MATTHIAS SCHÖRGHUBER[13] // MARTIN HUMENBERGER[2] // MARGRIT GELAUTZ[3]

matthias.schoerghuber@ait.ac.at // martin.humenberger@naverlabs.com // margrit.gelautz@tuwien.ac.at

1 AIT Austrian Institute of Technology GmbH, Center for Vision, Automation & Control, Autonomous Systems, Giefinggasse 4, 1210 Vienna, Austria.

2 Naver Labs Europe, Grenoble, France

3 Vienna University of Technology, Institute of Visual Computing and Human-Centered Technology, Vienna, Austria

CAMERA-BASED POSE ESTIMATION IN DYNAMIC ENVIRONMENTS – CONCEPT AND STATUS

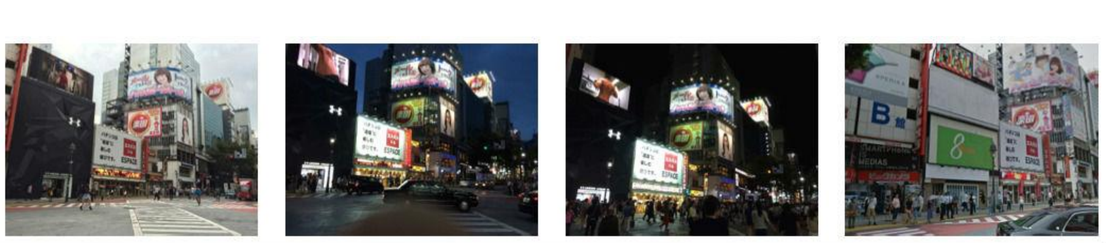
① Abstract

This PhD tackles the challenge of camera-based methods for navigation and environmental sensing in dynamic environments. The goal is to design a robust real-time localization and mapping algorithm which can reliably cope with dynamic (e.g. people, cars) and changing (e.g. structural changes, weather) environments.

We plan to introduce semantics into the traditional geometric processing cues, which allows for explicit treatment of dynamic and changing environments in order to improve mapping and, consequently, pose estimation. As a second goal we leverage the semantic information to introduce enhanced image retrieval techniques to improve the large-scale localization and map-maintenance in multi-session scenarios.

Real-world data contains dynamic elements

Exemplary dynamic scene with persons and cyclists from the Cityscapes dataset [3].



Images from the same place at different points in time. Source: [6].

② Approach

In this PhD we work on **improving visual pose estimation** in **dynamic environments**. Ⓐ-presents a VSLAM approach to combine the usage of an offline map with online mapping.

In Ⓑ we **address the challenge** of dynamic environments by **explicit detection of dynamic areas** during the mapping stage by using **semantic image segmentation**.

In the future work our approach seeks to **integrate semantic** scene **information into the map** representation of a VSLAM, to build maps for long-term usage that can be updated dynamically, and to further **improve robustness and accuracy**. Furthermore, we plan to **accelerate** the **initial localization** by enhancing image retrieval techniques with semantic information.

[1] M. Schörghuber, M. Wallner, M. Humenberger, and M. Gelautz, “Vision-based autonomous feeding robot,” OAGM Workshop, pp. 111-115, 2018

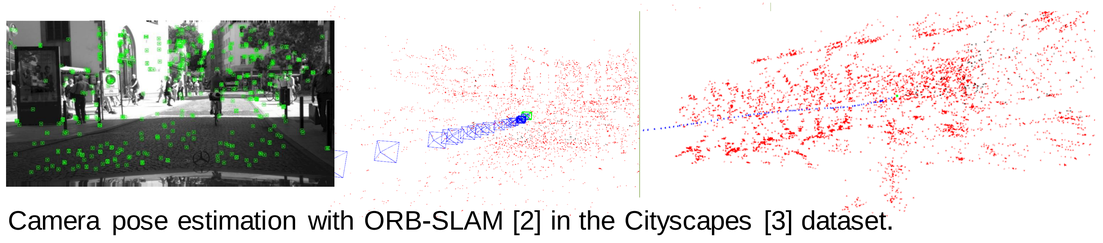
[2] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, “ORB-SLAM: A versatile and accurate monocular SLAM system,” IEEE Transactions on Robotics, vol. 31, no. 5, pp. 1147–1163, 2015

[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223, 2016

[4] F. Yu, V. Koltun, and T. Funkhouser, “Dilated residual networks,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 636–644, July 2017

[5] T. Scharwächter, M. Enzweiler, U. Franke, and S. Roth, “Stixmantics: A medium-level model for real-time semantic scene understanding,” European Conference on Computer Vision, pp. 533–548, 2014

[6] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla, “24/7 place recognition by view synthesis,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 1808-1817, 2015

