View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Directory of Open Access Journals

ACCURATE VISUAL LOCALIZATION IN OUTDOOR AND INDOOR ENVIRONMENTS EXPLOITING 3D IMAGE SPACES AS SPATIAL REFERENCE

D. Rettenmund [1], M. Fehr [1], S. Cavegn [1, 2], S. Nebiker [1]

[1] Institute of Geomatics, FHNW University of Applied Sciences and Arts Northwestern Switzerland, Muttenz, Switzerland - (daniel.rettenmund, markus.fehr, stefan.cavegn, stephan.nebiker)@fhnw.ch
[2] Institute for Photogrammetry, University of Stuttgart, Germany

Commission I, ICWG I/IV

KEY WORDS: Image-Based Localization, Visual Localization, ARCore, Image Orientation, Pose Estimation, 3D Image Spaces

ABSTRACT:

In this paper, we present a method for visual localization and pose estimation based on 3D image spaces. The method works in indoor and outdoor environments and does not require the presence of control points or markers. The method is evaluated with different sensors in an outdoor and an indoor test field. The results of our research show the viability of single image localization with absolute position accuracies at the decimetre level for outdoor environments and 5 cm or better for indoor environments. However, the evaluation also revealed a number of limitations of single image visual localization in real-world environments. Some of them could be addressed by an alternative AR-based localization approach, which we also present and compare in this paper. We then discuss the strengths and weaknesses of the two approaches and show possibilities for combining them to obtain accurate and robust visual localization in an absolute coordinate frame.

1. INTRODUCTION

Georeferenced collections of indoor or street level imagery covering large building complexes or entire cities provide great potential for accurate visual localization and pose estimation. In this paper, we present first results and insights from our work on image-based localization. Besides the description of our fully automated processing pipeline for single image orientation, we emphasize multiple techniques for reference image selection, and furthermore present evaluation results of our approach in both indoor and outdoor environments.
In our previous work, we introduced the concept of 3D image spaces (Nebiker et al., 2015). These collections of georeferenced RGB-D images provide an intuitive interface to digital models of urban areas. Capturing such 3D image spaces requires high quality mobile mapping systems. When it comes to keeping the data up-to-date, the cost for using high quality capturing systems would be enormous. Hence, there should be a solution for integrating images taken by consumer devices like smartphones, which do not contain precise positioning sensors. Additionally, there is a high demand for real-time device pose estimation for augmented reality applications, where 3D image spaces have a high potential for serving as reference data.

1.1 Related Work

The topic of visual localization is of interest for many different disciplines. Even Google has announced plans to enable pedestrian navigation using images in their visual positioning service (Cooper, 2018). Hence, there are several distinctive approaches under research. Sattler et al. (2018) distinguish the following categories: 3D structure-based, 2D image-based, sequence-based and learning-based localization.
The 3D structure-based approach, as in Schönberger et al. (2018) and Taira et al. (2018), uses 3D structures like point clouds, which serve as a reference for feature matching with the query image. Image-based localization works similar, but only uses one reference image. A big problem for both of these approaches are changes in viewpoint. Karpushin (2016) addresses this challenge with the use of a specialised RGB-D feature detector and descriptor, which leads to significantly improved results. Learning-based localization methods use a neural network for directly regressing the image pose. Examples for this approach are PoseNet by Kendall et al. (2015), VidLoc (Clark et al., 2017), which adds a LSTM to exploit image sequences and MapNet, where geometric constraints are included (Brahmbhatt et al., 2018). Other recent publications no longer use an end-to-end network approach but focus on training the individual pieces of the localization pipeline (Brachmann and Rother, 2018).

With augmented reality, it is possible to place virtual objects in the real world and create an illusion of realism. There are two techniques to determine the pose of the virtual object in the real world. The marker-based solution uses pre-calculated image descriptors, which are matched in real time with the current camera frame. Wagner et al. (2008) presented two approaches (SIFT and Ferns) for robust pose estimation of planar markers in real time on mobile phones. Wüest and Nebiker (2017) showed the feasibility of using image-based multi-markers for large-scale augmented reality applications, e.g. in museums.
On the other hand, there is the marker-less solution, where the pose of an object is estimated through tracking features over movement and time of the camera frame.
Lee and Hollerer (2008) described a hybrid feature tracking approach for marker-less augmented reality. Also in medicine field marker-less augmented reality is used. Kilgus et al. (2015) mounted a range camera on a tablet and estimated the camera pose based on depth data and surface registration of computer tomography.
Recently two new augmented reality framework entered the market: Google's ARCore (Gosalia, 2018) and Apple's ARKit (Apple, 2017). Both support marker-based and marker-less augmented reality solution.

