Evaluating SLAM Approaches for Microsoft Kinect

Corina Kim Schindhelm
Siemens AG – Corporate Research & Technologies
Munich, Germany
Email: corina.schindhelm@siemens.com

***Abstract*—The weak performances of GPS within buildings is the reason for a lot of different approaches for indoor positioning, e.g., by using WiFi or odometry. The current position of a person is crucial, for example, for location based services that increase not only outside of buildings. Navigation systems in subway stations is just one obvious example, where GPS fails to deliver the necessary information. Especially in the field of visual odometry, there are many approaches. But all of them are either based on normal 2d camera systems or on expensive 3d camera systems. In the presented approach, we use a Microsoft Kinect, as these systems are inexpensive and widespread. We evaluate how different state of the art techniques like RANSAC or ICP can be used in combination with the Kinect and how they perform in different indoor scenarios. Our evaluation shows that those techniques can be used for the Kinect but have their shortcomings in different scenarios. For that reason, a hybrid technique was developed which combines those methods using a Kinect specialized ICP weight function. In addition, we use a loop detection algorithm to further optimize the accuracy. Finally, we present our results obtained during tests in three different test environments. This paper presents the result of different SLAM approaches implemented on the Microsoft Kinect in order to calculate trajectories.**

Keywords-slam, kinect, odometry, indoor positioning.

I. INTRODUCTION

Many indoor positioning methods have been researched and some solutions found their way into consumer products. But there are still not many (public) buildings equipped with indoor positioning systems, even though it would add value to many public institutions (e.g., libraries, schools, universities) or other areas without satellite coverage (e.g., subway stations, tunnels). Mostly, indoor positioning solutions have been deployed into companies with sufficient funds to invest in expensive high precision technologies like Ultra Wide Band, since their businesses can directly benefit through use of indoor asset tracking [1].

A different approach to installing expensive indoor positioning solutions, which also often need a lot of calibration and maintenance, is to make use of a method known from the field of robotics called SLAM (simultaneous logging and mapping). The main idea there is to place a mobile robot at an unknown location in an unknown environment and let the robot incrementally build a consistent map of its environment while simultaneously determining its location within this map [2]. There exists a lot of different algorithms and solutions to solve this problem. We were interested in the question whether those approaches can also be applied to humans and everyday devices instead of robots equipped with high-end sensors.

This paper deals with the comparison of two different SLAM methods and a hybrid approach, which are applied to the Microsoft Kinect carried by a human being. We developed an evaluation platform which allows to compare different SLAM algorithms and their performance in different scenarios (test environments).

The reminder of this paper is structured as followed: Section II will introduce SLAM principals and list some reference work in this field. Section III describes the Microsoft Kinect, the concept and the three different test environments. Section V evaluates the implemented algorithms in respect of the test environments and Section VI concludes the paper.

II. FUNDAMENTALS OF SLAM

SLAM is a method usually applied by robots to create a map of the surrounding while at the same time estimate their location. Among the vast number of different SLAM methods the main principal remains the same: At the start there is no map of the environment, hence the position of the robot is the origin of the coordinate system and the measurement at this position is the initial measurement. From then on each subsequent measurement contains already known data and new unknown data. By comparing the current measurement with the data set the robot can find am overlapping, and therefore, calculates its new position. By including the new measured data into the map, the whole environment can be surveyed incrementally. Since the position shift between two measured data sets is not perfect, the map quality decreases over time. Tim Bailey and Hugh Durrant-Whyte offer two tutorials about SLAM, which deal with the SLAM problem and algorithms solving the problem [2][3]. As mentioned before, there exists a vast number of SLAM methods, e.g., algorithms using particle filters , the Extended Kalman Filter or graph based techniques.In this paper we will focus on two different approaches: the first is based on visual key points and the second one is based on point clouds. Further details will follow in Section III. In practice, there is a variety of systems based on SLAM that use different sensor equipment. A SLAM system using INS (inertial navigation systems) was developed by Robertson et al. INS sensors were installed to

 ISBN: 978-1-61208-203-5

pedestrians' feet to obtain 2D maps of large areas based on iterative processing of pedestrian odometry data [4]. A system using an Extended Kalman Filter and laser scanners was developed by Garulli et al. [5]. Multiple robots using landmarks to create independent maps, which have to be combined subsequently were used by Zhou [6]. A systems using cameras and SURF detectors was implemented by Engelhard et al. [7].

