Categorisation of 3D Objects in Range Images Using Compositional Hierarchies of Parts Based on MDL and Entropy Selection Criteria

Vladislav Kramarev, Krzysztof Walas, and Aleš Leonardis

Intelligent Robotics Laboratory, School of Computer Science,
University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom
{vvk201,walask,a.leonardis}@birmingham.ac.uk
http://www.cs.bham.ac.uk/research/groupings/robotics/

Abstract. This paper presents a new approach to object categorisation in range images using our novel hierarchical compositional representation of surfaces. The atomic elements at the bottom layer of the hierarchy encode quantized relative depth of pixels in a local neighbourhood. Subsequent layers are formed in the recursive manner, each higher layer is statistically learnt on the layer below via a growing receptive field. In this paper we mainly focus on the *part selection problem*, i.e. the choice of the optimisation criteria which provide the information on which parts should be promoted to the higher layer of the hierarchy. Namely, two methods based on *Minimum Description Length* and category based *entropy* are introduced.
The proposed approach was extensively tested on two widely-used datasets for object categorisation with results that are of the same quality as the best results achieved for those datasets.

Keywords: range images, object categorisation, compositional hierarchies, shape parts

1 Introduction

In recent years, the processing of range images has experienced a renaissance in the computer vision community. The topic was popular in the 1980s and early 1990s [3, 5], but attention waned until the advent of affordable range cameras. This has led to two approaches: i) ideas from the 1990s have been re-tested using new hardware; ii) recent techniques for intensity images have been adapted to range images.

This work aims at improved representations of 3D shape for the purposes of categorisation. Based on the successful implementation of compositional hierarchies [6, 16] and their use in categorisation tasks on intensity images, a proper hierarchical representation is sought for range data. However, this is not a trivial task as intensity and depth data are different in their properties. For example, the notion of scale variance for intensity images is associated with a projection from the 3D world onto a 2D image plane, whereas range images provide metric

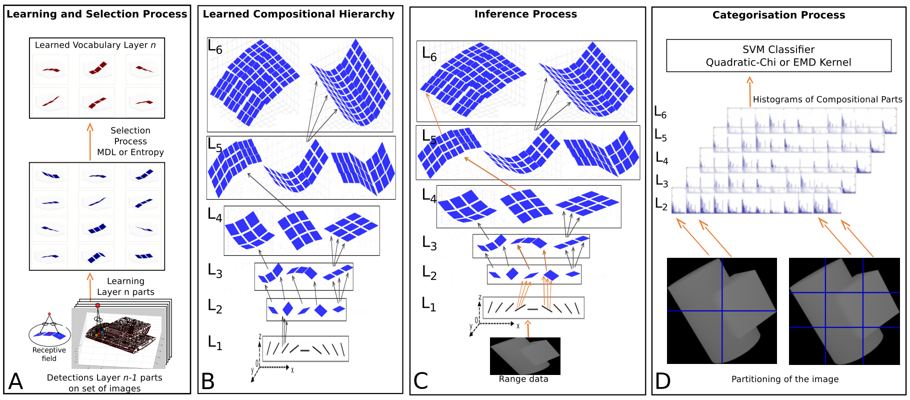

Fig. 1. Overview of the approach. Learning process (A) – a set of depth images is fed into the system. Detections of possible compositions for the next layer are found. Candidate parts undergo a selection process. A vocabulary for the selected layer is provided. Based on the learnt compositional hierarchy (B) an inference process is performed (C). The parts inferred for the specific object are fed into Histograms of Compositional Parts (D). Based on the discretised distribution of parts in partitioned images, categorisation of the object is performed.

data and the notion of scale variance is thus associated with the support size – the area for which a descriptor is computed.

An overview of the information flow in the system is shown in Fig. 1. The approach presented in this paper is based on computing small depth differences at the bottom layer. These atomic parts are then combined, using statistical learning (Fig. 1B), to form compositions at the second layer of the hierarchy. The compositions at each subsequent layer of the hierarchy are learnt from the previous layer. This representation is inherently surface based. In the learning process (Fig. 1A) a set of depth images is fed into the system, then detections of possible compositions are found. The rich set of candidate parts goes through a selection process that provides a constrained vocabulary for the selected layer. The compositions are promoted to the next layer if they fulfil the selected optimisation criterion. This work studies two selection criteria and their influence on object categorisation. The Learnt Compositional Hierarchy inference process, where specific nodes of the hierarchy are activated, is performed for an example object (Fig. 1C). Subsequently the active compositions are fed into a Histogram of Compositional Parts (Fig. 1D). Based on the discretised distribution of parts in the partitioned images, the object is categorised.

