2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Extracting Essential Local Object Characteristics for 3D Object Categorization

Marianna Madry   Heydar Maboudi Afkham   Carl Henrik Ek   Stefan Carlsson   Danica Kragic

***Abstract*— Most object classes share a considerable amount of local appearance and often only a small number of features are discriminative. The traditional approach to represent an object is based on a summarization of the local characteristics by counting the number of feature occurrences. In this paper we propose the use of a recently developed technique for summarizations that, rather than looking into the quantity of features, encodes their *quality* to learn a description of an object. Our approach is based on extracting and aggregating only the *essential* characteristics of an object class for a task. We show how the proposed method significantly improves on previous work in 3D object categorization. We discuss the benefits of the method in other scenarios such as robot grasping. We provide extensive quantitative and qualitative experiments comparing our approach to the state of the art to justify the described approach.**

I. Introduction

A meaningful representation should retain only information that is relevant for a specific task. This leads to the question: *What are the characteristics of an object that are essential for a task?* What makes it possible to grasp a pan and a knife in a similar way [7][6], what characteristics decide if an object affords drinking [9][20] and what makes a chair a chair [10]? These characteristics are often non-obvious, which is why they have been traditionally extracted by statistical supervised learning techniques.

Statistical learning is based on the assumption that it is possible to acquire a sufficient number of samples of the phenomenon to be modeled. However, in many scenarios this is not feasible due to the high-dimensionality of the data. A common approach to circumvent this is to look at information at a smaller scale where sufficient data can be acquired, such as a small neighborhood “patch”. This set of patches can then be *summarized* into a single representation [19], as shown in Figure 1 (top row). The traditional approach to represent a 3D object can be considered as a series of consecutive steps gradually increasing a level of summarization. First, a point cloud is extracted from sensory data, then local points are joined and summarized into a patch representation; these patches are finally summarized into an object representation. As such these summarizations can be seen on a continuum.

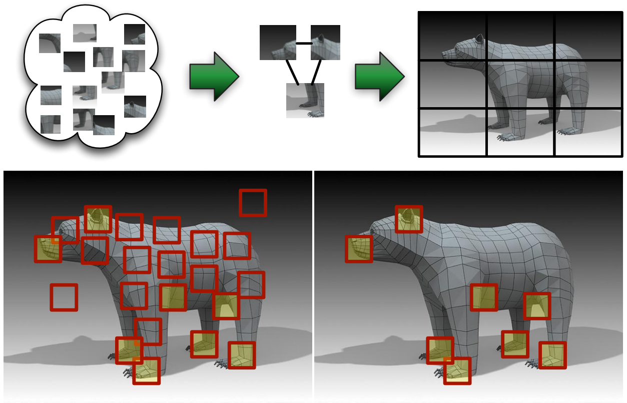


Fig. 1. We refer to the procedure of summarization as that of creating a single representation for a set of observations [19]. (Top) Illustration of the continuum of summarization steps that typically is a part of the pipeline of generating a single representation of an object for object classification. In the left a lagre number of very small patches are summarized while on the right side few large regions are summarized resulting in a single representation of the object. (Bottom) Illustration of the approach we take in this paper. The left image shows all the patches (red squares) extracted from the image with only a few being relevant for object description (yellow). Our descriptor creates a representation by only summarizing these *essential* patches.

Previously we have addressed the problem of summarizing local features to a global representation by incorporating object structure [16]. In this paper we will focus on summarization at a more local level and present an approach that extracts only the “essential” features needed to represent a specific object for a specific task, as explained in Figure 1 (bottom row). The intuition is that most of the local features of an object are irrelevant for most tasks and therefore using a summarization of all features will reduce the proportion of the variance that is relevant in the descriptor. For instance, take a cup and a can which share most of the same local geometry; in terms of object categorization it is the handle of the cup that is the essential characteristic while in terms of one’s ability to drink from the cup, it is the opening. Neither of these characteristics are dominant and are therefore likely to “disappear” in a representation constructed by summarizing all local features. To overcome these problems we will adapt a newly proposed summarization method which introduces the concept of a *Qualitative Summarization* [1]. The method facilitates supervision by creating a sparse interpretable feature space which extracts and summarizes only the essential characteristics of a class.

In this paper we will show that the applied methodology allows us to automatically discover essential and easily interpretable object characteristics. They are not only stable within each object category, but also specific to

M. Madry, H. Maboudi Afkham, C. H. Ek, S. Carlsson and D. Kragic are with the Centre for Autonomous Systems and the Computer Vision and Active Perception Lab, CSC, KTH Royal Institute of Technology, Stockholm, Sweden. Email: {madry,heydarma,chek,stefanc,danik}@csc.kth.se

This work was supported by the Swedish Foundation for Strategic Research, the Swedish Research Council (VR), the EU FP7 project TOMSY (IST-FP7-270436) and the EU FP7 project RoboHow.Cog (FP7-ICT-288533). We thank Golnaz Nozari for her help.

978-1-4673-6357-0/13/$31.00 ©2013 IEEE

it. In consequence, by encoding relevant object properties for a task, we significantly improve the categorization rate for real scenarios compared to the state of the art. Our approach opens doors for further enhancement of global representations based on the local features, such as the previously proposed Global Structure Histogram (GSH) [16]. Moreover, the information about a specific position of the essential features might in the future facilitate planning of robot actions required to verify an object class or a grasp hypothesis.

