Improving Generalization for 3D Object Categorization with Global Structure Histograms

Marianna Madry Carl Henrik Ek Renaud Detry Kaiyu Hang Danica Kragic

***Abstract*— We propose a new object descriptor for three dimensional data named the *Global Structure Histogram (GSH)*. The GSH encodes the structure of a local feature response on a coarse global scale, providing a beneficial trade-off between generalization and discrimination. Encoding the structural characteristics of an object allows us to retain low local variations while keeping the benefit of global representativeness. In an extensive experimental evaluation, we applied the framework to category-based object classification in realistic scenarios. We show results obtained by combining the GSH with several different local shape representations, and we demonstrate significant improvements to other state-of-the-art global descriptors.**

I. INTRODUCTION

The development of depth sensors has led to a wide availability and use of 3D sensory data [23], [26], [32]. This created a need for defining suitable data representations that facilitate detection, recognition and categorization of objects in natural settings. In this paper, we first analyze the desired characteristics of an object representation. Based on this analysis, we propose a 3D representation that encodes global object structure of local surface properties and can robustly generalize over different views with incomplete data.

How should objects be represented? The representation may be application-dependent but there are general characteristics that are desirable. A good representation is sensitive to inter-class variations allowing discrimination between instances deemed as different, while being able to generalize over examples that the task defines as being the same. Further, it should be robust such that small perturbations do not significantly affect the interpretation. To facilitate the above characteristics, a suitable similarity measure needs to be defined – a good representation maximizes the portion of relevant variance in the data which, in turn, simplifies the learning problem. This has not been the dominant strategy in data-driven learning as a number of methods seek strength in number of training examples and focus on quantity rather than quality. As an example, the focus of most approaches for the Pascal VOC challenge [2] have been to extract as much of the variance in the images as possible. This approach is justified, even though rarely explicitly, by the belief that given enough data, machine learning techniques will be capable to extract the right discrimination and generalization characteristics.

M. Madry, C. H. Ek, R. Detry, Kaiyu Hang 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,chek,detryr,kaiyuh,danik}@csc.kth.se

This work was supported by the Swedish Foundation for Strategic Research, the EU projects CogX (FP7-IP-027657) and TOMSY (IST-FP7-Collaborative Project-270436), and the Belgian National Fund for Scientific Research (FNRS).

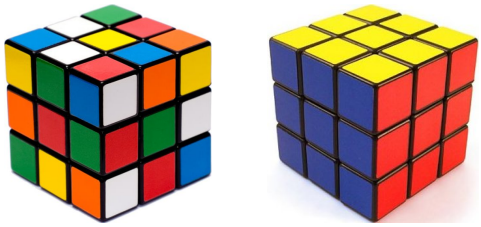


Fig. 1. Solving the classic Rubik's Cube implies arranging its tiles such that each face is populated by a single color. Two cubes in different states are shown: the left one in a unsolved state while the right is solved. What makes the cubes different? Each contains the same number of tiles and the same colors. The difference lies on a larger scale in the arrangement of the individual tiles. For Rubik's cubes it is not necessary to know the color that covers a specific face, it sufficient to know that each side is covered with tiles of only one color. This exemplifies the central motivation in this paper: by encoding structure we can achieve relevant generalization (in this case over all solved states) and by having a less detailed local description (no need for the absolute color of a tile) we can increase robustness.

The ideal characteristics of a representation is tightly entwined with application and task as these define the granularity that provides the means of interpreting variations in the data. In this paper, our focus is on a 3D object representation for category recognition. Our motivation for the work is twofold. First, it is an essential capability for a robot interacting with the environment. Second, compared to instance recognition and object pose estimation it requires a larger degree of generalization making it therefore a much more challenging problem.

Do we need an alternative feature descriptor? Common for the state-of-the-art descriptors is that the encoding is done on local characteristics and first-order statistics of the data. In this paper, we argue that the information with the right discrimination and generalization characteristics is related to the structure on a larger scale. Therefore, we wish to find a global representation that encodes the structure of the object. What we mean in formal terms is that: *an object is a two dimensional surface embedded in a three dimensional space which encapsulates a non-empty volume* [40]. We present a descriptor which is capable of robustly encoding local statistics of an object in such a manner that it reflects its global structure. Our idea is motivated in Fig. 1.

The remainder of the paper is structured as follows: Section II discusses our approach to defining a robust object representation describing $3D$ objects. Section III presents related work and Section IV introduces the details of the proposed method. Section V presents its comparison in terms of performance and generalization properties with other object representations. Section VI concludes the paper.

II. Our Approach

We propose a representation that respects the relevant structural granularity. We subscribe to the argumentation that, for most typical robotics tasks, achieving generalization is more challenging than discrimination when encoding data [11]. Performing a task such as pose and scale independent object category classification imposes significant challenges on the representation in terms of generalization. We argue that the relevant generalization is contained on a structural level and not in the local variations. By creating a representation based on this notion we achieve the following benefits:

1) Generalization: By generalizing over object pose, scale and instance specific properties, we can reduce the amount of training data required. This means, as we will show in Section V, that we can correctly classify objects seen from a viewpoint not present in the training phase. This allows for realistic applications as it is not feasible to acquire training data presenting all possible variations. In previous work, the problem of limited amount of training data has been addressed by generating multiple object views from synthetic data. However, in such case, sensor properties (e.g. noise) are not modeled and mismatch between ideal and real data is clearly visible [20], [39].

