Object Categorization from Range Images using a Hierarchical Compositional Representation

Vladislav Kramarev, Sebastian Zurek, Jeremy L. Wyatt and Aleš Leonardis
School of Computer Science
University of Birmingham
{vvk201, s.j.zurek, j.l.wyatt, 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 3D object recognition and categorization has been one of the central topics addressed by the computer vision community over decades. 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. Several important properties that make the principles of hierarchical compositionality beneficial for various image processing and computer vision applications have been identified [6]. For example, hierarchical approaches organize features on multiple layers built on top of one another, exploiting share-ability (also termed reusability) of parts. This improves the ability to generalize, as parts on each layer mainly capture common shape characteristics, while simultaneously tolerating some variance in relative positions of subparts. Since many parts are shared within the hierarchy, the size of a representation is usually relatively small and it grows logarithmically with the number of object categories. Another important property of compositional hierarchies is that they usually allow very fast inference through different indexing and matching schemes, restricting the search space when detecting features on each layer.

Most of the recent advances in hierarchical compositional representation have been focused on hierarchies of 2D features [7][6][8][9][10][11]. However, very little work has been done so far to address the formidable problem of true 3D categorization using compositional hierarchical approaches. This paper fills this gap and proposes a hierarchical compositional representation of 3D shapes that is a recursive compositional vocabulary of surface parts. The first layer of the hierarchy contains several pre-defined parts. All the parts from the layers above are learned and represent the most statistically significant compositions of several simpler shape parts from bottom layers.

In order to assess the quality of the representation learned by the layers of the compositional hierarchy, we introduce a 3D object categorization procedure that is based on histograms of compositional parts. (A similar idea for the 2D case was proposed by Tabernik et al. [12].) Each object category is represented by histograms reflecting the spatial distribution of the compositional parts that describe the object’s surface. Figure 1 sketches the process to obtain the features used as input to a classifier. We employ an SVM classifier with $\chi^2$ 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 competitive 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. [7][6][8] that address the problem of hierarchical compositional representation of 2D contour features. 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 to represent the shape of a whole object. In contrast to Fidler’s approach, our proposed method represents 3D shapes in a compositional hierarchy.

Savarese et al. [13] introduced a hierarchical framework for 3D object categorization and pose estimation. They extracted local features from images and grouped them into relatively large discriminative regions (called parts) that are pulled together to form a 3D category model. However, the authors built their representation on top of 2D hand-crafted features (SIFT and saliency detector), and did not use any features derived from range or stereo data.

Pratikakis et al. [14] 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. [15] 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 [16] 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. [17] 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 [18][19]. For comparison with our method on a standard benchmark, we show their results for object category recognition 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. For the remainder of this paper, the coordinate system is defined such that the x and y axes span the image, and that the z-axis encodes depth information.

A. Representation

The general scheme for the compositional hierarchy of parts is presented in Figure 2. Each compositional part, which represents a local surface patch, is built from several (sub)parts of the previous, i.e. lower, layer. Compositional parts exhibit a certain degree of shape variability, that is achieved by encoding the relative positions of subparts as a probability distribution.

We define a hierarchy of layers, where $L_0$ denotes the 0th layer of the hierarchy. The first layer $L_1$ 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 3 shows how these parts can be defined, i.e. we quantize relative depth measured from the training data into nine bins. The number of first layer parts can be chosen differently depending on the type of input data and required precision of the representation. To illustrate the action of the first layer, consider a range image of a mug. Figure 4 shows how the range data can be encoded in terms of the pre-defined vocabulary of 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 $P_{central}^{n-1}$:

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

Fig. 2. Scheme for the compositional hierarchy of parts.
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 training 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 simply by 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 five steps:

1) Express the training data in terms of parts of the layer $L_{n-1}$.

2) Perform local inhibition in the neighborhood of each part.

3) Construct statistical maps that characterize the 3D spatial relations between parts of the layer $L_{n-1}$ in the training data.

