Image and Vision Computing 30 (2012) 923–933

Contents lists available at SciVerse ScienceDirect

Image and Vision Computing

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

Viewpoint-aware object detection and continuous pose estimation ☆

Daniel Glasner [a,*], Meirav Galun [a], Sharon Alpert [a], Ronen Basri [a], Gregory Shakhnarovich [b]

[a] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Israel
[b] Toyota Technological Institute at Chicago, United States

ARTICLE INFO

Article history:
Received 31 May 2012
Received in revised form 7 September 2012
Accepted 30 September 2012

Keywords:
Viewpoint-aware
Object detection
Pose estimation
3D model
Viewpoint estimation
Structure from motion

ABSTRACT

We describe an approach to category-level detection and viewpoint estimation for rigid 3D objects from single 2D images. In contrast to many existing methods, we directly integrate 3D reasoning with an appearance-based voting architecture. Our method relies on a nonparametric representation of a joint distribution of shape and appearance of the object class. Our voting method employs a novel parameterization of joint detection and viewpoint hypothesis space, allowing efficient accumulation of evidence. We combine this with a re-scoring and refinement mechanism, using an ensemble of view-specific support vector machines. We evaluate the performance of our approach in detection and pose estimation of cars on a number of benchmark datasets. Finally we introduce the "Weizmann Cars ViewPoint" (WCVP) dataset, a benchmark for evaluating continuous pose estimation.

1. Introduction

The problem of category-level object detection has been at the forefront of computer vision research in recent years. One of the main difficulties in the detection task stems from variability in the objects' appearance due to viewpoint variation (or equivalently pose variation). While most existing methods treat the detection task as that of 2D pattern recognition, there is an increasing interest in methods that explicitly account for view variation and that combine detection with pose estimation. This paper presents an approach that integrates detection and pose estimation using 3D class models of rigid objects and demonstrates this approach on the problem of car detection.

Building a viewpoint-aware detector presents a number of challenges. The first question is how to acquire a 3D representation of a class. The recent availability of CAD models, 3D cameras, and robust structure-from-motion (SfM) software has simplified this problem. Using SfM methods or 3D cameras makes it straightforward to relate the available 3D representations to the appearance of objects in training images. Secondly, finding the pose of an object at test time requires search in the 6D space of possible Euclidean transformations. This can be accomplished by searching exhaustively through a discrete binning of this 6D space. An alternative is to use a combinatorial search e.g., RANSAC [1] procedure. Both options, however, can be computationally expensive. Finally, how should detection and pose estimation be integrated? Pose estimation can deteriorate significantly when detection is inaccurate. Can detection be improved if pose estimation is integrated into the process?

We suggest an approach that combines a nonparametric voting procedure with discriminative re-scoring for detection and pose estimation of rigid objects.

We address the challenge of building a 3D class model by applying state-of-the-art SfM reconstruction software [2,3] to a set of training images that we have collected. A 3D point cloud is reconstructed for each scene and the car of interest is manually segmented. The class model is composed by registering and merging these point clouds in 3D. Further details regarding the construction of the model can be found in Section 5. A notable advantage of using SfM is that it provides us with correspondences between an accurate 3D shape model (see Fig. 5) and 2D appearance in real images. We model the within class variability by using multiple exemplars in a non-parametric model. By performing a simple registration of the point clouds in 3D we are able to circumvent the difficult problem of finding correspondences between parts across different class instances. The registered point clouds can be seen in Fig. 4(d).

At test time we wish to infer the location and pose of a class instance or equivalently the most likely transformation which relates the 3D model to the 2D projection in the test image. We use a nonparametric voting procedure, by searching the database for patches similar to ones seen in the input image. Each match generates a vote, in the spirit of [4] and other methods inspired by Hough transform. However, here the vote implies not only a bounding box, as is done, e.g., in [4], but a full 6 DOF weak perspective transformation of the object. The 6 DOF voting mechanism relates detection and pose estimation starting at the very early stages of the visual processing. In addition to providing

☆ This paper has been recommended for acceptance by Thomas Brox.

* Corresponding author. Tel.: +972 89344443.
E-mail addresses: daniel.glasner@weizmann.ac.il (D. Glasner), meirav.galun@weizmann.ac.il (M. Galun), sharon.alpert@weizmann.ac.il (S. Alpert), ronen.basri@weizmann.ac.il (R. Basri), gregory@ttic.edu (G. Shakhnarovich).

0262-8856/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.imavis.2012.09.006

a natural framework to estimate viewpoint, this serves to improve the localization performance of the detector, by constraining the detections to be consistent with a specific viewpoint estimate. Specifically, if the detector proposes a hypothesized detection of a car seen from a particular viewpoint, we should expect that all the features that support it are consistent with that viewpoint. Intuitively, this is a built-in verification mechanism.

The voting mechanism outlined above, and described in detail in Section 3, serves as the first stage of detection. It can be thought of as an attention mechanism, generating a relatively small set of hypothesized object detections, each with a specific hypothesized 3D pose. In line with the previous work [5] we follow the voting phase by a discriminative stage. The second stage, refines the hypothesized detections, and ranks them by applying a set of support vector machines, each trained to score detections in a sector of viewpoints. At test time we use the pose estimation generated by the voting procedure to index the correct SVM. Thus, this final stage of verification is also viewpoint-aware. Overall, this process allows us to improve and refine our candidate detections.

