Guided Sparse Camera Pose Estimation

Fabian Schenk[1], Ludwig Mohr[1], Matthias Rüther[1], Friedrich Fraundorfer[1], and Horst Bischof[1]

Institute for Computer Graphics and Vision
Graz University of Technology, Austria
{*schenk,mohr1,ruether,fraundorfer,bischof*}@*icg.tu-graz.ac.at*

Abstract

In this paper, we present an idea for a sparse approach to calculate camera poses from RGB images and laser distance measurements to perform subsequent facade reconstruction. The core idea is to guide the image recording process by choosing distinctive features with the laser range finder, e.g. building or window corners. From these distinctive features, we can establish correspondences between views to compute metrically accurate camera poses from just a few precise measurements. In our experiments, we achieve reasonable results in building facade reconstruction with only a fraction of features compared to standard structure from motion.

1. Introduction

Structure from motion (SfM) has been an active research area in computer vision for decades as it is of interest in a wide range of practical applications such as robotic navigation and augmented reality. Common SfM approaches exploit a huge number of feature correspondences and finding suitable starting views poses a challenge, which is not necessarily simplified by the abundance of features. Often, this is tackled by assuming a set of ordered images or incorporating additional measurements for camera pose initialization. In a subsequent step, all the views are merged into a common global coordinate system, where the whole scene structure in 3D is calculated and refined together with the camera poses. The optimization of the final scene structure is computationally demanding due to the large amount of correspondences over multiple camera views.

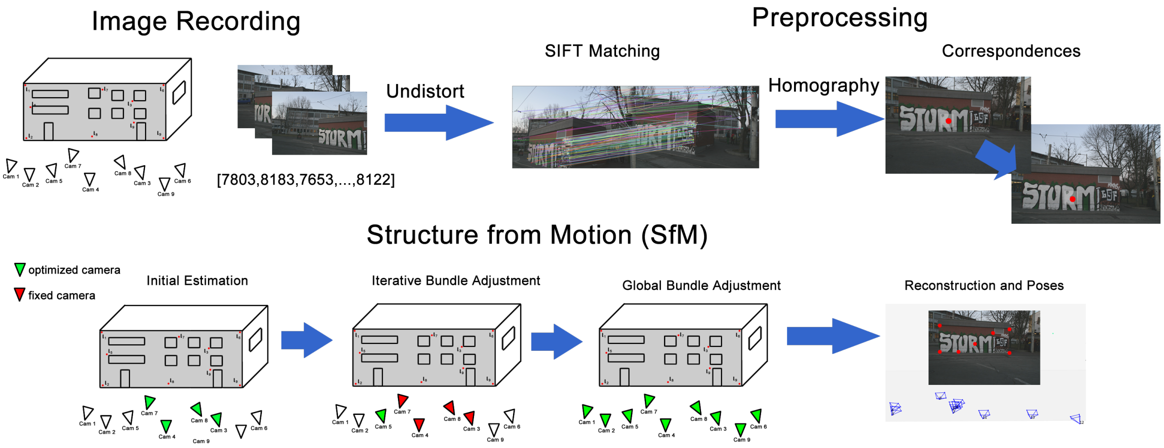

Figure 1: Our proposed sparse SfM approach, where we take RGB images and laser distance measures to estimate camera poses and a sparse point cloud.

In this work, we present an idea to estimate metrically correct camera poses with just a small number of features (see Fig. 1). Our hardware setup consists of an RGB camera and a laser range finder (LRF) (see Fig. 2 (a)). The LRF allows us to select highly distinctive features for pose estimation while at the same time obtaining their accurate distance. We focus on the reconstruction of facades, enabling us to utilize homographies instead of fundamental matrices for correspondence computation. For pose estimation, we use laser points with known distance from the camera and their respective matches in other views.

Finally, we compare our approach to the freely available SfM framework OpenMVG [10] and show that we achieve reasonable results for the camera poses with just a fraction of correspondences. This is of special interest for metric reconstruction on devices with constraints on computational power, e.g. mobile devices or UAVs.

2. Related Work

Most of the work related to the task of calibrating the extrinsic relationship of an LRF to a projective camera consider a setup with either a 2D [18, 7] or 3D LRF [14, 2]. Further, they rely on user input to establish correspondences between the laser measurements and the images taken by the camera.

