Going into depth: Evaluating 2D and 3D cues for object classification on a new, large-scale object dataset

Björn Browatzki
MPI for Biological Cybernetics
Tübingen, Germany
bjoern.browatzki@tuebingen.mpg.de

Jan Fischer
Fraunhofer IPA
Stuttgart, Germany
jan.fischer@ipa.fhg.de

Birgit Graf
Fraunhofer IPA
Stuttgart, Germany
birgit.graf@ipa.fhg.de

Heinrich H. Bülthoff
MPI for Biological Cybernetics
Tübingen, Germany
heinrich.buelthoff@tuebingen.mpg.de

Christian Wallraven
Korea University
Seoul, Korea
wallraven@korea.ac.kr

Abstract

Categorization of objects solely based on shape and appearance is still a largely unresolved issue. With the advent of new sensor technologies, such as consumer-level range sensors, new possibilities for shape processing have become available for a range of new application domains. In the first part of this paper, we introduce a novel, large dataset containing 18 categories of objects found in typical household and office environments—we envision this dataset to be useful in many applications ranging from robotics to computer vision. The second part of the paper presents computational experiments on object categorization with classifiers exploiting both two-dimensional and three-dimensional information. We evaluate categorization performance for both modalities in separate and combined representations and demonstrate the advantages of using range data for object and shape processing skills.

1. Introduction

The availability of consumer level range sensors such as the Microsoft® Kinect™ has opened up new possibilities for shape processing going beyond 2D color information. This hardware will lead to new applications in computer vision, interactive gaming, and also robotics. Many of these application domains require perceptual capabilities such as object recognition and categorization. In this paper, we have chosen to demonstrate the advantages of having access to *range data* in the robotics domain.

In robotics, it is already common to employ laser scanners or other ranging devices for tasks such as navigation and self-localization. Here, we want to study the effect of incorporating range, or 3D[1], data for the purpose of recognition and categorization of objects. The target platform for

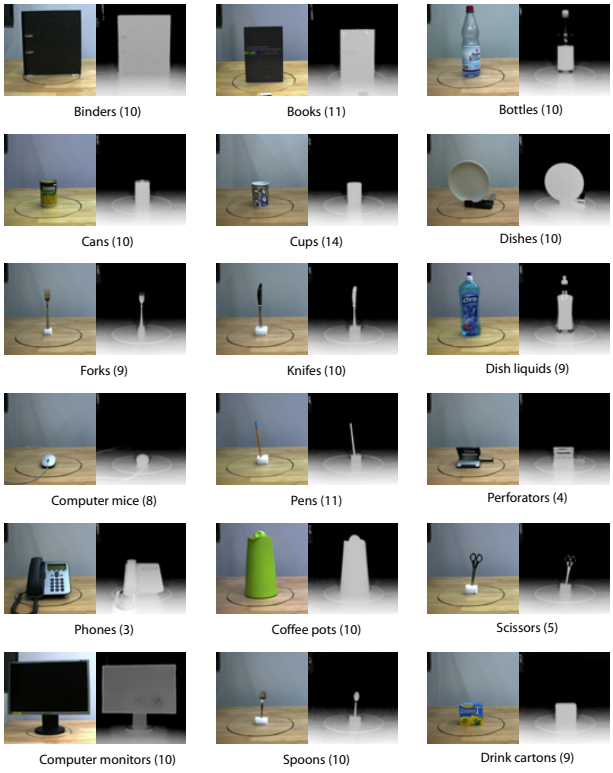
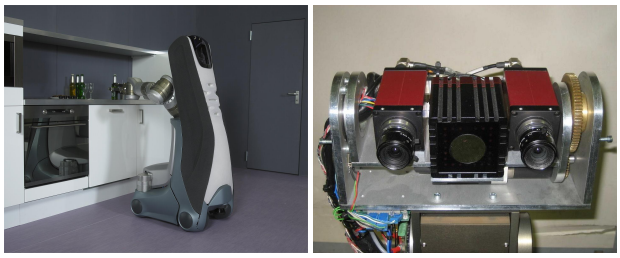


Figure 1: Color and depth images of categories in dataset. First view of first object for each category. Number of exemplars in parentheses.

