Towards Continuous Learned Semantic Representation through a Viewpoint-Dependent Observation Model

Yuri Feldman
Department of Computer Science, Technion, Israel Institute of Technology

Vadim Indelman
Department of Aerospace Engineering, Technion, Israel Institute of Technology

ANPL Autonomous Navigation and Perception Lab

TASP TECHNION AUTONOMOUS SYSTEMS PROGRAM

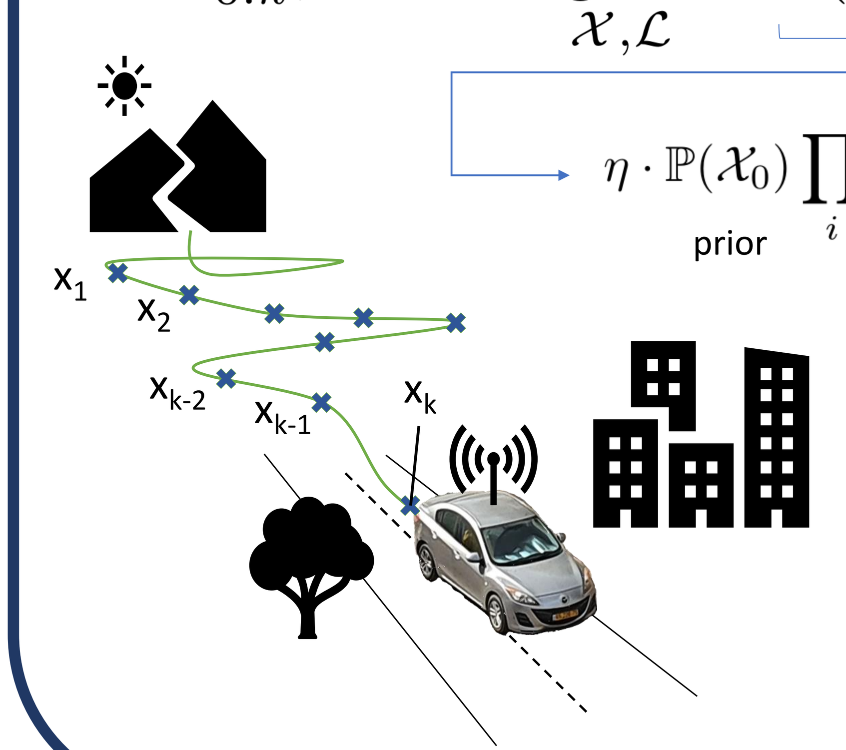
THE ISRAELI SMART TRANSPORTATION RESEARCH CENTER

1.a Background - Simultaneous Localization and Mapping (SLAM)

- SLAM is commonly formulated as joint max a-posteriori over agent poses $\mathcal{X}_{0:k}$ and landmarks $\mathcal{L}$ given history of observations and user controls.

$$\mathcal{X}^*_{0:k}, \mathcal{L}^* = \arg\max_{\mathcal{X},\mathcal{L}} \mathbb{P}(\mathcal{X}_{0:k}, \mathcal{L} \mid \mathcal{H}_k), \quad \mathcal{H}_k = \{\underbrace{\mathcal{Z}_{0:k}}_{\text{observations}}, \underbrace{\mathcal{U}_{0:k-1}}_{\text{user controls}}\}$$

$$\mathbb{P}(\mathcal{X}_{0:k}, \mathcal{L} \mid \mathcal{H}_k) = \eta \cdot \underbrace{\mathbb{P}(\mathcal{X}_0)}_{\text{prior}} \prod_i \underbrace{\mathbb{P}(\mathcal{X}_{i+1} \mid \mathcal{U}_i, \mathcal{X}_i)}_{\text{motion model}} \prod_{j \in \mathcal{M}_i} \underbrace{\mathbb{P}(\mathcal{Z}_j \mid \mathcal{L}_j, \mathcal{X}_i)}_{\text{observation model}}$$



© https://morioh.com/p/bd1b6fcgdgeb

See e.g. [1,2].