The remainder of the paper is organized as follows: Section II describes the related work that puts this paper in context. Section III explains the proposed approach. Section IV presents qualitative and quantitative results and Section V discusses other robotics application of the method. We conclude the paper and detail directions of future work in Section VI.

II. Related Work

Object representation methods aim to create a single object description from high dimensional sensory information. In order to facilitate efficient reasoning, it is desired to reduce the complexity of the representation and create a compact *summary* that encapsulates the key object properties. Moreover, in real applications, the representation needs to be robust to sensor noise, variations in object pose and scale as well as data incompleteness caused by occlusions and imperfect segmentation.

Object representation is often obtained by extracting a set of local object features and then defining an object model in terms of feature occurrence statistics, such as in the Bag-of-Words (BOW) model [11]. Recently, the increased accessibility of depth data has simulated development of 3D local descriptors. The majority of those representations encode local shape in the neighborhood of a point, for example the Fast Point Feature Histograms (FPFH) [21], Signature of Histograms of Orientations (SHOT) [27], and many more [12][26][8].

An object representation can be built incrementally where information is repetitively summarized at each step. It has been shown that incorporating information about object structure beyond the local properties significantly improves results [14][16]. Many 3D methods that incorporate object structure have been inspired by those proposed in the field of object modeling from 2D images. They usually first define a set of object parts based on local features and then encode their geometrical relationships.

One approach is to store coarse global spatial information by counting local feature occurrence at particular positions on an object, such as the 3D spatial pyramids [3][15]. However, since this quantitative approach relies on precise estimation of an object boundary or its center, it is not robust to imperfect segmentation or variations in object orientation. Another group of methods are those that directly add information about object structure to the local descriptor. For example, the methods from the Viewpoint Feature Histogram (VFH)-family [22][2] extend the FPFH by including estimation of a camera viewing direction and creating a global reference frame. However, the relation to the viewpoint makes them sensitive to object rotation. The problem of robustness to different object poses and scales has been addressed in [16]. We previously introduced the Global Structure Histogram (GSH) descriptor that obtains an incremental summarization by dividing an object surface based on its local characteristics into patches of different geometrical properties, and then encoding the distribution of distances between pairs of the patches. By implicitly representing a global ordering and position of the regions in an object internal reference frame, the GSH provides significant improvements to other state-of-the-art methods in realistic scenarios.

However, these methods summarize all local object features, whereas the intuition is that only a few are relevant for a specific task. In this paper we take an approach which we seek to extract only these essential features. This concept has been explored in computer vision using graphical models [13][24][29]. More generally, the approach is loosely related to interest point detection [18]. A few 3D keypoint detectors have been recently proposed and they are often motivated by similar work in the 2D domain [25][28].

In terms of representing object regions, the Clustered Viewpoint Feature Histogram (CVFH) [2] finds and describes all continuous surface patches in an object, as they are assumed to be less affected by noise which is predominantly associated with the object edges. Authors do not look into importance of different object regions for a given task in contrast to the method presented in this paper. Moreover, our method does not make any explicit assumptions about which information in the object is discriminative. Instead, it automatically discovers and extracts this information from data.

III. Methodology

Given a set of N objects $\mathcal{O} = \{O_i\}_1^N$ associated with class labels $\mathcal{L} = \{l_i\}_1^N$ from the set $l_i = \{c_m\}_1^M$, where M is the number of classes, we wish to find a vectorial representation $\mathbf{y}_i$ that is low-dimensional and robust to noise variations that are tied to those characteristics in the observations deemed relevant. Each object O_i is initially represented as a point cloud $\mathbf{o}_i$ extracted from the scene. From this point cloud a set of P_i local features $\mathbf{X}_i = \{\mathbf{x}_j^i\}_1^{P_i}$, where $\mathbf{x}_j^i \in \mathbb{R}^q$, can be extracted using one of the many local feature descriptors such as [21][12]. The focus of this paper is on how to summarize the set $\mathbf{X}_i$, where each object O_i can have a different cardinality, to a vector representation $\mathbf{y}_i$ with the same dimensionality.

The bag-of-words model [11] is a very popular approach to achieve such summarization. In that model, the first step is to obtain a discretization of the space of the local feature X by finding a set of key points often referred to as words. The words are usually found by clustering all local features from all objects. The notion is to direct words according to the underlying structure of the data. The final step in the summarization consists of associating each feature with a word through a similarity measure and using the distribution of associations as the feature space $X = \mathbb{R}^q$. However, if the set $\mathbf{X}_i$ is dominated by features $\mathbf{x}_j^i$ that are irrelevant or contain very little information about the class then $\mathbf{y}_i$ will not be a good representation of l_i. For example, a cup might be discriminated from a can by having a handle, but not by large cylindrical surfaces that are common for both objects. To avoid a representation that is dominated by irrelevant information, a summarization method based on the *quality* of a word rather than the *quantity* was proposed in [1]. This summarization is referred as the *Qualitative Vocabulary Based Descriptor* (QVBD) and is computed for a 3D point cloud data in the following three steps:

1) Estimate a local feature descriptor for each point and cluster the data
2) Compute the local classifiers for each word and each class
3) Describe the objects by max-pooling the responses obtained from the local classifiers

We will now proceed to describe this summarization method in detail and outline its specifics to 3D object representation.

