Camera Pan-Tilt Ego-Motion Tracking from Point-Based Environment Models

Jan Böhm
Institute for Photogrammetry, Universität Stuttgart, Germany

Keywords: Tracking, image sequence, feature point, camera, laser scanner

ABSTRACT

We propose a point-based environment model (PEM) to represent the absolute coordinate frame in which camera motion is to be tracked. The PEM can easily be acquired by laser scanning both indoors and outdoors even over long distances. The approach avoids any expensive modeling step and instead uses the raw point data for scene representation. Also the approach requires no additional artificial markers or active components as orientation cues. Using intensity feature detection techniques key points are automatically extracted from the PEM and tracked across the image sequence. The orientation procedure of the imaging sensor is solely based on spatial resection.

1. INTRODUCTION

Localization is an important problem studied in many fields of research. Applications range from pedestrian self-localization, augmented reality and autonomous robotics 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.

In the field of computer vision the problem is often confined to the problem of determining position and orientation, or orientation in short, from the information given by an imaging sensor, i.e. a camera, alone. The most prominent approaches today are for one structure from motion (SfM),[1] which puts the emphasis on the problem of deriving a representation of the scene and the actual localization is more of a by-product. Another well known approach is simultaneous localization and mapping (SLAM), or more specific visual SLAM (VSLAM),[2] which emphasizes more the localization problem. In their core these approaches use some variation of relative orientation and stereo matching algorithms to derive orientation and scene structure. These approaches are particularly suitable when the system is supposed to work in an unknown environment with no or limited information on the scene. In such cases pure local orientation relative to an arbitrary starting point is sufficient.

In contrast to these approaches we propose to separate the acquisition of the three-dimensional scene structure from the on-line orientation procedure. We propose a point-based environment model (PEM) to represent the absolute coordinate frame in which the motion is being estimated. The motivation for this approach is based on the expectation, that dense point clouds of large building complexes, industrial facilities and urban areas will become widely available in the next few years. The main drive behind the acquisition of such data is from the computer-aided facility management (CAFM) industry, preservation authorities and safety agencies. The main enabling factor is the recent wide spread availability of reliable sensor systems and service companies. Once this data has been acquired it can serve multiple purposes, our proposed application being only one in many.

The PEM can easily be acquired by up-to-date laser scanning systems both indoors and outdoors even over long distances. In contrast to model-based approaches we avoid any expensive modeling step and directly operate on the raw point data. Using intensity feature detection techniques we automatically extract key points from the point cloud and thus replace artificial markers or route marks by natural land marks. The on-line orientation procedure is carried out similar to standard SfM approaches. Using the orientation information obtained for one image we make a prediction on the location of the key points in a subsequent image in a sequence of images. A small area around the key point is matched within the current image for exact localization using the hypothesis on its local motion. In our case matching is performed based on a

jan.boehm@ifp.uni-stuttgart.de, www.ifp.uni-stuttgart.de/institut/staff/boehm.htm

linear motion model. This is valid since we have restricted the motion of the camera to a pan-tilt motion at an arbitrary but fixed position.

The proposed approach results in a robust yet simple sequence of image processing procedures, mainly consisting of off-line feature extraction and on-line intensity image matching. Using simple image processing blocks creates the possibility of using highly optimized image processing libraries opening the possibility for real-time capability of the approach. With respect to SfM and VSLAM the approach has the advantage of computing an absolute orientation with respect to the coordinate frame established by the PEM. Using the automatically extracted three-dimensional key points makes the approach more robust since it avoids the need of interleaved scene reconstruction and orientation, but is based on an already available reliable three-dimensional model.

For the experiments presented in this work, we restrict the camera motion to one common motion type known as pan-tilt motion, which consist of two rotations one about the vertical and one about the horizontal axis perpendicular to the optical axis, where the rotation center is the projection center of a camera. This restriction is made to simplify image feature tracking. However the framework is designed to make an extension to arbitrary motion possible.

2. POINT CLOUD

In our proposed approach a dense point cloud is used to describe the scene. This point-based environment model and the features extracted thereof shall serve as a navigational frame for subsequent image sensor orientation. 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 provide accurate control information for camera orientation.