III. SLAM WITH KINECT

This section offers hardware data of the Kinect, the SLAM algorithms and the information about the test settings.

A. The Kinect and quality of sensors

The technical components for the Kinect were developed by PrimeSense [8], which also published the open source API OpenNI together with WillowGarage [9] and Side-Kick [10]. PrimeSense patented Light Coding generates depth information based on a infrared laser projector and a monochrome CMOS sensor camera. The resolution of Kinect's depth image is 320 x 240 pixel, which is internally interpolated to the double size of 640 x 480. Objects can be recognized to a distance within the range of 0.8 meter to 6 meter. The horizontal field of view is 57 and the vertical 43 [11]. An additional RGB camera provides 640 x 480 pixel color images. Together with an audio channel, the micro processor offers a synchronized data stream of color, depths and audio information to a rate of 30 Hz [12].

Since SLAM algorithms are based on accurate sensor data we examined the error rate of Kinect's depths information. The test comprised a set of Kinect pictures of a simple wooden board placed parallel to the view of the Kinect. Measurements were taken from different distances. Figure 1 shows the result that with bigger distance the error of raw data grows significantly. A picture taken from four meters distance results in a maximum of 14.2 centimeters deviation, whereas with 80 centimeters distance the maximum deviation is only one centimeter. To reduce the errors which mainly result from signal noise we applied and examined different filters. Exemplary the results of a median filter [13] and a bilateral filter [14] with different parameters are depicted in Figures 1 and 2. The figures show that using filters can help minimizing the deviation.

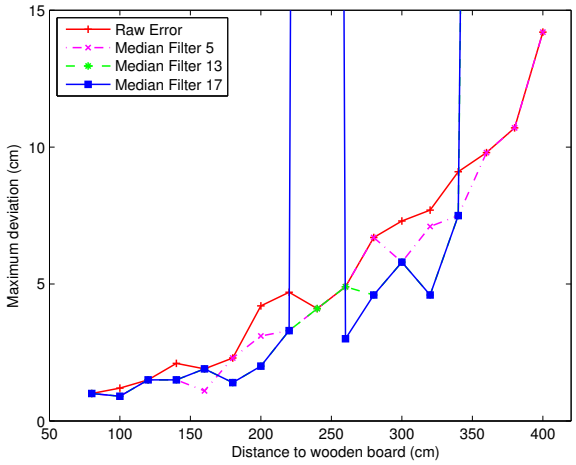

Figure 1. Medianfilter

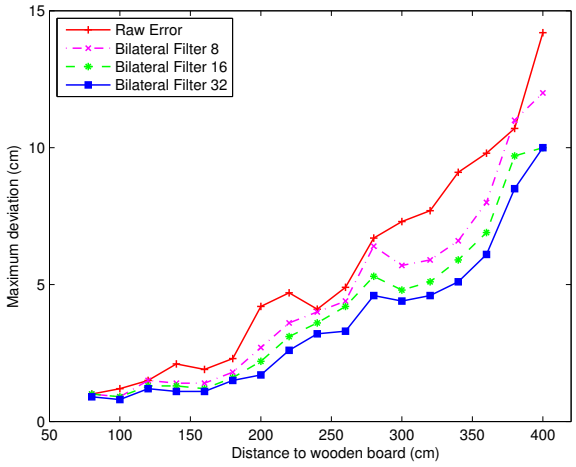

Figure 2. Bilateral filter

B. SLAM Algorithms

Figure 3 gives an overview of the algorithms that are implemented and examined: Visual Keypoints (upper part of Figure 3), Hybrid (middle part of Figure 3) and ICP (lower part of Figure 3).

