Orientation of Image Sequences in a Point-based Environment Model

Jan Böhm
Institute for Photogrammetry, Universität Stuttgart
jan.boehm@ifp.uni-stuttgart.de

Abstract

In this paper we propose a method to determine the exterior orientation of each frame of an intensity image sequence using prior knowledge of the scene stored in a point-based environment model (PEM). The orientation is performed by tracking landmarks across the image sequence acquired with a calibrated camera. The landmarks are intensity features, which are automatically extracted from the PEM. The PEM can easily be acquired by long range 3D sensors, such as terrestrial laser scanners. The orientation procedure of the imaging sensor is solely based on spatial resection.

1. Introduction

The problem of image orientation consists of the determination of the rotation and translation of the image frame with respect to some external coordinate frame. Image orientation from image sequences is a classical problem in photogrammetry and computer vision. Applications are widespread and can range from pedestrian self-localization, autonomous robotics and augmented reality to object recognition under camera motion. Many approaches addressing the problem have been proposed. Some use direct orientation sensor such as inertial measurement units (IMUs) and GPS systems to determine orientation and position, others use a combination of a variety of sensors, such as ranging, imaging and orientation sensors.

We propose a point-based environment model (PEM) to represent the absolute coordinate frame and to store the prior knowledge of the scene. A PEM is a dense point-wise sampling of the surface of the objects in a scene. Each sample consists of the three-dimensional coordinate of the location of the point and an associated intensity value. The novelty of the approach is the fact that we base the approach on a PEM that has been acquired from a seperate sensor system, than we use for the actual tracking.

The motivation for this approach is the expectation, that dense point clouds of large building complexes, industrial facilities and urban areas will become widely available during the next few years. The main enabling factor is the recent wide spread availability of reliable sensor systems and service companies. The main drive behind the acquisition of such data is from the computer-aided facility management (CAFM) industry, preservation authorities and safety agencies. Once this data has been acquired it can serve multiple purposes, our proposed application being only one in many. It is not the intention to acquire the point cloud specifically for the purpose of intensity image orientation. Other approaches (see below) would be more practical in that case.

The PEM provides landmarks which are used for intensity image orientation. The landmarks are feature points which are tracked across the image sequence. The core of the orientation procedure is based on ideas from structure from motion algorithms, tracking algorithms and generally visual navigation. Therefore we want to briefly introduce the related work in the following section. Section 3 details the acquisition and pre-processing of our PEM. The calibration of the imaging sensor is described in section 4. Section 5 describes our adoption of intensity image feature tracking and subsequent image orientation.

2. Related Work

Image orientation in intensity image sequences is a well-studied problem in computer vision. The solution to the problem is mainly associated with structure from motion (SfM) combined with feature tracking and SLAM [1, 2, 3, 4]. SfM assumes no prior knowledge of the scene. Therefore, no landmarks are available to represent an absolute reference frame. Instead, arbitrary (but well suited) feature points are tracked across the image sequence. The orientation procedure is based on relative orientation, also referred to as essential matrix computation in computer vision, when the interior orientation of the camera is known, i.e. the camera is calibrated. For an uncalibrated camera, the orientation can be performed by fundamental matrix computation. The SfM approach can also be used to reconstruct the scene. This can be done by forward intersection of tracked feature points using the recovered orientation. Obviously, the precision of the scene structure relies on the precision of the recovered orientation. Since the approach includes no prior information, the orientation and likewise the reconstructed scene are computed in an arbitrary coordinate frame, typically relative to the first frame of the sequence.

The orientation of single intensity images using a depth map has been studied for example in [5]. The process is formulated as a registration process. It is mainly useful for registering texture images which were acquired from a separate camera. The approach uses edge features extracted from both the intensity image and the range data to establish correspondences.

The case of image orientation of intensity image sequences when prior knowledge is available in form of range data has only recently received attention. In [6] the authors proposed an approach which basically combines SfM with the iterative closest point (ICP) algorithm. Initially images are oriented using normal SfM. Then the point cloud of the reconstructed scene is matched against the given range data using ICP. The approach has been demonstrated on a sequence of aerial images, where additional orientation sensors were included in the process.