[1] As we will use algorithms from computer graphics developed for 3D applications, we will refer to the range data as "3D" rather than 2.5D.



(a) Care-O-bot® 3 (b) Sensor head

Figure 2: Sensor setup of Care-O-bot® 3 for data acquisition. One stereo rig augmented with a range camera.

our evaluation is the service robot *Care-O-bot® 3* [20] developed by the Fraunhofer IPA (see Fig. 2a). It is equipped with two color cameras and a time-of-flight (TOF) camera. The TOF camera emits modulated infrared light and uses the phase shift of the reflected light to measure the distance to the reflection surface. The range data acquired by the setup is very similar in quality to that of the Microsoft® Kinect™. Motivated by the intended application area and the available sensors, we focus on the classification of unknown objects exploiting both modalities in the joint domain. Classification requires training data to learn the common attributes that describe a specific category. To acquire this data we recorded a dataset containing 154 object exemplars belonging to 18 categories, which occur in typical household and office scenarios[2]. We also evaluated classification performance on this dataset using data gathered by 2D and 3D sensors, and study how cue combination of both cues can lead to enhanced categorization results.

2. RELATED WORK

2.1. Datasets

There are many well-known image databases for the evaluation of object categorization algorithms based on *2D* information. Arguably one of the most popular ones in computer vision research is the Caltech-101 dataset [7]. It contains images from 101 categories with high intra-category variability. An extended version, the Caltech-256 dataset [10], contains 256 categories with a proper taxonomic structure. Other popular databases include the Graz-01 [18] database, the ETHZ shape dataset [8], or the PASCAL Visual Object Class database [6].

Most computer vision literature focuses exclusively on information obtained from 2D images. In that case the existing databases offer a suitable test-bed to evaluate algorithms and compare results. However, for 3D data obtained for example from laser scans, or as in our case range cameras, there are few comparable databases, or the data is rather sparse. Sun et al. [23] have collected a data set containing three object categories (mice, mugs, staplers) with 10 object instances each. They obtain depth information using a structured-light stereo camera. Lai et al. [17] have recently introduced the RGB-D dataset. This database contains color and depth information of 300 objects from 51 categories of household objects and is organized in a hierarchical structure. The dataset was used to train an object recognition system capable of detection 20 specific instances of objects as well as 4 object classes (bowl, cup, coffee mug, soda can) in cluttered scenes. Our paper extends the work of Lai et al. in that our goals are to present additional object categories, *and* to systematically study how the combination of multiple feature types across both modalities affects the classification performance of various object categories; that is, we want to demonstrate the benefits of combining 2D (color, and texture) with 3D (range) information.

2.2. Object categorization

As the literature on object categorization based on 2D information is vast, we chose to focus here on the state-of-the-art in 3D object processing. Most approaches based on 3D information deal with recognizing previously seen objects [12]. Often the focus lies on detection of specific objects in complex scenes using local surface descriptors on key points [14]. An extensive survey of 3D object recognition techniques is given by Campbell and Flynn in [5]. Classification tasks, posing the challenge of assigning class labels to unknown objects, have gained less attention so far. Ruiz-Correa et al. [21] introduced symbolic surface signatures to label surface regions in range scans. Objects were recognized by assigning regions to the object classes snowmen, rabbits, and dogs. A part-based classification approach was proposed by Huber et al. [13]. Eight classes of vehicles were separated into front, middle and back part. Based on Spin Images shape parts are recognized and the object class inferred using a generative model. The majority of literature on 3D object classification deals with cases in which the object geometry is available in the form of polygon meshes. A common scenario is similar shape retrieval from 3D object datasets [9]. Numerous approaches have been proposed to compare 3D models and calculate a measure of similarity. Bustos et al. provide an exhaustive overview of such shape matching approaches [4].

