Evaluation of Approaches Combining 2D and 3D Data for Object Recognition Developed for the Mobile Robot Lisa

Viktor Seib, Florian Polster, Dietrich Paulus
Active Vision Group (AGAS), University of Koblenz and Landau
Universitätsstraße 1, 56070 Koblenz, Germany
{vseib, fpolster, paulus}@uni-koblenz.de
http://agas.uni-koblenz.de

ABSTRACT

The image data that object recognition systems are designed for changes over time. As soon as a new imaging technology is developed or becomes affordable new algorithms are inspired or known algorithms are adapted. Thus, different object recognition algorithms were developed and used on our mobile robot Lisa. In this work we compare the different approaches and investigate how they can be combined to best use 2D and 3D data. The individual approaches as well as their combinations will be introduced. Evaluation is performed on a large public dataset and a dataset acquired during the RoboCup competition.

Keywords: Object Recognition, Implicit Shape Models, RGB-D data, Multi-modal object recognition

1. INTRODUCTION

Affordable RGB-D cameras provide fused color and depth data, which serves as inspiration for new algorithms. Well established features used in literature such as SIFT [1], SURF [2] and HOG [3] that work well on textured objects have been joined by new descriptors capable of representing an object's shape (e.g. FPFH [4], SHOT [5]). Additionally, fused data from RGB-D cameras led new features specifically designed to simultaneously represent shape and texture [6], [7]. Object recognition on a mobile robot is an active research topic in our group. Our service robot Lisa is the 2015's world champion in RoboCup@Home [8]. We developed several object recognition algorithms in recent years that all were tested on Lisa. The individual algorithms were motivated by different sensor setups, but also by the changing demands on object recognition during the RoboCup@Home competition. While many objects possessed a feature rich structure in the past, in recent years featureless objects are becoming standard. Approaches solely based on RGB data allow to recognize object instances (e.g. *John's mug* vs. *Mary's mug*). However, these approaches fail when applied to objects without texture. Combining 2D and 3D information not only allows to recognize previously learned object instances, but also provides information about geometric shapes of objects. Thus, even unknown object are classified as belonging to a certain class (e.g. *mug*, *can*).

In the presented work we evaluate the pool of different object recognition approaches that were developed for Lisa in recent years. Special focus is put on the combination of 2D and 3D data to improve object recognition. Sec. 2 introduces the different approaches that were developed for our robot. In the following Sec. 3 methods to combine the recognition algorithms will be presented. The datasets, performed experiments and obtained results are presented in Sec. 4. Finally, the paper is concluded in Sec. 5.

2. OVERVIEW OVER DIFFERENT ALGORITHMS

This sections provides a brief overview over the different object recognition approaches used so far on our robot Lisa.

Further author information: send correspondence to Viktor Seib, vseib@uni-koblenz.de

2.1 Hough Feature Clustering

The Hough Feature Clustering (HFC) is a visual classifier (i.e. employs 2D data) and was previously presented in [9]. HFC uses SURF features [2] and the generalized Hough-transform [10]. After extracting features of an input image, correspondences are established with features from the training phase. In a subsequent step correspondences are clustered in Hough-space. By selecting maxima in Hough-space most erroneous feature correspondences are discarded. A homography is used to filter out even more unlikely correspondences. This approach is similar to the Implicit Shape Model (ISM) formulation by Leibe et al. [11]. The HFC implementation used here differs in some aspects from the previous algorithm presented in [9]. The Hough-space was reduced to 2 dimensions (previously 4) and can now be used as discrete or continuous Hough-space. The confidence calculation for the presence of an object was refined and by additionally using color histograms the algorithms is able to disambiguate objects with a small number of features. Further, the current implementation is capable of multi instance recognition and is parallelized to run on multiple CPU cores.

2.2 Bag of Shape Features

The Bag of Shape Features (BSF) is a classifier based on geometric features extracted from point clouds (i.e. employs 3D data) [12]. During training a regular three-dimensional grid is superimposed on the object. SHOT features [5] are extracted from the grid cell centers and stored as the object's representation. In contrast to bag of words approaches [13], [14] we do not perform vector quantization, but store each extracted feature. To classify an object based on its shape, SHOT features are extracted on the same grid as was used in training. The extracted features are matched with the learned descriptors. Matching an input feature with k nearest learned features has been shown to work well in the Implicit Shape Model (ISM) formulation (referred to as *activation strategy*). We employ this method in our shape classification. Similar to bag of words approaches we build a histogram based on the extracted features of the test shape. However, we do not use codewords as bins, but object instance names that correspond to the matched k nearest features. The object instance with the most votes is selected as the final recognition result.