2.1. 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 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 on top.

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. For one the intensity values are recorded as the intensity of the reflected beam, thus intensities are recorded only at a very narrow bandwidth of the laser beam around 532 nm. Furthermore a camera depends on an external light source which is usually placed at an offset to the camera, e.g. natural sun light. In contrast the laser scanner uses its own laser beam for illumination. Thus the direction of illumination is always identical to the direction of observation. Therefore no shadowing can occur.

2.2. 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

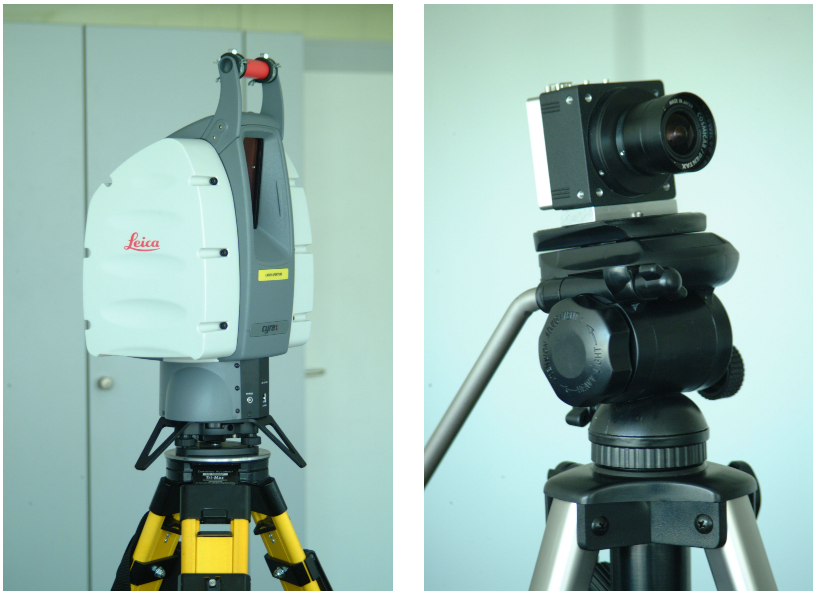

Figure 1. The sensors utilized for this study are a laser scanner and a machine vision camera. The laser scanner to the left is a Leica HDS 3000. The camera to the right is a Basler A302f. Both devices were mounted on a tripod during data acquisition.

function.[3] 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 by

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

and

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

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

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 2 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 and local thresholding to detect edge points in a fashion similar to that for feature points. The results are shown in figure 2 on the right. A simple three-dimensional edge model is obtained. However it shall be noted that the edges only consist of a collection of points, similar to edge pixels in an image.

