ENSEMBLE CLASSIFIER FOR JOINT OBJECT INSTANCE AND CATEGORY RECOGNITION ON RGB-D DATA

Viktor Seib, Raphael Memmesheimer, Dietrich Paulus

Active Vision Group (AGAS), University of Koblenz and Landau
Universitätsstraße 1, 56070 Koblenz, Germany
{vseib, raphael, paulus}@uni-koblenz.de

ABSTRACT

Sensors for RGB-D data have gained high popularity in the computer vision community. We present an efficient ensemble classifier that combines visual and depth data and achieves higher recognition rates than the individual classifiers or a classifier exploiting visual and depth data at the same time. The presented approach was evaluated in practice on a mobile robot during the RoCKIn robotics challenge in 2014.

***Index Terms*—** category and instance recognition, ensemble classifier, mobile robots

1. INTRODUCTION

Looking through the "eyes" of image capturing sensors the world has been colorful for many years. Recently, with affordable RGB-D cameras available for anyone it has become three dimensional. The pool of well established features used in literature such as Scale Invariant Feature Transform (SIFT, [1]), Speeded-up Robust Features (SURF, [2]) and Histograms of Oriented Gradients (HOG, [3]) that work well on textured objects has been extended by new descriptors capable of representing an object's shape. Among others, spin images (SI, [4]), fast point feature histograms (FPFH, [5]) and unique signatures of histograms (SHOT, [6]) allow for reliable 3D feature matching. Further, combinations of established features to describe texture and shape were proposed [7], as well as features specifically designed to represent shape and texture [8, 9]. This development in imaging techniques and feature descriptors has led to significant advances in 3D object recognition, shape retrieval and object manipulation.

To this day, matching local feature descriptors is still the most common methodology for object classification and related tasks. In this work we present an approach of joint object category and instance recognition. The presented ensemble classifier in Sec. 2 consists of 2 separate classifiers, one for visual data and a second for depth data. The results of both classifiers are combined to enhance recognition results. The proposed classifiers are evaluated on a challenging dataset and results are presented in Sec. 3. We compare the ensemble classifier with a classifier using features that simultaneously capture the object's shape and texture.

The advantage of separate classifiers is the ability to adapt the algorithm to the sensors available. Thus, the approach can be used with traditional RGB as well as *Kinect*-like RGB-D cameras. The presented classifiers enables us to determine the instance and class of an object in a short time. It is well suited to be applied on a mobile robot along all other components needed in an autonomous robotic system. We evaluated the presented approach on a mobile robot. Our team *homer@UniKoblenz* won the first place in the overall rating and the second place in the object perception benchmark at the RoCKIn[1] robotics competition 2014.

2. ENSEMBLE CLASSIFIER FOR RGB-D DATA

Given a set of object classes C, where each class $c \in C$ has several instances $i_c \in I$ we want to find the correct instance and class to a given test image s_t. Each instance i is trained from a set of sample views $s_i \in S$. From training we further know the mapping function $\mathcal{I} : S \rightarrow I$ that assigns an instance to each sample view and the function $\mathcal{C} : I \rightarrow C$ that maps each instance to a class.

2.1. Visual Classifier

The visual classifier is based on our previous work presented in [10]. We briefly summarize its concept to provide the reader with the most important information. The visual classifier uses SURF features and the generalized Hough-transform [11]. After extracting the features of an input image (containing a single segmented object or a scene without prior information), correspondences are established with features from the training phase. In a subsequent step correspondences are clustered in a four-dimensional Hough-space, representing a 2D position, scale and rotation. By selecting maxima in Hough-space most erroneous feature correspondences are discarded. This approach is similar to the Implicit Shape Model (ISM) formulation by Leibe et al. [12]. The most important

[1] RoCKIn website: `http://rockinrobotchallenge.eu`

differences are that we do not cluster the extracted features and that matching is not performed between the *k* most similar features. Rather, we only establish a correspondence if the distance ratio between the best and second best matching feature is below a certain threshold as suggested by Lowe et al. [1]. This prevents establishing random matches and yields more stable correspondences. In a final step the features from Hough-maxima are used to calculate a homography to project the outline of the object hypotheses into the query image. A hypothesis is confirmed if the ratio between the number of matching features and total number of features within the outline is high enough. The algorithms yields a list of detected and recognizes objects and their bounding boxes in the input image. This algorithm is robust towards clutter and enables us in detecting largely occluded objects. It is included in the visual classifier and extended as described below.

Feature descriptors like SURF perform well on feature-rich objects and are well suited for instance recognition. However, in case of homogeneous object textures or, likewise, low camera resolutions, only a small number of features can be extracted. This leads to small number of matches and thus to unreliable recognitions or no recognitions at all. In such cases the resulting object list may contain several objects with low detection confidence.

