Stereo-Based Ego-Motion Estimation Using Pixel Tracking and Iterative Closest Point

Annalisa Milella
Department of Mechanical Engineering,
Politecnico of Bari, viale Japigia 182,
70126 Bari, Italy
milella@poliba.it

Roland Siegwart
Autonomous Systems Lab,
Swiss Federal Institute of Technology
Lausanne (EPFL),
CH-1015 Lausanne, Switzerland
roland.siegwart@epfl.ch

Abstract

In this paper, we present a stereovision algorithm for real-time 6DoF ego-motion estimation, which integrates image intensity information and 3D stereo data in the well-known Iterative Closest Point (ICP) scheme. The proposed method addresses a basic problem of standard ICP, i.e. its inability to perform the segmentation of data points and to deal with large displacements. Neither a-priori knowledge of the motion nor inputs from other sensors are required, while the only assumption is that the scene always contains visually distinctive features which can be tracked over subsequent stereo pairs. This generates what is usually called Visual Odometry. The paper details the various steps of the algorithm and presents the results of experimental tests performed with an all-terrain mobile robot, proving the method to be as accurate as effective for autonomous navigation purposes.

1. Introduction

For a mobile robot to be autonomous, it must be able to self-localize while moving in its operational environment. Decades of research in the field of Automatic Vehicle Guidance have, in fact, largely proved that accurate localization is critical for most navigation-related tasks. Several methods have been developed that can be classified into two groups: relative positioning and absolute positioning.

The most commonly used relative positioning technique is known, in the robotics community, as dead-reckoning [16]. It mainly relies on odometry and Inertial Measurement Units (IMU). Active beacons, landmark-based navigation, and GPS are, instead, examples of absolute positioning systems [2].

In this paper, we deal with an emerging and promising localization method, usually referred to as Visual Odometry, i.e. motion estimate from visual input only [14]. The key idea of visual odometry is that of estimating the motion of the robot by visually tracking landmarks, opportunely selected in the environment, using an on-board camera. This technique, originally developed by Matthies [11], is nowadays considered as a "middle ground" between dead-reckoning and global localization [17].

In the very last years, a number of visual odometry algorithms have been proposed, using either single cameras [3, 4, 14, 18] or stereo vision [6, 10, 14, 17], which mainly differ depending on the feature tracking method and on the transformation applied for estimating the camera motion. For instance, in [14], robust visual motion estimation is achieved using preemptive RANSAC [13], followed by iterative refinement. In [17], odometry provides an estimation of the approximate robot motion that allows selecting a search area for improved feature tracking. A maximum-likelihood formulation is employed for motion computation. Finally, in [18], the visual module uses a variation of Benedetti and Perona's algorithm for feature detection, and correlation for feature tracking. Robustness is obtained integrating visual data and IMU by a Kalman filter.

The growing interest in vision-based navigation strategies is due to several reasons. First of all, video sensors allow the vehicle to self-localize while performing also other critical navigation tasks, such as detecting and avoiding obstacles or reaching a predefined target. At the same time, a huge amount of information about the environment can be acquired for exploration and map building, performing what is

Proceedings of the Fourth IEEE International Conference on Computer Vision Systems (ICVS 2006)
0-7695-2506-7/06 $20.00 © 2006 IEEE

usually referred to as Simultaneous Localization and Mapping (SLAM) [7, 21]. Like dead-reckoning, visual odometry accumulates error over time; nevertheless, it has been proved that it allows more accurate results for most sensor combinations if compared to dead-reckoning [17, 18]. In addition, video sensors are less expensive and more flexible than other sensors, such as laser scanners, traditionally employed in SLAM applications [15].

Here, an algorithm for real-time 6DoF ego-motion estimation is proposed, which integrates image intensity information and 3D stereo data in the well-known Iterative Closest Point (ICP) scheme.

The main application of ICP, as originally introduced by Besl [1], is the registration of digitized data from a rigid object with an idealized geometric model. The method is particularly suited for aligning point clouds where the correspondences are not known, and consists of a two-step kernel: the first step searches for corresponding points between the two point clouds, based on the nearest neighbors concept; the second step determines the transformation that minimizes the distance between the nearest neighbors [12]. The process is iterated until a convergence criterion is satisfied. This method has been extensively studied in literature and many variants have been proposed to both improve accuracy and reduce computational time [5, 19, 26]. Several applications have been developed in the field of surface registration and mapping, mainly based on laser scanners data. However, relatively little work has been published in the domain of ICP-based visual odometry [12].

