2nd year Final Report of Project NKFIH KH-17 #126513 "Visual appearance-based robot localization (Robot tájékozódás képi információk alapján)"

The project team has accomplished new scientific results and has advanced the state of the art in the field of computer vision and robotics in seven main topics by implementing the initially proposed project work plan. In this report, we summarize our main contributions in these topics.

Topic 1: Semantic segmentation and depth estimation with convolutional neural network

The goal of this topic was to research new deep learning methods to semantically segment the images captured on-board of aerial and ground vehicles in order to improve the visual localization and scene understanding in case of urban environments. To achieve these goals, we have proposed a new convolutional neural network (CNN) architecture [1] in order to advance the state-of-the-art in the field of single-image depth prediction. This work was also presented in Hungarian [1a]. The proposed architecture combines the advantages of deep residual nets [12] and the U-net design [13], since the proposed architecture contains additional side-to-side connections between the encoding and decoding branches of the network. We performed extensive experiments using the NYU Depth v2 benchmark dataset [14]. The quantitative evaluation and comparison with other state-of-the-art CNN based methods show the advantages of our hybrid method while keeping the complexity of the network and the number of variables roughly the same. In order to obtain reliable performance also in outdoor scenes the network was trained further using the Cityscapes Dataset (https://www.cityscapes-dataset.com/). We have achieved the semantic segmentation of the images using the same network architecture. Only the last layer was replaced to have a 7 channel output layer representing the 7 semantic classes used to categorize every pixel (flat surfaces, buildings, street furniture, vegetation, sky, humans, and vehicles). However, in this case the network was trained on the BDD100K dataset (http://bdd-data.berkeley.edu/). Furthermore, we performed the manual semantic labelling on a subset of the images from the Zurich Urban MAV Dataset [2], in order to evaluate the performance of the proposed network using the world's first dataset recorded on-board a camera equipped MAV flying within urban streets at low altitudes. Therefore, the dataset become now ideal to evaluate and benchmark appearance-based topological localization, monocular visual odometry, simultaneous localization and mapping (SLAM), and online 3D reconstruction algorithms also in the context of semantic recognition and understanding of the environment.

We made freely available the outcomes of this research (trained models, datasets and source code) to the general public at:
- Models and code of the paper [1]: https://github.com/karoly-hars/DE_resnet_unet_hyb
- Outdoor experiments: https://github.com/karoly-hars/3D_scene_rec_CNNs
- Zurich Urban Micro Aerial Vehicle Dataset [2]: http://rpg.ifi.uzh.ch/zurichmavdataset.html

Topic 2: Photogrammetric appearance-based localization in virtual 3D urban maps

Our second goal was to research accurate and robust visual appearance-based localization algorithm for ground and aerial robots operating in urban environments that can complement, or even replace, satellite-based GPS localization systems. In this case, our assumption was that there exists a dense 3D map (point cloud or mesh) of the environment captured either by the means of LIDAR sensors or computed form a large set of images with structure from motion technology. In our earlier publication underlying this KH proposal [3], we have introduced several methods to compute the topological location of MAV within large maps. Therefore, in order to achieve the goals of this task, we developed an algorithm that is able to further refine the position of the robot by matching the actual image captured on-board the vehicle to the most similar synthetic views obtained by moving a camera within the virtual world---namely the map---around the previously computed approximate (topological) location, c.f. figure below. In case of ground vehicles this search has three degree of freedom (along longitudinal and lateral axes, and around yaw angel). Conversely, in case of quadrotors MAVs the search has four degrees of freedom (along yaw axis additionally), since near hovering conditions are assumed [4]. In order to compute the best match between the actual and synthetic views the normalized mutual information (NMI) is used [15], since this metric is robust to multimodal image registration, changes in illumination, and changes in the scene (by dividing the image into regions). In order to reduce the computational complexity of the overall appearance-based localization algorithm the position of the vehicle is tracked using an altered ORB-SLAM algorithm [16], and the accurate image registration is performed only in keyframe locations.