A. Qualitative Features

Assume a feature set $\mathbf{X}_i$ representing each object O_i with the associated label l_i and a set of words $\mathbf{W} = \{\mathbf{w}_k\}_1^K$, where K is the number of words, that partition the space X. The qualitative feature summarization begins by associating each feature $\mathbf{x}_j^i$ with the most similar word $\mathbf{w}_k$ in $\mathbf{W}$. We will assume that similarity is encoded by proximity meaning that each feature is associated with its closest word. In the second step the aim is to try to recover the class-dependent structure of each word. To that end, a hyper plane $\mathbf{f}_{w_k}^{c_m}$ is found for each class c_m and each word w_k. The hyper-plane $\mathbf{f}_{w_k}^{c_m}$ is aimed at finding the best separation between each feature associated with word w_k that has class label c_m and all other features associated with word w_k. The intuition here is that a word for which each class has the same structure is irrelevant, but one with large separation is discriminative and contains important information about the class (see Figure 2). By employing such hyper-planes, we can generate a representation of each object $\mathbf{z}_i^{c_m} \in \mathbb{R}^{P_i \times K}$ with respect to the class c_m, where each element $\mathbf{z}_i^{c_m}(\cdot,\cdot)$ is a pseudo probability $0 \leq \mathbf{z}_i^{c_m}(\cdot,\cdot) \leq 1$ obtained from

$$\mathbf{z}_i^{c_m}(j,k) = \delta\left(w(\mathbf{x}_j^i), \mathbf{w}_k\right) L(\mathbf{x}_j^{i\,\mathbf{T}} \mathbf{f}_{w_k}^{c_m}). \qquad (1)$$

Here $w(\mathbf{x}_j^i)$ is the closest word to $\mathbf{x}_j^i$, $L(\cdot)$ is the logistic function which transfers the responses into a pseudo probability and $\delta(\cdot,\cdot)$ is the Dirac delta function. The value of $\mathbf{z}_i^{c_m}(j,k)$ is equal to zero for every word $\mathbf{w}_k \neq w(\mathbf{x}_j^i)$. Each element of this representation measures how well a given feature is representative of class c_m. Large separations are discriminative and contain important information about the class (see Figure 2).

Having $\{\mathbf{z}_i^{c_1}, \ldots, \mathbf{z}_i^{c_M}\}$ for each object instance, a final representation is calculated by summarizing them into one fixed dimensional matrix $\mathbf{y}_i \in \mathbb{R}^{M \times K}$ with

$$\mathbf{y}_i(m,k) = \max\{\mathbf{z}_i^{c_m}(j,k) : j \in [1,\ldots,P_i]\}. \qquad (2)$$

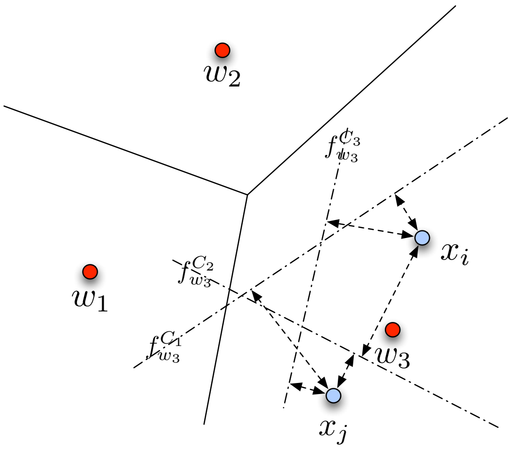


Fig. 2. Illustration of the qualitative summarization method from [1]. The method is a two stage approach. In the first step we estimate a set of *words* (shown in red) to partition the space of features. In the second step, side information (such as class label) is used to find a hyper-plane (dashed lines) that creates a soft partitioning of each word space with respect to class. This plane provides a measure of how representative a specific feature is to describe the concept used to supervise the procedure. The figure has been adopted with permission from [1].

The final representation contains the responses for the most representative features found on the object with respect to each word and each class, and is referred as the *Qualitative Vocabulary Based Descriptor* (QVBD).

The QVBD is a universal descriptor, which given an object point cloud, can extend any local feature descriptor to include qualitative information. In the next section, we evaluate the QVBD based on the popular FPFH [21] local descriptor by applying it to a 3D object database.

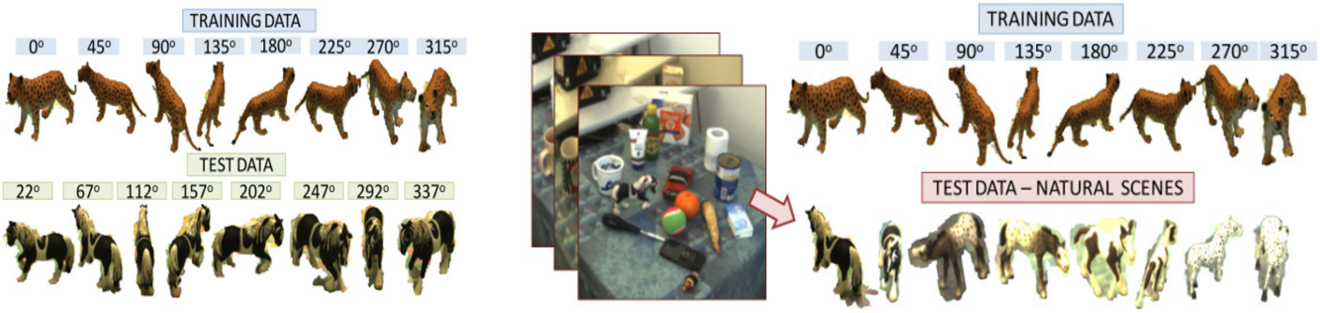
IV. EXPERIMENTAL EVALUATION

In this section, we present a qualitative and a quantitative evaluation of the QVBD descriptor for 3D object categorization. First, we thoroughly and systematically analyze its descriptive and discriminative properties, and demonstrate its ability to select essential 3D object characteristics. Second, we compare its performance with the state-of-the-art representations that use different types of summarization in real scenarios.