3. 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 are

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

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

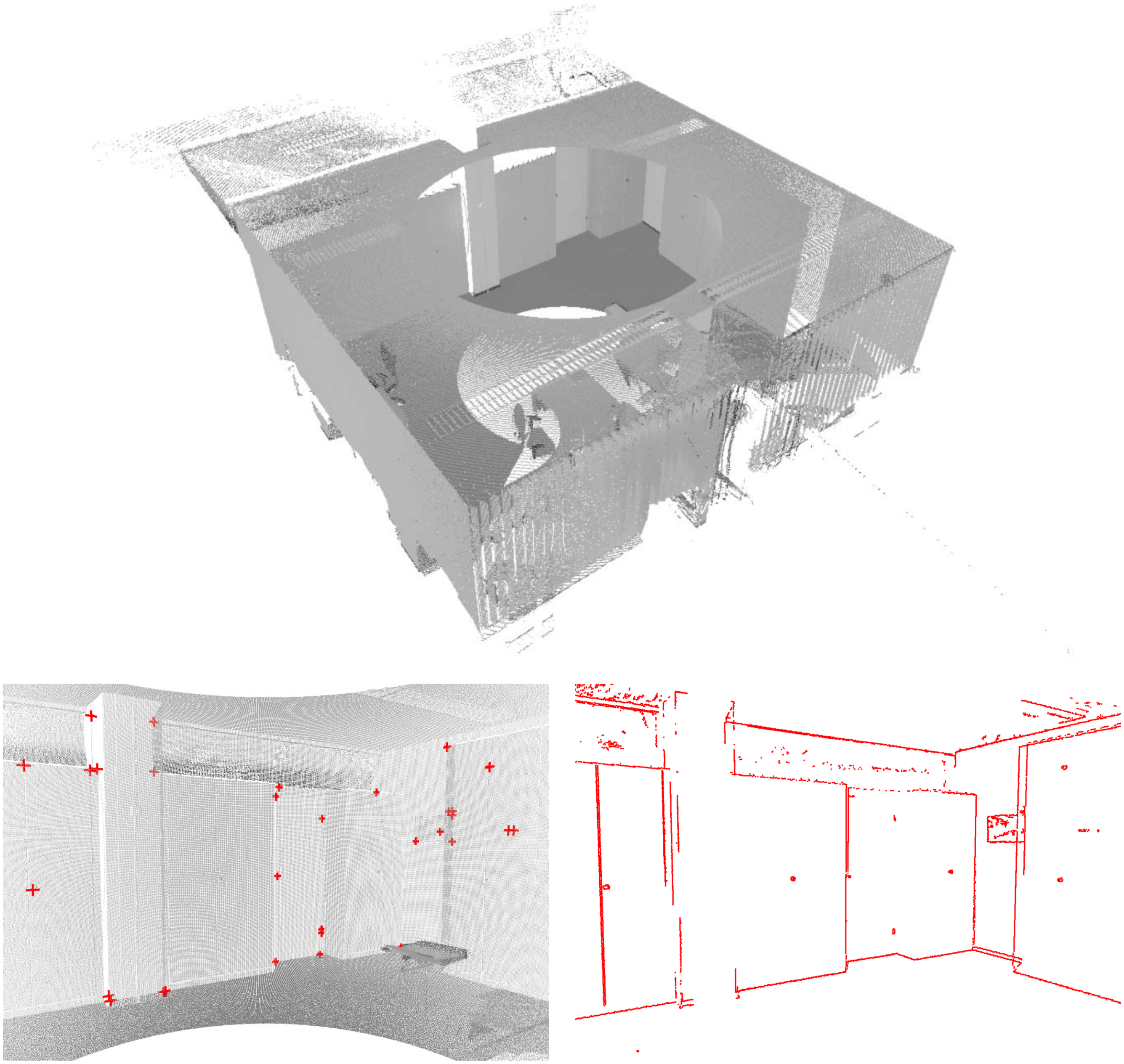

Figure 2. The point cloud data acquired by our laser scanner. The top figure shows an overview of the full scan of an office. The bottom left figure shows a part of the scan with automatically extracted feature points. The bottom right figure shows a reduced point cloud which contains only edge points.