2) Robustness: The feature we present encodes the structural relationship between local characteristics in the data. We argue that the relevant information is contained in the variations of the relationships and not in the local variations. Therefore a coarse local representation is sufficient. This will improve robustness as there is less information that needs to be "removed" at the modeling stage. In case of a statistical model (such as in this paper), we need less training data in order to sample the domain well.

III. Related Work

Object modeling for recognition and categorization is often approached by extracting a set of local object features, from one or multiple views of an object, then defining a model in terms of feature occurrence statistics. In 2D, features encode for instance corners [14], blobs [30], or more general photometric variations [21].

In the field of object modeling from imaging sensors, it has been shown that encoding object structure above the local properties significantly improves robustness. Usually, representations of global structure are done by defining a set of object *parts* and then encoding the geometrical relationships between them. Part-based representations can capture different amounts of information about object structure. Object parts can, for example, be treated as geometrically independent (*Bag-of-Words* BOW model [8]). Another approach may be to store only a coarse global spatial information (*spatial pyramids* [19]) or more explicit spatial information (*constellation models* [38], [12]) including methods based on probabilistic modeling [34].

A serious issue that limits use of 2D models is that variations introduced by projective geometry (geometrical transformations, self-occlusions) have to be robustly captured and handled by the model. These problems can be alleviated by learning models from multiple views [20], [29], [33]. However, this approach leads to very complex object representations that require enormous amounts of data to estimate a large number of model parameters. In consequence, the performance of the method depends on the amount of training data and its ability to scale over a large number of classes is limited.

In response to the shortcomings of 2D models discussed above, modeling the 3D shape of objects has become increasingly popular. Moreover, the emergence of cheap 3D sensors has renewed the interest for purely 3D approaches [17], [23], [26], [32], [36], [39]. Depth data are characterized by lower pose variations than images, since they do not suffer from projective transformations. However, a single 3D view is still affected by self-occlusions, which makes holistic 3D descriptors [16], [35] difficult to apply to real-life robotics scenarios. Likewise, histogram and transform-based approaches [37] rely on the precise definition of the center of mass of an object, which is difficult to obtain from a partial view. To the end of improving robustness to (self-)occlusions, several authors have developed models which encode 3D local shape in a close neighborhood of a point. Popular options are the Spin Images [17], Fast Point Feature Histograms (FPFH) [26], or Radius-based Surface Descriptor (RSD) [23], and many more [15], [32], [36]. Another group constitutes complex parametric shape descriptors such as superquadrics [4], [31] or local edge descriptors [10]. It is important to mention, several approaches that represent the spatial relation between the local 3D features were proposed [23], [27]. These models are suggested to be invariant with respect to translation and scale changes, and partial occlusions.

In this paper, we present a unified and flexible 3D object representation, called *Global Structure Histogram*, that can be learned using complete or incomplete information about an object, and instantiated in partial object views.

IV. Global Structure Histogram

For tasks such as object categorization it is important to look beyond local statistics and encode the global and structural information in the data. In this section we provide details of the proposed *Global Structure Histogram*, or *GSH* for short. The aim of GSH is to represent objects in a manner such that it can robustly generalize over different poses and views, and cope with incomplete data.

The GSH descriptor is computed from an object's point-cloud reconstruction, in a three-stage process. First, we compute local surface-shape characteristics from the object (Fig. 2 left). Then, based on local descriptor, point labels are generated by performing a vector quantization into n_C clusters using the k-means algorithm with the partial histogram metric based on the Jaccard similarity coefficient [7], [3]:

$$d(D,C) = \frac{\sum_{i=1}^{N} |max(D_i,C_i)| - \sum_{i=1}^{N} |min(D_i,C_i)|}{\sum_{i=1}^{N} |max(D_i,C_i)| + offset} \quad (1)$$

where D is a vector with the surface descriptor and C is a cluster centroid. This metric was chosen as a superior

