Robust Ego-Motion Estimation with ToF Cameras

David Droeschel* Stefan May‡ Dirk Holz* Paul Ploeger* Sven Behnke*†
*Fraunhofer IAIS, Schloss Birlinghoven, Sankt Augustin, Germany
†Autonomous Intelligent Systems Group, University of Bonn, Germany
‡INRIA, Sophia-Antipolis, France

***Abstract*— This paper presents an approach to estimate the ego-motion of a robot while moving. The employed sensor is a Time-of-Flight (ToF) camera, the SR3000 from Mesa Imaging. ToF cameras provide depth and reflectance data of the scene at high frame rates.**

The proposed method utilizes the coherence of depth and reflectance data of ToF cameras by detecting image features on reflectance data and estimating the motion on depth data. The motion estimate of the camera is fused with inertial measurements to gain higher accuracy and robustness.

The result of the algorithm is benchmarked against reference poses determined by matching accurate 2D range scans. The evaluation shows that fusing the pose estimate with the data from the IMU improves the accuracy and robustness of the motion estimate against distorted measurements from the sensor.

***Index Terms*— Ego-Motion Estimation, ToF Camera, Sensor Fusion**

I. Introduction

Time-of-flight (ToF) cameras are relatively new, compact, solid-state sensors that provide depth information at high frame rates. They employ an array of infrared LEDs which illuminate the environment with a continuous wave modulation. The reflected signal is received by a combined CCD/CMOS chip. Depth information is gained by measuring the phase shift of the reflected signal. The modulation signal is approximately sinusoidal, with frequencies in the order of some MHz. Measurements are performed in parallel for each pixel. The performance of distance measurements with ToF cameras is limited by a number of error sources. A detailed explanation of the working principle and a definition of an error model has been proposed by Lange [8] and by Schneider [18].

Compared to stereo vision, ToF cameras do not suffer from missing texture in the scene or bad lighting conditions with less computational expensiveness. The advantages of ToF cameras over laser scanners are the high frame rates and the compactness of the sensor. These advantages make them ideally suited for 3D perception and motion reconstruction.

The work presented in this paper utilizes the ToF camera for ego-motion estimation. Ego-motion estimation is solved by incorporating reflectance and depth data of the sensor. In order to increase the robustness of the motion estimate, the estimate based on the camera data is fused with an inertial measurement unit (IMU). The advantages of this approach and the contributions are:

Incorporation of reflectance and depth data: Incorporating both sensor modalities has an advantage over pure range image registration approaches in situations where the depth image has less structure but the reflectance image shows image structure, for example, when the sensor is moving orthogonal to a planar wall that shows texture.

Sensor fusion with IMU data improves accuracy and robustness: The fusion of the camera motion estimate with IMU data improves the accuracy and robustness of the estimate in situations where the camera information is distorted or subject to measurement errors.

The remainder of the paper is structured as follows: Section II summarizes the related work in this field. Section III and IV describe the main contribution: an approach to estimate the ego-motion from the camera data and a model to fuse this estimate with inertial measurements. Section V illustrates the experiments that have been carried out to benchmark the proposed method.

II. Related Work

The first robotics application of ToF cameras was published in 2004. Weingarten et. al. [22] used a CSEM ToF camera prototype for basic obstacle avoidance and local path planning. They evaluated and compared the results to a trajectory from 2D laser range-finders. Their experiments showed that path planning and obstacle avoidance based on the ToF camera data could prevent the robot from colliding with an obstacle that was not detected by the 2D laser range finder. The employed ToF camera was a Swisranger SR-2 from Mesa. Sheh et al. used a ToF camera for 3D mapping of a RoboCup Rescue environment [19]. Because of the low apex angle, they rotated the camera on a pan-tilt unit to gain a larger field of view. The robot stopped at every location and took 10 range images at different pan-tilt positions. The acquisition of one scan took 20 seconds. The registration of the acquired range images was assisted by a human operator. Ohno et. al. [15] use the ToF camera to estimate the robot's ego-motion. The ICP algorithm was used on a SR-2 camera from Mesa. The resulting trajectory was compared to a reference trajectory. The experiments involved almost straight trajectories with up to 6.5 m distance. The authors mentioned that in larger scenes with less structure the rotational error would be higher and that the use of a gyroscope could compensate this error. The above publications mainly use algorithms that have been successfully applied to laser range finder data. Applying these methods to the ToF camera is not straightforward mainly for two reasons:

- Compared to laser ranger finders, the measurement accuracy of todays ToF cameras is lower.
- Due to the larger field of view of laser range-finders, the registration of the range images is easier.