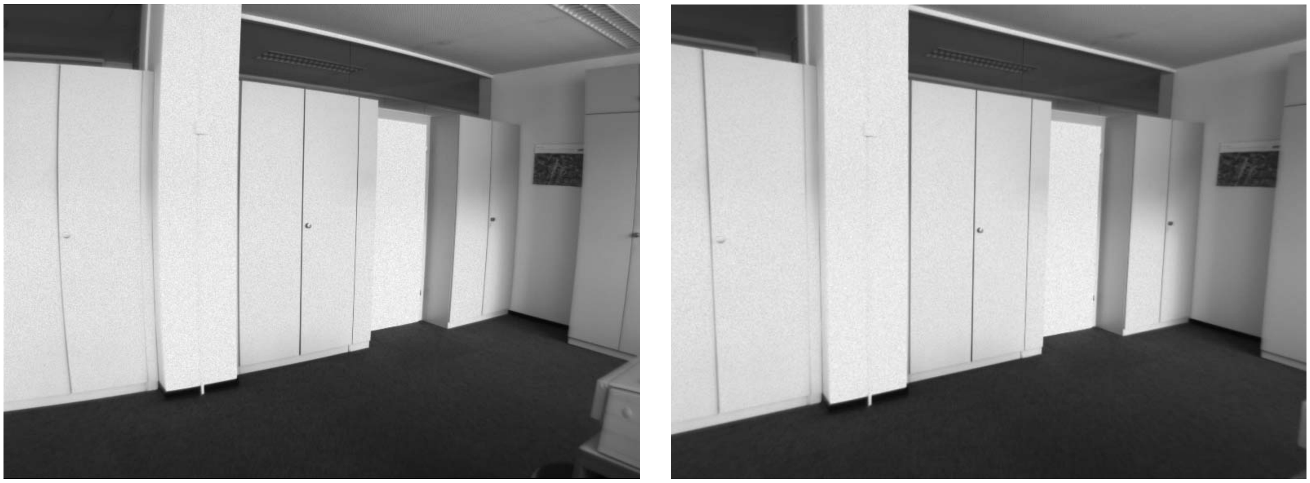

Figure 3. The lens distortion parameters and the coordinates of the principal point obtained during photogrammetric calibration are directly applied to the images. The left image shows an original frame from the sequence. The effects of lens distortions are particularly visible on the straight edges of the wall unit. The right image shows the processed image free of lens distortions.

3.1. Distortion Model

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[4] in a variation for digital cameras presented by C. S. Fraser.[5] 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' as follows

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

Where $r = \sqrt{\overline{x}^2 + \overline{y}^2}$ is 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 before-hand and the resulting parameters are stored. The corrections are then directly applied to the images resulting in distortion-free images. The result of camera calibration can be seen in figure 3. The advantage of this procedure is that we can use the simpler pin-hole camera model for the remaining computations, once the lens distortions have been removed.

4. CAMERA ORIENTATION

Most Structure from Motion (SfM) approaches rely on some form of relative orientation procedure in order to determine the orientation of the camera for a given image. This is necessary since no a priori information on the structure of the scene is available. As a result the orientation (and position) is only known in a local coordinates system relative to a second image. Creating a chain of relative orientations from the first image of a sequence to the last, one can determine the orientation of any image relative to the first. However we have to expect errors to propagate along that chain and thus a steady drift may occur. In addition since relative orientation can not determine the scale of the baseline in-between two images, the position of the cameras is determined without scale. This can create problems when trying to relate the local coordinate system to a global system.

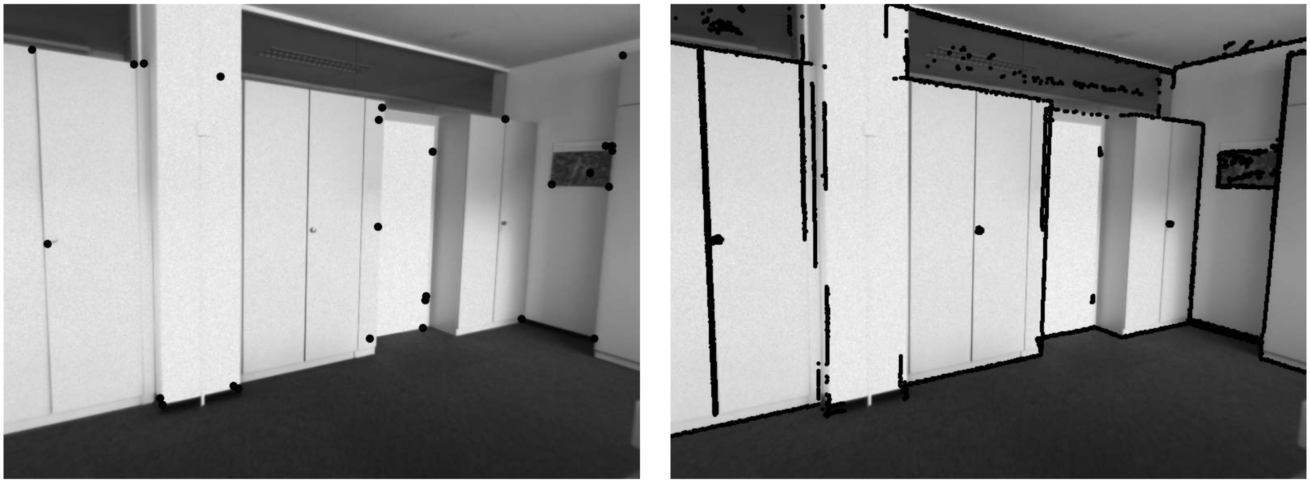

Figure 4. When the exterior orientation of the camera is computed, points measured in the laser point cloud can be projected into the image. The left image shows the laser corner points corresponding to figure 2 projected into the image. The right image shows the same for edge points.