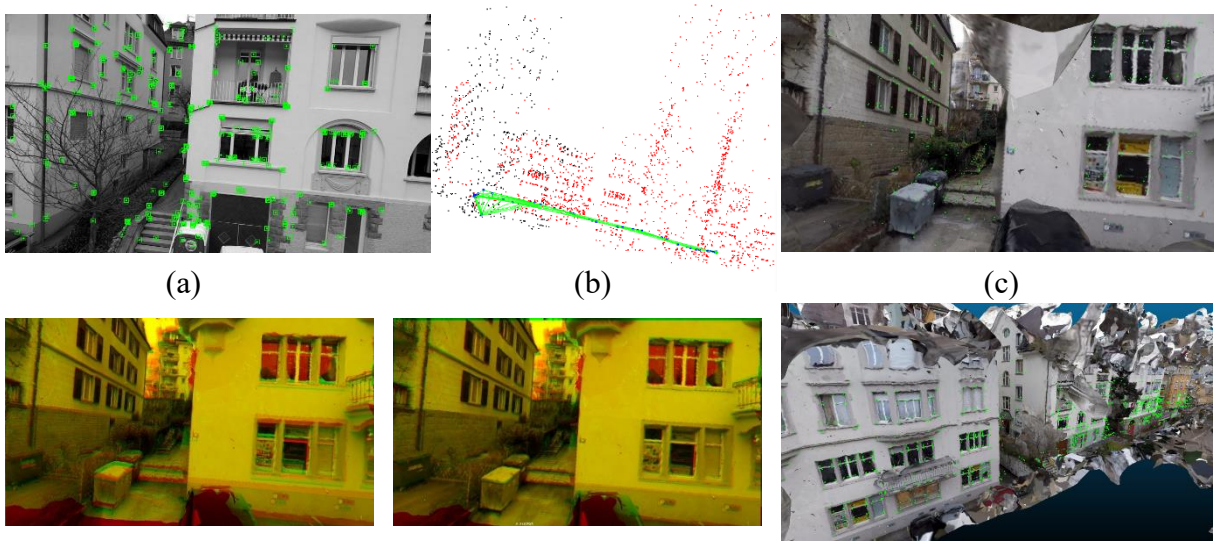


Fig. 1: the main steps of the proposed algorithm: (a) ORB keypoint feature (green dots) detection; (b) output of the ORB-SLAM algorithm: the keyframes are shown in blue, the pose-graph that connects them is green, the red and black dots represent the sparse feature points reconstructed in 3D; (c) synthetic view captured in the virtual 3D model; (d) overplayed image of the actual and the synthetic view (before NMI registration); (e) overplayed image after pose NMI based pose refinement; (f) registered and scaled sparse 3D feature point cloud (green) overplayed on the virtual 3D model.

Topic 3: MTA SZTAKI MIMO (MIcro aerial vehicle and MOtion capture) arena

We have developed and built the SZTAKI MIMO arena using a motion capture system with ten calibrated Optitrack (https://optitrack.com/) infrared cameras and palm sized Crazyflie (https://www.bitcraze.io/) miniature quadrotor MAVs. A motion capture system enables the tracking of retroreflective markers with high precision and data rate. Therefore, the MIMO arena is an ideal tool to record ground truth data to evaluate the performance of different on-board computer vision algorithms, e.g., visual odometry, 3D mapping, simultaneous localization and mapping, and tracking of objects. Additionally, we implemented a hardware-software system for the execution of high-speed autonomous MAV maneuvers within the arena, where the state of the MAV is precisely tracked by the motion capture system [5]. Thus, the arena is suitable for high-speed MAV races and is a research testbed for vision based state estimation and control methods.

