Int J Comput Vis (2012) 98:243–262
DOI 10.1007/s11263-011-0507-2

Object Recognition by Sequential Figure-Ground Ranking

João Carreira · Fuxin Li · Cristian Sminchisescu

Received: 19 February 2011 / Accepted: 8 November 2011 / Published online: 19 November 2011

Abstract We present an approach to visual object-class segmentation and recognition based on a pipeline that combines multiple figure-ground hypotheses with large object spatial support, generated by bottom-up computational processes that do not exploit knowledge of specific categories, and sequential categorization based on continuous estimates of the spatial overlap between the image segment hypotheses and each putative class. We differ from existing approaches not only in our seemingly unreasonable assumption that good *object-level segments* can be obtained in a feed-forward fashion, but also in formulating recognition as a regression problem. Instead of focusing on a one-vs.-all winning margin that may not preserve the ordering of segment qualities inside the non-maximum (non-winning) set, our learning method produces a *globally consistent* ranking with close ties to segment quality, hence to the extent entire object or part hypotheses are likely to spatially overlap the ground truth. We demonstrate results beyond the current state of the art for image classification, object detection and semantic segmentation, in a number of challenging datasets including Caltech-101, ETHZ-Shape as well as PASCAL VOC 2009 and 2010.

Keywords Object recognition · Semantic segmentation · Learning and ranking

The first two authors contributed equally.

J. Carreira · F. Li · C. Sminchisescu (✉)
University of Bonn, INS, Wegelerstrasse 6, Bonn 53115, Germany
e-mail: cristian.sminchisescu@ins.uni-bonn.de

1 Introduction

Recognizing and localizing different categories of objects in images is essential for scene understanding. Approaches to object-category recognition based on sliding windows have recently been demonstrated convincingly in difficult benchmarks (Viola and Jones 2001; Felzenszwalb et al. 2010; Vedaldi et al. 2009). By scanning the image at multiple locations and scales, recognition is phrased as a binary decision problem for which many powerful classifiers exist. Recent developments have shown that scanning hundreds of thousands of windows efficiently can be feasible for certain types of features and classifiers (Vedaldi et al. 2009; Blaschko and Lampert 2008). The bounding box approach to recognition has proven successful for object categories with stable features that can 'fill' the correct window significantly, like faces or motorbikes, it nevertheless tends to be unsatisfactory for objects with more complex appearance and geometry, or for advanced tasks such as pose prediction and action recognition where the knowledge of an object's shape is also important.

This motivates the focus on *semantic segmentation*, where the objective is to both identify the spatial support of objects, and to recognize their category. In semantic segmentation, the brute-force sliding windows approach to generic category recognition may not be feasible. Consider Fig. 1(a). A reliable object detector might locate the person and place a bounding box around her. However, the non-canonical pose may impose a large bounding box, or alternatively a large search space if different rotations of the bounding box are scanned, still leaving a non-trivial contour hypothesis space to be explored, even inside the correct bounding box, *e.g.* Fig. 1(b).

The semantic segmentation problem could be approached top-down (Borenstein and Ullman 2002; Leibe et al. 2008),