Int J Comput Vis
DOI 10.1007/s11263-013-0663-7

Probabilistic Joint Image Segmentation and Labeling by Figure-Ground Composition

Adrian Ion · João Carreira · Cristian Sminchisescu

Received: 14 March 2013 / Accepted: 30 September 2013

Abstract We propose a layered statistical model for image segmentation and labeling obtained by combining independently extracted, possibly overlapping sets of figure-ground (FG) segmentations. The process of constructing consistent image segmentations, called *tilings*, is cast as optimization over sets of maximal cliques sampled from a graph connecting all non-overlapping figure-ground segment hypotheses. Potential functions over cliques combine unary, Gestalt-based figure qualities, and pairwise compatibilities among spatially neighboring segments, constrained by T-junctions and the boundary interface statistics of real scenes. Building on the segmentation layer, we further derive a joint image segmentation and labeling model (JSL) which, given a bag of FGs, constructs a joint probability distribution over *both* the compatible image interpretations (*tilings*) composed from those segments, *and* over their *labeling* into categories. The process of drawing samples from the joint distribution can be interpreted as first sampling tilings, followed by sampling labelings conditioned on the choice of a particular tiling. We learn the segmentation and labeling parameters jointly, based on maximum likelihood with a novel estimation procedure we refer to as incremental saddle-point approximation. The partition function over tilings and labelings is increasingly more accurately approximated by including incorrect configurations that are rated as probable by candidate models during learning. State of the art results are reported on the Berkeley, Stanford and Pascal VOC datasets, where an improvement of 28 % was achieved for the segmentation task only (tiling), and an accuracy of 47.8 % was obtained on the test set of VOC12 for semantic labeling (JSL).

Keywords Image segmentation · Image labeling · Semantic segmentation · Statistical models · Learning and categorization

A. Ion
Faculty of Informatics, Vienna University of Technology, Vienna, Austria
e-mail: ion@prip.tuwien.ac.at

J. Carreira
Institute of Systems and Robotics, University of Coimbra, Coimbra, Portugal
e-mail: joaoluis@isr.uc.pt

C. Sminchisescu (✉)
Department of Mathematics, Faculty of Engineering, Lund University, Lund, Sweden
e-mail: cristian.sminchisescu@math.lth.se

C. Sminchisescu
Institute of Mathematics of the Romanian Academy, Bucharest, Romania

1 Introduction

One of the main goals of scene understanding is the semantic segmentation of images: label a diverse set of object properties, at multiple scales, while at the same time identifying the spatial extent over which such properties hold, a process also called image labeling. For instance, an image may be segmented into things (man-made objects, people or animals) or their parts, into amorphous regions or stuff like grass or sky, or main geometric properties like the ground plane or the vertical planes corresponding to buildings in the scene. It appears to be now well understood that a successful identification of such properties requires models that can make inferences over adaptive spatial neighborhoods that span well beyond small homogeneous regions around individual pixels (superpixels) and that may overlap. Indeed, if image regions could be extracted so they would at least partly overlap the projections of visible surfaces in the scene, it would be conceivable

Published online: 17 November 2013