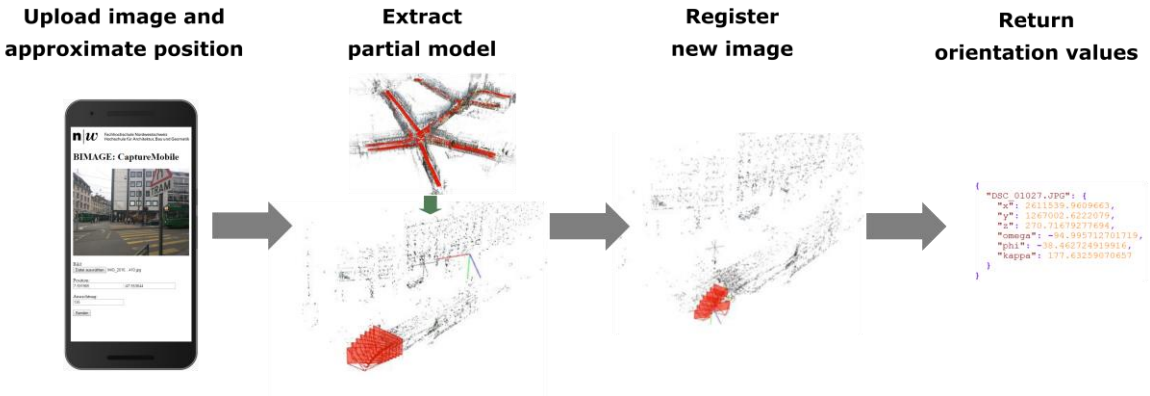

Figure 1. Automated web service-based processing pipeline of our visual localization approach that registers a single image to reference 3D image spaces and thus computes its pose

2. IMAGE ORIENTATION PIPELINE BASED ON 3D IMAGE SPACES

2.1 Overview of our Visual Localization Approach

Our method enables determining the pose of single images using a database of georeferenced RGB-D imagery. Our processing pipeline is based on the image registration functionality of COLMAP (Schönberger and Frahm, 2016) and is shown in Figure 1. It requires a pre-built model as reference, which features precise relative orientations between the reference images. Using this reference model speeds up the process since feature extraction and matching only need to be done for the new image. To avoid feature matching with a large number of reference images, we extract a local sub-model before processing. The query image is then registered to the existing image bundle of this sub-model using the corresponding feature points. As shown in Figure 1, our image orientation tool is integrated into a web service, which enables the user to upload an image together with its approximate position. The query image is then oriented and the service returns the pose of the image.
Since we use COLMAP for the image registration, the crucial part of this process is the selection of the reference images, which requires a reasonably accurate initial value for the position of the query image. We subsequently discuss our approach for these challenges in sections 2.2 and 2.3.

2.2 Determination of Initial Image Pose

The search for similar images with techniques like bag of visual words (Nistér and Stewénius, 2006) is difficult in urban and indoor environments, since there are many repetitive structures (Kendall et al., 2015).
Therefore, we use the current position to narrow down the number of potential reference images. When using a smartphone, the current pose can be requested from the device's positioning sensors. These provide the position determined by satellite navigation, and the rotation of the device. However, these values are not always available. When entering urban canyons, the position accuracy degrades and inside of buildings, it is missing completely. Hence, there need to be other approaches to obtain an initial pose. Our current solution relies on some information given by the user. In indoor use cases, we ask for the current room number, from which we derive the coordinate values. Additionally there is a possibility to set the current position and viewing direction on an overview map of the study area.
As an alternative to the single image approach, we developed a prototype application to make use of the most recent augmented reality frameworks. It allows initializing its position using control points with known coordinates and then tracks the movement of the device. We discuss this in detail in section 3.3.

2.3 Selection of Reference Images

We have implemented different strategies for the selection of reference images based on the initial pose. The basic approach uses the approximate position and selects the spatially nearest neighbours in the reference model. If we know the device's orientation, we filter the candidate images based on the viewing direction in order to reduce unsuitable reference images.
The more sophisticated approach searches for reference images by exploiting the sparse 3D point cloud of the COLMAP model. In a first step, we project the assumed field of view into the model. We then select points that could possibly be visible in the query image. Afterwards we return the images that contain most of these points.
For speeding up the selection of corresponding images in the indoor use case, where we obtain the room number from the user, we segmented the COLMAP model based on the buildings floorplan. Thus, the query execution for all images in a specific room becomes very efficient.

3. ACQUISITION SYSTEMS AND CORRESPONDING STRATEGIES FOR EVALUATION OF VISUAL LOCALIZATION APPROACH

For the evaluation of our method, we used three different sensor systems with four types of image sensors. As shown in Table 1, the specifications of these sensors differ significantly. Furthermore, not all systems supply a ground truth pose and hence there is need for different evaluation methods. In the following sections, we present the three systems and the respective evaluation strategies.

