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

Abstract

Fine-grained categorization of object classes is receiving increased attention, since it promises to automate classification tasks that are difficult even for humans, such as the distinction between different animal species. In this paper, we consider fine-grained categorization for a different reason: following the intuition that fine-grained categories encode metric information, we aim to generate metric constraints from fine-grained category predictions, for the benefit of 3D scene-understanding. To that end, we propose two novel methods for fine-grained classification, both based on part information, as well as a new fine-grained category data set of car types. We demonstrate superior performance of our methods to state-of-the-art classifiers, and show first promising results for estimating the depth of objects from fine-grained category predictions from a monocular camera.

1 Introduction

The recognition of basic-level object categories [25] in natural images has made remarkable progress over the last decade, both in image-level categorization and bounding box localization settings [6]. More recently, the recognition of finer-grained, subordinate categories is receiving increased attention [2, 3, 7, 18, 20, 28, 30, 32, 33]. The problem of fine-grained categorization 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

http://dx.doi.org/10.5244/C.26.36

strong local appearance cues (such as random color image patches for birds [32]) or global image statistics (such as color histograms for flowers [20]). In this setting, computer vision techniques could be shown to facilitate fine-grained categorization tasks that are difficult even for humans due to the sheer number and diversity of subordinate categories [3, 20, 28].

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. Following this line of argumentation, we focus our attention on rigid, geometric objects that can provide, if correctly categorized, reliable metric size estimates, and introduce a novel dataset of fine-grained car types as a test bed for our approach. This data set is annotated with 2D bounding boxes, viewpoint estimates, car types, and additionally includes metric object sizes (length, width, and height) for use in geometric reasoning.

The second way our work departs from previous work [20, 32] is that we design a fine-grained object class representation that captures variations in object shape and geometry rather than appearance, 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 deformable part model (DPM [10]). The first variant is based on part detections provided by a pre-trained, generic detector for the object class. Similar in spirit to object-bank [16], it generates features from (part) detector responses by spatial pooling, and feeds them into a classifier for categorization. Relying on existing detectors, this first variant is computationally cheap, and outperforms state-of-the-art classifiers on our data set. The second variant uses the DPM directly for fine-grained categorization, by reformulating it as a structured output prediction problem [24], and directly optimizing a multi-class loss function. While this variant is computationally more demanding, it significantly improves over the first, since part detectors are now directly optimized for the task at hand. It outperforms state-of-the-art classifiers by a large margin.

In summary, our paper makes the following contributions. i) we introduce a novel data set of fine-grained car types that can serve as a test bed for future research on categorization of geometric objects as well as training data for scene-level reasoning methods based on fine-grained categories. ii) we propose two different variants of utilizing part detections for fine-grained categorization of geometric objects, and demonstrate superior performance compared to the state-of-the-art, and iii) to our knowledge, we are the first to attempt the application of fine-grained category prediction for the benefit of 3D scene-level reasoning. In particular, we show first results for the task of estimating the depth of objects relative to a calibrated monocular camera based on fine-grained category predictions.

2 Related Work

Understanding visual scenes in their entirety has been an important focus of computer vision research since its early days [4, 17, 19, 23]. It offers the prospect of removing false positive predictions by imposing additional constraints on the layout of objects in the scene, either in the form of scene priors (such as ground plane affinity [14, 31], mechanics [12], or human-centric functions [13]) or the likelihood of observations, given the current hypothesis (such as the agreement between predicted and observed object poses [1]). In this paper, we aim to expand upon the latter aspect, by providing fine-grained category predictions as additional input cues to scene-level reasoning. These cues complement existing observations, such as 2D object bounding boxes and viewpoint estimates. To our knowledge, no such use of fine-grained categorization has been reported in the literature so far.

According to cognitive psychology [27], the presence or absence of parts is related to the formation of basic-level object categories (a car has wheels, a chair does not), while specific properties of parts are indicative of subordinate categories (a sports car has a different trunk than a sedan). Several attempts have been made to exploit this principle for categorization, ranging from discriminative part features over generative constellations [2], over histograms of poselets [18], pose-normalized appearance features from geometric primitives [7], 3D shape models [33], to part localization with humans in the loop [3]. Our approach goes beyond these works in several ways: i) our part-based representation is based on the best performing object class detector to date, the DPM, ii) as a result, we explicitly encode spatial information ([18] does not), and iii) we do not require part annotations ([3, 7, 18, 33] do).