Because of the lower measurement accuracy of ToF cameras, many groups addressed error modeling and calibration. Lindner et al. [9] as well as Kahlmann et al. [6] estimate intrinsic parameters of a ToF camera using the reflectance image of a checkerboard and a planar test field with Near-Infra-Red (NIR) LEDs, respectively. A per-pixel precision of at least 10mm was achieved.

Regarding the registration of range images, the *Iterative Closest Point* (ICP) algorithm is the most popular approach [4]. ICP iteratively estimates the transformation between two point clouds, the *model point set* and the *scene point set*. In every iteration, the point correspondences between model and scene are determined by a nearest neighbor search and the transformation between the point correspondences is estimated by a least squares minimization. The mean squared error of the estimated transformation applied to the scene is determined in every iteration. The algorithm iterates until the error converges or a maximum number of iterations is reached. There are many variations of the ICP algorithm. The application of the ICP to ToF camera data has also been studied [12]. A practical problem in the application of the ICP algorithm is the convergence to a local minimum. This is particularly the case in scenes with low structure. These situations occur especially often with the smaller field of view of ToF cameras. Sheh et. al. [19] handled this problem by using a pan-tilt unit which results in a low data acquisition rate. In scenes where the structure is low but the texture of the objects is high, image features from the reflectance image of the camera could contribute to a better motion estimate. Combining registrations based on depth data and reflectance data has been proposed by Swadzba et. al. [21] and Huhle et. al. [5]. The approaches employ feature tracking on reflectance data and range image registration on depth data. Additionally the fusion with higher-resolution cameras has been proposed [16, 17].

III. Ego-Motion Estimation

To estimate the cameras motion between two consecutive frames, image features in the reflectance image of the ToF camera are used to assign point correspondences between the frames. To detect image features, the *Scale Invariant Feature Transform* (SIFT) [10, 11] is used. SIFT features are invariant in rotation and scale and are robust against noise and illumination changes. The SIFT algorithm has been shown to outperform other feature extraction methods [14]. Various refinements of the basic SIFT algorithm have been proposed: PCA-SIFT [7], GLOH [13] and SURF [3] count as the most important. Bauer et. al. [2] compares recent implementations of SIFT and SURF. They show that SIFT yields the best results regarding the *ratio* of incorrect and correct matches and the total number of correct matches.

In order to estimate the camera motion between two camera frames, the features from the two frames are matched against each other. As described in [11], the best match is the *nearest neighbor* in the keypoint descriptor space. To determine the *nearest neighbor*, the Euclidean distance is used. To measure the quality of a match, the *nearest neighbor* and the *second-nearest neighbor* are searched and the distance between them is determined. If they are too close to each other, the match is rejected. Being too close to each other means that the distance of the *nearest neighbor* times c_r is larger than the distance of the *second-nearest neighbor*, where c_r is a suitable value in [0, 1]. Hence only features that are unambiguous in the descriptor space are considered as matches. Experiments have shown that $c_r = 0.6$ results in the best rejection rates in our case.

Figure 1 (a) and (b) show the reflectance image of two consecutive frames. The red and green dots show the detected features from the two images. Figure 1 (c) shows the matching results of the two images. The green dots are the features from image (a) and the red dots are the matched features from frame (b). The white lines, connecting the red and green dots indicate the displacement of a feature over two consecutive frames. 245 features of frame (a) are successfully matched to features from frame (b).

