38th German Conference on Pattern Recognition (GCPR), Hannover, September 2016.

Efficient Single-view 3D Co-segmentation using Shape Similarity and Spatial Part Relations

Nikita Araslanov, Seongyong Koo, Juergen Gall, and Sven Behnke

Computer Science Institute, University of Bonn, Germany
{araslanov@ | koosy@ais | gall@iai | behnke@cs}.uni-bonn.de

Abstract. The practical use of the latest methods for supervised 3D shape co-segmentation is limited by the requirement of diverse training data and a watertight mesh representation. Driven by practical considerations, we assume only one reference shape to be available and the query shape to be provided as a partially visible point cloud. We propose a novel co-segmentation approach that constructs a part-based object representation comprised of shape appearance models of individual parts and isometric spatial relations between the parts. The partial query shape is pre-segmented using planar cuts, and the segments accompanied by the learned representation induce a compact Conditional Random Field (CRF). CRF inference is performed efficiently by A^*-search with global optimality guarantees. A comparative evaluation with two baselines on partial views generated from the Labelled Princeton Segmentation Benchmark and point clouds recorded with an RGB-D sensor demonstrate superiority of the proposed approach both in accuracy and efficiency.

1 Introduction

As humans, we generally feel comfortable interacting with objects of diverse shapes that can belong to the same semantic category. Mugs, for example, take a variety of shapes, though their functions of containing liquid and drinking are not impeded. It takes us little effort to associate handles of different mugs despite these shape variations. This ability allows us to seamlessly generalise our limited experience to all other objects of similar type that we encounter later in life.

We define the object correspondence problem in terms of co-segmentation. In contrast to a pointwise correspondence, co-segmentation seeks to establish a semantic correspondence between object *parts* by modelling the object structure based on part appearance and topological part relations. We understand parts to fulfill a certain function within the working of the whole shape, such as legs of a chair for stability, or a handle of a vase for grasping. We argue that this formulation lends itself well for many practical applications where high shape discrepancies between same-category objects and partial views make it difficult to estimate a full deformation model.

2 Related Work

3D shape co-segmentation is closely related to shape correspondence with some of the classical approaches surveyed by van Kaick et al. [38].

The cornerstone of supervised co-segmentation methods [13,15] is the representation of the object surface mesh with a Conditional Random Field (CRF), inspired by similar models in computer vision [29,33]. The unary data terms in the CRF model geometric similarity of individual faces in the mesh, whereas the pairwise term is learned to differentiate between segment boundaries and their interior. Van Kaick et al. [13] also added an "intra-edge" term to distinguish between different shape parts based on their geometric similarity. Good performance of these methods hinges on the size and diversity of the training data.

In an unsupervised setting, a coherent segmentation is sought over a group of shapes simultaneously. The scores of normalised cuts guided agglomerative clustering [8,20] and a tree structure was used [14] to exploit the hierarchy of object structures. Like [14], Sidi et al. [34] computed diffusion maps, but applied spectral clustering instead. Single features or a concatenation thereof were used to establish correspondence between mesh faces [8,13,15], segments [34], supervoxels [20] or corresponding indicator functions [11,23,40], while Hu et al. [9] also clustered each feature independently in their own subspace and fused the result. A rigid pre-alignment was crucial to establish the initial correspondence in [8,14], whereas non-rigid variability of the reference model [1,25,44] or a template structure of primitive shapes [16,42,45] drove the co-analysis itself, very much by analogy with deformable models in images [37]. Huang et al. [10] formulated the problem as a quadratic integer program that jointly optimises over individual segmentation and its consistency with the other shapes in the group. They and Hu et al. [9] applied relaxation techniques, but combinatorial optimisation [13,15], alternating schemes [20] and greedy strategies [16] were also employed to efficiently solve the non-convex objective.

All discussed methods exhibit a number of practical limitations. Most notably, it is the heavy reliance on holistic and contextual features which make them suitable only for closed manifold models [11,13,15,31,32]. At the same time, the potential of structural constraints remains largely untapped [16,21]. Also, unsupervised co-analysis is inherently unable to exploit the ground-truth segmentation: The user might want to identify specific parts of a known model on the novel shape. Kim et al. [16] and the semi-supervised approach of Wang et al. [41] allow mechanisms for a progressive refinement of the segmentation, albeit by means of manual intervention of the user.

Encouraged by the success of part-based models in the context of object detection [2,7], as well as recent advances in classification [5,26,28] and shape retrieval [22,24,35], we propose to revise the classical co-segmentation approach. We learn a part-based representation from a single CAD model or a physical object and expect the query shape to be provided only as a partial point cloud. Our model is also flexible enough for encoding structural constraints between parts in a natural manner. In this work, we demonstrate one such possibility.