Topic 4: Robust camera-based localization and SLAM beyond keypoint matching

Traditionally in vision based localization and SLAM algorithms keypoint features are applied, i.e., ORB features [16]. In order to research more accurate, robust and computationally efficient algorithms Local Affine Frame (LAF) could be used. LAFs aside the center of the feature point contains a 2D affine transformation that describes the local shape and orientation of the surrounding image region. In [6] we proposed a closed-form solution, optimal in the least-squares sense, for correcting the parameters of multi-view affine correspondences represented as a set of LAFs. The algorithm was validated both in synthetic experiments and on publicly available real-world datasets. The results clearly show that the method almost always improves the input LAFs. Consequently, the proposed correction improves homography, surface normal and relative motion estimation via improving the input of these methods. Likewise, image keypoint descriptors encode the scale and orientation that should be exploited. In [7] we propose a theoretically justifiable interpretation of the angles and scales which the orientation- and scale-covariant feature detectors, e.g. SIFT or ORB, provide. Consequently, two new general constraints are proposed for covariant features. These constraints are then exploited to derive two new formulas for homography estimation. Using the derived equations, a solver is proposed for estimating the homography from only two correspondences. The new solver is numerically stable and easy to implement. Moreover, it leads to results superior in terms of geometric accuracy in many synthetic tests and on publicly available real-world datasets consisting of thousands of image pairs.

We made freely and publicly available the source code of the proposed algorithms:
- Source code of paper [6]: https://github.com/eivan/multiview-LAFs-correction
- Source code of paper [7]: https://github.com/danini/homography-from-sift-features

Topic 5: Initial pose-graph estimation for fast and robust SLAM back-ends

Modern Light Detection and Ranging (LIDAR) and image-based visual odometry and SLAM systems use the pose-graph representation to solve the underlying concurrent localization and mapping problem. Therefore, there is a great need for robust and efficient pose-graph optimization back-ends. Generally, the main task of a pose-graph based SLAM back-end is to minimize the accumulated error of the consecutive measurements with the restrictions gained by different loop closures in the movement. The effectiveness of the pose-graph optimization is largely determined by the initial guess from which the optimization is initiated. In [8] we proposed a novel algorithm to compute the initial structure of pose-graph that uses all the previously computed nodes of the pose-graph to estimate a new location. Therefore, we name our method Multi-Ancestor Spatial Approximation Tree (MASAT). In detail, we perform a Breadth-First Search (BFS) on the graph in order to obtain multiple votes regarding the location of a certain robot position from all of its previously processed neighbors. Next, we define the initial location of a pose as the average of the multiple alternatives. By adopting the proposed initialization approach, the number of iterations needed for optimization is significantly reduced while the computational complexity remains lightweight. The quantitative evaluation on various 2D and 3D benchmark datasets demonstrate the advantages of the proposed method.

We made freely available the source code of the proposed algorithm and the dataset used in this work. These are publicly available at:
- C++ source code of paper [8]: https://github.com/karoly-hars/MASAT_IG_for_SLAM
- MASAT noisy datasets: http://mplab.sztaki.hu/masat_slam/masat_slam_data.zip

