EM-Fusion: Dynamic Object-Level SLAM With Probabilistic Data Association

Michael Strecke and Jörg Stückler

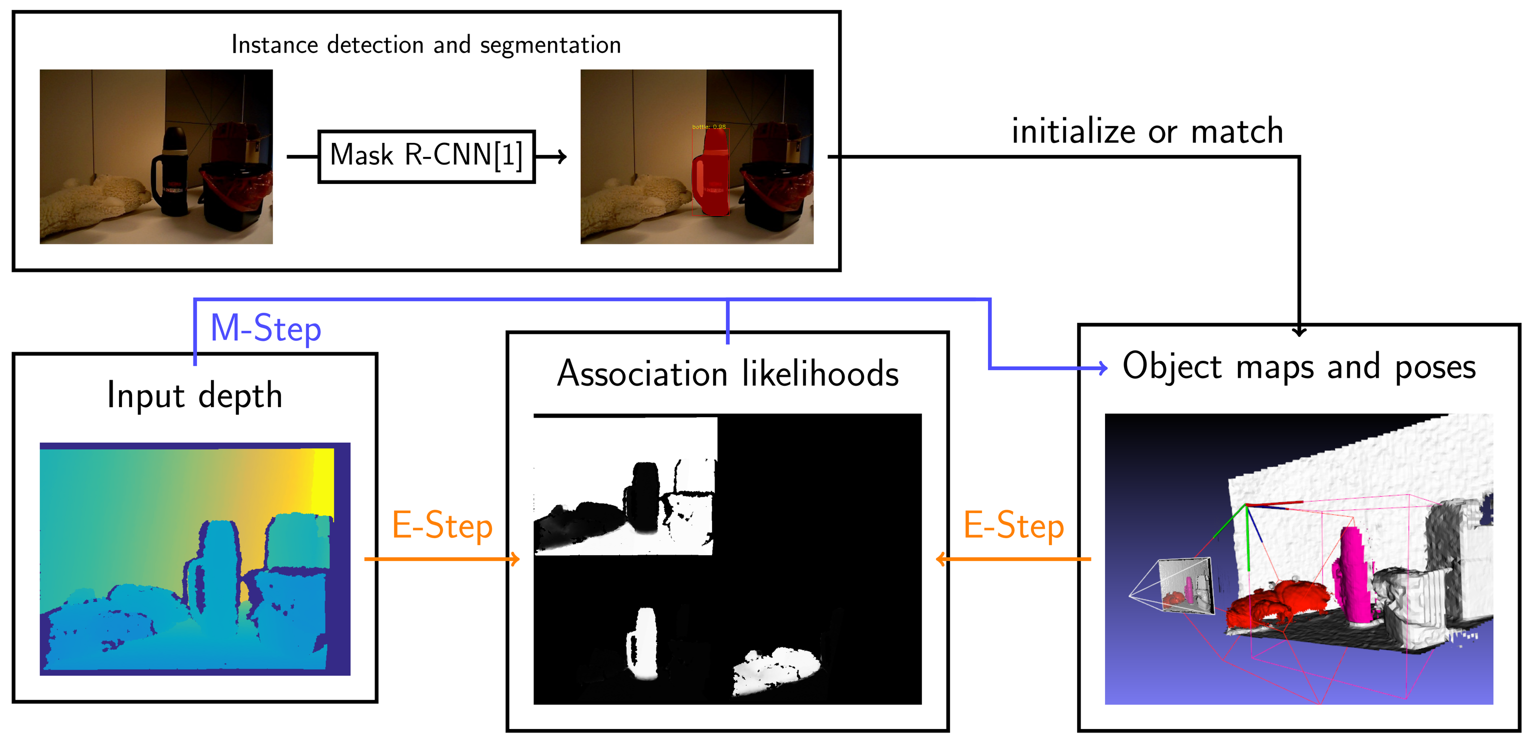
Embodied Vision Group, Max Planck Institute for Intelligent Systems, Tübingen, Germany

Motivation

Aim: Vision-based 4D representations of dynamic environments for control and planning.

Overview

Problem: Segmentation, mapping and pose estimation of moving objects and background from RGB-D images.