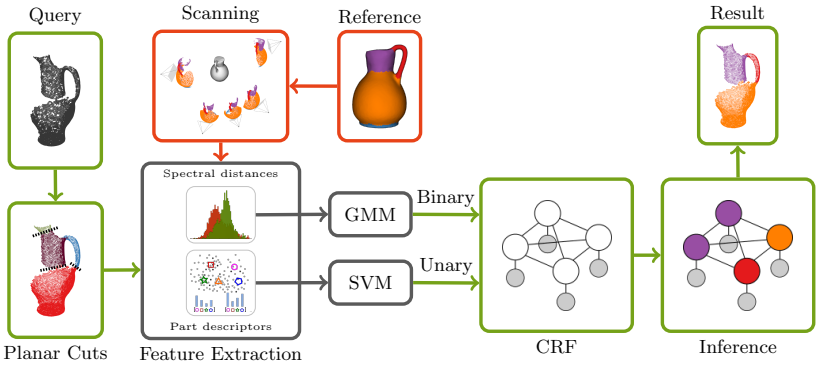


Fig. 1: Overview of our approach. A segmented mesh of the reference shape is used for training. The query is a point cloud of a similar partially observed shape.

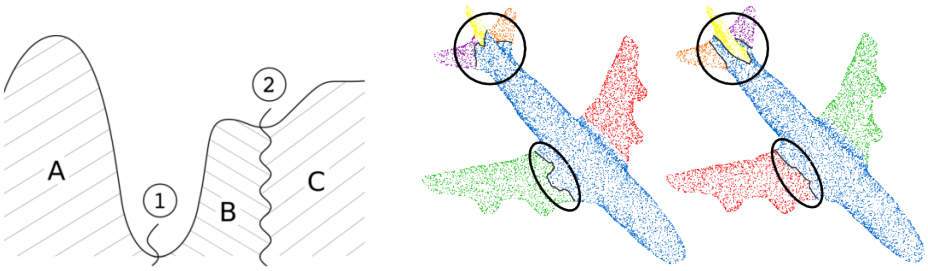
3 Method

In the real world, objects can be observed only partially from any given view angle of the sensor. Therefore, we train a model of the given reference shape using the feature vectors extracted from a set of single views (see Fig. 1). The provided query shape is pre-segmented in an unsupervised fashion to obtain part candidates. Note that the partitioning may oversegment object parts. Our choice of the pre-segmentation algorithm is driven by the intention to avoid fine over-segmentation while keeping track of potential model parts at the same time. This makes it compatible with the learned model and allows structure constraints to be incorporated at the level of shape parts as opposed to segment boundaries. The co-segmentation problem ultimately reduces to an efficiently solved inference in a moderately-sized Conditional Random Field.

3.1 Model

In the context of pointwise shape correspondence, Bronstein et al. [3] developed a shape embedding framework based on the notions of *intrinsic* and *extrinsic* similarity. While isometric deformations do not affect the intrinsic similarity, it is prone to topology changes. By contrast, extrinsic similarity is topologically stable, yet does not exhibit isometric invariance. We integrate these notions of similarity in a part-based representation by minimising the discrepancy in the part shape appearance and the inter-part isometric distortion.

We assume the reference $\mathcal{S} := \bigcup_i \mathcal{S}_i$ and the query $\mathcal{T} := \bigcup_i \mathcal{T}_i$ shapes to be collections of segments formed by points in the Euclidean space $\mathbb{E}^3$. We define a label function $\ell : \mathcal{S} \to L_{\mathcal{S}}$ mapping shape segments $\mathcal{S}_i$ to the label space $L_{\mathcal{S}} \subset \mathbb{Z}$ and let ℓ_i denote the label of segment $\mathcal{S}_i$ for short. The probability $p(\ell_i \mid \mathcal{T}_j)$ is related to appearance similarity of the segment $\mathcal{T}_j$ with the segments in $\mathcal{S}$ labelled ℓ_i. Similarly, we model probability $p(\ell_i, \ell_j \mid \mathcal{T}_i, \mathcal{T}_j)$ to measure the degree



(a) (b)

Fig. 2: Our modification of the CPC segmentation: **(a)** Illustration of the problem; **(b)** Example segmentation without (left) and with (right) modification.

of isometric distortion between each pairwise assignment. Our objective can be formulated as a maximum likelihood estimate of the form:

$$\underset{\ell}{\text{minimize}} \quad -\sum_i \overbrace{\log p(\ell_i \mid \mathcal{T}_i)}^{\text{extrinsic similarity}} -\sum_{i,j} \overbrace{\log p(\ell_i, \ell_j \mid \mathcal{T}_i, \mathcal{T}_j)}^{\text{intrinsic similarity}}. \tag{1}$$

3.2 Segmentation

We base the construction of segment candidates on the recently introduced Constrained Planar Cuts (CPC) method [30]. The algorithm finds planar cuts through concave regions that define segment boundaries. In our preliminary experiments, we found that multiple cuts in the regions with high concentration of concave points frequently yield a number of small fragments which are subsequently merged with neighboring segments according to size. The suboptimality of this approach is illustrated in Fig. 2a. Consider an imaginary object profile segmented with cuts (1) and (2) into parts $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ such that $|\mathbf{B}| < |\mathbf{C}| < |\mathbf{A}|$, where $|\cdot|$ is a segment size measure (e.g. segment area). If segment $\mathbf{B}$ is small enough to be merged, the CPC algorithm will assign it to segment $\mathbf{A}$ since $|\mathbf{A}| > |\mathbf{C}|$. However, cut (1) exhibits a more pronounced concavity than cut (2) and, hence, merging $\mathbf{B}$ with $\mathbf{C}$ will be more visually cohesive.