Topic 6: Advanced model estimation for robust outlier detection

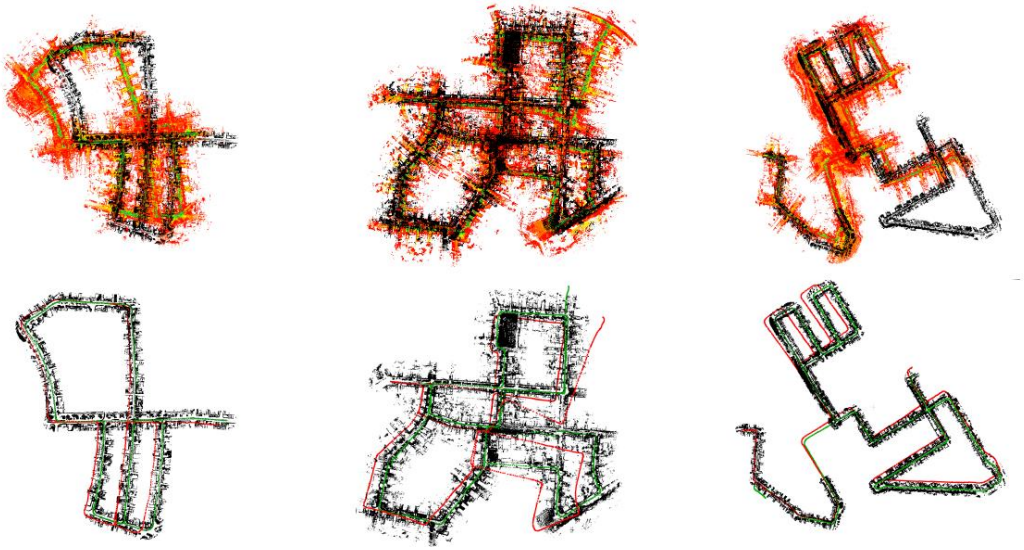
The measurements in computer vision and robotics are always affected by noise, therefore efficient and robust model estimation is essential. The RANdom SAmple Consensus (RANSAC) algorithm [17] proposed by Fischler and Bolles in 1981 has become the most widely used robust estimator in computer vision. Since many variants were proposed, in [9] we proposed a robust approach, called σ–consensus for eliminating the need of a user-defined threshold by marginalizing over a range of noise scales. In addition, due to not having a set of inliers, a new model quality function and termination criterion were introduced. The algorithm is superior to other state-of-the-art methods in terms of geometric accuracy on publicly available real-world datasets for epipolar geometry (F and E) and homography estimation. Furthermore, in [10] we proposed an algorithm for geometric multi-model fitting. The method interleaves sampling and consolidation of the current data interpretation via repetitive hypothesis proposal, fast rejection, and integration of the new hypothesis into the kept instance set by labeling energy minimization. Due to exploring the data progressively, the method has several beneficial properties. First, a

clear criterion, adopted from RANSAC, controls the termination and stops the algorithm when the probability of finding a new model with a reasonable number of inliers falls below a threshold. Second, it is an any-time algorithm. Thus, whenever is interrupted, e.g. due to a time limit, the returned instances cover real and, likely, the most dominant ones. The method is superior to the state-of-the-art in terms of accuracy in both synthetic experiments and on publicly available real-world datasets for homography, two-view motion, and motion segmentation.

We made freely available the source code of the proposed algorithms. These are publicly available at:
- Source code of the MAGSAC algorithm [9]: https://github.com/danini/magsac
- Source code of the Progressive-X algorithm [10]: https://github.com/danini/progressive-x

Topic 7: Geometric localization of vehicles in urban 3D point cloud maps

The goal of this task was to tackle the challenges of globally localizing a LIDAR equipped vehicle driving in urban environments, where a premade target 3D map point cloud exists to localize in. In our work [11] we proposed a novel Lidar-only Odometry and Localization (LOL) algorithm, where in order to correct the accumulated drift of the Lidar-only odometry we apply a place recognition method to detect geometrically similar locations between the online 3D point cloud and the a priori offline map.



(a) (b) (c)

Fig. 2: **Top row:** the result of the Loam mapping algorithm, tested on various length Kitti [20] datasets: (a) Drive 18, 04:36 minutes, approx. 2200 meter (m) (b) Drive 27, 07:35 minutes, approx. 3660 m (c) Drive 28, 08:38 minutes, approx. 4125 m. The ground truth map is visualized with black points, while the self built map points on the green-red scale according to the vertical height. **Bottom row:** the results obtained with the proposed LOL method (green line) with respect to the baseline Loam algorithm (red line) using the different datasets.

