The challenge of simultaneous object detection and pose estimation: a comparative study

Daniel Oñoro-Rubio, Roberto J. López-Sastre, Carolina Redondo-Cabrera, Pedro Gil-Jiménez

PII: S0262-8856(18)30158-6
DOI: doi:10.1016/j.imavis.2018.09.013
Reference: IMAVIS 3722

To appear in: *Image and Vision Computing*

Received date: 12 January 2018
Accepted date: 14 September 2018

The challenge of simultaneous object detection and pose estimation: a comparative study

Daniel Oñoro-Rubio, Roberto J. López-Sastre, Carolina Redondo-Cabrera and Pedro Gil-Jiménez

GRAM, University of Alcalá, Alcalá de Henares, 28805, Spain

Abstract

Detecting objects and estimating their pose remains as one of the major challenges of the computer vision research community. There exists a compromise between localizing the objects and estimating their viewpoints. The detector ideally needs to be view-invariant, while the pose estimation process should be able to generalize towards the category-level. This work is an exploration of using deep learning models for solving both problems simultaneously. For doing so, we propose three novel deep learning architectures, which are able to perform a joint detection and pose estimation, where we gradually decouple the two tasks. We also investigate whether the pose estimation problem should be solved as a classification or regression problem, being this still an open question in the computer vision community. We detail a comparative analysis of all our solutions and the methods that currently define the state of the art for this problem. We use PASCAL3D+ and ObjectNet3D datasets to present the thorough experimental evaluation and main results. With the proposed models we achieve the state-of-the-art performance in both datasets.

Keywords: Pose estimation, viewpoint estimation, object detection, deep learning, convolutional neural network

1. Introduction

Over the last decades, the category-level object detection problem has drawn considerable attention. As a result, much progress has been realized, leaded mainly by international challenges and benchmarking datasets, such as the PASCAL VOC Challenges [1] or the ImageNet dataset [2]. Nevertheless, researchers soon identified the importance of not only localizing the objects, but also estimating their poses or viewpoints, *e.g.* [3, 4, 5, 6]. This new capability results fundamental to enable a true interaction with the world and its objects. For instance, a robot which merely knows the location of a cup but that cannot find its handle, will not be able to grasp it. In the end, the robotic solution needs to know a viewpoint estimation of the object to facilitate the inference of the visual affordance for the object. Also, in the augmented reality field, to localize and estimate the viewpoint of the objects, is a crucial feature in order to project a realistic hologram, for instance.

Technically, given an image, these models can localize the objects, predicting their associated bounding boxes, and are also able to estimate the relative pose of the object instances in the scene with respect to the camera. Figure 1 shows an example, where the viewpoint of the object is encoded using just the azimuth angle. In the image, the target objects are the sofa and the bicycle. Their locations are depicted by their bounding boxes (in green), and their azimuth angles are represented by the blue arrow inside the yellow circle.

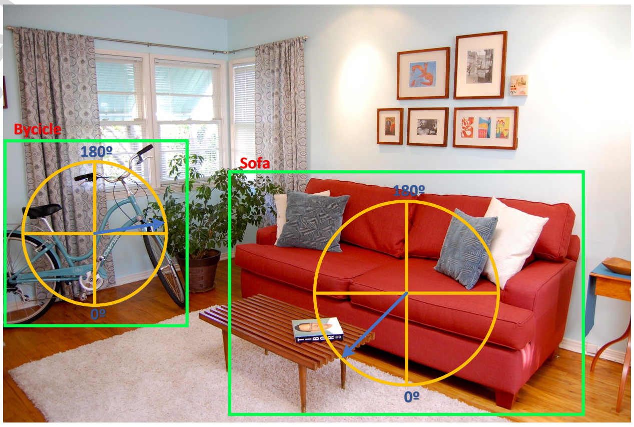

Figure 1: Object category detection and pose estimation example. In the image, the sofa and the bicycle are localized by the green bounding boxes. The blue arrow inside the yellow circles shows the azimuth angles of the objects, which is a form of viewpoint annotation.

The computer vision community rapidly detected the necessity of providing the appropriate annotated datasets, in order to experimentally validate the object detection and pose estimations approaches. To date, several datasets have been released. Some examples are: 3D Object categories [4], EPFL Multi-view car [7], ICARO [8], PASCAL3D+ [9] or ObjectNet3D [10].

Thanks to these datasets, multiple models have been experimentally evaluated. It is particularly interesting to observe how all the published approaches can be classified in two groups. In the first one, we find those models that decouple both problems (*e.g.* [11, 12, 13]), making first a location of the object, to later estimate its pose. In the second group we identify the approaches that solve both tasks simultaneously (*e.g.* [9, 14, 15]), because they understand that to carry out a correct location re-

Email address: daniel.onoro@edu.uah.es, robertoj.lopez@uah.es, carolina.redondoc@edu.uah.es and pedro.gil@uah.es (Daniel Oñoro-Rubio, Roberto J. López-Sastre, Carolina Redondo-Cabrera and Pedro Gil-Jiménez)