A. Database

Evaluation is performed on the challenging *Stereo Object Category (SOC)* database [17] that contains RGB-D data (images and point clouds) collected using the 7-joint Armar III robotic head equipped with two foveal and peripheral cameras. Objects are separated from the background using an active segmentation method [4]. The database contains 14 object categories: *ball, bottle, box, can, car-statuette, citrus, cup, 4-legged animal-statuette, mobile, screwdriver, tissue, toilet paper, tube* and *root-vegetable*, with 10 different object instances per category.

The SOC database consists of two datasets. In the first one, for each object, the data are collected from 16 views



(a) Illustration of the first experimental setup of the SOC database where rotation of single objects differs; 8 views per object are used for training an object model (top row) and other 8 views for evaluation (bottom row).

(b) Illustration of the second experimental setup of the SOC database for testing object representations in real conditions. Models trained on the data from the previous setup are tested on examples from 10 natural scenes where an object pose and scale, and a degree of occlusions vary significantly.

Fig. 3. Experimental setups and examples of objects from the Stereo Object Category database [17]. Object representations are evaluated only on 3D portion of the database. We use images of the objects here for better visualization. Data for all objects and natural scenes can be viewed at our web site `http://www.csc.kth.se/~madry/research/stereo_database/index.php`.

uniformly spaced around the object (every 22.5°), see Figure 3(a). In the second, the data are extracted from 10 natural scenes where 10 to 15 object instances from 14 different categories are randomly placed on a table. This dataset has 235 object point clouds that are characterized by significant variations in the objects poses, scale and degree of occlusion typical for real scenarios, see Figure 3(b).

B. Experimental Setup

In this paper, we used the same setup as in [16]. We performed cross-validation with the data divided into a training and test set with ratio 60:40%. Each experiment was repeated three times for randomly chosen object instances in order to average the results. Moreover, an object instance used for the training phase was never again used for an evaluation.

The state-of-the-art descriptors compared in this paper model the distribution of different features using the histogram-based representation. Following the previous successful results of applying the SVM classifier to this type of data [5], we employ the same strategy. We report the categorization results for a linear kernel for our approach, and the best of linear, RBF, χ^2 and histogram intersection kernels for the other discussed methods.

C. Qualitative Evaluation

As stated, our aim is to identify object characteristics that are essential for categorization. It is desired that the selected features are consistent for different object instances within one category and remain stable over variations in object pose and scale, and are unaffected by changes in data quality. Moreover, they need to be specific to each category to enable discrimination between different object classes. We first analyze these properties qualitatively for our method on the SOC database for both the single object dataset and the natural scenes (Figures 3).

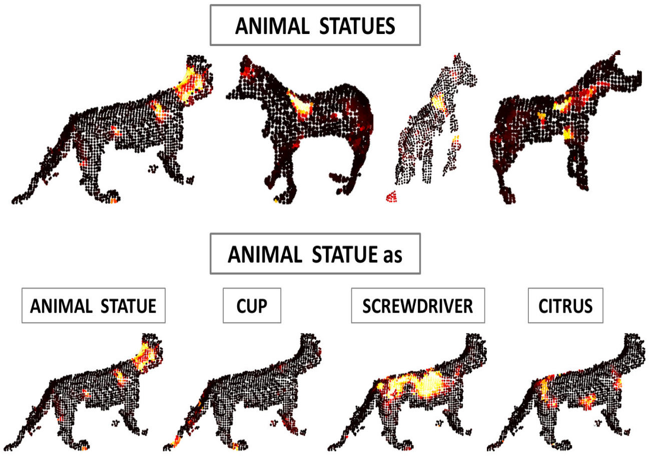


Fig. 4. Estimation of *essential* characteristics of an object for a given category. The heat maps visualize values of pseudo-probabilities at each object point (the brighter the point, the higher the probability). The bright regions consist of features that are crucial for identification of an object category. (Top) Four *animal* object instances in different poses interpreted by the *animal* category model. Our method is able to identify consistent feature regions for objects within one category. (Bottom) A *tiger* object interpreted by four different category models. The non-animal models find features on the tiger that correspond to their own class.

To this end, we first extract a local surface descriptor for each 3D point. We used the established FPFH [21] for feature extraction, since it is related with the further discussed (semi)-global descriptors. Then, we cluster the features using K-means with the Euclidean distance to obtain words and assign the features to their closest word. We computed the local linear classifiers for each word and each object category, Thus, the final number of the local classifiers is equal to $K \times M$, where K is the number of words and M is the number of object categories (see Section III). Next, we measured the response of each feature vector to the corresponding M local classifiers ($\mathbf{f}^{c_m}_{w_k}$ hyper-planes). This gives a pseudo-probability value for each 3D point and each category. Thus, we could interpret an arbitrary object point cloud using a category-specific model. Figures 4 and 6 visualize the probabilities obtained on different point clouds as heat maps.