3. A 2D & 3D object dataset

Our first task consisted of building a resource for testing object categorization using 2D and 3D information. For this, we recorded 18 categories of objects that are likely to

[2]The dataset is available for download at http://www.kyb.mpg.de/~browatbn.

be encountered by a robot operating in a household environment. Each category contains between 3 and 14 objects. In Fig. 1 all categories are depicted including the respective number of exemplars. Each object was put on a step-motor-controlled turntable in their default orientation (except for silverware objects and scissors, which were propped up on a stand) and we recorded views every 10° around the vertical axis, yielding 36 views per object. In total, the dataset contains 154 objects with $154 \times 36 = 5544$ views. Every view consists of two high-resolution (1388×1038 px) 2D color images and a range scan obtained from a PMD[TM] CamCube 2.0 time-of-flight camera. The resolution of the range images is 204x204 px with an accuracy of approximately $\pm 1\%$ with respect to the measured distance. Compared to the consumer version of the Microsoft® Kinect™, the horizontal and vertical dimensions of the depth image are a little lower, and those of the RGB image a little higher—the combined image, however, is of similar resolution.

4. 2D and 3D features and cue combination

4.1. Features

We extract four 2D descriptors from the color images as well as four 3D descriptors from the range scans. In both cases the set of descriptors is intended to exploit different properties, aimed at providing complementary information. For the 2D data we took Speeded Up Robust Features (SURF) [1], Pyramids of histograms of oriented gradients [3], Self Similarity Features [22] and color histograms (CIELAB color space).

The 2D descriptors are widely used in current computer vision research. For 3D data, the choice of feature descriptors is not as large as in the 2D case. We believe that the following selection of descriptors covers a suitable range of shape characteristics[3]: 3D Shape Context (SC3D) [16], Depth Buffer [11], Shape-Index Histograms [15] and MD2 Shape Distributions [19].

Most of the features listed above are descriptive enough for being used to build strong classifiers but still fast enough to compute so that real time application is feasible. Extraction times range between <1 ms for color histograms and $\approx$250 ms for Depth Buffers on a standard desktop PC with a 3GHz dual-core CPU and 2GB RAM.

SURF, Self Similarity, and SC3D are local feature descriptors. To transform the local features into a global descriptor we employ the common Bag of Words method. A collection of feature vectors is taken from various objects across all object classes and clustered in the respective feature space. The resulting set of cluster points (vocabulary) is used to quantify the local features. A histogram is created that represents the local feature distribution in respect to the vocabulary entries. We used a vocabulary size of 50 throughout all experiments. We did not alter this parameter as we did not notice a significant sensitivity to changes in the quantization process.

[3]We also experimented with incorporating absolute object size (in meters) as an additional cue. This, however, did not lead to an increase in classification performance.

4.2. Training of single classifiers

After extracting features we train one support vector machine (SVM) for each feature type and each class. If n is the number of classes and d the number of feature types, we create $n \times d$ classifiers in total. SVM parameters are optimized through cross-validation on the training set. Each of these classifiers predict whether an unknown sample is likely to be a member of a certain object class based on specific feature type. It is obvious that we do not always obtain a consistent prediction across different feature types for a certain class. One class might be more dependent on information supplied by a specific feature than a different class.

4.3. Ensemble prediction

To obtain a joint prediction from the different classifiers we train a multilayer perceptron (MLP) for each object class. The outputs of the single classifiers are used as training samples. In our case a training sample for the perceptron would be an eight-dimensional vector produced by the eight different descriptors. To evaluate the performance of either 2D or 3D data independently, the MLPs are trained with the four-dimensional input vector retrieved from the four descriptors of one modality. We do not define the structure of the network a priori. The free parameters of the MLP, that is, the number of layers as well as the number of nodes per layers are found through cross-validation on the training set. Cross-validation is done for each class separately.

5. Evaluation

As a baseline for further evaluations, we tested the 2D part of our approach on the Caltech-101 database. Using a standard training setup of 30 training samples, we obtained a recognition rate of 60.1% by relying on the combination of the four 2D features. The performance stays somewhat below results reported in current literature (e.g. [2]), which might be due to more extensive parameter tuning and more optimized classifier kernels that these papers employ. Nevertheless, for simple 2D features without any additional shape/configuration modeling of the categories, these performance levels are encouraging—especially as the features can be evaluated in near-real-time.