In [7] the authors have demonstrated a different approach, where the prior knowledge is available in form of a digital elevation map (DEM). They use the DEM to reduce the complexity of the orientation computation, basically by cutting off image rays, where they intersect the DEM. This avoids explicit reconstruction of the scene. Again the approach has been demonstrated on an aerial image sequence where additional sensor information on the camera motion is available.

It is interesting to observe that [6] attempts to reconstruct the scene from the image sequence even though the three-dimensional information is already available. The correspondence of range data to image data is performed in three-dimensional space after or during reconstruction. Consequently the orientation procedure relies on some form of relative orientation. In contrast to this our proposed method does not attempt to reconstruct the scene. It solely relies on the PEM to represent the scene. Also we do not use any form of relative orientation procedure, but solely rely on absolute orientation. In addition to reduced complexity, this has the advantage that the image is directly oriented within the absolute coordinate frame and there is no propagation of errors along a chain of relative orientations.

3. Point-based Environment Model

The PEM and the landmarks extracted from it, serve as a navigational frame for subsequent image orientation. The PEM mainly consists of a dense point cloud with associated intensity values. Many approaches for the acquisition of dense point clouds are known. Some are triangulation-based, either active or passive. Others are based on the time-of-flight (TOF) principle. The advantage of using a proven TOF scanner is that the points can be determined with great accuracy and reliability. The rigid geometry of the point cloud is the key to providing accurate control information for camera orientation.

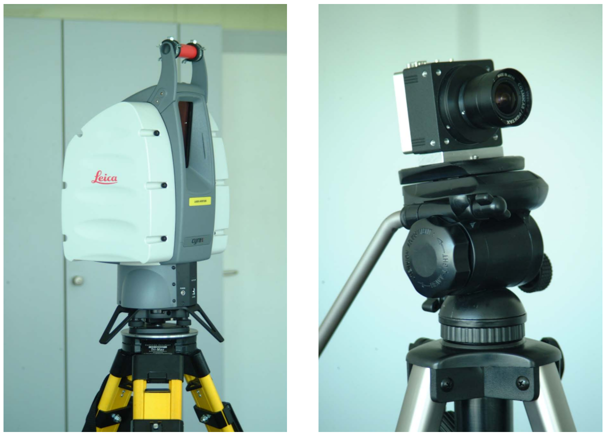

Figure 1. The sensors utilized for this study are a laser scanner and a machine vision camera. The laser scanner to the left is a Leica HDS 3000. The camera to the right is a Basler A302f. Both devices were mounted on a tripod during data acquisition.

3.1. Data acquisition

A Leica HDS 3000 was used to perform the laser scanner measurements. The scanner is shown in figure 1 on the left. The HDS 3000 is a pulsed laser scanner operating at a wavelength of 532 nm. Distance is measured by TOF measurement along a laser beam. The beam is deflected using a mirror about two axes. The resulting polar coordinates are converted to Cartesian coordinates centered at intersection point of the scanners horizontal and vertical axis. The scanner is able to acquire a scene with a field of view of up to 360° horizontal and 270° vertical in a single scan. The typical standoff distance is 50 to 100 meters, but measurements from as close as 1 meter are also possible. The manufacturer specifies the accuracy of a single point measurement with 6 mm. But when averaging over surfaces, the accuracy on modeled surfaces is about 2 mm.

For testing purposes the experiments were set up in an office environment. The scanner is placed in the middle of the room. A single scan captures the full room, with little occlusions. The resolution on the surfaces was chosen to 5

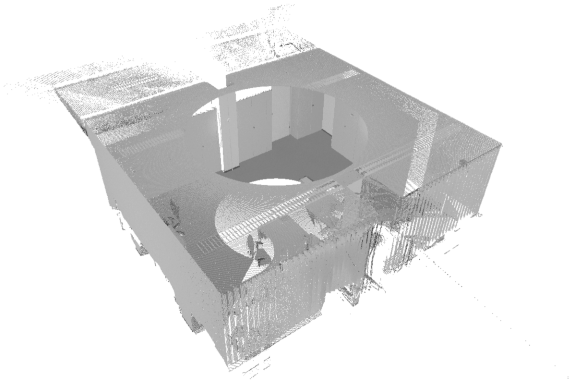

Figure 2. The point cloud data acquired by our laser scanner. The figure shows an overview of the full scan of an office.

mm on average. The point cloud consists of over 1.5 million points, which were acquired in about 15 minutes. In addition to the x, y and z coordinates, the scanner also records the intensity of the reflected laser pulse. An overview of the collected point cloud can be seen in figure 2.