Sensor	Resolution [px]	Pixel Size [µm]	Focal Length [mm]
MM: AVT	4008 x 2672	9.0	21
MM: Basler	1920 x 1080	7.4	8
DSLR D7000	4928 x 3264	4.8	18
Galaxy S8	2220 x 1440	2.8 (1.4)	4.4

Table 1. Overview of the sensors used in our evaluation

3.1 Image Sequences from Mobile Mapping

We used data that we captured using our vehicle-based multi-sensor stereovision mobile mapping system, which was presented in several of our previous publications, including Cavegn et al. (2018). It consists of three stereo systems featuring industrial cameras with CCD sensors and a GNSS/INS positioning system. As shown in Figure 2, we mounted all sensors on a rigid frame that guarantees a stable relative orientation of all stereo systems and the positioning system.
The stereovision system facing forward consists of two 11 MP AVT cameras and has a calibrated stereo base of 905 mm. The cameras have a resolution of 4008 x 2672 pixels at a pixel size of 9.0 µm, a focal length of 21 mm and a resulting field of view of 81° in horizontal and 60° in vertical direction. In addition, there are two stereovision systems pointing left and back-right respectively. They include Basler HD cameras with a resolution of 1920 x 1080 pixels, a pixel size of 7.4 µm, a focal length of 8 mm and a field of view of 83° x 53°. The base lengths of these systems are 779 mm (back-right) and 949 mm (left).

Figure 2. Sensor configuration of our vehicle-based mobile mapping system (Cavegn et al., 2018)

With its included positioning sensors, this system delivers the pose of the images, which can be further improved by post-processing the trajectory and including ground control points. Hence, we treated these image poses as known reference values, when orienting single images of the sequences. To determine the accuracy of our newly calculated poses, we simply computed the difference between the target and the actual values. Since the definition of rotations in three-dimensional space is ambiguous, we combined the individual Euler angles to a single spatial angle difference.

3.2 Roundshot Images

To evaluate the precision of the calculated projection centres, we created a second evaluation dataset, using a DSLR camera (Nikon D7000) mounted on a RoundShot VR Drive panoramic tripod, as shown in Figure 3. We will subsequently refer to these pictures as ‘roundshot images’. The camera has a 23.6 x 15.6 mm CMOS sensor, which records images with a resolution of 4928 x 3264 pixels. We used a Nikkor zoom lens as objective and mechanically fixed its focal length to 18 millimetres. By using the panoramic tripod, we achieve that all images recorded from the same station have an identical projection centre. This allows us to evaluate the precision of the projection centres calculated by our orientation pipeline. For this purpose, we calculated the standard deviation of all projection centres per location. To get a measure for the overall precision, we calculated the differences from all projection centres to the centre points of the corresponding station in order to compute the standard deviation over the whole dataset.
In addition to evaluating the projection centre positions, we also analysed the accuracy of the pose angles. This was done by re-projecting checkpoints with known coordinates into the image plane using the computed pose. By comparing the re-projected point with its real position, we could determine the residuals in image space. These differences provide a measure for the quality of the image pose.

3.3 Smartphone with ARCore Application

We furthermore used new augmented reality techniques to track the position and orientation of a mobile phone. For this purpose, we used the augmented reality framework ARCore by Google released in February 2018 (Gosalia, 2018).
ARCore estimates the pose of the mobile device with concurrent odometry and mapping (COM). COM uses feature points to compute its change in location over move and time, in combination with accelerometer and gyroscope measurements from the device’s sensors (Google, 2018). Through clustering the feature points, ARCore builds a map of its environment and – as a recent feature – detects horizontal and vertical surfaces.
Our prototype application is able to capture images and get the corresponding camera position and orientation from the estimated ARCore pose. We developed the app with Unity3D and used the ARCore version 1.2.0. We implemented two different approaches for geolocalisation. To use these methods, we have to introduce control points in two coordinate systems. The control points are known in the target coordinate system and have to be measured in the local coordinate system. To calculate the local position of the control points, our app detects surfaces (planes) over the known control point with help of the ARCore technology. To determine the local coordinates, we use a ray cast and calculate its intersection with the detected plane. The first geolocalisation method requires one known control point and one known direction to translate and rotate the local coordinate system into the national coordinate system. The second approach uses a 2D Helmert similarity transformation. Before we can start capturing images with the device, we have to measure two known points. During measurement periods, we can dynamically add new control points to stabilize the transformation. It is also possible to delete former points, which are too far away and deteriorate the result. To calculate the translation of the height, we used the mean of the differences.

Figure 3. Nikon D7000 with Roundshot panoramic tripod (left) and our ARCore application (right)

For our investigations, we used a Samsung Galaxy S8 and captured images with a resolution of 2220 x 1440 pixels. The camera sensor built into the phone has a pixel size of 1.4 µm. As we save the images with a resolution reduced to half, this leads to a virtual pixel size of 2.8 µm. In a calibration process, we determined a focal length of 4.4 millimetres. For the evaluation of the image poses, we re-projected checkpoints into the images as described in section 3.2.

4. EVALUATION IN AN OUTDOOR ENVIRONMENT

4.1 Test Site and Reference Data

Our outdoor test site is located at the Bankverein, a busy road junction in the city centre of Basel. In addition to numerous tramlines and the corresponding stops, there are also five roads, where stereo image sequences of several mobile mapping campaigns are available. Furthermore, there are independent control points as well as a point cloud, captured using a total station and a terrestrial laser scanner respectively.
Our reference dataset comes from a survey in July 2014, where we used our mobile mapping system described in section 3.1, with its full sensor configuration. The positions of the images are shown in Figure 4 as purple dots. In total, the dataset contains 3387 images, which were combined into a COLMAP model. To align the locally oriented models to national coordinates, we performed a 3D similarity transformation, using the projection centres with known coordinates as tie points.