The SLAM method based on visual key points (see Figure 3 upper part) works as follows: In the first step striking key points have to be detected and categorized (e.g., SURF and Shi Tomasi). The SURF(Speeded Up Robust Feature) method [15] is an enhancement of the SIFT(Scale-invariant feature transform) method [16]. The goals of both is to robustly identify key points among disordered data with a descriptor invariant to uniform scaling, orientation, and partially invariant to distortion and illumination changes. The advantage of the SURF is the higher speed which is achieved for example by replacing the Gaussian filter with a Box filter. Shi and Tomasi detectors are based on Harris and Moravec detectors. The goal of those approaches is to detect corners, whereas a corner is defined as a point with low self similarity. Afterwards, in a second step, homologous key points in two subsequent picture frames must be found. Key points between two pictures found with SIFT/SURF detectors and descriptors can be matched with the minimal Euclidean distance. For key points found with the approach of Shi and Tomasi, the optical flow is applied. In a final, step homologous key point pairs are used to calculate the position

 ISBN: 978-1-61208-203-5

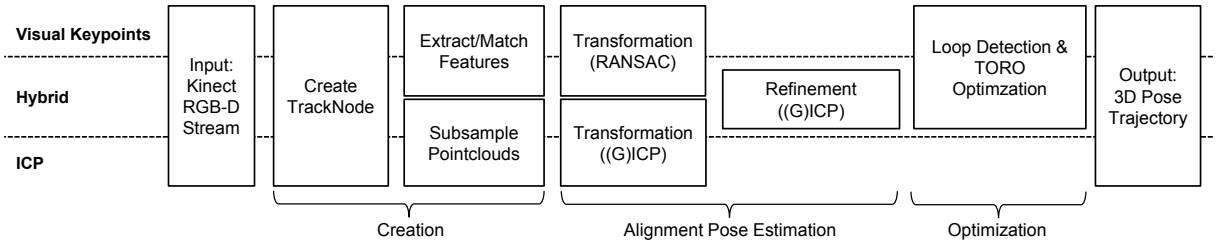

Figure 3. Overview of application flow of visual key point, hybrid and ICP approaches

transformation (RANSAC [17]). The goal of the RANSAC algorithm is to find a suitable model that describes the position transformation best. The algorithm can be described in 4 steps: 1. Select randomly sufficient homologous key point pairs. 2. Define a possible characteristic of the model. 3. Apply this model to all key points of the first picture. The key point pairs fitting this model are defined as inliers. 4. Calculate the quality of the model and decide whether the new model with the number of inliers is better than the current model. If so, the new model is now the best model. This procedure is repeated a prior defined fixed number of times, each time producing either a model which is rejected because too few inlier points were found or a better model with lower error measurement. The RANSAC is very robust to noise and measurement errors and outliers, but the number of iteration has to be limited since it is a non-deterministic approach, which may result in an imprecise or even incorrect model. Finally, the TORO (Tree-based network optimizer) optimization is performed [18]. The resulting graph of RANSAC underlies the general problem of all SLAM methods. The errors in sensor measurement cumulates over time and results in a deviation that also increases over time. In case a position is passed twice, pictures and key points can be recognized and a loop is detected. The goal of TORO is now to minimize errors of the calculated positions, which might have occurred since the time when the position was passed the first time.

The second method (see Figure 3 lower part), the ICP (Iterative Closest Point) Application Flow, uses point clouds as input to calculate position transformations. The Generalized ICP [19] takes two partly overlapping or completely identical point clouds and aligns them until they match. The algorithm works in two steps: Find correspondences between both sets of point clouds and iteratively revises the translation and rotation needed to minimize the distance between the two sets. The correspondences can be weighted either with a Point-to-Point Minimization [20] or a Point-to-Plane Minimization [21].

We also examined a hybrid application flow (see Figure 3 middle part), which works in the beginning like the visual key point application flow, but performs a refinement with the ICP in the Alignment Pose Estimation Phase. In the case not enough homologous key point pairs could be found, the algorithm immediately switches to the ICP calculation, which ensures that even in situations where visual SLAM fails a position can be calculated and gaps in the output graph prevented.