We on the other hand want to solve the task of extrinsic calibration of a 1D LRF to a camera without user interaction. We require the 3D world position of the plane whose distance is measured to be inferable from the images as well as the laser point produced by the LRF to be visible within the image. In contrast to [13], where they jointly perform geometric camera and LRF calibration, we do not refine the intrinsic calibration of the camera using the LRF measurements but expect the intrinsic calibration to be done beforehand and to be of sufficient quality.

SfM algorithms for 3D reconstruction and camera pose estimation from unstructured data usually only capture the scene up to scale. In [1, 15] the authors perform large scale 3D scene reconstruction from Internet photos. Their work examines 3D modeling from unstructured data, yet the reconstruction can be only performed up to scale due to inherent lack of metric information. In [3], the authors first solve the relative motion on a local scale among just a few images, and then use these local relations as initialization for the global solution.

Methods solving the metric reconstruction problem with the SfM paradigm often rely on either an underlying structure of data (sequential image capturing, constant acquisition frame rate) in connection with registered motion estimations using GPS or inertial measurements as in [16, 3] or rely on direct geometry measurements with 3D LRFs and subsequent registration of the resulting point clouds [6, 7].

The approach presented in [12] is the one most similar to ours. However, in addition to 1D laser measurements corresponding to images of the scene, they leverage motion estimations between images through IMU data as well as interactive gestures for semantic cues aiding in reconstruction.

We propose an approach for metric camera pose estimation from unstructured images. Instead of searching for dense point correspondences among all images, we restrict ourselves to a sparse wireframe model with each image contributing just a single point (the location of the laser distance measurement). This allows us to ensure robust reconstruction by choosing distinct and easily matchable locations on the facade during data acquisition.

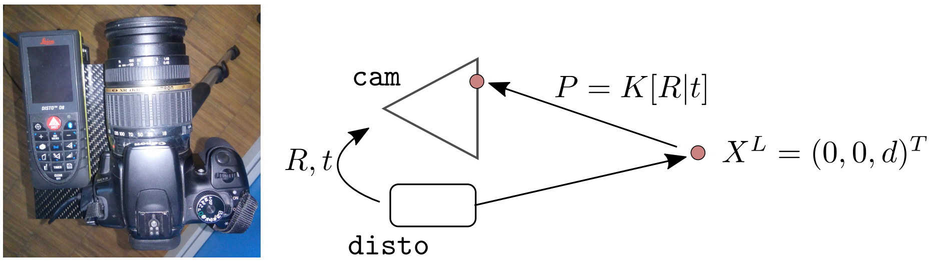

Figure 2: Left: The setup of the DSLR and laser range finder on the carbon panel. Right: The schematics of the camera and laser range finder setup and the idea behind calibration.

3. Camera Setup and Calibration

Our hardware setup consists of a standard DSLR mounted onto a rigid carbon panel next to a 1D LRF (see Fig. 2), which we control remotly via USB and Bluetooth for easy acquisition of images and distance measurements. We perform intrinsic calibration with a modified version of the Bouguet toolbox and a custom target as proposed in [4]. In the remaining part of the paper we expect the camera to be calibrated and the geometric distortions introduced by the camera and lens assembly to be removed from the images. This enables us to infer the real world line of sight $\mathbf{l}_{los}$ with respect to the center of projection of the camera of any given pixel in an image I_i as:

$$\mathbf{l}_{los,i} = \mathbf{K}^{-1} \cdot \mathbf{l}_{2D,i,i}, \tag{1}$$

where $\mathbf{K}$ denotes the camera matrix and $\mathbf{l}_{2D,i,i} = [x, y, 1]^T$ is the 2D position of the laser point i in I_i in homogeneous coordinates.

In theory, provided an intrinsic camera calibration $\mathbf{K}$ and a known plane in 3D, the rotation $\mathbf{R}_{LRF}$ and translation $\mathbf{t}_{LRF}$ of the laser range finder relative to the camera can be estimated using two measurements only. Yet, in order to obtain a more robust estimation, we take several measurements $M = \{d_i, I_i\}_1^N$ of distances d_i with corresponding images I_i. Since the application is 3D facade reconstruction and we expect the facade measurements to be taken nearly fronto-parallel to the image sensor, we restrict the extrinsic calibration sequence to a fronto-parallel movement of the target relative to the camera, ensuring that the position at which the LRF takes its measurement is well within the calibration target.