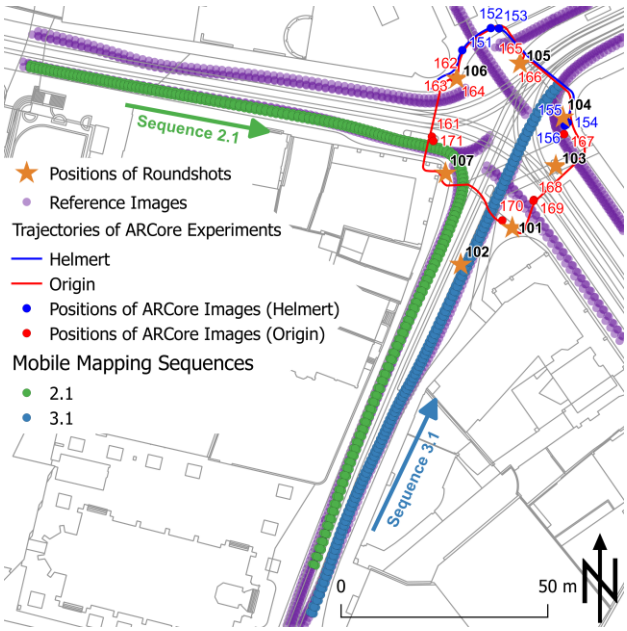

Figure 4. Map of the outdoor test site including the image positions (Source: Geodaten Kanton Basel-Stadt)

4.2 Evaluation of Mobile Mapping Images

First, we tried to orient images from a later survey using our pipeline. We used our mobile mapping system described in section 3.1 to capture image sequences in August 2015. In this test, we only used the images from the front system. This leads to a dataset of 307 images, whose poses we calculated in a bundle adjustment using some control points with Agisoft PhotoScan. These image sequences are described in detail by Cavegn et al. (2016). As approximate pose, we used the values from the direct sensor orientations.

Sequence	Rate	2D [cm]	H [cm]	3D [cm]	Angle [°]
Seq. 2.1	177/191	12	5	14	0.21
Seq. 3.1	103/116	15	1	15	0.20

Table 2. Median differences between reference orientation parameters and calculated values

We processed each image independently, intentionally not using the sequential information, to examine the possibility of single image localization. As shown in Table 2, our approach was able to align successfully a total number of 280 images (i.e. 91% of the test data set). This table lists the median values, whereby the influence of outliers is reduced. It is evident that the accuracy of the angle and the height is significantly better than the position.
There are explanations for the 27 images, for which the orientation process failed. For example, the mobile mapping vehicle drove next to a tram for a while. Consequently, the images from this segment mainly contain the side panel of the tram and the rest of the image shows a façade, which was covered by scaffolding on the reference images. The few images with unsatisfying poses can be justified too, since they show the aforementioned façade as well as another building with reflective glass cladding.
The significantly better results for the height in sequence 3.1 are due to the terrain in the study area. While sequence 2.1 includes a segment with a slight slope, the whole sequence 3.1 is approximately at the same level.

4.3 Evaluation of Roundshot Images

For determining the reliability of our system, we captured panoramic image series using the setup described in section 3.2. Overall, we created seven image series from different locations all around the junction. In Figure 4, orange stars indicate these locations numbered from 101 to 107. Each series consists of 30 to 40 images. This means that there is an intermediate angle of around 10 degrees between two consecutive images.
Table 3 shows the rate of successfully oriented images and the standard deviation of the calculated projection centres per location. When looking at the numbers, the poor results of location 101 immediately attract the attention. Beside the very low proportion of only four successfully oriented images, this location shows by far the largest standard deviations. Except location 106, all of the remaining series have a similar success rate of around two thirds to three quarters. The values of the standard deviations show, that the height component is far more accurate than the location. Especially when comparing the combined values for 2D and 3D, it is obvious that the major part of the total error is due to the uncertainties of the 2D position.

Location	Rate	E [cm]	N [cm]	H [cm]	2D [cm]	3D [cm]
101	4/31	34	248	25	251	252
102	25/37	68	84	17	108	109
103	27/41	57	114	20	127	129
104	21/34	70	104	22	125	127
105	25/33	59	53	11	79	80
106	16/33	111	84	15	139	140
107	21/31	71	89	26	114	116
Total	139/240	50	72	13	87	88

Table 3. Standard deviations of projection centres per location for our outdoor environment

The poor results of location 101 can be explained by its location, where several difficulties occur at once. First, the sidewalk it is located on has dramatically changed its appearance. Then the images facing the north-eastern direction all show the façade that was heavily scaffolded on the reference images, which additionally are covered by a tram that was passing by during the mobile mapping campaign. Furthermore, the location is too close to the building in the southeast, so that its characteristic façade is cut off, if it is on the pictures at all. Finally, the selection of the reference images was, suboptimal. Due to the directionally separated lanes of the nearest road segment, the building to the north is not showing up in the reference images as all.
Many of the images from location 106 show the north-eastern façade of the building to the west of the junction. The appearance

of this façade changed since the acquisition of the reference data, due to a refurbishment. Unfortunately, some logos of the company look quite the same as before, but have been changed. This leads to a very strange behaviour of COLMAP when aligning the image.