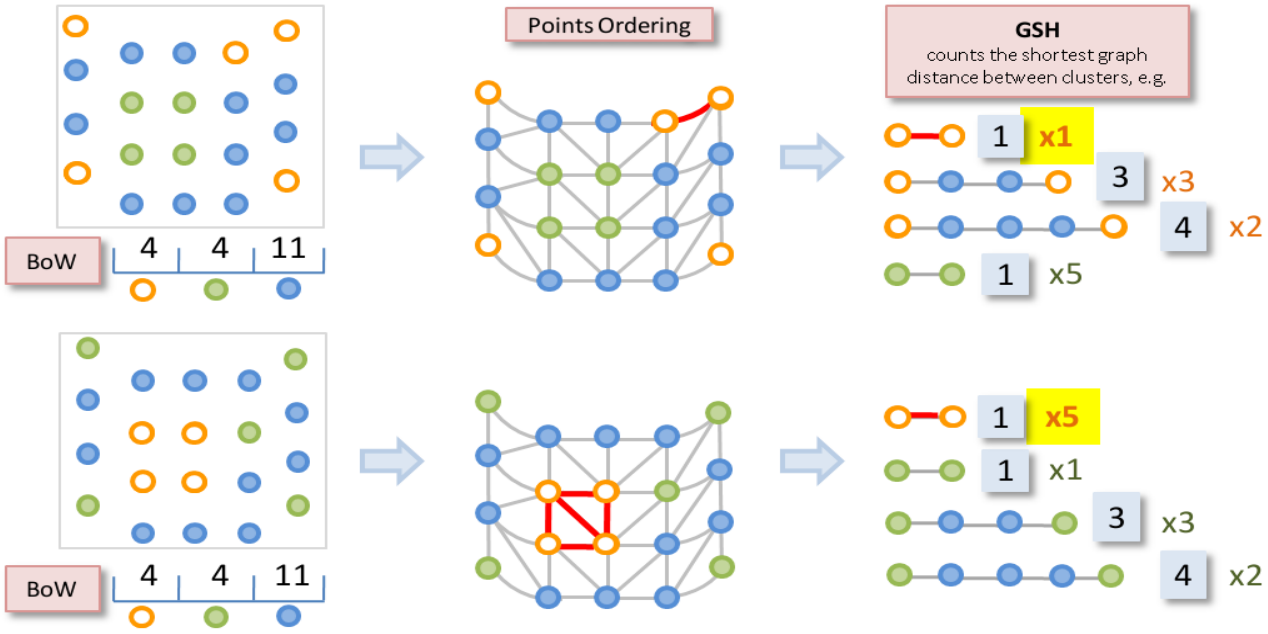


Fig. 2. *The figure illustrates schematically the proposed GSH object descriptor. The first and the second row depict two different instances of objects that we wish to encode. In the leftmost column a local descriptor has been used to associate each point as belonging to one of three classes, here represented by color. The discriminating information is contained in the relationship between the different points which would not be captured by a simple Bag-of-Words representation. The middle column describes the second step in the GSH descriptor where an approximation of the object surface has been computed. Based on it, we can compute the relationship between the local point classes as defined by this surface. The right column shows the distribution of paths of different length for the two objects. These distributions clearly separate the two objects.*

to other preliminary tested distances, such as City-block, Euclidean, or Hamming. The Jaccard distance limits the influence of isolated changes in a histogram originating in noise and enhances the importance of the repeating changes resulting from the object structure. At this stage, the Bag-of-Words (BOW) model can be obtained by estimating a distribution of the points over the n_C clusters. An example of dividing object points into different clusters based on the local surface descriptor is presented in Fig. 3, where each cluster is specified by a different color.

The abstraction from points to class labels "grounds" each part of the object and translates it to a common frame such they can be related to each other. An object can be defined as a two-dimensional surface embedded in a three dimensional space which encapsulate a non-empty volume [40]. This implies that given a point on the object, one can travel to any other point belonging to the object by traversing this enclosing surface. We wish the GSH descriptor to be sensitive to the structure of this encapsulating surface. To that end we use this path and encode the object using the structure of local characteristics when traversing along its surface (Fig 2, left). Being represented using a point cloud we first perform fast triangulation of unordered points to approximate the encapsulating surface. The chosen method [24] offers an adequate balance between computational efficiency and quality of a reconstruction. Once triangulation has been performed we can traverse the surface between two points by computing the shortest path using the Floyd-Warshall algorithm [13]. However, it is possible to use a method with a lower computational complexity [25].

Once we can compute the ordering of points with respect to the surface of the object the third and final stage is how to encode this structure in a robust manner. This structure or ordering is a combinatorial characteristic of the object. Encoding such information in a robust manner is often very challenging. To that end, instead of encoding the exact ordering we represent the object as the distribution of distances along the surface between each combination of two point classes. For each of such combinations, the distribution of distances is modeled as a histogram with B bins. This means that using n_C classes will result in a descriptor with $\frac{n_C \times (n_C+1)}{2} \times B$ elements (Fig 3, right). Experimentally estimated computational complexity of the algorithm that computes the GSH representation is equal to $O(log(n_C))$.

To obtain the final form of the GSH, we normalize: (1) each distance histogram for a combination of two point classes with an amount of point pairs in that histogram, and (2) the distance distribution matrix with respect to the maximum distance between points in the graph (it is equal to a total number of object points). An example of the GSH representation for several objects is presented in Fig. 3, where each row of the matrix represents distribution of a geodesic distance between two points in a graph that belongs to two different clusters.

In summary, the GSH descriptor is a global descriptor for encoding 3D data. Given a point cloud of an object, it is applicable to extend any local feature descriptor to include information which encodes the surface structure of the object. The descriptor is computed using following three steps:

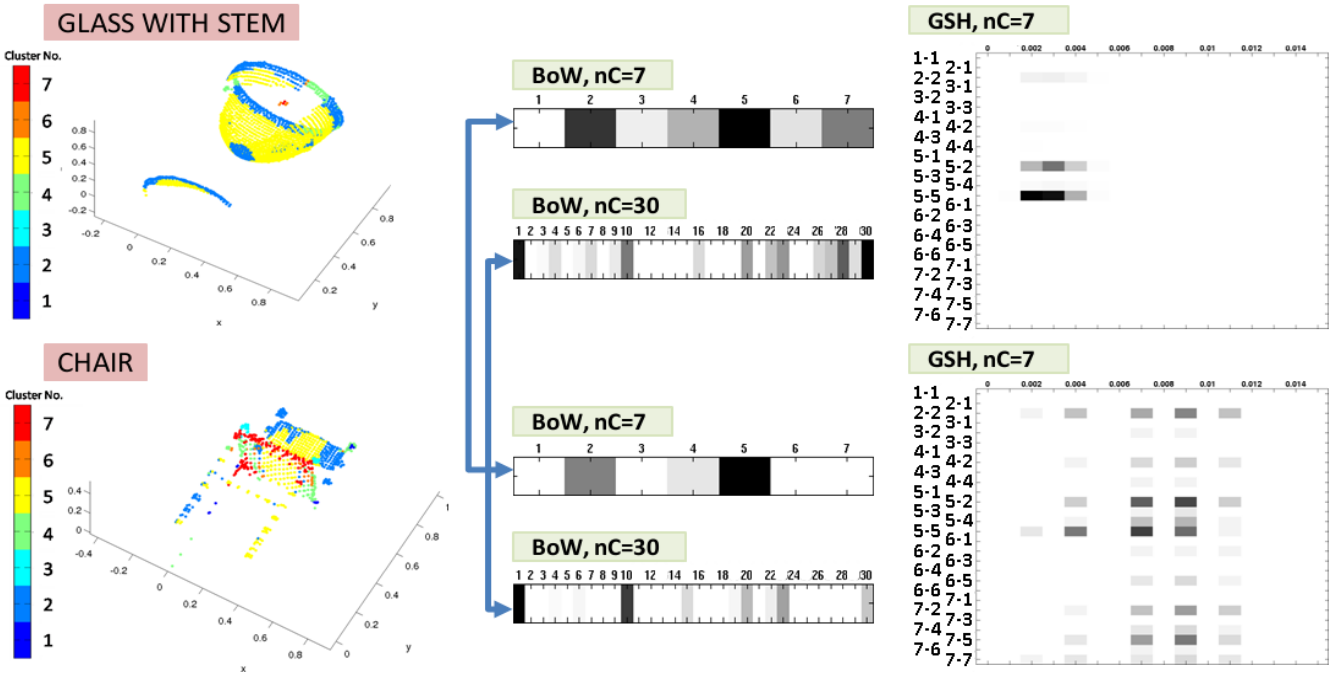


Fig. 3. *Example partial views of glass with stem (top row) and chair (bottom row) objects represented by seven types of surfaces ($n_C = 7$) estimated using FPFH local descriptor [26]. For both objects, the Bag-of-Words (BOW) representation is difficult to differentiate even for a large number of clusters ($n_C = 30$). Adding the structure information, even for a small number of clusters ($n_C = 7$), leads to the easily distinguishable Global Structure Histogram (GSH) representation. The resultant GSH Global Structure Histogram is illustrated as a matrix where each row represents a histogram of distances between all points belonging to the surface type j and surface type k, i.e., the first row is the distance histogram between points of type 1 and 1 (marked as 1-1), the second row is the distance histogram between points of type 1 and 2 (marked as 1-2) etc. The images are best viewed in color.*

1) Estimate local feature descriptor and approximate type of an object surface for all points
2) Determine ordering of points along the surface
3) Represent object as distribution of paths along a surface

In the next section, we analyze the performance of the proposed representation by applying it to benchmarks of 3D object databases.

V. EXPERIMENTAL EVALUATION

In this section, we present an evaluation of the GSH for object categorization. We systematically and exhaustively compare its performance and generalization properties with other local and global object representations on databases that differ in quality and amount of available training examples.

A. Databases

Let us first present two databases on which the experimental evaluation was performed:

1) Princeton Shape Benchmark: We collected complete object models from the Princeton Shape Benchmark (PSB) [1] for seven object categories of complex shapes: *bottle, car, chair, glass with stem, handgun, ice cream, vase*, each with 8 different object instances per category. Further, incomplete/partial point clouds were acquired as a subset of points of the complete model that are visible from a given viewpoint. We selected 8 different camera positions in two elevations $\alpha = \{30°, 60°\}$ and four horizontal direction separated by $90°$. The visibility of each point was determined using the Hidden Point Removal (HPR) operator [18].

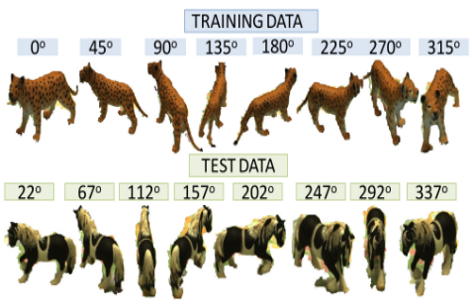
2) Stereo Object Category Database: The Stereo Object Category (SOC) database [22] contains RGB-D data collected using the 7-joint Armar III robotic head equipped with two foveal and peripheral cameras. To differentiate an object and background, an active segmentation method was used [5]. The database includes 14 object categories, each with 10 different object instances per category. For each object, both 2D (RGB image) and 3D (point cloud) data were collected from 16 different views around the object (separated by $22.5°$), see Fig. 4(a). In this paper, we use only the 3D portion of the database.

Additionally, there is a choice of data collected in the realistic scenarios. A few subjects were asked to randomly place between 10 to 15 objects from 14 different categories on a table. As a result, objects poses, scale and degree of occlusion vary significantly. In this way, data for 10 natural scenes (235 object point clouds) were obtained, see Fig. 4(b).