In a first step, we detect the target position and orientation in 3D relative to the camera's center of projection, as well as the target position in 2D within the image. We detect the laser point as brightest object on the calibration target using adaptive thresholding and then take its center of mass as the 2D position $\mathbf{l}_{2D,i,i} = [x_i, y_i, 1]^T$ of the laser point i in I_i. We then calculate the position $\mathbf{l}_{3D,i}$ in 3D of a laser point by intersecting the line of sight $\mathbf{l}_{los,i}$, on which the laser point lies, with the target plane.

When holding the camera positions fixed and moving the calibration target plane relative to it, all points $\mathbf{l}_{3D,i}$, $i \in \{1, \ldots, N\}$ lie on a straight line. This line corresponds to the viewing direction $\mathbf{l}_{LRF}$ of the LRF, which we calculate by fitting a line to the measurements using singular value decomposition on the 3D points stacked to a matrix $\mathbf{L}_{3D} = [\mathbf{l}_{3D,1}, \cdots, \mathbf{l}_{3D,N}]$. The right-singular vectors obtained by SVD correspond to the orthogonal directions of maximum variance within the data. Thus, the right-singular vector corresponding to the largest singular value of $\mathbf{L}_{3D}$ coincides with the viewing direction of the LRF, provided that the uncertainty of the estimation of the LRF's origin is sufficiently smaller than the relative movement of the calibration target.

We obtain several noisy estimates for the position $\mathbf{t}_{LRF}$ of the LRF through the correspondence

$$\mathbf{t}_{LRF,i} = \mathbf{l}_{3D,i} - d_i \cdot \mathbf{l}_{LRF}, \tag{2}$$

where $\mathbf{l}_{LRF}$ has been normalized to unit length. We obtain the final estimate for the position $\mathbf{t}_{LRF}$ of the LRF by taking the median of all noisy estimates. The rotation $\mathbf{R}_{LRF}$ is given by the angle between the viewing direction $\mathbf{l}_{LRF}$ of the laser range finder and the cameras optical axis in the plane spanned by the optical axis of the camera and $\mathbf{l}_{LRF}$.

4. Sparse Pose Estimation and 3D Scene Reconstruction

The proposed approach is structured in steps typical to SfM pipelines: image recording, preprocessing, relative pose and motion estimation between views and ultimately 3D reconstruction. Since it is aimed at the reconstruction of building facades, which can to a large extent be modeled as a set of flat surfaces, it is sufficient to reconstruct the building as a wire-frame model using surface vertices together with a few supporting points on the walls. We compute SIFT matches to estimate homographies between image pairs (I_i, I_j), which can be used to establish correspondences based on the known laser point $\mathbf{l}_{3D,i}$ in I_i and its respective 2D position $\mathbf{l}_{2D,i,j}$ in I_j.

Using an initial set of 4 images with full correspondences and the laser measurements, we are able to initialize and calculate an early estimate for our model and the relative camera poses. Then we iteratively add the remaining cameras and distance measurements and finally refine the poses with a global bundle adjustment. Since we know the respective distance information to each camera pose, this estimation is accurate in its scale.

4.1. Image Recording

Since we perform sparse camera pose estimation and reconstruction, the accuracy of the solution depends upon a few, yet highly significant features which are easily found in images taken from different perspectives. For a good reconstruction, the significant features should be chosen in a way such that they lie on the facade and are well-distributed over its surface including the corner points, e.g. vertices of walls and corners of windows. Figure 1 depicts the data recording process, where we take RGB images from various view points while measuring the distance of a single point in the respective image with the LRF.

4.2. Preprocessing and Feature Extraction

To keep our approach as flexible as possible and to reduce the complexity during manual data acquisition, we assume no particular order of the images. Initially all possible image pairs are added to a working set. We extract SIFT features [9] from gray-scale versions of the images and establish point correspondences using a FLANN-based matcher [11] followed by Lowe's ratio test to filter outliers. With these correspondences, we robustly estimate a homography between each image pair using RANSAC [5] with a threshold of 1 px. We only want to keep image pairs with a certain overlap in the working set, thus we filter out all with less than $n = 10$ inliers according to the RANSAC estimate and a ratio of inliers to number of matches of $< 50\%$. As a measure for the quality of the remaining image correspondences, we define an error $E_{i,j}$ for an image pair (I_i, I_j) using the 2D positions $P_{SIFT,i}$ and $P_{SIFT,j}$ of their matched features as follows:

$$E_{i,j} = mean(||P_{SIFT,i} - P_{SIFT,j}||_2), \forall i, j \in N, i \neq j. \tag{3}$$

The idea is that the Euclidean distance between SIFT matches of an image pair taken spatially closer together is lower than when taken from positions farther apart. E is used to find the images to initialize the algorithm and to find subsequent images to iteratively extend the model.

