Efficient 3D shape co-segmentation from single-view point clouds using appearance and isometry priors

Master thesis

Nikita Araslanov

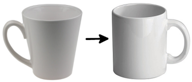
Compute Science Institute VI
University of Bonn

March 29, 2016

Contents

Motivation

Learning a new shape building upon knowledge acquired from similar shapes.



Many applications in robotics would profit from transfer of shape knowledge:

- determining grasping pose of unknown object having seen a similar one;
- human body tracking using a single body model;
- translating human body pose onto a (humanoid) robot for teleoperation or learning from demonstration.

Modelling as deformation with appropriate model:

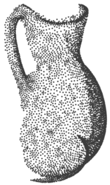
- articulated objects;
- easily deformable objects from soft materials.

Problem statement

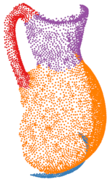
Co-segmentation problem:

- *Given*
  - union of the reference shape segments $\mathcal{S} = \bigcup \mathcal{S}_i$;
  - label mapping $\ell : \mathcal{S} \to L$;
  - query shape $\mathcal{T} := \{t_i \mid t_i \in \mathbb{R}^3\}$ as a point cloud.
- *Task*: find segmentation $\bigcup \mathcal{T}_i = \mathcal{T}$ with a mapping $\ell^\star : \mathcal{T} \to L$ such that $\ell^\star(\mathcal{T}_j) = \ell(\mathcal{S}_i)$ if and only if segments $\mathcal{S}_i$ and $\mathcal{T}_j$ represent semantically corresponding parts.

(a) Reference



(b) Query



(c) Ground truth

Previous work

Supervised: segment labels are provided

- Kalogerakis et al. (2010)[1].
- Kaick et al. (2011)[2].

Unsupervised: segmentation over an unlabelled object category

- Huang et al. (2011)[3].
- Sidi et al. (2011)[4].
- Meng et al. (2013)[5].

[1] Evangelos Kalogerakis, Aaron Hertzmann, and Karan Singh. "Learning 3D mesh segmentation and labeling". In: **ACM Transactions on Graphics (TOG)**. vol. 29. 4. ACM. 2010, p. 102.

[2] Oliver van Kaick et al. "Prior knowledge for part correspondence". In: **Computer Graphics Forum**. Vol. 30. 2. Wiley Online Library. 2011, pp. 553–562.

[3] Qixing Huang, Vladlen Koltun, and Leonidas Guibas. "Joint shape segmentation with linear programming". In: **ACM Transactions on Graphics (TOG)**. vol. 30. 6. ACM. 2011, p. 125.

[4] Oana Sidi et al. **Unsupervised co-segmentation of a set of shapes via descriptor-space spectral clustering**. Vol. 30. 6. ACM, 2011.

[5] Min Meng et al. "Unsupervised co-segmentation for 3D shapes using iterative multi-label optimization". In: **Computer-Aided Design** 45.2 (2013), pp. 312–320.

Previous work: Supervised

- Kalogerakis et al. (2010)[6].
- Kaick et al. (2011)[7].

Main idea: Learn Conditional Random Field (CRF):

$$E(\mathbf{x}) = \sum_i \phi(x_i) + \sum_{i,j} \phi(x_i, x_j),$$

where

- $\phi(x_i)$ models geometrical similarity of a single face by means of shape descriptors;
- $\phi(x_i, x_j)$ models segment boundaries.

Similarities:

- Shape descriptors (unary term);
- JointBoost classifier.

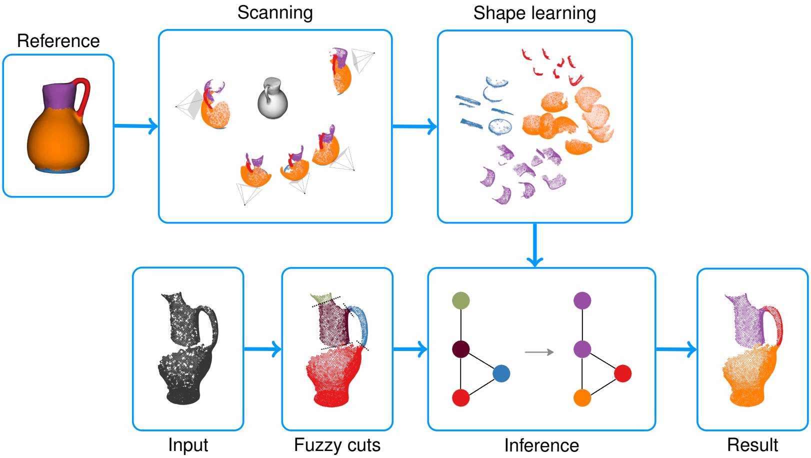
Differences:

- Inference (alpha expansion and alpha-beta swap).
- Pairwise features.