One match constitutes a point correspondence between two frames. By knowing the depth of every pixel, a point correspondence in 3D is known. The set of points from the current frame is called the *scene* point set, and the set of corresponding points in the previous frame is called the *model* point set. The scene is translated and rotated by the sensors ego motion. Thus, the sensor ego-motion can be deduced by finding the best transformation that maps the scene points to the model points. A common way in estimating a rigid transformation is described in [1]. It uses a closed form solution for estimating the 3×3 rotation matrix $\mathbf{R}$ and the translation vector $\vec{t}$, which is based on singular value decomposition (SVD).

The distances between corresponding points, after applying the estimated transformation, forms the *Root Mean Square* (RMS) error. The RMS error is often used in range registration to evaluate the scene-to-model consistency. It can be seen as a measure for the quality of the match: if the RMS error is significantly high, the scene-to-model registration can not be consistent. On the other hand, a low RMS error does not imply a consistent scene-to-model registration, since this is also depending on the number and distribution of the point correspondences.

The translation vector $\vec{t}$ is composed of $(\Delta x, \Delta y, \Delta z)^T$, which is the translational change of the camera between two camera frames. The rotation matrix $\mathbf{R}$ is the change of the camera orientation between two frames. From the rotation matrix the three Euler angles can be calculated.

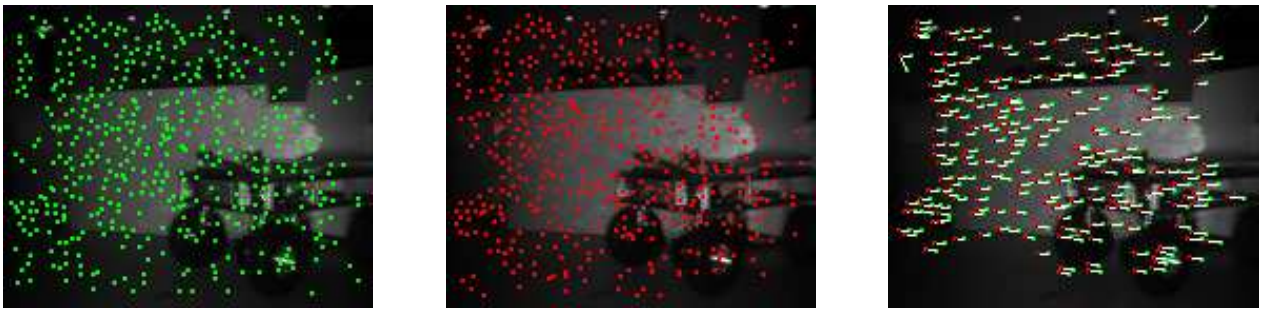
Since the robot is moving on planar ground, the position estimate can be simplified to 2D space. Hence, the translation $(\Delta x, \Delta y)^T$ and the rotation around the vertical (yaw) axis $\Delta\theta$ can be considered as the transformation that describes the camera's motion between two frames.

From the $(\Delta x_k, \Delta y_k, \Delta\theta_k)^T$ at frame k the trajectory of the camera can be built incrementally. The pose $(x_k, y_k, \theta_k)^T$ at frame k can be calculated by

$$(x_k, y_k)^T = (x_{k-1}, y_{k-1})^T + \mathbf{R}(\theta_{k-1})(\Delta x_k, \Delta y_k)^T \quad (1)$$

and

$$\theta_k = \theta_{k-1} + \Delta\theta_k, \quad (2)$$



(a) (b) (c)

Fig. 1: SIFT feature extraction and matching applied on a ToF reflectance image. The scene shows a robot in the pavilion at the Fraunhofer IAIS. Images (a) and (b) show the detected SIFT features on two consecutive camera frames.The number of detected features are 475 (a) and 458 (b). Image (c) shows the matching result: 245 features from image (a) are matched to features from image (b).

where $k-1$ is the previous frame and $\mathbf{R}(\Delta\theta_{k-1})$ is the 2D rotation matrix of θ_{k-1}.

IV. Fusion of Motion Estimates

The motion estimation described in the previous section provides a translational $(\Delta x, \Delta y)^T$ and rotational change $\Delta\theta$ of the camera between two camera frames. By knowing the time between two frames Δt, the translational and rotational velocity is known. This is considered as observation

$$z_k = \frac{1}{\Delta t}(\Delta x, \Delta y, \Delta\theta)^T \tag{3}$$

at time step $\mathbf{k}$.