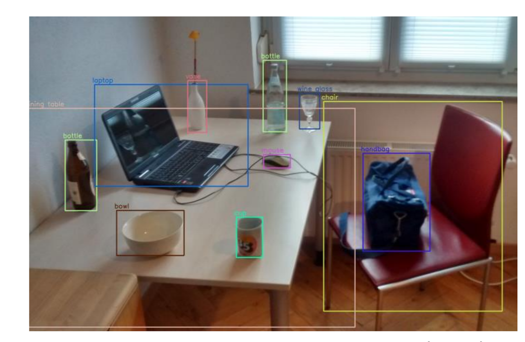
1.b Background – Object-Level SLAM

- In *Object SLAM* (e.g. [3-8]) mapping is done on the level of objects.

$$\arg\max_{\mathcal{X},\mathcal{C},\mathcal{O}} \mathbb{P}(\underbrace{\mathcal{X}_{0:k}}_{\text{robot track}}, \underbrace{\mathcal{C}}_{\text{object categories (discrete!)}}, \underbrace{\mathcal{O}}_{\text{object geometry}} \mid \underbrace{\mathcal{H}_k}_{\text{measurement and control history}})$$

$$= \arg\max_{\mathcal{X},\mathcal{C},\mathcal{O}} \underbrace{\mathbb{P}(\mathcal{X}_{0:k}, \mathcal{O} \mid \mathcal{C}, \mathcal{H}_k)}_{\text{Continuous hypothesis}} \cdot \underbrace{\mathbb{P}(\mathcal{C} \mid \mathcal{H}_k)}_{\text{Hypothesis weight}}$$

$$\mathbb{P}(\mathcal{X}_{0:k}, \mathcal{O} \mid \mathcal{C}, \mathcal{H}_k) = \eta \cdot \mathbb{P}(\mathcal{X}_0) \prod_i \underbrace{\mathbb{P}(\mathcal{X}_{i+1} \mid \mathcal{U}_i, \mathcal{X}_i)}_{\text{motion model}} \prod_{j \in \mathcal{M}_i} \underbrace{\mathbb{P}(\mathcal{Z}_j \mid \mathcal{O}_j, \mathcal{X}_i, \mathcal{C}_j)}_{\text{observation model (one per class!)}}$$



© Wikipedia

Drawbacks:

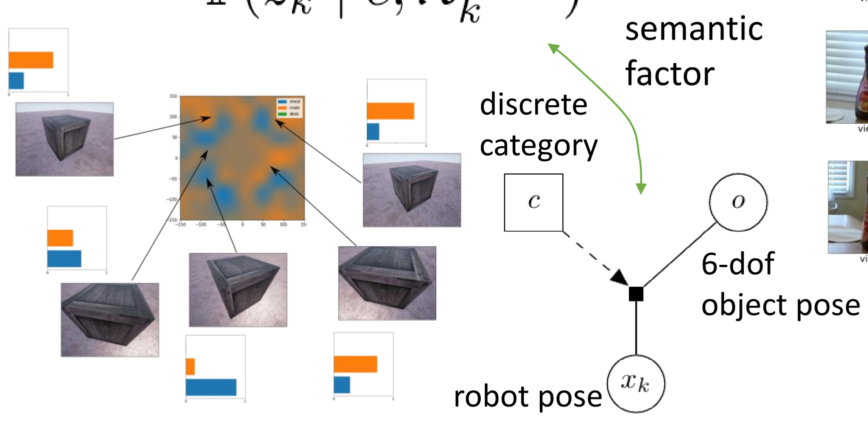
- Per-class models required
- Mixed inference, exponential number of hypotheses:

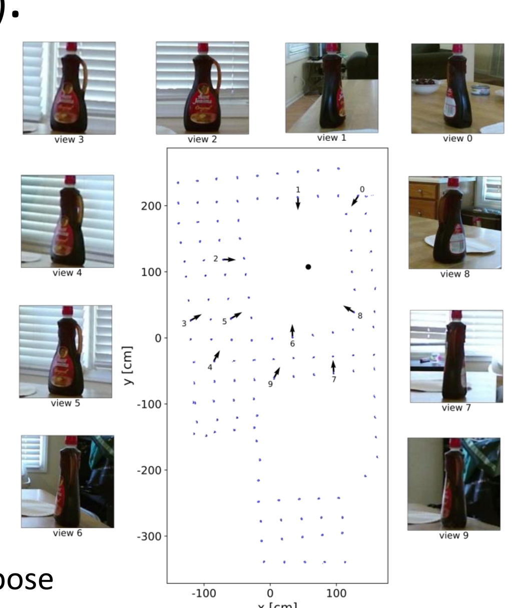
$$c \in \{1, \dots, N\} \qquad |\{(c_1, \dots, c_m)\}| = N^m$$

1.c Viewpoint-Dependent Semantic Models

- Viewpoint-dependent models couple inference of geometry and semantics, making them mutually beneficial (see [9-11]).