To address this problem, we merge segments in the ascending order of concavity scores computed as the fraction of concave points on the boundary. Recall that for neighbouring supervoxels with centroids $\mathbf{x}_1$ and $\mathbf{x}_2$ and normals $\mathbf{n}_1$ and $\mathbf{n}_2$ the connection is concave if $\mathbf{n}_1 \cdot \mathbf{d} - \mathbf{n}_2 \cdot \mathbf{d} < 0$, where $\mathbf{d} = (\mathbf{x}_1 - \mathbf{x}_2)/\|\mathbf{x}_1 - \mathbf{x}_2\|$. A comparative example in Fig. 2b demonstrates that an arbitrary order of merging the shape fragments leads to the jagged segment boundaries on the wing and tail of the airplane, whereas agglomeration of the fragments in the increasing order of concavity results in a more natural segmentation. We quantitatively summarise the effectiveness of our modification in comparison with the original version in Table 1.

3.3 Shape Appearance

We represent appearance of each object part using "shape words" derived from a generative model of the underlying feature space spanned by the local shape descriptors. In this work, we contrast the frequency-counting Bag-of-Words (BoW) paradigm with the second-order statistics gathered by the Fisher vectors (FV) [12].

For each view v and shape part with label ℓ, we extract a set of 3D point clusters $\mathcal{P}_{\ell,v}$ with uniformly sampled centres. From every set of point clusters we draw an equal number of randomly sampled fixed-sized subsets $\mathcal{P}_{\ell,v,i} \subset \mathcal{P}_{\ell,v}$, or *feature packets* for short. By construction, the feature packet does not rely on complete visibility of the shape part present in the training data. Furthermore, the effect of the disparity in the surface area of each shape part is mitigated, since every feature packet comprises the same number of point clusters.

Let $M_{\ell,v} = \{\mathbf{m}_{\ell,v,t} \mid \mathbf{m}_{\ell,v,t} \in \mathbb{R}^D, \forall t = 1, ..., T\}$ denote the set of T shape feature vectors with label $\ell \in L_{\mathcal{S}}$ visible from view angle $v \in V$. We model the union of the feature vectors over all labels and views $\bigcup_{\ell \in L_{\mathcal{S}}, v \in V} M_{\ell,v}$ by the Gaussian mixture model (GMM):

$$p(\mathbf{m}_{\ell,v,t}) = \sum_{i=1}^{K} w_i \mathcal{N}(\mathbf{m}_{\ell,v,t} | \boldsymbol{\mu}_i, \Sigma_i), \tag{2}$$

where $\mathcal{N}(\mathbf{m}_{\ell,v,t} | \boldsymbol{\mu}_i, \Sigma_i)$ is a multinomial normal distribution with mean $\boldsymbol{\mu}_i$ and diagonal covariance matrix Σ_i.

A BoW vector $\mathbf{f}_{BoW}(p_{\ell,v}) \in \mathbb{R}^K$ for each cluster in the feature packet $\rho_{\ell,v,i} \in \mathcal{P}_{\ell,v,i}$ can be constructed as:

$$f^{(k)}_{\mathrm{BoW}}(\rho_{\ell,v,i}) = \frac{w_k}{|\rho_{\ell,v,i}|} \sum_{t} \mathcal{N}(\mathbf{m}_{\ell,v,t} | \boldsymbol{\mu}_k, \Sigma_k), \tag{3}$$

where $\mathbf{m}_{\ell,v,t} \in \rho_{\ell,v,i}$, $|\rho_{\ell,v,i}|$ is the number of low-level feature descriptors and $f^{(k)}_{\mathrm{BoW}}(\cdot)$ is the kth dimension of vector $\mathbf{f}_{\mathrm{BoW}} \in \mathbb{R}^K$. We vectorise each feature packet by taking the average over the BoW vectors of the clusters it contains.

To construct the Fisher vector, we compute the gradients $G_{\boldsymbol{\mu}_k}(\rho_{\ell,v,i}) := \frac{\partial \log p(\rho_{\ell,v,i}|\lambda)}{\partial \boldsymbol{\mu}_k}$ and $G_{\boldsymbol{\sigma}_k}(\rho_{\ell,v,i}) := \frac{\partial \log p(\rho_{\ell,v,i}|\lambda)}{\partial \boldsymbol{\sigma}_k}$ for every point cluster $\rho_{\ell,v,i}$ of the feature packet $\mathcal{P}_{\ell,v,i}$ extracted from a segment with label ℓ in view v:

$$G_{\boldsymbol{\mu}_k}(\rho_{\ell,v,i}) = \frac{1}{|\rho_{\ell,v,i}|\sqrt{\omega_k}} \sum_{t=1}^{|\rho_{\ell,v,i}|} \gamma_{\ell,v,t}(k) \left(\frac{\mathbf{m}_{\ell,v,t} - \boldsymbol{\mu}_k}{\sigma_k} \right), \tag{4}$$

