GENERATION AND WEIGHTING OF 3D POINT CORRESPONDENCES FOR IMPROVED REGISTRATION OF RGB-D DATA

K. Khoshelham [a,*], D. R. Dos Santos [b], G. Vosselman [a]

[a] Faculty of Geo-Information Science and Earth Observation, University of Twente, Netherlands - (k.khoshelham, george.vosselman)@utwente.nl
[b] Federal University of Paraná, Curitiba, Brazil - danielsantos@ufpr.br

Commission III/WG III-2

KEY WORDS: Alignment, Indoor Mapping, Kinect, Loop Closing, Point Cloud, RGB-D, SLAM.

ABSTRACT:

Registration of RGB-D data using visual features is often influenced by errors in the transformation of visual features to 3D space as well as the random error of individual 3D points. In a long sequence, these errors accumulate and lead to inaccurate and deformed point clouds, particularly in situations where loop closing is not feasible. We present an epipolar search method for accurate transformation of the keypoints from 2D to 3D space, and define weights for the 3D points based on the theoretical random error of depth measurements. Our results show that the epipolar search method results in more accurate 3D correspondences. We also demonstrate that weighting the 3D points improves the accuracy of sensor pose estimates along the trajectory.

1. INTRODUCTION

Since their recent introduction to the market, RGB-D cameras, such as the Kinect (Microsoft, 2010), have gained a lot of popularity for indoor mapping, modelling and navigation. The Kinect sensor captures depth and colour images at a rate of 20~30 frames per second, which can be combined into a coloured point cloud, also referred to as RGB-D data. Compared to laser scanning, Kinect RGB-D data have lower accuracy and resolution (Khoshelham, 2011). However, the high data acquisition rate and the great flexibility of the Kinect make it an attractive sensor for mapping and modelling indoor environments.

A primary step in mapping by RGB-D data is the registration of successive frames. The common approach is based on visual features, i.e. point correspondences extracted from the colour images by keypoint extraction and matching methods such as SIFT (Lowe, 2004) and SURF (Bay et al., 2008). These point correspondences are transformed to 3D space by using the depth data, and are then used to estimate the rotation and translation between every pair of frames.

The pairwise registration is prone to error due to the random error of individual points but also the transformation from the colour space to the depth space. In a long sequence, the pairwise registration errors accumulate and lead to deformation in the resulting point cloud. To cope with registration errors, loop closing has been used (May et al., 2009; Du et al., 2011; Endres et al., 2012; Henry et al., 2012). A loop in the trajectory of the sensor can be detected when the sensor returns to a scene that is previously observed. Loop closing is essentially a global adjustment of the sensor pose (position and rotation) simultaneously for all frames in a sequence.

Loop closing is not always feasible, for example when mapping a long narrow corridor, or when the two frames at the closing do not have sufficient overlap or reliable keypoint matches. In such situations, improvement of the pairwise registrations is important as it can reduce the error and deformations in the final point cloud.

In this paper, we look into two sources of error in pairwise registration based on visual features: the error in the transformation from the RGB space to the depth space, and the random error of individual points in the 3D space. We present a method for accurate transformation of point features from the RGB space to the depth space, and propose a weighting scheme to adjust the contribution of the 3D point correspondences in the estimation of the registration parameters. Our experiments show the role of relative orientation in the accuracy of the 3D point correspondences. We also demonstrate that weighting point correspondences based on their theoretical random error improves the registration accuracy.

The paper proceeds with a review of related literature in Section 2. In Section 3, the methods for the generation and weighting of 3D point correspondences are described. Section 4 describes the experiments and results of registration using weighted point correspondences. The paper concludes with final remarks in Section 5.

2. RELATED WORK

The popular approach to registering point clouds is the iterative closest point (ICP) algorithm (Besl and McKay, 1992; Chen and Medioni, 1992). Izadi et al. (2011) showed real-time registration of Kinect depth images using a GPU implementation of the ICP algorithm. The method of Fioraio and Konolige (2011) was also based on ICP, but could integrate features from the colour image.

