Pattern Recognition 48 (2015) 1844–1853

Contents lists available at ScienceDirect

Pattern Recognition

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

A coarse-to-fine approach for fast deformable object detection

Marco Pedersoli [a,*,1], Andrea Vedaldi [b], Jordi Gonzàlez [a], Xavier Roca [a]

[a] Computer Vision Center and Universitat Autònoma de Barcelona, Edifici O, Campus UAB, 08193 Bellatera, Spain
[b] Department of Engineering Science, Oxford University, Oxford OX1 3PJ, UK

ARTICLE INFO

Article history:
Received 13 November 2013
Received in revised form
17 October 2014
Accepted 8 November 2014
Available online 15 November 2014

Keywords:
Object recognition
Object detection

ABSTRACT

We present a method that can dramatically accelerate object detection with part based models. The method is based on the observation that the cost of detection is likely dominated by the cost of matching each part to the image, and not by the cost of computing the optimal configuration of the parts as commonly assumed. To minimize the number of part-to-image comparisons we propose a multiple-resolutions hierarchical part-based model and a corresponding coarse-to-fine inference procedure that recursively eliminates from the search space unpromising part placements. The method yields a ten-fold speedup over the standard dynamic programming approach and, combined with the cascade-of-parts approach, a hundred-fold speedup in some cases. We evaluate our method extensively on the PASCAL VOC and INRIA datasets, demonstrating a very high increase in the detection speed with little degradation of the accuracy.

1. Introduction

In the last few years the interest of the object recognition community has moved from image classification and orderless models such as bag-of-words [1] to sophisticated representations that can explicitly account for the location, scale, and deformation of the objects [2–5]. By reasoning about geometry instead of discarding it, these models can extract a more detailed description of the image, including the object location, pose, and deformation, and can result in better detection accuracy.

A major obstacle in dealing with deformable objects is the combinatorial complexity of the inference. For instance, in the pictorial structures pioneered by Fischler and Elschlager [6] an object is represented as a collection of P parts, connected by springs. The time required to find the optimal part configuration to match a given image can be as high as the number L of possible part placements to the power of the number P of parts, *i.e.* $O(L^P)$. This cost can be reduced to $O(PL^2)$ or even $O(PL)$ by imposing further restrictions on the model ([2], Sections 2, 3.1), but is still significant due to the large number of possible part placements L. For instance, just to test for all possible translations of a part, L can be as large as the number of image pixels. This analysis, however, does not account for several aspects of typical part based models, such as the fact that useful object deformations are not very large and that, with appearance descriptors such as histograms of oriented gradients (HOG) [7], locations can be sampled in a relatively coarse manner.

The first contribution of this paper, an extension of our prior work [8,9], is a new analysis of the cost of part based models (Section 3.1) which better captures the bottlenecks of state-of-the-art implementations such as [7,3,10]. In particular, we show that the cost of inference is likely to be dominated by the cost of *matching each part to the image* rather than by the cost of determining the optimal part configuration. This suggests that accelerating inference requires minimizing the number of times the parts are matched.

Reducing the number of part evaluations can be obtained by using a *cascade* [11], a method that rejects quickly unpromising object hypotheses based on cheaper models. For deformable part models two different types of cascades have been proposed (Sections 2, 3.1). The first one, due to Felzenszwalb et al. [12], matches parts sequentially, comparing the partial scores to learned thresholds in order to reject object locations as soon as possible. The second one, due to Sapp et al. [13], filters the part locations by thresholding marginal part scores obtained from a lower resolution model.