$$G_{\boldsymbol{\sigma}_k}(\rho_{\ell,v,i}) = \frac{1}{|\rho_{\ell,v,i}|\sqrt{2\omega_k}} \sum_{t=1}^{|\rho_{\ell,v,i}|} \gamma_{\ell,v,t}(k) \left(\frac{(\mathbf{m}_{\ell,v,t} - \boldsymbol{\mu}_k)^2}{\sigma_k^2} - 1 \right), \tag{5}$$

where vector division is element-wise, $\sigma_i := \mathbf{diag}(\Sigma_i)$ and

$$\gamma_{\ell,v,t}(k) = \frac{\omega_k u_k(\mathbf{m}_{\ell,v,t})}{\sum_{j=1}^{K} \omega_j u_j(\mathbf{m}_{\ell,v,t})}, \quad \mathbf{m}_{\ell,v,t} \in \rho_{\ell,v,i} \tag{6}$$

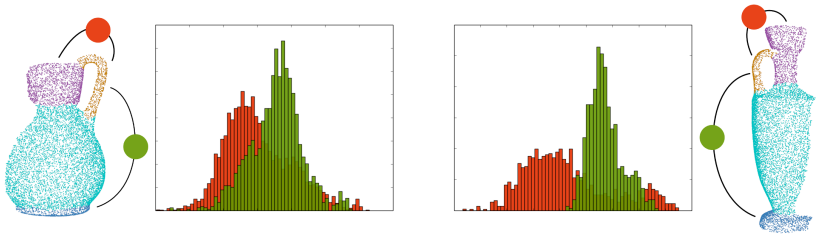


Fig. 3: **Comparison of spectral distances between two pairs of segments.** The distribution of the commute time distances between pairs of points on the base and the handle (green) and on the handle and the neck (red) are shown. Despite shape discrepancy, the histograms still capture the key features of the shape topology: The base is "farther away" to the handle than the neck.

is the soft assignment of descriptor $\mathbf{m}_{\ell,v,t}$ to the Gaussian centre k. The Fisher vector is formed by concatenating the gradients (4) and (5) of each Gaussian centre:

$$\mathbf{f}_{\mathrm{FV}}(\rho_{\ell,v,i}) = (G^T_{\boldsymbol{\mu}_1}(\rho_{\ell,v,i}), ..., G^T_{\boldsymbol{\mu}_K}(\rho_{\ell,v,i}), G^T_{\boldsymbol{\sigma}_1}(\rho_{\ell,v,i}), ..., G^T_{\boldsymbol{\sigma}_K}(\rho_{\ell,v,i}))^T. \quad (7)$$

The inherent sparsity of $2KD$-dimensional Fisher vectors is detrimental to their discriminative properties when used in conjunction with the common L2-distance. As a remedy, we apply the power normalisation that uniformly rescales the vector by applying the following function element-wise [26]: $f(z) = sign(z)|z|^{\alpha}$.

The computed vectors $\mathbf{f}_{BoW}(\cdot)$ and $\mathbf{f}_{FV}(\cdot)$ along with the corresponding labels of the shape parts they represent, form the training dataset of the shape appearance model. We use the Support Vector Machine (SVM) with an RBF kernel for BoW vectors and a linear SVM for Fisher vectors. The trained classifier is used for label prediction of each segment on the query shape.

3.4 Isometric Spatial Relations

The distribution of spectral distances can be used as a measure of shape similarity [4]. Moreover, we could also apply this principle to segment pairs of the same shape, i.e we can extract the distances between a pair of point sets and compare the distribution to that derived from another pair. To support our intuition, an example with histograms of two partially visible vases is shown in Fig. 3. In addition to providing insights into the shape topology, the histograms also insinuate the normal distribution.

In order to account for distance variation in partial views and intrinsic shape symmetries, we propose to learn a multinomial distribution of distances extracted from every pair of shape parts $D_S(\ell_i, \ell_j) := \{d_S(s_i, s_j) \mid s_i \in \mathcal{S}_i, s_j \in \mathcal{S}_j\}$. Note, that by construction $D_S(\ell_i, \ell_j) = D_S(\ell_j, \ell_i)$ and we allow $i = j$ since the distance distribution is also informative within a single segment.

