LONG-TERM VISUAL LOCALIZATION IN LARGE SCALE URBAN ENVIRONMENTS EXPLOITING STREET LEVEL IMAGERY

J. Meyer [1, *], D. Rettenmund[1], S. Nebiker[1]

[1] Institute of Geomatics, FHNW University of Applied Sciences and Arts Northwestern Switzerland, Muttenz, Switzerland – (jonas.meyer, stephan.nebiker)@fhnw.ch, daniel.rettenmund@gmail.com

Commission II, WG II/1

KEY WORDS: Visual Localization, Image-Based Localization, Long-Term Matching, Image Orientation, Pose Estimation, Benchmark, Georeferencing, Ubiquitous Positioning

ABSTRACT:

In this paper, we present our approach for robust long-term visual localization in large scale urban environments exploiting street level imagery. Our approach consists of a 2D-image based localization using image retrieval (NetVLAD) to select reference images. This is followed by a 3D-structure based localization with a robust image matcher (DenseSfM) for accurate pose estimation. This visual localization approach is evaluated by means of the 'Sun' subset of the RobotCar seasons dataset, which is part of the Visual Localization benchmark. As the results on the RobotCar benchmark dataset are nearly on par with the top ranked approaches, we focused our investigations on reproducibility and performance with own data. For this purpose, we created a dataset with street-level imagery. In order to have independent reference and query images, we used a road-based and a tram-based mapping campaign with a time difference of four years. The approximately 90% successfully oriented images of both datasets are a good indicator for the robustness of our approach. With about 50% success rate, every second image could be localized with a position accuracy better than 0.25 m and a rotation accuracy better than 2°.

1. INTRODUCTION

Modern vehicle-based and portable mobile mapping systems with multi-camera sensor systems combined with state-of-the-art georeferencing techniques enable a large-scale acquisition of accurate street level imagery. The resulting georeferenced collections of indoor or street level imagery covering large building complexes, entire cities or even states provide a powerful basis for urban infrastructure management. They furthermore bear a great potential for accurate visual localization and 6DOF pose estimation – even in areas with no or only poor GNSS coverage. Such a universally applicable visual localization would, for example, enable highly accurate Augmented Reality (AR) applications with robust absolute (re-)localization in large-scale indoor and outdoor environments without a need for an additional positioning infrastructure, such as GNSS or WiFi. Furthermore, visual localization could be used to significantly improve existing inaccurate positioning in street canyons.

In our previous work, we had discussed the concept and exploitation of 3D image spaces (Nebiker et al., 2015), consisting of collections of accurately georeferenced RGB-D images. Capturing such 3D image spaces requires high quality mobile mapping systems and advanced georeferencing techniques (Cavegn et al., 2018). When it comes to keeping the data up to date, the cost of exclusively using high-quality capturing systems would be enormous. Hence, there should be a solution for integrating images captured by non-geospatial experts with consumer devices such as smartphones. However, these consumer devices do not contain precise positioning sensors. This so far limited georeferencing accuracies to a few meters in outdoor environments and even prevented reliable indoor positioning. Visual localization using existing 3D image spaces as a reference, not only promises to address the task of sensor positioning but also the task of determining the sensor pose. Only if both tasks can be solved reliably and accurately, can the new imagery be integrated into the existing database and used for measurement and asset management tasks. In addition to database updating and asset management, there is a high demand for real-time device pose estimation for augmented reality applications, where 3D image spaces have a great potential for serving as reference data. First investigations of visual localization using 3D image spaces showed that large temporal differences and the associated changes in scene content and appearance are one of the main challenges in long-term visual localization (Rettenmund et al., 2018).

In this paper, we investigate and demonstrate the capabilities of state-of-the art visual localization methods in large-scale urban environments. For this, we first introduce our processing pipeline, which is built on top of our highly scalable street level imagery database. We then introduce our long-term visual localization approach emphasising robust and accurate long-term matching. We subsequently evaluate our approach, first using the 'RobotCar Seasons' dataset of the long-term visual localization benchmark and second using our own Basel Bench50 dataset. We finally discuss the results, which demonstrate the capability of our visual localization approach to reliably and accurately determine 6DOF image poses in urban spaces.