The employed IMU is a XSens MTi motion tracker, calibrated with the vendor's calibration toolbox. It provides measurements for the rotational velocity v_θ and translational acceleration on the x and y axis, $(\vec{a}_x, \vec{a}_y)$. The *mean* rotational acceleration $\vec{a}_{\theta,k}$ at time step k can be calculated by the difference between the velocity at time steps k and $k-1$:

$$\vec{a}_{\theta,k} = v_{\theta,k} - v_{\theta,k-1}. \tag{4}$$

A Kalman filter predicts the system velocity estimate $(v_x, v_y, v_\theta)^T$. The motion estimate from the camera is considered as observation, whereas the IMU data is considered as control input to the system. The RMS error and its individual components of the estimated transformation reflects the certainty of the observation and is therefore used as an approximation of the observation covariance (similar to [20]).

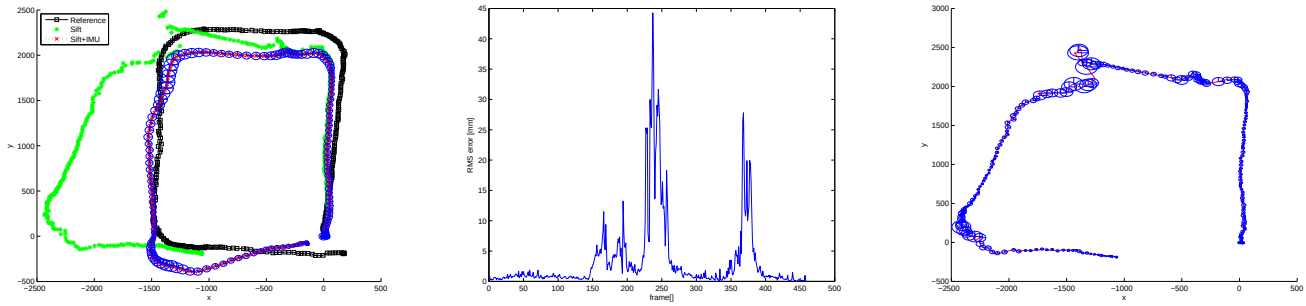
V. Experiments and Results

The following experiments demonstrate the accuracy and robustness of the proposed procedure. A Sick LMS200 laser range finder was used to incrementally construct an accurate and consistent 2D map and to compute a reference trajectory. To generate a reference trajectory, the ICP algorithm is applied.

Figure 2 shows the scene of the first experiment. The experiment was carried out in the Robotic Pavilion at Fraunhofer IAIS, Sankt Augustin. The image shows a wooden staircase with a robot, some posters and a calibration pattern. The robot moved along a square with 120 cm side length. Applying the described motion estimation method to the sensor data results in the estimated trajectory depicted in Figure 3a. The black trajectory shows the reference trajectory based on the 2D laser range finder. The green trajectory shows the SIFT-based motion estimate. On the upper left corner the trajectory is distorted. The application of the sensor fusion is depicted in the red trajectory. Comparing the green and the red trajectories visually shows that the red trajectory is less distorted than the green trajectory, especially in situations where the RMS error is high, e.g. in curves. The blue ellipses on the red trajectory depict the *a posteriori* system covariance of the Kalman filter.

Fig. 2: The scene of the first experiment carried out in the Robotic Pavilion at Fraunhofer IAIS, Sankt Augustin. The scene consists of a wooden staircase with a robot on it, some posters and a calibration pattern. The robot moved on a square with 120 cm side length.

Figure 3b depicts the RMS error of the estimated transformation applied to the matched point pairs. The RMS error was considered as a measure for the quality of the match. The first 150 frames show a relative low RMS error compared to the peak at frame 245. To visualize the correlation of the RMS error and the distorted trajectory, Figure 3c shows the



(a) (b) (c)

Fig. 3: (a) Estimated trajectories. The black trajectory shows the reference data based on the 2D laser range finder. The green trajectory shows the SIFT-based motion estimate. On the upper left corner, the trajectory is distorted. The application of the sensor fusion is depicted in the red trajectory. (b) RMS error of the estimated transformation applied to the matched point pairs for each frame. The first 150 frames show a relative low RMS error compared to the peak at frame 245. (c) Correlation of the RMS error and the distorted trajectory.

estimated trajectory as well as the RMS error distribution. The trajectory is plotted by a red line and the RMS error is visualized by blue ellipses, where the magnitude of the RMS error correlates to the size of the ellipsis. The figure shows high RMS errors at those poses that deviate from the reference trajectory.