Since ICP is a fine registration method requiring a close approximation of the registration parameters, it has been often used to refine an initial coarse registration. In the work of Henry et al. (2010), the initial registration parameters were estimated

---

* Corresponding author.

from SIFT key points (Lowe, 2004) extracted from and matched across the colour images, where outliers were removed using RANSAC (Fischler and Bolles, 1981). Du et al. (2011) followed a similar approach but allowed user interaction. The RGB-D SLAM method (Engelhard et al., 2011; Endres et al., 2012) and the method of Bachrach et al., (2012) were both based on the idea of initial registration using visual feature points, although they used different feature extraction operators. Dryanovski et al. (2012) performed the initial registration based on edge features extracted from the colour images. Steinbrucker et al. (2011) adopted an energy minimization approach to registering RGB-D data.

For loop closing several methods have been used. Graph-based optimization methods (Olson et al., 2006; Grisetti et al., 2007; Kummerle et al., 2011) represent the poses and their constraints as nodes and edges of a graph, and apply an optimization method such as gradient decent to minimize the error. Sparse bundle adjustment (Lourakis and Argyros, 2009) involves least-squares (re-)estimation of pose parameters by minimizing the re-projection error in the image space.

Other types of correspondences, such as planes (Brenner and Dold, 2007; Khoshelham and Gorte, 2009; Khoshelham, 2010), point-planes (Sande et al., 2010; Grant et al., 2012) and lines (Bucksch and Khoshelham, 2013; dos Santos et al., 2013), have been used for registering laser scanner point clouds. Taguchi et al. (2012) combined points and planes for the registration of RGB-D data. Dou et al. (2013) combined planes with visual features in both pairwise registration and global adjustment. A comparison of RANSAC and Hough transform for plane extraction and mapping using RGB-D data is presented by Nasir et al. (2012).

3. GENERATION AND WEIGHTING OF 3D POINT CORRESPONDENCES

In this paper, we follow the concept of initial pairwise registration using point features extracted from the colour images. We focus on two aspects in this approach: transformation of the colour image features to the depth image for the generation of 3D point correspondences, and weighting of the 3D point pairs based on the theoretical random error of individual points.

3.1 3D point correspondences from 2D keypoints

We use SURF (Bay et al., 2008) to extract and match keypoints in successive colour images as it is considerably faster than similar algorithms. The keypoints are defined in the 2D coordinate system of the colour image. For the estimation of the pairwise registration parameters the 2D points should be transformed to 3D space by using the depth data. We define the 3D coordinate system of the point cloud with its origin at the centre of the infrared camera, the Z axis perpendicular to the image plane, the X axis perpendicular to the Z axis in the direction of the baseline between the infrared camera centre and the laser projector, and the Y axis orthogonal to X and Z making a right handed coordinate system.

To generate 3D correspondences from the 2D keypoints, in some previous works it has been assumed that a shift of the depth image pixels (applied within the driver) is sufficient to align the depth image with the colour image (Engelhard et al., 2011; Endres et al., 2012; Henry et al., 2012). As we will show, there are cases where the shift between the coordinates of

conjugate points in the colour image and the depth image has a large variance, even when the image coordinates are corrected for lens distortions.

A more proper way to transform the coordinates from the colour image to the depth image is by using the relative orientation parameters (three rotations and three translations – different from photogrammetric relative orientation which involves five parameters) between the two cameras. This of course requires that the relative orientation parameters are estimated in a previous calibration procedure. For the estimation of relative orientation parameters stereo calibration with a calibration grid has been used (Khoshelham and Elberink, 2012). This method provides relative orientation parameters but with relatively low accuracy due to the short length of the baseline between the two cameras in proportion to the distance to the calibration grid. Another approach is by using a 3D calibration field with markers that can be measured in the depth image as well as in the colour image. By measuring the markers in the depth image the 3D coordinates of the points are obtained in the infrared camera coordinate system. Using the 3D coordinates in the infrared frame and the corresponding 2D coordinates in the RGB frame the transformation between the two frames can be obtained by a least-squares space resection procedure.