Since the scanner acquires the points in an almost regular raster in the two angles of deflection, the recorded intensity values can be interpreted as an intensity image. However we shall note that the acquisition of this intensity image is fundamentally different from an image acquired by a camera. The intensity values are recorded as the intensity of the reflected beam, which locally illuminates the surface at a very narrow bandwidth of the laser beam around 532 nm. In contrast a camera depends on an external light source which is usually placed at an offset to the camera, e.g. natural sun light or room light. Therefore shadowing frequently occurs in camera images, which can lead to unstable feature points. It is advantageous that this can not occur for the PEM data. However we might miss some good feature points, which are not visible at the narrow bandwidth of the light source.

3.2. Feature extraction

If we disregard these differences we can use any standard image processing algorithm on the intensity values recorded by the laser scanner. In order to detect prominent points in the scene we use the well known Harris-Stephens corner response function [8]. The function is given by $R = \det M - k(\text{trace}M)^2$, where the matrix $M(a, b)$ for an image I is given in table 1 using

$$w(u, v) = e^{-(u^2+v^2)/(2\sigma^2)}$$

The experimentally derived value for k of 0.04 is also adopted.

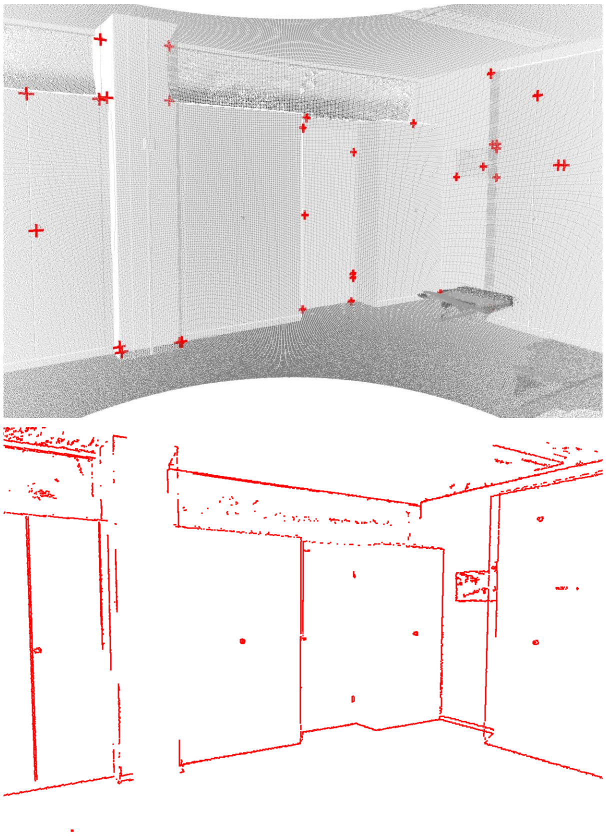

Figure 3. The processed point cloud. The left figure shows a part of the scan with automatically extracted feature points. The right figure shows a reduction of the point cloud which contains only edge points.

To extract individual points, which can be sufficiently differentiated from their neighbors, we use a non-maximum suppression scheme, where only points are selected which have a corner response value larger than any of their neighbors. For the points selected by their intensity values the corresponding x, y and z values are extracted as well and thus full three-dimensional feature points are obtained. The results of this processing is shown in figure 3 on the left, where the extracted three-dimensional feature points are highlighted by crosses in the point cloud.