2. RELATED WORK

Pushed forward by innovations such as augmented reality and autonomous driving, the field of visual localization is in rapid

* Corresponding author

evolution. In order to establish a possibility for comparing the results of different visual localization approaches Sattler et al. (2018) created a benchmark with several datasets, each with some distinct characteristics. They also give an overview on the various strategies for visual localization. Coarsely, they classify the approaches into 3D-structure based, 2D-image based, sequence based and learning based visual localization. In 3D-structure based localization, there is a three-dimensional representation of the environments such as a point cloud or a 3D model. By searching corresponding points in the structure and the image, basic geometric principles can be applied to calculate the image pose. However, the big challenge of this approach is to determine matching points over long time periods and in varying conditions. For speeding up the process, there are several methods, that prioritize points close to a reliable match (Sattler et al., 2012) or augment feature points with additional visibility information (Svärm et al., 2017). However, all these approaches rely heavily on the existence of a sufficient number of matching points. When using 2D-image based methods, the goal is to determine the most similar from the collection of reference images. Because it is quite likely that two similar images have been captured from the same location, this pose is being used as a result. There are methods that use hand-crafted features for describing the image's contents such as DenseVLAD (Torii et al., 2015), while others include some neural networks as NetVLAD (Arandjelovic et al., 2016). The drawback of this method is the requirement of a big number of reference images with different viewpoints to reach good results.

To reduce the false positive rates of single image localization approaches Sattler et al. (2018) propose the use of multiple images in the form of a sequence in the correct order. To estimate the relative poses of the images visual odometry or visual SLAM algorithms can be applied. Current visual odometry and visual SLAM algorithms use feature-based methods such as ORB-SLAM (Mur-Artal et al., 2015) and ORB-SLAM2 (Mur-Artal, Tardós, 2017) or direct methods such as LSD-SLAM (Engel et al., 2015) and DSO (Wang et al., 2017). Known relative poses allow modelling the cameras of the image sequences as a generalized camera (Pless, 2003), i.e. as a camera with multiple centres of projections. The absolute pose from 2D-3D matches can be estimated by using approaches for multi-camera systems (Lee et al., 2015) and camera trajectories (Camposeco et al., 2016).

Learning based localization methods were first introduced by Kendall et al. (2015). The main idea is to train neural networks, so that they directly regress the pose of an image. While this looks promising in some small-scale experiments, this approach is hard to scale to real-world problems. Mueller et al. (2018) showed that it is possible to improve the performance by integrating synthetically generated views of the test site in the training dataset. However, this requires a quite detailed 3D model for rendering these views. Furthermore, Sattler et al. (2019) point out, that the pose regression of these networks is very similar to image retrieval followed by applying a slight pose offset.

In recent publications, those getting the best results combine the strengths of the different approaches, e.g. Sarlin et al. (2019). Instead of using neural networks as a "magic black-box", they are just used as parts of the localization pipeline, where they actually generate some benefits. Thus, network architectures such as L2-Net (Tian et al., 2017; Tian et al., 2019), D2-Net (Dusmanu et al., 2019) or SuperPoint (DeTone et al., 2018) are used to generate point descriptors, which help to generate better matches for the use in structure-based image orientation tools such as COLMAP (Schönberger, Frahm, 2016). Widya et al. (2018) skip the step of keypoint detection by just using an intermediate layer of a convolutional neural network as feature map. This generates a regular grid of feature vectors. By using image-retrieval, the number of image pairs for matching can be reduced. This helps to minimize the computational costs.