4.3. Laser Point Correspondence Computation

With the image pairs (I_i, I_j) remaining in the working set (see Sec 4.2.), we establish correspondences for the measurements of the LRF. We then compute the 2D laser point $\mathbf{l}_{2D,i,i}$ by projecting $\mathbf{l}_{3D,i}$ into its respective image using the extrinsic calibration (see Sec. 3.). As we typically deal with planar structures like facades, we estimate a homography and transform $\mathbf{l}_{2D,i,i}$ into image I_j to get $\mathbf{l}_{2D,i,j}$. This approach proves to be fairly robust in our experiments, however due to the highly repetitive nature of many facades, false positives still pose challenge.

4.4. Structure from Motion (SfM)

Our structure from motion (SfM) approach consists of three successive steps: finding an initial set for model initialization (i), iteratively adding one image at a time (ii) and one final bundle adjustment over all pairs (iii).

Model Initialization

As for all iterative SfM systems, finding a good set of starting images is challenging. Due to the many available features in common approaches, only one image pair is necessary to initialize camera pose estimations. In contrast, our approach needs a larger initial set to account for its sparse nature. Each camera has 6 degrees of freedom (3 for rotation, 3 for translation), hence we need at least 6 equations to estimate its pose. In the previous step we obtained for each image pair (I_i, I_j) two 3D-2D correspondences $\mathbf{l}_{3D,i} \Leftrightarrow \mathbf{l}_{2D,i,j}$ and $\mathbf{l}_{3D,j} \Leftrightarrow \mathbf{l}_{2D,j,i}$, i.e. 4 equations for each given image pair. For a set of at least 4 images, the resulting equation system is solvable with 6 different image pairs resulting in 4 equations each. The initial set I_{init} is chosen as the set of 4 images with the smallest sum of mutual errors $E_{i,j}$.

We solve the task of finding relative rotations $\mathbf{R_i}$ and translations $\mathbf{t_i}$ in 3D for each camera by minimizing the reprojection error $C(\cdot)$ of a laser measurement $\mathbf{l}_{3D,i}$ and its 2D correspondences $\mathbf{l}_{2D,i,j}$. with bundle adjustment. The reprojection error is defined as:

$$C(I_{init}) = min||\pi(\mathbf{R}_j(\mathbf{R}_i^{-1}\mathbf{l}_{3D,i} - \mathbf{t_i}) + \mathbf{t_j}) - \mathbf{l}_{2D,i,j}||_2^2, \forall i,j \in I_{init}, i \neq j, \tag{4}$$

with $\mathbf{R}$ and $\mathbf{t}$ the rotation and translation of each respective view and $\pi(\cdot)$ the projection. This formulation first projects a laser measurement $\mathbf{l}_{3D,i}$ in 3D from its respective camera coordinate system i into a common world coordinate system and subsequently reprojects it to the camera coordinate system j. We solve the minimization problem of bundle adjustment with a Levenberg-Marquardt [17] least-squares solver. We denote the resulting set of rotations and translations of all camera views currently involved as our current model M_{curr}.

Iterative Bundle Adjustment

In the next step, we extend our model M_{curr} by adding a new image I_k from the pool of candidates. We find I_k by summing up the error $E_{curr,k}$ of all possible image pairs (I_{curr}, I_k) and take the one with the most correspondences and the lowest overall error. We also set the initial rotation $\mathbf{R}_k$ and translation $\mathbf{t}_k$ of the newly added camera equal to the parameters of the closest camera, i.e. the one with the lowest $E_{i,k}, i \in M_{curr}$. For each image pair, i. e. 3D-2D correspondence, we get two independent equations for the x- and y-position, thus we need at least 3 correspondences to solve the 6 degrees of freedom given by $\mathbf{R}_k$ and $\mathbf{t}_k$. As mentioned in Section 4.3., outliers (wrong matches) are possible, thus in practice we use at least 4 correspondences.