C. Evaluation platform and test environments

The evaluation platform offers several features to ensure consistent and comparable results: All algorithms must have the same input data (Kinect data stream). Hence, the platform offers a record function, where each walking path is stored into an ONI file. Each algorithm can be applied separately on that ONI file. Therefore consistent input data can be guaranteed and the performance of the SLAM approaches can be compared for one particular scenario. When algorithms are applied, duration and load are measured. Together with the results, the platform offers the functionality of exporting this data. Finally a modular comparison can be performed. Additionally, the position transformation are visualized in 2D and 3D.

To calculate the accuracy of the algorithms in different test environments, the paths are marked with tape and when passing one of those marks, the picture frame number is logged. Later the calculated position by the algorithms and the real position can be compared. To enable similar conditions between the test environments, the test person carrying the Kinect tries to hold the Kinect in the same manner for all walked paths in all scenarios.

IV. Evaluation results

We chose three different test environments to evaluate the performance of the algorithms in different scenarios and situations. The first test environment was a 7 room apartment, the second environment was an office building with connected rooms and the third environment was a subway station in Munich.

A. Test environments

In the **apartment scenario**, the visual key points approach was evaluated first. By comparing the SURF with the optical flow/KLT approach, the SURF approach outperforms the KLT (compare Figure 4). The effect of changes in the maximum distance of inliers to the model of the RANSAC algorithm were examined next. Comparing SURF and KLT, both approaches show similar effects. The best results are