3. PROCESSING PIPELINE

3.1 Overview

The reference images of previous mobile mapping campaigns are stored in a large-scale cloud-based architecture and accessible through an applications programming interface (API), which serves the image metadata. By querying the database with the approximate position from the navigation sensors, we get the spatially nearest neighbours. For each image, the pose, an URL to download the actual image and the camera's intrinsics are returned. If the raw orientation of the image is known, the resulting list of images can be filtered further by removing the images, whose projection centres are near to the assumed position, but point to the opposite direction (Figure 1).

Because image matching is the part of the processing workflow, that consumes most of time and resources, we search for the reference images that are most similar to the query image. NetVLAD (Arandjelovic et al., 2016) uses a neural network with the VGG-16 architecture (Simonyan, Zisserman, 2015) to create a global descriptor for each image. Comparing these descriptor vectors is much faster than matching all feature vectors for all keypoints in an image.

Once we have identified the reference images, that are most similar to the query image, we can perform feature matching on a much smaller number of image pairs (Figure 1). To get better geometric conditions, it is important to have the keypoints evenly distributed over the whole image. The DenseSfM approach by Widya et al. (2018), achieves this by using an intermediate feature map of the VGG-16 network as descriptors. Hence, there is a regular grid of feature vectors, that spans all of the image. To increase the accuracy, the features get relocalized into the full-size image by searching for the pixels, that had the biggest influence on the respective descriptor.

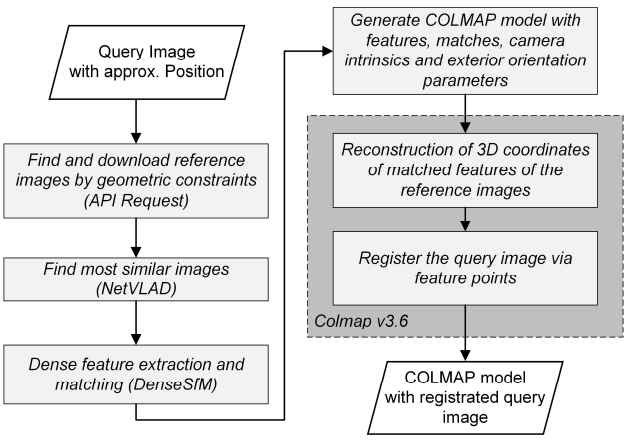

Figure 1. Our processing workflow

For the actual image orientation process, we use COLMAP v3.6 (Schönberger, Frahm, 2016), which allows us to use the known exterior orientation parameters of the reference images. Hence, we get a COLMAP model, where only the query image needs to be aligned with respect to the fixed reference images. First, we generate a COLMAP model with the processed features and

matches as well as the camera specifications and the exterior orientation parameters (EOP) of the reference images. We then use the functions of COLMAP to reconstruct the 3D coordinates of the feature matches by fixing the known EOPs of the reference images and to register the query image via the matched feature points (Figure 1).

3.2 Robust Long-Term Matching

While trying to achieve better results using the traditional SIFT features (Lowe, 2004), we reached a limit of handcrafted feature descriptors for long-term matching. While these features are designed to be invariant for slight changes in illumination and orientation, they fail miserably in robustly establishing long-term correspondences. Especially the changing appearance of the environment caused by seasonal changes such as snow covered or wet roads as well as shadows or strong sunlight leads to serious problems (Figure 2).

Figure 2. Four images of the Robot Car Seasons dataset depicting the same location in different conditions

By matching the same images multiple times with different parametrization of the matcher, we showed that relaxing the restrictions for outlier filtering results in more matches. However, the additional matches have a high probability of getting eliminated during the bundle adjustment, which in other words just means they are outliers that got rejected for some reason.

To overcome these limitations, feature descriptors should incorporate some semantic information. When a neural network is applied to create image features, semantics get somewhat implicitly integrated into the descriptor vector. Thus, we compared different types of trained descriptors. Among LF-Net (Ono et al., 2018), SuperPoint (DeTone et al., 2018), DenseSfM (Widya et al., 2018) and SOSNet (Tian et al., 2019), DenseSfM achieved the best results on our dataset.

