LEARNING TO SEGMENT DYNAMIC OBJECTS USING SLAM OUTLIERS

Adrian Bojko, Romain Dupont, Mohamed Tamaazousti and Hervé Le Borgne
Paris-Saclay University, CEA, List, France

adrian.bojko@cea.fr - romain.dupont@cea.fr
mohamed.tamaazousti@cea.fr - herve.le-borgne@cea.fr
1. CONTEXT

1. SLAM: Simultaneous Localization and Mapping in a static environment

2. Dynamic SLAM: SLAM extended to dynamic environments

3. Some applications:

   - Autonomous vehicles
   - Robotics
   - Augmented reality

 Images adapted from Wikimedia.
1. FOCUS: FEATURE-BASED SLAM

We use ORB-SLAM 2 [Mur-Artal et al., 2016].

Feature-based SLAM: generate and track keypoints (in green) across images to compute camera trajectory while reconstructing the environment.
1. PROBLEM: CONSENSUS INVERSION

Consensus Inversion: implicit use of a frame of reference that is not the ground when the motion of dynamic objects is dominant.

Example of false start (a type of consensus inversion):
the camera is static but ORB-SLAM 2 (monocular) computes a fake trajectory.
1. PROBLEM: PRIORS ON DYNAMIC OBJECTS CAN BE WRONG

Priors on dynamic objects (e.g. people) can be completely wrong. The train station is moving, not the people nor the train.

*Scene from "Top Secret!" (1984, Paramount Pictures)*
2. STATE OF THE ART: DYNAMIC SLAMS

General principle of Dynamic SLAM: filter interest points on dynamic objects

→ Critical step: dynamic object detection
### 2. STATE OF THE ART: DYNAMIC SLAMS

#### Approaches

1. **SLAM + geometry:**
   - Uses: optical flow, depth maps...
   - *An Accurate Localization Scheme* [Chen et al., 2018]

2. **SLAM + semantic masks:**
   - Uses: Mask R-CNN, ...
   - *Mask-SLAM* [Kaneko et al., 2018]

3. **Hybrid:**
   - a) **SLAM + geometry (runtime) + semantic masks (runtime)**
     - *DynaSLAM* [Bescos et al., 2018], *SLAMANTIC* [Schorghuber et al., 2019]
   - b) **SLAM + geometry (training) + semantic masks (runtime)**
     - *Driven to distraction* [Barnes et al., 2018]

#### Limits

- **Vulnerable to consensus inversions**
- **Limited by the scope of the training databases**
- **Vulnerable to consensus inversions**
- **Requires a lot of training data**
  - (several traversals of the same location)
3. CONTRIBUTIONS

Our main contribution is a Dynamic SLAM:
• Based on self-supervised learning of masks
  (we use outliers i.e. keypoints rejected during optimization)
• Supports consensus inversions
• That only requires one learning sequence per dynamic object

Additional contributions:
1) Database Consensus Inversion
2) SLAM Robustness metrics (Penalized ATE RMSE and Success Rate)
3.1 METHOD

Hypothesis: dense outliers that appear suddenly characterize dynamic objects in sequences with no consensus inversion.

Dynamic SLAM = SLAM + semantic filter of keypoints
3.1 METHOD: MASK CREATION

- SLAM inlier / outlier collection
- Mask database creation:
  a) Search for dense outliers using sliding windows + creation of bounding boxes
     • We look for drops in the inlier/outlier ratio inside the sliding window.
     • We then merge overlapping boxes that have inlier/outlier ratio drops. The result is bounding boxes enclosing dynamic boxes.

  Image $n$: before the car moves. (inliers in green, outliers in blue)
  Image $n + 3$: after the car moves.

  Image $n$: after merging windows with inlier/outlier ratio drops.

  b) Creation and propagation of masks across sequences using video segmentation tools:
     COSNet [Lu et al, 2019] and SiamMask [Ventura et al., 2019]
3.2 METHOD: NETWORK TRAINING

a) Train single-object models using the created mask database, DeepLabv3+ [Chen et al., 2018] architecture
b) Infer masks with each model and superimpose the result per sequence
c) Train a global model with the superimposed masks

Single-object models: mask objects separately

Global model: mask all objects simultaneously
3.3 ADDITIONAL CONTRIBUTIONS

- **Our dataset "Consensus Inversion"** contains sequences with consensus inversion, rarely present in SLAM datasets.

- **Metrics to measure SLAM robustness:**
  
  - **SLAM failure:** Tracking Rate too low (compared to perfectly masking moving objects and consensus inversions) or ATE RMSE above a fixed threshold (e.g. 10cm).
  