The estimated orientation parameters allow the transformation of 3D points to the colour image (back projection), whereas we need to transform the 2D keypoints to the 3D space. This is an ill-posed problem. To overcome that, we make use of the epipolar geometry in the following procedure:

Given a keypoint in the RGB frame:
1. calculate the epipolar line in the depth frame using the relative orientation parameters;
2. define a search band along the epipolar line using the minimum and maximum of the range of depth values (0.5 m and 5 m respectively);

For all pixels within the search band:
1. calculate 3D coordinates and re-project the resulting 3D point back to the RGB frame;
2. calculate and store the distance between the re-projected point and the original keypoint;

Return the 3D point whose re-projection has the smallest distance to the keypoint.

Note that interior orientation parameters (including lens distortion) are used in both frames to transform back and forth between pixel coordinates and image coordinates. When the distance between the keypoint and the nearest re-projected point is larger than a threshold (e.g. 2 pixels) the keypoint is flagged as not having a valid 3D correspondence. Figure 1 illustrates the procedure in a test scene.
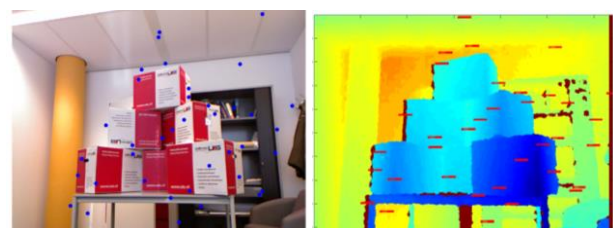

Figure 1. Finding 3D points in the depth image (right) corresponding to 2D keypoints in the colour image (left) by searching along epipolar lines (red bands).

3.2 Definition of weights

Pairwise registration involves the estimation of a rotation matrix **R** and a translation vector **t** between two sets of corresponding points, which minimize the error:

$$E_j = \sum_{i=1}^{n} w_i \left\| X_{i,j-1} - \mathbf{R} X_{i,j} - \mathbf{t} \right\| \tag{1}$$

where $X_{i,j-1}$ and $X_{i,j}$ are the 3D coordinates of point i in frames j-1 and j respectively, and w_i is the weight associated to the point pair i. Since points in the Kinect point clouds do not have a uniform precision (Khoshelham, 2011), it makes perfect sense to weight the points according to their random errors.

As Kinect depth images are captured typically at a frame rate of 20 to 30 fps, resulting in small rotation and translation parameters between successive frames, we can approximate our observation equations with $\mathbf{v}_i = X_{i,j-1} - X_{i,j}$, for which the weight can be defined inversely proportional to the variance of the observation:

$$w_i = \frac{k}{\sigma_{\mathbf{v}_i}^2} = \frac{k}{\sigma_{X_{i,j-1}}^2 + \sigma_{X_{i,j}}^2} \tag{2}$$

where σ_X^2 is the variance of point X and k is an arbitrary constant.

We define the weights for every pair of corresponding points based on the theoretical random error of their depth values (Z) only. This is because weighting based on the error of X, Y coordinates would reduce the contribution of the points with increasing distance from the centre of the point cloud, which is counter-intuitive as off-centre points are expected to play a more important role in the correct alignment of two surfaces. It has been shown that the variance of the depth σ_Z^2 has the following relation with the variance of the measured disparity σ_d^2 (Khoshelham and Elberink, 2012):

$$\sigma_Z^2 = c_1^2 \sigma_d^2 Z^4 \tag{3}$$

where c_1 is a depth calibration parameter. This gives us the following equation for the weight of a point pair:

$$w_i = \frac{k c_1^{-2} \sigma_d^{-2}}{Z_{i,j-1}^4 + Z_{i,j}^4} \tag{4}$$

3.3 Pairwise registration

Once the corresponding 3D points and their associated weights are obtained the point clouds of two successive frames can be registered. The common approach, which is also used here, is to combine the least-squares estimation method with RANSAC to eliminate the outliers (Hartley and Zisserman, 2003). To speed up the registration we use Horn's closed-form solution (Horn, 1987) to estimate the registration parameters for each random sample within RANSAC. Once the inliers are identified, a final iterative least-squares estimation using weighted inlier points is performed to obtain the registration parameters.