The high probability regions (indicated by bright color) consist of features that play a significant role in identification of an object category. Our method does not use all the features in the high probability regions, but chooses the best candidates that support a given hypothesis. The visualizations of these regions in terms of heat maps show the robustness of our method in selecting similar features across different examples. Figure 4 (top row) presents the heat maps for a few different *animal* objects that have been interpreted by the *animal* category model. We can observe that our method consistently identifies important regions (for the *animal* category around the neck and ears) generalizing over object instance specific characteristics and various poses.

What links the tiger with an animal, a cup, a screwdriver or a citrus? To find the answer, we applied the models of different categories to the tiger instance, as presented in Figure 4 (bottom row). In line with one's intuition, our results suggest that the regions close to the tiger's neck, claws and tail are important to perceive it as an *animal*, the shape of the tail links it with a *cup*, its body with a *screwdriver*, the round sides of the body are common with a *citrus* and it does not have characteristics that are essential for a *box*. This shows that the QVBD allows us to automatically discover human-interpretable properties of a given object that are common with objects from other categories. This is key to define a suitable object representation for autonomous agents.

In order to achieve discrimination between different categories, a representation should capture unique characteristics of each class. As presented in Figure 6, a *cup*, a *can*, a *bottle* and a *toilet paper* all have large cylindrical surfaces. This makes them difficult to distinguish for quantitative descriptors (see the confusion matrix for the BOW in Figure 7). In contrast, our qualitative descriptor extracts essential properties for each category and considers repetitive structures irrelevant. It identified as distinctive the regions close to: (a) the rim and the handle of a *cup*, (b) the flat top of a *can*, and (c) the narrowing of the neck of a *bottle*. For any category, the cylindrical areas were not seen as important. We will show in Section IV-D that this property significantly improves the results compared to the methods based on other types of summarization.

Figure 5 presents the most important results for applying a model to a specific category to all objects in a scene. As can be seen, the high probability regions are consistent with the previous results. For example, the method finds important features close to the neck of an *animal*, the rim and the handle of a *cup*. This confirms that it is capable of generalizing over large variations in object appearance, pose and scale. Since the pseudo-probabilities are calculated using local classifiers with no information about the global structure of the object, high probability regions can appear on object instances from other categories. For example, for a *box* additional objects that have box-like features were detected (the mobile phone or the tissue package). In this framework it is left to the final classifier to pick which responses are needed to identify each object class.

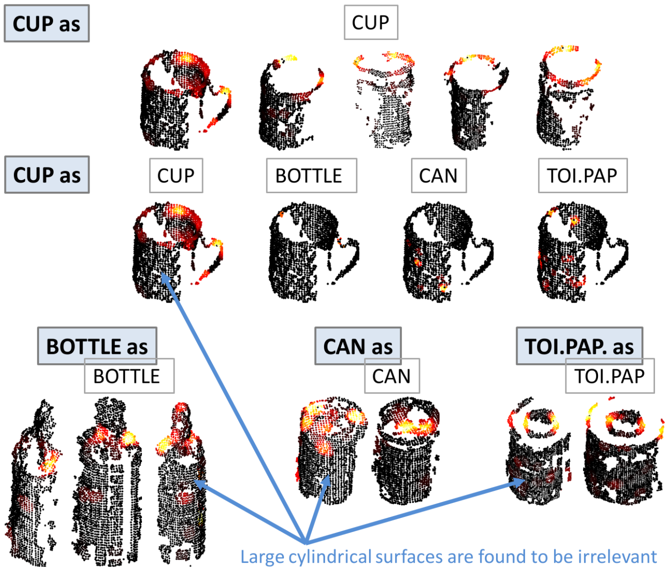


Fig. 6. Estimation of unique characteristics for geometrically similar object categories. The high probability regions (indicated by bright color) consist of features that play a significant role in identification of object categories. In the first row, the *cup* model is applied to different instances of this category and it can be seen that distinctive features of the cups are robustly extracted across several instances. Meanwhile, models coming from other categories do not produce high probability regions on the cup instance (second row). For objects that are similar to a *cup* (third row), their corresponding models highlight regions that are unique for those categories and not shared across the other categories.

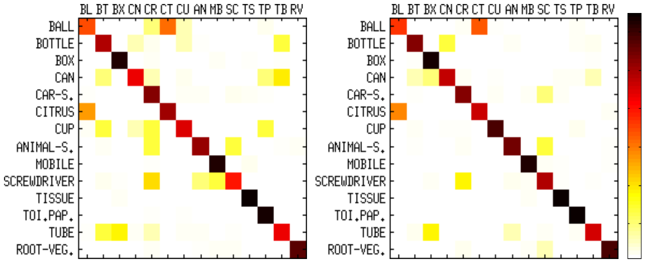


Fig. 7. Confusion matrices obtained for: (left) the Bag-of-Words (BOW) and (right) the Qualitative Vocabulary Based Descriptor (QVBD) for the first setup of the Stereo Object Category (SOC) database [17] presented in Section IV-C. The images are best viewed in color.