[6] Evangelos Kalogerakis, Aaron Hertzmann, and Karan Singh. "Learning 3D mesh segmentation and labeling". In: **ACM Transactions on Graphics (TOG)**. vol. 29. 4. ACM. 2010, p. 102.

[7] Oliver van Kaick et al. "Prior knowledge for part correspondence". In: **Computer Graphics Forum**. Vol. 30. 2. Wiley Online Library. 2011, pp. 553–562.

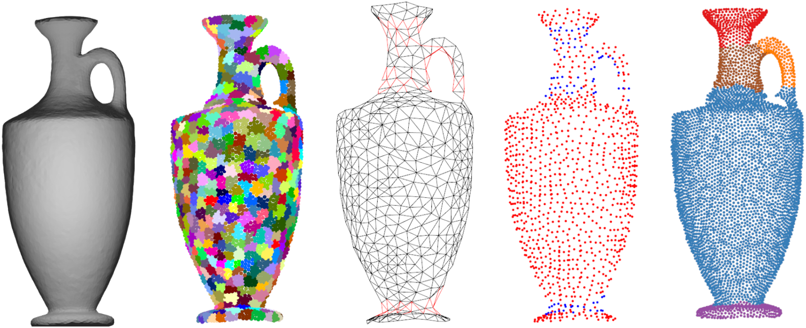
Method: Overview



Pre-segmentation

Based on the Constrained Planar Cuts segmentation[8].

1. Supervoxel segmentation
2. Construct edge cloud (induced by the edges of the supervoxels)
3. Classify points concave/convex
4. Cut concave points with RANSAC



[8] Markus Schoeler, Jeremie Papon, and Florentin Wörgötter. "Constrained Planar Cuts-Object Partitioning for Point Clouds". In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. 2015, pp. 5207–5215.

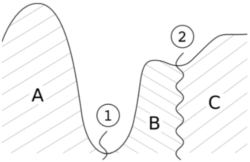
Pre-segmentation

Issue:

- Merging small segments to larger ones

Solution:

- Merging in the order of decreasing concavity



A > C > B

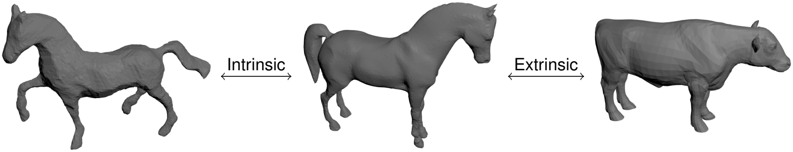
Algorithm 1: Modified CPC algorithm

```
// Original CPC
...
Initialise EdgeQueue from VoxelClusters and EdgesCut;
while EdgeQueue ≠ ∅ do
    (V1, V2) ← EdgeQueue.pop();
    if Score(V1, V2) < ScoreThreshold or
      |V1| < SizeThreshold or |V2| < SizeThreshold then
        MergeNodes (V1, V2), update EdgeQueue;
    end
end
...
```

Method: Model

Two groups of deformations[9]:

- extrinsic → shape part appearance;
- intrinsic → isometric.



- Part appearance modelled by $p(\ell_i \mid \mathcal{T}_i)$.
- Degree of isometric distortion $p(\ell_i, \ell_j \mid \mathcal{T}_i, \mathcal{T}_j)$.

$$\underset{\ell}{\text{maximize}} \quad \prod_{i,j} p(\ell_i \mid \mathcal{T}_i) p(\ell_j \mid \mathcal{T}_j) p(\ell_i, \ell_j \mid \mathcal{T}_i, \mathcal{T}_j),$$

[9] Alexander M Bronstein et al. "A Gromov-Hausdorff framework with diffusion geometry for topologically-robust non-rigid shape matching". In: **International Journal of Computer Vision** 89.2-3 (2010), pp. 266–286.