Figure 4 shows the translational error in mm (4a), the rotational error in degrees for every frame (4b), and the cumulated rotational error, up to every frame (4c) for the unfiltered and the filtered motion estimate. The blue dashed lines illustrate the error of the unfiltered motion estimate. The red line illustrates the Kalman-filtered motion estimate. The Kalman-filtered motion estimate improves up to 1006 mm on the translational error and up to 25.4 degree on the rotational error.

Figure 5 shows the estimated trajectories of a second experiment. The experiment involved a larger scene with up to 8m diameter. Figure 6 depicts the translational and rotational error of the applied methods, comparing the ego-motion estimate based solely on the camera data to the fused ego-motion estimate. The rotational error of the fused ego-motion estimate improves up to 28.6 degree. The improvement of the translational error in Figure 6a is up to 669mm.

Figure 7 shows the resulting 3D maps based on the estimated ego-motion. Figure 7a shows the unfiltered motion estimate. The resulting map is squeezed on the end of the trajectory due to the error in the ego-motion estimate. In contrast, Figure 7b shows the improved map based on the fused ego-motion estimate.

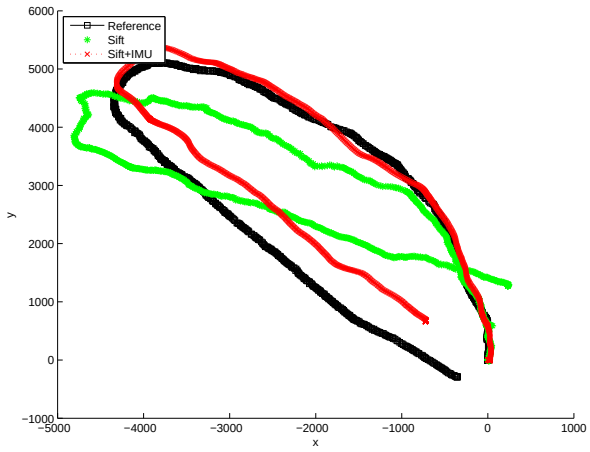


Fig. 5: Estimated trajectories. The black trajectory shows the reference data from the 2D laser range finder. The green trajectory shows the SIFT based motion estimate. The red trajectory shows the Kalman-filtered motion estimate with the fused IMU data. The red trajectory is less distorted than the green one.