To disambiguate similar results and strengthen the classifier for objects with homogeneous textures we use color signatures. If depth data is available, objects are segmented prior to recognition. The signature is calculated from the color data of the segmented object. Without depth data signatures can be calculated from the bounding box of the object determined after homography calculation. The color signature is calculated from the HSV color space with 8 bins from hue and 3 bins from saturation and value with individual bin sizes.

In case of several hypotheses inside a segmented object mask the histogram intersection kernel [13] is used to determine the similarity between a learned sample view h and a query image g. We evaluate several strategies. From the first part of the visual classifier we obtain an instance hypotheses list $H_i \subseteq I$ and a class hypotheses list $H_c \subseteq C$. To find the most likely instance we obtain a set $S_{H_i} = \{s | \mathcal{I}(s) \in H_i\}$ containing all sample views from training belonging to all instances from the instance hypotheses list. We also obtain $S_{H_c} = \{s | \mathcal{C}(\mathcal{I}(s)) \in H_c\}$, a set containing all sample views of all classes on the hypotheses list. The strategy in Eq. 1 decides in favor of the object instance with the *highest* color signature similarity to the query image g

$$\underset{x}{\operatorname{argmax}}\, \mathrm{f}(h_x, g) = \sum_{i}^{b} \min(\mathcal{S}(h_x)_i, \mathcal{S}(g)_i), \forall h_x \in S_H. \tag{1}$$

Here, $\mathcal{S}(h)$ is a function to determine the color signature with b bins of an image h. After obtaining x the classifiers returns the resulting instance $\mathcal{I}(h_x)$. The strategy in Eq. 2 decides in favor of the object instance that's *average* signature similarity is highest to the query image

$$\underset{x}{\operatorname{argmax}}\, \mathrm{f}(h_x, g) = \frac{1}{m_x} \sum_{j}^{m_x} \sum_{i}^{b} \min(\mathcal{S}(h_x)_{ij}, \mathcal{S}(g)_i), \quad \forall h_x \in S_H \tag{2}$$

where m_x is the number of sample views belonging to each instance $\mathcal{I}(h_x)$. In total, we test four strategies by setting S_H in Eq. 1 and Eq. 2 to S_{H_i} and S_{H_c}, respectively.

2.2. Shape Classifier

The second classifier uses depth data only. In contrast to the visual classifier that accepts single object images as well as scene images for classification, the shape classifier works only with segmented objects. During training a regular three-dimensional grid is superimposed on the object. SHOT features [6] are extracted from the grid cell centers and stored as the object's representation. We use a grid size and SHOT radius of 0.1 m. In contrast to bag of words approaches [14, 15] we do not perform vector quantization, but store each extracted feature. Vector quantization is a common technique that works well for 2D recognition tasks and is also applied for 3D data. However, recent experiments [16, 17] show that clustering 3D feature descriptors leads to a considerable loss in their descriptiveness and worse recognition results.

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. Thus, each feature l from the test shape votes for the k nearest features in the learned dictionary $\mathcal{F}$ as indicated by Eq. 3

$$\mathcal{F}_k = \{f_j \in \mathcal{F} \,|\, j = 1, 2, \ldots, k \wedge \mathrm{d}(l, f_j) \leq \mathrm{d}(l, f_{j+1})\}, \tag{3}$$

with the distance function $\mathrm{d}(f_1, f_2) = \|f_1 - f_2\|_2$.

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. We evaluate this classifier with $k \in \{1, 2, 3, 4\}$.

2.3. Combined Classifier

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

Fig. 1. Example objects from the evaluation dataset [7]. Each object is a different instance from the dataset.

Table 1. Classification results of the presented classifiers on the partial RGB-D dataset [7]

joint class and instance classification				
	visual	shape	combined	CSHOT
instance	39.2	26.9	42.1	36.4
class	46.4	41.2	55.9	51.0
time	4.7 s			0.4 s
unknown instance classification				
	visual	shape	combined	CSHOT
class	23.5	53.7	54.3	47.0
time	4.6 s			0.4 s

both classifiers as described in the following. The shape classifier determines the candidate classes and the visual classifier chooses the best matching instances from theses classes. Four different versions of this strategy are evaluated: taking only the best, respectively the three best classes from the activation with $k = 2$ and $k = 4$. If the category of an *unknown* instance needs to be found, the visual classifier only considers classes that are most similar to learned instance views.

This combination of classifiers is hand-crafted. To better assess the quality of this rule bases combination of classifiers, we compare it with a classifier that uses CSHOT [8] features. The CSHOT feature directly combines visual and shape cues and does not rely on rules to combine individual classifiers. Apart from the feature descriptor used, the CSHOT classifier works exactly as the shape classifier described above.