In a first step, we compute the reprojection error of all correspondences C (4) using the initialized camera pose and the mean reprojection error $\bar{C}$. Then we take all correspondences with an error smaller than $1.5\bar{C}$ or a threshold ϵ_C. We then optimize with the same cost function (4) as for the initial bundle adjustment with the major difference that only the pose of the newly added camera k is optimized, while the rest of the system M_{curr} is fixed. After optimization, we again filter bad correspondences with the same approach as described above and perform a second optimization, which is usually very fast due to the already good estimation. We iterate through all images until no more can be added, i.e. do not fulfill any of the conditions.

Global Bundle Adjustment
While keeping the camera system M_{curr} fixed and only optimizing the new camera k is very fast and gives an estimate of the model structure, it does not replace a global optimization approach. We perform a final global bundle adjustment step, where all the camera poses are optimized. In this case, we take all the correspondences used during the iterative bundle adjustment step and initialize the camera poses with the previously computed rotation and translation. Here, we also use the cost function presented in (4). Similar to the iterative bundle adjustment, we again filter outliers with the reprojection error after optimization, but instead impose that the error must be smaller than $< 1.2\bar{C}$. After this final filtering step, we perform one last global bundle adjustment.

5. Results and Discussion

In this section, we present early results of our guided sparse camera pose estimation. We evaluate our approach on two datasets from different buildings, one with a well-textured facade and one with redundant structures. As reference, we use the open-source SfM framework OpenMVG, which can achieve an accuracy of around 1 cm in ideal cases [10]. It utilizes SIFT features and many correspondences to estimate camera poses and a point cloud. OpenMVG chooses the starting views randomly and in our evaluation we had to start SfM multiple times to get a reconstruction. We evaluate the distance between the camera centers generated by the two approaches. OpenMVG estimates the reconstruction up to scale, thus to metrically measure the distance between cameras, we transform its world coordinate system to ours using a robust similarity transform.

Figure 3 shows a histogram, where the bins show the number of distances between camera centers in the respective range in cm. Cameras in the first few bins are closer to OpenMVG, while the cameras in the last bin are farther away. Especially in the first experiment we achieve reasonable results and that with only 90 correspondences compared to OpenMVG’s 2969, which is a reduction by a factor of 30. In the second experiment, we only use 56 correspondences compared to 1880 in the reference. The histograms show that we are centimeters away from OpenMVGs reconstruction even though we still achieve visually appealing results when reprojecting the laser points into the images (see Fig. 4). Due to the sparse correspondences, even one unfiltered outlier can decrease the final result significantly.

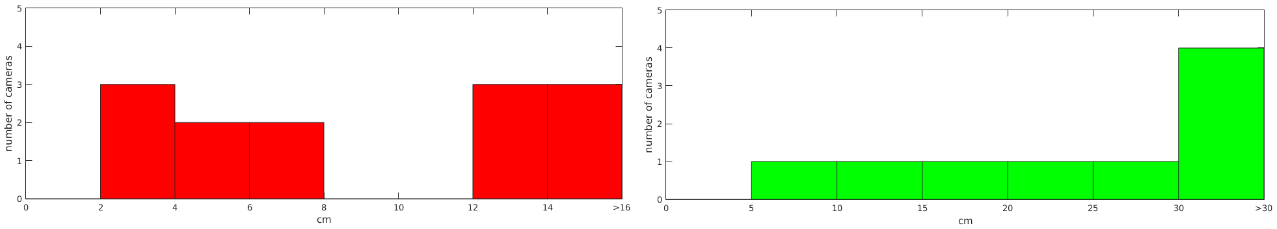

Figure 3: The two experiments and the distances between our camera positions and OpenMVG's. The bins of the histogram show the number of distances between camera centers in the respective range [cm].

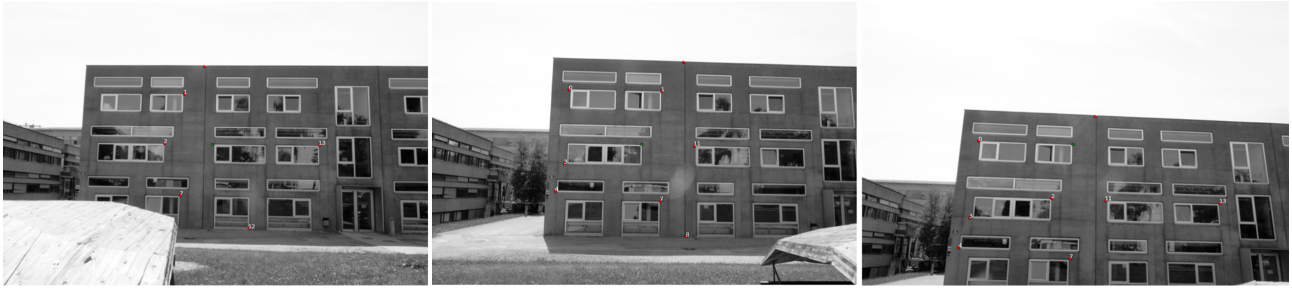

Figure 4: The reprojected laser point from the computed camera poses.