For display purposes we need to compute a reduced point cloud which still sufficiently represents the scene. We use a simple differentiation filter on the intensity values and perform local thresholding to detect edge points in a fashion similar to that for feature points. The results are shown in figure 3 on the right. A simple three-dimensional edge model is obtained. However, it shall be noted that the edges extracted only consist of a collection of points, similar to

$$M(a,b) = \left[\begin{array}{cc} \sum_{(u,v)} w(u,v) \left(\frac{\partial I(a+u,b+v)}{\partial x} \right)^2 & \sum_{(u,v)} w(u,v) \frac{\partial I(a+u,b+v)}{\partial x} \frac{\partial I(a+u,b+v)}{\partial y} \\ \sum_{(u,v)} w(u,v) \frac{\partial I(a+u,b+v)}{\partial x} \frac{\partial I(a+u,b+v)}{\partial y} & \sum_{(u,v)} w(u,v) \left(\frac{\partial I(a+u,b+v)}{\partial y} \right)^2 \end{array} \right]$$

Table 1. Matrix for Harris-Stephens corner response function.

edge pixels in an image and are not edges in a CAD sense.

4. Camera Calibration

We use the collinearity equations, well known from photogrammetry, to describe the relation of object point coordinates to image coordinates. Let c be the principal distance of the camera, (X_0, Y_0, Z_0) the coordinates of the projection center in object space, (X, Y, Z) the coordinates of the object point, (x_0, y_0) the principal point and (x, y) the coordinates of the corresponding image point. Then the collinearity equations [9] are

$$\begin{array}{l} x = x_0 - c\frac{(X-X_c)r_{11}+(Y-Y_c)r_{12}+(Y-Y_c)r_{13}}{(X-X_c)r_{31}+(Y-Y_c)r_{32}+(Y-Y_c)r_{33}} \\ y = y_0 - c\frac{(X-X_c)r_{21}+(Y-Y_c)r_{22}+(Y-Y_c)r_{23}}{(X-X_c)r_{31}+(Y-Y_c)r_{32}+(Y-Y_c)r_{33}} \end{array} \quad (1)$$

where r_{ij} are the elements of a rotation matrix.

While this basic pin-hole camera describes the geometric relations in an ideal case, additional parameters are used for a more complete description of the imaging device. The following parameters follow the physically motivated approach of D. C. Brown [10] in a variation for digital cameras presented by C. S. Fraser [11]. Three parameters K_o, K_1 and K_1 are used to describe the radial distortion, also known as pin cushion distortion. Two parameters P_1 and P_2 describe the descentering distortions. Two parameters b_1 and b_2 describe a difference in scale in-between the x- and y-axis of the sensor and shearing. To obtain the corrected image coordinates x, y the parameters are applied to the distorted image coordinates x', y' using the formulas in table 2. The variable $r = \sqrt{\overline{x}^2 + \overline{y}^2}$ denotes the radial distance from the principal point.

The camera's parameters are determined in a bundle adjustment using a planar test field. The camera is calibrated beforehand and the resulting parameters are stored. The corrections are then directly applied to the images resulting in distortion-free images. The advantage of this procedure is that we can use the simpler pinhole camera model for the remaining computations and we can directly superimpose edge points. Figure 4 shows the initial frame of a sequence acquired with a Basler A302f. The camera is able to acquire up to 30 frames at a resolution of 780x582 pixels. It is equipped with a 4.8 mm wide-angle lens. The result of camera calibration is shown on the left.