 ISBN: 978-1-61208-203-5

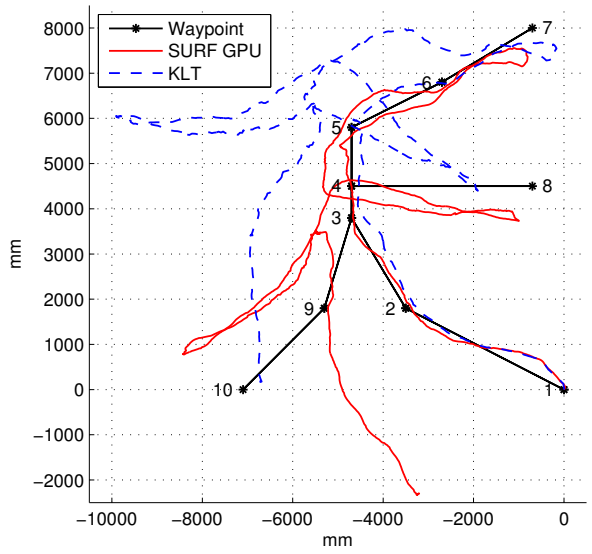

Figure 4. Resulting graphs of SURF and KLT

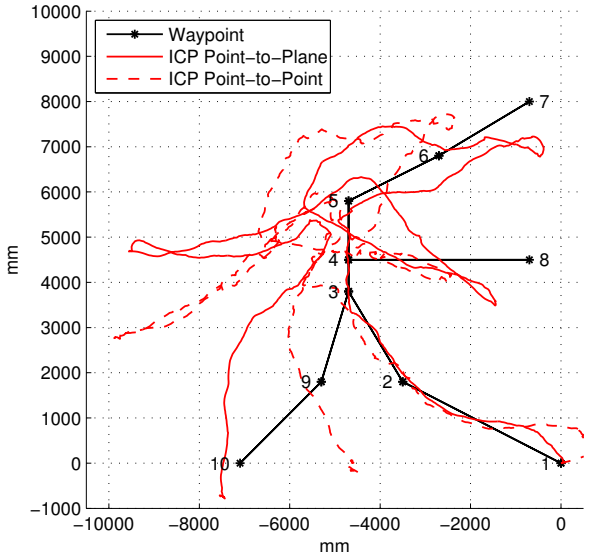

Figure 5. Point-to-Plane and Point-to-Point

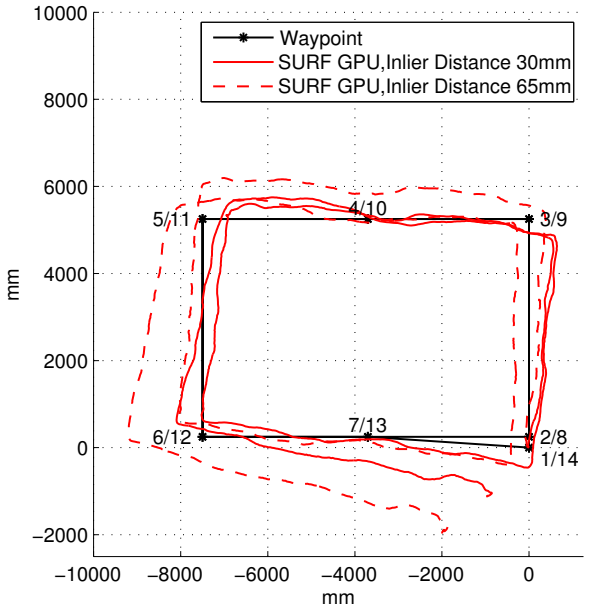

Figure 6. Two best Inlier distance values

achieved with a distance of 65 mm, higher or lower maximum values result in less accurate graphs. Applying the loop detection algorithm and TORO (where every 40th tracknode is compared to the new added) results in an enhancement of the graph. The enhancement for the KLT approach is higher than for the SURF approach. If more tracknodes are considered no significant enhancement could be measured.

Within the ICP method we compare the Point-to-Point method and the Point-to-Plane method. In Figure 5 the Point-to-Point method underlies a strong drift from the beginning on, whereas the Point-to-Plane method performs very well until the fifth waypoint. Afterwards, we examined the effect of different sizes of point clouds. Smaller variations of the size do not influence the Point-to-Plane method, whereas the Point-to-Plane method is sensitive to changes of the size. In comparison, the Point-to-Plane method is more robust and calculates good results with smaller point clouds.

In the **office scenario** the rooms were connected and the path walked outlines a closed rectangle. By varying the distance of the Inlier to the model for the RANSAC algorithm, similar results to the apartment scenario are calculated. The best two values for the maximum distance are depicted in Figure 6. Applying the loop detection algorithm and TORO results also in a enhancement of the graph. Interesting in this case is that the reduction of the track node distance from 40 to 20 in combination with the KLT and TORO, do not result in significant changes of graph accuracy. The evaluation of ICP algorithms (compare Figure 7) shows similar results to the apartment scenario.

The **subway station scenario** depicts a special scenario which differs in various aspects from the two previous scenarios. Subway stations consist of large areas and big halls. Since the range of the Kinect is limited, the test scenario was adjusted and the way the camera was positioned changed. To allow the Kinect to at least gather some depths information the Kinect was tilted towards the floor. Furthermore, bright illumination causes a lot of reflexions, which disturb the algorithms. After testing both visual methods and the ICP methods, the only approach which could calculate positions at all was the SURF approach. Both KLT and ICP method failed in the environment of the subway station.

B. Conclusions

Concluding the visual approaches, the SURF approach performed better than the KLT in all scenarios and test envi-

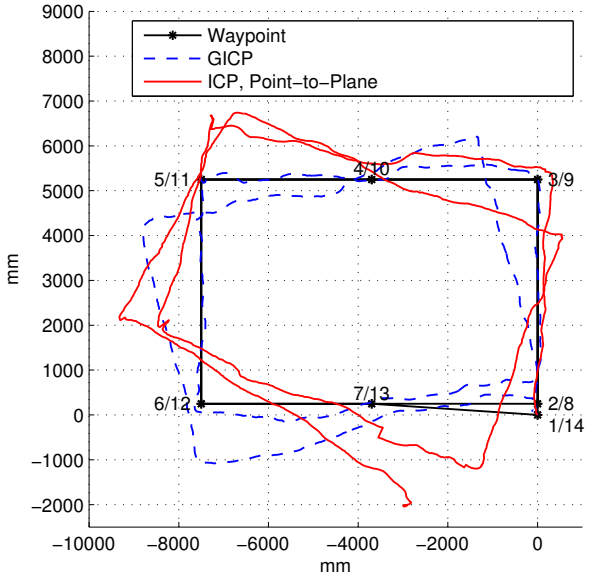

Figure 7. ICP approaches

ronments. In the test environment of the subway station, the KLT approach failed entirely because of variations of lighting, homogenous surfaces and missing depth information. TORO enhances both approaches in the apartment scenario, whereas in the office scenario SURF could be enhanced more with TORO than KLT. Varying the maximal distance between inliers and the model for the RANSAC algorithm enhanced both approaches. A standard configuration that performs equally well for all scenarios could not be found. For the SURF method, a maximal distance between 30 and 65 mm is feasible and for the KLT method between 30 and 50 mm.