Denoting by $\ell_{i\sim i'}$ the assignment of label i' to segment $\mathcal{T}_i$ of the query shape, we compute the likelihood estimate of the data given a pairwise assignment as

$$p\big(D_S(\mathcal{T}_i,\mathcal{T}_j) \mid \ell_{i\sim i'},\ell_{j\sim j'}\big) = \sum_n \sum_k \omega_k^{i'j'} \mathcal{N}(\mu_k^{i'j'},\sigma_k^{i'j'} \mid d_{\mathcal{S}}(t_{in},t_{jn})). \quad (8)$$

Assuming any given pairwise assignment to be equiprobable, its probability estimate is computed using the Bayes rule:

$$p\big(\ell_{i\sim i'},\ell_{j\sim j'} \mid D_{\mathcal{S}}(\mathcal{T}_i,\mathcal{T}_j)\big) = \frac{p\big(D_{\mathcal{S}}(\mathcal{T}_i,\mathcal{T}_j) \mid \ell_{i\sim i'},\ell_{j\sim j'}\big)}{\sum_{i'',j''} p\big(D_{\mathcal{S}}(\mathcal{T}_i,\mathcal{T}_j) \mid \ell_{i\sim i''},\ell_{j\sim j''}\big)}. \quad (9)$$

To compute the distance $D_{\mathcal{S}}(\cdot,\cdot)$, we use the eigenfunctions of the Laplace-Beltrami operator applied directly to the point cloud. We estimate it using the Moving Least Squares (MLS) approximation [17] and define $d_S(x,y)$ as the commute time distance [4], $d_S^2(x,y) = \sum_i \frac{1}{\lambda_i}(\phi_i(x)-\phi_j(y))^2$, where λ_i and $\phi_i(\cdot)$ is the i-th eigenvalue and eigenfunction of the operator.

3.5 Inference

Our resulting model is a small to medium-sized CRF with fully connected label nodes and the energy defined by (1). We extract the same number of feature packets on the query shape $\bigcup_j \mathcal{T}_j$ whose segments were generated in the pre-segmentation step. The scores obtained from predictions of the individual feature packets are averaged over complete segments and the result of the prediction of the label assignment ℓ_i to the segment $\mathcal{T}_j$ is naturally interpreted as $p(\ell_i \mid \mathcal{T}_j)$. We let $p(\ell_i,\ell_j \mid \mathcal{T}_i,\mathcal{T}_j) := p\big(\ell_{i\sim i'},\ell_{j\sim j'} \mid D_{\mathcal{S}}(\mathcal{T}_i,\mathcal{T}_j)\big)$ define the distance measure between the two segments. We also add "hard" constraints that penalise part neighbourhoods not observed in the reference shape [13,15].

We observe that our compact model is not dissimilar to the one used by [2] for object part detection. Their study revealed that for small graphs, A^*-based inference often outperformed other algorithms, such as (Loopy) Belief Propagation [43] and the Tree Reweighted Belief Propagation [39] not only in the optimality but also in the runtime. This insight and the equivalence of our models supports our choice of the $A^\star$-search for the inference technique.

4 Evaluation

In two experiments, we evaluate our approach and compare its performance to the baseline derived from the state-of-the-art [13] and [15]. In our implementation of these methods, we only omitted volumetric and global feature descriptors which do not scale to partial shapes, also corroborated by the failure of the author's original C++/Matlab implementation of [15]. In the first quantitative experiment, we evaluated variants of our approach on the Labelled PSB dataset [15]. In the second qualitative experiment, we used point cloud data of two watering cans recorded with an RGB-D camera. This experiment demonstrates the practical aspects and efficiency of our approach in real-world scenarios.

Table 1: Average accuracy on the LPSB dataset, in percent.

	Ant	Airplane	Armadillo	Bearing	Bird	Bust	Chair	Cup	Fish	Fourleg	Glasses	Hand	Human	Mech	Octopus	Plier	Table	Teddy	Vase	Overall
[13]	58.8	62.7	35.2	43.2	58.1	43.5	59.6	81.6	84.2	**60.1**	78.1	52.2	41.3	81.3	82.0	33.7	71.6	71.9	64.3	61.2
[15]	58.9	62.0	35.6	43.4	57.0	43.2	59.6	81.8	**84.4**	59.4	**78.6**	52.7	41.6	81.7	**82.8**	32.5	70.9	71.1	65.5	61.2
SHOT+BoW	66.2	59.2	42.2	52.1	57.4	43.8	60.6	90.0	72.1	51.1	75.4	53.4	35.8	82.4	76.5	70.5	**88.9**	64.5	70.6	63.8
FPFH+BoW	69.9	57.5	48.9	55.7	55.3	40.8	54.5	88.9	75.9	51.6	60.7	54.1	38.0	80.2	66.0	69.5	84.1	72.5	69.2	62.8
SHOT+BoW+ISO	65.6	57.0	39.9	50.5	52.0	43.6	56.7	87.6	71.7	48.1	75.5	46.8	34.2	84.4	75.0	57.3	87.5	69.4	65.3	61.5
SHOT+BoW+ISO*	63.9	58.8	41.2	52.1	53.4	44.0	57.4	89.6	73.2	48.3	74.0	49.7	37.0	84.3	76.8	70.5	86.5	68.1	69.4	63.0
FPFH+FV	**77.7**	64.0	**54.9**	52.2	58.5	44.6	60.2	88.7	78.4	54.9	74.1	56.0	43.7	84.1	69.6	**71.9**	85.4	76.4	70.3	66.6
FPFH+FV (CPC)	77.2	60.4	43.8	49.4	57.8	40.6	58.2	**90.9**	77.9	50.3	69.2	**56.2**	41.2	**85.8**	71.3	70.8	85.5	75.3	68.9	64.8
FPFH+FV (L2)	74.6	63.3	52.2	54.5	57.2	44.8	60.3	88.2	77.8	54.2	73.8	56.0	41.6	82.3	68.2	68.2	85.7	76.4	66.2	65.6
SHOT+FV	72.1	**64.7**	52.5	**57.6**	**60.3**	42.5	**64.1**	90.4	79.0	59.0	75.6	56.0	**45.7**	79.0	79.9	71.8	87.9	76.5	**72.8**	**67.8**
FPFH+FV+ISO	74.1	60.0	51.5	51.2	53.6	**45.0**	55.5	87.5	77.7	50.6	73.4	49.6	40.4	84.6	69.8	58.8	84.1	**77.0**	63.8	63.6