4. EXPERIMENTS

To show the effect of relative orientation on the transformation of keypoints from the RGB space to the depth space we made a test scene with markers that could be measured manually in both the depth image and the colour image. The markers were captured and measured in seven pairs of images. Figure 2 shows one of the seven pairs. The coordinates of the markers were then transformed from the colour image to the depth image using the epipolar search method as described in Section 3.1.

The transformation was done using two sets of relative orientation parameters. The first set was obtained by a standard stereo calibration procedure using a calibration grid. The second set was obtained by the space resection method using a 3D calibration field similar to the scene shown in Figure 2. The discrepancies between the manually measured coordinates of the markers in the depth image, and the transformed coordinates obtained by each of the two sets of relative orientation parameters provide an indication of the error in transforming the keypoints from the 2D space to the 3D space.

Figure 3 (a) shows first the difference between the colour image coordinates and the depth image coordinates (both corrected for lens distortion using the model of Brown (1971)) of the markers to test whether the transformation is only a shift. Clearly, there is a large variance in the shift between the two sets of coordinates. Figure 3(b) shows the discrepancies between the measured and the transformed coordinates of the keypoints, where the relative orientation parameters from the stereo calibration are used for the transformation. Figure 3(c) shows the discrepancies between the measured and the transformed coordinates of the keypoints, where the relative orientation parameters from the space resection method are used. It can be seen in Figure 3(c) that the transformed points have a variance of about 1 pixel. This shows that transforming the points by the epipolar search method and using the relative orientation parameters from the space resection is more accurate and reliable than the other methods.
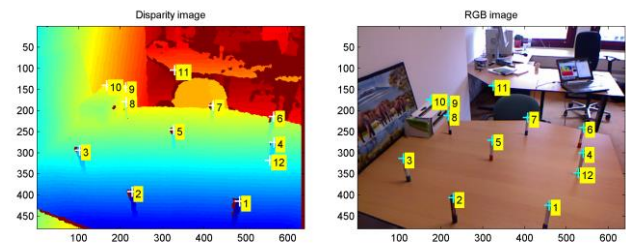

Figure 2. Manually measured markers in the disparity (left) and colour image (right).
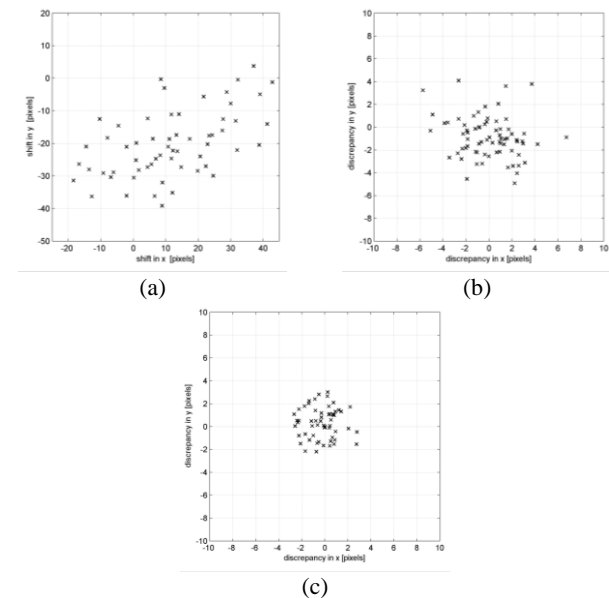

(a)

(b)

(c)

Figure 3. Discrepancies between the manually measured and transformed coordinates of the markers using only a shift (a), using parameters from stereo calibration (b) and using parameters from space resection (c).