The paper begins with a description of related work in the field of categorisation and detection using range images and compositional representations. In section 2, the proposed compositional hierarchy for range images is presented. Section 3 provides a study of the part selection process. The theoretical description is then followed in sections 4 and 5 by an evaluation based on two widely available datasets [7, 13]. Finally, concluding remarks and plans for future work are given in section 6.

1.1 State of the Art

Most vision categorisation systems, either for intensity or depth images, are based on feature extraction and have three main sub-systems. i) keypoints (local extrema of a saliency measure) detector ; ii) keypoints descriptor; iii) feature matching algorithm, which performs the actual object recognition process. This survey examines all the aforementioned aspects of feature-based categorisation systems.

Keypoint Detection An extensive survey of keypoint detection algorithms for depth data was presented in [9]. The authors divided these techniques into two categories: Fixed-Scale and Adaptive-Scale detectors. An evaluation of different keypoint detection algorithms was presented in [22], and results on the repeatability of keypoints (under slight object transformations) were reported.

Local Shape Descriptors The detected keypoints should be equipped with a good descriptor, which maintains a high level of distinctiveness and separability between descriptors so as to make the matches between keypoints unique. A recent survey of local surface descriptors by Guo et al. [9] listed 38 different descriptor types and divided approaches into signature, histogram, and transform based methods.

Global Shape Descriptors In contrast to local shape descriptors, global descriptors are designed to describe a whole object with a single descriptor – see [2] for a survey. The four main classes of descriptor are histogram, transform, 2D view and graph based. The latest approach reported in [20] used covariance descriptors. Another method is to use a Hough Transform for 3D shape recognition [24].

Compositional Hierarchies Local shape descriptors are suitable for detection tasks as they are robust to clutter and viewpoint changes, whereas global shape descriptors are better for shape retrieval tasks – coping with intra-class variability. One concept to link these two worlds is that of compositional hierarchy.

Obtaining 3D object templates by hierarchical quantisation [15] is one way of building the compositional hierarchy for 3D data using a volumetric representation. An alternative approach is to use a Deep Neural Network that learns, in an unsupervised manner, the representation of the provided range data [18]. An approach using depth images and building the compositional representation from surface patches was described in our previous work [11]. In this paper we go beyond our previous results and provide a study of the mechanisms for establishing part importance, and hence how to obtain better classification performance.

Part Selection This process is understood here as an optimisation technique for selecting appropriate parts which are then combined in the subsequent layer. It is strongly dependent on the actual task; e.g., for shape retrieval the parts ought to be generic, whereas for pose estimation the parts are required to be discriminative. The problem of part importance for 2D recognition and reconstruction task was considered in [8], as well as the more general problem of finding interesting, salient points [1] or ranking 3D features [23]. Additionally, the problem of how to learn to detect repeatable parts in depth data was addressed in [10]. In the context of 3D compositional hierarchies, our previous work [11] used a frequency-based part selection strategy, i.e. the probability of including a candidate part in the vocabulary depended on the frequency of observing that part in the training data.

Histogram Similarity Measures. In our approach, object categories are represented as stacked histograms of compositional parts (see Fig. 1D). Histograms have widely been used for image classification, as shown in study [26]. χ^2 and histogram intersection are very popular histogram similarity measures. However, these measures fail to take into account cross-bin similarities, which are very important for compositional hierarchical approaches, since different compositional parts may represent similar surface types.

Cross-bin relations are tackled with Earth Mover's Distance (EMD), defined as a minimal cost that must be paid to transform one histogram into the other. Zamolotskikh et al. [25] show three ways in which EMD kernels may be used with an SVM classifier. The Quadratic-Chi histogram distance measure [17] not only takes into account cross-bin relations, but also performs normalisation such that a difference between small bins becomes as important as a difference of large ones.