Method: shape appearance

Feature encoding based on random sampling:

1. *Feature packet*: a number of point clusters sampled from a sparse uniform grid (for each part and viewpoint);
2. Extract feature descriptors contained in each cluster;
3. Encode each cluster with **Bag-of-Words** or **Fisher** vector;
4. Average the vector encoding over all clusters in the packet.

Bag-of-Words

1. Extract feature descriptors $\mathbf{m}_{\ell,v,t}$ (SHOT) from each view v and part ℓ.
2. Fit Gaussian mixture model (GMM) $(w_k, \boldsymbol{\mu}_k, \Sigma_k)_k$
3. Encode

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

   for each cluster $|\rho_{\ell,v,i}|$.
4. Vectorise each feature packet $\mathcal{P}_{\ell,v,i}$ by taking the average over the clusters it contains:

$$f_{\text{BoW}}(\mathcal{P}_{\ell,v,i}) = \frac{1}{|\mathcal{P}_{\ell,v,i}|} \sum_{\rho_{\ell,v,i}} f_{\text{BoW}}(\rho_{\ell,v,i}), \quad \rho_{\ell,v,i} \in \mathcal{P}_{\ell,v,i}$$

5. Train with an RBF-kernel SVM;

Fisher vectors

1. Encode (FPFH):

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

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

where

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

2. Concatenate the gradients of each Gaussian centre:

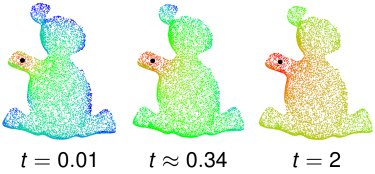
$$\mathbf{f}_{\mathrm{FV}}(\rho_{\ell,v,i}) = (G^T_{\boldsymbol{\mu}_1}(\rho_{\ell,v,i}), \ldots, G^T_{\boldsymbol{\mu}_K}(\rho_{\ell,v,i}), G^T_{\boldsymbol{\sigma}_1}(\rho_{\ell,v,i}), \ldots, G^T_{\boldsymbol{\sigma}_K}(\rho_{\ell,v,i}))^T.$$

3. Normalise[10] with $f(z) = sign(z)|z|^{\alpha}$.
4. Train with a linear SVM;

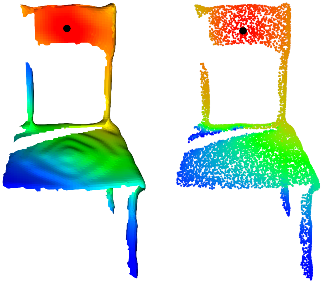
[10] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. "Improving the fisher kernel for large-scale image classification". In: **Computer Vision–ECCV 2010**. Springer, 2010, pp. 143–156.

Method: isometry prior

Diffusion distance: $d_t^2(x,y) = \sum_i K^{2t}(\lambda_i)(\phi_i(x) - \phi_j(y))^2$,
Commute time distance:$d_{\mathrm{CT}}^2(x,y) = \sum_i \frac{1}{\lambda_i}(\phi_i(x) - \phi_j(y))^2$,
where $\phi_i(\cdot)$ and λ_i are eigenfunctions and eigenvalues of the Laplace-Beltrami operator[11].



(a) Diffusion distance

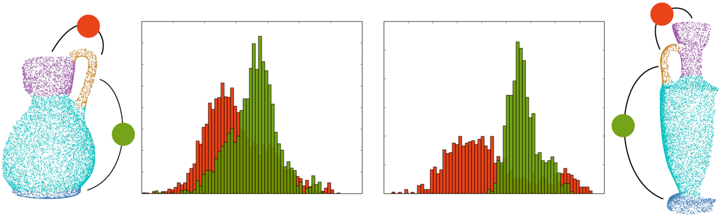


(b) Geodesic (left) and commute time distance (right)

[11] Jian Liang et al. "Geometric understanding of point clouds using laplace-beltrami operator". In: **Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on**. IEEE. 2012, pp. 214–221.

