A Comparison of Geometric and Energy-Based Point Cloud Semantic Segmentation Methods



Mathieu Dubois¹, Paola K. Rozo², Alexander Gepperth¹, Fabio A. González O.² and David Filliat¹

¹ENSTA ParisTech - INRIA Flowers Team

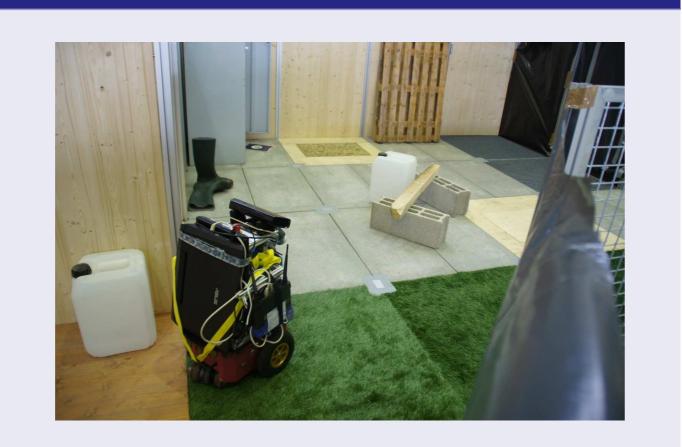
²Universidad Nacional de Colombia mathieu.dubois@ensta.fr



Context

Semantic navigation for indoor robots: mapping recognize objects, rooms, *etc.* Low-cost RGB-D cameras: use depth information

CAROTTE competition http://www.defi-carotte.fr

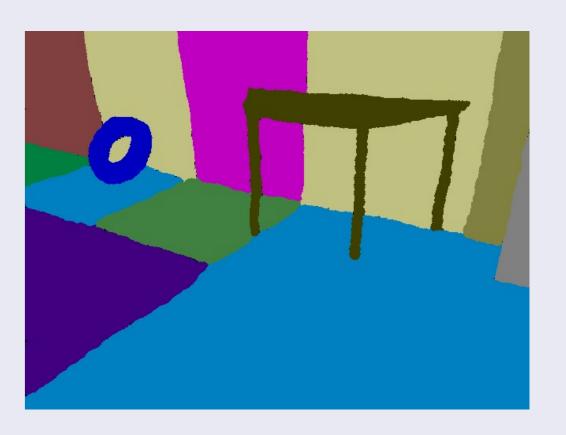


Problem

Object recognition at the category level is difficult:

 \Rightarrow focus on segmentation: try to recognize the structure of the environment (walls, ground) and the presence of objects Multimodal segmentation:

 $\hat{\boldsymbol{x}} = \arg \max P(\boldsymbol{x} | \boldsymbol{A}, \boldsymbol{S}, \boldsymbol{G})$



Baseline system

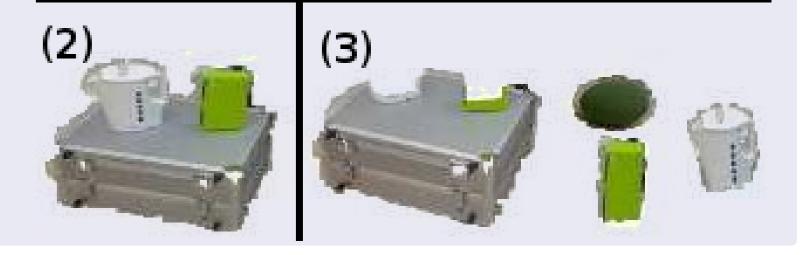
Developed for the PACOM system (Filliat et al. 2012), inspired by Rusu et al. 2009

Purely geometric

- Detect the ground plane and remove the points
- Detect walls *i.e.* planes perpendicular to ground and remove the points (1)
- Project remaining points and group them: objects (2)
- Decompose objects into planes and regroup them (3)
- + No training, very good performance, decompose objects
- Many parameters (including robot specific)