1.2 Contributions

The main contribution of this paper lies in a successful conception, design and implementation of a learnt compositional hierarchy of surface parts, which are used for 3D object categorisation. The representation optimises two different part selection criteria which results in state-of-the-art categorisation performance on two publicly available datasets. We also tested three different affinity measures to produce kernel functions for SVM classification.

2 Hierarchical Compositional Representation of Surfaces

This section outlines the compositional hierarchical representation for depth data using surface patches. The description is divided into two parts. First, the learning process is described; second, the inference procedure is specified.

The coordinate system for the remainder of this paper is as follows: the x and y axes span the image plane and the z-axis encodes depth information.

The general scheme of the proposed compositional hierarchy is shown in Fig. 2d. The representation is defined as a hierarchy of layers L_n where n denotes the layer number. The vocabulary of each layer L_n contains a set of parts representing surface patches with different geometrical properties.

The first layer L_1 consists of pre-defined parts. Following [11] we use quantized differences of depth between pixels at a fixed distance from each other (along the x-axis). An example of the vocabulary of the first layer parts is depicted in Fig. 2a, while an example of the surface, represented in terms of the first layer parts is shown in Fig. 2b. Each part P_i^n of the higher layers $L_n, \forall n > 1$ describes spatial configurations of the parts from the previous layer L_{n-1}:

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

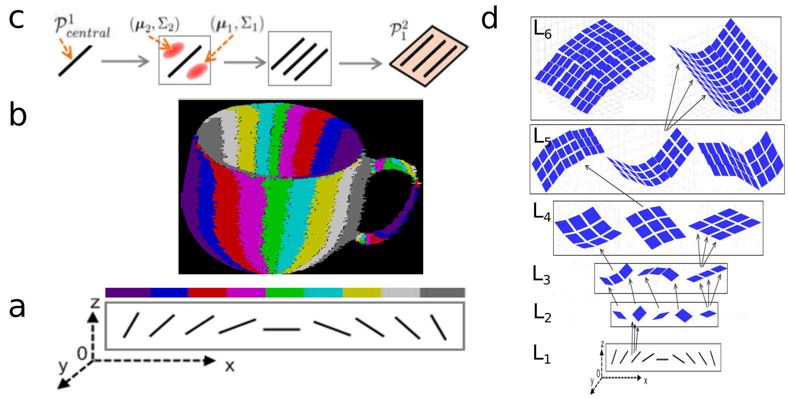

Fig. 2. Quantization of depth differences into nine bins (a). Realization of L_1 parts on the example object – colour coded depth differences (b). Three adjacent depth differences form a small oriented surface patch of layer L_2 (c). A scheme of the hierarchical shape vocabulary (d).

where: $\mu_j = (x_j, y_j, z_j)$ is the mean relative position of the sub-part P_j^{n-1} w.r.t. the central sub-part $P_{central}^{n-1}$, and Σ_j is the covariance matrix representing variability of possible relative positions. These spatial configurations are described in a local neighborhood (termed *receptive field*) of the central sub-part $P_{central}^{n-1}$.

The process of constructing second layer parts is depicted in Fig. 2c. In the current implementation of our framework the design choice is to build parts of the odd layers describing relative positions of sub-parts adjacent in x-direction, while parts of even layers are formed by describing relative positions of sub-parts that are adjacent in y-direction.

2.1 Learning the Vocabulary of Parts

The hierarchical shape vocabulary for each layer L_n, $\forall n > 1$ is learnt using the procedure described in detail in our previous work [11]. In the current paper, only those steps of the algorithm that remained unchanged are outlined. Here the emphasis is on the proposed novel approaches to the part selection problem. The entire vocabulary learning procedure for the layer L_n, $\forall n > 1$ comprises the following steps:

1. Represent the given training range images in terms of parts of layer L_{n-1}. As a result we are given some instances (*part realizations*) of each vocabulary part of layer L_{n-1} that reside in certain locations in the training images. Let $R^n_{i_k}$ be the k-th part realization of part P^n_i of layer L_n. Usually each vocabulary part has many realizations in the training images.
2. In the local neighborhood of each part realization perform *local inhibition*, i.e. suppress strongly overlapped part realizations.
3. Build *statistical maps* characterizing the 3D spatial relations between part realizations of the layer L_{n-1} in training images.