In our system, we integrated a state-of-the-art LIDAR odometry algorithm (Loam [18]) with a recently proposed 3D point segment matching method ([19]) by complementing their advantages. We also proposed a set of enhancements: (i) a RANSAC-based geometrical verification to reduce the number of false matches between the online point cloud and the offline 3D map; and (ii) a fine-grained ICP alignment to refine the re-localization accuracy whenever a good match is detected. The utility of the proposed LOL system is demonstrated on several Kitti Vision Benchmark datasets [20] of different lengths and environments (shown on the figure above), where the re-localization accuracy and the precision of the vehicle's trajectory were significantly improved in every case, while still being able to maintain real-time performance.

We made freely available the source code of the proposed algorithm:
- C++ source code of the paper [11]: https://github.com/RozDavid/LOL
- A video demonstration is available at: https://youtu.be/ektGb5SQGRM

Further scientific activities and events related to the project

1) The project was presented to the general public within the framework of the European Researchers' Night 2018 on September 28, title: Sensing drones in action, link of the event: https://www.kutatokejszakaja.hu/esemeny/erzekelo-dronok-akcioban/

Selected media appearances covering this event:
a) MTVA, hirado.hu: https://www.hirado.hu/belfold/kozelet/cikk/2018/09/22/erdeklik-a-dronok-ide-menjen-a-kutatok-ejszakajan
b) Indamedia Network, divany.hu: https://divany.hu/szuloseg/2018/09/27/kutatok-ejszakaja-2018/
c) European Robotics Week 2018: https://www.eu-robotics.net/robotics_week/events/erzekel-dronok-akcioban--1-3---r.-night-mta-sztaki.html

2) Benjamin Berta (supervised BSc students) won 1st prize in 2017 Scientific Student Conference (TDK, http://tdk.bme.hu/Browse/Users/Berta-Benjamin) with the work entitled: Navigation of small-scale aircraft in urban environments using deep-learning methods, in Mechatronics Section at Budapest University of Technology and Economics (BME), Faculty of Mechanical Engineering. He was delegated to the national level conference, namely the National Conference of Student Research Societies (OTDK), where he obtained 2nd place in the Mechatronics Section (http://otdk34muszaki.bme.hu/content/34_otdk_muszaki_eredmenyek.pdf).

3) The project was presented to the general public within the framework of the European Robotics Week 2019 on November 22, title: Open house at the Machine Perception Laboratory, link of the event:
https://www.eu-robotics.net/robotics_week/events/erw-2019-events/open_house_at_the_machine_perception_laboratory.5096.html

Our publications cited in the report:

[1] K. Harsányi, A. Kiss, **A. Majdik**, T. Sziranyi, "A Hybrid CNN Approach for Single Image Depth Estimation: A Case Study", Multimedia and Network Information Systems, MISSI 2018, Advances in Intelligent Systems and Computing, vol 833. Springer, 2019. DOI: https://doi.org/10.1007/978-3-319-98678-4_38

[1a] K. Harsányi, A. Kiss, **A. Majdik**, T. Szirányi, "Hibrid CNN hálózat egyetlen kép alapú mélység becsléséhez: egy esettanulmány", Képfeldolgozók és Alakfelismerők Társaságának 12. országos konferenciája, NJSZT KÉPAF, 2019.

[2] **A. Majdik**, Ch. Till, D. Scaramuzza, "The Zurich Urban Micro Aerial Vehicle Dataset", The International Journal of Robotics Research, vol. 36, no. 3, 2017. DOI: https://doi.org/10.1177/0278364917702237, IF: 4.047

[3] **A. Majdik**, D. Verda, Y. Albers-Schoenberg, D. Scaramuzza, "Air-ground Matching: Appearance-based GPS-denied Urban Localization of Micro Aerial Vehicles", J. Field Robotics, 32: 1015-1039, 2015. DOI: https://doi.org/10.1002/rob.21585, IF: 2.059

[4] S. Gazdag, **A. Majdik**, "Accurate appearance-based localization in 3D maps", research report (paper draft) to be submitted to IEEE Robotics and Automation Letters journal, RA-L with IROS 2020 option, deadline: February 24, 2020, IF: 2016 was the first year of publication, therefore the impact factor is expected by the end of 2019.