The second contribution of the paper is a different cascade design (Section 3.2). Similar to [11,13], our method is coarse-to-fine. However, we note that, by thresholding scores independently, standard cascades propagate to the next level clusters of nearly identical hypotheses (as these tend to have similarly high scores). Instead of thresholding, we propose to reject all but the hypothesis whose score is *locally maximal*. This is motivated by the fact that looking for a locally optimal hypothesis at a coarse resolution often predicts well the best hypothesis at the next resolution level

* Corresponding author. Tel.: +32163 21095; fax: +32163 21723.
[1] Present address: KU Leuven, ESAT-PSI-VISICS/iMinds Kasteelpark Arenberg 10, 3001, Leuven, Belgium.
E-mail address: marco.pedersoli@esat.kuleuven.be (M. Pedersoli).

http://dx.doi.org/10.1016/j.patcog.2014.11.006
0031-3203/

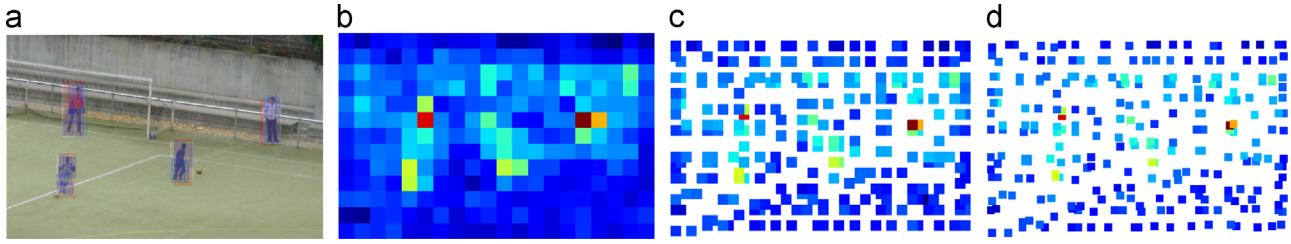

Fig. 1. Coarse-to-fine inference. We propose a method for the fast inference of multi-resolution part based models. (a) example detections; (b) scores obtained by matching the lowest resolution part (root filter) at all image locations; (c) scores obtained by matching the intermediate resolution parts, only at location selected based on the response of the root part; (d) scores obtained by matching the high resolution parts, only at locations selected based on the intermediate resolution scores. A white space indicates that the part is not matched at a certain image location, resulting in a computational saving. The saving *increases with the resolution*.

(Section 3.2). As suggested in Fig. 1, and as showed in Sections 3.2–3.4, this results in an *exponential saving, which has the additional benefit of being independent of the image content.* Experimentally, we show that this procedure can be more than ten times faster than the distance transform approach of [2,3], while still yielding excellent detection accuracy.

Compared to using global thresholds as in the cascade of parts approach of Felzenszwalb et al. [12], our method does not require fine tuning of the thresholds on a validation set. Thus it is possible to use it not just for *testing*, but also for *training* the object model, when the thresholds of the cascade are still undefined (Section 3.5). More importantly, the cascade of parts and our method are based on complementary ideas and can be combined, yielding a *multiplication the speed-up factors.* The combination of the two approaches can be more than two order of magnitude faster than the baseline dynamic programming inference algorithm [2] (Section 4).

2. Related work

In object category detection the goal is to identify and localize in images natural objects such as people, cars, and bicycles. Formally, we regard this as the problem of mapping an image $\mathbf{x}$ to a label or *interpretation* $\mathbf{y}$ that specifies whether an instance of the object is contained in the image and, if so, a bounding box enclosing it.

In order to simplify analysis as well as learning, the map $\mathbf{x} \mapsto \mathbf{y}$ is usually represented indirectly by a *scoring function* $S(\mathbf{x}, \mathbf{y})$, expressing how well an interpretation $\mathbf{y}$ describes an image $\mathbf{x}$. The advantage is that the scoring function can have a simple form, often linear in a vector of parameters $\mathbf{w}$, *i.e.* $S(\mathbf{x}, \mathbf{y}) = \langle \mathbf{w}, \Psi(\mathbf{x}, \mathbf{y}) \rangle$. *Inferring* the interpretation $\mathbf{y}$ from the image $\mathbf{x}$ reduces then to finding which interpretations have a sufficiently large score, typically by computing the maximizer $\mathbf{y} = \arg\max_{\hat{\mathbf{y}} \in \mathcal{Y}} S(\mathbf{x}, \hat{\mathbf{y}})$. Unfortunately, maximizing the scoring function is often computationally quite challenging. Next, we briefly review the main ideas that have been explored to address this issue.