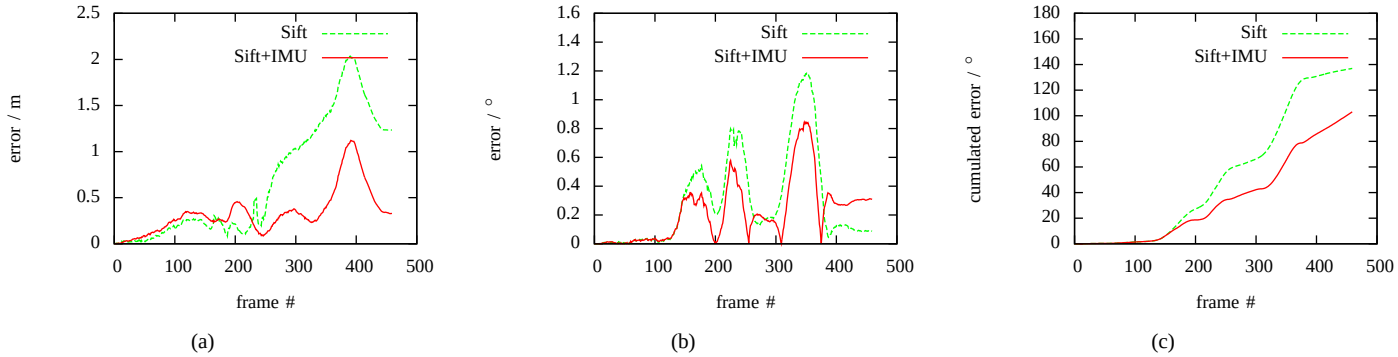
VI. Conclusions

This paper presented a way to estimate a robot's ego-motion while moving. An application of this motion estimate is to map an unknown environment based on the sensor data. The employed sensor is a ToF camera, the SR3000 by Mesa Imaging. ToF cameras provide depth and reflectance data of the scene at a high frame rate. They suffer from a set of error sources which make them difficult to handle. The proposed method utilizes the coherence of depth and reflectance data of ToF cameras by detecting image features on the reflectance data and estimating the motion on the depth data.

The visual motion estimate is fused with the IMU measurements to gain higher accuracy and robustness. The result of the algorithm is benchmarked against reference poses from a 2D laser range finder. The evaluation shows that fusing the pose estimate with the data from the IMU improves the translational error up to 1006 mm and the rotational error up to 28.6 degree. Hence, the proposed method

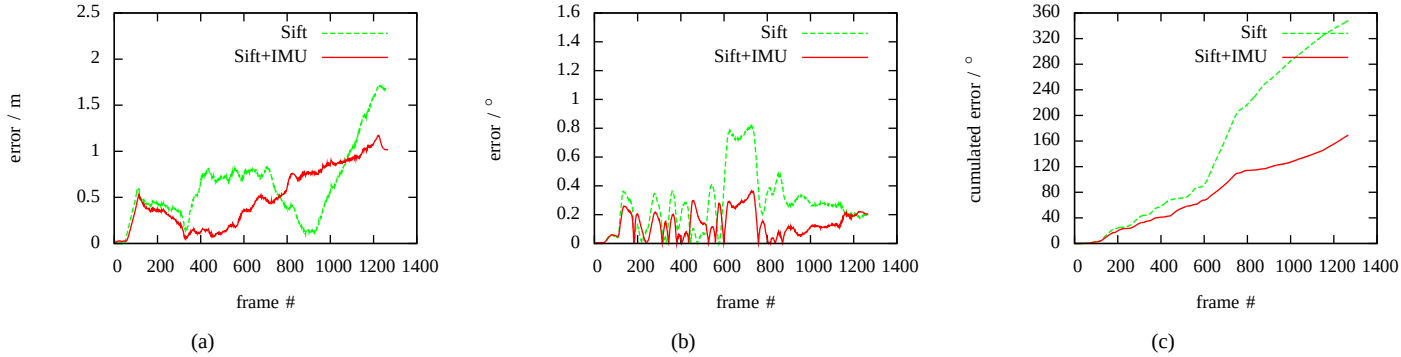
- **Improves the accuracy** of the motion estimate compared to a reference pose from a 2D laser range finder.
- **Improves the robustness** of the motion estimate against distorted measurements from the sensor.

In the first setup the system was used in 3DOF. The limitation of the robot, moving on a planar ground, and of the 2D laser range finder as reference system are the main reasons.



(a) (b) (c)

Fig. 4: (a) Translational error of the unfiltered (dashed green) and filtered (red) motion estimate compared to the reference data from the 2D laser range finder. (b) The rotational error in degrees for every frame. (c) The cumulated rotational error in degrees up to every frame.



(a) (b) (c)

Fig. 6: (a) Translational error of the unfiltered (dashed green) and filtered (red) motion estimate compared to the reference data from the 2D laser range finder. (b) Rotational error in degrees for every frame. (c) Cumulated rotational error in degrees up to every frame.

Future work will concentrate on the extension to 6DOF.

Another important point is the determination of the observation covariance. Here, the RMS error was used as an estimate. In the future work, a camera specific error model has to be considered.

Acknowledgment

This work was supported by the B-IT foundation, Applied Sciences Institute, a cooperation between Fraunhofer IAIS and University of Applied Sciences Bonn-Rhein-Sieg.

References

[1] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-d point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(5):698–700, 1987.