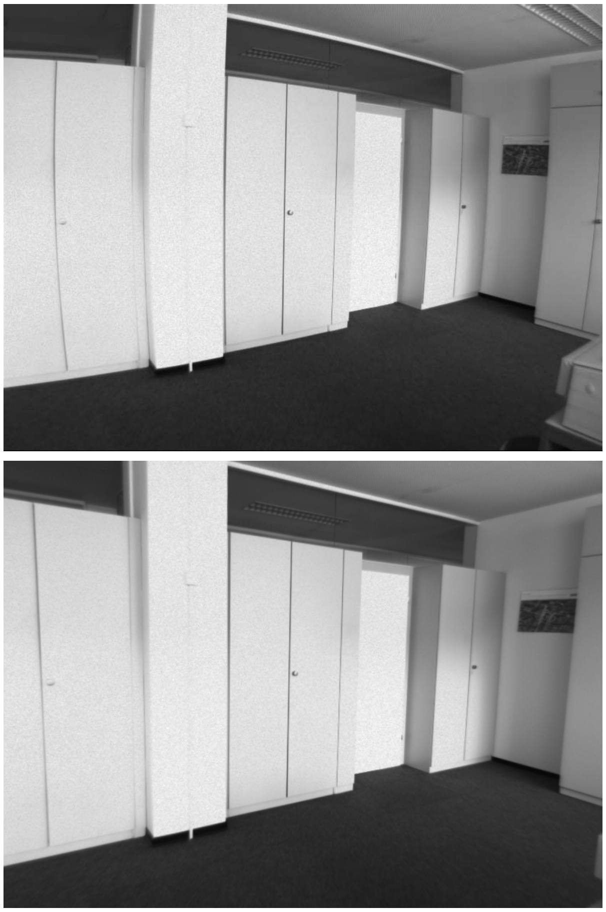

Figure 4. The lens distortion parameters and the coordinates of the principal point obtained during photogrammetric calibration are directly applied to the images. The left image shows an original frame from the sequence. The effects of lens distortions are particularly visible on the straight edges of the wall unit. The right image shows the processed image free of lens distortions.

$$\begin{array}{l}
\overline{x} = x' - x_0 \\
\overline{y} = y' - y_0 \\
\triangle x = \overline{x} r^2 K_1 + \overline{x} r^4 K_2 + \overline{x} r^6 K_3 + (2\overline{x}^2 + r^2) P_1 + 2 P_2 \overline{xy} + b_1 \overline{x} + b_2 \overline{y} \\
\triangle y = \overline{y} r^2 K_1 + \overline{y} r^4 K_2 + \overline{y} r^6 K_3 + 2 P_1 \overline{xy} + (2\overline{y}^2 + r^2) P_2 \\
x = \overline{x} + \triangle x \\
y = \overline{y} + \triangle y
\end{array}$$

Table 2. Formulas for lens distortions.

5. Feature Tracking and Image Orientation

As we have stated in the introduction the image orientation is based on the tracking of intensity features. For the design of a tracking algorithm several key components can be identified [12]. Among them are the feature extraction algorithm, the motion model, the image matching algorithm and a template similarity measurement.

5.1. Tracking Strategy

Within the proposed framework we do not select feature points from the intensity information given in the image. Rather we rely on the landmarks projected into the image. The algorithm randomly selects a given number of key points from all landmarks whose projection falls within the current frame. The number of points is a trade-off in-between reliability and processing speed. For the experiments presented within this paper we chose the number of key points to be tracked to ten.

We use a simple sum of absolute differences operator to perform the matching of local templates across a sequence of images. This is a very fast operation and worked sufficiently well in our tests. We use a linear motion model to detect outliers in the matching process. We adopt the random sample consensus (RANSAC) [13] strategy for outlier detection. We randomly select single key points and check on the consensus of their planar motion to that of all other key points. When outliers are detected these key points have to be replaced to guarantee that the number of key points will not degenerate. The same is true when a key point moves outside the image boundaries. A key point is replaced by randomly selecting any other landmark whose projection falls within the image boundaries. For any landmark that has been tracked the image template is stored for later re-use, shall the point be tracked again. This adds additional stability to feature tracking.

Figure 5 shows the initial frame introduced above with the projected landmarks marked by squared boxes. To the bottom we see the template patches extracted at these locations.