4.1 Quantitative Results

For each category in the Labelled PSB dataset [15], we generated a dataset of *valid* random views. To retain object diversity, a random view was considered valid if at least 20% of each shape part is visible. We created a uniform grid of view points on a sphere enclosing the shape and proved each for the validity criterion. In order to obtain distinctive shapes, we selected randomly only eight viewpoints with the maximum spread. The dataset is publicly available[1].

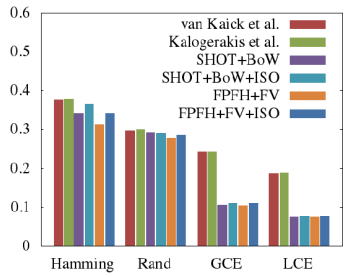
For every category, we ran "one-vs-all" co-segmentation scheme with every *compatible* pair of shapes. The pair counts as compatible if the query doesn't contain labels not present on the reference shape. Hence, we performed co-segmentation for at most $20 \times 8 \times 19 = 3040$ object pairs in each category (the runs against own partial views were excluded).

For benchmarking, we adopted the evaluation metrics proposed by [6], namely the Hamming distance, Rand index[2], the Global and Local Consistency Error (GCE and LCE) and the accuracy. We compared a number of configurations of our approach based on either Bag-of-Words (BoW) or Fisher vector (FV) representations. Two shape descriptors were used for the underlying feature space: FPFH [27] and SHOT [36]. To assess the influence of the binary term, we also ran experiments with and without the isometric spatial relations abbreviated ISO.

The accuracy results per category are summarised in Table 1. We observe that our approach shows higher accuracy in 14 out of 19 categories and on average. However, our isometric context did not improve the results. In fact, the

[1] `http://www.ais.uni-bonn.de/data/alroma`

[2] Reported as one minus Rand index, by convention



(a) Evaluation on the criteria [6].

	van Kaick et al.	FPFH+FV
Training	259.6	581.0
Learning CRF	506.5	-
Total	766.1	**581.0**
Pre-segmentation	-	34.2
Inference	290.15	16.1
Total	290.15	**50.3**

(b) Average time per object pair, in seconds.

Fig. 4: Quantitative evaluation results.

accuracy is worse on average and per category, with the exception of Bust, Mech and Teddy. We attribute this fact to a better pre-segmentation quality of these shapes which did not lead to a significant distortion of the distribution of spectral distances. In some categories, the hard constraints further exacerbated the performance. Dropping them (configuration SHOT+BoW+ISO*) particularly improved the segmentation of Plier, where the failure to segment out the pivot, previously led to a violation of the hard constraint "handles–nose".

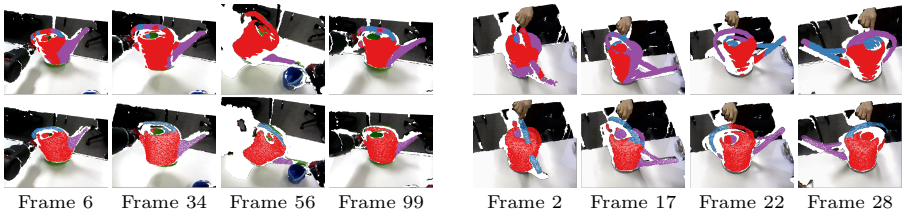
Another insight is the better accuracy obtained with SHOT descriptors wrt. FPFH. However, the memory demands become impractical for commodity hardware if Fisher vectors are used (∼10 times more than FPFH). The L2-normalisation of Fisher vectors [26] did not improve the results (FPFH+FV (L2)) which is in line with the expectation that the original motivation for using it does not apply (i.e. there is no background to neglect). Also, the original CPC algorithm (FPFH+FV (CPC)) yielded a lower accuracy on average. This is expected since our modification aimed only at refining segment boundaries.

Our method exhibits a sharp decrease of GCE and LCE as seen in Fig. 4a. While the baselines tend to produce segmentations with many local inconsistencies, our method assigns labels to a few large segments. The results in Fig. 4a also agree with those in Table 1: The configuration based on the Fisher vectors achieves best scores overall while the isometric context exacerbates the performance.

4.2 Qualitative Results

In the second experiment, we evaluated our configuration FPFH+FV on real data by comparing its efficiency and qualitative accuracy with the baseline [13].