As Figure 6 shows, the size of reprojection errors varies greatly. On some images, all points have very small residuals, while others show large differences. The median of all residuals is at 9.3 pixels. By visually inspecting the residual plots, it is obvious that the reprojection errors are not homogeneous over a whole image. The pattern of the error vectors often makes it evident, which parts of an image contain most feature points and therefore had the biggest impact on the image orientation process.

In general, the results of the outdoor evaluations show the big challenges for visual localization. In ideal cases, images can be aligned successfully, but there are some major difficulties. Since the environment itself is changing over time, it becomes trickier to align new images to a reference dataset, as the reference data gets older. In our case, with four-year-old images of a busy city centre, where changes occur rapidly, many attempts fail. In Figure 5, there is a rare example of an image that could be aligned despite the reference images showing mainly scaffolding present at the time. The vectors indicating the reprojection error are scaled by a factor of five to enhance the visibility. The small vectors on the left side of the image show, that this part has mainly been used for aligning, while the residuals get larger on the façade that has been scaffolded.

Figure 5. Example of a successfully oriented image (bottom) overlaid with residuals of reprojection (scaled 5x). This picture could be aligned, even though in the reference images an entire building façade was covered by scaffolding (top).

4.4 Evaluation of ARCore Image Series

To test our ARCore app prototype, we captured image series using both alignment strategies, which we described in section 3.3. Figure 4 shows the trajectories and image positions of these series. For the first series using the origin method, the alignment was done on the traffic island in the west of the junction. After walking around the junction in clockwise direction, it ended at the same location. During this series we captured 11 images. In the series using Helmert transformation, we initialized using points on the crosswalk in the north-west and completed the series after capturing seven images in the opposite side of the junction, as shown in Figure 4.

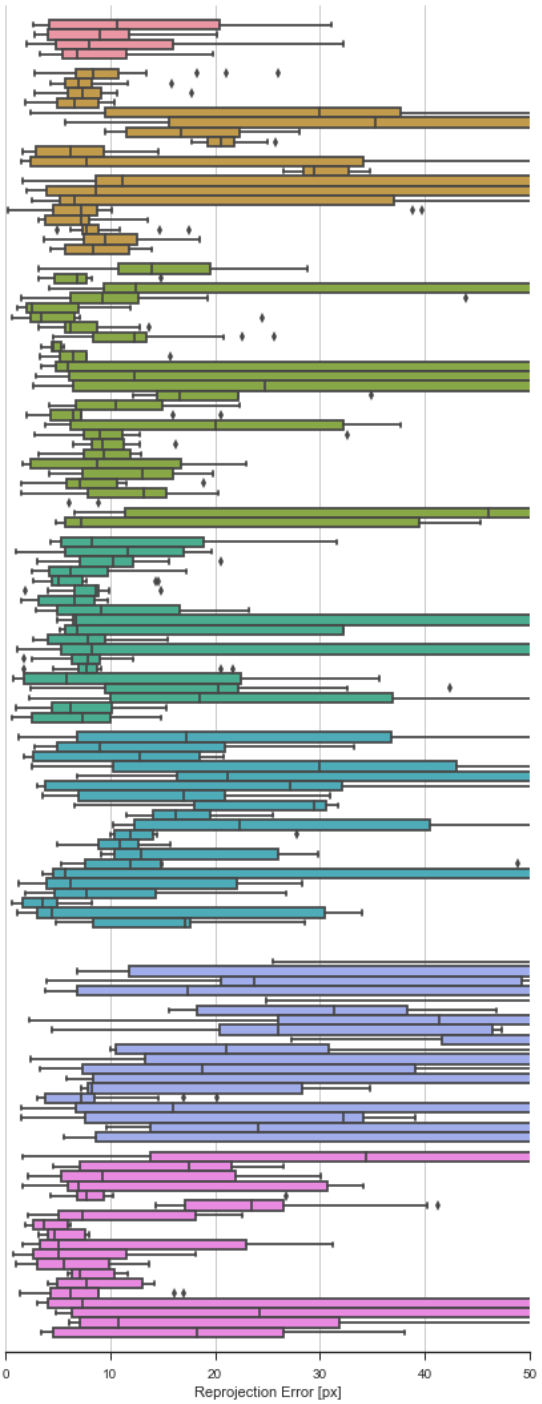

Figure 6. Boxplots for reprojection errors of check points in our outdoor scene. Each location is indicated by a different colour, from 101 on top to 107 on bottom. X axis is cut off at 50 pixels to prevent extreme distortion due to outliers.