$$\mathbb{P}(\underbrace{z_k}_{\text{current observation}} \mid \underbrace{c}_{\text{object class}}, \underbrace{\mathcal{X}_k^{(rel)}}_{\text{viewpoint relative to object}})$$



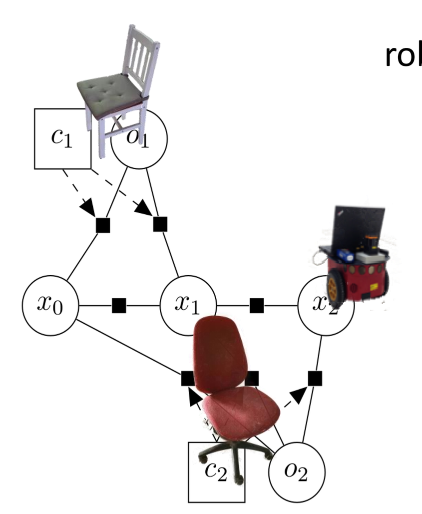


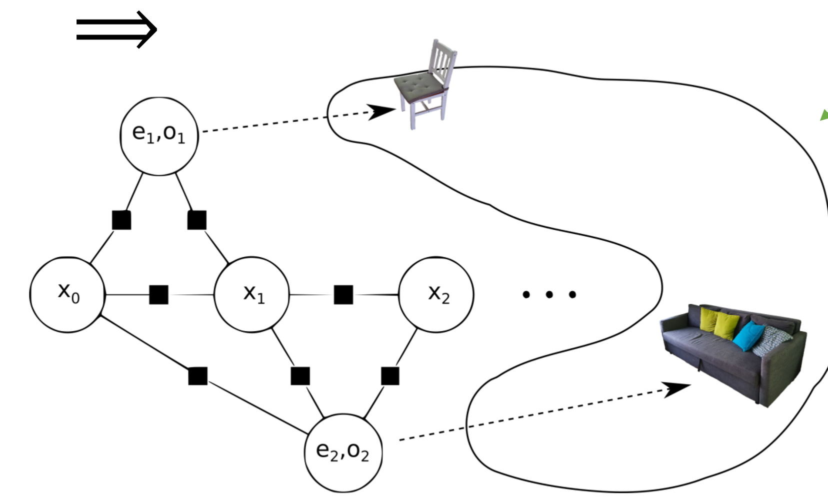
Feldman and Indelman. "Spatially-dependent Bayesian semantic perception under model and localization uncertainty." Autonomous Robots (2020): 1-29.

2. Continuous Semantic Representation

- Idea: learn a continuous semantic representation

$$\mathbb{P}(\underbrace{\mathcal{X}_{0:k}}_{\text{robot track}}, \underbrace{\mathcal{C}}_{\text{object categories (discrete)}}, \underbrace{\mathcal{O}}_{\text{object geometry}} \mid \underbrace{\mathcal{H}_k}_{\text{measurement and control history}}) \Rightarrow \mathbb{P}(\mathcal{X}_{0:k}, \underbrace{\mathcal{E}}_{\text{continuous representation}}, \mathcal{O} \mid \mathcal{H}_k)$$





Advantages:

- Continuous inference
- Relaxed ground-truth requirements
- Learned semantic representation

2 (cont'd). Approach

- Training:

$$\arg\max_{\theta,\mathcal{E}_{1:n}} \mathbb{P}\left(\underbrace{\mathcal{Z}_{0:k}}_{\text{observations}} \mid \underbrace{\mathcal{E}_{1:n}}_{\text{representation variables for n objects}}, \underbrace{\mathcal{X}_{0:k}^{(rel)}}_{\text{corresponding relative poses}}, \underbrace{\beta_{0:k}}_{\text{data association}}\right)$$

$$= \arg\max_{\theta,\mathcal{E}_{1:n}} \sum \underbrace{\log \mathbb{P}_\theta\left(\mathcal{Z}_i \mid \mathcal{E}_{\beta_i}, \mathcal{X}_i^{(rel)}\right)}_{\text{viewpoint - dependent observation model}} + \log \mathbb{P}(\mathcal{E}_{1:n})$$

- Inference:

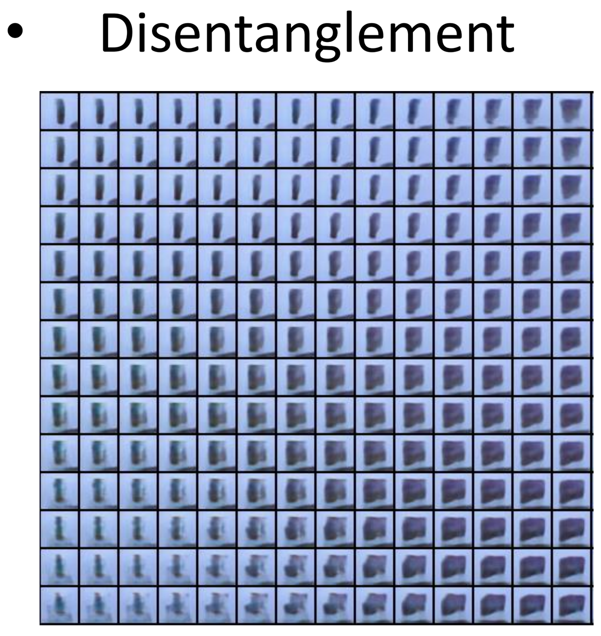
$$\arg\max_{\mathcal{X}_{0:k},\mathcal{O},\mathcal{E}} \mathbb{P}(\underbrace{\mathcal{X}_{0:k}}_{\text{robot track}}, \underbrace{\mathcal{O}}_{\text{object geometry}}, \underbrace{\mathcal{E}}_{\text{object representations}} \mid \mathcal{H}_k)$$

$$= \arg\max_{\mathcal{X}_{0:k},\mathcal{O},\mathcal{E}} \mathbb{P}(\mathcal{X}_0)\, \mathbb{P}(\mathcal{E}) \prod_i \mathbb{P}(\mathcal{X}_{i+1} \mid \mathcal{U}_i, \mathcal{X}_i) \prod_{j \in \mathcal{M}_i} \mathbb{P}(\mathcal{Z}_j \mid \mathcal{O}_j, \mathcal{X}_i, \mathcal{E}_{\beta_j})$$

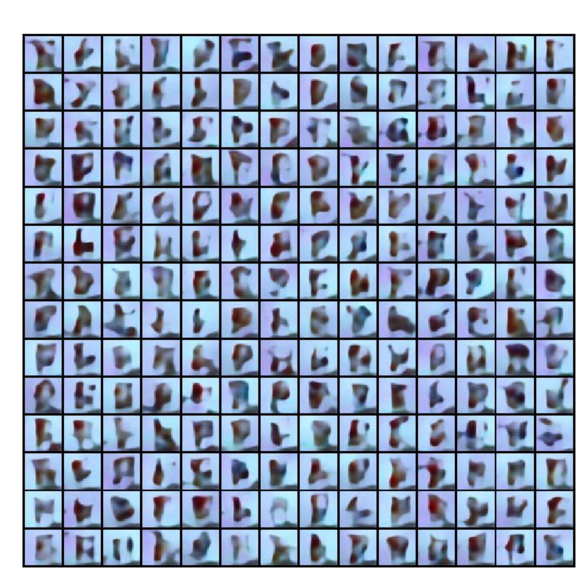
More details in Feldman and Indelman 20' Workshop on Self-Supervised Robot Learning, in conjunction with Robotics: Science and Systems (RSS), and future publications.