2.3 Implicit Shape Models

Recently, we adapted the Implicit Shape Model (ISM) formulation of Leibe et al. [11] to 3D data [15]. We have shown in [15] that our approach is able to achieve higher classification rates than similar methods (e.g. [16], [17]). In our approach we extract SHOT features [5] on a dense grid to represent individual objects. Unlike approaches in related work, we do not construct a dictionary of codewords, but rather use the features as they are to achieve higher discrimination. For classification features are extracted on the query object and the votes are matched with the trained object descriptions. Since no codebook is created, generalization from learned shapes is achieved by a k-NN activation strategy during classification (instead of training as in many approaches). Finally, object locations are acquired by analyzing the voting space for maxima using Mean-shift mode estimation [18]. A key difference between our ISM formulation and other ISM adaptations to 3D data is that we use a continuous voting space. Our experiments show that a continuous voting space is beneficial for correct object classification.

3. COMBINING APPROACHES

In this section we shortly introduce and discuss one previous method that we used to combine 2D and 3D data. Further, we introduce two different approaches to combine these data that will be evaluated in the next section.

3.1 Point Feature Histograms and Hough Clustering

We presented a method to combine 2D and 3D data in [19]. In this early approach the 2D data was handled by the HFC algorithm as described above and introduces in [9]. However, the 3D data was learned in 2 separate steps. First, FPFH features [4] were used to estimate the local geometry around each point and a conditional random field (CRF) was trained to classify local structures. In a second step a global geometric model was obtained from the CRF classification to train a SVM model to represent object classes. This approach had several limitations such as the need to train 2 different classifiers (CRF and SVM) and the limited number of different classes the SVM model could handle. Further, training and recognition times were too long to be used on a mobile robot platform. For these reasons we decided to investigate other approaches to combine 2D and

(a) RGBD-Washington dataset (b) RoboCup Hefei dataset

Figure 1. Example objects from the datasets used for evaluation

3D data. This early approach will not be evaluated in the following as its design prohibits its application to the used datasets (mainly because of the inherently fixed number of object classes this approach can handle).

3.2 Combined Feature Descriptor

Several feature descriptors to simultaneously describe shape and texture of an object were proposed recently [6], [7]. Since we already have obtained good results in the past with using the SHOT descriptor we decided to use CSHOT (color SHOT) [6] for our experiments. Basically, the CSHOT descriptor is formed by concatenating the known SHOT descriptor with a newly introduced descriptor for colored textures. Before obtaining the descriptor a local reference frame is computed for each interest point in the same manner as is done for the SHOT descriptor.

3.3 Rule-based Combination of Classifiers

Intuitively, a shape classifier is better suited for class recognitions since objects of different classes usually have different shapes. On the other hand, a visual classifier is better suited to tell apart instances of the same class since instances usually have similar forms, but different textures. We combine two classifiers as described in the following. A shape classifier determines the candidate classes and a visual classifier chooses the best matching instances from these classes. Although this combination of classifiers is hand-crafted, it has already shown its strength in the past [12].

4. EVALUATION

This section describes the combinations of approaches and the datasets that were used for evaluation. Further, it presents and discusses the obtained results.

4.1 Datasets

RGBD Washington: The RGBD Washington dataset was introduced in [20]. This dataset comprises 300 objects in 51 categories (RGB and depth data). Examples are shown in Fig. 1(a). The data in the set was recorded from 3 different camera angles: 30 °, 45 ° and 60 °. For training the 30 ° angle is used, while the other angles are used for testing. To have a uniform distribution of instances per objects, we chose the first 3 objects from each category (153 objects). Training was performed with 15 different object views, while 40 views were used for testing.

RoboCup Hefei: The RoboCup Hefei dataset was created on the RoboCup world championship in Hefei [8]. It consists of 29 objects in 5 different categories with 20 to 62 object views per object (RGB and depth data). Example objects are shown in Fig. 1(b). The objects are divided into a training set and a recognition set. The training set consists of 17 views for each object, while the remaining views form the test set.