4. ROBOTCAR BENCHMARK

To test the performance of our approach, we chose to process a dataset of the long-term visualization benchmark by Sattler et al. (2018). This benchmark provides various image collections along with the corresponding orientation values, which can be used as reference data, as well as some images, where no further data is provided, which are used as query images.
Of the datasets provided, 'RobotCar Seasons' is the one, which is most similar to our imagery. It consists of image sequences captured during the Oxford RobotCar experiment by Maddern et al. (2017) and has similar characteristics to street level imagery acquired by mobile mapping systems.

4.1 Evaluation Strategy

The long-term visual localization benchmark uses a joint evaluation of position and rotation. The calculated poses are compared to the ground truth poses as follows: For the positions, the Euclidean distance is used. And for the rotations, the minimal angle required to align the two rotations is computed. The formulae used for the calculation can be found in Sattler et al. (2018).

The poses are assigned to three precision classes: High, Medium and Coarse. The criteria for a class are matched, if the differences for both position and rotation are below a certain threshold. These thresholds vary depending on the dataset. The threshold values given for RobotCar Seasons are shown in Table 1.

	Position [m]	Rotation [deg]
High	0.25	2
Medium	0.5	5
Coarse	5	10

Table 1. Threshold values for RobotCar Seasons

4.2 Reference Data

The Oxford RobotCar platform was used to capture a long-term dataset for autonomous driving use cases. The RobotCar is equipped with different cameras, 2D and 3D LiDAR, as well as an inertial and GNSS navigation system (Maddern et al., 2017). The imagery selected for the RobotCar Seasons benchmark dataset, was captured at intervals of one Meter using three synchronized global shutter Point Grey Grasshopper 2 cameras. The intrinsics of the cameras as well as their relative poses are known (Sattler et al., 2018). The cameras with a resolution of 1024 x 1024 pixels (1MP) were mounted to the left, rear and right of the car. A more detailed description of the configuration and specification of the cameras can be found in Maddern et al. (2017).

Sattler et al. (2018) created 49 non-overlapping local 3D models from a reference traversal by using bundle adjustment. The query images were obtained by gathering all images within 10 m of a reference position in each 3D model. The Dataset consists of 26121 reference images and 11934 query images of nine traversals covering different seasonal and illumination conditions. The poses of the reference images are available in local COLMAP model coordinates. For competitive reasons, neither the exact nor approximated poses of the query images are provided.

4.3 Test Site

As the whole RobotCar Seasons dataset contains 11'934 query images, we decided to process only one traversal. As the

traversals cover various weather conditions, we choose the 'sun' traversal, which is most similar to typical street level imagery and to the images we intend to process in this pipeline. The sun subset of RobotCar Seasons consists of 1380 query images.

4.4 Collection of Approximate Values

In our use case we always have at least very coarse approximate values for the image poses, e.g. from the last GNSS position fix or from WiFi or cellular network IDs. Therefore, we first had to derive initial values from the available original RobotCar data (Maddern et al., 2017). For the calculation of the query image poses in local COLMAP model coordinates, we only used the rear camera, because it has the smallest offset to the inertial and GNSS navigation system. We purposely did not aim at exact approximate values for our use case. For this reason, the offsets of the lever arms and relative orientations of the cameras were not considered. The datasets of the different sensors of the RobotCar are not time synchronous. Therefore, for each image pose of the rear camera the GPS position with the smallest time difference was searched and assigned to the images. A 2D similarity transformation was then calculated between the corresponding images in the global coordinate system and the COLMAP model coordinates. The transformation parameters were applied to the RobotCar positions of the query images in order to obtain the approximate image poses in local COLMAP model coordinates.

4.5 Results

Our results for the investigated 'Sun' subset of the RobotCar Seasons dataset are shown in Table 2. With 89.3% in the Coarse accuracy class (position error < 5 m and orientation error < 10°), 70.2% in the Medium class (< 0.5 m and < 5°), and 47.0% in the High class (< 0.25 m and < 2°), our results are on a competitive level with those of HF-Net (Sarlin et al., 2019), the leading approach at the time of writing. Our results are significantly superior to exclusively neural network based global descriptors, such as NetVLAD (Arandjelovic et al., 2016), which in itself is part of our processing pipeline.