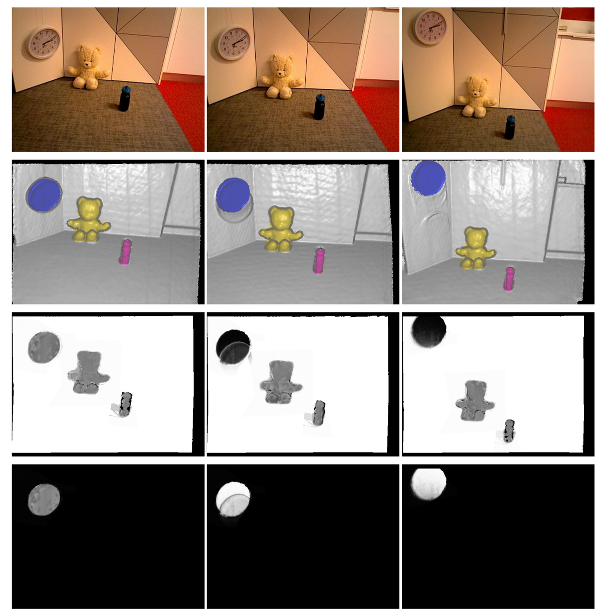
Approach:

- Deep-learning based instance detection and segmentation for initialization of object instances.
- Representation by individual discretized volumetric signed distance functions $\psi(\mathbf{p}) : \mathbb{R}^3 \to \mathbb{R}$ for each object.
- Key contribution: probabilistic formulation for associating pixels to objects based on the expectation-maximization (EM) framework.
  1. E-Step: Infer latent association c_t of pixels in images $\mathbf{z}_{1:t}$ to objects:
  $$\arg\max_{q(c_t)} \sum_{c_t} q(c_t) \ln p(\mathbf{z}_t, c_t \mid \mathbf{m}, \boldsymbol{\xi}_t) = \frac{p(\mathbf{z}_t \mid c_t, \mathbf{m}, \boldsymbol{\xi}_t)}{\sum_{c'_t} p(\mathbf{z}_t \mid c'_t, \mathbf{m}, \boldsymbol{\xi}_t)} \tag{1}$$
  2. M-Step: Find maximum a posteriori estimate for maps $\mathbf{m}$ and poses $\boldsymbol{\xi}_t \in SE(3)$ given images $\mathbf{z}_{1:t}$:
  $$\arg\max_{\mathbf{m},\boldsymbol{\xi}_t} p(\mathbf{m}, \boldsymbol{\xi}_t \mid \mathbf{z}_{1:t}) = \arg\max_{\mathbf{m},\boldsymbol{\xi}_t} p(\mathbf{z}_t \mid \mathbf{m}, \boldsymbol{\xi}_t)\, p(\mathbf{m} \mid \mathbf{z}_{1:t-1})\, p(\boldsymbol{\xi}_t) \tag{2}$$

Instance Detection and Segmentation

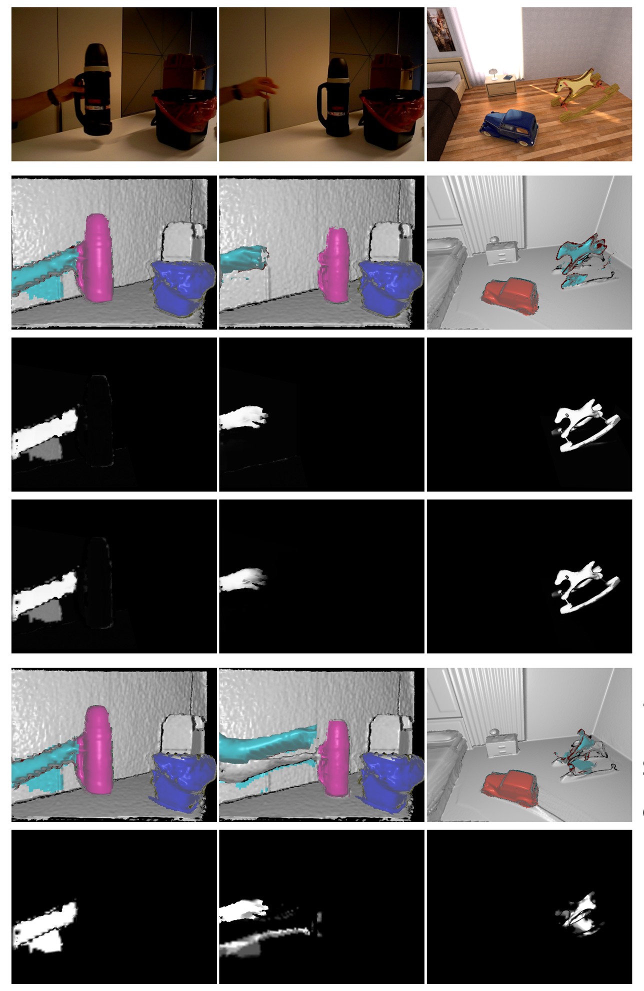
- Similar to [3], detect objects based on Mask R-CNN [1] (deep instance segmentation).
- Maintain recursive estimate of foreground probability $p_{fg}(\mathbf{p} \mid i) = Fg_i(\mathbf{p})/(Fg_i(\mathbf{p}) + Bg_i(\mathbf{p}))$ of points $\mathbf{p}$ through counts in corresponding voxels for each object i.
- Associate Mask R-CNN detections with existing objects based on segment IoU.
- Maintain an existence probability $p_{ex}(i) = Ex(i)/(Ex(i) + NonEx(i))$ through counts. Delete objects when $p_{ex}(i) < 0.1$.