5.1. Overall categorization performance

For our evaluations, the database was randomly split into training and test sets[4]. 6 objects per class were used for

[4]For the evaluations, we excluded the classes *perforator* and *phone* due to the low number of exemplars. Furthermore, we combined *forks*, *knives*,

Table 1: Classification performance

Descriptor	Performance
SURF	42.4%
PHOG	69.9%
Self-Similarity	41.7%
Color	26.6%
2D only	66.6%
Shape Distributions	25.4%
Shape Index	34.6%
Shape Context 3D	55.2%
Depth Buffer	72.9%
3D only	74.6%
2D + 3D	82.8%

training, the remaining objects were put into the test set. For each object, we selected 18 views for training and 18 views for testing. The views were equally distributed across the 36 available views, i.e. one view every 20°. The training set consisted of 82 objects with a total of 1476 views. The test set contained 74 objects with 1332 views. Features were extracted for each view and classifiers were trained according to Sec. 4. Classification results are listed in Table 1. The values represent average classification accuracy normalized by class size. In case only 2D descriptors are used, we obtain 66.6% correct categorization, whereas the 3D descriptors yield 74.6% correct results. Combining both 2D and 3D descriptors, the performance increases significantly to 82.8% correct.

We also tested combinations of all descriptors from one modality (2D or 3D), which did not lead to a significant improvement of the results. In the 3D case, the joint performance is 2.3% higher than the performance of the best single descriptor (Depth Buffer). The performance of all 2D descriptors combined is around 3.3% lower than the best single one (PHOG). However, if the two modalities are fused, the results improve clearly, which is due to the different object characteristics that the features latch on to.

In Fig. 3 categorization performance is shown for each class separately. This data shows that some classes are more sensitive to 2D information (e.g. *silverware*), whereas for other classes 3D cues are more effective (e.g. *drink cartons*). For almost all classes, however, the combination of both information pathways leads to an increase in performance.

5.2. ROC curves and confusion matrices

The rate of correct results is only one part of the story. First of all, one might wish to specifically set acceptance thresholds for the different classes, in order to control the number of false alarms versus hits. The ROC curves for the 2D, 3D and combined cases are shown in Fig. 5. These results show that the 3D cues are always better than the 2D cues, and that the combined case always provides a clear increase over the single modalities. From these curves, the EER (Equal-Error-Rates), which provide a good indication of the trade-off between false alarms and hits are determined as: 2D (12.0%), 3D (8.7%), Combined (6.2%).

In some cases, it might be interesting to look at the pattern of confusions to determine, for example, the degree of generalizability of the features, or to provide a different, more effective clustering of categories. Fig. 4 shows the confusion matrices for all features, as well as separately for the 2D and 3D features. Consistent with the previous categorization results, predictions are more consistent if based on 3D features than on 2D features (cf. number of non-zero off-diagonal elements in each matrix). Furthermore, if we examine the categories that are often confused in the combined case (e.g. *dish liquids* and *bottles*), we see that those contain the classes that also are confused for both 2D and 3D data. If in contrast at least one modality is able to clearly distinguish between two classes, the result in the combined case is determined by the more descriptive modality. This does not only suggest that the two modalities capture different class properties but also that they can be combined very effectively.

Finally, the remaining confusions after cue combination include categories that are hard to distinguish not only for a computational system. We observe high confusion rates for *cans* and *drink cartons*, as well as *dish liquids* and *bottles*. Especially if only 2D data is used *drink cartons* are very likely to be confused with *cans*, as in the 2D images both objects appear as rectangular items with varying texture. Even here, however, the addition of 3D features (which will add the curvature of cans) reduces the number of confusions substantially. As an additional observation, frequently confused categories might rather be distinguished by their functionality than by their appearance/shape. Although the generalizability seems already quite good with the number of exemplars currently in the database, we would expect to be able to capture the inherent category structure even better with a larger number of exemplars in the difficult categories.