4. Detect peaks in the statistical maps of occurrences of parts in a local neighbourhood. These peaks indicate the most frequently observed spatial configurations. Then we fit the area around each peak with 3-dimensional Gaussian, and establish its parameters μ_j and Σ_j.
5. Based on the peaks in statistical maps build *candidate parts* of the layer L_n as shown in Eq. (1) and insert these parts into $\mathcal{S}$, the set of candidate parts.
6. Perform a part selection procedure, i.e. select a subset $\mathcal{S}' \subseteq \mathcal{S}$ that satisfies some optimality criteria (*selection criteria*). The subset $\mathcal{S}'$ becomes a learnt vocabulary of the layer L_n. In this paper we propose, discuss and experimentally evaluate two novel part selection strategies, presented in section 3.

2.2 Inference

The main goal of the inference process is to detect vocabulary parts in input range image using a previously learnt hierarchical compositional shape vocabulary. Inference is sequentially performed for all layers starting from L_1. Since the parts of the layer L_1 represent quantized differences of depth measured in direction of the x axis, the inference process for this layer can be done as follows:

1. Convolve the range image with Gaussian-derivative filter (for x-derivative). The parameter σ can be chosen according to the image size and noise level. For our experiments we use $\sigma = 1.0$ for the SHREC07 dataset and $\sigma = 1.5$ for the Washington RGBD dataset.
2. Quantize the filter responses at each pixel and assign each of them to the bin that corresponds to the closest first layer part.

For each subsequent layer $L_n, \forall n > 1$ the inference process is done as follows:

1. Consider a local neighborhood (termed *receptive field*) around each part realization $R_{i_k}^{n-1}$. For this receptive field we call the part realization $R_{i_k}^{n-1}$ the *central part realization* and denote it as $R_{central}^{n-1}$.
2. Extract all other part realizations R_j^{n-1} present in this receptive field and their relative positions $\Delta X_j = [\Delta x_j, \Delta y_j, \Delta z_j]$ with respect to the the central part realization $R_{central}^{n-1}$.
3. Spatial configurations of $R_{central}^{n-1}$ and neighboring part realizations extracted from receptive field are then matched to vocabulary parts of the layer L_n. For instance, if we observe a spatial configuration: $R_i^n \equiv (R_{central}^{n-1}, P_j^{n-1}, \{\Delta X, \}_j)$ we have to match it to the vocabulary parts of layer L_n defined as shown in Eq. (1). If the match is found, we say that we detected a part realization of the layer L_n in the location of $R_{central}^{n-1}$.

As we proceed to higher layers of the hierarchy, the size of the receptive field increases. At the end of the whole procedure, the input depth image is described in terms of parts from the learnt vocabulary for each layer.

3 Part Selection Criteria

In our previous work [11] the part selection problem was described as follows. For the rich set of candidate parts $\mathcal{S} = \{P_i^n : i = 1..N\}$ for a given layer L_n, we want to obtain an efficient representation with a relatively small number of parts in the vocabulary. Additionally, to facilitate generalization using our representation, similar surfaces should be encoded by the same vocabulary element.

In order to fulfil these requirements we specify a procedure that selects a subset $\mathcal{S}' \subseteq \mathcal{S}$, solving the following optimization problem:

$$E(\mathcal{S}') = \sum_{i=1}^{N} d(P_i, P'(P_i))\, \nu_i \; + \; \alpha\, |\mathcal{S}'| \tag{2}$$

where ν_i is the frequency of occurrence of realizations of the i-th candidate part P_i^n in the training data, $d(\cdot,\cdot)$ is a distance function that quantifies the similarity between two parts (from the same layer), and $P'(P_i)$ is the part in $\mathcal{S}'$ that is closest to P_i^n. Finally $\alpha \in \mathbb{R}^+$ is a meta-parameter that regulates the trade-off between precision of the representation and the number of selected parts.

Two main disadvantages of this approach are: i) a selected vocabulary strongly depends on the frequency of parts' observation in the training images; e.g., three surface patches constituting a corner and surfaces of high curvature are usually observed less frequently than planar surfaces; ii) a vocabulary selected in this manner is not specifically designed for any particular task, i.e. a specific measure of part importance for solving different computer vision tasks is missing. Therefore, we introduce two different selection criteria.