In this paper, the potentialities of ICP for visual odometry are investigated, using stereo vision. Specifically, two basic problems of ICP are addressed: the susceptibility to gross outliers, and the failure when dealing with large displacements. As an extension of these issues, another drawback of ICP is its inability to segment input data [1]. Typical solutions use odometry information for predicting the displacement between consecutive frames and providing initial motion estimate before ICP registration [23]. Conversely, the method described here allows overcoming both problems, using the information deriving from a single stereo device, without previous knowledge of the motion. The only assumption is that the scene always contains visually distinctive features which can be tracked over subsequent images.

The method can be summarized as follows. First of all, for each acquired stereo pair, a dense disparity map is generated, employing an area correlation algorithm [8]; then, interesting pixel points are selected in the left image, based on the Shi-Tomasi feature detector [22]. Only the visual landmarks with an associated high stereo-confidence level 3D point are retained. Potential matches between two consecutive frames are established using image intensity information and are exploited to obtain approximate motion estimate. Finally, Zhang's implementation of ICP [26] is applied for refinement.

A similar approach is employed in [7] for camera motion estimation prior to 3D environment modeling. Iterative methods combining intensity and 3D information can also be found in [20] for map building and in [25] for the registration of 3D partial surface models.

This work focuses instead on the visual odometry issue. Various image processing and 3D registration techniques are efficiently combined for improving outlier rejection in both stereo matching and feature tracking, so that accurate motion estimates can be achieved though using a few interesting points and preserving real-time constraints.

Experimental results obtained with an all-terrain rover, the Shrimp mobile robot [9], equipped with a Videre Design stereo head, are presented. Tests were performed both on a flat surface in a typical office-like indoor environment and on a simulated rocky soil, proving the effectiveness of the method for autonomous navigation in different contexts.

The rest of the paper is structured as follows. Section 2 details the various steps of the proposed algorithm, illustrating a sample case. Section 3 shows experimental results with the Shrimp robot. Section 4 contains the conclusions of the presented work.

2. Description of the method

In this section, an algorithm for real-time 6DoF ego-motion estimation is presented, which enables a robot to self-localize using only the data acquired by a stereo head, mounted on-board.

The method combines intensity and 3D information in the well-known Iterative Closest Point (ICP) scheme and allows overcoming two basic problems of ICP: the susceptibility to gross statistical outliers, and the failure when dealing with large displacements. As an extension of these issues, another drawback of ICP is addressed, i.e. its inability to perform the segmentation of input data points: if data points from two shapes are intermixed and matched against the individual shapes, registration fails [1].

These limitations are intrinsic in ICP basic concept and become particularly restrictive for robot self-localization and navigation purposes, as, while the sensor moves, different parts of the scene become occluded and, conversely, new objects may appear.

Proceedings of the Fourth IEEE International Conference on Computer Vision Systems (ICVS 2006)
0-7695-2506-7/06 $20.00 © 2006 IEEE

Therefore, vast regions may be present in only one of two consecutive point clouds, and, if an outlier region is too close to a valid region, there is no possibility for ICP to perform a correct matching process [12].

The method presented in this work involves three main phases: 1) Feature selection, 2) Feature tracking, and 3) Motion estimation. In the remainder of this section, each phase is discussed in detail. Results of a test case are also shown to illustrate how the various steps work.

2.1. Approach

Feature selection. The algorithm starts by acquiring a stereo pair and generating a dense disparity map to obtain 3D points. The SRI Stereo Engine algorithm is employed [8]. It consists of an area correlation-based matching process, followed by a post-filtering operation that uses a combination of a confidence filter and left/right check to reject areas with insufficient texture, where bad matches are very likely to appear.

The Shi-Tomasi feature detector [22] is then applied to the left image to select interesting pixel points. Only the pixels with an associated high stereo-confidence level 3D point are retained for further processing.

Two point clouds are in the end available for each stereo pair: the pixel point cloud and its associated 3D point cloud.

Feature tracking. The tracking of visual landmarks between consecutive frames is performed using a normalized cross-correlation-based algorithm.

Let us denote with {L1} and {L2} the visual landmarks detected in two successive left images. Each point in {L1} is paired with the point in {L2} that generates the maximum normalized cross-correlation coefficient in a 5x5 pixels window centered at the point. To speed up and improve the searching process, only features within a certain pixel distance from each other are matched. A minimum value for the correlation coefficient is also established.

False matches are then rejected using two strategies: the mutual consistency check and robust statistics. The former consists in applying the cross-correlation-based pairing from both {L1} to {L2} and {L2} to {L1}; only pairs that mutually have each other as preferred mate are accepted as valid matches [14] and are stored together with their correlation value. A final selection is accomplished based on the median [7] and the standard deviation from median of the computed correlation coefficients; pairs whose correlation differs from the median by more than two times the standard deviation from median are rejected.