Concluding the ICP approach, the Point-to-Plane Minimization method outperforms the Point-to-Plane method in the apartment scenario. An interesting aspect is the size of the point cloud. It was not the biggest point cloud that obtained more favorable results. In the Point-to-Plane alternative 3000 points achieved the best results.

Visual approaches could be further enhanced by inserting a refinement via ICP. The KLT approach reaches in each scenario the best performance in combination with the ICP, whereas for the SURF approach the data set and the test environment are crucial whether ICP can enhance the approach further more or not.

C. Overview of error rate

For the overview of the deviation in Figure 8, the best results from the apartment and the office test scenarios were accumulated and an average calculated. The scenario of the subway station was left out, since not all methods could provide feasible results.

An overview of accuracy, calculation time and robustness is given in Table I. The results of the subway scenario

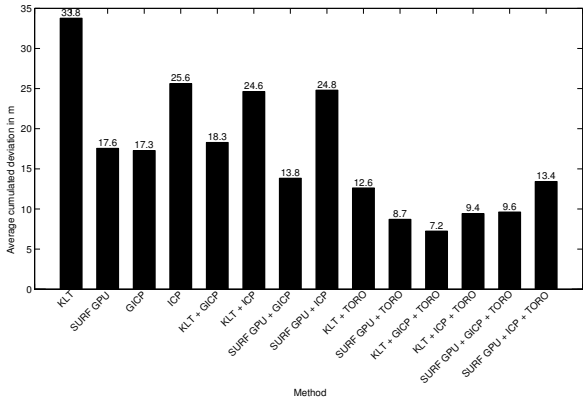

Figure 8. Average deviation (apartment and office tests)

were included in this overview. The subway scenario shows the weakness of the visual approaches. Since in subway stations the conditions are harsh (lighting changes extremely, homogenous surfaces and reflexions) key points could not always be found. The vast areas and big halls furthermore hamper SLAM methods using the Kinect.

V. Conclusion

In this work, we have shown that the Microsoft Kinect can be used for visual odometry and therefore is suitable for indoor positioning solutions in public buildings. For this purpose we tested the aptitude of state of the art techniques like SURF, RANSAC and ICP in combination with the Kinect in different scenarios. The results showed that every approach has some flaws, depending on the scenario.

For that reason, we developed a hybrid approach which makes use of visual methods as well as ICP. In order to do this, we use a customized RANSAC and then enhanced the results by additionally applying the ICP. For this purpose, we used a weight function customized for the Microsoft Kinect. All approaches were tested with an evaluation software which enabled us to test the approaches in real life environments and allowed us to record those environments for evaluations.

The results show that each the ICP and the hybrid approach usually outperform the pure visual methods inside of buildings. The scenario of the subway stations depicted a very harsh environment, where the sensors of the Kinect delivered weak data and only the SURF approach could estimate positions at all.

References

[1] C. Schindhelm, F. Gschwandtner, and M. Banholzer, “Usability of apple iphones for inertial navigation systems,” in *Proceedings of the 22nd Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Toronto, Canada, September 2011.

 ISBN: 978-1-61208-203-5

Method	∅ cummulated deviation (in m)	∅ calculating time (in ms)	robustness
KLT	33.7755	146.6980	-
SURF GPU	17.5606	106.5613	++
GICP	17.2786	273.3400	+
ICP	25.6255	200.6345	+
KLT + GICP	18.2869	342.7750	+
KLT + ICP	24.6325	310.9405	+
SURF GPU + GICP	13.8349	294.6580	++
SURF GPU + ICP	24.8067	282.1535	++
KLT + TORO	12.6179	635.1590	-
SURF GPU + TORO	8.7115	280.6925	++
KLT + GICP + TORO	7.2415	941.1915	+
KLT + ICP + TORO	9.4268	903.1700	+
SURF GPU + GICP + TORO	9.6246	538.9260	++
SURF GPU + ICP + TORO	13.4291	478.8190	++

Table I
OVERVIEW OF EVALUATION RESULTS

[2] H. Durrant-Whyte and T. Bailey, "Simultaneous localisation and mapping (slam): Part i the essential algorithms," *IEEE Robotics and Automation Magazine*, vol. 2, pp. 1–9, 2006.