Minimum Description Length. This part selection strategy finds the minimal subset of candidate parts which is required to represent most of the input range data in terms of the shape vocabulary. It is implemented as a greedy algorithm similar to the one used in [14]. Let us introduce two definitions first. The total coverage $Cover(P_i^n)$ of the part P_i^n is the sum of areas (expressed in the number of pixels) in training images where the part P_i is active.

The second measure is $Cover(P_i^n|\mathcal{S}')$, which is the *conditional coverage* of part P_i^n given that the subset $\mathcal{S}'$ is already included in the vocabulary. It measures the total area in the training images where the part P_i^n is detected, excluding those locations where realizations of parts from $\mathcal{S}'$ are detected. As part realizations may overlap with each other (in terms of area of images where they reside), conditional coverage is a non-increasing measure: As the subset $\mathcal{S}'$ grows, the conditional coverage of each non-selected part either remains unchanged or is lowered by the influence of an already selected part.

Consider learning a vocabulary for layer L_n. The part selection algorithm can be described as follows:

1. Assume the set of selected parts $\mathcal{S}'$ is empty. All candidate parts are stored in the set $\mathcal{S}$.
2. Consider the part P_i^n in $\mathcal{S}$ which has the largest $Cover(P_i^n)$; remove this part from $\mathcal{S}$ and insert it into $\mathcal{S}'$. Candidate parts P_j^n which are similar to P_i^n (if $d(P_i^n, P_j^n) < Thresh$)) are also removed from the set $\mathcal{S}$.

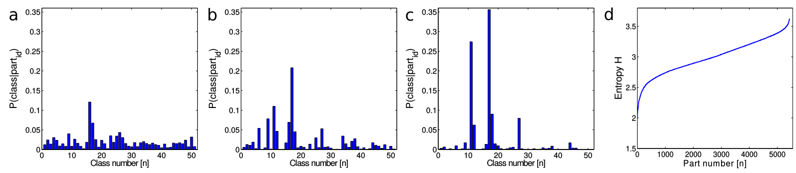

Fig. 3. Entropy based on the probability of the class given individual part $P(class|P_i^n)$. Histogram for part with: high entropy(a), mean entropy(b), low entropy(c). Sorted entropy values for the candidate parts at layer three (d).

3. Perform the following steps: in each iteration find in the set $\mathcal{S}$ the part P_k^n which has the largest $Cover(P_k^n|\mathcal{S}')$, remove this part from $\mathcal{S}$, and add it to the S'; Parts that are similar to P_k^n, are also removed from $\mathcal{S}$.
4. Proceed until the largest value of $Cover(P_j^n|\mathcal{S}')$ becomes less than a predefined threshold value T.
5. The resulting vocabulary is $\mathcal{S}'$.

In contrast to the part selection criteria defined by the cost function in Eq. (2), this part selection algorithm allows selection of parts with high curvature as long as their total coverage is above the threshold value T.

Entropy-based part selection In this part selection strategy we incorporate some discriminative information into the vocabulary learning procedure. The decision to include a candidate part to the vocabulary is based on the distribution of part realizations across object categories. If realizations of a candidate part P_i^n reside in objects of one category only, we consider this part quite category-specific, while if realizations of a candidate part appear in objects of many categories, we assume this part would be less helpful for categorization.

The category probability given a candidate part P_i^n $P(category|P_i^n)$ is approximated using histograms computed for all candidate parts throughout the whole training dataset. The entropy of each individual part is computed from the histogram using:

$$H = -\sum_{i=1}^{n} p_i log(p_i). \tag{3}$$

The $P(category|P_i^n)$ for the parts with high, mean and low entropy are shown in Fig. 3 a,b,c respectively. A graph presenting the sorted entropy H computed for all the the compositions detected at layer L_3 is shown in Fig. 3d. Our proposed strategy is to select a certain number of candidate parts with the highest entropy at each layer of the hierarchy.

