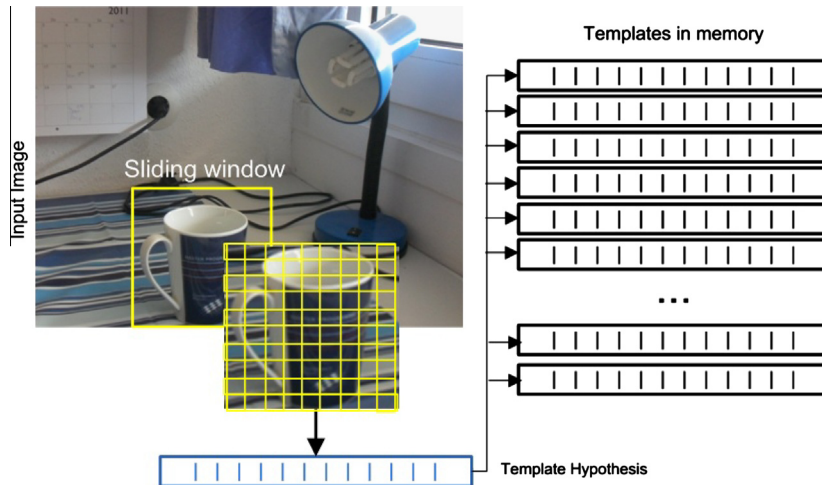
Fig. 1. Template matching. A sliding window is run over the image obtaining a template representation for each position of the window. Then it is evaluated against each of the templates stored in memory.

but only *a priori* known instances of the *same* object. Another more robust method was introduced recently by the same authors, called LINE2D [5]. It improves over DOT, but suffers from the same limitations.

In this paper, we investigate whether binary Dominant Orientation Templates (DOT) can also be used successfully for learning to detect instances of an object *class*, without sacrificing efficiency. This requires a method that can generalize beyond the examples seen during training, so as to also recognize previously unseen instances of the object class. Given the within-class variability, a single template per viewpoint is no longer sufficient, and the greedy selection of templates as in [4] is suboptimal (see Section 5.2). Increasing the amount of jittering would allow for more variability, but also reduces the discriminative power leading to more false positives. Instead, starting from multiple training images possibly covering different object poses, we need a method for template selection, as well as a scheme for combining the outputs of the different templates, that effectively allows to generalize beyond the individual examples (templates). To this end, we propose to learn masks for the DOT templates and to use these as weak classifiers in an Adaboost framework [6]. We also show, that these masks can be applied to template based methods for specific object detection, showing up to 17% increment in performance.

1.1. Main contribution

Our main contribution is the construction of an efficient, category-level object detector based on a computationally efficient, yet discriminative template representation. We propose to learn a binary mask for each template, based on feature selection using a linear support vector machine. We also show how the masks can be used for meta data transfer.

By learning the masks we refine the bounding box annotation to a more representative template, including relevant context cues (e.g. the shadow beneath a car) while at the same time removing background clutter (e.g. the buildings behind the car) or areas on the object itself that do not generalize well (e.g. the prints on mugs). This turns out to be a crucial step for obtaining a highly accurate system. We show that, by learning appropriate binary masks and using the masked templates as weak classifiers in a boosting framework, the output of several templates can be combined into a powerful detector. Moreover, it is inherently fast thanks to the use of bitwise operations only.

Additionally, we propose a method based on the masks, to optimize template tables for specific object detection, incrementing

the performance without sacrificing detection speed, at the cost of a very small training time (ms). The same scheme can be applied to other binary based template methods such as LINE2D or LINE-MOD [5] to improve their performance.

1.2. Efficient template based object detection

While the Adaboost framework is a standard classifier in vision applications, it has mostly been applied in combination with localized weak classifiers (e.g. [7,8]).

Here we show it can also be applied successfully in combination with global features (templates). This results in a strong classifier that combines multiple DOT templates at classification time. Note how this is significantly different from applying several templates one by one as in [4] or using a mixture model with latent SVM [9]: We do not pick (implicitly or explicitly) one template for a given test image, but instead always evaluate multiple templates, combining their scores with the weights determined by Adaboost. We believe that this is essential to make the scheme generalize well to previously unseen object instances.

Initially, the mask learning scheme we propose seemed to be limited to the case of well aligned training data (e.g. a single viewpoint). By unsupervised clustering of the training data, we show it can also deal with multiview settings. Here it is important to stress that this clustering is only used for learning the masks. Unlike e.g. [10] we do not learn separate viewpoint specific classifiers, but instead run a single detector that has learnt to deal with the viewpoint changes.

The remainder of the paper is structured as follows: We first discuss related work in Section 2. Then we briefly explain the dominant orientation templates [4] in Section 3. In Section 4, we show how to adapt DOT to work as features (weak classifiers) and present our method for object class learning, including the Adaboost template selection and weighting, as well as the learning of the masks. Section 5 describes our experimental results, Section 6 presents discussion and future work and Section 7 concludes the paper.

2. Related work

Here we focus on the most relevant work on shape oriented methods, efficient multiview, template-based as well as boosting-based methods. Examples of *template-based* category-level object detection methods are rare. Probably the best known example is the method proposed by Gavrilla and Philomin [1], which was