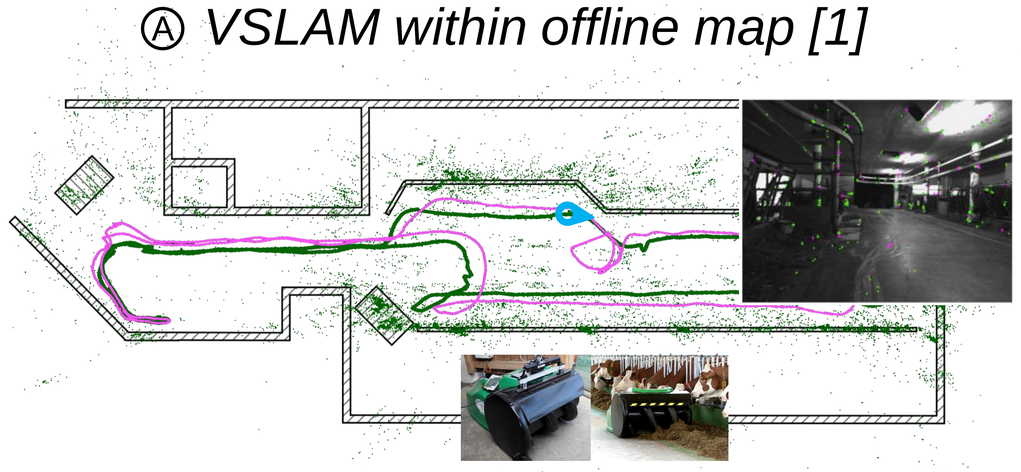


Camera pose estimation with ORB-SLAM [2] in the Cityscapes [3] dataset.

② Problem Description

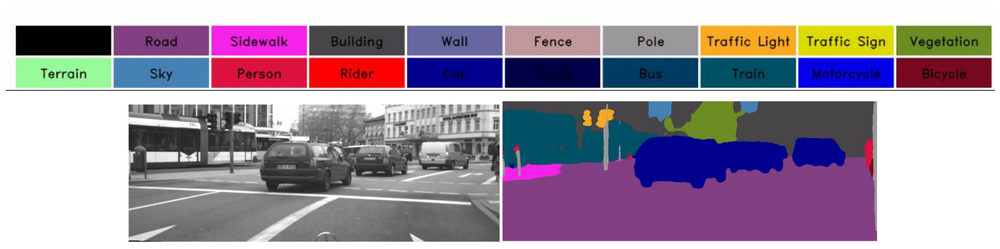
The **accuracy** of the estimated pose in a **VSLAM** algorithm relies on **valid geometric representations** of the observed environment. During operation, a VSLAM algorithm extends the map (3D scene model) by adding new 3D measurements which are generated by estimating the depth of image points or areas captured from different viewpoints. If an **object has moved** in between these points in time, the **triangulation** of image points representing the object does not relate to the correct distance to the camera and therefore **does not lead to the correct camera pose**. Hence, with these methods, **reliable** results can only be achieved in **static** and distinctive (in terms of texture and structure) **environments**. This shortcoming is the main challenge for VSLAM algorithms **in real-world scenarios**.

Ⓐ *VSLAM within offline map [1]*

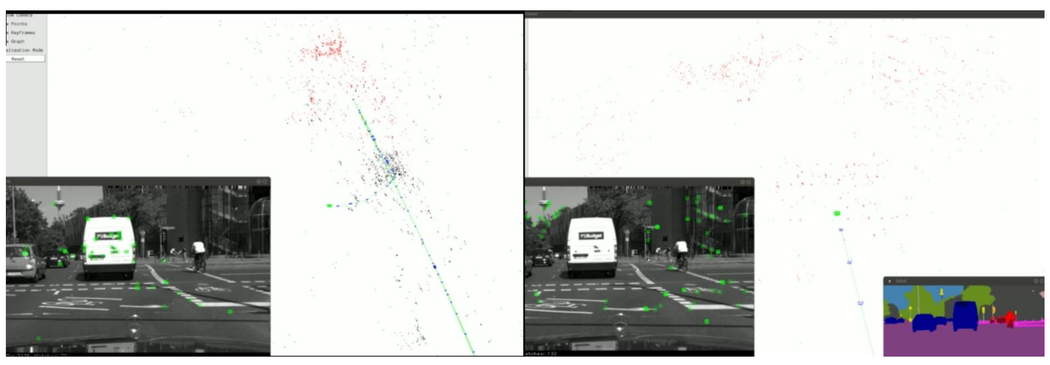


VSLAM approach for navigation of an autonomous feeding robot in a dynamic environment. We combine an offline map (green elements) with online mapping (purple elements) to improve the accuracy and robustness of the pose estimation. The increased accuracy is achieved by incorporating reliable 3D points from the offline map in order to robustly triangulate new accurate 3D points. Consequently, the new 3D points are inherently aligned to the mission-specific world coordinate frame. Furthermore, in the case of lost pose tracking, the robot can relocalize itself in the offline map and continue its mission. See [1] for details.

Ⓑ *Leverage semantic information*



Semantic labeling example using [4] (top: label bar). Scene from [5].



The image shows our preliminary result on a problematic traffic light scene of the Cityscapes [3] dataset. (Approach to a waiting truck.)

Left: Original ORB-SLAM [2] fails due to relying on 3D points originating from the truck. When the truck moves, the static assumption is violated.

Right: In contrast to detecting dynamic objects by using outliers from pose estimation, we use semantic image segmentation to suppress 3D point generation on dynamic labeled objects which allows successful tracking of this scene.