4 Experiments

In order to test the performance of our hierarchical compositional representation two publicly available datasets were used. We focus our attention on the task of object categorization from range images. The RGB-D Washington Dataset [13] contains real world data registered with a Kinect-like device – 300 objects are

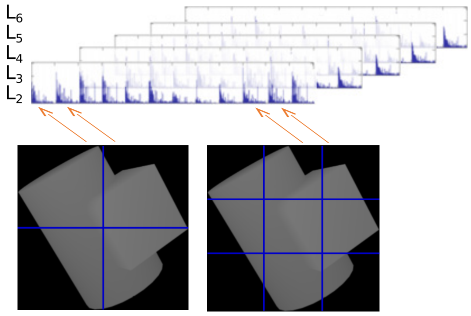

Fig. 4. A histogram of Compositional Parts – the histograms from all sub-regions and all layers detections are stacked together to form a category signature.

split into 51 categories. We used only 20% of the images of each object instance (depth channel only). To partition the data into train and test sets, the exact list of object instances left out provided in [13] was used.

The second dataset used was SHREC07 [7], which contains 400 meshes representing objects that are split into 20 categories. We rotated these mesh models about the z-axis with a step of 18 degrees, thus producing $400\times20 = 8000$ depth images. We evaluated performance using leave-one-out cross-validation.

Learning of Hierarchical Shape Vocabulary We learnt layers 2-6 of the compositional hierarchical shape vocabulary on each dataset separately. In our first experiment we learnt a compositional hierarchy using the MDL principle. In the second experiment we additionally included the most category-specific parts to the vocabulary of each layer (parts with low entropy). All the experiments were conducted in a way to avoid data snooping, i.e. we extracted category-specific parts from the training data only.

Classification The procedure started with the inference of parts of layers 1-6 from range images. Then, following [11], we built stacked histograms of parts as shown in Fig. 4, i.e. we partitioned each object image into 14 sub-regions (1 for the whole object, plus 2×2 and 3×3 partitioning). The distribution of parts of layers 2-6 within each sub-region was represented as a histogram; histograms of each sub-region were then concatenated to form a HOCP (Histogram of Compositional Parts) surface descriptor. These descriptors were fed into an SVM classifier.

Experiments with kernels for SVM Three different types of kernels for SVM were tested in our experiments. χ^2 kernels are implemented in standard packages for SVM. The EMD measure which does not produce a positive semi-definite kernel matrix was modified [25]. The following transformation: $K(S,Q) = exp(-EMD(S,Q))$, where S and Q are histograms, was applied. Additionally, a ground distance metric D_{ij} which denotes distance between bins i and j of the histogram was defined based on geometric similarity of parts. Note that we consider part similarities within each sub-region, and do not take into account cross sub-region relations. For Quadratic-Chi measure [17], it is required to define a bin-similarity matrix A_{ij}. We do it in the same way as in case of EMD kernels, i.e. only similarities of parts within each sub-region are

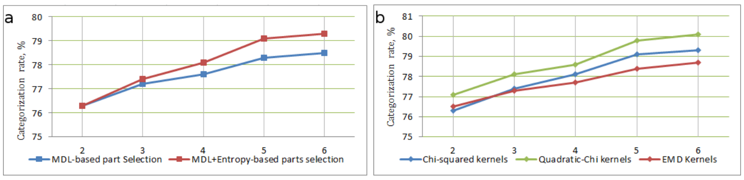

Fig. 5. Categorisation accuracy achieved on Washington RGBD dataset for: different part selection strategies – SVM with χ^2 kernel (a); different types of SVM kernels(b).

taken into consideration. The Quadratic-Chi histogram distance is defined as:

$$QC_m^A(P,Q) = \sqrt{\sum_{ij} \left(\frac{(P_i - Q_i)}{(\sum_c (P_c + Q_c) A_{ci})^m} \right) \left(\frac{(P_j - Q_j)}{(\sum_c (P_c + Q_c) A_{cj})^m} \right) A_{ij}} \quad (4)$$

where: P and Q are two non-negative bounded histograms $P, Q \in [0, U]^N$;
A is a non-negative symmetric bounded bin-similarity matrix;
$A \in [0, U]^N \times [0, U]^N$ and $\forall_{i,j}\ A_{ii} \leq A_{ij}$.