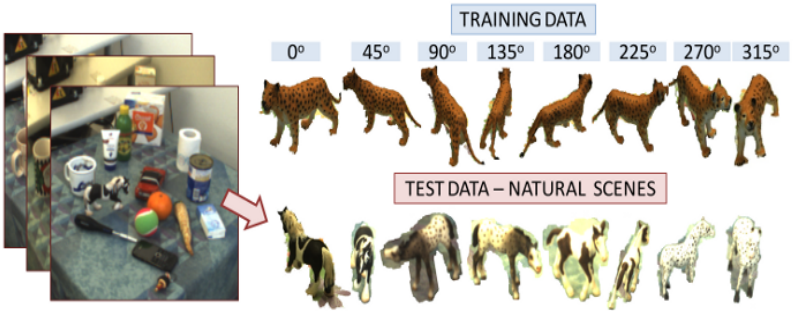
B. Experimental Setup

For each experiment, we performed cross-validation with data divided into a training and test set with ratio 50:50% of models per category per set for PSB and 60:40% for SOC database. To average the results, each experiment was repeated three times for randomly chosen object instances. Due to the fact that the aim was to test the performance of the system for object categorization and not object instance recognition, an object used for the training phase was never again used for an evaluation.

1) Classification: Descriptors evaluated in this paper, model the distribution of different features. Several works have shown that χ^2 kernels are good representations for



(a) *An experimental protocol where a rotation of objects differs, i.e. 8 views per object are selected to train an object model (top row) and other 8 views for its evaluation (bottom row).*



(b) *An experimental protocol for testing object representations in real conditions. Models trained on the data from the previous protocol are tested on examples from 10 natural scenes where an object pose and scale vary significantly.*

Fig. 4. *Experimental protocols and examples of objects from the Stereo Object Category database. Object representations are evaluated only on 3D portion of the database. We use here images of the objects 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`.

such data [9]. To that end, we employ the same strategy and perform object classification by applying an SVM in the space induced by the χ^2 kernel [6].

C. Experimental Results

We performed a thorough experimental evaluation comparing several state-of-the-art local and global representations for a number of parameters. Here, we present selected results highlighting the most significant and relevant properties of these representations for the use under challenging real word conditions. For each experiment, we report the average categorization rate and standard deviation (σ).

1) Selection of a Local Representation: GSH can be used in combination with any local surface descriptor, for example the Radius-based Surface Descriptor (RSD) [23] or Fast Point Feature Histograms (FPFH) [26]. Our representation requires object points to be divided into several groups (clusters) depending on surface properties in a close neighborhood of a point. Examples of assignment of points to different groups are shown in Fig. 3 and 7 where each of the clusters is marked by a different color.

To cluster the data, we use the k-means algorithm with n_C centers. Once each point is assigned with a cluster label, we compute the BOW representation by estimating a distribution of local descriptors over the n_C clusters. An optimal number of clusters needs to be selected in order to preserve a balance between discrimination and generalization properties of the representation; using few clusters compresses the data reducing the discrimination power of the representation, and many clusters decreases generalization within the clusters resulting in high sensitivity to small variations in the data. Not surprisingly, in Fig. 5 the classification rate increases together with the number of clusters until the point where it saturates and adding more clusters reduces the performance. We selected the optimal number of clusters for a BOW representation experimentally using cross-validation. Note, that an optimal n_C for other representations may differ from the one for BOW.

Fig. 5 shows that BOW_{FPFH} achieved a significantly higher categorization rate than BOW_{RSD}. We note that

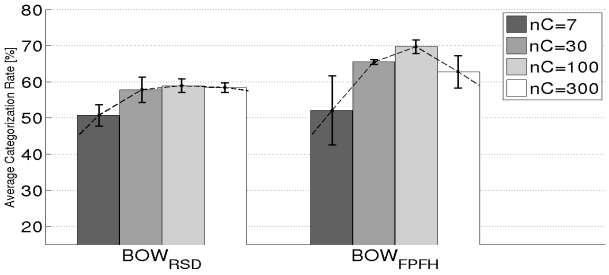
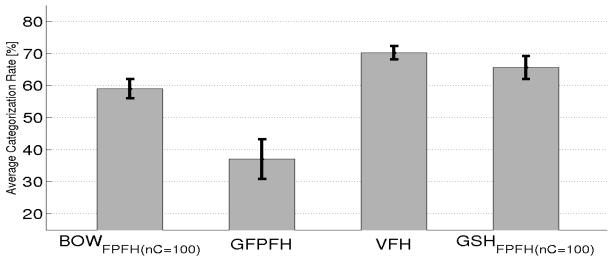


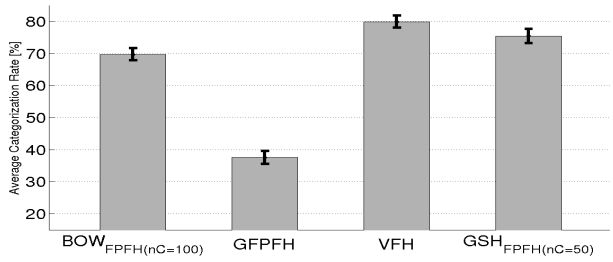
Fig. 5. *Comparison of RSD and FPFH local features encoded using the Bag-of-Words representation: BOW_{RSD} and BOW_{FPFH} for a different number of cluster $n_C = \{7, 30, 100, 300\}$ (using a higher number of n_C does not improve the results) on the Stereo Object Category database. As a result of a lower dimension of feature space for RSD than FPFH the optimal $n_{C_{RSD}} < n_{C_{FPFH}}$.*

FPFH outperforms RSD for all datasets irrespective of the use of BOW or GSH encoding. Thus, for the sake of clarity, we present only the results using FPFH as the local descriptor in the following sections. The categorization rate of BOW_{FPFH} constitutes a baseline for further evaluation of different global descriptors.