6. Conclusion and Outlook

In this paper, we presented a first evaluation of our idea to utilize a combination of RGB camera and LRF for sparse camera pose estimation, where we can select significant features during the image recording process with the LRF. We showed that metrically accurate camera pose estimation with just a few correspondences is possible. In future work, we plan a more sophisticated way to redetect significant features in the other views without the use of SIFT matches and homography estimation, which would also enable the reconstruction of more complex 3D structures. Additionally, we plan to improve the accuracy by robustifying our approach against outliers during bundle adjustment.

An interesting direction for future work is inspired by Li et al. [8], where they address the SfM problem by estimating up-to-scale edge lengths of a rigid graph constructed from 3D features and their respective image rays. We would like to investigate whether the laser range measurements can be directly used as edge lengths for this approach.

References

[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building rome in a day. In *ICCV*, pages 72–79, Sept 2009.

[2] H. Aliakbarpour, P. Nunez, J. Prado, K. Khoshhal, and J. Dias. An efficient algorithm for extrinsic calibration between a 3d laser range finder and a stereo camera for surveillance. In *ICAR*, pages 1–6, June 2009.

[3] H. Aliakbarpour, K. Palaniappan, and G. Seetharaman. Fast structure from motion for sequential and wide area motion imagery. In *Computer Vision Workshop (ICCVW)*, pages 1086–1093, Dec 2015.

[4] David Ferstl, Christian Reinbacher, Gernot Riegler, Matthias Rüther, and Horst Bischof. Learning depth calibration of time-of-flight cameras. In *BMVC*, pages 102.1–102.12. BMVA Press, September 2015.

[5] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[6] Christian Früh and Avideh Zakhor. An automated method for large-scale, ground-based city model acquisition. *IJCV*, 60(1):5–24, 2004.

[7] Ji Hoon Joung, Kwang Ho An, Jung Won Kang, Myung Jin Chung, and Wonpil Yu. 3d environment reconstruction using modified color icp algorithm by fusion of a camera and a 3d laser range finder. In *IROS*, pages 3082–3088. IEEE, 2009.

[8] Hongdong Li. Multi-view structure computation without explicitly estimating motion. In *CVPR*, pages 2777–2784. IEEE, 2010.

[9] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[10] P. Moulon, P. Monasse, and R. Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *ICCV*, pages 3248–3255, Dec 2013.

[11] Marius Muja and David G Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1)*, 2:331–340, 2009.

[12] Thanh Nguyen, Raphael Grasset, Dieter Schmalstieg, and Gerhard Reitmayr. Interactive syntactic modeling with a single-point laser range finder and camera. In *ISMAR*, pages 107–116. IEEE, 2013.

[13] Thanh Nguyen and Gerhard Reitmayr. Calibrating setups with a single-point laser range finder and a camera. In *IROS*, pages 1801–1806. IEEE, 2013.

[14] D. Scaramuzza, A. Harati, and R. Siegwart. Extrinsic self calibration of a camera and a 3d laser range finder from natural scenes. In *IROS*, pages 4164–4169, Oct 2007.

[15] Noah Snavely, Steven M Seitz, and Richard Szeliski. Modeling the world from internet photo collections. *IJCV*, 80(2):189–210, 2008.

[16] P. Tanskanen, K. Kolev, L. Meier, F. Camposeco, O. Saurer, and M. Pollefeys. Live metric 3d reconstruction on mobile phones. In *ICCV*, pages 65–72, Dec 2013.

[17] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment-a modern synthesis. In *Vision algorithms: theory and practice*, pages 298–372. Springer, 1999.

[18] Qilong Zhang and Robert Pless. Extrinsic calibration of a camera and laser range finder (improves camera calibration). In *IROS*, volume 3, pages 2301–2306. IEEE, 2004.