Method: isometry prior

Idea: model isometric distortion between shape parts with a distribution of diffusion distances[12]



1. extract CT distances between each pair of shape parts (including itself);
2. fit Gaussian mixture model;
3. for a new pre-segmented shape

$$p(\ell_{i\sim i'}, \ell_{j\sim j'} \mid D_{CT}(\mathcal{T}_i, \mathcal{T}_j)) = \frac{p(D_{CT}(\mathcal{T}_i, \mathcal{T}_j) \mid \ell_{i\sim i'}, \ell_{j\sim j'})}{\sum_{i'',j''} p(D_{CT}(\mathcal{T}_i, \mathcal{T}_j) \mid \ell_{i\sim i''}, \ell_{j\sim j''})}$$

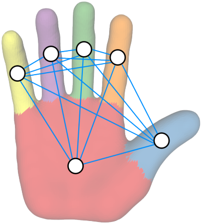
4. CRF parameter λ: $\sigma' := (1+\lambda)\sigma$ (learned using pre-segmentation).

[12] Michael M Bronstein and Alexander M Bronstein. "Shape recognition with spectral distances". In: **IEEE Transactions on Pattern Analysis & Machine Intelligence** 5 (2010), pp. 1065–1071.

Method: CRF

Re-formulate the objective:

$$\underset{\ell}{\text{minimize}} \quad -\sum_i \log p(\ell_i \mid \mathcal{T}_i) - \sum_{i,j} \log p(\ell_i, \ell_j \mid \mathcal{T}_i, \mathcal{T}_j).$$



Features:

- Complete graph;
- Moderate size (max. 30 nodes).

Inference with A*:

- Convergence to a global optimum (with an admissible heuristic);
- More efficient than belief propagation[13].

[14] Martin Bergtholdt et al. "A study of parts-based object class detection using complete graphs". In: **International Journal of Computer Vision** 87.1-2 (2010), pp. 93–117.

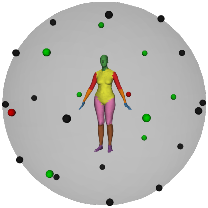
Experiment I: Dataset

Dataset: Labelled Princeton Segmentation Benchmark[15].

- 19 (15 selected) categories derived from Princeton Segmentation Benchmark[16].
- Manual ground-truth labelling based on average human segmentation.

Generating random views

- uniform grid on a sphere;
- *valid* if at least 20% of each the shape part visible;
- select at most 8 viewpoints with maximum spread.



[15] Evangelos Kalogerakis, Aaron Hertzmann, and Karan Singh. "Learning 3D mesh segmentation and labeling". In: **ACM Transactions on Graphics (TOG)**. vol. 29. 4. ACM. 2010, p. 102.

[16] Xiaobai Chen, Aleksey Golovinskiy, and Thomas Funkhouser. "A benchmark for 3D mesh segmentation". In: **ACM Transactions on Graphics (TOG)**. vol. 28. 3. ACM. 2009, p. 73.

Experiment I: Criteria

- **Accuracy**: % of area labelled correctly.
- **Hamming distance**: the average of the missing rate and false alarm rate:

$$R_m(\mathcal{S},\mathcal{T}) = \frac{D_H(\mathcal{S} \Rightarrow \mathcal{T})}{\|\mathcal{T}\|} \quad R_f(\mathcal{S},\mathcal{T}) = \frac{D_H(\mathcal{T} \Rightarrow \mathcal{S})}{\|\mathcal{S}\|},$$

where $D_H(\mathcal{S} \Rightarrow \mathcal{T}) := \sum_{\mathcal{S}_i \sim \mathcal{T}_j} \|\mathcal{T}_j \setminus \mathcal{S}_i\|$ is the Directional Hamming Distance.

- **Rand index**: the likelihood that a pair of faces is either in the same or different segments in two segmentations: $R = \binom{N}{2}^{-1}(a+b)$, where
  - the number of pairs of faces a in the same segment;
  - the number of pairs of faces b in different segments.
- **Local Consistency Error (LCE)**:

$$LCE(\mathcal{S},\mathcal{T}) = \frac{1}{N}\sum_i \min\left\{E_i(\mathcal{S},\mathcal{T}), E_i(\mathcal{T},\mathcal{S})\right\}.$$

