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

Michael Strecke and Jörg Stückler
Embodied Vision Group, Max Planck Institute for Intelligent Systems
{michael.strecke,joerg.stueckler}@tue.mpg.de

	det. frames	non-det. frames	Overall
ToyCar3	789	241	259
Room4	561	124	139
PlaceItems	611	89	107
TeddyHandover	772	142	164
SlidingClock	759	188	208

Table 1. Average runtimes per frame in ms on the dynamic sequences of the Co-Fusion dataset [1]. The runtimes on detection frames include not only Mask R-CNN inference but also creation and updates of object maps.

In this document, we provide additional performance numbers for our method as well as an explanation of the accompanying video.

1. Runtimes

In Table 1, we report average runtimes per frame on the dynamic sequences from Co-Fusion [1]. One can clearly see that the runtime for detection frames is much higher than on the frames where only tracking and mapping with our probabilitic data association is run. The overhead in detection frames amounts to more than just the inference runtime of Mask R-CNN (which has been reported around 200ms/frame or 5Hz [2]), since we also allocate new object volumes in these frames or match and update existing ones. Note that this implementation is not yet tuned for computational efficiency. Note further, that the *ToyCar3* dataset has a resolution of 960×540 pixels while all other datasets are capture at a resolution of 640×480 pixels. The higher runtime in *ToyCar3* is thus partly due to the fact that more data needs to be processed per frame.

2. Varying detection rates

Table 2 shows an ablation study of how varying the detection rate for Mask R-CNN affects trajectory accuracy, trajectory coverage, and the number of detected non-moving objects. The coverage is computed as the percentage of frames in the sequence for which our approach maintained a model for the object. Note that by this measure, 100% cannot be achieved for most objects since they are not visible in all frames. Note further, that in *ToyCar3*, the static airplane is always detected and instantiated as an object which is why there is always at least one non-moving object detected in this scene.

The clear and intuitive tendency is that trajectory coverage improves with increased detection rate but this also creates more spurious detections instantiating objects. If the detection frequency is too low, some objects might be missed. This happenes for the second car in *ToyCar3* and the horse in *Room4* when we run the detection only every 60 frames (s. Table 2 (a), (b)).

Interestingly, while a larger trajectory coverage can induce higher AT-RMSE (since more frames can deviate from the ground truth), we do not observe this as a clear tendency. For most objects, the AT-RMSE remains at a very similar level. In some cases, such as for Car2 in *ToyCar3*, the AT-RMSE even improves with increased trajectory coverage.

3. Accompanying Video

The accompanying video shows several results on dynamic scenes. It displays RGB images for reference, the object detections that are used to instantiate new objects, the rendered output maps with gray background model and color-coded object instances, as well as association weights for a selection of objects. Note that the color frames are not directly used as input for tracking and mapping, but only for generation of instance masks. Tracking and mapping only uses depth images from the RGB-D data.

The first five scenes show result on sequences from the Co-Fusion dataset [1]. Additionally we show results on *f3w_xyz* from the TUM RGB-D benchmark [3] as a proof of concept for robust camera tracking (visualization of "person" models disabled). Note that persons violate the rigidness assumption of our approach. Thus, the models that are maintained for them are not very accurate. This leads to some artifacts and floating surfaces being integrated into the background volume. These are removed once valid background depth data is available.

		1	15	30	60
ToyCar3	Static Bg	0.94	0.94	0.95	0.96
	Car1	0.80	0.83	0.77	0.85
	Car2	0.13	0.17	0.18	-
Room4	Static Bg	1.35	1.37	1.37	1.40
	Airship	0.34/	0.56/	0.56/	0.56/
		3.87/	1.35/	1.41/	1.43/
		0.75	0.75	0.75	0.75
	Car	2.34	2.13	2.10	2.10
	Horse	2.19	3.57	3.57	-

(a)

		1	15	30	60
ToyCar3	Car1	82.6%	81.1%	81.1%	62.3%
	Car2	21.4%	6.5%	6.5%	0%
Room4	Airship	7.4%/	9.2%/	9.2%/	9.2%/
		11.1%/	6.1%/	6.3%/	6.3%/
		10.9%	5.7%	5.7%	5.7%
	Car	12.0%	9.7%	7.9%	7.9%
	Horse	19.5%	18.5%	18.5%	0%

(b)

	1	15	30	60
ToyCar3	6	2	2	1
Room4	10	1	0	0

(c)

Table 2. Ablation study with varying detection rates. (a) Average trajectory (AT) RMSE when running detections every 1, 15, 30, or 60 frames. (b) Trajectory coverage for these setups. (c) Number of non-moving objects detected in these cases.

References

[1] Martin Rünz and Lourdes 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.

[2] Martin Rünz, Maud Buffier, and Lourdes 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.

[3] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 573–580, Oct 2012.