Looking at the templates we see, that they are not neces-

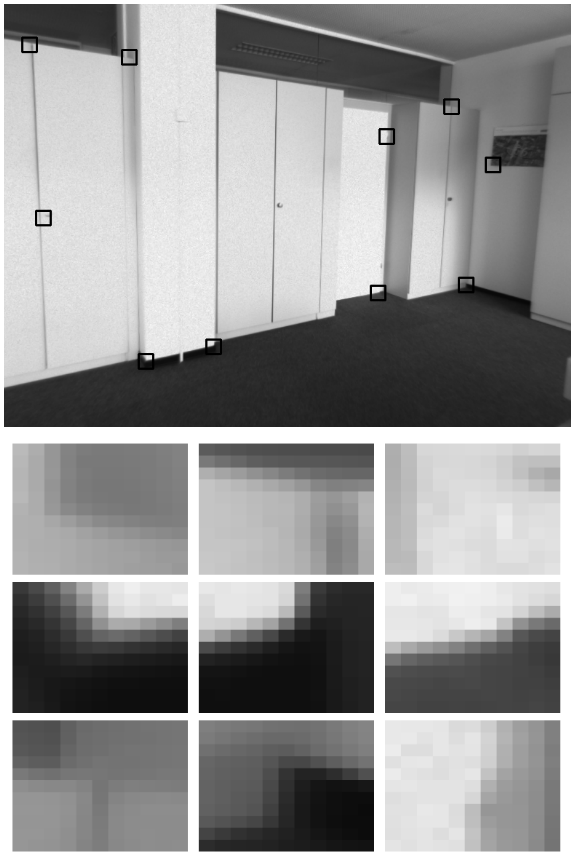

Figure 5. From the available feature points a fixed number of points are randomly selected for tracking, in this case ten are selected. The top figure shows the selected areas. The bottom figure is a composition of the image templates cut out in these areas.

sarily optimal in the sense of Shi and Tomasi [12]. This is a penalty we pay for the fact that our approach does not extract feature points based on the information of the given image. Rather we rely on the assumption that a three-dimensional landmark identified based on the intensity information of the laser scanner also is a suitable two-dimensional feature point in the image. This assumption is not always valid. For one this is due to the principal differences in the image formation process already discussed above. A further reason is the perspective discrepancy created by the offset in-between the laser scanner station and the camera station, which is unavoidable for a freely moving camera. Nevertheless figure 6 in the left column shows the successful tracking of the key points over a sequence of images with arbitrary motion, mostly traverse to the left.

In order to initialize the tracking process the user has to identify three landmarks in the first frame of the sequence. These correspondences are used to compute the orientation of the first frame via spatial resection. This allows for the projection of all other features into the image, which can then be tracked across the next frames.

5.2. Spatial Resection

For the orientation of each frame in the absolute coordinate system we use a photogrammetric technique known as spatial resection. Spatial resection involves the determination of the six parameters of the exterior orientation of a camera station. Several solutions both closed-form and iterative have been proposed in the literature [14].

Since we are working on image sequences where little change in the exterior orientation is to be expected in-between frames, we use an over-determined iterative solution, where the results of the previous epoch serve as initial values for the current computation. For an iterative solution, the collinearity equations have to be linearized. This is standard in photogrammetry. The partial derivatives $\frac{\partial x}{\partial X}, \frac{\partial y}{\partial X}, \frac{\partial x}{\partial Y}, \frac{\partial y}{\partial Y}, \frac{\partial x}{\partial Z}, \frac{\partial y}{\partial Z}, \frac{\partial x}{\partial \omega}, \frac{\partial y}{\partial \omega}, \frac{\partial x}{\partial \phi}, \frac{\partial y}{\partial \phi}, \frac{\partial x}{\partial \kappa}, \frac{\partial y}{\partial \kappa}$ need to be formed from equation 1, where ω, ϕ and κ are three Euler angles parameterizing the rotation matrix. This procedure is well known in photogrammetry.

From at least three points known in three-dimensional space and observed by the camera, we can determine the unknowns X, Y, Z, ω, ϕ and κ. Up to four solutions exist in theory, but since we initiate the iterative estimation process close to its final solution, this ambiguity is irrelevant. One condition for the success of the computation is, that the points selected do not lie on a straight line in three-dimensional space. Furthermore the process will be more reliable when the points are well distributed in image space.

After the orientation has been performed successfully all three-dimensional features can be projected into the image plane. Figure 6 in the right column shows the projection of the extracted edges onto the image. Currently we can not provide reference measurements for direct checking the accuracy of the recovered orientations. We can only estimate accuracy by observing image residuals at the control points, which is the quantity minimized by spatal resection. Visually the good alignment of extracted edge points projected onto the edges in the image indicate successful orientation of the camera.

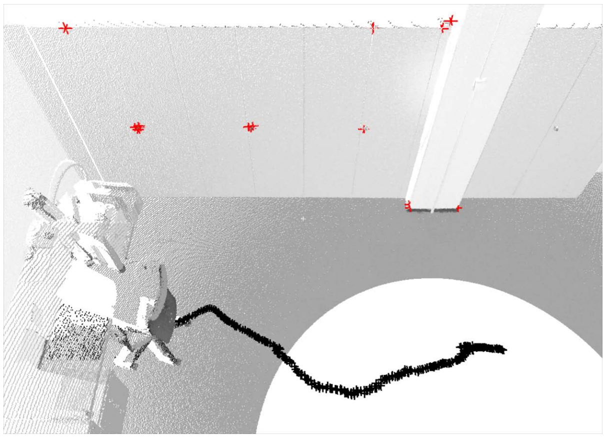

Figure 7. Recovered trajectory over a sequence of 400 frames with arbitrary camera motion.