The orientation process was successful for eight images of the origin series and for five of the Helmert one. The failure of the other images can be explained in the same way as the failed examples of the roundshot series. The positions for most of the images seamed plausible, since they were close to those determined by our ARCore app. Only for image 164 we got an insufficient result. This is also visible in the boxplot of the point reprojections, which are depicted in Figure 7. All the other images show good results that outperform those of the roundshot experiment. When comparing with the results of Figure 6, one needs to bear in mind, that pixels in an ARCore image are approximately twice as large as those of the D7000.

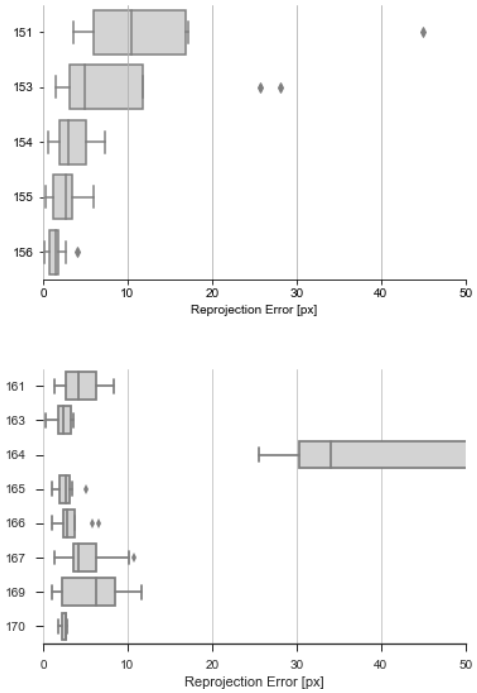

Figure 7. Boxplots of reprojection error for ARCore outdoor series using Helmert transformation (top) and origin method (bottom) respectively

5. EVALUATION IN AN INDOOR ENVIRONMENT

5.1 Test Site and Reference Data

As an indoor evaluation environment, we chose our institute’s offices in the FHNW campus in Muttenz/Basel. There are control points throughout the hallway as well as in rooms, whose 3D coordinates we determined using total stations. Additionally there are terrestrial laser scanning point clouds that we collected independently. As reference images we used a dataset, which we acquired using a mobile mapping backpack. This capturing system is described in detail by Blaser et al. (2018). It mainly consists of a panoramic camera PointGrey Ladybug5 and two Velodyne VLP-16 lidar profile scanners for positioning. The panoramic camera has six camera heads with a resolution of 2448 x 2048 pixels and a focal length of 4.3 millimetres each.
The reference dataset contains images from two different epochs. One part was acquired in November 2017 and is described in Cavegn et al. (2018). The second part was captured in March 2018 and used in the work of Blaser et al. (2018). In Figure 8, the positions of the reference images recorded in November 2017 are depicted. In the second campaign, we only captured images in the hallway and one laboratory.

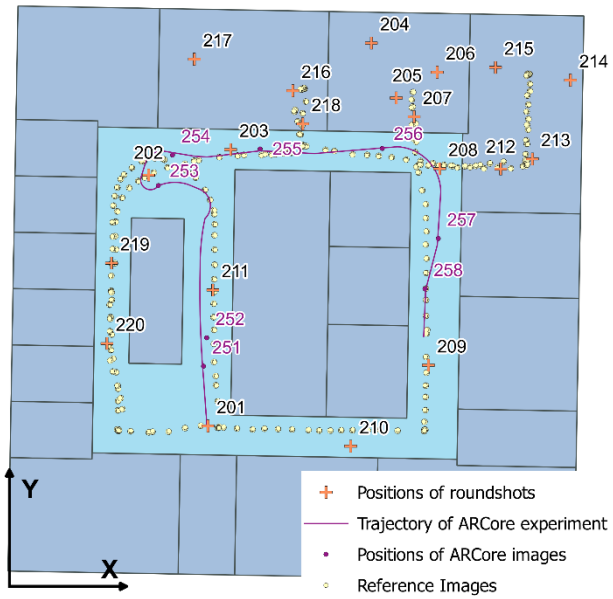

Figure 8. Map of our indoor test site showing positions and trajectories

5.2 Evaluation of Roundshot Images

Similar to the outdoor environment, we captured panoramic roundshot image series using the system described in section 3.2. We show the positions of the 20 series in Figure 8, where they are labelled with numbers from 201 to 220. Because the camera is much closer to objects in our indoor environment, we raised the intermediate angle between two consecutive images, leading to a number of around 10 to 15 images per location.
The values in Table 4 show the standard deviations of the projection centres per location. The absence of the locations 204 and 217 in this table is due to the fact that none of the images from these locations were aligned successfully. When comparing the success rates, it is apparent that (with few exceptions) the locations in the hallway reach better values. Similar to the values from the outdoor evaluation, the height is generally more accurate than the position. The value of the overall precision for the three combined dimensions amounts to 18 centimetres. However, as the results of location 203 show, it is possible to achieve a precision of 3 centimetres. The median of the reprojection error over all images is at 15.1 pixels, which is higher than in the outdoor environment. This is caused by some images with poor positioning accuracy.
As it is visible in Figure 8, locations 204 and 217 lie on the opposite side of the rooms when compared to the reference images. Hence, it would be surprising if the orientation showed good results. As expected, the differences of the viewing directions were too large to allow an alignment of the images. The results in the other rooms were as expected, too. These are in use for classes, so the furniture is constantly moved around, leading to changing appearance of the room. The locations with lower precision in the hallway are located in a very narrow part with glass showcases (locations 219 and 220) and in regions where the only distinct features are located on poster walls that can be moved around easily. In general, the low rate of oriented images is not surprising, as many of the images just show a part of a wall with very little texture on it.