Data Association (E-Step)



- Model data likelihood of pixel $\mathbf{u}$ in object c_t with mixture distribution,
$$p(\mathbf{u} \mid c_t, \boldsymbol{\theta}) = \alpha \frac{1}{2\sigma} \exp\left(-\frac{|\psi_{c_t}(\mathbf{p}_{c_t})|}{\sigma}\right) p_{fg}(\mathbf{p}_{c_t} \mid c_t) + (1-\alpha)\, p_{\mathcal{U}}(\mathbf{p}_{c_t}), \tag{3}$$
where ψ_{c_t} is the SDF of object c_t and $\mathbf{p}_i := \mathbf{T}(\boldsymbol{\xi}_i)\, \pi^{-1}(\mathbf{u}, D(\mathbf{u}))$.
- Association likelihood as data likelihood normalized over all models:
$$p(c_t \mid \mathbf{u}, \boldsymbol{\theta}) = \frac{p(\mathbf{u} \mid c_t, \boldsymbol{\theta})}{\sum_{c'_t} p(\mathbf{u} \mid c'_t, \boldsymbol{\theta})} \tag{4}$$

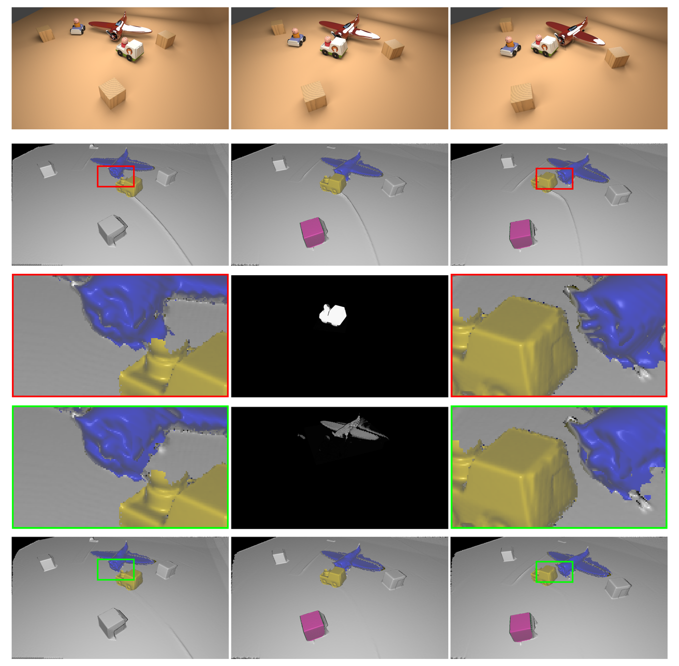
Tracking (M-Step)



- Minimize distance of measured points to implicit surface represented by object SDF,
$$E(\boldsymbol{\xi}) = \frac{1}{2} \sum_{\mathbf{u} \in \boldsymbol{\Omega}} q(c_u) \left|\psi\left(\mathbf{T}(\boldsymbol{\xi})\, \mathbf{p}(\mathbf{u})\right)\right|_\delta, \tag{5}$$
where $\mathbf{p}(\mathbf{u}) := \pi^{-1}(\mathbf{u}, D(\mathbf{u}))$.
- Use Huber norm with threshold δ to achieve robustness with regard to outliers.
- Weigh residuals by map confidence
$$W(\mathbf{T}(\boldsymbol{\xi})\, \mathbf{p}(\mathbf{u})) / \max_{\mathbf{u}' \in \boldsymbol{\Omega}} W\left(\mathbf{T}(\boldsymbol{\xi})\, \mathbf{p}(\mathbf{u}')\right), \tag{6}$$
where W is accumulated integration weight in map.

Top to bottom: RGB images, our 3D reconstruction with reprojected object segmentation, association likelihoods and tracking weights for the hand/horse object, 3D reconstruction with foreground probability instead of the association likelihood, tracking weights with foreground probability instead of association likelihood.

