An Outdoor Ground Truth Evaluation Dataset for Sensor-Aided Visual Handheld Camera Localization

Daniel Kurz* metaio GmbH
Peter Georg Meier† metaio GmbH
Alexander Plopski‡ Osaka University
Gudrun Klinker§ Technische Universität München

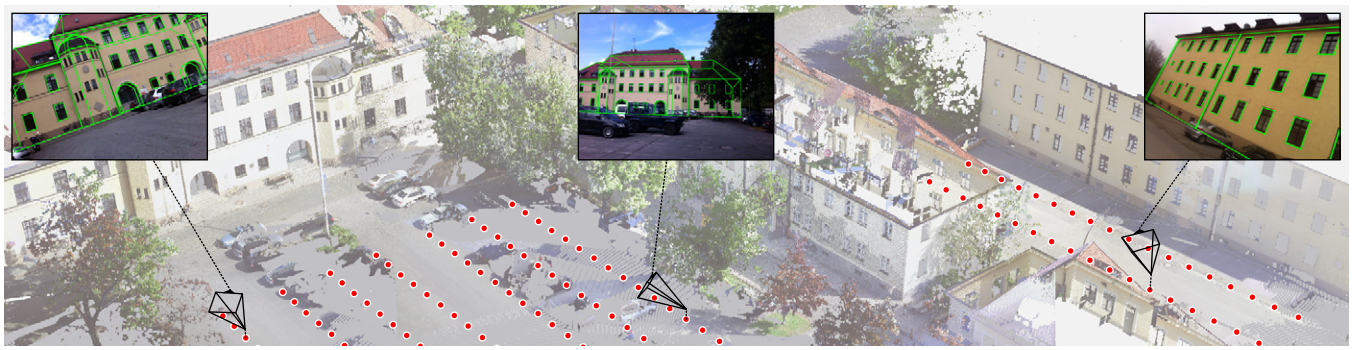

Figure 1: Our dataset comprises a precise environment model and over 45,000 camera images with sensor values and 6DoF ground truth poses.

Abstract

We introduce the first publicly available test dataset for outdoor handheld camera localization comprising over 45,000 real camera images of an urban environment captured under natural camera motions and different illumination settings. For all these images the dataset not only contains readings of the sensors attached to the camera, but also ground truth information on the geometry and texture of the environment and the full 6DoF ground truth camera pose.

This poster describes the extensive process of creating this comprehensive dataset that we have made available to the public. We hope this not only enables researchers to objectively evaluate their camera localization and tracking algorithms and frameworks on realistic data but also stimulates further research.

1 Introduction and Related Work

One of the most important challenges towards the everyday usage of video-see-through handheld Augmented Reality (AR) is precise and robust camera localization in wide area outdoor environments. Pose estimation based only on information from sensors such as GPS, compass and inertial sensors, is currently used for example in AR browsers such as junaio but the results are not accurate enough for pixel-precise registry of overlays in the camera image.

Visual localization and tracking is very well suited to provide accurate registration and is frequently used for camera tracking in desktop-sized environments. In particular simultaneous localization and mapping (SLAM) systems are commonly used that reconstruct a sparse 3D map of the environment while tracking. If the application however is to overlay landmarks with a known absolute 3D position, the camera pose needs to be estimated with respect to a known model of the environment.

*e-mail: daniel.kurz@metaio.com
†e-mail: peter.meier@metaio.com
‡e-mail: plopski@lab.ime.cmc.osaka-u.ac.jp
§e-mail: klinker@in.tum.de

Recently, methods that enable precise camera localization outdoors on mobile devices were proposed, e.g. [1]. In addition to information from the camera image, this approach exploits readings of sensors attached to the camera. A quantitative evaluation of such 6DoF localization methods under real-world conditions requires a realistic dataset with ground truth information. For a set of given input data, i.e. a camera image, intrinsic camera parameters, and sensor readings, the expected 6DoF pose is needed. However, it is a tedious task to determine the ground truth pose for real camera images – particularly outdoors.

Irschara *et al.*[2] do not have ground truth information and therefore measure the effective number of inliers to rate if a localization succeeded or not. Ventura and Höllerer [5] synthesize camera images as unwarped parts of omnidirectional images. They use as ground truth the position of the omnidirectional camera determined in the SfM process to create the reference map. Similarly [1] simulate online-created panoramic images as subsets of existing full panoramas for evaluation and manually set the ground truth camera position. There are datasets of real handheld camera images with corresponding 6DoF ground truth poses but these either only contain planar tracking templates [3] or 3D objects captured indoors [4] and without any associated sensor readings. Existing datasets for outdoor environments come from the robotics domain, e.g. [6], and consequently do not contain handheld camera motion.

2 Outdoor 6DoF Ground Truth Dataset

In the following, we describe our extensive procedure to create the first outdoor 6DoF ground truth dataset[1] comprising:

- a highly accurate, geo-referenced, and textured 3D model of an urban environment spanning about 10,000 square meters,
- video sequences containing over 45,000 individual images of the environment with realistic handheld camera motion taken from different locations with an off-the-shelf mobile phone,
- the sensor readings of GPS, compass, and the gravity vector for each image of the sequences mentioned above, and
- an accurate 6DoF ground truth pose for every camera image.