In summary, employing 3D descriptors gives better results than 2D descriptors. However, by combining both cues we obtain an even better overall performance with significant increases for many object classes.

5.3. Generalizability across views and number of objects

The classification results with respect to the number of objects used for training are plotted in Fig. 6. The number of views per object was kept constant. As before, a view was selected every 20°, resulting in 18 views for each ob-

and *spoon* into the joint category *silverware*.

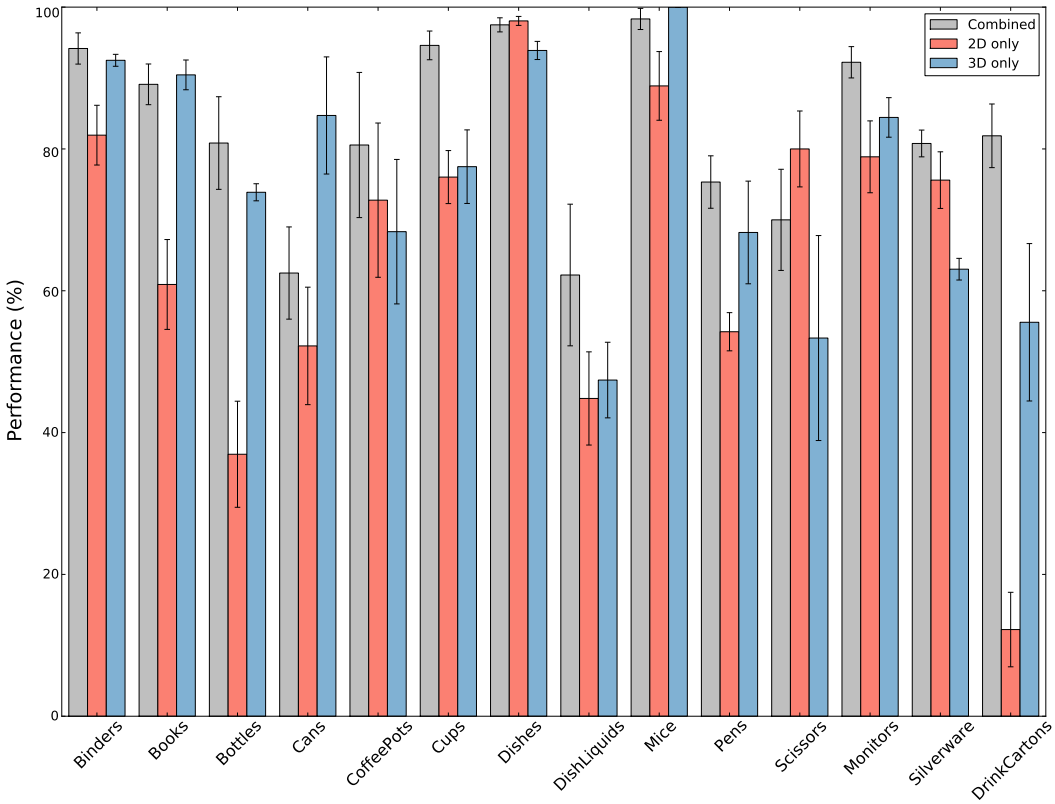


Figure 3: Classification results of single classes.