Location	Rate	x [cm]	y [cm]	z [cm]	2D [cm]	3D [cm]
201	13/13	7	13	2	15	15
202	10/13	26	14	2	29	30
203	9/11	2	2	2	3	3
205	2/10	2	0	0	2	2
206	4/11	11	44	6	45	46
207	3/10	10	41	2	42	42
208	8/9	36	20	1	42	42
209	6/10	21	24	29	31	38
210	3/10	4	2	0	4	4
211	10/12	6	37	26	37	44
212	7/13	4	4	1	6	6
213	8/14	55	37	8	66	67
214	5/11	13	14	5	19	19
215	6/12	24	8	1	25	25
216	11/21	12	9	6	15	16
218	4/16	4	12	6	13	13
219	3/13	4	87	6	87	87
220	6/13	22	37	26	43	49
Total	118/224	12	13	5	18	18

Table 4. Standard deviations of projection centres per location for our indoor environment

In our indoor environment, there are many trapdoors for a robust image orientation. Besides the mentioned movable elements like furniture or poster walls, doors have proven to be tricky. In some edge cases, an opened door is depicted in an angle, that it is accidentally well aligned with the frame. Figure 9 show an example, where the result of the orientation was wrong because of this. The plotted residuals show, that the alignment is based on features around the door, since this is where the differences are the smallest.

Figure 9. Extreme case where an opened door leads to wrong results. In the reference image (left) the door is closed, in query image (right) it is open. The red lines in the query image indicate the residuals from reprojection.

5.3 Evaluation of ARCore Image Series

In the indoor environment, we evaluated our ARCore app with an experiment where we walked through the hallway and occasionally captured an image. We performed the geolocalisation using the origin method. The path walked during the experiment and the positions of the eight images are shown in Figure 8. The boxplots of reprojection errors in Figure 10 show that the accuracy in general was good. Overall, the median of the reprojection errors amounts to 3.7 pixels and the mean, which is distorted by large values in image 258, is 9.3 pixels. The large differences in image 258 are plausible, since the image shows a glass cabinet, whose contents changed as well as a door that was closed in the reference images but open in the query image. The point indicated in red in Figure 11 has a reprojection error of around 10 pixels. On the right image of Figure 11, one can see that with an accurate initial pose it is even possible to align images with few distinctive points.

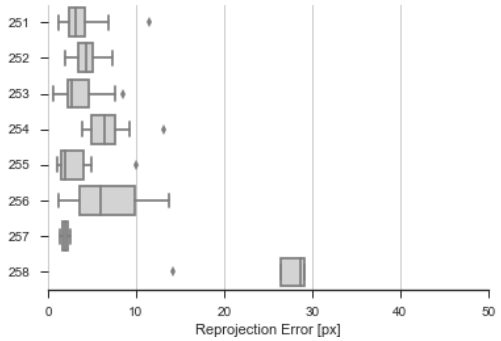

Figure 10. Boxplots of reprojection error for ARCore indoor series

Figure 11. Images 255 (left) and 257 (right) overlaid with the reprojected checkpoints. Images converted to greyscale for better visibility of points.

6. CONCLUSIONS AND OUTLOOK

In summary, our first results demonstrate the viability of accurate visual localization based on large-scale 3D image spaces – both in indoor and outdoor environments. Absolute accuracies at the decimetre level for outdoor applications and below 5 cm for indoor applications can be reached in ideal cases. However, this requires ideal conditions. Especially the up-to-dateness of the reference images has a big impact on the results. Furthermore, the accuracy gets much better, if the query image is similar to the reference images. This implies the same lighting and seasonality as well as a comparable point of view. The mobile mapping images from subsequent campaigns show an example where these conditions were met. Hence, the results are quite accurate. In our other experiments, we revealed some weaknesses of our approach, especially when there are big changes compared to the reference images. With a reference dataset containing images showing the current situation from different viewpoints, a globally uniform localisation quality can be achieved.