We supplied both algorithms with a manually labelled reference shape obtained from an RGB-D sensor. Since the baseline [13] only works with meshes, we computed a fast triangulation [19] from a representative partial view cloud. Our approach, by contrast, was able to learn the model from a small number of labelled single-view point clouds extracted from a sequence of training frames.



(a) (b)

Fig. 5: Test sequences from an RGB-D sensor. **Top row:** van Kaick et al. [13]; **Bottom row:** Ours (FPFH+FV). (a) Same query shape as the reference; (b) A novel query shape.

The first test sequence is a recording of the original model in a previously unseen action. From a representative selection of frames in Fig. 5a, we conclude that although our approach failed to detect the handle in Frame 99, it still performed well in other frames. By comparison, the baseline method [13] identified only patches of the handle throughout the sequence.

In a more challenging setup, both algorithms had to co-segment a novel instance of a watering can. The baseline method misclassified a large fraction of the container in the first two frames and confused the spout with the handle in the last two (Fig. 5b). Our approach mixed up the parts in Frame 2 and detected only part of the handle in Frame 22, but otherwise performed well.

Time benchmarking was conducted in the first part of the experiment using a laptop with Intel Core i7 CPU and 8GB RAM. The code was parallelised for face- and pointwise operations (e.g. normals and curvatures). The timing results are summarised in Figure 4b. Despite the additional pre-segmentation step, our co-segmentation was almost six times faster than the baseline implementation.

5 Conclusions

We presented a new approach to the co-segmentation problem that addresses practical limitations of the existing state-of-the-art methods. Our algorithm is readily applicable to point clouds captured from real sensors and does not require a complete object model both for the reference and the query shape. The generality of our pipeline suggests a number of configurations and we have investigated only a subset of them. In future, we plan to experiment with other feature encoding schemes, such as spatial sensitive Bag-of-Words [24] and improve the contextual features using other approximations of diffusion distances [3,18] and spatial relations.

6 Acknowledgements

This work was supported by German Research Foundation (DFG) under grant BE 2556/12 ALROMA in priority programme SPP 1527 Autonomous Learning.

References

1. Alhashim, I., Xu, K., Zhuang, Y., Cao, J., Simari, P., Zhang, H.: Deformation-driven topology-varying 3D shape correspondence. TOG 34(6), 236 (2015)
2. Bergtholdt, M., Kappes, J., Schmidt, S., Schnörr, C.: A study of parts-based object class detection using complete graphs. IJCV 87(1-2), 93–117 (2010)
3. Bronstein, A.M., Bronstein, M.M., Kimmel, R., Mahmoudi, M., Sapiro, G.: A Gromov-Hausdorff framework with diffusion geometry for topologically-robust non-rigid shape matching. IJCV 89(2-3), 266–286 (2010)
4. Bronstein, M.M., Bronstein, A.M.: Shape recognition with spectral distances. PAMI 33(5), 1065–1071 (2010)
5. Chatfield, K., Lempitsky, V.S., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: BMVC (2011)
6. Chen, X., Golovinskiy, A., Funkhouser, T.: A benchmark for 3D mesh segmentation. In: TOG. vol. 28, p. 73. ACM (2009)
7. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. PAMI 32(9), 1627–1645 (2010)
8. Golovinskiy, A., Funkhouser, T.: Consistent segmentation of 3D models. Computers & Graphics 33(3), 262–269 (2009)
9. Hu, R., Fan, L., Liu, L.: Co-segmentation of 3D shapes via subspace clustering. In: CGF. vol. 31, pp. 1703–1713 (2012)
10. Huang, Q., Koltun, V., Guibas, L.: Joint shape segmentation with linear programming. In: TOG. vol. 30, p. 125. ACM (2011)
11. Huang, Q., Wang, F., Guibas, L.: Functional map networks for analyzing and exploring large shape collections. TOG 33(4), 36 (2014)
12. Jaakkola, T., Haussler, D., et al.: Exploiting generative models in discriminative classifiers. NIPS pp. 487–493 (1999)
13. van Kaick, O., Tagliasacchi, A., Sidi, O., Zhang, H., Cohen-Or, D., Wolf, L., Hamarneh, G.: Prior knowledge for part correspondence. In: CGF (2011)
14. van Kaick, O., Xu, K., Zhang, H., Wang, Y., Sun, S., Shamir, A., Cohen-Or, D.: Co-hierarchical analysis of shape structures. TOG 32(4), 69 (2013)
15. Kalogerakis, E., Hertzmann, A., Singh, K.: Learning 3D mesh segmentation and labeling. In: TOG. vol. 29, p. 102. ACM (2010)
16. Kim, V.G., Li, W., Mitra, N.J., Chaudhuri, S., DiVerdi, S., Funkhouser, T.: Learning part-based templates from large collections of 3D shapes. TOG 32(4), 70 (2013)
17. Liang, J., Lai, R., Wong, T.W., Zhao, H.: Geometric understanding of point clouds using Laplace-Beltrami operator. In: CVPR. pp. 214–221 (2012)
18. Liu, Y., Prabhakaran, B., Guo, X.: Point-based manifold harmonics. VCG 18(10), 1693–1703 (2012)
19. Marton, Z.C., Rusu, R.B., Beetz, M.: On Fast Surface Reconstruction Methods for Large and Noisy Datasets. In: ICRA. Kobe, Japan (May 12-17 2009)
20. Meng, M., Xia, J., Luo, J., He, Y.: Unsupervised co-segmentation for 3D shapes using iterative multi-label optimization. CAD 45(2), 312–320 (2013)
21. Mitra, N.J., Wand, M., Zhang, H., Cohen-Or, D., Kim, V., Huang, Q.X.: Structure-aware shape processing. In: Eurographics Report. ACM (2014)
22. Ohbuchi, R., Osada, K., Furuya, T., Banno, T.: Salient local visual features for shape-based 3D model retrieval. In: SMA. pp. 93–102. IEEE (2008)
23. Ovsjanikov, M., Ben-Chen, M., Solomon, J., Butscher, A., Guibas, L.: Functional maps: a flexible representation of maps between shapes. TOG 31(4), 30 (2012)