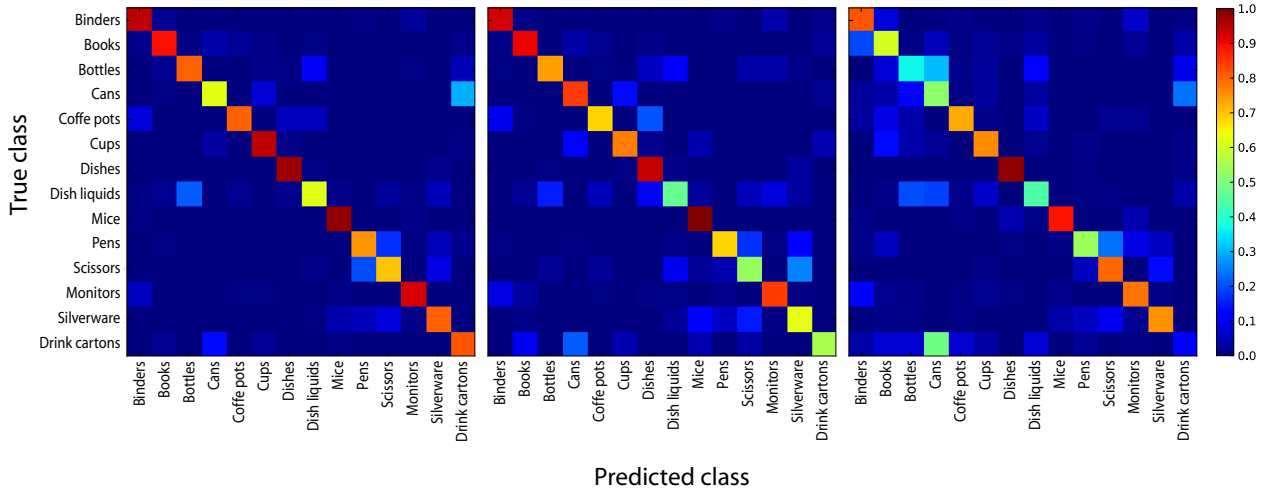


Figure 4: Confusion matrices: Combined cues (left), 3D only (middle), 2D only (right)

ject. Evaluation was carried out on the remaining objects in the dataset. For small classes we made sure to retain at least one object in the test set. As a result, for the class *scissors* the number of training objects is limited to 4. The evaluation for each object count was repeated 30 times with random splits into training and test set. It is not surprising to see the performance rise as the number of objects is increased—again, one cannot expect generalization of such variable categories to happen from only one exemplar.

Fig. 7 shows the classification performance with respect to the number of views per object. The evaluation procedure remained the same with the difference of selecting 6 train-

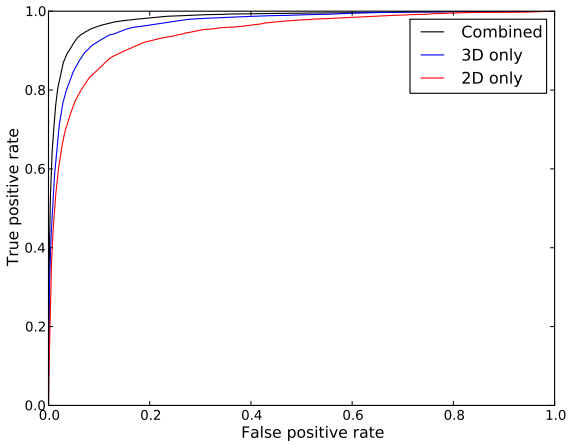


Figure 5: ROC curves for both modalities separately and in the combined case.

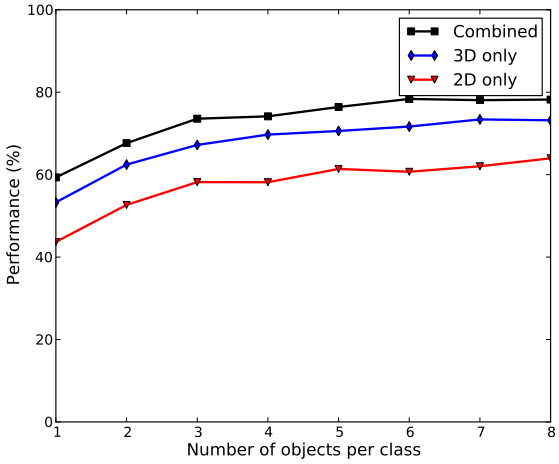


Figure 6: Classification results for number of training objects. 18 views per object.

ing objects for each category and altering the number of views per object. We see that in contrast to object count the number of views seems to plays a minor role, as long as a minimum amount of object orientations is covered. With 3-4 views—for 2D and 3D data combined—we achieve near optimal performance. Since views are equidistantly distributed, 3 views results in one view every 120°. Our data suggests that already a small number of views can be sufficient to obtain an adequate sampling of the object surface. In contrast, when using only 2D or 3D information, more views are needed to reach peak performance. This seems reasonable as less information is encoded in each view.