In figure 7 we finally show the trajectory recovered with our approach in the example office environment. The sequence consist of 400 frames with the camera mentioned above. Three of the frames were already shown in figure 6, including the tracked key points. The camera motion is unconstrained. The landmarks used for the orientation process are also shown.

6. Summary

We have presented a method for the orientation of images in a sequence within the absolute coordinate frame given by a point-based environment model. The method effectively aligns the image stream with a three-dimensional point cloud. A contribution of this work is the observation, that intensity features extracted from laser scanner point clouds provide sufficient landmarks for the orientation of intensity images. This enables the use of separately acquired point cloud data (possibly collected for completely different purposes) for self-localization and navigation tasks. The benefit of this approach is that no artificial landmarks, beacons, etc. have to be place in the environment. Still the mobile agent needs to be equipped with only an intensity camera.

Currently a shortcoming of the method is the fact that

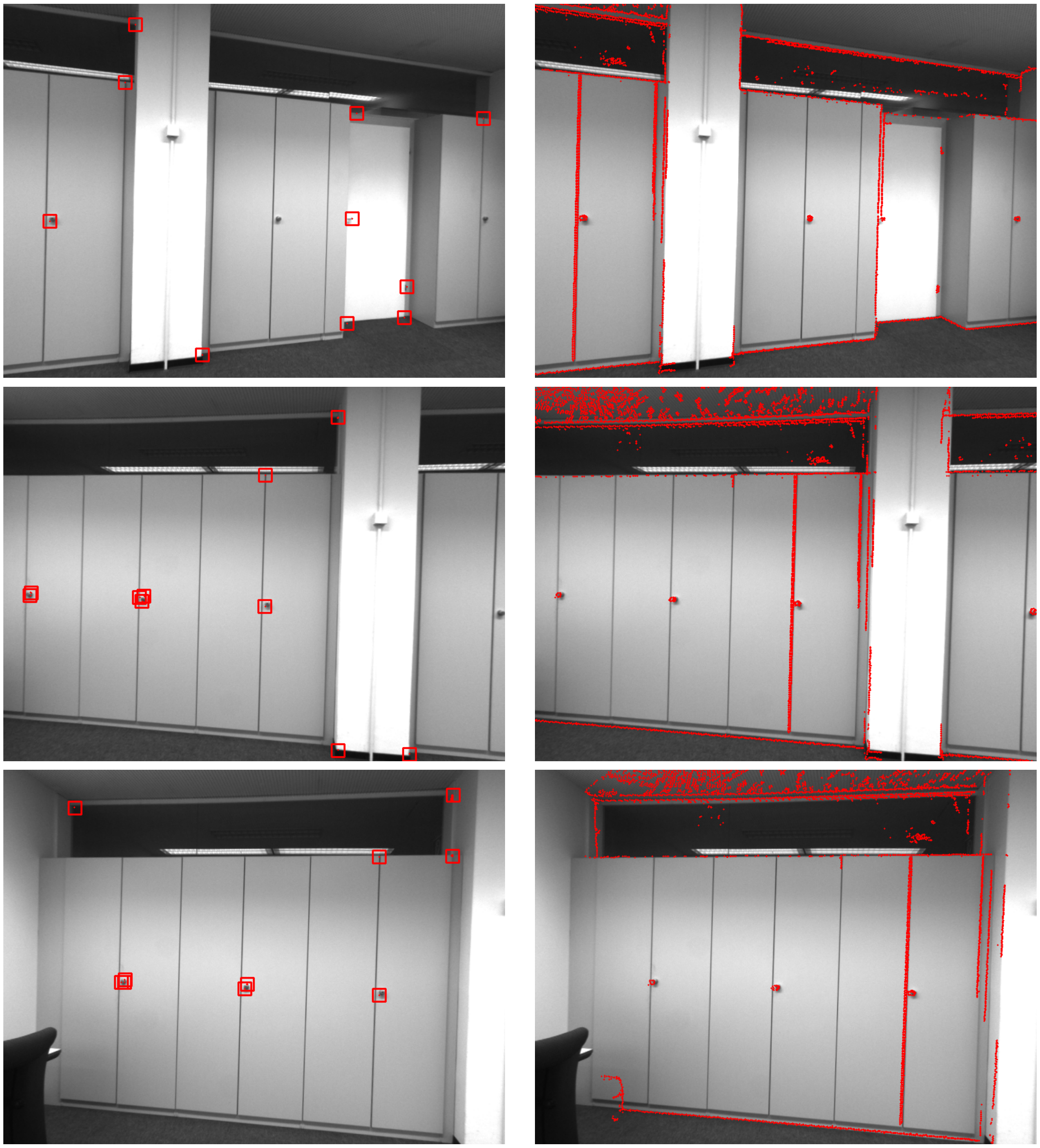

Figure 6. Three frames of a sequence each 100 frames apart showing the successful tracking of landmarks. They are part of a sequence of 400 frames taken from an office environment. The order of the sequence is from top to bottom. The left column shows the features selected for tracking which are marked with a box. The right column shows the projection of the edge points extracted from the laser data onto the images to verify the computed camera orientation.

we need to manually initialize the tracking process by identifying a minimal number of landmarks in the first frame. In the future this could be replaced by providing a collection of more descriptive landmarks, for example using the SIFT operator [15], which provides a local feature description. This can possibly automate the initial identification of landmarks and could also serve as a recovery strategy, when tracking is interrupted.

References

[1] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method," *Int. J. Comput. Vision*, vol. 9, no. 2, pp. 137–154, 1992.