We focus our experiments on cars and apply our algorithm to four datasets: Pascal 2007, Stanford 3Dpose dataset [6], EPFL car data set [7] and a new benchmark introduced here the "Weizmann Cars ViewPoint" dataset.

2. Background

A common approach for coping with viewpoint variability is to use multiple, independent 2D models. In this multiview approach one describes the appearance of an object class at a discrete set of representative viewpoints. These algorithms (e.g., [8,9,7,10]) implicitly assume that the 2D appearance of an object near the representative views varies smoothly and that local descriptors are robust enough to handle these appearance variations.

Extensions to the popular Deformable Part Model (DPM) [9] which allow for viewpoint estimation have been suggested in [11] and in the recent work of Pepik et al. [12]. Both works consider a model in which the mixture components are identified with a discrete set of viewpoints. [11] further suggest continuous viewpoint estimation by learning a linear model around each of the discrete representative views. [12] extend the DPM by modeling continuous part locations in 3D. Their method produces state-of-the-art detection and pose estimation results on the Stanford 3Dpose dataset in which the task is to classify test images into one of a discrete set of viewpoints. They do not report results on a continuous viewpoint estimation task.

Another line of studies [6,13,14] approaches the problem of view variation by building 2D multi-part representations and establishing correspondences among the parts across different class views. The resulting model accounts for a dense, multiview representation and is capable of recognizing unseen views.

Many of the algorithms which explicitly model 3D shape utilize 3D CAD models [15–18,12,19]. These works take different approaches to generate correspondences between the 3D CAD models and the 2D appearances which can be matched to the test images.

The approach in [15,16,18] is to generate non-photorealistic renderings of the CAD models from which they extract features (e.g., edges) or learn 2D detectors which can then be used to find matches in real images. The rendering approach maintains an exact correspondence between 3D shape and 2D appearance, but the resulting 2D appearance models are not as powerful as those learned from real 2D images.

The authors of [17,12,19] also take advantage of real images as part of their training data. The approach described in [17] is to learn 3D shape models and 2D appearance models separately. The 3D and 2D models are then linked by a rough correspondence which is established between bounding boxes extracted from real images and from ones which are rendered using the 3D CAD. The bounding boxes are partitioned into blocks centered on a regular grid and correspondence is determined by the grid coordinates.

In their 3D parts model the authors of [12] suggest a Deformable Part Model with 3D volumetric parts. These are parameterized as axis aligned fixed size 3D bounding cubes. The 2D appearance templates are learned from non-photorealistic renderings of CAD models. Their framework also allows for the inclusion of real training images as part of training.

The work described in [19] suggests modeling 3D objects using "aspect parts". These are defined as a portion of the object whose entire surface is either visible or occluded from any viewpoint. Correspondence between 3D configurations of aspect parts and their appearance in the image is modeled as a conditional random field. Rather than using multiple appearance templates the authors suggest applying homographies to rectify parts to a canonical viewpoint. To generate correspondences between the 3D aspect-parts and their appearances in 2D images the authors rely on extensive hand labeled part annotations of the training data.

In other work, Arie-Nachimson and Basri [20] construct a 3D model by employing an SfM process on the entire training set of class images. An advantage of this approach is that the SfM provides correspondences between 3D shape and 2D appearance. However, their method requires finding correspondences between parts as they appear in different class instances.

Villamizar et al. [21] present a two-step approach. In the first step a Hough-based pose estimator identifies candidate detections along with their pose estimates. In the second step these are validated using pose-specific classifiers. In both steps their method uses a common set of Random Fern features which are also shared between the classifiers.

Sun et al. [22] suggest the use of depth information, and train models using depth maps acquired with a range camera. Detections are generated by depth-encoded voting. Pose estimation is then achieved by registering the inferred point cloud and a 3D CAD model.

Payet and Todorovic [23] learned a collection of 2D shape templates describing the appearance of the object contours at a discrete set of viewpoints. At test time, the learned templates are matched to the image contours while allowing for an arbitrary affine projection. The matching problem is approximated using a convex relaxation. The parameters of the chosen affine projection serve as a continuous estimate of the pose in 3D while the best matching template identifies a discrete estimate.

Finally, a hybrid 2D–3D model is suggested in [24]. The model consists of stick-like 2D and 3D primitives. The learning selects 3D primitives to describe viewpoint varying parts and 2D primitives where the appearance is viewpoint invariant.

In contrast to related work, we propose a simple and flexible method to construct rich, nonparametric 3D class models. State-of-the-art SfM software allows us to model 3D shape without requiring a library of CAD models. It also provides us with accurate correspondence between 3D shape and real-world 2D appearances. Our method does not need costly manual part annotation, in-fact it manages to bypass the difficult problem of finding correspondences between parts in 2D, by solving an easy global registration problem in 3D. Finally, unlike other works which are restricted to discrete pose classification, our method optimizes over a continuous 6D transformation space and is able to generate accurate continuous pose-estimates.

3. Nonparametric detection and pose estimation

We approach the problem of object detection and pose estimation in two stages. First we apply nonparametric voting to produce a bank of candidate detections along with their estimated poses. Then we apply a discriminative re-scoring procedure designed to improve the detection and pose estimation results. In this section we describe the