To study the effect of weighting 3D point correspondences in pairwise registration a set of six RGB-D sequences from an office environment was acquired. Since obtaining ground truth trajectories was difficult, the sequences were acquired such that the first and the last frame of each sequence had sufficient overlap and could be registered to form a closed loop. This allowed the calculation of the closing error for each trajectory based on the following equation:

$$\Delta = \begin{bmatrix} \delta_R & v \\ 0^T & 1 \end{bmatrix} = \mathbf{H}_2^1 \ldots \mathbf{H}_n^{n-1} \mathbf{H}_1^n \qquad (5)$$

where $\mathbf{H}_i^j$ denotes the transformation from frame i to frame j, and Δ is a residual transformation matrix containing a closing translation vector v and a closing rotation matrix δ_R. From these we calculated two error metrics to evaluate the accuracy of each trajectory: a closing distance from v and a closing angle as the sum of (absolute) rotation angles in δ_R.

Figure 4 shows the closing distances and closing angles for the six sequences after the pairwise registration with and without weights. The sequences were sorted in order of increasing length, and the horizontal axes show sequence length. It can be seen that both the closing distances and closing angles are improved as a result of using weights in pairwise registrations. Table 1 shows the average closing distance and closing angle over all sequences registered with and without weights.

Figure 5 compares for one of the sequences the trajectory obtained by weighted registration (blue curve) with that obtained by registration without weights (red curve). The black curve is the closed loop obtained by a global adjustment of the sensor poses. It can be seen that the trajectory from the weighted registration follows more closely the globally adjusted trajectory. Example point clouds of an office environment obtained by the weighted registration of RGB-D sequences are shown in Figure 6.
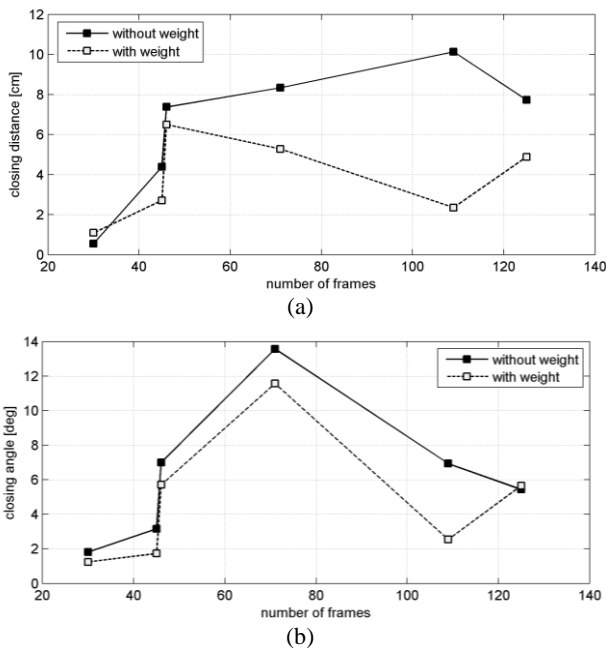
Registration	Average closing distance [cm]	Average closing angle [deg]
without weight	6.42	6.32
with weight	3.80	4.74

Table 1. Average closing errors for registrations with and without weight.
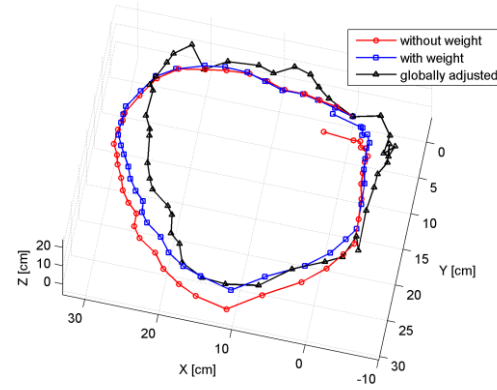

Figure 5. Trajectory obtained by weighted registration of an RGB-D sequence (in blue) compared with the trajectory obtained by registration without weights (in red) and one obtained by global adjustment (in black).

Figure 6. Example point clouds of an office environment obtained by weighted registration of RGB-D sequences.

(a)

(b)

Figure 4. Closing distance (a) and closing angle (b) for six RGB-D sequences registered with and without weights.