Mapping (M-Step)



- Recursively integrate depth maps into background and object maps,
$$\psi(v) \leftarrow \frac{W(v)\psi(v) + q(c_u)\, d(v)}{W(v) + q(c_u)}, \qquad W(v) \leftarrow \min(W_{max}, W(v) + q(c_u)), \tag{7}$$
- Incorporate the association likelihood $q(c_u)$ of the pixel $\mathbf{u}$ which passes through voxel v.
- $d(v)$ is measured depth difference of the voxel towards integrated depth image.
- Cap on $W(v)$ prevents the model from becoming overconfident in SDF estimate and allows for faster adaptation in case of inaccurate or missing segmentations of dynamic objects

Results

		Kintinuous [7]	ElasticFusion [8]	Co-Fusion [4]	MaskFusion [5]	**Ours**
ToyCar3	Static Bg	**0.10**	0.59	0.61	20.60	0.95
	Car1	-	-	7.78	1.53	**0.77**
	Car2	-	-	1.44	0.58	**0.18**
Room4	Static Bg	**0.16**	1.22	0.93	1.41	1.37
	Airship	-	-	0.91/ 1.01	13.62/ 2.29/ 3.46	**0.56**/ 1.41/ 0.75
	Car	-	-	**0.29**	2.66	2.10
	Horse	-	-	5.80	-	**3.57**

Object and background tracking: AT-RMSEs (in cm).
Our method achieves competitive results with a static SLAM system (EF) for the static background and outperforms other dynamic SLAM approaches (CF, MF) on most of the objects.

	VO-SF[2]	SF[6]	CF[4]	MF[5]	MID-F[9]	**Ours**
f3s static	2.9	1.3	1.1	2.1	1.0	**0.9**
f3s xyz	11.1	4.0	**2.7**	3.1	6.2	3.7
f3s halfsphere	18.0	4.0	3.6	5.2	**3.1**	3.2
f3w static	32.7	**1.4**	55.1	3.5	2.3	**1.4**
f3w xyz	87.4	12.7	69.6	10.4	6.8	**6.6**
f3w halfsphere	73.9	39.1	80.3	10.6	**3.8**	5.1

(a) Absolute trajectory (AT) RMSE (in cm)

	VO-SF[2]	CF[4]	SF[6]	MF[5]	**Ours**
f3s static	2.4	1.1	1.1	1.7	**0.9**
f3s xyz	5.7	2.7	2.8	4.6	**2.6**
f3s halfsphere	7.5	**3.0**	**3.0**	4.1	**3.0**
f3w static	10.1	22.4	1.3	3.9	**1.2**
f3w xyz	27.7	32.9	12.1	9.7	**6.0**
f3w halfsphere	33.5	40.0	20.7	9.3	**5.1**

(b) Relative pose (RP) RMSE (cm/s)

Robust background tracking by representing people explicitly as dynamic objects

Conclusions

- EM formulation for dynamic object-level SLAM with RGB-D cameras
- Probabilistic treatment of data associations key ingredient to robust tracking and mapping in dynamic scenes

Outlook:

- Use RGB image for tracking
- Use for interactive perception of objects
- Global graph optimization and more efficient data structures

References

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, Oct 2017.

[2] M. Jaimez, C. Kerl, J. Gonzalez-Jimenez, and D. Cremers. Fast odometry and scene flow from RGB-D cameras based on geometric clustering. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3992–3999, May 2017.

[3] J. McCormac, R. Clark, M. Bloesch, A. J. Davison, and S. Leutenegger. Fusion++: Volumetric object-level SLAM. In *International Conference on 3D Vision (3DV)*, pages 32–41, Sep. 2018.

[4] M. Rünz and L. Agapito. Co-Fusion: Real-time segmentation, tracking and fusion of multiple objects. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4471–4478, May 2017.

[5] M. Rünz, M. Buffier, and L. Agapito. MaskFusion: Real-time recognition, tracking and reconstruction of multiple moving objects. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 10–20, Oct 2018.

[6] R. Scona, M. Jaimez, Y. R. Petillot, M. Fallon, and D. Cremers. StaticFusion: Background reconstruction for dense RGB-D SLAM in dynamic environments. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–9, May 2018.

[7] T. Whelan, M. Kaess, M. F. Fallon, H. Johannsson, J. J. Leonard, and J. B. McDonald. Kintinuous: Spatially extended KinectFusion. In *RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras*, Sydney, Australia, Jul 2012.

[8] T. Whelan, S. Leutenegger, R. S. Moreno, B. Glocker, and A. Davison. ElasticFusion: Dense SLAM without a pose graph. In *Proceedings of Robotics: Science and Systems (RSS)*, Rome, Italy, July 2015.

[9] B. Xu, W. Li, D. Tzoumanikas, M. Bloesch, A. Davison, and S. Leutenegger. MID-Fusion: Octree-based object-level multi-instance dynamic SLAM. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

Acknowledgements:
We acknowledge support from the BMBF through the Tuebingen AI Center (FKZ: 01IS18039B) and Cyber Valley. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Michael Strecke.

Project website:

emfusion.is.tue.mpg.de