3. EXPERIMENTS AND RESULTS

We evaluate our classifiers on the RGB-D dataset presented in [7]. This dataset comprises 300 objects in 51 categories. From each object 3 complete 360° image sequences were recorded by RGB-D cameras mounted at elevations of 30°, 45° and 60° degrees at a distance of 1 m from the objects. The dataset includes a broad variety of object classes, ranging from easy distinguishable food boxes to fruits and vegetables that even humans have difficulties to tell apart (see Fig. 1). The number of instances per category varies between 3 and 14. To have a uniform distribution of instances per category, we choose the first 3 instances from each class (153 objects). Further, to decrease training and classification time we reduce the number of training images and point clouds to one third of the data used in [7] (taking each 15th sample view).

We perform 2 types of experiments: joint instance and category classification, as well as classification of unknown instances. For joint instance and category classification we use 2 of the image sequences for training and the third for testing. For unknown instance classification we use 2 of 3 instances of each category for training and the remaining unknown instance for testing. Each of these tests is performed 3 times, once for each image sequence and instance, respectively, that is left out in training. The results are averaged and reported in Table 1. In total, close to 7000 sample views need to be classified in each of these tests.

Table 1 shows that using an ensemble classifier that exploits both, visual and shape data, improves classification results in both types of experiments. In joint class and instance classification the combined classifier gains 2.9% for instance and 9.5% for class recognition, if compared to the best individual classifier. However, in case of the unknown instance classification the performance gain when combining the classifiers is only 0.6%. This is explained by the lack of expressiveness of visual data of unknown instances.

Both individual classifiers show different characteristic behaviors. While the visual classifier determines 46.4% of all classes correctly in the joint class and instance classification, the rate drops to 23.5% for unknown instance classification. Contrary to the visual classifier, the shape classifier shows its full strength when applied to unknown instances. Compared to the joint class and instance classification it gains 12.5% in classifying unknown instances. This is explained by the design of the experiments: in the first evaluation each object is learned from 2 image sequences captured with different camera elevations. However, in the second test all 3 image sequences are used, thus better capturing the shape appearance of the whole object.

The CSHOT-classifier uses features that combine visual and shape information. However, Table 1 indicates that the proposed classifier outperforms the CSHOT-classifier in both experimental setups. For joint class and instance recognition the correct classification rate of the proposed classifier is 5.7% (instances) and 4.9% (classes) higher than with the CSHOT-classifier. In the classification of unknown instances the proposed classifier outperforms the CSHOT classifier by 7.3%.

Our goal in designing the presented classifiers was to obtain a good trade-off between classification results and runtime to use these approaches on a mobile robot. Efficient clas-

Table 2. Comparisons of strategies used for joint instance and class recognition

joint class and instance classification				
visual classifier				
strategy	average instance	average class	best instance	best class
instance	31.7	31.6	39.2	38.2
class	33.9	39.9	42.4	46.4
shape classifier				
strategy	act. k=1	act. k=2	act. k=3	act. k=4
instance	25.0	26.5	26.9	26.9
class	37.8	40.5	41.1	41.2
combined classifier				
strategy	n=1, k=2	n=1, k=4	n=3, k=1	n=3, k=4
instance	42.0	42.1	33.2	30.9
class	55.8	55.9	44.6	41.2
CSHOT classifier				
strategy	act. k=1	act. k=2	act. k=3	act. k=4
instance	34.6	36.3	36.4	36.3
class	48.5	50.6	51.0	50.9

Table 3. Comparisons of strategies used for unknown instance classification

unknown instance classification				
visual classifier				
strategy	average instance	average class	best instance	best class
class	7.2	7.5	23.1	23.5
shape classifier				
strategy	act. k=1	act. k=2	act. k=3	act. k=4
class	51.6	53.3	53.7	53.4
combined classifier				
strategy	n=1, k=2	n=1, k=4	n=3, k=1	n=3, k=4
class	42.6	42.1	54.3	50.4
CSHOT classifier				
strategy	act. k=1	act. k=2	act. k=3	act. k=4
class	45.9	47.0	47.1	47.1

sification is crucial for mobile robot systems since numerous components are executed in parallel, leaving only little computational power for each individual component. All experiments were performed on a notebook with an Intel Core i7 processor and 12 GB RAM. Table 1 shows that the CSHOT classifier needs 0.4 s to determine the class and instance of an object, while the combined classifier takes 4.6 s. The reason for the high runtime in the latter case is the visual classifier that is not sufficiently parallelized.