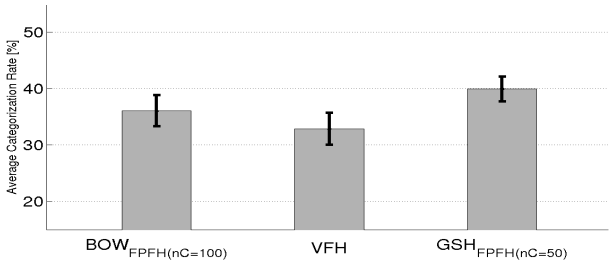
2) Encoding 3D Object Structure: The motivation behind the GSH is to construct a representation that encodes the global structure of an object. This notation has been previously exploited for a 3D object representation by the Global Fast Point Feature Histogram (GFPFH) descriptor [28]. This descriptor encodes the relation between local patches along rays. In Figures 6(a) and 6(b), we see that the GSH outperforms GFPFH for both synthetic and real stereo data. This may seem surprising given that the local feature (FPFH) is the same in both cases. However, there is a fundamental difference between the two approaches in the way they interpret structure of an object, and in consequence, encode the relationship between local patches. We base our approach on a standard definition, recapitulating from Section I that: *an object is a 2D surface embedded in a 3D space which encapsulates a non-empty volume* [40]. When the geodesic structure of local information in the GSH respects the above definition, the rays inducing the structure underpinning GFPFH are independent of the surface, i.e. 3D curvature of an object.



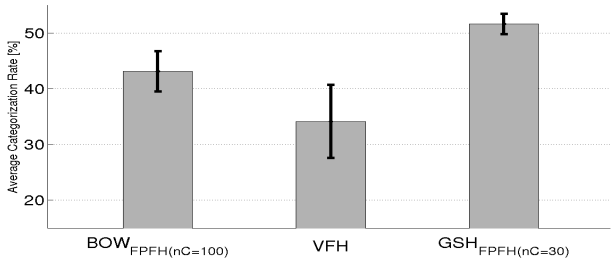
(a) *Results for synthetic incomplete object models collected form PSB database where training and test data match in terms of an object pose and scale.*



(b) *Results for single objects from SOC database (real stereo data) where training and test data differ in an object rotation. Experimental data protocol is illustrated in Fig. 4(a).*



(c) *Results for objects from 10 natural scenes in SOC database (real stereo data). Training and test data differ significantly in object an pose and scale. Experimental data protocol is illustrated in Fig. 4(b).*



(d) *Results for synthetic models collected form PSB database where only a single example per object instance is used for training (complete models) and multiple examples of objects in various poses are used for testing (incomplete models).*

Fig. 6. *Comparison of several state-of-the-art local and global 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$_{FPFH}$ - Fast Point Feature Histograms [26] encoded using the Bag-of-Words; GFPFH- Global Fast Point Feature Histogram [28]; VFH - Viewpoint Feature Histogram [27]; GSH$_{FPFH}$ - our Global Structure Histogram representation based on the FPFH local features.*

An altogether different approach is the one taken in [27], where the presented Viewpoint Feature Histogram (VFH) descriptor directly encodes information about both object shape and a camera viewpoint in the feature vector. We will present comparison of the VFH and our descriptor on data that differ in quality and amount of available training examples in the next section.

3) Four Scenarios – Towards Real Word Conditions: The choice of an object representation is crucial for achieving a robust categorization system. Ideally, the object representation should have high discrimination and generalization power, be invariant to object pose and scale variations, sensor noise as well as occlusions and partial visibility of an object (incomplete views). Another important aspect is the ability of the algorithm to learn a 3D object model from a small number of examples.

We address these problems one at a time by gradually increasing the complexity of the experimental data. We start with a relatively simple scenario where object representations are tested on synthetic partial models. For this scenario, a comparably large amount of data is available for training. Then, we increase the difficulty of the problem by performing experiments for single objects from SOC database (real stereo data) in which object segmentation is imperfect and sensor noise is present. We also vary a rotation of object examples used for training and testing, as presented in Fig. 4(a). Next, in order to evaluate object representations in realistic conditions, models trained on data from the previous protocol are tested against objects from 10 natural scenes (SOC database) where an object pose and scale differ significantly, see Fig. 4(b). In the end, the representations are evaluated in the scenario where only a single example per object instance is used for training (complete models from PSB database) and multiple examples of objects in various poses are used for the evaluation (incomplete models from PSB database). It is the toughest scenario in which generalization properties of the representations can be well reflected. It is due to the fact that: (a) a small amount of data is used for training, and (b) the test data (incomplete models) contain a highly limited amount of information compared to the training data (complete models).

Hereafter, we present a final comparison of the local and global representations such as the BOW$_{FPFH}$, GSH and VFH, for the four described scenarios and present the most important conclusions in terms of usability in real applications.

(1) As presented in Fig. 3, adding structure information leads to a significantly more descriptive GSH representation than BOW. This observation is confirmed by the higher performance of the GSH for all four scenarios as it is shown in Fig. 6. Additionally, when comparing a number of clusters required to obtain an optimal categorization rate for the FPFH and GSH, we can see that a higher number of clusters is needed for local features. This is consistent with our expectations that a simpler local representation can be used when global information is encoded.
(2) In the top row of Fig. 6, we can see that the VFH feature is the best performing representation. However, for these

data all object poses used at test time were also available at training. This means that for these data the representation does not need to generalize over different views. Generating datasets containing every possible view is not realistic and will not scale when the number of objects increases.
To that end a more realistic experiment are the ones depicted in Fig. 6(c) where an object pose and scale vary across data used for training and testing. Further, in Fig. 6(d) we can see the results when training on a complete 3D model of an object and testing on partial views. In both these experiments the GSH feature outperforms the VFH indicating that the latter does not generalize over views to the same extent as the proposed representation. We showed that the GSH requires less training data to model the object without the need of generating a large number of redundant partial views for training.