First experiments using the PoseNet approach of Kendall et al. (2015) showed promising results. When applying the trained model to a test dataset from the same campaign, we could achieve accuracies similar to those of Kendall et al. (2015). However, the positioning quality decreases rapidly, when the locations of query images move away from the traffic lanes, where the reference images had been captured. If we succeed to get an initial pose

solely based on the image, we get rid of the need for positioning sensors such as GNSS, and it becomes possible to initialise the image orientation even in areas with poor or no satellite visibility. Our tests with ARCore showed that the use of augmented reality tracking techniques increases the robustness over a single-image approach for visual localization. Its major weakness is the current need for control points or some other kind of marker for absolute position initialisation. These control points have to be close to the region of interest, as otherwise there is an extrapolation. When starting at some point and walking away this approach has an error propagation similar to open traverses. Hence, one has to return to the starting position to achieve a loop closure.
In our future work, we aim at combining the strengths of both approaches, i.e. the robustness of AR tracking with the capability of markerless absolute visual localization using large-scale 3D image spaces as reference.

ACKNOWLEDGEMENTS

This work was co-funded by the Swiss Innovation Agency (Innosuisse, formerly CTI) as part of the BIMAGE project (No. 18493.2 PFES-ES). The investigations on Pose Estimation using Deep Learning were 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".

REFERENCES

Apple, 2017. iOS 11 brings Powerful New Features to iPhone and iPad this Fall. https://www.apple.com/newsroom/2017/06/ios-11-brings-new-features-to-phone-and-ipad-this-fall/ (10 July 2018).

Blaser, S., Cavegn, S. & Nebiker, S., 2018. Development of a Portable High Performance Mobile Mapping System Using the Robot Operating System. In: *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, Karlsruhe, Germany, (accepted).

Brachmann, E. & Rother, C., 2018. Learning Less is More - 6D Camera Localization via 3D Surface Regression. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, pp. 4654–4662.

Brahmbhatt, S., Gu, J., Kim, K., Hays, J. & Kautz, J., 2018. Geometry-Aware Learning of Maps for Camera Localization. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, pp. 2616–2625.

Cavegn, S., Blaser, S., Nebiker, S. & Haala, N., 2018. Robust And Accurate Image-Based Georeferencing Exploiting Relative Orientation Constraints. In: *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, Vol. IV-2, pp. 57–64.

Cavegn, S., Nebiker, S. & Haala, N., 2016. A Systematic Comparison Of Direct And Image-Based Georeferencing In Challenging Urban Areas. In: *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, Vol. XLI-B1, pp. 529–536.

Clark, R., Wang, S., Markham, A., Trigoni, N. & Wen, H., 2017. VidLoc: A Deep Spatio-Temporal Model For 6-DoF Video-Clip Relocalization. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, USA, pp. 2652–2660.

Cooper, D., 2018. Google Shows Off Its Camera-Enabled Navigation System. In: *Engadget*. https://www.engadget.com/2018/05/08/g/ (13 July 2018).

Google, 2018. ARCore - Fundamental Concepts. In: *Google Dev*. https://developers.google.com/ar/discover/concepts (11 July 2018).

Gosalia, A., 2018. Announcing ARCore 1.0 and new Updates to Google Lens. In: *Google Blog*. https://blog.google/products/arcore/announcing-arcore-10-and-new-updates-google-lens/ (11 July 2018).

Karpushin, M., 2016. Local Features for RGBD Image Matching Under Viewpoint Changes. PhD Thesis, ParisTech, Paris, FR.

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

Kilgus, T., Heim, E., Haase, S., Prüfer, S., Müller, M., Seitel, A., Fangerau, M., Wiebe, T., Iszatt, J., Schlemmer, H.P., Hornegger, J., Yen, K. & Maier-Hein, L., 2015. Mobile Markerless Augmented Reality And Its Application In Forensic Medicine. In: *Int. J. Comput. Assist. Radiol. Surg.*, Vol. 10, pp. 573–586.

Lee, T &, Hollerer, T., 2008. Hybrid Feature Tracking and User Interaction for Markerless Augmented Reality. In: *IEEE Virtual Reality Conference*, Reno, USA, pp. 145–152.

Nebiker, S., Cavegn, S. & Loesch, B., 2015. Cloud-Based Geospatial 3D Image Spaces - A Powerful Urban Model for the Smart City. In: *ISPRS Int. J. Geo-Information*, Vol. 4, pp. 2267–2291.

Nistér, D. & Stewénius, H., 2006. Scalable Recognition With a Vocabulary Tree. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, New York, USA, pp. 2161–2168.

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. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, pp. 8601–8610.

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

Schönberger, J.L., Pollefeys, M., Geiger, A. & Sattler, T., 2018. Semantic Visual Localization. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, pp. 6896–6906.

Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T. & Torii, A., 2018. InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, pp. 7199–7209.

Wagner, D., Reitmayr, G., Mulloni, A., Drummond, T. & Schmalstieg, D., 2008. Pose Tracking from Natural Features on Mobile Phones. In: *IEEE/ACM International Symposium on Mixed and Augmented Reality (ISMAR)*, Cambridge, UK, pp. 125–134.

Wüest, R. & Nebiker, S., 2017. Geospatial Augmented Reality for the Interactive Exploitation of Large-Scale Walkable Orthoimage Maps in Museums. In: *International Cartographic Conference (ICC)*, Washington D.C, USA.