3 Fine-Grained Categorization with Deformable Parts

Our approach to fine-grained categorization is applicable for the wide range of object classes that are characterized by shape and geometry rather than appearance. It 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 build upon the deformable part model (DPM [10]), but represent part layout information differently.

3.1 Bank of Part Detectors

The basis for our first model is an existing DPM detector for the (basic-level) object class of interest. For example, if the fine-grained task at hand is to distinguish between different car types, the basis for our model is a car detector. While our method could be applied in combination with any detector capable of generating dense response maps of part detections, we chose the DPM since it has proven superior to other detectors for a variety of different object classes, including the rigid ones that we are focusing on [6].

Assuming that the detector has been run on an input image, we propose to form features from the generated part response maps, similar in spirit to object-bank [16]. Note that object-bank uses responses of (massive amounts of) entire object class detectors, lending itself to scene-classification problems that provide enough spatial support in terms of image area. In contrast, we focus on fine-grained classification of individual objects, which are likely to cover only small image regions, and expect to capture more fine-grained information by using responses of individual part detectors. Furthermore, using only part detectors is more efficient in terms of computation, since reasoning about pairwise deformation costs can be spared. Concretely, we compute spatial pyramid (SP) [15] representations (1×1, 2×2, and 4×4 cells) at different scales over the response maps of all parts, over all components of the DPM. For each SP cell, we memorize min and max responses (pooling), concatenate all values into a single feature vector, and train a linear SVM with L2 loss and regularizer. In the following, we refer to this model as part-bank (PB).

3.2 Multi-Class Deformable Part Model

The second model constitutes a proper extension of the DPM [10], which we implement based on its reformulation as a structured output prediction problem proposed by [24], from which we borrow the notation in the following (we omit recapitulating the well-known original DPM formulation [10]). Specifically, we phrase the DPM as a (latent) linear multi-class SVM that can be coherently optimized for the multi-class problem, without the need for a posteriori output coding, such as 1-vs-all or 1-vs-1 schemes [22]. In the following, we refer to this model as structDPM.

The structDPM is trained from a set $\{x_i, y_i\}$ of images x_i and class labels $y_i \in \{1, ..., K\}$. Similar to [10], each class y is represented in the model with a set of n components $\{c^y\}$, where n is a free parameter of the model. The structDPM is the union of components across all classes, $\{c^1\} \cup \{c^2\} \cup \ldots \cup \{c^K\}$. The mapping of training examples to components is latent, with the constraint that for every training example x_i, only components of class y_i can be assigned to it. Each component c is composed of a dedicated root p_c^0 and a set of deformable parts p_c^k, the positions of which are aggregated in latent variables $h = \{p_c^k\} \cup c$, together with the component assignment c. Each part is characterized by a HOG [5] template F_c^k and a spatial deformation cost w.r.t. the root d_c^k. For notational convenience we first stack all model parameters in a single vector for each component c, $\beta_c = \left(F_c^0, F_c^1, \ldots, F_c^n, d_c^1, \ldots, d_c^n, b_c\right)$, where b_c is a bias term, and further into a single vector for an entire model $\beta = (\beta_1, \ldots, \beta_M)$. The features are stacked accordingly: $\Psi(x,y,h) = (\psi_1(x,y,h), \ldots, \psi_M(x,y,h))$, with $\psi_k(x,y,h) = [c = k]\psi(x,y,h)$ ($[\cdot]$ is Iverson bracket notation) being the features computed for component k, where $k \in \{c^y\}$. The vector $\Psi(x,y,h)$ is zero except at the c'th position, i.e., $\langle \beta, \Psi(x,y,h) \rangle = \langle \beta_c, \psi_c(x,y,h) \rangle$. During training, we optimize the following latent structured SVM objective:

$$\begin{array}{ll} \min\limits_{\beta,\xi \geq 0} & \frac{1}{2}\|\beta\|^2 + C\sum\limits_{i=1}^{N}\xi_i \\ \text{sb.t.} & \forall i, \bar{y} \neq y_i : \max\limits_{h_i}\langle \beta, \Psi(x_i, y_i, h_i)\rangle - \max\limits_{h}\langle \beta, \Psi(x_i, \bar{y}, h)\rangle \geq \Delta(y_i, \bar{y}) - \xi_i \end{array}$$

where Δ is a loss function, which we instantiate as $\Delta(y, \bar{y}) = [y \neq \bar{y}]$. For both training and test, we allow the root part to move inside the object bounding box by considering all hypotheses which have an overlap of at least 0.4. At test time, we solve $\text{argmax}_{(y,h)}\langle \beta, \Psi(x,y,h)\rangle$.

4 Experiments

In the following, we carefully analyze the performance of our models. To that end, we introduce a novel data set of fine-grained *car-types*, and conduct experiments in two different settings: first, we evaluate fine-grained categorization in isolation, as a standard multi-class classification task (Sect. 4.2), comparing to state-of-the-art classifiers. Second, we explore fine-grained categorization in the context of 3D scene understanding, showing promising results for estimating object depth from fine-grained category predictions (Sect. 4.3).

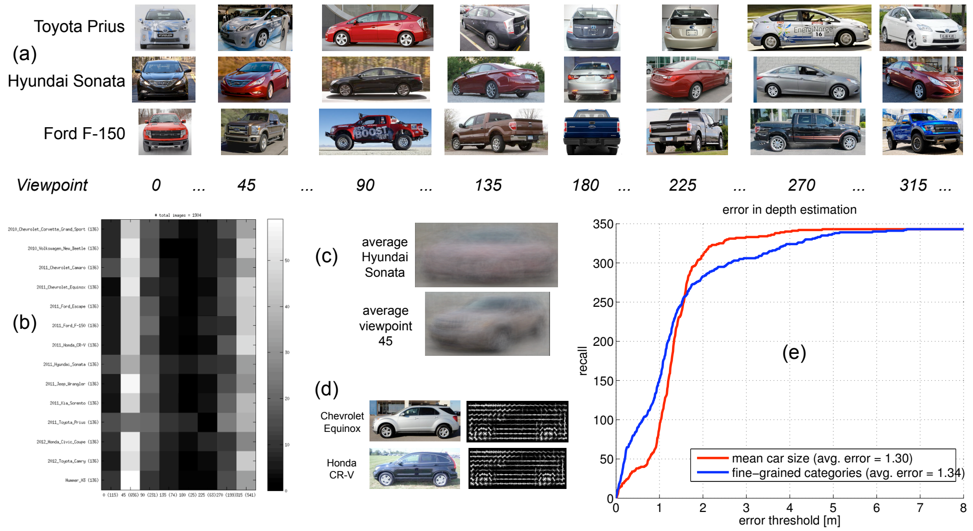

Figure 1: Our novel *car-types* data set (Sect. 4.1): (a) example images, (b) statistics, (c) average images, (d) HOG features. (e) Comparison of depth estimation error (Sect. 4.3). This figure is best viewed in the electronic version, with magnification.

4.1 Novel Fine-Grained Car Data Set

We introduce a novel data set of fine-grained *car-types*, which we will make publicly available upon publication (Fig. 1) [1]. It consists of 1904 images of cars from 14 different categories (Fig. 1 (b)), downloaded from the internet. In particular, we queried google image search with terms corresponding to the most frequently appearing sedans, SUVs, sports cars, and compact cars, according to a car trading website. Downloaded images were manually filtered for those that depict at least one car of the queried category in a prominent position. Images are annotated with category labels, 2D bounding boxes, and a viewpoint estimate in the form of the azimuth angle, binned to 5 degrees (we report results on the standard 45 degree binning in the experiments of Sect. 4.2).