[2] Johannes Bauer, Niko Sünderhauf, and Peter Protzel. Comparing several implementations of recently published feature detectors. In *Proceedings of the International Conference on Intelligent and Autonomous Systems, IAV*, 2007.

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded-up robust features. In *9th European Conference on Computer Vision*, Graz, Austria, 2006.

[4] P.J. Besl and N.D. McKay. A method for Registration of 3–D Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[5] B. Huhle, P. Jenke, and W. Straßer. On-the-fly scene acquisition with a handy multisensor-system. In *Proceedings of the Dynamic 3D Imaging Workshop in Conjunction with DAGM (Dyn3D)*, Heidelberg, Germany, 2007.

[6] T. Kahlmann, F. Remondino, and H. Ingensand. Calibration for increased accuracy of the range imaging camera swissranger. In H.-G. Maas and D. Schneider, editors, *Proceedings of the ISPRS Commission V Symposium 'Image Engineering and Vision Metrology'*, volume XXXVI, pages 136–141, Dresden, Germany, 2006.

[7] Yan Ke and Rahul Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:506–513, 2004.

[8] Robert Lange. *3D time-of-flight distance measurement with custom solid-state image sensors in CMOS/CCD-technology*. PhD thesis, University Siegen, 2000.

[9] Marvin Lindner and Andreas Kolb. Lateral and depth calibration of pmd-distance sensors. In *Advances in Visual Computing*, volume 2, pages 524–533. Springer, 2006.

[10] D. G. Lowe. Object recognition from local scale-invariant features. volume 2, pages 1150–1157 vol.2, 1999.

[11] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Distinctive Image Features from Scale-Invariant Keypoints*, 60(2):91–110, 2004.

[12] Stefan May. *3D Time-of-Flight Ranging for Robotic Perception in Dynamic Environments*. PhD thesis, Universitt Osnabrck, 03 2009.

[13] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10):1615–1630, Oct. 2005.

[14] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation

(a)

(b)

Fig. 7: (a) Top view of the resulting 3D map based on the estimated ego-motion. The map is squeezed at the end of the trajectory due to the error in the ego-motion estimate. (b) Improved map based on the fused ego-motion estimate.

of local descriptors. In *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 257–263, June 2003.

[15] K. Ohno, T. Nomura, and S. Tadokoro. Real-time robot trajectory estimation and 3d map construction using 3d camera. pages 5279–5285, Oct. 2006.

[16] A. Prusak, O. Melnychuk, H. Roth, I. Schiller, and R. Koch. Pose estimation and map building with a pmd-camera for robot navigation. In *Proceedings of the Dynamic 3D Imaging Workshop in Conjunction with DAGM (Dyn3D)*, Heidelberg, Germany, 2007.

[17] Leila Sabeti, Ehsan Parvizi, and Q.M. Jonathan Wu. Visual Tracking Using Color Cameras and Time-of-Flight Range Imaging Sensors. *Journal of Multimedia*, 3:28–36, 2008.

[18] Bernd Schneider. *Der Photomischdetektor zur schnellen 3D-Vermessung für Sicherheitssysteme und zur Informationsübertragung im Automobil*. PhD thesis, Universität-Gesamthochschule Siegen, 2003.

[19] Raymond Sheh, M. Waleed Kadous, and Claude Sammut. On building 3d maps using a range camera: Applications to rescue robotics. Techreport, ARC Centre of Excellence for Autonomous Systems - School of Computer Science and Engineering - The University of New South Wales, Sydney Australia, Sydney, Australia, April 2006.

[20] Andrew J. Stoddart, S. Lemke, Adrian Hilton, and T. Renn. Estimating Pose Uncertainty for Surface Registration. In *Proceedings of the British Machine Vision Conference (BMVC)*, 1996.

[21] Agnes Swadzba, Bing Liu, Jochen Penne, Oliver Jesorsky, and Ralf Kompe. A comprehensive system for 3d modeling from range images acquired from a 3d tof sensor. In *The 5th International Conference on Computer Vision Systems*, 2007.

[22] J.W. Weingarten, G. Gruener, and R. Siegwart. A state-of-the-art 3d sensor for robot navigation. volume 3, pages 2155–2160 vol.3, Sept.-2 Oct. 2004.