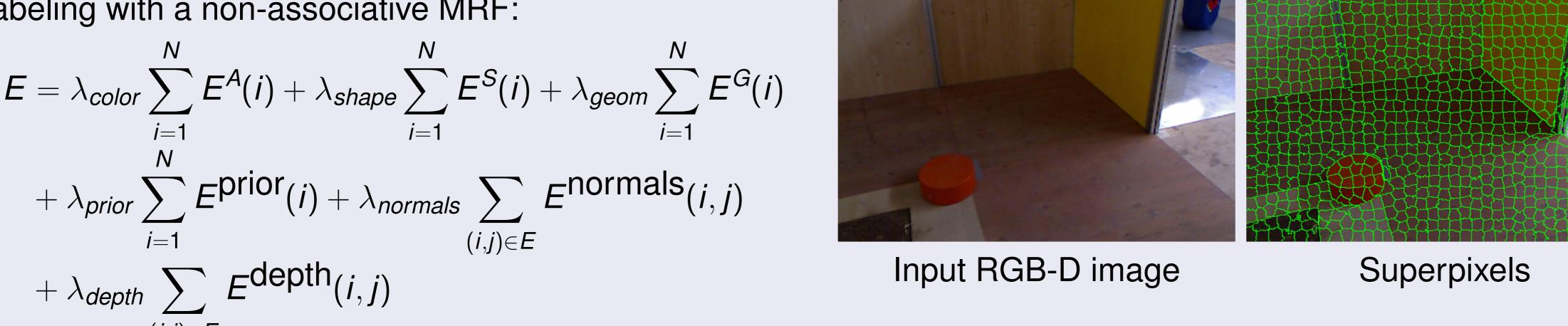


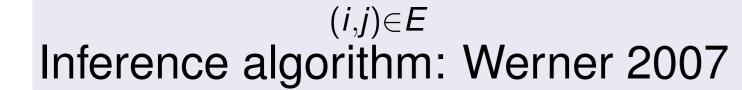


Labeling

MRF-based system

Multimodal over-segmentation with an extension of SLIC (Achanta et al. 2012) 2 Labeling with a non-associative MRF: $E = \lambda_{color} \sum_{i=1}^{N} E^{A}(i) + \lambda_{shape} \sum_{i=1}^{N} E^{S}(i) + \lambda_{geom} \sum_{i=1}^{N} E^{G}(i)$





Multimodal SLIC Superpixels

Adaptation of the k-means algorithm with local search (linear complexity)

Distance function: given 2 pixels *i* and *j*:

 $D(i,j) = \sqrt{d_c(i,j)^2 + \frac{m^2}{S^2} d_s^2(i,j)}$

where:

$$egin{aligned} & d_c(i,j) = \sqrt{(l_j - l_i)^2 + (a_j - a_i)^2 + (b_j - b_i)^2} \ & d_s(i,j) = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2} \end{aligned}$$

Large *m* enforce compact superpixels, small *m* enforce adherence to image boundaries

Energies (green = learned parameter)

Unary:

- $\blacksquare E^A(\ell) = -\log P(w^A|\ell)$ where w^A is the quantized SIFT descriptor $= E^{S}(\ell) = -\log P(w^{S}|\ell)$ where w^{S} is the quantized depth descriptor (Shotton et al.
- 2011) $\mathbf{E}^{G}(\ell) = -\log \mathcal{N}(\boldsymbol{g}|\boldsymbol{\mu}_{\ell}, \boldsymbol{\Sigma}_{\ell})$ where
- $\boldsymbol{g} = [\boldsymbol{x}, \boldsymbol{y}]^T$ is 2D position
- $= E^{\text{prior}}(\ell) = -\log P(\ell)$
- \Rightarrow Necessary for unbalanced dataset

Easy to learn: discrete or Gaussian PDF

Binary:

 $= E^{\text{normals}}(\ell_1, \ell_2) = -\log P(\ell_1, \ell_2 | \overline{\phi}) \text{ where } \overline{\phi} \text{ is the }$ quantized angle between normals \Rightarrow Enforce detection of different surfaces

Database

Acquired during competition

Autonomous robot navigation

using the baseline method

- 100 manually labeled point clouds
- 3 classes (wall, ground, object) decomposed into detailed classes (9 walls, 8 grounds, 16 objects) Unbalanced
- + Usable for segmentation and recognition

Results: 5-fold cross-validation							
Algorithm	Precision			Recall			Overall
	Walls	Ground	Objects	Walls	Ground	Objects	Overall
Baseline	93.3	97.8	65.0	87.7	91.2	98.1	94.9
MRF strong regul. ($\lambda_{normal} = \lambda_{depth} = 1.0$)	96.4	89.5	46.8	77.6	79.5	41.6	76.0
MRF weak regul. ($\lambda_{normal} = \lambda_{depth} = 0.2$)		89.4	88.1	82.8	80.4	23.5	77.86
MRF no regul ($\lambda_{normal} = \lambda_{depth} = 0$)	94.7	88.8	64.8	81.1	78.5	32.1	76.8

MRF is less good than domain-specific algorithm but gives interesting results MRF models have several advantages:

 \Rightarrow We use only 3 classes http://cogrob.ensta.fr/pacom

More generic Less tuning

■ Use appearance Probabilistic output

References

R. Achanta et al. "SLIC Superpixels Compared to State-of-the-art Superpixel Methods". In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2012).

- D. Filliat et al. "RGBD object recognition and visual texture classification for indoor semantic mapping". In: Proceedings of the 4th International Conference on Technologies for Practical Robot Applications (TePRA). 2012, pp. 127–132.
- R. B. Rusu et al. "Close-range scene segmentation and reconstruction of 3D point cloud maps for mobile manipulation in domestic environments". In: *IROS*. IEEE, 2009, pp. 1–6.
- J. Shotton et al. "Real-time human pose recognition in parts from single depth images". In: CVPR. IEEE, 2011, pp. 1297–1304.
- T. Werner. "A linear programming approach to max-sum problem: A review". In: IEEE PAMI 29.7 (2007), pp. 1165–1179.