The nearly 90% of successful image localizations in the Coarse class and only 10% 'failed' localizations are a good indicator for the robustness of our approach. Improving the results in the Medium and High class proved to be a challenging undertaking, requiring careful attention to calibration parameters and error-free source code. However, with a 47% success rate in the High class, nearly every second single image is localized with a position accuracy better than 0.25 m and a rotation accuracy of better than 2°.

	RobotCar Seasons - Sun		
	High	Medium	Coarse
HF-Net [%]	52.0	74.3	93.3
Ours [%]	47.0	70.2	89.3
ActiveSearch [%]	29.6	57.4	84.1
NetVLAD [%]	5.7	16.5	86.7

Table 2. Results for the RobotCar Seasons Benchmark ('Sun' subset)

5. BASEL DATASET

As our RobotCar Benchmark results are nearly on par with the top-ranked approaches, we further investigated the reproducibility and performance of our approach with our own street level imagery data.

5.1 Test Site and Data

5.1.1 Acquisition System: The street level data used in the subsequent investigations was captured using our vehicle-based multi-view stereovision mobile mapping system, which was presented in several of our previous publications, including Cavegn et al. (2018). Depending on the system setup, there are several stereo camera systems, a panoramic camera and a GNSS/INS positioning system. All sensors are mounted on a rigid frame that guarantees a stable relative orientation of all stereo systems and the positioning system. With its included positioning sensors, this system delivers the pose of the images by means of direct georeferencing. This 'standard' georeferencing can be improved by post processing the trajectory and including ground control points. An additional improvement can be achieved by image-based georeferencing using bundle adjustment (Cavegn et al., 2016). We treated the image poses from advanced direct georeferencing as known reference values, when visually localizing single images of the sequences.

5.1.2 Used Datasets: We subsequently used two series of mobile mapping imagery that had been captured in the city of Basel (Switzerland) in two independent campaigns, which were four years apart:

a) a *road-based mapping campaign* in Summer 2014 using a car-based mobile mapping system and
b) a *rail-based mapping campaign* in Summer 2018, where the system had been mounted on a tramway.

As can be seen in Figure 3, the Basel dataset covers a dense urban area with multi-storey buildings, narrow streets, partly dense vegetation, overpasses etc., which makes accurate georeferencing a challenge. The two campaigns were georeferenced using state-of-the-art GNSS-based direct sensor orientation. No additional image-based co-registration between the image sequences of the two mapping campaigns was applied. As shown by Cavegn et al. (2018), we thus can expect the trajectory and subsequent absolute pose accuracy of our reference data sets to be in the order of a few decimetres.

In order to have independent reference and query image datasets, we chose the street level imagery from the rail-based campaign b) as reference data set and the imagery from the road-based campaign a) as query dataset. Query images were selected by using spatial operations to filter road segments, that are situated next to tramlines, or do even have tram lines included in the lanes (e.g. see Figure 3, top right, bottom left and bottom right). Then we randomly selected images of the image sequences on these street segments. With a visual verification, we removed the images, that are impossible to localize. Reasons for this could be, that the image is only showing a single wall without any characteristic features or that there are other vehicles that block the view on the environment. We selected 50 query images for our Basel Bench50 dataset (Figure 3). The geometric resolution of the images depends on the sensor. The reference images have resolutions between 5 and 12 MP, while all query images have a resolution of 2 MP. Other than in the RobotCar benchmark, in the Basel Bench50 reference poses are known to the authors, which subsequently enables a more sophisticated and detailed evaluation.

