Object Categorization from Range Images using a Hierarchical Compositional Representation

Vladislav Kramarev
School of Computer Science
University of Birmingham
vvk201@cs.bham.ac.uk

Sebastian Zurek
School of Computer Science
University of Birmingham
s.j.zurek@cs.bham.ac.uk

Jeremy L. Wyatt
School of Computer Science
University of Birmingham
j.l.wyatt@cs.bham.ac.uk

Aleš Leonardis
School of Computer Science
University of Birmingham
a.leonardis@cs.bham.ac.uk

***Abstract*—This paper proposes a novel hierarchical compositional representation of 3D shape that can accommodate a large number of object categories and enables efficient learning and inference. The hierarchy starts with simple pre-defined parts on the first layer, after which subsequent layers are learned recursively by taking the most statistically significant compositions of parts from the previous layer. Our representation is able to scale because of its very economical use of memory and because subparts of the representation are shared. We apply our representation to 3D multi-class object categorization. Object categories are represented by histograms of compositional parts, which are then used as inputs to an SVM classifier. We present results for two datasets, Aim@Shape [1] and the Washington RGB-D Object Dataset [2], and demonstrate the competitive performance of our method.**

***Keywords*—3D object representation, 3D object categorization, compositional hierarchy, classification.**

I. Introduction

Reliable object recognition and categorization has been one of the central topics addressed by the computer vision community over decades. The methods based on visual words are widely used to solve 3D object categorization and shape retrieval problems. Some authors, for example Toldo et al. [3] and Fehr et al. [4], use a Bag-of-Words (BoW) strategy where an object is represented by a set of local features. Others, e.g. Madry et al. [5], also introduce data structures describing spatial relations of local features.

Compositional hierarchies have recently become a popular topic in computer vision. Principles of *hierarchical compositionality* allow one to develop generalizable category representation and recognition frameworks, where new categories can be added efficiently to the system. However, most of the recent advances in this area have been focused on hierarchies of 2D features [6][7][8][9][10][11]. Very little work has been done so far to address the formidable problem of true 3D categorization using compositional hierarchical approaches.

In this paper we shed some light on this problem and propose a hierarchical compositional representation of 3D shapes that is a recursive compositional vocabulary of surface parts represented by a directed graph [7] (see Figure 1).

The first layer L_1 of the hierarchy contains several pre-defined parts. All the parts from the above layers are learned and represent the most statistically significant compositions of several simpler shape parts from bottom layers.

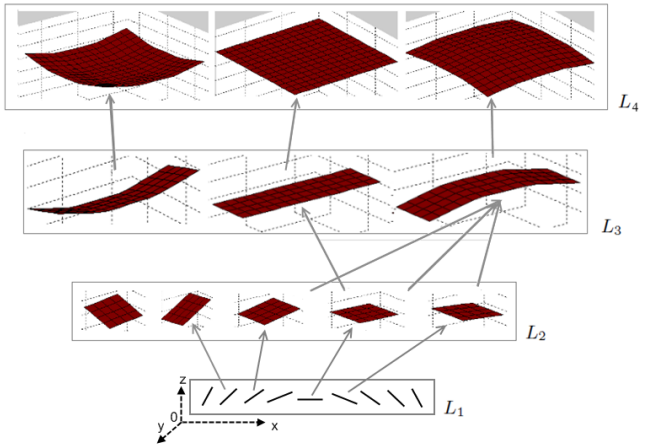

Fig. 1. 3D compositional hierarchy of parts.

In order to examine the learned layers of the compositional hierarchy, we introduce a new 3D object categorization method which is based on histograms of compositional parts. Each object category is represented by histograms reflecting the spatial distribution of the compositional parts that describe the object's surface. We employ an SVM classifer with χ^2 kernels for categorization. We tested our method on the Aim@Shape dataset [1] containing 20 object categories and achieved 95.6% success rate for categorization. We also obtained promising results for the larger Washington RGB-D Object Dataset [2].