[5] D. Rozenberszki, **A. Majdik**, "The MTA SZTAKI micro aerial vehicle and motion capture arena", Conference of Hungarian Association for Image Analysis and Pattern Recognititon, January 28-31, 2019.

[6] I. Eichhardt, D. Baráth, "Optimal Multi-view Correction of Local Affine Frames", 30th British Machine Vision Conference, Cardiff, UK, BMVC 2019.

[7] D. Baráth, Z. Kukelova, "Homography from two orientation-and scale-covariant features", Proceedings of the IEEE International Conference on Computer Vision, pp. 1091-1099, ICCV 2019.

[8] K. Harsányi, A. Kiss, T. Szirányi, **A. Majdik**, "MASAT: A fast and robust algorithm for pose-graph initialization", Pattern Recognition Letters, Vol. 129. pp. 131-136, ISSN 0167-8655, 2020. DOI: http://doi.org/10.1016/j.patrec.2019.11.010, IF: 2.81 (as of 2018)

[9] D. Baráth, J. Matas, and J. Noskova. "MAGSAC: marginalizing sample consensus", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10197-10205, CVPR 2019.

[10] D. Baráth J. Matas, “Progressive-X: Efficient, Anytime, Multi-Model Fitting Algorithm”, Proceedings of the IEEE International Conference on Computer Vision, pp. 13780- 3788, ICCV 2019.

[11] D. Rozenberszki, **A. Majdik**, “LOL: Lidar-only Odometry and Localization in 3D point cloud”, IEEE International Conference on Robotics and Automation (ICRA), Paris, France, May 31 - June 4, 2020, submitted (under review).

Other references:

[12] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, N. Navab, “Deeper depth prediction with fully convolutional residual networks“, Conference on 3D Vision (3DV), 2016.

[13] O. Ronneberger, P. Fischer, T. Brox, “U-net: Convolutional networks for biomedical image segmentation”, International Conference on Medical image computing and computer-assisted intervention, Springer 234–241, 2015.

[14] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, “Indoor segmentation and support inference from RGBD images“, European Conference on Computer Vision, Springer 746–760, 2012.

[15] R. W. Wolcott, R. M. Eustice, "Visual localization within LIDAR maps for automated urban driving", 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 176-183, 2014. DOI: 10.1109/IROS.2014.6942558

[16] R. Mur-Artal, J. M. M. Montiel, J. D. Tardós, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System", IEEE Transactions on Robotics, vol. 31, no. 5, pp. 1147-1163, 2015. DOI: 10.1109/TRO.2015.2463671

[17] M. A. Fischler, R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography", Communications of the ACM, 1981.

[18] J. Zhang and S. Singh, “Loam: Lidar odometry and mapping in realtime”, Robotics: Science and Systems, RSS 2014.

[19] R. Dube, A. Cramariuc, D. Dugas, J. I. Nieto, R. Siegwart, C. Cadena, “SegMap: 3D segment mapping using data-driven descriptors”, Robotics: Science and Systems, RSS 2018.

[20] A. Geiger, P. Lenz, R. Urtasun, "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite", Proceedings of the IEEE Conference on Conference on Computer Vision and Pattern Recognition, CVPR 2012.

Our publications not strictly related to this project:

[21] T. Szirányi T, A. Kriston, T. Csilling, **A. Majdik**, L. Tizedes, "Fusion Markov Random Field Image Segmentation for a Time Series of Remote Sensed Images", Progress in Industrial Mathematics at ECMI 2018, 20th European Conference on Mathematics for Industry, Springer, pp. 621-629., 2019.

[22] T. Szirányi T, A. Kriston, T. Csilling, **A. Majdik**, L. Tizedes, "Fusion Markov Random Field Image Segmentation for a Time Series of Remote Sensed Images", Képfeldolgozók és Alakfelismerők Társaságának 12. országos konferenciája, NJSZT KÉPAF, 2019.