4) Produce a list of candidate parts by constructing compositions based on the statistical maps.

5) 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.

(1) The goal of this step is to represent the objects (scenes) in terms of parts from the compositional hierarchical shape vocabulary of the layer $L_{n-1}$. This is done during an inference procedure which is explained in Section III-C.

(2) The next 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 $P^j_k$, we remove the parts that reside in a (small) neighborhood of $P^j_k$ and have a large intersection with $P^j_k$ 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 $P^j_k$. The procedure is illustrated in Figure 7 for $L_2$ parts.

(3) Next, we measure the co-occurrence statistics of the parts of layer $L_{n-1}$ in the training data and hence obtain
Fig. 7. Implementation of local inhibition: a) Detected parts of layer $\text{L}_1$ lying on a surface; b) Derived part $\text{P}_2^2$ of the $\text{L}_2$ layer; c) Other $\text{L}_2$ parts that have intersection with $\text{P}_2^2$ are to be removed (e.g. part $\text{P}_6^2$); d) Surface patch covered by $\text{L}_2$ parts after performing local inhibition.

**statistical maps** that describe relative positions of parts in 3D space. The maps for layer $\text{L}_n$ are functions $f$:

$$f(P_{i-1}^n, P_{j-1}^n, x, y, z) \rightarrow [0, 1]$$

(2)

that are defined for each pair of elements $P_{i-1}^n$ and $P_{j-1}^n$ in layer $\text{L}_{n-1}$, and a 3D offset $(x, y, z) \in \mathbb{R}^3$. The maps encode the probability of observing a part $P_{i-1}^n$ displaced by $(x, y, z)$ relative to a central part $P_{i-1}^n$. 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 $\mu_j$ and covariance matrix $\Sigma_j$. Figure 8 shows an example of such a statistical map depicting co-occurrences of some second layer part $P_{241}^2$ with itself.

(4) Parts $P_i^m$ of the layer $\text{L}_n$ are constructed from the previous layer $\text{L}_{n-1}$ using $\mu_j$ and $\Sigma_j$ as shown in equation (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 or more complex configurations encoding spatial relations of three or more parts from the previous layer. These configurations become **candidate parts** that will reside in layer $\text{L}_n$. Figure 2 depicts how parts are formed for the third and fourth layers.

(5) Typically the set of candidate parts $S = \{P_i^m : i = 1..N\}$ for a given layer $\text{L}_n$ 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 $E(S')$ 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:

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

(3)

where $\nu_i$ is the frequency of occurrence of the i-th candidate part $P_i^m$, $d(\cdot, \cdot)$ 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.

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 $\sigma$ associated with the filter depends on the noise level of the images and was chosen from within the range $[0.5, 2.0]$. Then we 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: maximize the surface coverage, (ideally the entire object surface should be covered by parts from the vocabulary), minimize the overlaps between detected parts, and minimize the reconstruction error. To fulfill all these requirements we have to select a subset $S_{sel} \subset S_{pot}$. Following e.g. Leonardis et al. [20] 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 the following procedure for each layer. 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 form a certain spatial configuration which has to be matched against vocabulary elements at layer $L_n$ of the compositional hierarchy. If a match is found it yields a detection. Note, that we have to match the observed spatial configuration only to the vocabulary parts with central part $\{P_i^{n-1}\}$ (see equation 1). This dramatically reduces the number of hypotheses to be tested and therefore enables very fast inference.

3) The described procedure leads to redundant (strongly overlapped) detections, 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.

4) We eliminate parts to minimize the reconstruction error, maximize coverage and minimize overlap using an optimization function similar to that described by Leonardis et al. [20].

IV. EXPERIMENTS

Each of our experiments consisted of two main phases: (1) learning a compositional hierarchical shape vocabulary, and (2) applying this vocabulary to an object categorization task on a standard benchmark dataset. We conducted experiments with two benchmarks: the Aim@Shape dataset [1], and the Washington RGB-D Object Dataset [2].