The rest of the paper is organized as follows. In Section II we discuss related work. In Section III we give a detailed description of our method. Section IV describes our experiments and results. Section V concludes the paper.

II. Related work

The proposed method is related to the works of Fidler et al. [6][7][8]. They have introduced a framework for learning a hierarchical compositional shape vocabulary for multi-class object representation. Each part in the hierarchy is composed of *less complex* parts according to statistical properties of their spatial configurations. At each layer, parts are recursively combined into more complex compositions, each exerting a high degree of shape variability. At the top layer of the hierarchical vocabulary, the compositions are sufficiently complex

Preprint submitted to 22nd International Conference on Pattern Recognition.
Received December 21, 2013.

to represent the shape of a whole object. The main difference between our proposed work and Fidler et al. is that they provided a mechanism to learn a 2D shape vocabulary (contour parts), while our proposed method represents 3D shapes of objects in a compositional hierarchy.

Savarese et al. [12] introduced a hierarchical framework for 3D object categorization and recognition. They extracted local features from the images and grouped them into relatively large discriminative regions (called parts) that are pulled together to form a 3D category model. Whilst our compositional hierarchy models 3D shape independently of 2D image context, Savarese et al. built a hierarchy grouping both 2D local features and 3D shape features This precludes their approach for applications where only 3D data is provided (e.g. Kinect data, or haptic data in robotics applications).

Pratikakis et al. [13] proposed a 3D compositional model in which point clouds are decomposed into sections that are represented by a predefined set of primitives, e.g. cone, torus, sphere or cylinder. Their method has a limitation in that it deals with the simplest shapes (mainly hand-made objects) and hence is not suitable for general multi-class category detection.

Detry et al. [14] proposed a hierarchical object representation framework that encodes probabilistic spatial relations between 3D features using Markov networks. Features extracted at the base layer of the hierarchy are bound to local 3D descriptors. Higher-level features recursively encode probabilistic spatial configurations of the features obtained from previous layers. However, their approach does not involve statistical learning of a single 3D shape vocabulary that is shared by objects of different categories.

Recently Fox and colleagues have published a series of papers in which they introduced several algorithms for object classification (at both category and instance level) and evaluated these on their RGB-D image dataset. In [2], after partitioning the depth image within a 3D bounding box, they computed spin image [15] histograms that were used to form *efficient match kernel* (EMK) features. After dimensionality reduction with PCA, these features (around 2700) were used to train a classifier, such as a gaussian-kernel SVM. In subsequent work Lai et al. [16] developed a new classifier, based on the *instance distance learning* (IDL) technique and data sparsification, that was able to improve categorization performance. By using kernel descriptors and *hierarchical matching pursuit* to build feature hierarchies, further gains in categorization accuracy were achieved [17] [18]. For comparison with our method, we show their results for object category recogntion from range images in Table II.

III. Compositional Hierarchy of 3D Parts

In this section we describe our compositional hierarchical representation of 3D object shape, how the representation is learned and how to perform inference using it.

A. Representation

We define our coordinate system such that the x and y axes span the image, and the z-axis encodes depth information. We define a hierarchy of layers, where L_n denotes the n-th layer of the hierarchy. The first layer L_1 of the hierarchy contains several pre-defined, rather than learned, features or parts. First layer parts encode quantized differences of depth (relative depth) between pixels at a fixed distance from each other in the x-axis direction. Figure 2 shows one way in which these parts can be defined. In this case we quantize all possible values of relative depth into nine bins. However, the number of first layer parts can be chosen differently depending on the type of input data and required precision of the representation.

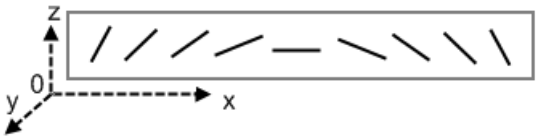

Fig. 2. Parts of the first layer.