In addition, please note that although a chosen mesh reconstruction method [24] does not provide an optimal solution and suffers from the quality of the data, the GSH descriptor is capable of handling resultant distance length variations and provides a stable representation. Using of a more accurate reconstruction method may open perspectives for the further improvement of a categorization rate.

Finally, the GSH representation allows for using a single framework to represent complete and incomplete views of an object, as shown for synthetic models in Fig 7. On the right side of each object, the resultant Global Shape Descriptor for $n_C = 7$ is illustrated as a matrix where each row represents a histogram of distances between points belonging to two clusters (description of a matrix can be found in the caption of Fig. 3). In Fig. 8 we can observe that for real noisy data, objects with similar types of surfaces such as a *bottle* and *toilet paper* can be discriminated thanks to encoding of global structure.

VI. CONCLUSIONS

We have presented the Global Structure Histogram (GSH) descriptor for object representations using 3D sensory data. It is used to represent both complete and incomplete (partial) point cloud information. We have shown that the descriptor significantly improves object category classification compared to the state-of-the-art in realistic scenarios. Exploiting object structure allows us to achieve significantly better results from a less discriminative local features. This is beneficial as it makes object recognition less sensitive to small differences in the local appearance.

Simultaneous encoding of an object category and pose, such as in [27], [39], suffers from the problem of scaling over increasing number of object poses and classes. Our representation is capable to improve generalization over various object poses and scales in relation to object category. This is essential for large scale object categorization in real world applications. Moreover, the GSH opens the possibility for modeling object properties based on a small amount of training examples. This is currently under investigation in our work.

REFERENCES

[1] *Princeton Shape Benchmark*. http://shape.cs.princeton.edu/benchmark, Last visited: Nov 2011.
[2] The PASCAL Visual Object Classes Challenge 2010 (VOC2010).
[3] 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 *ICCV: 3dRR Workshop*, 2011.
[4] G. Biegelbauer and M. Vincze. Efficient 3D object detection by fitting superquadrics to range image data for robot's object manipulation. In *ICRA*, 2007.
[5] M. Bjorkman and D. Kragic. Active 3D scene segmentation and detection of unknown objects. In *ICRA*, May 2010.
[6] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM*, 2011.
[7] F. Chierichetti, R. Kumar, S. Pandey, and S. Vassilvitskii. Finding the jaccard median. In *ACM-SIAM*, 2010.
[8] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV: Statistical Learning in Computer Vision Workshop*, 2004.
[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.
[10] R. Detry, N. Pugeault, and J. Piater. A probabilistic framework for 3D visual object representation. *TPAMI*, 2009.
[11] C. H. Ek and D. Kragic. The importance of structure. *ISRR*, 2011.
[12] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003.
[13] R. W. Floyd. Algorithm 97. In *Comm. ACM*, 1962.
[14] D. Forsyth and J. Ponce. *Computer vision: a modern approach*. Prentice Hall Professional Technical Reference, 2002.
[15] A. Frome, D. Huber, R. Kolluri, T. Bulow, and J. Malik. Recognizing objects in range data using regional point descriptors. In *ECCV*, 2004.
[16] N. Iyer, S. Jayanti, K. Lou, Y. Kalyanaraman, and K. Ramani. Three-dimensional shape searching: state-of-the-art review and future trends. *CAD*, 2005.
[17] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *TPAMI*, 1999.
[18] S. Katz, A. Tal, and R. Basri. Direct visibility of point sets. In *SIGGRAPH*, 2007.
[19] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.
[20] J. Liebelt, C. Schmid, and K. Schertler. Viewpoint-independent object class detection using 3D feature maps. In *CVPR*, 2008.
[21] G. D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.
[22] M. Madry, D. S. Song, and D. Kragic. From Object Categories to Grasp Transfer Using Probabilistic Reasoning. In *ICRA*, 2012.
[23] Z.-C. Marton, D. Pangercic, N. Blodow, J. Kleinehellefort, and M. Beetz. General 3D modelling of novel objects from a single view. In *IROS*, 2010.
[24] Z. C. Marton, R. B. Rusu, and M. Beetz. On Fast Surface Reconstruction Methods for Large and Noisy Datasets. In *ICRA*, 2009.
[25] M. Potamias, F. Bonchi, C. Castillo, and A. Gionis. Fast shortest path distance estimation in large networks. In *CIKM*, 2009.
[26] R. B. Rusu, N. Blodow, and M. Beetz. Fast Point Feature Histograms (FPFH) for 3D Registration. In *ICRA*, May 2009.
[27] R. B. Rusu, G. Bradski, R. Thibaux, and J. Hsu. Fast 3D recognition and pose using the Viewpoint Feature Histogram. In *IROS*, 2010.
[28] R. B. Rusu, A. Holzbach, G. Bradski, and M. Beetz. Detecting and segmenting objects for mobile manipulation. In *S3DV*, 2009.
[29] S. Savarese and L. Fei-Fei. View synthesis for recognizing unseen poses of object classes. In *ECCV*, 2008.
[30] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *PAMI*, 19(5):530–535, 1997.
[31] F. Solina and R. Bajcsy. Recovery of parametric models from range images: The case for superquadrics with global deformations. *TPAMI*, 1990.
[32] B. Steder, G. Grisetti, M. Van Loock, and W. Burgard. Robust on-line model-based object detection from range images. In *IROS*, 2009.
[33] H. Su, M. Sun, L. Fei-Fei, and S. Savarese. Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. In *ICCV*, 2009.
[34] M. Sun, H. Su, S. Savarese, and L. Fei-Fei. A multi-view probabilistic model for 3d object classes. In *CVPR*, 2009.