24. Ovsjanikov, M., Bronstein, A.M., Bronstein, M.M., Guibas, L.J.: Shape Google: a computer vision approach to isometry invariant shape retrieval. In: ICCV Workshops. pp. 320–327. IEEE (2009)
25. Ovsjanikov, M., Li, W., Guibas, L., Mitra, N.J.: Exploration of continuous variability in collections of 3D shapes. TOG 30(4), 33 (2011)
26. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: ECCV, pp. 143–156. Springer (2010)
27. Rusu, R.B., Blodow, N., Beetz, M.: Fast point feature histograms (FPFH) for 3D registration. In: ICRA. pp. 3212–3217. IEEE (2009)
28. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the fisher vector: Theory and practice. IJCV 105(3), 222–245 (2013)
29. Schnitman, Y., Caspi, Y., Cohen-Or, D., Lischinski, D.: Inducing semantic segmentation from an example. In: ACCV, pp. 373–384. Springer (2006)
30. Schoeler, M., Papon, J., Wörgötter, F.: Constrained planar cuts-object partitioning for point clouds. In: CVPR. pp. 5207–5215 (2015)
31. Shapira, L., Shalom, S., Shamir, A., Cohen-Or, D., Zhang, H.: Contextual part analogies in 3D objects. IJCV 89(2), 309–326 (2010)
32. Shapira, L., Shamir, A., Cohen-Or, D.: Consistent mesh partitioning and skeletonisation using the shape diameter function. VC 24(4), 249–259 (2008)
33. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. IJCV 81(1), 2–23 (2009)
34. Sidi, O., van Kaick, O., Kleiman, Y., Zhang, H., Cohen-Or, D.: Unsupervised cosegmentation of a set of shapes via descriptor-space spectral clustering. In: TOG. vol. 30. ACM (2011)
35. Toldo, R., Castellani, U., Fusiello, A.: Visual vocabulary signature for 3D object retrieval and partial matching. In: Proc. of the 3DOR Conf. pp. 21–28 (2009)
36. Tombari, F., Salti, S., Di Stefano, L.: Unique signatures of histograms for local surface description. In: ECCV. pp. 356–369. Springer (2010)
37. Toshev, A., Shi, J., Daniilidis, K.: Image matching via saliency region correspondences. In: CVPR. pp. 1–8 (2007)
38. Van Kaick, O., Zhang, H., Hamarneh, G., Cohen-Or, D.: A survey on shape correspondence. In: CGF. vol. 30, pp. 1681–1707 (2011)
39. Wainwright, M.J., Jaakkola, T.S., Willsky, A.S.: Map estimation via agreement on trees: message-passing and linear programming. IEEE Trans. on Inf. Theory 51(11), 3697–3717 (2005)
40. Wang, F., Huang, Q., Ovsjanikov, M., Guibas, L.J.: Unsupervised multi-class joint image segmentation. In: CVPR. pp. 3142–3149 (2014)
41. Wang, Y., Asafi, S., van Kaick, O., Zhang, H., Cohen-Or, D., Chen, B.: Active co-analysis of a set of shapes. TOG 31(6), 165 (2012)
42. Xu, K., Li, H., Zhang, H., Cohen-Or, D., Xiong, Y., Cheng, Z.Q.: Style-content separation by anisotropic part scales. TOG 29(6), 184 (2010)
43. Yedidia, J.S., Freeman, W.T., Weiss, Y.: Constructing free-energy approximations and generalized belief propagation algorithms. IEEE Trans. on Inf. Theory 51(7), 2282–2312 (2005)
44. Zhang, H., Sheffer, A., Cohen-Or, D., Zhou, Q., Van Kaick, O., Tagliasacchi, A.: Deformation-driven shape correspondence. In: CGF. vol. 27, pp. 1431–1439 (2008)
45. Zheng, Y., Cohen-Or, D., Averkiou, M., Mitra, N.J.: Recurring part arrangements in shape collections. In: CGF. vol. 33, pp. 115–124 (2014)