Figure 3. Four representative query images of our Basel Bench50 dataset

5.2 Evaluation

In order to evaluate the results of the investigations on our own data, we used the same precision classes as for the RobotCar dataset (Table 1). We computed the Euclidean distance for the positions and the minimal angle for the rotations between the processed and the ground-truth poses. In addition, we can examine our results for systematic deviations based on the known image poses. Therefore, we generated Scatterplots where the positional differences were plotted against the rotation differences.

5.3 Results

The results of the Basel Bench50 test are shown in Figure 3. With 92% of the localized images in the Coarse accuracy class (position error < 5 m and orientation error < 10°) and 56% in the High class (< 0.25 m and < 2°) the results on our own data outperform the results made on the RobotCar Seasons dataset. The Medium class (< 0.5 m and < 5°) with 60% of oriented images shows a drop of 10% compared to the results on the RobotCar Seasons dataset. The reasons for the better results on our own data in the classes High and Coarse could be due to the better image quality, better geometric resolution, the known camera intrinsics and calibration parameters. However, it should be noted that all images that were verified as 'impossible' to localize were previously removed from the dataset.

	Basel Bench50		
	High	Medium	Coarse
Ours [%]	56	60	92

Table 3. Results on Own Data

There is no indicator of systematic deviations to explain the decrease in the Medium class (Figure 4). The most obvious reason is that our dataset contains local differences between the two campaigns (road and rail-based campaign) because no co-registration of the two campaigns had been done. An indication of this is the accumulation of image poses with positional differences of about 0.8 m and rotational differences of about 0.5° (Figure 4).

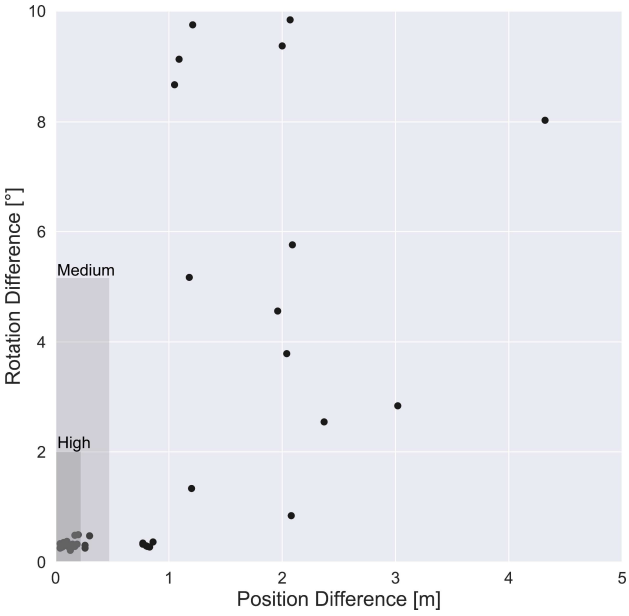

Figure 4. Scatterplot of the positional and rotation differences between processed image poses and ground truth poses

6. CONCLUSION

In this paper we presented our approach for long-term visual localization in urban environments. Our approach combines a 2D-image based and a 3D-structure based visual localization strategy. We first use image retrieval (NetVLAD) to find the most similar images. Then we extract and match densely distributed features by DenseSfM. We subsequently use the SfM-software COLMAP to reconstruct a sparse point cloud of the matched features of the reference images and register the query image to these feature points. We subsequently evaluated our Visual Localization approach using two test data sets. On the RobotCar Seasons dataset of the long-term visual localization benchmark, we achieved results that are nearly on-par with the top ranked methods with 89.3% in the Coarse class (position error < 5 m and rotation error < 10°), 70.2% in the Medium class (< 0.5 m and < 5°) and 47.0% in the High class (< 0.25 m and < 2°). On our own street level data, we were able to outperform the results from the RobotCar dataset in the precision classes High and Coarse by 10% and 3% respectively. The success rate for the Medium class was around 10% lower than the RobotCar dataset.