Fig. 7. *Examples of synthetic object point clouds from the Princeton Shape Benchmark. Data are divided into $n_C = 7$ clusters, marked by different colors. Results are shown for six different object instances from a "bottle" category (PSB database), for both the complete (top row) and incomplete views (bottom rows). On the right side of each object, the resultant Global Structure Histogram is illustrated as a matrix described in the caption of Fig. 3. Images are best viewed in color.*

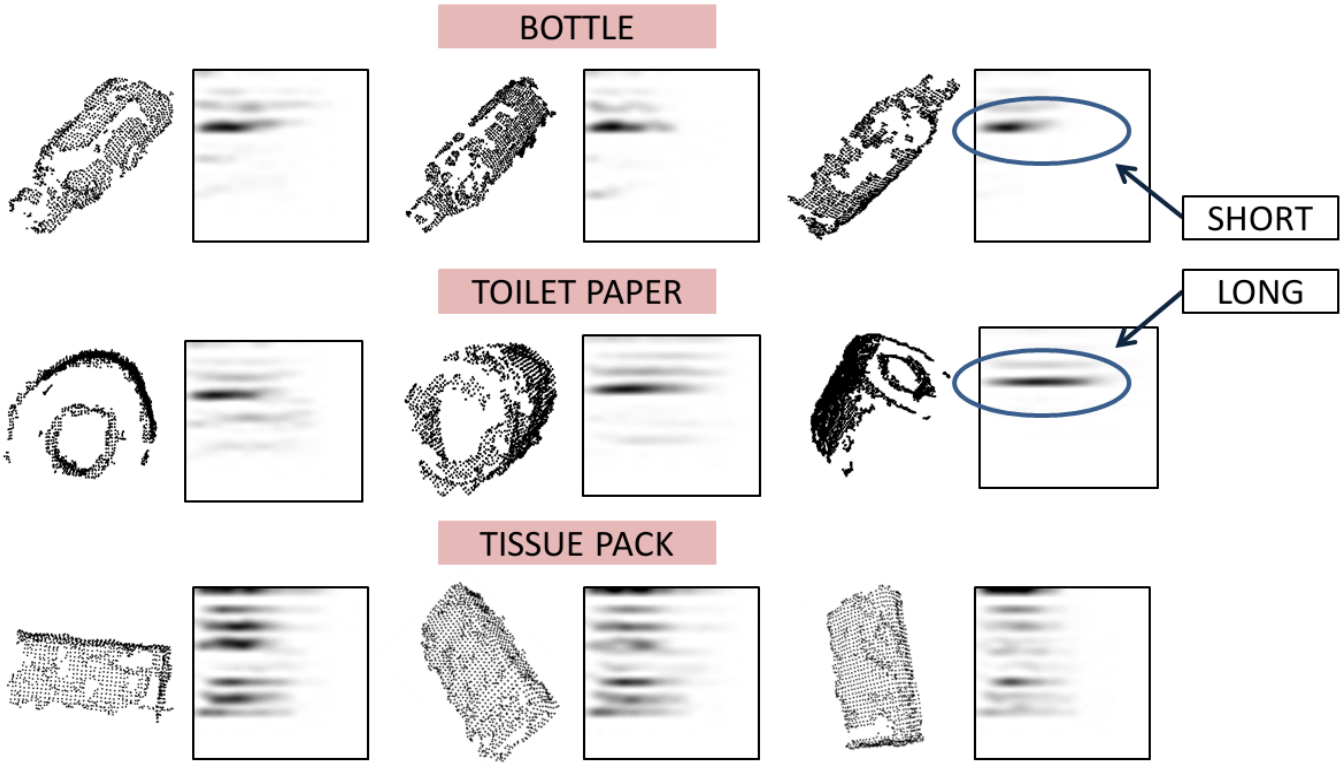


Fig. 8. *Examples of real object point clouds from the Stereo Object Category database. On the right side of each object, the resultant Global Structure Histogram for $n_C = 7$ is illustrated as a matrix described in the caption of Fig. 3. We can observe that the objects of relatively similar types of surfaces such as a "bottle" and "toilet paper" can be discriminate due to encoding global structure of an object.*

[35] J. W. H. Tangelder and R. C. Veltkamp. A survey of content based 3D shape retrieval methods. In *SMI*, 2004.

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

[37] D. V. Vranic and D. Saupe. 3D shape descriptor based on 3D fourier transform. In *ECMCS*, 2001.

[38] M. Weber, M. Welling, and P. Perona. Towards automatic discovery of object categories. In *CVPR*, volume 2, pages 101–108, 2000.

[39] W. Wohlkinger and M. Vincze. Shape-based depth image to 3D model matching and classification with inter-view similarity. In *IROS*, 2011.

[40] R. A. Zlatanova, S. and W. Shi. Topological models and frameworks for 3D spatial objects. *CG*, 2004.