- **Global Consistency Error (GCE)**:

$$GCE(\mathcal{S},\mathcal{T}) = \frac{1}{N}\min\left\{\sum_i E_i(\mathcal{S},\mathcal{T}), \sum_i E_i(\mathcal{T},\mathcal{S})\right\}$$

Experiment I: Results

Category	van Kaick et al.	Kalogerakis et al.	BoW	BoW+ISO	FV	FV+ISO
Ant	58.8	58.9	66.2	65.6	**77.7**	74.1
Airplane	62.7	62.0	59.2	57.0	**64.0**	60.0
Bird	58.1	57.0	57.4	52.0	**58.5**	53.6
Chair	59.6	59.6	**60.6**	56.7	60.2	55.5
Cup	81.6	81.8	**90.0**	87.6	88.7	87.5
Fish	84.2	**84.4**	72.1	71.7	78.4	77.7
Fourleg	**60.1**	59.4	51.1	48.1	54.9	50.6
Hand	52.2	52.7	53.4	46.8	**56.0**	49.6
Human	41.3	41.6	35.8	34.2	**43.7**	40.4
Mech	81.3	81.7	82.4	84.4	84.1	**84.6**
Octopus	82.0	**82.8**	76.5	75.0	69.6	69.8
Plier	33.7	32.5	70.5	57.3	**71.9**	58.8
Table	71.6	70.9	**88.9**	87.5	85.4	84.1
Teddy	71.9	71.1	64.5	69.4	76.4	**77.0**
Vase	64.3	65.5	**70.6**	65.3	70.3	63.8
Average	64.2	64.1	66.6	63.9	**69.3**	65.8

Figure : Average accuracy on the LPSB dataset used in Experiment I (in percent)

Experiment I: Results

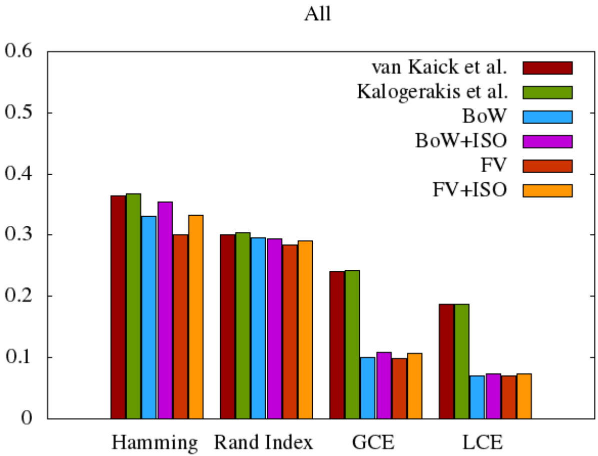
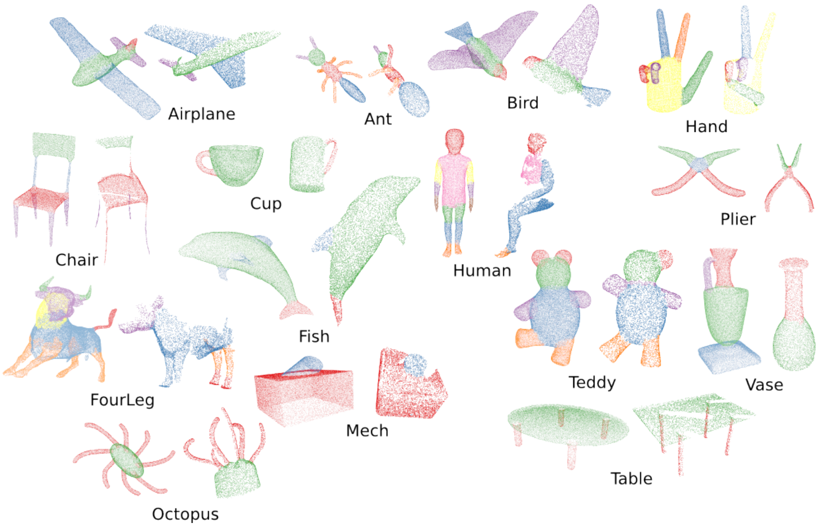


Figure : The average performance of different co-segmentation algorithms for **all** categories used in Experiment I

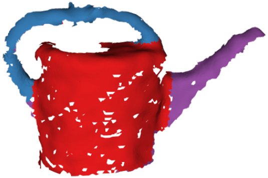
Experiment I: Results



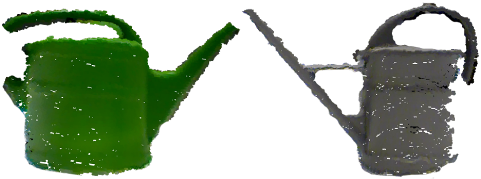
Experiment II: Setup

Experiment with real point cloud data recorded with ASUS Xtion sensor.