For the approach presented in this paper we follow a different orientation scheme since a full point-based environment model is available to describe the scene. The algorithm to determine the orientation and position of a camera suited for this approach is known as spatial resection.

4.1. 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.[6] 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.

4.2. Feature projection

Obviously we chose the three-dimensional feature points obtained from the processing described in section 2.2 as control points for the resection. If we can find the position of these points in the image we can determine the exterior orientation of the camera and thus project all remaining features into the image as well using equation 1. The results of such a projection are shown in figure 4, where both the projection of individual corner points and projection of the set of edge points is shown. Due to the calibration of the camera carried out before-hand, a high accuracy of the projection can be expected.

5. FEATURE TRACKING AND MOTION ESTIMATION

As we have stated in the introduction the actual tracking process is based on ideas developed in the context of SfM approaches. For the design of a tracking algorithm several key components can be identified.[7] Among them are the feature extraction algorithm, the motion model, the image matching algorithm and a template similarity measurement. Within the proposed framework we do not select feature points from the intensity information given in the image. Rather we rely on the key points projected into the image. The algorithm randomly selects a given number of these points from all key points

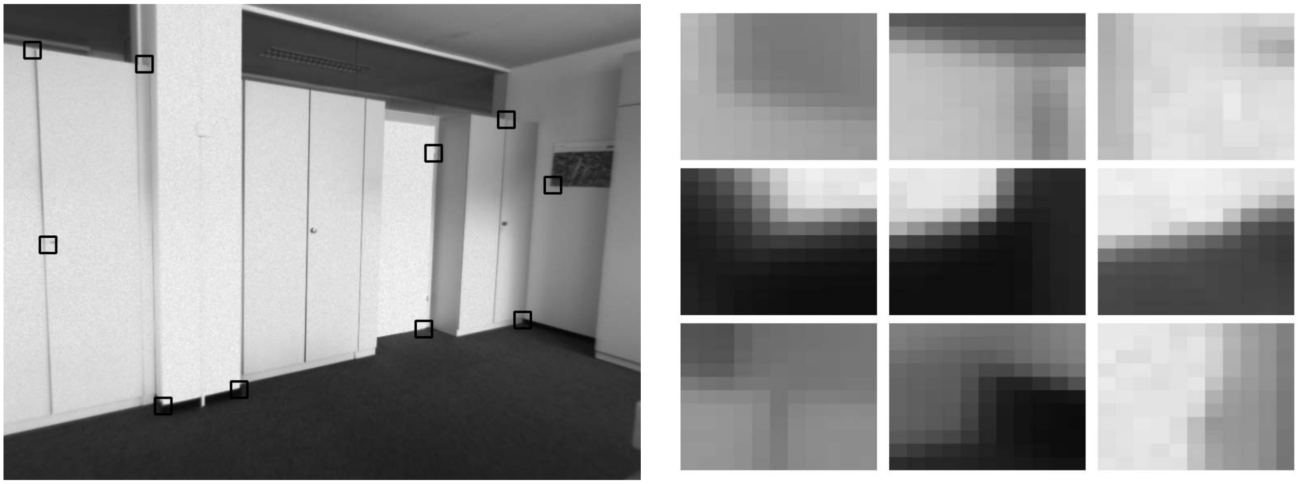

Figure 5. From the available feature points a fixed number of points are randomly selected for tracking, in this case ten are selected. The left figure shows the selected areas. The right figure is a composition of the image templates cut out in these areas.

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 have restricted the camera motion to pan-tilt motion. This was done mainly to make tracking of features simpler. When no change in perspective occurs local image patches remain similar over a long sequence of images. 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 given the constraints works sufficiently well in our tests.