[2] M. Pollefeys, R. Koch, M. Vergauwen, and L. Van Gool, "Automated reconstruction of 3d scenes from sequences of images," *ISPRS Journal Of Photogrammetry And Remote Sensing*, vol. 55, no. 4, pp. 251–267, 2000.

[3] M. Montemerlo and S. Thrun, *FastSLAM: A Scalable Method for the Simultaneous Localization and Mapping Problem in Robotics*. Springer, 2007.

[4] A. Davison, "Real-time simultaneous localisation and mapping with a single camera," in *Proc. International Conference on Computer Vision, Nice*, Oct. 2003.

[5] R. Kurazume, K. Nishino, M. D. Wheeler, and K. Ikeuchi, "Mapping textures on 3d geometric model using reflectance image," *Systems and Computers in Japan*, vol. 36, no. 13, pp. 92–101, 2005.

[6] W. Zhao, D. Nister, and S. Hsu, "Alignment of continuous video onto 3d point clouds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1305–1318, 2005.

[7] R. Lerner, E. Rivlin, and H. P. Rotstein, "Pose and motion recovery from feature correspondences and a digital terrain map," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1404–1417, 2006.

[8] C. Harris and M. Stephens, "A combined corner and edge detector," in *Fourth Alvey Vision Conference*, pp. 147–151, 1988.

[9] C. McGlone, ed., *Manual of Photogrammetry*. American Society for Photogrammetry and Remote Sensing, 2004.

[10] D. C. Brown, "Close-range camera calibration," *Photogrammetric Engineering*, vol. 37, no. 8, pp. 855–866, 1971.

[11] C. S. Fraser, "Digital camera self-calibration," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 52, pp. 149–159, 1997.

[12] J. Shi and C. Tomasi, "Good features to track," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)*, (Seattle), June 1994.

[13] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applicationsto image analysis and automated cartography," *Communications of the ACM*, vol. 24, pp. 381–393, June 1981.

[14] R. M. Haralick, C.-N. Lee, K. Ottenberg, and M. Nöelle, "Review and analysis of solutions of the three point perspective pose estimation problem," *Int. J. Comput. Vision*, vol. 13, no. 3, pp. 331–356, 1994.

[15] D. G. Lowe, "Object recognition from local scale-invariant features," in *International Conference on Computer Vision*, (Corfu, Greece), pp. 1150–1157, 1999.