Fig. 1 (a) gives sample images from 3 categories and 8 different viewpoint bins, cropped to approximately the object bounding box. Fig. 1 (b) gives the number of images for each category and viewpoint, together with the corresponding marginals (in parentheses). We note that the data set is heavily biased w.r.t. viewpoints, which reflects the availability of images we encountered during data collection. It proved almost impossible to collect more than a handful of images for certain combinations of car-type and viewpoint. Fig. 1 (c) and (d) highlight the challenge of the fine-grained classification problem: images of all categories from a certain viewpoint are better aligned than images from a certain car-type across all viewpoints (c), and differences in HOG feature space are hard to spot even visually (d). For evaluation, we split the data set into 50% *train*, 25% *val*, and 25% *test* images.

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

setup		training	car-type	car-type × vp	car-type × vp	car-type × vp
		test	car-type	car-type	vp	car-type × vp
		# categories	14	14	8	104
method	i)	HOG [5]	77.5	81.3	87.8	75.6
		LLC [29]	84.5	82.6	84.2	72.9
	ii)	PB(DPM) *(ours)*	84.0	84.9	88.0	77.1
		PB(mvDPM) *(ours)*	85.3	87.0	88.2	79.4
		PB(structDPM) *(ours)*	89.9	85.5	87.6	77.7
	iii)	structDPM *(ours)*	**93.5**	**88.2**	**88.4**	**79.8**
	iv)	HOG+LLC+PB(mvDPM) *(ours)*	89.1	**88.9**	**89.9**	**81.3**
		HOG+LLC+structDPM *(ours)*	**90.3**	86.3	88.9	79.4

Table 1: Comparison of classification accuracy on the *car-types* data set in %, including HOG [5] and LLC [29]. Best individual and combined methods are shown in bold font.

4.2 Fine-Grained Categorization

We first evaluate our fine-grained categorization in isolation, as a standard multi-class classification task. We train on the designated *train* data defined by our data set, use *val* for parameter optimization, and test on *test*. For training and test, classifiers are provided images as well as ground truth object bounding boxes, since the task is classification, not detection.

Methods. Tab. 1 gives the results for fine-grained categorization on our *car-types* data set, measuring the accuracy of classification as the fraction of correctly classified instances in the test set. It compares four different groups of approaches in its sections: i) baselines, ii) part-bank, iii) structDPM, and iv) combinations of i), ii) and iii). As baselines (i), we consider a HOG [5] template with a linear SVM, and locality constrained linear coding (LLC [29]), which is one of the most powerful image-level classifiers to date (among the state-of-the-art on Caltech-101 [8] and -256 [11] classification benchmarks). For ii), we compare our part-bank (PB) computed on response maps of the DPM [10] car detector as provided by the authors [9] (PB(DPM)), and part-bank computed on response maps of the bank of 8 viewpoint-dependent DPMs proposed by [1] (PB(mvDPM)). Since the latter explicitly distinguishes between different viewpoints, we expect the corresponding part response maps to be more informative than the ones of PB(DPM). We also add part-bank computed on response maps of our structDPM (PB(structDPM)). For iii), we train our structDPM with 2 components per fine-grained category. For iv), we consider stacking-based combinations of the baselines with the best performing part-bank method PB(mvDPM) (HOG+LLC+PB(mvDPM)) and structDPM (HOG+LLC+structDPM).

Settings. Columns of Tab. 1 correspond to different evaluation settings, characterized by the set of class labels provided to the different methods during training: we distinguish *car-type* (col. 1) and both *car-type* and *viewpoint* (col. 2-4). We do the same for testing, which ranges from predicting only *car-type* (col. 1, 14 class problem) to predicting both *car-type* and *viewpoint* (col. 4, 104 classes). Col. 2 and 3 marginalize the predictions of col. 4 over *viewpoint* (col. 2, 14 classes) and *car-type* (col. 3, 8 classes), respectively. Note that the data set does not contain enough images for 8 particular combinations of car-type and viewpoint, which leaves us with 104 classes for the car-type × vp setting.

Car-type. In Tab. 1 col. 1, we observe a clear ordering of performance. While HOG performs moderately (77.5%), it is outperformed by LLC (84.5%) by a large margin (7%). Equally, our PB(DPM) improves over HOG by 6.5%, performing on par with the state-of-the-art LLC. Enriching part-bank with viewpoint information in fact improves performance by 1.3% (PB(mvDPM), 85.3%), and is significantly increased (5.9%) by using parts optimized for the classification problem (PB(structDPM), 89.9%). Using the structDPM end-to-end further increases performance to a striking 93.5%, which is a 9.0% improvement to the best baseline method LLC, and can not be attained by either of the combined methods.