Figure 3c shows the range image of the mug and demonstrates how the range data can be encoded in terms of the pre-defined first layer parts. In Figure 3d locations of parts depicted in Figure 3a are represented using color coding given in Figure 3b.

Fig. 3. (a) Pre-defined first layer parts. (b): Color coding of the first layer parts. (c): Range image of the mug. (d): Encoding of the mug with first layer parts.

In general, the higher layers L_n, $\forall n > 1$, are learned using joint statistical properties of parts from the layer below. Each part P_i^n in L_n is a composition of subparts, that is a list of subparts and a description of the spatial relations between these constituent subparts. We say that a composition P_i^n consists of a central part $P_{central}^{n-1}$ and other subparts that reside at some positions relative to $P_{central}^{n-1}$:

$$P_i^n \equiv (P_{central}^{n-1}, \{P_j^{n-1}, \mu_j, \Sigma_j\}_j) \quad (1)$$

where $\mu_j = (x_j, y_j, z_j)$ is the mean relative position of the subpart P_j^{n-1}, and Σ_j is the covariance matrix expressing the variability of possible relative positions. In this paper, we specialize this scheme by assuming that all compositions consist of three subparts.

Second layer parts can be regarded as very small surface patches, that are constructed out of three L_1 parts. Figure 4 sketches the construction of second layer parts (in the general scheme), and Figure 5 illustrates several examples of learned parts.

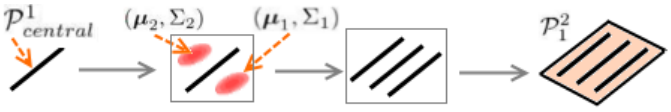

Fig. 4. Construction of second layer parts.

Preprint submitted to 22nd International Conference on Pattern Recognition.
Received December 21, 2013.

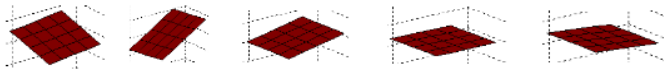

Fig. 5. Examples of second layer parts.

Third layer parts are assembled from triplets of L_2 parts, which are adjacent along the x-direction. Similarly fourth layer parts consist of three adjacent L_3 parts aligned vertically (i.e. along the y-direction).

In this and future work we intend to demonstrate that the representation proposed above has the following desirable properties:

Efficient use of memory: In current state-of-the-art 3D object categorization, objects are mainly represented by salient surface patches or discriminative local features such as spin images [15], or by statistical moments. Very large numbers of patches and local features must be collected and stored in order to achieve the best results for multi-class categorization. In compositional hierarchical approaches, parts are shared throughout the hierarchy. More complex parts are described in terms of simpler parts from the previous layers, therefore a simpler part at one layer can be used to describe many parts in higher layers. This re-usability results in a very compact representation of the vocabulary.

Unsupervised learning: Compositional parts in the hierarchy are learned in an unsupervised manner. This learned "vocabulary of parts" captures and compactly represents the most statistically relevant regularities in the dataset.

Fast, incremental learning: The proposed method enables new object categories to be learned efficiently, i.e. with less computational complexity than batch schemes. Moreover its efficiency increases with the amount of data already learned by the system. In fact, new objects or object categories can be added to the representation by simply pulling together a small number of appropriate parts.

B. Learning a vocabulary of parts

The goal of the learning procedure is to construct compositions of parts that encode the most statistically significant spatial relations between parts of the layer below. The collection of compositions from all layers in a trained compositional hierarchy is termed a vocabulary. In general, each composition has to be flexible in that it should tolerate some variability in the relative spatial position of elements.

The learning process for each layer L_n, $\forall n > 1$, can be summarized in four steps:

1) Perform *local inhibition* in the neighborhood of each part.
2) Construct statistical maps that characterize the 3D spatial relations between parts of the previous layer.
3) Produce a list of *candidate parts* by constructing compositions based on the statistical maps.
4) Optimize the list of candidate parts to form the vocabulary, i.e. select a subset of parts that satisfies some optimality criterion

We now describe each step in more detail.