D. Quantitative Evaluation

In this section, we present a comprehensive quantitative comparison of our method with several state-of-the-art 3D descriptors adapting different methods of summarization, such as Bag-of-Words (BOW) based on Fast Point Feature Histograms (FPFH) [21], and the (semi-)global descriptors such as the Global Fast Point Feature Histograms (GFPFH) [23], the Viewpoint Feature Histogram (VFH) [22], the Clustered Viewpoint Feature Histogram (CVFH) [2] and the Global Structure Histogram (GSH) [16]. We selected results that demonstrate the most important properties of these representations for use under real-world conditions. For each experiment, we report the average categorization rate and the standard deviation (σ). We refer to results presented in [16].

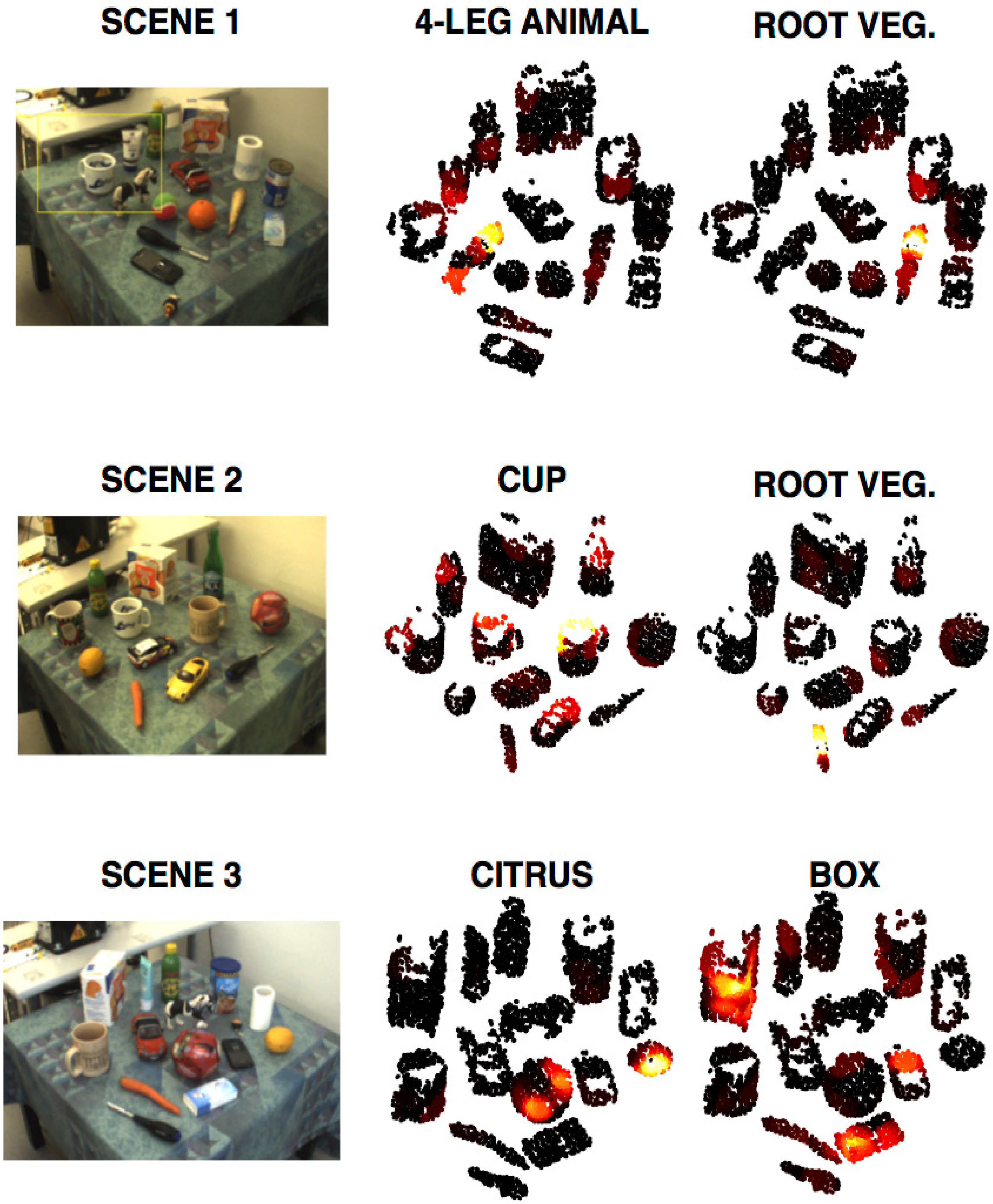
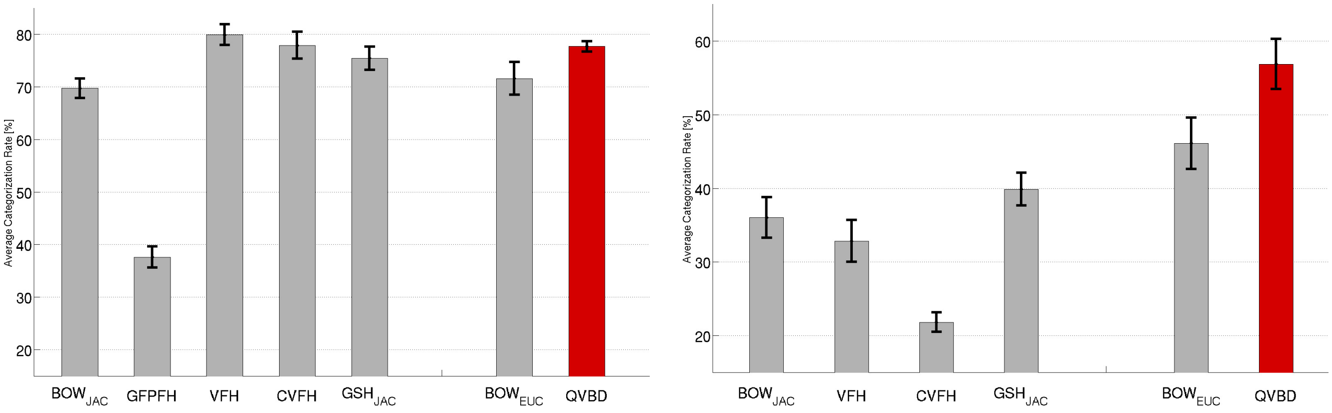


