Int J Comput Vis (2015) 112:131–132
DOI 10.1007/s11263-015-0807-z

Guest Editorial: Scene Understanding

Derek Hoiem · James Hays · Jianxiong Xiao · Aditya Khosla

Published online: 10 March 2015

Scene understanding is the ability to visually analyze a scene to answer questions such as: What is happening? Why is it happening? What will happen next? What should I do? For example, in the context of driving safety, the vision system would need to recognize nearby people and vehicles, anticipate their motions, infer traffic patterns, and detect road conditions. So far, research has focused on providing complete (e.g., every pixel labeled) or holistic (reasoning about several different scene elements) interpretations, often taking into account scene geometry or 3D spatial relationships.

Accordingly, in this issue, several papers offer improvements to image segmentation and labeling through use of region classifiers, detectors, and object and scene context:

"Indoor Scene Understanding with RGB-D Images: Bottom-up Segmentation, Object Detection and Semantic Segmentation" (doi:10.1007/s11263-014-0777-6) by Gupta et al. addresses problems of interpreting indoor scenes from a paired RGB and depth image. The method infers whether observed contours are due to depth, normal, or albedo changes and uses the contours to produce a hierarchical scene segmentation, in which non-adjacent regions can also be grouped to account for occlusion. Several geometry-based region features are proposed to classify regions into category labels, and object detectors using histograms of depth gradients and height are applied to localize individual objects. The paper demonstrates best-reported performance on segmentation, pixel labeling, and detection tasks in the NYU Depth v2 dataset.

"Scene Parsing with Object Instance Inference Using Regions and Per-exemplar Detectors" (doi:10.1007/s11263-014-0778-5) by Tighe et al. describes an approach to use a combination of region classification and exemplar-based detection to label pixels into categories, segment object instances, and infer a depth ordering. The approach particularly emphasizes the problem of training and inference in "open universe" datasets that contain rare objects and are continually evolving.

"Labeling Complete Surfaces in Scene Understanding" (doi:10.1007/s11263-014-0776-7) by Guo and Hoiem shows that inferring occluded background regions provides a more complete interpretation and improves accuracy in categorizing visible surfaces by enabling better context and shape priors. The method is applied to label pixels in images and to infer support surfaces at varying heights in rooms depicted in RGBD images.

Other papers focus on 3D geometric reasoning to improve segmentation, object detection, layout estimation, and to enable further reasoning about safety and stability:

"Towards Scene Understanding with Detailed 3D Object Representations" (doi:10.1007/s11263-014-0780-y) by Zia et al. models cars as deformable 3D wireframes, improving occlusion reasoning, and jointly infers the orientation and position of the ground plane and objects. The approach can recover 3D location and pose of objects from monocu-

D. Hoiem (✉)
Computer Science Department, University of Illinois, 201 N Goodwin Ave, Urbana, IL 61801, USA
e-mail: dhoiem@illinois.edu

J. Hays
Computer Science Department, Brown University, 115 Waterman, Providence, RI 02912, USA
e-mail: hays@cs.brown.edu

J. Xiao
Computer Science Department, Princeton University, 35 Olden Street, Princeton, NJ 08540, USA
e-mail: xj@princeton.edu

A. Khosla
32 Vassar St, D428, Cambridge, MA 02139, USA
e-mail: khosla@mit.edu

lar images with known intrinsic parameters. Experiments on the KITTI dataset demonstrate the helpfulness of layout and occlusion reasoning in estimating object 3D pose and layout.

"Indoor Scene Understanding with Geometric and Semantic Contexts" (doi:10.1007/s11263-014-0779-4) by Choi et al. describes a 3D Geometric Phrases representation for encoding 3D spatial relationships of objects. A scene is modeled as a hierarchical graph with a scene category, groups of objects that belong together such as a table and chairs, and the individual objects. 3D scene and object layout is inferred from a single RGB image, and the use of geometric and semantic context is shown to improve scene classification, object detection, and layout estimation.

"Scene Understanding by Reasoning Stability and Safety" (doi:10.1007/s11263-014-0795-4) by Zheng et al. describes an innovative method to model physical stability of objects from a 3D point cloud. The method recovers 3D volumetric primitives from voxels and groups primitives into objects to improve model stability, reasoning that objects in a static scene should be physically stable. The paper also demonstrates using the recovered scene model to evaluate safety and stability of objects under disturbances such as human activity, wind, or earthquakes.

Finally, "Graph-Based Discriminative Learning for Location Recognition" (doi:10.1007/s11263-014-0774-9) by Cao and Snavely addresses the problem of recognizing the location from which a photograph was taken. The key idea is to represent places as graphs that encode relations among images and train classifiers to identify similar places in local subgraphs. These classifiers are then used to assign a query photo to likely subgraphs, from which the location is recognized. The method outperforms the standard bag-of-words retrieval approach and performs similarly to more expensive direct feature matching techniques.