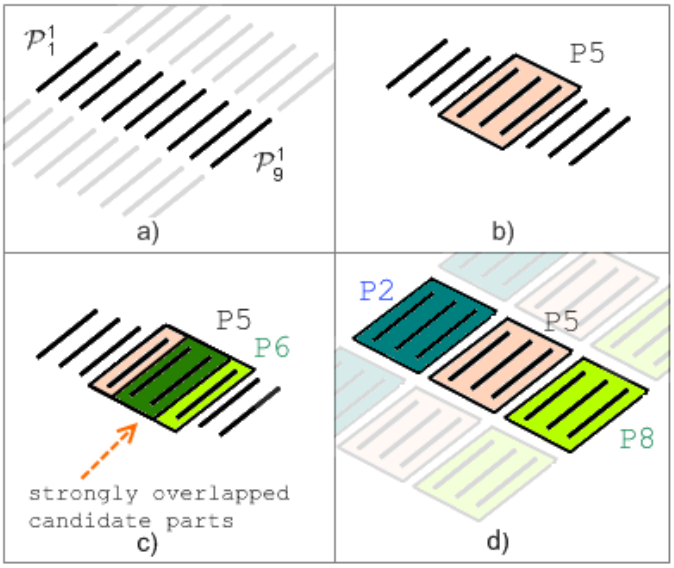

Fig. 6. Implementation of local inhibition: a) Detected parts of layer L_1 lying on a surface; b) Derived part P_1^2 of the L_2 layer; c) Other L_2 parts that have intersection with P_1^2 are to be removed (e.g. part P_4^2); d) Surface patch covered by L_2 parts after performing local inhibition.

The important first step is local inhibition which helps to avoid unnecessary redundancy in coding. Assume that we are given a range image that is encoded in terms of parts $\{P_j^n\}$ at layer L_n. For each part P_k^n, we remove the parts that reside in a (small) neighborhood of P_k^n and have a large intersection with P_k^n in terms of L_{n-1} parts.

This step can be considered as the removal of those local surface features that are already partially encoded by P_k^n. The procedure is illustrated in Figure 6 for L_2 parts.

Next, we construct statistical maps that describe relative positions of parts in 3D space. The maps for layer L_n are functions f:

$$f(P_i^{n-1}, P_j^{n-1}, x, y, z) \rightarrow [0, 1] \qquad (2)$$

that are defined for each pair of elements P_i^{n-1} and P_j^{n-1} in layer L_{n-1}, and a 3D offset $(x, y, z) \in \mathbb{R}^3$. The maps encode the probability of observing a part P_j^{n-1} displaced by (x, y, z) relative to a central part P_i^{n-1}. A natural way to visualize the collected co-occurrence statistics is to project the 5-dimensional function f into 3 dimensions by fixing the first and second parameters.

After the co-occurrence statistics of parts are computed, we detect peaks in the spatial maps, and fit the data in surrounding regions by a Gaussian distribution with mean μ_j and covariance matrix Σ_j. Figure 7 shows an example of such fitting of the statistical map depicting co-occurrences of the second layer parts P_{41}^2 and P_{42}^2.

Parts P_i^n of the layer L_n are constructed from the previous layer L_{n-1} using μ_j and Σ_j as shown in (1). This procedure is implemented in two steps. First, we construct pairs, i.e. elements comprising two parts from the previous layer. Next, we group them into triples encoding spatial relations of three parts from the previous layer. Triples become *candidate parts*

Preprint submitted to 22nd International Conference on Pattern Recognition.
Received December 21, 2013.

