Scene Cut: Class-specific object detection and segmentation in 3D scenes

Jan Knopp, Mukta Prasad, Luc Van Gool

VISICS, KU Leuven, Belgium
BIWI, ETH Zurich, Switzerland
The goal of the work is to find objects of a specific class in the scene.
Why is it difficult?

The problem of a successful scene segmentation comes with the following challenges:

- Collision of objects

- Global context of local parts

- Real-life data is often noisy and uncomplete
Why is it important?

Successful segmentation opens doors for the following applications:

• Scene understanding

• Robot navigation, range maps

• Create layers of the scene automatically
Motivation

SOTA vs. ours

Method

Results & Summary
Related work

How to detect and segment objects in the scene?


• Kalogerakis et. al. “Learning 3D mesh segement...”, In ACM Trans. On Grph., 2010
We combined a voting-based ISM class recognition approach with Graph-Cuts segmentation to find class-specific objects in 3D scenes.

- **GraphCut**: segmentation as a graph optimization problem
- **Detection**: find the object's location
- **Back-projection**: model the object characteristics
Motivation

SOTA vs. ours

Method

Results & Summary
MRF based segmentation

1) foreground obeys object class characteristics
2) foreground/background segments should be smooth
MRF based segmentation

The goal is to find the labelling $\mathbf{f}$ that maximizes the distribution $p$ in the scene $S$ of vertices $\mathbf{v}$ with descriptors $d$ and edges $E$,

$$p(\mathbf{f}|S, \theta) = \frac{1}{Z(S, \theta)} \exp\left(-Q(\mathbf{f}; S, \theta)\right)$$

$$Q(\mathbf{f}; S, \theta) = \sum_{i=1}^{N_v} \sum_{j|(i,j) \in E} \left( \phi(\mathbf{v}_k; f_k, \theta) + \Lambda_\psi \cdot \psi(f_i, f_j; \theta) + \Lambda_\chi \cdot \chi(f_i, f_j, d_i, d_j; \theta) \right)$$

1) foreground obeys object class characteristics
2) foreground/background segments should be smooth
MRF based segmentation

The goal is to find the labelling $f$ that maximizes the distribution $p$ in the scene $S$ of vertices $v$ with descriptors $d$ and edges $E$,

$$p(f|S, \theta) = \frac{1}{Z(S, \theta)} \exp \left( -Q(f; S, \theta) \right)$$

$$Q(f; S, \theta) = \sum_{i=1}^{N_v} \left( \sum_{j|(i,j) \in E} \left( \Lambda_{\psi} \cdot \psi(f_i, f_j; \theta) + \Lambda_{\chi} \cdot \chi(f_i, f_j, d_i, d_j; \theta) \right) + \phi(v_k; f_k; \theta) \right)$$

**Unary**: is the specific vertex/node foreground or not?
MRF based segmentation

The goal is to find the labelling $f$ that maximizes the distribution $p$ in the scene $S$ of vertices $v$ with descriptors $d$ and edges $E$,

$$p(f|S, \theta) = \frac{1}{Z(S, \theta)} \exp\left(-Q(f; S, \theta)\right)$$

$$Q(f; S, \theta) = \sum_{i=1}^{N_v} \left( \sum_{j|(i,j) \in E} \left( \Lambda_\psi \cdot \psi(f_i, f_j; \theta) + \Lambda_\chi \cdot \chi(f_i, f_j, d_i, d_j; \theta) \right) + \phi(v_k; f_k, \theta) \right)$$

**Unary**: is the specific vertex/node foreground or not?

**Ising pairwise prior**: neighbours have same label.
MRF based segmentation

The goal is to find the labelling $\mathbf{f}$ that maximizes the distribution $p$ in the scene $\mathcal{S}$ of vertices $\mathbf{v}$ with descriptors $d$ and edges $\mathbf{E}$,

$$p(\mathbf{f}|\mathcal{S}, \theta) = \frac{1}{Z(\mathcal{S}, \theta)} \exp\left(-Q(\mathbf{f}; \mathcal{S}, \theta)\right)$$

$$Q(\mathbf{f}; \mathcal{S}, \theta) = \sum_{i=1}^{N_v} \left( \sum_{j|(i,j)\in\mathbf{E}} \phi(\mathbf{v}_k; f_k, \theta) + \psi(f_i, f_j; \theta) + \chi(f_i, f_j, d_i, d_j; \theta) \right)$$

**Unary**: is the specific vertex/node foreground or not?

**Pairwise Prior**: neighbours have same label.

**Data-dependent Pairwise**: similar vertex characteristics should have the same label.
The power of the detection method was used to define the unary term. The descriptors of the training shapes are clustered. ISM represents shapes as a collection of these cluster centres (visual words).

The **ISM** [1,2] is the collection of visual words together with their position related to the object centre:

---

Using the trained ISM model, we find the centre of the object of interest.
ISM based unary term

The centre of the object is then used to back-project the training data and to construct the unary term.

**Unary term:**

\[
\frac{1}{\rho} \sum_{i=1}^{\rho} \exp \left( -\frac{||v_k - b_k^i||^2}{\alpha^2} \right) + f_c(v_k)
\]

where \( b \) is \( i \)-th closest back-projection of \( k \)-th vertex \( v \). \( f \) models whether the vertex voted correctly for the object centre or not.
**Pairwise term**

**Ising smoothness prior:** neighbours should have the same label.

**Pairwise data-dependent:**

1. **Vertex similarity**

   Based on the difference of descriptors of vertices:

   $$\chi_v(i, j) = \exp \left( \frac{-\|d_i - d_j\|^2}{\beta^2} \right)$$

2. **“Plane-like” surface**

   Edge smoothness measured using the angle between faces.
Half of the TOSCA [1] dataset was used for training. Models excluded from training composed a test scene. Additionally, the Light contest scene [2] together with the TOSCA shape was used. In all experiments, we used 3DSURF [3] as the shape descriptor.

**References:**


Half of TOSCA [1] dataset used for training. Models excluded from training were merged to create a test scene.

<table>
<thead>
<tr>
<th></th>
<th>Ising</th>
<th>Ising + Vert. sim.</th>
<th>Ising + Plane-like</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>VW</strong></td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td><strong>VW-tfidf</strong></td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td><strong>Multi-VW</strong></td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td><strong>Desc-NN</strong></td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td><strong>ISM</strong></td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
</tbody>
</table>

**ISM:**

Experiments real-life data

Process of obtaining data:
- Set of images

Training data:
- Arc3D

SfM algorithm

Reconstructed model

Results:
- Set of images
- SfM algorithm
- Reconstructed model
Summary

• Segmentation using the power of the ISM based detection approach gives an important improvement.

• No additional scene processing such as background plane filtering is required.

• Segmentation of clean shapes as well as noisy and incomplete models from SfM.
Thank you for your attention!