Table 1. Recognition results for the RGBD-Washington dataset

	HFC	BSF-SHOT	ISM	BSF-CSHOT	BSF-SHOT + HFC	BSF-CSHOT + HFC
true positive rate	8.0 %	35.1 %	45.0 %	42.1 %	27.2 %	31.6 %
false positive rate	47.7 %	292.3 %	788.9%	228.1 %	64.3 %	61.2 %
false negative rate	44.3 %	0.0 %	0.0%	0.0 %	8.4 %	7.2 %
precision	0.14	0.11	0.05	0.16	0.30	0.34
recall	0.15	1	1	1	0.76	0.81
average time per model	3.8 s	0.4 s	1.0 s	0.4 s	0.7 s	0.7 s

Table 2. Recognition results for the RoboCup Hefei dataset

	HFC	BSF-SHOT	ISM	BSF-CSHOT	BSF-SHOT + HFC	BSF-CSHOT + HFC
true positive rate	48.6 %	80.4 %	75.7 %	-	77.5 %	-
false positive rate	2.8 %	88.8 %	18.7 %	-	16.8 %	-
false negative rate	48.6 %	0.0 %	0.0 %	-	5.6 %	-
precision	0.94	0.48	0.80	-	0.82	-
recall	0.50	1	1	-	0.93	-
average time per model	2.79 s	0.4 s	0.8 s	-	0.6 s	-

4.2 Experiments and Results

The aforementioned datasets will be tested with 3 algorithms using only 2D or only 3D data (unimodal). These results will then be compared with 3 algorithms combinations using 2D and 3D data at the same time (multi-modal). The unimodal algorithms will be the ones presented in Sec. 2: HFC for 2D data and BSF and ISM with the SHOT feature for 3D data. The multi-modal algorithms will be the combinations presented in Sec. 3 applied to the unimodal algorithms from Sec. 2. We test the combined descriptor CSHOT with the BSF algorithm. Further, the rule based approach will be tested by combining the BSF (with SHOT) and HFC. Finally, we will also evaluate a rule based approach with a combined descriptor: BSF with CSHOT combined with HFC.

The reported results include all recognized objects, even if multiple object were recognized on the same spot. This evaluation method is the most general one and does not take any assumptions on the input data or object segmentation. However, this also means that the false positive rate can go beyond 100 % simply because one single object can be recognized as multiple different objects. The evaluation results are presented in Tab. 1 and in Tab. 2. Please note that in Tab. 2 no results for algorithms using the CSHOT descriptor could be reported, because no registered RGB and point cloud data was available for this dataset.

We observe that HFC by itself is well suited for object detection on feature-rich objects as were present in the RoboCup Hefei dataset (Tab. 2). However, with very similar looking objects or object with only a small number of features the performance decreases (Tab. 1). The unimodal approaches using 3D data only (BSF-SHOT and ISM) also perform better on the RoboCup Hefei dataset. The reason here is again that the Washington dataset has lots of similar looking and similar shaped objects that are hard to tell apart. This also leads to an enormous number of false positive object recognitions if only the 3D information is considered. A direct comparison between BSF-SHOT and BSF-CSHOT shows that the additional color information in the descriptor is indeed helpful. The true positive rate increases while at the same time less false positives are found. Still, the false positive rate stays on a very high level. The rule-based combination BSF-SHOT + HFC has similar results on both datasets: the true positive rate is almost as good as in the BSF-SHOT algorithm if used individually, while at the same time the false positive rate is strongly reduced. A direct comparison between BSF-SHOT + HFC and BSF-CSHOT + HFC again shows that the additional color information in the descriptor further improved the results.

5. SUMMARY AND CONCLUSION

In this work we have given an overview over different approaches for object recognition that were developed for our mobile robot platform Lisa. These algorithms use 2D and 3D data and can be combined in a meaningful way to improve recognition results. Our evaluation was performed without any assumptions on the number of objects or object segmentation to represent the true performance of the algorithms. In practice these results can be improved: in many cases segmented RGBD object data can be obtained and therefore only the most probable recognition result per object considered. However, this was not the focus of the present work. Instead we wanted

to present different means of 2D and 3D data combinations to show its benefits on the true algorithm performance. In future we plan to investigate further means of data combinations, e.g. by fusing the Hough-voting spaces of the HFC and ISM algorithm.

REFERENCES

[1] Lowe, D. G., "Distinctive image features from scale-invariant keypoints," *Int. Journal of Computer Vision* **60**(2), 91–110 (2004).