5. CONCLUSIONS

When registering long RGB-D sequences, pairwise registration errors accumulate and lead to inaccurate and deformed point clouds, particularly in situations where loop closing is not feasible. We showed that accurate transformation of keypoints from the RGB space to the depth space using an epipolar search method results in more accurate 3D point correspondences. We also showed that assigning weights based on the theoretical random error of the depth measurements improves the accuracy of pairwise registration and sensor pose estimates along the trajectory.

Using weighted observations in pairwise registration allows the estimation of covariance matrices for the estimated pose vectors. These can be used to weight pose vectors in the global adjustment, and further improve the sensor pose estimates in a closed loop.

A drawback of registration by using visual features is the influence of synchronization errors between the RGB camera shutter and the IR camera shutter on the transformation of keypoints to the 3D space. This emphasises the importance of a fine registration step using point- and plane correspondences extracted from the depth images to generate accurate point clouds from RGB-D data.

REFERENCES

Bachrach, A., Prentice, S., He, R., Henry, P., Huang, A.S., Krainin, M., Maturana, D., Fox, D., Roy, N., 2012. Estimation, planning, and mapping for autonomous flight using an RGB-D camera in GPS-denied environments. *The International Journal of Robotics Research* 31(11), 1320-1343.

Bay, H., Ess, A., Tuytelaars, T., Gool, L.V., 2008. SURF: Speeded Up Robust Features. *Computer Vision and Image Understanding* 110(3), 346-359.

Besl, P.J., McKay, N.D., 1992. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14(2), 239-256.

Brenner, C., Dold, C., 2007. Automatic relative orientation of terrestrial laser scans using planar structures and angle constraints, ISPRS Workshop on Laser Scanning 2007 and SilviLaser 2007, Espoo, Finland, pp. 84-89.

Brown, D.C., 1971. Close-range camera calibration. *Photogrammetric Engineering* 37(8), 855-866.

Bucksch, A., Khoshelham, K., 2013. Localized Registration of Point Clouds of Botanic Trees. *IEEE Geoscience and Remote Sensing Letters* 10(3), 631-635.

Chen, Y., Medioni, G., 1992. Object modeling by registration of multiple range images. *Image and Vision Computing* 10(3), 145-155.

dos Santos, D.R., Dal Poz, A.P., Khoshelham, K., 2013. Indirect Georeferencing of Terrestrial Laser Scanning Data using Control Lines. *The Photogrammetric Record* 28(143), 276-292.

Dou, M., Guan, L., Frahm, J.-M., Fuchs, H., 2013. Exploring high-level plane primitives for indoor 3d reconstruction with a hand-held RGB-D camera. Proceedings of the 11th international conference on Computer Vision - Volume 2, Daejeon, Korea.

Dryanovski, I., Jaramillo, C., Jizhong, X., 2012. Incremental registration of RGB-D images, IEEE International Conference on Robotics and Automation (ICRA), St. Paul, Minnesota, pp. 1685-1690.

Du, H., Henry, P., Ren, X., Cheng, M., Goldman, D.B., Seitz, S.M., Fox, D., 2011. Interactive 3D modeling of indoor environments with a consumer depth camera. Proceedings of the 13th international conference on Ubiquitous computing, Beijing, China.

Endres, F., Hess, J., Engelhard, N., Sturm, J., Cremers, D., Burgard, W., 2012. An evaluation of the RGB-D SLAM system, IEEE International Conference on Robotics and Automation (ICRA), St. Paul, Minnesota, pp. 1691-1696.

Engelhard, N., Endres, F., Hess, J., Sturm, J., Burgard, W., 2011. Realtime 3D visual SLAM with a hand-held RGB-D camera. RGB-D Workshop on 3D Perception in Robotics at the European Robotics Forum, Vasteras, Sweden.

Fioraio, N., Konolige, K., 2011. Realtime visual and point cloud slam. RGB-D Workshop on Advanced Reasoning with Depth Cameras at Robotics: Science and Systems Conf. (RSS), University of Southern California.