The motion of features within the image resulting from a pan-tilt camera motion is purely linear or in other words is restricted to a translation in row and column direction. We use this motion model to detect outliers in the matching process. We adopt the random sample consensus (RANSAC) strategy for outlier detection. We randomly select key points and check on the consensus of its row and column motion to that of all other key points. When outliers are detected the key point has to be replaced to guarantee 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 key point whose projection falls within the image boundaries. For any key point 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 of a sequence of 500 images acquired with a Basler A302f shown in figure 1 on the left. The camera is able to acquire up to 30 frames at a resolution of 780×582 pixels. It is equipped with a 4.8 mm wide angle lens. In the left image of figure 5 we see the template patches extracted from the image at the location of the projected feature points extracted from the PEM corresponding to the areas marked in the left image. Looking at the templates we see, that they are not necessarily optimal in the sense of Shi and Tomasi.[7] This is a penalty we pay for the fact that our approach does not extract feature points based on the information of the image given. Rather we rely on the assumption that a three-dimensional feature point 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 in section 2.1. A further reason is the perspective discrepancy created by the offset in-between the laser scanner station and the camera station. We purposely set the camera tripod to a different location than the laser scanner in order to test whether arbitrary but fixed camera positions can be successfully processed. If we look back at figure 4 we can see that a feature point is projected on an uniform area of the white column. This point is a feature point extracted on the occluded back-side of the column, but projected onto the column due to the difference in perspective . Our experiments have shown that such problematic cases are easily detected and discarded by our motion model, since they result in arbitrary matches and stray motion.

Figure 6 finally shows a set of four images from the sequence introduced above. The images are not consecutive but are several frames apart. The left column shows the projection of three-dimensional feature points and the search areas for the image template matching. A single ray passing through the projection center, the matched location of the feature point in the image and the corresponding three-dimensional feature point itself, gives the necessary information to compute the angles of the pan tilt motion by simple trigonometric computation. We rather choose to re-use our orientation procedure introduced above, and thus compute a spatial resection from all tracked locations of the key points. This adds accuracy as it is a least square adjustment of all observation. Furthermore it takes into consideration the fact that the camera does not exactly rotate about its principle point as it is attached to a two-axis tripod head, which rotates off-axis. Therefore a slight movement in the position is inevitable.

The right column of figure 6 shows the projection of three-dimensional edge points. This serves as a visual test for the correctness of the tracking and the estimation of the orientation by resection. The optimal alignment of extracted laser edge points to the edges in the image indicates successful orientation of the camera.

6. SUMMARY

We have presented a point-based environment model which represents the scene and is used for orientation of an imaging sensor. The point cloud was acquired using a proven laser scanning system. The point cloud was automatically processes using image processing techniques applied to the intensity channel. Key points extracted from the PEM were successfully used for tracking across an image sequence. From the known three-dimensional location of these points and their respective location in the image, the motion of the camera was reliably estimated. This shows that SfM technology can readily be applied when three-dimensional scene information is available a-priori and processing benefits from this information since orientation procedures are simplified. The natural extension of the proposed method is to lift the restrictions on the motion type and allow for free motion. The spatial resection algorithm chosen for orientation computation already allows for such motion. Adaptations need to be made to the motion model of the tracker. This shall be further investigated in subsequent studies and experiments.

REFERENCES

1. 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* **55**(4), pp. 251–267, 2000.
2. A. Davison, “Real-time simultaneous localisation and mapping with a single camera,” Oct. 2003.
3. C. Harris and M. Stephens, “A combined corner and edge detector,” in *Fourth Alvey Vision Conference*, pp. 147–151, 1988.
4. D. C. Brown, “Close-range camera calibration,” *Photogrammetric Engineering* **37**(8), pp. 855–866, 1971.
5. C. S. Fraser, “Digital camera self-calibration,” *ISPRS Journal of Photogrammetry and Remote Sensing* **52**, pp. 149–159, 1997.
6. 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* **13**(3), pp. 331–356, 1994.
7. J. Shi and C. Tomasi, “Good features to track,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR’94)*, (Seattle), June 1994.