If we again look at the results for single classes, we see that the benefit of additional views depends strongly on the shape characteristics of the respective class. The comparison of Fig. 8 and Fig. 9 shows that classes of asymmetrical objects such as *binders* or *drink cartons* exhibit a much steeper increase in classification accuracy than classes of more round and symmetrically shaped objects such as *cups* and *bottles*. Interestingly, in the case of *drink cartons*, this increase is only visible in the 3D domain and the combined case — this is due to the fact, that the drink cartons as a category contain very different labels and therefore 2D appearance measures will have problems with a clear identification of the overall category, whereas the 3D shape is much better defined for this category.

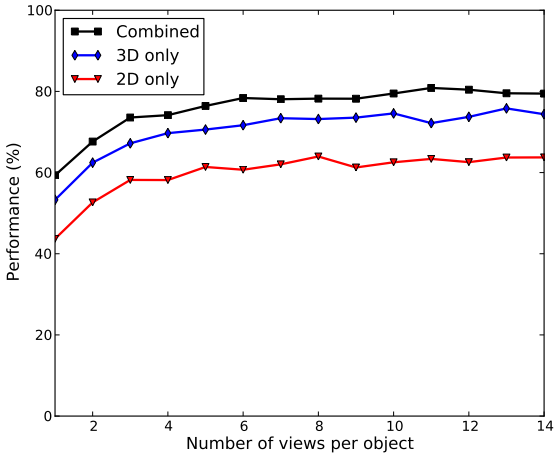
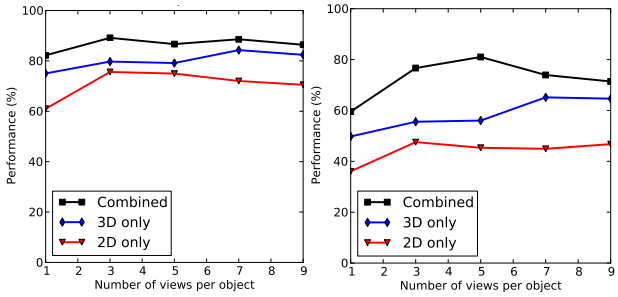


Figure 7: Classification results for number of training views per object. 6 objects per class.

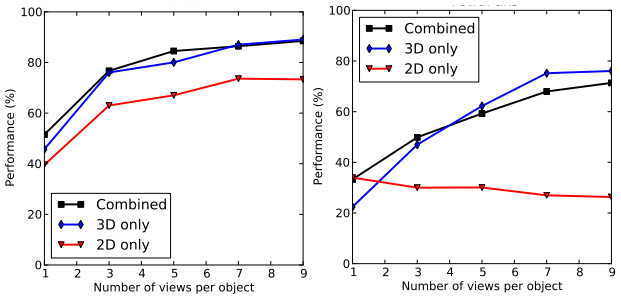


(a) Cups (b) Bottles

Figure 8: Classification rates for two symmetrical objects. Symmetical objects show low sensitivity to view count.

6. Conclusions

In this work we have introduced a new dataset for joint 2D and 3D object categorization containing data of real-world objects. The dataset was designed to allow for good generalization to real-world applications, which are based on Microsoft® Kinect™, for example. Using this dataset, we have demonstrated that the incorporation of range data is highly beneficial for object categoriza-



(a) Binders (b) Drink cartons

Figure 9: Classification rates for two asymmetrical objects. Asymmetrical objects benefit from additional views.

tion tasks. The fact that combination of multiple sensory inputs leads to increased recognition performance is not surprising—however, the performance gains are sometimes substantial.