Fischler, M.A., Bolles, R.C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* 24(6), 381-395.

Grant, D., Bethel, J., Crawford, M., 2012. Point-to-plane registration of terrestrial laser scans. *ISPRS Journal of Photogrammetry and Remote Sensing* 72(0), 16-26.

Grisetti, G., Stachniss, C., Grzonka, S., Burgard, W., 2007. A tree parameterization for efficiently computing maximum likelihood maps using gradient descent. Proc. of Robotics: Science and Systems (RSS), Atlanta, GA, USA.

Hartley, R., Zisserman, A., 2003. *Multiple view geometry in computer vision, 2nd edition*. Cambridge University Press, Cambridge, UK.

Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D., 2010. RGB-D Mapping: Using Depth Cameras for Dense 3D Modeling of Indoor Environments. Proc. of International Symposium on Experimental Robotics (ISER), Delhi, India.

Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D., 2012. RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments. *International Journal of Robotics Research* 31(5), 647-663.

Horn, B.K.P., 1987. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America a-Optics Image Science and Vision* 4(4), 629-642.

Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., Fitzgibbon, A., 2011. KinectFusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera. ACM Symposium on User Interface Software and Technology, Santa Barbara, California.

Khoshelham, K., 2010. Automated Localization of a Laser Scanner in Indoor Environments Using Planar Objects. International Conference on Indoor Positioning and Indoor Navigation (IPIN), Zürich, Switzerland.

Khoshelham, K., 2011. Accuracy analysis of kinect depth data. ISPRS workshop laser scanning 2011, Calgary, Canada.

Khoshelham, K., Elberink, S.O., 2012. Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications. *Sensors* 12(2), 1437-1454.

Khoshelham, K., Gorte, B.G., 2009. Registering point clouds of polyhedral buildings to 2D maps. Proceedings of the 3rd ISPRS International Workshop 3D-ARCH 2009: "3D Virtual Reconstruction and Visualization of Complex Architectures", Trento, Italy.

Kummerle, R., Grisetti, G., Strasdat, H., Konolige, K., Burgard, W., 2011. g2o: A general framework for graph optimization, IEEE International Conference on Robotics and Automation (ICRA), Shanghai, China, pp. 3607-3613.

Lourakis, M.I.A., Argyros, A.A., 2009. SBA: A Software Package for Generic Sparse Bundle Adjustment. *ACM Transactions on Mathematical Software* 36(1), 1-30.

Lowe, D.G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60(2), 91-110.

May, S., Droeschel, D., Holz, D., Fuchs, S., Malis, E., Nüchter, A., Hertzberg, J., 2009. Three-dimensional mapping with time-of-flight cameras. *Journal of Field Robotics* 26(11-12), 934-965.

Microsoft, 2010. Kinect for Windows sensor components and specifications. http://msdn.microsoft.com/en-us/library/jj131033.aspx (26 June 2013).

Nasir, A.K., Hille, C., Roth, H., 2012. Plane Extraction and Map Building Using a Kinect Equipped Mobile Robot. IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012, Workshop on Robot Motion Planning: Online, Reactive, and in Real-time, Vilamoura, Algarve, Portugal.

Olson, E., Leonard, J., Teller, S., 2006. Fast iterative optimization of pose graphs with poor initial estimates. IEEE International Conference on Robotics & Automation (ICRA), Orlando, Florida.

Sande, C.v.d., Soudarissanane, S., Khoshelham, K., 2010. Assessment of Relative Accuracy of AHN-2 Laser Scanning Data Using Planar Features. *Sensors* 10(9), 8198-8214.

Steinbrucker, F., Sturm, J., Cremers, D., 2011. Real-time visual odometry from dense RGB-D images, IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, pp. 719-722.

Taguchi, Y., Jian, Y.D., Ramalingam, S., Feng, C., 2012. SLAM using both points and planes for hand-held 3D sensors. . IEEE International Symposium on Mixed and Augmented Reality, Atlanta, USA.