Car-type × vp. In Tab. 1 col. 4, we observe a general drop in performance compared to col. 1, due to the increased difficulty of the classification problem (104 vs. 14 classes). The performance of the baselines is reversed – the rigid HOG (75.6%) apparently benefits more from the viewpoint alignment of the training data than LLC (72.9%). Both baselines are consistently outperformed by all variants of part-bank. Again, adding viewpoint information helps (increase from 77.1% for PB(DPM) to 79.4% for PB(mvDPM)). PB(structDPM) performs on par (77.7%). As in col. 1, the best performance for a single method is achieved by structDPM (79.8%), which is remarkable for a 104 class problem. Combining methods improves marginally (to 81.3% for HOG+LLC+PB(mvDPM)).

Marginalizing over *viewpoints* (col. 2), we observe an increase in performance compared to directly predicting the *car-type* (col. 1) for some methods (HOG +3.8%, PB(DPM) +0.9%, PB(mvDPM) +1.7%), and a decrease for others (LLC -1.9%, PB(structDPM) -4.4%, structDPM -5.3%, HOG+LLC+PB(mvDPM) -0.2%, HOG+LLC+structDPM -4.0%).

Marginalizing over *car-types* (col. 3), the performance largely follows the ordering of col. 4. Both baselines (HOG 87.8%, LLC 84.2%) are consistently outperformed by our part-bank classifiers (PB(DPM) 88.0%, PB(mvDPM) 88.2%, PB(structDPM) 87.6%), topped by our structDPM (88.4%) and the combined classifiers (HOG+LLC+PB(mvDPM) 89.9%, HOG+LLC+structDPM 88.9%). In comparison to an existing data set for viewpoint classification into 8 azimuth angle bins [26], where classification is tied to an even more difficult detection setting, the best achieved accuracies on our new data set are considerably worse (89.9% vs. 97.9% [24]). This suggests that our data set is also a more challenging test bed for viewpoint classification.

Summary. We conclude that both part-bank and structDPM outperform the baselines HOG and LLC by significant margins, in both *car-type* and the even more challenging *car-type* × *vp* settings. While the computationally more expensive structDPM shows a clear benefit in the former setting, PB(mvDPM) offers a good compromise between computational efficiency at training time (since it relies on pre-trained detectors) and performance, in particular for the latter setting, where it loses only 0.4% compared to structDPM. Combining methods hardly improves, suggesting that our methods are not complementary to HOG and LLC, but rather subsume information encoded by either of them.

4.3 3D Geometric Reasoning

While Sect. 4.2 evaluates our fine-grained categorization in isolation, we now move on to the more challenging task of applying it in the context of a 3D scene understanding task, on a recently proposed street scene data set [1, 21]. To that end, we design an idealized experiment,