[1]The dataset is available at: http://www.metaio.com/research

Figure 2: Sequence recording at a known camera position using a lead weighted string and measured survey points on the ground.

Model Acquisition We chose an office park in Munich as testing environment, which comprises a large parking lot, different buildings, some parking lanes, and small streets. As we aim at creating a very precise and detailed model of the environment, we used a FARO Focus 3D laser scanner to create nine high-precision panoramic laser scans with texture information for different parts of the environment. The individual scans were then registered to a common coordinate system using proprietary software. The merged model has finally been geo-referenced based on the known latitudes and longitudes of building corners. The full model with color information is shown in figure 1 rendered from a bird's-eye view.

Sequence Recording at Known Camera Positions There are two important aspects to keep in mind when recording sequences for testing. These are *relevance* for the targeted application and *universality*. It is important to use a capturing device and camera motions similar to those that can be expected to be used in real applications. And it is crucial, that the dataset comprises a high variance in parameters such as the camera position, the weather, and partial occlusions by parking cars and pedestrians for the data to be considered universal and representative.

We decided to use an iPhone 4 mobile phone because it is a very common device and allowed to obtain the GPS position, compass heading and a measurement of the gravity vector for every image. The image resolution is set to (480×360) pixels and the camera's intrinsic parameters were calibrated offline. We use a custom-made application to capture image sequences with all relevant sensor readings at a framerate of $\sim$25 Hz and save them to files.

Because installing an external tracking system that would measure the 6DoF pose of a phone anywhere in the environment with a precision that can be considered ground truth is nearly impossible, we limit ourselves to a set of 156 discrete camera positions which are spread all over the area and are chosen such that placing a camera to these positions is easy to achieve. Our test environment comprises a parking lot and two parking lanes which are divided into individual cells by white markings on the ground. We use the crossings and end points of these markings as survey points and measure their precise 3D positions with a total station (Trimble 3603 DR). Figure 1 displays these survey points as red circles.

We then divide the process of obtaining sequences with 6DoF ground truth poses into two steps. In the first step we capture sequences at known 3D camera positions and recover the corresponding 3DoF orientation in a second step. Attaching a lead weighted string of known length to the phone's camera, makes it easy to precisely move the device to a known 3D position. As shown in figure 2, we hold the mobile phone directly over a survey point $\mathbf{s}_i$ on the ground with a string of known length h_i to be sure the camera is located at $\mathbf{s}_i + h_i \cdot \mathbf{z}$.

We use strings at a lengths of 1 meter and 1.8 meters which can be considered to represent the actual heights users hold their mobile devices as well. While capturing and recording sequences (some frames of two exemplary sequences are shown in figure 2), the camera only undergoes rotational movements and does not change its position. Since users usually prefer standing while using the screen of a mobile phone, we believe that our sequences have realistic kinds of camera motion for handheld AR applications.

6DoF Ground Truth Recovery For the second step of the ground truth acquisition process we prepared an edge model of the environment based on the highly precise ground truth model (laser scans). Using the coarse camera orientation obtained from the sensor readings and the accurately known 3D position of the camera, we project the edge model into the camera image and find the orientation for which the model best fits gradients in the the camera image using exhaustive search in a neighborhood around the initial orientation estimate. Finally, the recovered 3DoF camera orientation together with the 3DoF known ground truth position, make the 6DoF ground truth pose. To account for potential errors in labeling or recovery of the rotation, the ground truth poses of all images have been manually verified by rendering a wireframe model onto the video stream. Figure 1 displays the recovered 6DoF ground truth pose for three exemplary images of the dataset in green.

3 Conclusions and Future Work

We presented the first outdoor dataset comprising 125 sequences of real camera images and sensor readings captured with a mobile phone with known 6DoF ground truth poses.

The dataset can be used to evaluate any localization and tracking method including those relying on color features, edge features, or even model-based approaches that require a dense and textured reference model. As capturing additional sequences does not require any hardware except a capturing device and a lead weighted string, we plan to expand the dataset by more sequences taken with different devices, by different users, and under different weather conditions.

References

[1] C. Arth, A. Mulloni, and D. Schmalstieg. Exploiting Sensors on Mobile Phones to Improve Wide-Area Localization. In *Proc. ICPR*, 2012.

[2] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *Proc. IEEE CVPR*, 2009.

[3] S. Lieberknecht, S. Benhimane, P. Meier, and N. Navab. A dataset and evaluation methodology for template-based tracking algorithms. In *Proc. IEEE ISMAR*, 2009.

[4] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A Benchmark for the Evaluation of RGB-D SLAM Systems. In *Proc. IEEE/RSJ IROS*, 2012.

[5] J. Ventura and T. Höllerer. Wide-area scene mapping for mobile visual tracking. In *Proc. IEEE ISMAR*, 2012.

[6] O. Wulf, A. Nuchter, J. Hertzberg, and B. Wagner. Ground truth evaluation of large urban 6D SLAM. In *Proc. IEEE/RSJ IROS*, 2007.