Multi-modal object representations that integrate cues beyond visual information, such as haptic, proprioceptive or auditive cues might offer even more potential to capture the variety of features defining the objects in our environment. How to acquire and employ such rich representations has to be studied in future work. We believe that multi-sensory 3D object categorization is a research topic that will gain even more importance in the future as range sensors have become a consumer-level product. We hope that by supplying the research community with our data set, and by demonstrating the merits of sensor combination, more work in this domain will follow.

7. Acknowledgements

The authors gratefully acknowledge that this work was funded as part of the "Pool Projekt 2009-22" from the Centre for Integrative Neuroscience in Tübingen, Germany. Part of this research was also supported by the WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R31-2008-000-10008-0).

References

[1] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *CVIU*, 110(3):346–359, 2008. 3

[2] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *CVPR*, pages 1–8. IEEE, 2008. 3

[3] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *CIVR*, pages 401–408. ACM, 2007. 3

[4] B. Bustos, D. Keim, D. Saupe, T. Schreck, and D. Vranić. An experimental effectiveness comparison of methods for 3D similarity search. *IJDL*, 6(1):39–54, Feb. 2006. 2

[5] R. Campbell. A Survey Of Free-Form Object Representation and Recognition Techniques. *CVIU*, 81(2):166–210, Feb. 2001. 2

[6] M. Everingham, L. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, 88(2):303–338, Sept. 2009. 2

[7] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *CVIU*, 106(1):59–70, Apr. 2007. 2

[8] V. Ferrari, T. Tuytelaars, and L. Van Gool. Object detection by contour segment networks. *ECCV*, 3953(section 4):14–28, 2006. 2

[9] T. Funkhouser, P. Min, M. Kazhdan, J. Chen, A. Halderman, D. Dobkin, and D. Jacobs. A search engine for 3D models. *ACM Transactions on Graphics*, 22(1):83–105, Jan. 2003. 2

[10] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Mar. 2007. 2

[11] M. Heczko, D. Keim, and D. Saupe. Methods for similarity search on 3D databases. *Datenbank-Spektrum*, 2002. 3

[12] G. Hetzel, B. Leibe, P. Levi, and B. Schiele. 3D object recognition from range images using local feature histograms. In *CVPR*, volume 2, pages 394–399. IEEE, 2001. 2

[13] D. Huber, A. Kapuria, R. Donamukkala, and M. Hebert. Parts-based 3d object classification. In *CVPR*, volume 2. IEEE, 2004. 2

[14] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *PAMI*, 21(5):433, 1999. 2

[15] J. Koenderink and A. van Doorn. Surface shape and curvature scales. *Image and vision computing*, 10(8):557–564, 1992. 3

[16] M. Körtgen, G. Park, M. Novotni, and R. Klein. 3D shape matching with 3D shape contexts. In *CESCG*, volume 3. Citeseer, 2003. 3

[17] K. Lai, L. Bo, X. Ren, and D. Fox. A Large-Scale Hierarchical Multi-View RGB-D Object Dataset. In *ICRA*, 2011. 2

[18] A. Opelt and M. Fussenegger. Weak hypotheses and boosting for generic object detection and recognition. *ECCV*, 2004. 2

[19] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin. Shape distributions. *ACM Transactions on Graphics*, 21(4):807–832, Oct. 2002. 3

[20] U. Reiser, C. Connette, J. Fischer, J. Kubacki, A. Bubeck, F. Weisshardt, T. Jacobs, C. Parlitz, M. Hägele, and A. Verl. Care-O-bot 3: creating a product vision for service robot applications by integrating design and technology. In *IROS*, pages 1992–1998. IEEE Press, 2009. 2

[21] S. Ruiz-correa, L. G. Shapiro, and M. Meil. A new paradigm for recognizing 3-D objects from range data. *ICCV*, pages 1126–1133 vol.2, 2003. 2

[22] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *CVPR*, pages 1–8, 2007. 3

[23] M. Sun, G. Bradski, B.-X. Xu, and S. Savarese. Depth-encoded hough voting for coherent object detection, pose estimation, and shape recovery. In *ECCV*, 2010. 2