5 Results and Discussion

The results achieved on the Washington dataset are shown in Fig. 5. Our experiments show that combined MDL and entropy-based selection performs better than MDL alone (see Fig. 5a), especially for higher layers where parts become more category-specific. Additionally, the comparison of performance of different kernels used in the classification procedure is shown in Fig. 5b. The best results were achieved for the Quadratic-Chi kernel, which takes into account cross-bin relations and performs normalisation as well, such that the difference between small bins becomes as important as the difference of large ones. This turns out to be a substantial property: all parts in the histograms are distributed non-uniformly and some non-frequent but discriminative parts may have a larger impact on category detection. This observation might also explain the poorer performance of the EMD kernel as pointed out in [17]. The comparison of our results with other approaches is given in Table 1, which shows that our approach is almost as good as the best results achieved for this dataset. Additionally we have also provided the results for the artificial data which are given in the dataset from the Shape Retrieval Contest SHREC07 [7]. Table 1 shows our approach is also almost reaching the top score.

6 Conclusion

The research presented in this paper is focused on the part selection problem and its influence on the performance of hierarchical compositional representations in the object categorisation task. The presented results show that an appropriate selection criterion can boost task performance. Compared to our previous approach, the gain for the RGB-D dataset is about 4%. Additional tests with the

Washington dateset		SHREC'07 dataset	
Method	Depth data only	Method	Depth data
Random Forest [13]	66.8±2.5	ISM 1-NN [19]	79.0
HCR [11]	75.6	BoW [21]	87.3
HMP3D [12]	76.5±2.2	HCR [11]	95.6
CNN-RNN [18]	78.9±3.8	ISM 2-NN [19]	100.0
SP+HMP [4]	81.2±2.3		
Ours	80,1±2.2	Ours	96.4

Table 1. Results for Washington RGB-D dataset [13] and SHREC'07 dataset [7].

histogram similarity measure show that an additional one percent was added to the final result. The use of compositional hierarchies could not only improve performance, but being generative models, they can also furnish additional properties such as better scalability and robustness to partial occlusions. Moreover, they enable performing detection tasks without using sliding windows. A thorough investigation of all these aspects of compositional hierarchies is envisioned as future work.

Acknowledgments. We gratefully acknowledge the support of EU-FP7-IST grant 600918 (PaCMan). The authors would also like to thank Jeremy Wyatt and Sebastian Zurek for their helpful comments.

References

1. Aanæs, H., Dahl, A., Steenstrup Pedersen, K.: Interesting Interest Points. Int. J. of Computer Vision 97(1), 18–35 (2012)
2. Akgul, C., Sankur, B., Yemez, Y., Schmitt, F.: 3D Model Retrieval Using Probability Density-Based Shape Descriptors. Pattern Analysis and Machine Intell., IEEE Trans. on 31(6), 1117–1133 (2009)
3. Besl, P.J.: Surfaces in Range Image Understanding. Springer Series in Perception Engineering, Springer-Verlag (1988)
4. Bo, L., Ren, X., Fox, D.: Unsupervised Feature Learning for RGB-D Based Object Recognition. In: J.P. Desai et al. (ed.) Experimental Robotics, Springer Tracts in Advanced Robotics, vol. 88, pp. 387–402. Springer Int. Publishing (2013)
5. Faugeras, O.: Three-Dimensional Computer Vision. A geometric view point. MIT Press (1993)
6. Fidler, S., Leonardis, A.: Towards Scalable Representations of Object Categories: Learning a Hierarchy of Parts. In: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conf. on. pp. 1–8 (2007)
7. Giorgi, D., Biasotti, S., Paraboschi, L.: SHape REtrieval Contest 2007: Watertight Models Track
8. Guo, G., Wang, Y., Jiang, T., Yuille, A., Fang, F., Gao, W.: A Shape Reconstructability Measure of Object Part Importance with Applications to Object Detection and Localization. Int. J. of Computer Vision 108(3), 241–258 (2014)
9. Guo, Y., Bennamoun, M., Sohel, F., Lu, M., Wan, J.: 3D Object Recognition in Cluttered Scenes with Local Surface Features: A Survey. Pattern Analysis and Machine Intell., IEEE Trans. on 36(11), 2270–2287 (2014)