The 92% of successful image localizations in the Coarse class and only 8% 'failed' localizations are a good indicator for the robustness of our approach. Improving the results in the Medium and High class proved to be a challenging undertaking, requiring careful attention to calibration parameters, consistent and accurate reference data, as well as error-free source code. With a 56% success rate in the High class, more than every second single street-level image is localized with a position accuracy better than 0.25 m and a rotation accuracy of better than 2°.

These results demonstrate the enormous potential of long-term visual localization in combination with accurately georeferenced street level imagery. In this combination, visual localization not only provides accurate and ubiquitous positioning but a powerful 6DOF pose determination method. This could make visual localization an ideal absolute positioning backend for future Augmented Reality applications – in outdoor and indoor environments alike. In order to make such an ubiquitous and

instant 6DOF positioning service a reality, our main current limitation, the very high computational cost and required time for processing an accurate image pose, needs to be overcome.

7. OUTLOOK

In our future work we will test our visual localization workflow with a large (500 images), representative and co-registered dataset without previously removing images classified as 'impossible' to localize. This should show the full accuracy potential of our approach. We will also address the current processing power and time requirements of our feature extraction and matching approach. For this, we will be investigating other feature descriptors with the goal of enabling real-time applications in the longer run.

With regard to the use of our approach in very challenging environments, such as railway tracks, the robustness has to be increased even further. For this purpose, we intend to use sequential information of consecutive images. We expect that the use of image sequences will lead to significantly more robust visual localization results than from single images only.

ACKNOWLEDGEMENTS

This work was funded by the following partners: The fundamental research on visual localization in large-scale outdoor environments was funded by the Swiss National Science Foundation (SNSF) as part of EVAC project (No. 407540_167278) within the National Research Programme NFP75 on "Big Data". The applied research on visual localization exploiting street level imagery was co-funded by the Swiss Innovation Agency (Innosuisse) and by iNovitas AG as part of the cloudIO project (No. 32411.1 IP-ICT).

REFERENCES

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J., 2016: NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5297–5307. Las Vegas, USA.

Camposeco, F., Sattler, T., Pollefeys, M., 2016: Minimal Solvers for Generalized Pose and Scale Estimation from Two Rays and One Point. *Proceedings of European Conference on Computer Vision*, 202–218. Amsterdam, The Netherlands.

Cavegn, S., Blaser, S., Nebiker, S., Haala, N., 2018: Robust And Accurate Image-Based Georeferencing Exploiting Relative Orientation Constraints. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV–2, 57–64. doi.org/10.5194/isprs-annals-IV-2-57-2018.

Cavegn, S., Nebiker, S., Haala, N., 2016: A SYSTEMATIC COMPARISON OF DIRECT AND IMAGE-BASED GEOREFERENCING IN CHALLENGING URBAN AREAS. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B1, 529–536. doi.org/10.5194/isprsarchives-XLI-B1-529-2016.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018: SuperPoint: Self-Supervised Interest Point Detection and Description. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 224–236. Salt Lake City, USA.

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T., 2019: D2-Net: A Trainable CNN for Joint Description and Detection of Local Features. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8092–8101. Long Beach CA, USA.

Engel, J., Stückler, J., Cremers, D., 2015: Large-scale direct SLAM with stereo cameras. *IEEE International Conference on Intelligent Robots and Systems*, 1935–1942. Hamburg, Germany. doi.org/10.1109/IROS.2015.7353631.

Kendall, A., Grimes, M., Cipolla, R., 2015: PoseNet : A Convolutional Network for Real-Time 6-DOF Camera Relocalization. *IEEE International Conference on Computer Vision (ICCV)*, 2938–2946. Santiago, Chile. doi.org/10.1109/ICCV.2015.336.

Lee, G. H., Li, B., Pollefeys, M., Fraundorfer, F., 2015: Minimal solutions for the multi-camera pose estimation problem. *The International Journal of Robotics Research*, *34*(7), 837–848.
Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2), 91–110.

Maddern, W., Pascoe, G., Linegar, C., Newman, P., 2017: 1 Year, 1000 km: The Oxford RobotCar Dataset. *International Journal of Robotics Research*, 36(1), 3–15. doi.org/10.1177/0278364916679498.