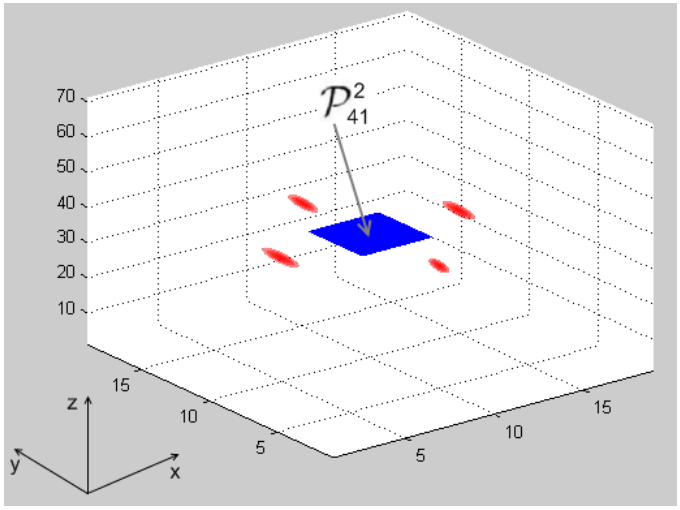

Fig. 7. Statistical map depicting co-occurrences of parts P^2_{41} and P^2_{42}. The size of the local neighborhood was chosen to be $17 \times 17 \times 70$.

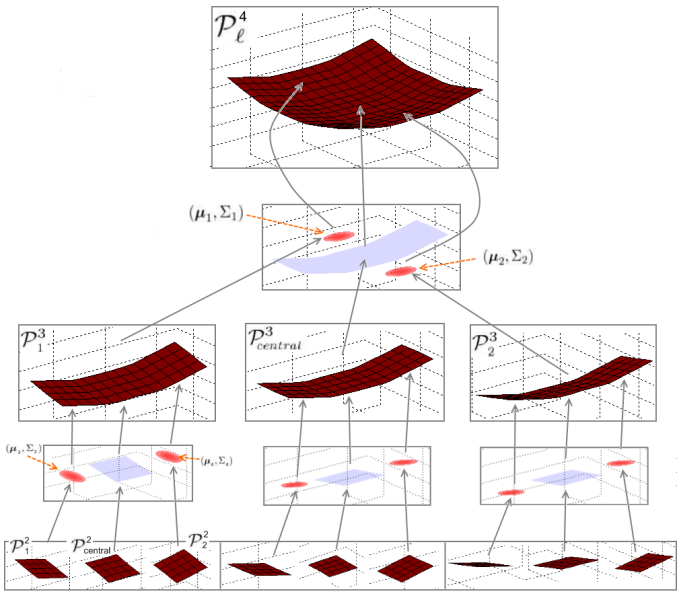

Fig. 8. General learning scheme for layers L_3 and L_4 in the compositional hierarchy.

that will reside in layer L_n. Figure 8 depicts how triples are formed for the third and fourth layers.

Our experiments have shown that, for the first layers of the hierarchy, steps 2 and 3 in the above algorithm can be approximated. It is sufficient to consider only the z component that has a predefined spatial relationship in the x and y directions defined by the object size. Hence, for the experiments in this paper, we assumed that parts (in a given layer) had a predefined spatial relationship in the x and y directions, and we collected only a quantized z component of the 3D offset. This simplification did not significantly affect the part selection process and therefore the categorization accuracy. However, it yielded a significant improvement in terms of processing time. We note that for learning layers beyond L_4 this simplification may not be suitable, as more complex parts may have more complex spatial configurations.

Typically the set of candidate parts $S = \{P^n_i : i = 1..N\}$ for the given layer L_n is rather large, and contains many parts that represent very similar surface types. In order to maintain a manageable number of parts in the vocabulary and to facilitate generalization, we specify a procedure that selects a somewhat smaller subset $S' \subseteq S$. This selection is performed by approximately solving the following optimization problem. The cost function E which is minimized measures the reconstruction error, i.e. how well the set of candidate parts can be represented by the vocabulary. We also include a term that penalizes vocabularies with more parts, so that the cost E takes the following form:

$$E(S') = \min_{S' \in S} \sum_{i=1}^{N} d(P_i, P'(P_i))\, \nu_i \; + \; \alpha\, |S'| \qquad (3)$$

where ν_i is the frequency of occurrence of the i-th candidate part P^n_i, $d(\cdot,\cdot)$ is a distance function that quantifies the similarity between two parts (from the same layer), and $P'(P_i)$ is the part in S' that is closest to P_i. Also $\alpha \in \mathbb{R}^+$ is a meta-parameter that regulates the trade-off between precision of the representation and number of selected parts. In addition, we have explored adding further penalty terms to influence part selection according to the geometric properties of parts.

C. Inference

This section describes the inference process that generates features from a range image, using a given vocabulary. These features can then be used for object category (or instance) recognition.

Our method performs part detection layer by layer, starting from the first layer. Assume we are given a range image of an object (or scene), where each pixel value encodes depth. The goal is to represent the object in terms of parts from the compositional hierarchical shape vocabulary.

The first stage is to represent the object in terms of first layer parts. We convolve an oriented Gaussian-derivative filter (aligned along the x-axis) with the range image. The variance parameter σ associated with the filter depends on the noise level of the images and was chosen from within the range [0.5,2.0].

The next stage is to quantize the filter response at each pixel by assignment to the bin that corresponds to the closest first layer part. A reconstruction error E_i is computed as a distance to the closest bin center divided by the size of the bin. This procedure gives us a set of *potential parts* $S_{pot} = \{P_1, P_2, ..., P_m\}$, that can be detected at certain locations with corresponding reconstruction errors $E_1, E_2, ..., E_m$, where m is the number of detected potential parts.

However, such a strategy leads to a redundant representation, as all the detected potential parts are significantly overlapped with each other. To proceed, we specify several criteria that our inference process should jointly optimize:

1) Maximize the surface coverage. Ideally the entire object surface should be covered by parts from the vocabulary.
2) Minimize the overlaps between detected parts.

Preprint submitted to 22nd International Conference on Pattern Recognition.
Received December 21, 2013.

3) Minimize the reconstruction error.

To fulfil all the above requirements we have to select a subset S_{sel} of potential parts S_{pot}. Following e.g. Leonardis et al. [19] we define an energy function that incorporates all three criteria, and then solve the associated optimization problem.

When part detection for the first layer is completed, we do inference at subsequent layers L_n, $\forall n > 1$, performing essentially the same procedure. Suppose we have a range image represented in terms of parts at layer L_{n-1}. Then the inference algorithm for parts at layer L_n can be described as follows:

1) Consider a local neighborhood around each part $\{P_i^{n-1}\}$. For this neighborhood, the part $\{P_i^{n-1}\}$ is referred to as a *central part*.
2) Extract parts located in this neighborhood and their relative positions with respect to the central part ($\{P_i^{n-1}\}$). The central part together with other neighboring parts can be represented as a *potential compositional part* $\{P_{pot}^n\}$ of layer L_n, as described in equation (1).
3) This potential compositional part is matched against vocabulary elements of layer L_n, and if found yields a detection. The matching process can be implemented in a very efficient manner.
4) This procedure leads to a redundant representation, since we attempt to detect a layer L_n part in all positions of detected L_{n-1} parts. Since parts in higher layers are always larger, the potential parts will overlap.
5) We eliminate parts to minimize the reconstruction error, maximize coverage and minimize overlap using an optimization function similar to that described for the first layer.

IV. Experiments

A. Method

In order to evaluate the compositional hierarchical representation we constructed a classifier to perform category-level object recognition from range images. Given a dataset of these images we learn a vocabulary, layer by layer up to L_4. Then we perform multi-class object categorization using histograms of compositional parts, obtained from a training subset of the dataset. Each range image is partitioned into 4 (2×2) and 9 sectors (3×3 – see Figure 9) which together with the original image comprise 14 subimages from which histograms of parts are computed. The histograms are stacked to form a large descriptor which is used as the input vector for each image to a χ^2 kernel SVM classifier.