[2] Bay, H., Tuytelaars, T., and Van Gool, L., "Surf: Speeded up robust features," in [*Computer Vision–ECCV 2006*], 404–417, Springer (2006).

[3] Dalal, N. and Triggs, B., "Histograms of oriented gradients for human detection," in [*Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conf. on*], **1**, 886–893, IEEE (2005).

[4] Rusu, R. B., Blodow, N., and Beetz, M., "Fast point feature histograms (fpfh) for 3d registration," in [*Robotics and Automation, 2009. ICRA'09. IEEE Int. Conf. on*], 3212–3217, IEEE (2009).

[5] Tombari, F., Salti, S., and Di Stefano, L., "Unique signatures of histograms for local surface description," in [*Proc. of the European conference on computer vision (ECCV)*], *ECCV'10*, 356–369, Springer-Verlag, Berlin, Heidelberg (2010).

[6] Tombari, F., Salti, S., and Di Stefano, L., "A combined texture-shape descriptor for enhanced 3d feature matching," in [*Image Processing (ICIP), 2011 18th IEEE Int. Conf. on*], 809–812, IEEE (2011).

[7] Zaharescu, A., Boyer, E., Varanasi, K., and Horaud, R., "Surface feature detection and description with applications to mesh matching," in [*Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conf. on*], 373–380, IEEE (2009).

[8] Seib, V., Memmesheimer, R., Manthe, S., Polster, F., and Paulus, D., "Team homer@unikoblenz: Approaches and contributions to the robocup@ home competition," *RoboCup 2015: Robot World Cup XIX* **9513**, 83–94 (2016).

[9] Seib, V., Kusenbach, M., Thierfelder, S., and Paulus, D., "Object recognition using hough-transform clustering of surf features," in [*Workshops on Electronical and Computer Engineering Subfields*], 169 – 176, Scientific Cooperations Publications (2014).

[10] Ballard, D. H., "Generalizing the hough transform to detect arbitrary shapes," *Pattern Recognition* **13**(2), 111–122 (1981).

[11] Leibe, B., Leonardis, A., and Schiele, B., "Combined object categorization and segmentation with an implicit shape model," in [*ECCV' 04 Workshop on Statistical Learning in Computer Vision*], 17–32 (2004).

[12] Seib, V., Memmesheimer, R., and Paulus, D., "Ensemble classifier for joint object instance and category recognition on rgb-d data," in [*Image Processing (ICIP), 2015 IEEE Int. Conf. on*], 143–147, IEEE (2015).

[13] Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C., "Visual categorization with bags of keypoints," in [*Workshop on statistical learning in computer vision, ECCV*], **1**(1-22), 1–2 (2004).

[14] Toldo, R., Castellani, U., and Fusiello, A., "A bag of words approach for 3d object categorization," in [*Computer Vision/Computer Graphics CollaborationTechniques*], 116–127, Springer (2009).

[15] Seib, V., Link, N., and Paulus, D., "Pose estimation and shape retrieval with hough voting in a continuous voting space," in [*Pattern Recognition*], Gall, J., Gehler, P., and Leibe, B., eds., *LNCS* **9358**, 458–469, Springer Int. Publishing (2015).

[16] Knopp, J., Prasad, M., Willems, G., Timofte, R., and Van Gool, L., "Hough transform and 3d surf for robust three dimensional classification," in [*ECCV (6)*], 589–602 (2010).

[17] Salti, S., Tombari, F., and Di Stefano, L., "On the use of implicit shape models for recognition of object categories in 3d data," in [*ACCV (3)*], *Lecture Notes in Computer Science*, 653–666 (2010).

[18] Cheng, Y., "Mean shift, mode seeking, and clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17**(8), 790–799 (1995).

[19] Seib, V., Christ-Friedmann, S., Thierfelder, S., and Paulus, D., "Object class and instance recognition on rgb-d data," in [*Sixth Int. Conf. on Machine Vision (ICMV 2013)*], Verikas, A., Vuksanovic, B., and Zhou, J., eds., 90670J–90670J–7 (2013).

[20] Lai, K., Bo, L., Ren, X., and Fox, D., "A large-scale hierarchical multi-view rgb-d object dataset," in [*Robotics and Automation (ICRA), 2011 IEEE Int. Conf. on*], 1817–1824, IEEE (2011).