Mueller, M., Metzger, A., Jutzi, B., 2018: CNN-Based Initial Localization Improved By Data Augmentation. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV–1, 117–124. doi.org/10.5194/isprs-annals-IV-1-117-2018.

Mur-Artal, R., Montiel, J. M. M., Tardós, J. D., 2015: ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics*, 31(5), 1147–1163. doi.org/10.1109/TRO.2015.2463671.

Mur-Artal, R., Tardós, J. D., 2017: ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Transactions on Robotics*, 33(5), 1255–1262.

Nebiker, S., Cavegn, S., Loesch, B., 2015: Cloud-Based Geospatial 3D Image Spaces—A Powerful Urban Model for the Smart City. *ISPRS International Journal of Geo-Information*, 4(4), 2267–2291. doi.org/10.3390/ijgi4042267.

Ono, Y., Trulls, E., Fua, P., Yi, K. M., 2018: LF-Net: Learning Local Features from Images. *32nd Conference on Neural Information Processing Systems (NIPS)*. Montreal, Canada.

Pless, R., 2003: Using many cameras as one. *IEEE International Computer Society Conference an Computer Vision and Pattern Recognition*. Madison, USA.

Rettenmund, D., Fehr, M., Cavegn, S., Nebiker, S., 2018: Accurate Visual Localization In Outdoor And Indoor Environments Exploiting 3D Image Spaces As Spatial Reference. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII–1, 355–362. doi.org/10.5194/isprs-archives-XLII-1-355-2018.

Sarlin, P.-E., Cadena, C., Siegwart, R., Dymczyk, M., 2019: From Coarse to Fine: Robust Hierarchical Localization at Large Scale. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 12716–12725. Long Beach CA, USA.

Sattler, T., Leibe, B., Kobbelt, L., 2012: Improving Image-Based Localization by Active Correspondence Search. *European Conference on Computer Vision (ECCV)*, 752–765. Florence, Italy. doi.org/10.1007/978-3-642-33718-5_54.

Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., Kahl, F., Pajdla, T., 2018: Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8601–8610. Salt Lake City, USA. doi.org/10.1109/CVPR.2018.00897.

Sattler, T., Zhou, Q., Pollefeys, M., Leal-Taixe, L., 2019: Understanding the Limitations of CNN-based Absolute Camera Pose Regression. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach CA, USA.

Schönberger, J. L., Frahm, J.-M., 2016: Structure-from-Motion Revisited. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4104–4113. Las Vegas, USA. doi.org/10.1109/CVPR.2016.445.

Simonyan, K., & Zisserman, A., 2015: Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learing Representations (ICLR)*. San Diego, USA.

Svärm, L., Enqvist, O., Kahl, F., Oskarsson, M., 2017: City-Scale Localization for Cameras with Known Vertical Direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7), 1455–1461. doi.org/10.1109/TPAMI.2016.2598331.

Tian, Y., Fan, B., Wu, F., 2017: L2-Net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, USA.

Tian, Y., Yu, X., Fan, B., Wu, F., Heijnen, H., Balntas, V., 2019: SOSNet: Second Order Similarity Regularization for Local Descriptor Learning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach CA, USA.

Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., Pajdla, T., 2015: 24/7 Place Recognition by View Synthesis. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1808–1817. Boston, USA.

Wang, R., Schworer, M., Cremers, D., 2017: Stereo DSO: Large-Scale Direct Sparse Visual Odometry with Stereo Cameras. *Proceedings of the IEEE International Conference on Computer Vision*, 3923–3931. Venice, Italy. doi.org/10.1109/ICCV.2017.421.

Widya, A. R., Torii, A., Okutomi, M., 2018: Structure-from-Motion using Dense CNN Features with Keypoint Relocalization. *IPSJ Transactions on Computer Vision and Applications*, 10(1).

Revised May 2020