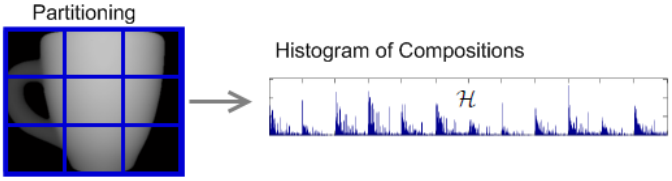

Fig. 9. Partitioning of the object to build a histogram of compositional parts.

We applied this evaluation method to two benchmark datasets: Aim@Shape [1], and the Washington RGB-D Object Dataset [2].

B. Evaluation on Aim@Shape Dataset

From the Aim@Shape dataset we rendered a set of range images presenting all the 3D models at different scales and under different viewing angles. Since the first layers of the hierarchy contain very generic parts which are shared by many categories, only 50-100 of randomly selected range images were required to learn the vocabulary of the second layer, and 200-300 images to learn the third layer.

To be able to compare our approach with other methods we used leave-one-out cross-validation to measure performance. In this experiment we used eight viewing angles and three scales per model to train the SVM classifer.

Figure 10 shows how using features from more layers improves performance.

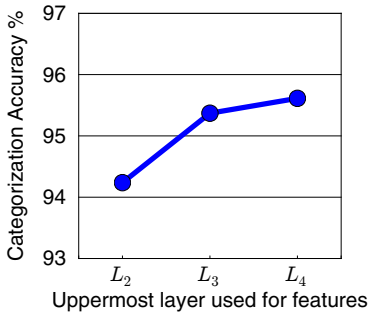

Fig. 10. For the Aim@Shape dataset, the categorization accuracy improves as more layers are used to provide features to the classifier.

In table I we see that the categorization accuracy of our method is comparable with the state-of-the-art, when using features obtained from all layers up to L_4 in the vocabulary.

TABLE I. Results for Aim@Shape Dataset

Method	Accuracy %
Toldo et al. [3]	87.3
Salti et al. [20] using 1-NN for codebooks	79
Salti et al. [20] using 2-NN for codebooks	100
This work (up to L_4)	95.6

C. Evaluation on Washington RGB-D Object Dataset

For the Washington RGB-D Object Dataset we learned a vocabulary up to L_4 from around 2,000 images selected randomly from the whole dataset of about 250,000 images. A remarkable fact is that our shape vocabulary was stored in less than 50kB of memory, demonstrating the memory efficiency of our approach.

To train the SVM classifier, we used only 10% of the available training data. As in [2] we estimated performance with leave-one-out cross-validation.

Table II shows that our method improves upon the earlier work of Lai et al. [2][16] and compares favourably with the accuracies of individual depth kernel descriptors (detailed in [17]).

Preprint submitted to 22nd International Conference on Pattern Recognition.
Received December 21, 2013.

TABLE II. RESULTS FOR RGB-D OBJECT DATASET (USING RANGE IMAGES ONLY)

Method	Accuracy %
Spin Images & 3D Bounding Boxes [2]	64.7
Sparse Distance Learning [16]	70.2
RGB-D Kernel Descriptors [17]	80.3
Hierarchical Matching Pursuit [18]	81.2
This work (up to L_2)	72.7
This work (up to L_3)	73.8

V. CONCLUSION AND FUTURE WORK

We have presented a 3D learning and recognition framework built on the principle of hierarchical compositionality. The framework accommodates a large number of object categories, and since parts are shared, the size of the representation grows logarithmically with the number of learned object categories. The framework provides mechanisms for *transfer of knowledge* that enables its use in a variety of computer vision and robotics applications, such as object grasping and manipulation.