Fig. 5. Estimation of *essential* characteristics of objects in three real scenes. Each of the scenes is analyzed for two different categories and the results are aligned row-wise. All objects in the scenes are interpreted by the same category model, for example 14 objects in the first scene by the *animal* model (top row, middle column). The high probability regions (indicated by bright color) indicate features that are crucial for a given category.



(a) First setup - Results for single objects from the SOC database where training and test data differ in object rotation. Experimental setup is shown in Figure 3(a).

(b) Second setup - Results for objects from 10 natural scenes in the SOC database where training and test data differ significantly in an object pose and scale. Experimental setup is shown in Figure 3(b).

Fig. 8. Comparison of several quantitative and qualitative object representations in terms of average categorization rate performed on data that differ in quality and amount of available training examples. Abbreviations used for representations: BOW-JAC - Bag-of-Words based on the Jaccard distance with 100 words; GFPFH - Global Fast Point Feature Histogram [23]; VFH - Viewpoint Feature Histogram [22]; CVFH - Clustered Viewpoint Feature Histogram [2]; GSH - Global Structure Histogram [16]; BOW-JAC - Bag-of-Words based on the Euclidean distance with 1000 words; QVBD - Qualitative Vocabulary Based Descriptor (QVBD) with 1000 words. The BOW-JAC, GFPFH, VFH and GSH have been evaluated in [16].

In order to systematically study the properties and robustness of the method, we formulated two experimental setups of increasing complexity. First, we perform experiments on the first part of the SOC database in which we vary rotation of single objects used for training and testing, as presented in Figure 3(a). The quality of the data is also influenced by imperfect segmentation and real sensory noise. Second, the models obtained for the first setup are tested against objects extracted from 10 natural scenes, as presented in Figure 3(b). In this setup, the difficulty of the problem is considerably increased by significant variations in the object pose, scale, and data resolution as well as data incompleteness due to object occlusions.

(1) *Importance of qualitative information:*While comparing the confusion matrices for the BOW and the QVBD (Figure 7), we can observe that incorporating qualitative information helps to discriminate between object categories that share a substantial portion of the local appearance and only differ in a small amount of features. As reasoned in the previous section, the QVBD is able to correctly discriminate between such categories whereas BOW easily confuses them. For example, the *cup* category is well discriminated from the *bottle*, *can* and *toilet paper*. The hardest to recognize are the *ball* and *citrus*. Since in 3D they both contain almost solely one type of features (describing round surfaces), the QVBD discovers one kind of essential regions and in consequence does not improve the results.

Figure 8 compares the categorization rates of the QVBD with the BOW representations based on the Euclidean distance (BOW-EUC) and the Jaccard distance (BOW-JAC), where the latter is adapted from [16]. The QVBD clearly outperforms the BOW representations for both analyzed setups, confirming the importance of high quality features.

(2) *Summarization method:* We compare QVBD with the state-of-the-art 3D descriptors that are based on different types of summarization. For the first setup (Figure 3(a)), the VFH descriptor slightly outperforms other descriptors. Nevertheless, the results show that the QVBD performs equally well as the three quantitative descriptors - the VFH, CVFH, GSH (σ=1.8%). However, in this setup all object poses used at the test time were also available during training and there is relatively small variance in the amount of different types of features between the training and testing views. Collecting data that represent all possible object orientations present in real scenarios is very expensive or even unfeasible.

Therefore, more realistic results are those obtained for the second setup (Figure 3(b)). We tested the descriptors on the real scenes where an object pose, scale and data resolution significantly differ. The QVBD massively outperforms all the quantitative descriptors encoding global object structure (by 17% comparing to GSH) and the BOW (by 11%). Finally, while comparing two closely related descriptors, namely the VFH and the CVFH, we can see that describing only the continuous regions is highly effected by the data incompleteness. These results demonstrate the great importance of encoding and properly selecting the relevant object characteristics.

V. Discussion

In this work we showed how to extract the essential object characteristics for discriminating between different object classes. However, we believe that the benefits of the proposed method expand beyond object classification. First and foremost, the method can learn a representation given labels of any kind, not necessarily corresponding to human-defined object categories. As such, one possible application is grasping, for which our method would permit learning relationships between the local characteristics of an object and the grasp parameters. This would facilitate transfer of grasps between locally similar objects.