Exhaustive and greedy search. If the interpretation space is sufficiently small, an inference algorithm can *score exhaustively* all interpretations $\mathbf{y} \in \mathcal{Y}$ and pick the best one. Sometimes this strategy can be applied even to continuous interpretations spaces up to discretization. A notable example are *sliding-window detectors* such as Dalal and Triggs [7]. A candidate interpretation $\mathbf{y}$ obtained from a discretized model can be further improved by a sequence of local greedy modifications, similar to gradient ascent. Unfortunately local search can easily get stuck in local optima. In less trivial cases, such as deformable part models, the interpretation space $\mathcal{Y}$ is far too complex for such simple strategies to suffice.

Sampling. By interpreting the score $S(\mathbf{x}, \mathbf{y})$ as a posterior probability $p(\mathbf{y}|\mathbf{x})$ on the interpretations, inference can be reduced to the problem of drawing samples $\mathbf{y}$ from $p(\mathbf{y}|\mathbf{x})$ (because the most likely interpretations are also the ones with larger scores). Sampling ideas have been explored in the context of sliding-window object detectors in [14] demonstrating a two fold speed-ups over exhaustive search. Similar in spirit, but based on prior knowledge about the general shape of an object, are selective search [15] and objectness [16]. The main speed-up of these methods is again due to a reduced set of samples. However, in this case the samples are category independent (.i.e the same bounding boxes are used to represent different categories) so that the feature encoding can be computed only once for all categories.

Branch-and-bound. It is sometimes possible to compute efficiently upper bounds on the scores of large subsets $\mathcal{Y}' \subset \mathcal{Y}$ of interpretations at once. If a better interpretation is found somewhere else, then the whole subset $\mathcal{Y}'$ can then be removed without further consideration. *Branch-and-bound* methods apply this idea to a recursive partition of the interpretation space $\mathcal{Y}$. If the splits are balanced and the bounds sufficiently tight, these strategies can find the optimal interpretation very quickly. This idea has been popularized in the recent literature on sliding-window object detectors by Lampert and Blaschko [17].

Dynamic programming (*DP*). Sometimes interpretations are obtained by combining smaller interpretations of portions of the image. For example, in pictorial structures [6] an object is an arrangement $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ of N object parts (*e.g.*, the head, torso, arms, and legs of a person), where $\mathbf{y}_i$ is the location of the corresponding part in the image. While there is a combinatorial number of such arrangements, in constellation models [18], the score decomposes as $S(\mathbf{y}_0, \mathbf{y}_1) + S(\mathbf{y}_0, \mathbf{y}_2) + \cdots + S(\mathbf{y}_0, \mathbf{y}_N)$, where $\mathbf{y}_0$ is a reference part connected in a star to the other parts. Hence the optimal arrangement can be obtained by finding the optimal position of each part $\arg\max_{\mathbf{y}_i} S(\mathbf{y}_0, \mathbf{y}_i)$ relative to the reference part $\mathbf{y}_0$, and then optimizing over the location of the reference. Efficient inference extends to more complex topologies such as trees and can be further improved under certain assumptions on the scores, yielding to the efficient pictorial structures of [2] (Section 3.1).

Cascades. A *cascade* considers cheaper scoring functions along with $S(\mathbf{x}, \mathbf{y})$ and uses them to prune quickly unpromising interpretations $\mathbf{y}$ from consideration. Applied to an exhaustive search of the possible object locations, this yields the well-known cascade approach to sliding-window object detection [19]. The idea has been popularized by its application to AdaBoost [20–23] and has remained popular through the years, including applications to multiple kernels detectors [24]. The same idea has been applied directly to part-based models to either prune object locations by visiting only a small number of parts [12] or by finding plausible placements of the parts based on scoring functions with a lower degree of part dependencies [25] or lower resolution parameters [13]. Section 3.2 introduces an alternative coarse-to-fine cascade design. A more general analysis of other problems related with fast detection can be found in [26].

Recent methods. In parallel with the submission of this work and during the revision period several new methods for speeding