[3] T. Bailey and H. Durrant-whyte, "Simultaneous localisation and mapping (slam): Part ii state of the art," *Computational Complexity*, vol. 13, no. 3, pp. 1–10, 2006.

[4] P. Robertson, M. Puyol, and M. Angermann, "Collaborative pedestrian mapping of buildings using inertial sensors and footslam," in *Proc. of the 24th International Technical Meeting of The Satellite Division of the Institue of Navigation (ION GNSS 2011)*, Portland, Oregon, September 2011, pp. 1366–.

[5] A. Garulli, A. Giannitrapani, A. Rossi, and A. Vicino, "Mobile robot slam for line-based environment representation," in *44th IEEE Conference on Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC'05.*, Seville, Spain, December 2005, pp. 2041–2046.

[6] X. S. Zhou and S. I. Roumeliotis, "Multi-robot slam with unknown initial correspondence: The robot rendezvous case," in *in Proceedings of IEEE International Conference on Intelligent Robots and Systems*, Beijing, China, October 2006, pp. 1785–1792.

[7] N. Engelhard, "Real-time 3d visual slam with a hand-held rgb-d camera," *Pattern Recognition*, vol. 2, no. c, 2011.

[8] The PrimeSense website. [Accessed Apr. 27, 2012]. [Online]. Available: http://www.primesense.com/

[9] The Willow Garage website. [Accessed Apr. 27, 2012]. [Online]. Available: http://www.willowgarage.com/

[10] The SideKick website. [Accessed Apr. 27, 2012]. [Online]. Available: http://www.sidekick.co.il/

[11] L. Gallo, A. P. Placitelli, and M. Ciampi, "Controller-free exploration of medical image data: Experiencing the kinect," in *Proceedings of the 24th IEEE International Symposium on Computer-Based Medical Systems, 27-30 June, 2011, Bristol, United Kingdom*. IEEE, 2011, pp. 1–6.

[12] M. Tolgyessy and P. Hubinsky, "The kinect sensor in robotics education," in *Proc. of 2nd International Conference on Robotics in Education (RiE 2011)*, R. Stelzer and K. Jafarmadar, Eds. Vienna, Austria: INNOC - Austrian Society for Innovative Computer Sciences, September 2011, pp. 143–146.

[13] W. K. Pratt, *Digital image processing*. New York, NY, USA: John Wiley & Sons, Inc., 1978.

[14] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proceedings of the Sixth International Conference on Computer Vision*, ser. ICCV '98. Washington, DC, USA: IEEE Computer Society, January 1998, pp. 839–.

[15] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer Vision–ECCV 2006*, pp. 404–417, 2006.

[16] D. Lowe, "Object recognition from local scale-invariant features," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 2. Kerkyra, Greece: IEEE, September 1999, pp. 1150–1157.

[17] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[18] G. Grisetti, C. Stachniss, and W. Burgard, "Non-linear constraint network optimization for efficient map learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, pp. 428–439, 2009.

[19] A. Segal, D. Haehnel, and S. Thrun, "Generalized-icp," in *Proc. of Robotics: Science and Systems*, 2009.

[20] G. Godin, M. Rioux, and R. Baribeau, "Three-dimensional registration using range and intensity information," in *Proceedings of SPIE*, vol. 2350, Boston, Massachusetts, November 1994, p. 279.

[21] Y. Chen and G. Medioni, "Object modeling by registration of multiple range images," in *Robotics and Automation, 1991. Proc., 1991 IEEE International Conference on*. Sacramento, California, USA: IEEE, April 1991, pp. 2724–2729.

 ISBN: 978-1-61208-203-5