We describe these two phases in the following subsections.

A. Learning of the hierarchical shape vocabulary

Since features at the lower layers of the hierarchy contain generic parts that are shared by many categories, we employed an unsupervised learning scheme to obtain the shape vocabulary, i.e. we did not use information about object categories during the learning process. Furthermore, we did not require a large amount of data in order to learn the first layers of the hierarchy – our experiments have shown that the co-occurrence statistics, collected and encoded in statistical maps, converge after processing 300-500 images for the third layer and 700-1000 images for the fourth layer.

From the Aim@Shape dataset we rendered a set of range images presenting all the 3D models at different scales and viewing angles. To form the training data, range images were randomly selected from each object category. Vocabularies for both datasets were learned separately and did not affect each other. Figures 9 and 10 show some example parts from layers $L_3$ and $L_4$.

![Fig. 9. Examples of third layer ($L_3$) parts learned from the Aim@Shape (brown) and Washington (blue) datasets.](image1)

![Fig. 10. Examples of fourth layer ($L_4$) parts learned from the Washington dataset.](image2)

A remarkable fact is that our shape vocabulary was stored in less than 50KB of memory, demonstrating the memory efficiency of our approach.

B. Object categorization based on histograms of compositional parts

In order to evaluate the learned compositional hierarchical shape vocabularies, we constructed a classifier to perform multi-class object categorization from range images. We used histograms of compositional parts (HoPs) to represent each category model. HoPs were computed as follows: first we performed inference over each range image to obtain compositional parts up to layer $L_4$. Then the images, with parts detected, were partitioned into 4 $(2 \times 2)$ and 9 sectors $(3 \times 3)$ as shown in Figure 11, which together with the original image yielded 14 subimages. All the compositional parts of layers $L_2 - L_4$ detected in these subimages were used to compute histograms of parts for each subimage. Finally the 14 histograms were stacked to form a large descriptor (one per image) that was used as the input vector to a $\chi^2$ kernel SVM classifier.

![Fig. 11. Partitioning of the object to build a histogram of compositional parts.](image3)

To be able to compare our approach with other methods we used leave-one-out cross-validation to measure performance.
In the experiment on Aim@Shape we used 20 viewing angles and three scales per model to train the SVM classifier. For the Washington dataset, we used only 20% of the available data.

Figure 12 shows that using features from more layers improved performance for both datasets. In Table I and Table II 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

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toldo et al. [3]</td>
<td>87.3</td>
</tr>
<tr>
<td>Salti et al. [21] using 1-NN for codebooks</td>
<td>79</td>
</tr>
<tr>
<td>Salti et al. [21] using 2-NN for codebooks</td>
<td>100</td>
</tr>
<tr>
<td>Our method (up to ( L_4 ))</td>
<td>95.6</td>
</tr>
</tbody>
</table>

### Table II. Results for RGB-D Object Dataset (Using Range Images Only)

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spin Images &amp; 3D Bounding Boxes [2]</td>
<td>64.7</td>
</tr>
<tr>
<td>Sparse Distance Learning [17]</td>
<td>70.2</td>
</tr>
<tr>
<td>RGB-D Kernel Descriptors [18]</td>
<td>80.3</td>
</tr>
<tr>
<td>Hierarchical Matching Pursuit [19]</td>
<td>81.2</td>
</tr>
<tr>
<td>Our method (up to ( L_3 ))</td>
<td>74.3</td>
</tr>
<tr>
<td>Our method (up to ( L_4 ))</td>
<td>75.6</td>
</tr>
</tbody>
</table>

### V. Conclusion and Future Work

We have presented a 3D learning and category recognition framework built on the principle of hierarchical compositionality. The framework accommodates a large number of object categories, and since parts are shared, the entire representation has a very small memory footprint, making it particularly suited for mobile phone applications. 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. Additionally we intend to investigate other part selection methods based on localization, discrimination and reconstruction criteria.

### 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