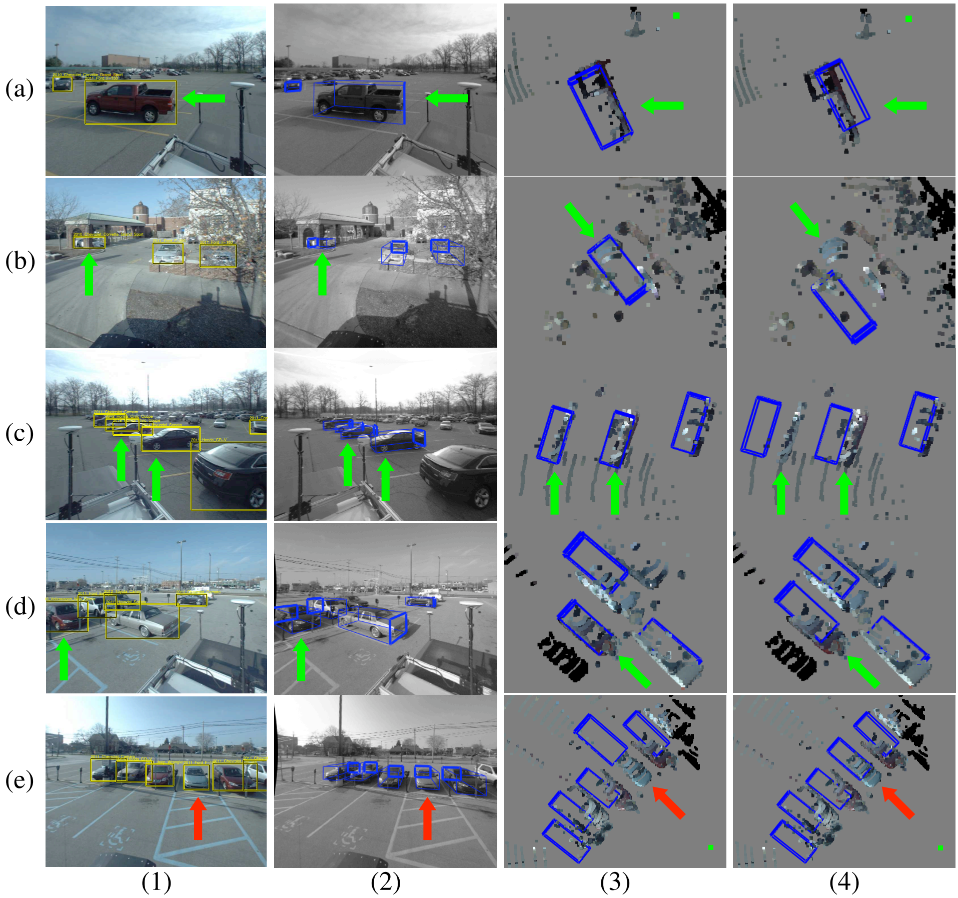

Figure 2: Depth estimation results. (1) 2D GT BBs with predicted fine-grained category labels, (2) estimated 3D BBs when using fine-grained category information, (3) point cloud top view for fine-grained, (4) for mean metric sizes. Green: improvement, red: failure. This figure is best viewed in the electronic version, with magnification.

in which we isolate the contribution of fine-grained category information from possible deficiencies of other system components (such as object localization). While this experiment constitutes a best case evaluation, it highlights that fine-grained category information has the potential to provide valuable constraints in a scene-level reasoning context.

Data set. We use the Ford campus vision and lidar data set [1, 21] for testing, as it provides calibrated camera images as well as registered point cloud data that can serve as the basis for metric 3D evaluation. Applying fine-grained categorization on this data set is challenging, as its statistics deviate largely from our *car-types* data set used for training, both w.r.t. the imagery (images are taken from an omni-directional camera mounted on a car roof, resulting in image distortions despite correction, and more elevated views of nearby objects) and the objects depicted (cars are not restricted to the types in our data set, and they appear at largely varying, often tiny, scales and are heavily occluded). The data set consists of a number of distinct street scenes, from which we use the test set defined by [1], consisting of 141 images in total. Fig. 2 col. (1) shows examples. We manually annotate the corresponding point clouds with 3D bounding boxes for all visible car objects above a certain size.

Task. We consider the task of predicting the depth of a given object (its distance from the camera) from a single view of the calibrated camera. For that purpose, we are given the object's *ground truth* 2D bounding box (which we derive from our 3D annotations), its *ground truth* viewpoint (its azimuth angle), and its *estimated* physical extent (length, width, and height) as an input. This task is based upon the fact that the depth of an object is a function of its extent, once the other parameters (projection to the image plane and rotation) are fixed.

To identify this depth, we cast a ray from the camera center through the center of the 2D object bounding box in the image plane. We then instantiate a 3D bounding box along that ray, aligned to the ground plane (which is enabled by the camera calibration including the up-vector of the ground plane). We then size the box according to our fine-grain category estimate, and rotate it to match the ground truth azimuth. Finally, we slide it along the ray, such that the overlap between its 2D projection and the ground truth 2D bounding box is maximized. Maximization is done via exhaustive search over discrete positions on the ray. Fig. 2 visualizes 2D object bounding boxes (col. (1)) together with their estimated 3D bounding boxes (col. (2)).