10. Holzer, S., Shotton, J., Kohli, P.: Learning to Efficiently Detect Repeatable Interest Points in Depth Data. In: A. Fitzgibbon et al. (ed.) Computer Vision – ECCV 2012, Lecture Notes in Computer Science, vol. 7572, pp. 200–213. Springer (2012)
11. Kramarev, V., Zurek, S., Wyatt, J.L., Leonardis, A.: Object Categorization from Range Images using a Hierarchical Compositional Representation. In: Pattern Recognition (ICPR), 2014 22nd Int. Conf. on (2014)
12. Lai, K., Bo, L., Fox, D.: Unsupervised feature learning for 3D scene labeling. In: Robotics and Automation (ICRA), 2014 IEEE Int. Conf. on. pp. 3050–3057 (2014)
13. Lai, K., Bo, L., Ren, X., Fox, D.: A large-scale hierarchical multi-view RGB-D object dataset. In: Robotics and Automation (ICRA), 2011 IEEE Int. Conf. on. pp. 1817–1824 (2011)
14. Leonardis, A., Bischof, H., Maver, J.: Multiple eigenspaces. Pattern Recognition 35(11), 2613–2627 (2002)
15. Li, B., Wu, T., Zhu, S.C.: Integrating Context and Occlusion for Car Detection by Hierarchical And-Or Model. In: D. Fleet et al. (ed.) Computer Vision – ECCV 2014, Lecture Notes in Computer Science, vol. 8694, pp. 652–667. Springer (2014)
16. Ommer, B., Sauter, M., Buhmann, J.M.: Learning Top-Down Grouping of Compositional Hierarchies for Recognition. In: Proceedings of the 2006 Conf. on Computer Vision and Pattern Recognition Workshop. pp. 194–211. CVPRW '06, IEEE Computer Society, Washington, DC, USA (2006)
17. Pele, O., Werman, M.: The Quadratic-Chi Histogram Distance Family. In: K. Daniilidis et al. (ed.) Computer Vision – ECCV 2010, Lecture Notes in Computer Science, vol. 6312, pp. 749–762. Springer Berlin Heidelberg (2010)
18. Richard Socher and Brody Huval and Bharath Bhat and Christopher D. Manning and Andrew Y. Ng: Convolutional-Recursive Deep Learning for 3D Object Classification. In: Advances in Neural Information Processing Systems 25 (2012)
19. Salti, S., Tombari, F., Di Stefano, L.: On the Use of Implicit Shape Models for Recognition of Object Categories in 3D Data. In: R. Kimmel et al. (ed.) Computer Vision – ACCV 2010, Lecture Notes in Computer Science, vol. 6494, pp. 653–666. Springer Berlin Heidelberg (2011)
20. Tabia, H., Laga, H., Picard, D., Gosselin, P.H.: Covariance Descriptors for 3D Shape Matching and Retrieval. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conf. on. pp. 4185–4192 (2014)
21. Toldo, R., Castellani, U., Fusiello, A.: A bag of words approach for 3d object categorization. In: Computer Vision/Computer Graphics Collaboration Techniques, pp. 116–127. Springer (2009)
22. Tombari, F., Salti, S., Di Stefano, L.: Performance Evaluation of 3D Keypoint Detectors. Int. J. of Computer Vision 102(1-3), 198–220 (2013)
23. Tuzel, O., Liu, M.Y., Taguchi, Y., Raghunathan, A.: Learning to Rank 3D Features. In: D. Fleet et al. (ed.) Computer Vision – ECCV 2014, Lecture Notes in Computer Science, vol. 8689, pp. 520–535. Springer Int. Publishing (2014)
24. Woodford, O., Pham, M.T., Maki, A., Perbet, F., Stenger, B.: Demisting the Hough Transform for 3D Shape Recognition and Registration. Int. J. of Computer Vision 106(3), 332–341 (2014)
25. Zamolotskikh, A., Cunningham, P.: An Assessment of Alternative Strategies for Constructing EMD-Based Kernel Functions for Use in an SVM for Image Classification. In: Content-Based Multimedia Indexing, 2007. CBMI '07. Int. Workshop on. pp. 11–17 (2007)
26. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. Int. J. of Computer Vision 73(2), 213 238 (2007)