Thus far we have examined learning for the first four layers of the hierarchy and applied our method to multi-class object categorization with promising results. In future we plan to learn further layers of the compositional hierarchy and to test our method on other 3D categorization and shape retrieval datasets. In particular we intend to investigate how the size of representation changes with the number of object categories. Given the very small memory footprint, the representation may be particularly suited for mobile phone applications.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of EU-FP7-IST grant 600918 (PaCMan). The authors would also like to thank Mete Özay for helpful discussions.

REFERENCES

[1] R. C. Veltkamp and F. B. ter Haar, "SHREC2007: 3D Shape Retrieval Contest," Utrecht University, Tech. Rep. UU-CS-2007-015, 2007.

[2] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," in *IEEE International Conference on Robotics and Automation, ICRA*, 2011, pp. 1817–1824.

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

[4] J. . Fehr, A. Streicher, and H. Burkhardt, "A bag of features approach for 3d shape retrieval," in *Advances in Visual Computing*. Springer, 2009, pp. 34–43.

[5] M. Madry, C. H. Ek, R. Detry, K. Hang, and D. Kragic, "Improving generalization for 3d object categorization with global structure histograms," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012*. IEEE, 2012, pp. 1379–1386.

[6] S. Fidler and A. Leonardis, "Towards scalable representations of object categories: Learning a hierarchy of parts," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[7] S. Fidler, M. Boben, and A. Leonardis, "Learning hierarchical compositional representations of object structure," in *Object Categorization: Computer and Human Vision Perspectives*, S. Dickinson, A. Leonardis, B. Schiele, and M. Tarr, Eds. Cambridge University Press, 2009. [Online]. Available: vicos.fri.uni-lj.si/data/alesl/chapterLeonardis.pdf

[8] ——, "Optimization framework for learning a hierarchical shape vocabulary for object class detection." in *BMVC*, 2009, pp. 1–12.

[9] S. C. Zhu and D. Mumford, *A stochastic grammar of images*. Now Publishers Inc, 2007, vol. 2, no. 4.

[10] B. Ommer and J. M. Buhmann, "Learning the compositional nature of visual objects," in *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07.* IEEE, 2007, pp. 1–8.

[11] L. L. Zhu, C. Lin, H. Huang, Y. Chen, and A. Yuille, "Unsupervised structure learning: Hierarchical recursive composition, suspicious coincidence and competitive exclusion," in *Computer Vision–ECCV 2008*. Springer, 2008, pp. 759–773.

[12] S. Savarese and L. Fei-Fei, "3d generic object categorization, localization and pose estimation," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[13] I. Pratikakis, M. Spagnuolo, T. Theoharis, and R. Veltkamp, "Learning the compositional structure of man-made objects for 3d shape retrieval," in *Eurographics Workshop on 3D Object Retrieval (2010)*, 2010.

[14] R. Detry, N. Pugeault, and J. Piater, "A probabilistic framework for 3d visual object representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1790–1803, 2009.

[15] A. E. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3D scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 5, pp. 433–449, 1999.

[16] K. Lai, L. Bo, X. Ren, and D. Fox, "Sparse distance learning for object recognition combining RGB and depth information," in *IEEE International Conference on Robotics and Automation, ICRA*, 2011, pp. 4007–4013.

[17] L. Bo, X. Ren, and D. Fox, "Depth kernel descriptors for object recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, 2011, pp. 821–826.

[18] ——, "Unsupervised feature learning for RGB-D based object recognition," in *Experimental Robotics - The 13th International Symposium on Experimental Robotics, ISER*, 2012, pp. 387–402.

[19] A. Leonardis, H. Bischof, and J. Maver, "Multiple eigenspaces," *Pattern Recognition*, vol. 35, no. 11, pp. 2613–2627, 2002.

[20] S. Salti, F. Tombari, and L. D. Stefano, "On the use of implicit shape models for recognition of object categories in 3d data," in *Computer Vision–ACCV 2010*. Springer, 2011, pp. 653–666.

Preprint submitted to 22nd International Conference on Pattern Recognition.
Received December 21, 2013.