Methods. We compare two different methods for estimating the physical extent of an object, which serves as the basis for computing its depth. For the first one, we determine the metric sizes of all *car-types* in our data set (length, width, height) from internet product information. We then apply our fine-grained categorization (structDPM) to all 2D ground truth bounding boxes in the test set, and chose the size of an instantiated 3D object bounding box according to the metric information for the predicted fine-grained category. The second one is our baseline: it ignores fine-grained categories, and instantiates all 3D object bounding boxes with the mean over all metric sizes in our *car-types* data set.

Results. Fig. 1 (e) gives the results for depth estimation, comparing the performance of using fine-grained category information (blue) with using the mean over all metric sizes (red). It plots the recall of objects with correctly estimated depth according to an error threshold (in meters) vs. that threshold. We observe that using fine-grained category information in fact results in a noticeable improvement in the high precision region of the curve, up to an error of 1.5m (the blue curve stays consistently above the red curve). Beyond that point, the mean over car sizes proves to be more robust than our fine-grained category predictions. This is understandable, given that the test set is quite different from our *car-types* data set used for training, in particular w.r.t. the occurring car-types. Nevertheless, the total average error for fine-grained category predictions is only 4 cm larger than for the mean car sizes.

Fig. 2 visualizes example results. Green arrows highlight improved depth estimates resulting from fine-grained category information, red arrows mark failure cases. In (a), we correctly predict a Ford F150, which is considerably larger than the mean car size, leading to a more accurate depth estimate. (b) shows the same effect with a Chevrolet Corvette Grand Sport. In (c), we correctly predict smaller cars than the mean (Hyundai Sonata and Honda Civic Coupe), also in (d), where we predict a VW New Beetle (which is wrong, but the actual car is small, and can be mistaken for a Beetle). In (e), we mistake the marked car for being an F150, leading to an overestimated size and hence depth.

5 Conclusion

We have considered fine-grained categorization of geometric object classes, aiming to use fine-grained category predictions in a 3D scene-understanding context. We introduced two different methods that utilize part detectors to encode category-specific information, which we have shown to outperform baseline classifiers on a newly proposed car-types data set by a significant margin. We further showed first results on using fine-grained category predictions for estimating object depth, which we consider a valuable starting-point for future research.

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.

References

[1] S.Y. Bao and S. Savarese. Semantic structure from motion. In *CVPR*, 2011.

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

[3] 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.

[4] R. A. Brooks. Symbolic reasoning among 3-d models and 2-d images. *Artificial Intelligence*, 17(1-3):285–348, 1981.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[6] 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.

[7] 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.

[8] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *PAMI*, 28 (4):594–611, 2006.

[9] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 4. http://people.cs.uchicago.edu/ pff/latent-release4/.

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

[11] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007. URL http://authors.library.caltech.edu/7694.

[12] A. Gupta, A. A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *ECCV*, 2010.

[13] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3d scene geometry to human workspace. In *CVPR*, 2011.

[14] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. *IJCV*, 80(1):3–15, 2008.

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[16] 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.

[17] D. G. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31(3):355–395, 1987.

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

[19] D. Marr and H. K. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. Roy. Soc. London B*, 200(1140):269–194, 1978.

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

[21] 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.

[22] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011.

[23] A. P. Pentland. Perceptual organization and the representation of natural form. *Artificial Intelligence*, 28(3):293–331, 1986.

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

[25] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes Braem. Basic objects in natural categories. *Cognitive Psychology*, 1976.

[26] S. Savarese and L. Fei-Fei. 3D generic object categorization, localization and pose estimation. In *ICCV*, 2007.

[27] B. Tversky and K. Hemenway. Objects, parts, and categories. *Journal of Experimental Psychology: General*, 1984.

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

[29] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[30] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[31] C. Wojek, S. Roth, K. Schindler, and B. Schiele. Monocular 3d scene modeling and inference: Understanding multi-object traffic scenes. In *ECCV*, 2010.

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

[33] M. Z. Zia, M. Stark, K. Schindler, and B. Schiele. Revisiting 3d geometric models for accurate object shape and pose. In *3rd International IEEE Workshop on 3D Representation and Recognition*, 2011.