  - **Penalized ATE RMSE** = \( \max(L) \cdot (1 + \tau) \), if SLAM failure
    
    \[ \text{ATE RMSE otherwise} \]

  - Defined within a SLAM benchmark.
  - \( L \) is the set of ATE RMSEs of all benchmarked SLAMs that were successful and \( \tau \) the penalty factor.

  - **Success Rate:** % of sequences that are successfully processed by the SLAM.
3.4 QUALITATIVE RESULTS

- Example on TUM RGB-D [Sturm et al., 2012] (a popular SLAM database) in RGB-D
- Moving people cause a consensus inversion

Sequence fr3_walking_xyz (RGB-D), tracked keypoints in green.
3.4 QUANTITATIVE RESULTS

- Evaluation on TUM RGB-D (dynamic sequences) and Consensus Inversion
- Also tested network integration in LDSO [Gao et al, 2018], a direct SLAM

- Results (partial):

<table>
<thead>
<tr>
<th>Test set</th>
<th>State-of-the-Art</th>
<th>ORB-SLAM 2 + ...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DynaSLAM</td>
<td>SLAMANTIC</td>
</tr>
<tr>
<td>Consensus Inversion / Dyn. - Mono</td>
<td>0.0693</td>
<td>0.0692</td>
</tr>
<tr>
<td>TUM RGB-D / Dyn. - Mono</td>
<td>0.1108</td>
<td>0.1101</td>
</tr>
<tr>
<td>Consensus Inversion / Dyn. - Stereo</td>
<td>0.0627</td>
<td>0.0699</td>
</tr>
<tr>
<td>TUM RGB-D / Dyn. - RGB-D</td>
<td>0.0206</td>
<td>0.0173</td>
</tr>
</tbody>
</table>

Average Penalized ATE RMSE (m)

<table>
<thead>
<tr>
<th>Test set</th>
<th>State-of-the-Art</th>
<th>ORB-SLAM 2 + ...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DynaSLAM</td>
<td>SLAMANTIC</td>
</tr>
<tr>
<td>Consensus Inversion / Dyn. - Mono</td>
<td>63.6%</td>
<td>63.6%</td>
</tr>
<tr>
<td>TUM RGB-D / Dyn. - Mono</td>
<td>62.5%</td>
<td>62.5%</td>
</tr>
<tr>
<td>Consensus Inversion / Dyn. - Stereo</td>
<td>72.7%</td>
<td>63.6%</td>
</tr>
<tr>
<td>TUM RGB-D / Dyn. - RGB-D</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Success Rate (%)

Results are better or equal than the state of the art in Mono / Stereo / RGB-D.

LDSO + ...

LDSO on Consensus Inversion / Dyn.

Avg. Penalized ATE RMSE (m) 0.0581
Success Rate (%) 63.6%
CONCLUSION

Contributions:
1. A novel method that learns to segment dynamic objects
   • No manual labelling.
   • Uses only one monocular sequence per dynamic object.
   • Supports consensus inversions.

2. The first dataset for Consensus Inversion evaluation.
3. The first robustness metrics that integrate SLAM failures.

Results:
• We improved ORB-SLAM 2 monocular/stereo/RGB-D as well as LDSO and achieved top results in very challenging scenarios.
BIBLIOGRAPHY

"ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras" (Mur-Artal et al., IEEE Transactions on Robotics 2017)

"An Accurate Localization Scheme for Mobile Robots Using Optical Flow in Dynamic Environments" (Cheng et al., ROBIO 2018)

"Mask-SLAM: Robust Feature-Based Monocular SLAM by Masking Using Semantic Segmentation" (Kaneko et al., CVPRW 2018)

"DynaSLAM: Tracking, Mapping, and Inpainting in Dynamic Scenes" (Bescos et al., IEEE Robotics and Automation Letters, 2018)

"Slamantic - leveraging semantics to improve vslam in dynamic environments" (Schorghuber et al. ICCV Workshops, 2019).

"Driven to Distraction: Self-Supervised Distractor Learning for Robust Monocular Visual Odometry in Urban Environments" (Barnes et al., ICRA 2018)

"See More, Know More: Unsupervised Video Object Segmentation With Co-Attention Siamese Networks" (Lu et al., CVPR 2019)

"Fast Online Object Tracking and Segmentation: A Unifying Approach" (Wang et al., CVPR 2019)

"Rvos: End-to-end recurrent network for video object segmentation" (Ventura et al., CVPR 2019)

"Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation" (Chen et al., ECCV 2018)

"LDSO: Direct Sparse Odometry with Loop Closure" (Gao et al., IROS 2018)