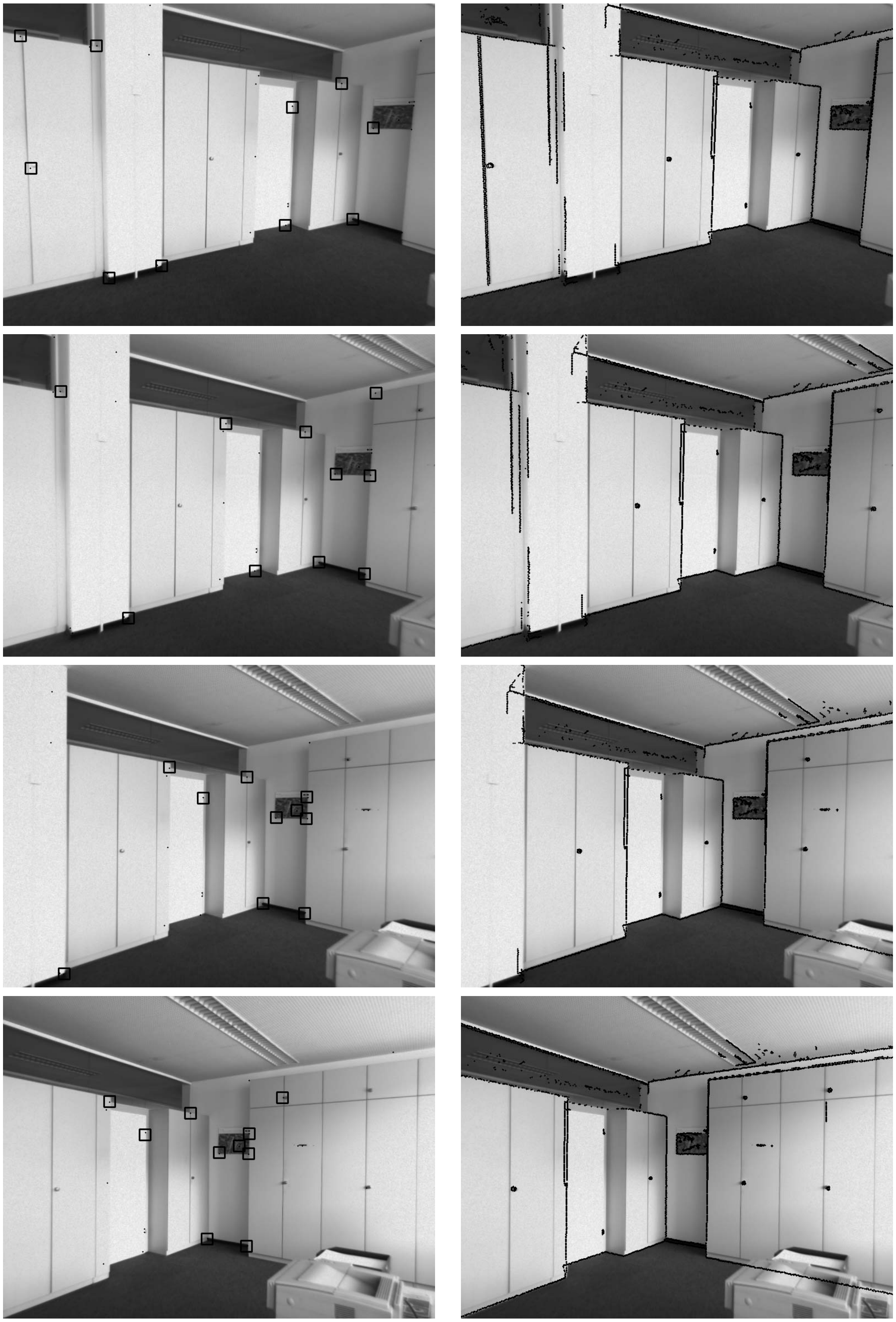

Figure 6. Four frames out of a sequence of 500 frames taken from an office environment. The four frames show a pan motion from left to right. The order of the sequence is from top to bottom. The left column shows the projected feature points, which were extracted from the laser range data. The features selected for tracking are marked with a box. The right columns shows the projection of the edge points extracted from the laser data onto the images.