Fine-Grained Categorization for 3D Scene Understanding

Michael Stark[1]
mst@cs.stanford.edu

Jonathan Krause[1]
jkrause@cs.stanford.edu

Bojan Pepik[2]
bpepikj@mpi-inf.mpg.de

David Meger[3]
dpmeger@cs.ubc.ca

James J. Little[3]
little@cs.ubc.ca

Bernt Schiele[2]
schiele@mpi-inf.mpg.de

Daphne Koller[1]
koller@cs.stanford.edu

[1] Computer Science Department
Stanford University
Stanford, CA, USA

[2] Max Planck Institute for Informatics
Saarbrücken, Germany

[3] Computer Science Department
University of British Columbia
Vancouver, BC, Canada

Basic-level object category recognition has made remarkable progress over the last decade, both in image-level categorization and bounding box localization settings [3]. More recently, the recognition of finer-grained, subordinate categories is receiving increased attention [1, 2, 4, 7, 8, 11, 12]. It is deemed challenging due to the need to capture subtle appearance differences between categories while at the same time maintaining robustness to intra-category variations induced by changes in pose and viewpoint. As a consequence, the focus of previous work has been mostly on object categories *and* methods that favor discrimination by strong local appearance cues (such as random color image patches for birds [12]) or global image statistics (such as color histograms for flowers [8]).

Our paper goes beyond previous work on fine-grained categorization in two ways. First, in addition to exploring the task of fine-grained categorization itself, we suggest the use of fine-grained category predictions as an input for higher-level reasoning. This is based on the observation that fine-grained categories can encode, among other aspects, information about metric object sizes, which can in turn provide geometric constraints for scene-level reasoning. Accordingly, we focus our attention on rigid, geometric objects that can provide, if correctly categorized, reliable metric size estimates, and introduce a novel dataset [1] of fine-grained car types as a test bed for our approach (Fig. 1). This data set is annotated with 2D bounding boxes, viewpoint estimates, car types, and additionally includes metric object sizes (length, width, and height) for geometric reasoning.

Secondly, we design a fine-grained object class representation that captures variations in object shape and geometry rather than appearance [8, 12], in order to match the object class of interest. To that end, we introduce two different variants of utilizing part detections as indicators of object geometry, of varying complexity. Both are based on the best-performing object class detector to date, the DPM [5].

Novel car-types data set. We introduce a novel data set of fine-grained *car-types*, consisting of 1904 images of cars from 14 different categories (Fig. 1), annotated with category labels, 2D bounding boxes, and a viewpoint estimate (azimuth angle binned to 5 degrees).

Fine-grained categorization. Our approach follows the intuition that object geometry, and hence, category affiliation, can be encoded in the layout of its constituent parts. We thus design two different models that capture part layout, both building upon the DPM [5]: i) *part-bank*, a feature derived from response maps of a basic-level object class detector, similar in spirit to object-bank [6], and ii) *structDPM*, a multi-class variant of the DPM [10] that directly optimizes for fine-grained categorization. Our experiments show that both models outperform state-of-the-art classifiers by significant margins in fine-grained categorization. *structDPM* outperforms *part-bank*, at the cost of higher computational complexity.

3D Geometric reasoning. We demonstrate the potential of fine-grained category predictions to aid 3D geometric reasoning in a first, idealized experiment: the task is to predict the depth of a given object (its distance from the calibrated camera) from a single view, based on its 2D bounding box and metric size information derived from its predicted fine-grained category (Fig. 2). Our experiments on a public data set [9] confirm the benefit of these predictions over a baseline in the high precision domain.

[1] The data set will be publicly available under https://www.d2.mpi-inf.mpg.de/datasets.

Figure 1: Example images from our novel *car-types* data set with fine-grained category and viewpoint (azimuth angle) annotations.

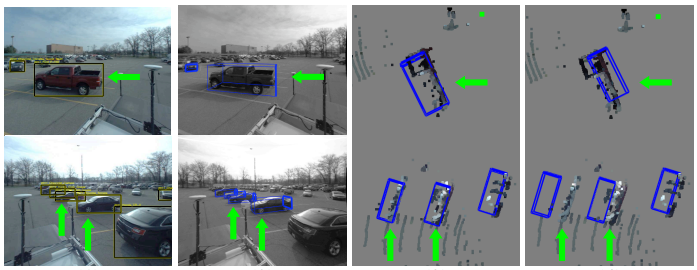

Figure 2: Depth estimation results. (1) 2D GT BB's with predicted fine-grained categories, (2) estimated 3D BBs for fine-grained categories, (3) point cloud top view for fine-grained, (4) for baseline. Green arrows: improvement. Best viewed in the electronic version, with magnification.

Acknowledgements. This material is based upon work supported by the Max Planck Center for Visual Computing and Communication and the Defense Advanced Research Projects Agency under Contract No. FA8650-10-C-7020.

[1] A. Bar-Hillel and D. Weinshall. Subordinate class recognition using relational object models. In *NIPS*, 2006.

[2] S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *ECCV*, 2010.

[3] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[4] R. Farrell, O. Oza, N. Zhang, V. I. Morariu, T. Darrell, and L. S. Davis. Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In *ICCV*, 2011.

[5] P. F. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010.

[6] L.-J. Li, Hao Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010.

[7] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011.

[8] M. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

[9] G. Pandey, J. R. McBride, and R. M. Eustice. Ford campus vision and lidar data set. *International Journal of Robotics Research*, 30(13):1543–1552, November 2011.

[10] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Teaching 3d geometry to deformable part models. In *CVPR*, 2012.

[11] C. Wah, S. Branson, P. Perona, and S. Belongie. Multiclass recognition and part localization with humans in the loop. In *ICCV*, 2011.

[12] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *CVPR*, 2011.