3. (Initial) Results

- Disentanglement

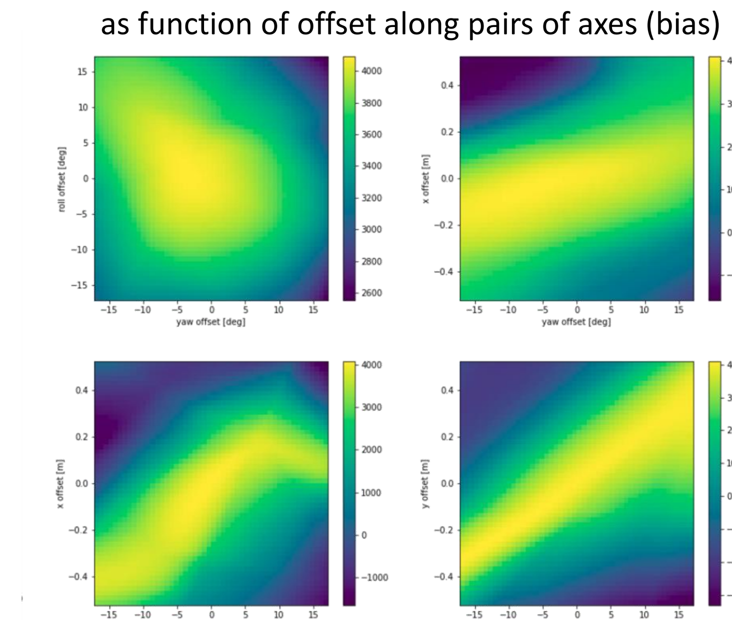


predictions for varying X

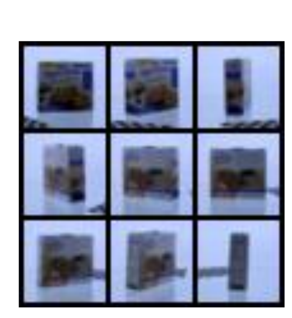


predictions for randomized e

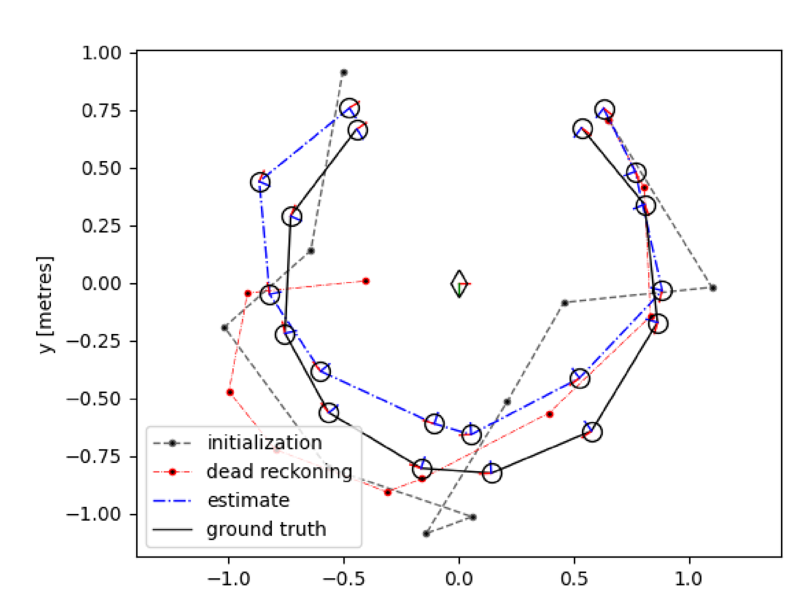
- Learned likelihood
  as function of offset along pairs of axes (bias)



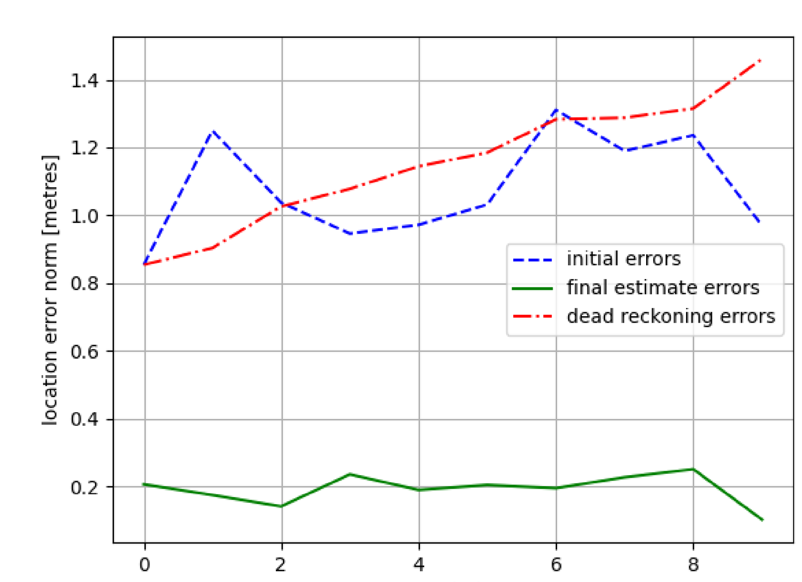
- Inference



Object views



Ground truth and estimated tracks



Evolution of estimation error

Bibliography

[1] Kaess et al. 08' TRO (iSAM)
[2] Kaess et al. 12' IJRR (iSAM2)
[3] Salas-Moreno et al. 13' CVPR
[4] Choudhary et al. 14' IROS
[5] Bowman et al. 17' ICRA
[6] McCormack et al. 18 3dv
[7] Nicholson et al. 19' RAL
[8] Yang and Scherer 19' TRO
[9] Feldman and Indelman 20' ARJ
[10] Tchuiev et al. 19' IROS
[11] Kopitkov and Indelman 18' IROS

DEPARTMENT OF AEROSPACE ENGINEERING | TECHNION Israel Institute of Technology

TECHNION | The Henry and Marilyn Taub Faculty of Computer Science