As can be seen from the qualitative results, the response of the feature is consistent over both object instances and viewpoints. This information can be exploited in order to perform other types of reasoning about objects. In particular, imagine a scenario where a robot located in a kitchen is tasked with finding an object to drink from; our representation would allow the robot to search for characteristics resembling objects that are known to afford drinking. Furthermore, knowledge about essential features would allow the robot to plan for observing the viewpoints that are most relevant for discriminating the object of interest.

Finally, the approach generates a sparse representation in which an object is described by only a subset of the extracted local features. We believe this to be a very useful property for extending the global summarization methods, such as the GSH descriptor [16], where focusing on the *essential* local characteristics should further improve the object categorization performance. Moreover, the sparsity of the representation should significantly reduce the computational complexity. This is where we intend to focus our future work.

VI. Conclusions

Building representations of objects is traditionally based on incremental summarization of local features. In this paper we have shown that local statistics of 3D features contain a large portion of variations that are not discriminative for the class. As a solution, we proposed the use of a qualitative summarization approach which generates a representation by automatically selecting only the essential characteristics of an object given the task. Further, the features are consistent over various instances of the same class, and most importantly, form interpretable structures which could be used in many robotics applications. We demonstrated significant improvements for 3D object categorization on a challenging dataset with respect to other state-of-the-art methods.

References

[1] H. M. Afkham, C. H. Ek, and S. Carlsson. Qualitative Vocabulary Based Descriptor. In *ICPRAM*, Feb. 2013.

[2] A. Aldoma, N. Blodow, D. Gossow, S. Gedikli, R. B. Rusu, M. Vincze, and G. Bradski. CAD-Model Recognition and 6 DOF Pose Estimation using 3D Cues. In *3dRR-ICCV*, 2011.

[3] M. Ankerst, G. Kastenmuller, H.-P. Kriegel, and T. Seidl. 3D shape histograms for similarity search and classification in spatial databases. In *SSD*, 1999.

[4] M. Bjorkman and D. Kragic. Active 3D scene segmentation and detection of unknown objects. In *ICRA*, May 2010.

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

[6] R. Detry, C. H. Ek, M. Madry, and D. Kragic. Learning a dictionary of prototypical grasp-predicting parts from grasping experience. In *ICRA*, 2013.

[7] R. Detry, C. H. Ek, M. Madry, J. Piater, and D. Kragic. Generalizing grasps across partly similar objects. In *ICRA*, 2012.

[8] B. Drost, M. Ulrich, N. Navab, and S. Ilic. Model globally, match locally: Efficient and robust 3D object recognition. In *CVPR*, 2010.

[9] J. Gall, A. Fossati, and L. van Gool. Functional categorization of objects using real-time markerless motion capture. In *CVPR*, 2011.

[10] H. Grabner, J. Gall, and L. Van Gool. What makes a chair a chair? In *CVPR*, 2011.

[11] Harris, Z. Distributional structure. *Word*, 1954.

[12] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *PAMI*, 1999.

[13] D. Larlus and F. Jurie. Latent mixture vocabularies for object categorization and segmentation. *IVC*, 2009.

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

[15] R. J. López-Sastre, A. García-Fuertes, C. Redondo-Cabrera, F. J. Acevedo-Rodríguez, and S. Maldonado-Bascón. Evaluating 3D spatial pyramids for classifying 3D shapes. *CG*, 2013.

[16] M. Madry, C. H. Ek, R. Detry, K. Hang, and D. Kragic. Improving generalization for 3D object categorization with Global Structure Histograms. In *IROS*, 2012.

[17] M. Madry, D. S. Song, and D. Kragic. From object categories to grasp transfer using probabilistic reasoning. In *ICRA*, 2012.

[18] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 2004.

[19] B. Mirkin. *Core Concepts in Data Analysis: Summarization, Correlation and Visualization*. UTCS. Springer London, 2011.

[20] A. Pieropan, C. H. Ek, and H. Kjellstrom. Functional object descriptors for human activity modeling. In *ICRA*, 2013.

[21] R. B. Rusu, N. Blodow, and M. Beetz. Fast Point Feature Histograms (FPFH) for 3D Registration. In *ICRA*, 2009.

[22] R. B. Rusu, G. Bradski, R. Thibaux, and J. Hsu. Fast 3D recognition and pose using the Viewpoint Feature Histogram. In *IROS*, 2010.

[23] R. B. Rusu, A. Holzbach, G. Bradski, and M. Beetz. Detecting and segmenting objects for mobile manipulation. In *S3DV*, 2009.

[24] P. Schnitzspan, S. Roth, and S. Bernt. Automatic discovery of meaningful object parts with latent CRFs. In *CVPR*, 2010.

[25] I. Sipiran and B. Bustos. A robust 3D interest points detector based on Harris operator. In *EV3DOR*, 2010.

[26] B. Steder, G. Grisetti, M. Van Loock, and W. Burgard. Robust on-line model-based object detection from range images. In *IROS*, 2009.

[27] F. Tombari, S. Salti, and L. Di Stefano. Unique signatures of histograms for local surface description. In *ECCV*, 2010.

[28] F. Tombari, S. Salti, and L. Di Stefano. Performance evaluation of 3D keypoint detectors. *IJCV*, 2013.

[29] L. Yang, R. Jin, R. Sukthankar, and F. Jurie. Unifying discriminative visual codebook generation with classifier training for object category recognition. *CVPR*, 2008.