- Comparison of FV with the method of van Kaick et al. (2011)[17].
- **Given**:
  - manually labelled watercan (from partial views);
- **Query**:
  1. single views of the same watercan (new sequence);
  2. single views of a different watercan.



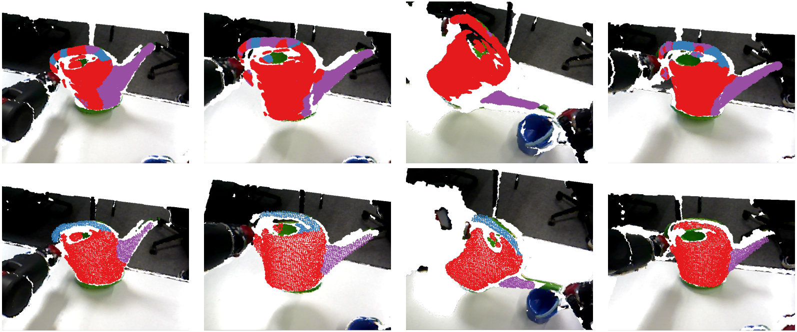
(a) Reference shape



(b) Query shapes

[17] Oliver van Kaick et al. "Prior knowledge for part correspondence". In: Computer Graphics Forum. Vol. 30. 2. Wiley Online Library. 2011, pp. 553–562.

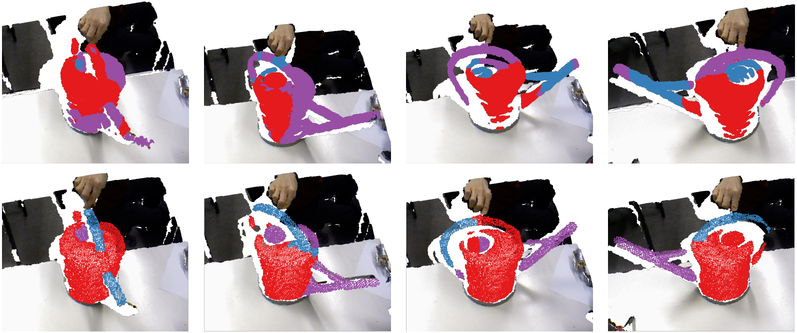
Experiment II: Results (1)



Frame 6 Frame 34 Frame 56 Frame 99

Figure : Test sequence with the same query shape as the reference. **Top row:** van Kaick et al.; **Bottom row:** Ours (FV).

Experiment II: Results (2)



Frame 2 Frame 17 Frame 22 Frame 28

Figure : Test sequence with a novel query shape. **Top row:** van Kaick et al.; **Bottom row:** Ours (FV).

Experiment II: Results (3)

	van Kaick et al.	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**

Figure : Average time per object pair in Experiment II (in seconds)

- Hardware: Intel Core i7, 8GB RAM.
- C++ implementation, OpenMP for face- and pointwise operations (e.g. normal estimation).
- Feature computation is included in the "Inference" step.
- FV is almost x6 times faster.

Limitations & Future work

Limitations:

- weak link between pre-segmentation and inference
  - pre-segmentation provides an upper-bound on overall performance.
- concavity is not the only cue of the segment boundaries and it can be occluded in partial views.
- limited use of the proposed isometry prior.

Future work:

- Improvement of the context features (isometric distortion):
  - Other Laplace-Beltrami approximations exist for point clouds.
  - Diffusion distance can be approximated with Euclidean distance and a Gaussian kernel.
- Other feature encoding schemes, such as spatial sensitive Bag-of-Words, may improve the performance.

Conclusions

- new co-segmentation approach;
- can be applied to single frames of point clouds captured with RGB-D sensor;
- does not require a complete model
  - be learned from a sequence of partial views).
- efficient inference with strong optimality guarantees.

Thank you!