The presented classification rates are below the state of the art on this dataset. Lai et al. [7] achieve a rate of 74.8% for instance classification and 83.3% in classifying unknown instances. Bo et al. [18] report even 92.8% (instance classification) and 87.5% (unknown instances). Both approaches achieve these results on the whole dataset (300 objects), while we use 153 objects. However, to minimize training and recognition time we use only one third of the training data per object as opposed to the state of the art. Further, due to the reduced dataset we only have 2 training objects per category to recognize unknown instances. In state of the art $n-1$ objects are used where n is the number of objects per category and varies between 3 and 14.

Tables 2 and 3 present a comparison of different strategies evaluated in the design of the classifiers. The visual classifier was evaluated with different disambiguation strategies as introduced in Eq. 1 (best) and Eq. 2 (average). In both equations sample views S_{H_i} (instances) and S_{H_c} (classes) were tested. The strategies that decide in favor of the single best matching instance perform better in both experiments than strategies choosing the average best matching instance. This emphasizes the strength of the visual classifier in matching similar appearances of objects. The average strategies, however, suffer from very distinct appearances of the same object instance in different sample views. The shape and the CSHOT classifiers were tested with different values for k for feature matching. Matching only the closest feature, as well as too many, leads to numerous erroneous correspondences. In our experiments the best value was $k = 3$. Finally, the combined classifier takes the best $n \in \{1, 3\}$ classes from the shape classifier and activation with $k \in \{2, 4\}$ and the visual classifier disambiguates using the *best class* strategy. For joint instance and category recognition the performance mostly depends on the choice of n: both strategies with $n = 1$ were superior to $n = 3$, while the choice of k had almost no influence. These results are reversed in classifying unknown instances. In this case it is favorable to choose higher n and lower k.

4. CONCLUSION

We presented an ensemble classifier that benefits from combining 2 individual classifiers that use visual and depth data, respectively. The combination leads to higher recognition rates for joint instance and category recognition, as well as for classification of unknown instances. Further, the proposed ensemble classifier achieves higher classification rates than a single classifier that simultaneously exploits visual and shape features. We do not outperform the state of the art on the evaluated dataset, though. However, our classifiers are efficient and were tested to perform well on a mobile robot during the RoCKIn robotics competition where we won the second place in the object perception benchmark. Our future work will concentrate on evaluating different strategies for classifier combination and further experiments with classifiers that simultaneously use shape and visual data.

5. REFERENCES

[1] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," in *Computer Vision–ECCV 2006*, pp. 404–417. Springer, 2006.

[3] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 886–893.

[4] Andrew E. Johnson and Martial Hebert, "Using spin images for efficient object recognition in cluttered 3d scenes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, no. 5, pp. 433–449, 1999.

[5] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. IEEE, 2009, pp. 3212–3217.

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

[7] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1817–1824.

[8] Federico Tombari, Samuele Salti, and Luigi Di Stefano, "A combined texture-shape descriptor for enhanced 3d feature matching," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011, pp. 809–812.

[9] Andrei Zaharescu, Edmond Boyer, Kiran Varanasi, and Radu Horaud, "Surface feature detection and description with applications to mesh matching," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 373–380.

[10] Viktor Seib, Michael Kusenbach, Susanne Thierfelder, and Dietrich Paulus, "Object recognition using hough-transform clustering of surf features," in *Workshops on Electronical and Computer Engineering Subfields*. 2014, pp. 169 – 176, Scientific Cooperations Publications.

[11] D. H. Ballard, "Generalizing the hough transform to detect arbitrary shapes," *Pattern Recognition*, vol. 13, no. 2, pp. 111–122.

[12] Bastian Leibe, Ales Leonardis, and Bernt Schiele, "Combined object categorization and segmentation with an implicit shape model," in *ECCV' 04 Workshop on Statistical Learning in Computer Vision*, 2004, pp. 17–32.

[13] Annalisa Barla, Francesca Odone, and Alessandro Verri, "Histogram intersection kernel for image classification," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*. IEEE, vol. 3, pp. III–513.

[14] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, 2004, vol. 1, pp. 1–2.

[15] Roberto Toldo, Umberto Castellani, and Andrea Fusiello, "A bag of words approach for 3d object categorization," in *Computer Vision/Computer Graphics CollaborationTechniques*, pp. 116–127. Springer, 2009.

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

[17] Viktor Seib, Norman Link, and Dietrich Paulus, "Implicit shape models for 3d shape classification with a continuous voting space," in *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2015, to appear.

[18] Liefeng Bo, Xiaofeng Ren, and Dieter Fox, "Unsupervised feature learning for rgb-d based object recognition," in *Experimental Robotics*. Springer, 2013, pp. 387–402.