This process takes two principal advantages: first of all, features which do not belong to both frames are discarded, i.e. a segmentation of the input data is performed; furthermore, a set of corresponding 3D points is selected which can be used for the successive motion estimation stage.

Motion estimation. The problem of estimating the motion that the camera has undergone between two consecutive stereo acquisitions can be expressed as finding the 3D transformation matrix T that minimizes the mean-squares objective function:

$$F(T) = \frac{1}{N}\sum_{i=1}^{N}\left\| TP_2^i - P_1^i \right\|^2 \qquad (1)$$

where P_1^i and P_2^i indicate corresponding 3D points at two successive time instants, and N is the number of pairs. A first estimate of T is performed based on the correspondences found in the cross-correlation pairing process. This provides initial approximate motion estimation [7].

Finally, ICP registration is applied. At each iteration, 3D point pairs are generated based on point-to-point distance metric. The rejection scheme proposed by Zhang [26] is employed, which allows setting adaptively the value of the maximum distance between corresponding points using the statistics of the distances. Least-squares rotation and translation are computed using the dual number quaternion method [24]. The process stops when the change in motion estimate between two successive iterations is less than 1%.

2.2. A sample case

Here, results for a test case are reported as an example. In this experiment, the algorithm is applied to 320x240 px stereo images, after the camera has undergone a pan rotation of 10°.

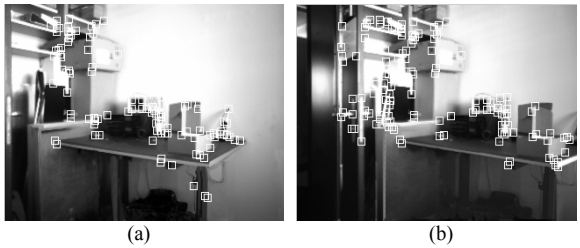
Figure 1.a and 1.b show the left frames of the two stereo pairs with the detected visual landmarks. Each feature has an associated 3D point. Once the features in two consecutive stereo pairs have been selected, the problem of finding corresponding points has to be solved. This is done using both pixel intensity and 3D stereo information.

In Figure 2.a, the left image before rotation is shown along with the correspondences determined using intensity information. Features at a distance of 100 pixels are matched and a correlation threshold of 0.85 is fixed. False matches are still present; that indicates the necessity of a refinement process. Nevertheless, correspondences established based on

pixel intensity information can be employed to obtain a first motion estimate. Stereo data are then used, applying ICP. Final pairs are plotted in the image plane in Figure 2.b.

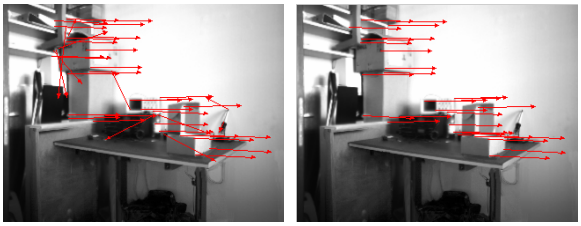
After seven iterations, the absolute position errors remain stable at 0.78 cm along the pan axis (x), 2.8 cm along the tilt axis (y), and 0.68 cm along the swing axis (z), while the absolute errors in rotation are of 1.10°, 0.39°, and 0.04° for pan, tilt and swing angles, respectively.

In Figure 3, final selected 3D pairs are displayed, before and after registration. Conversely, Figure 4 reports the result obtained by applying ICP directly to the 3D point clouds, without previous processing. Evidently, no good motion estimate would be achieved.



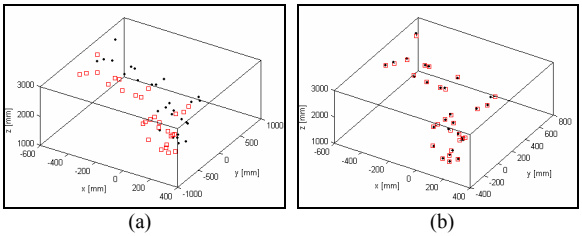
(a) (b)

Figure 1. Left images before (a) and after (b) rotation, with selected features superimposed



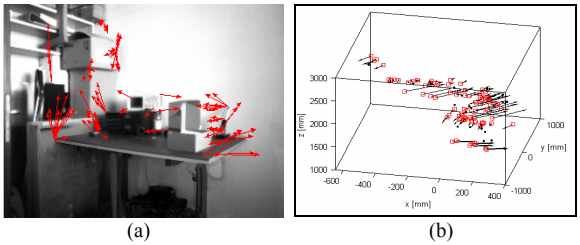
(a) (b)

Figure 2. Corresponding features after correlation-based tracking (a), and at the end of ICP refinement (b)



(a) (b)

Figure 3. Final pairs in 3D space before (a) and after (b) registration, using correlation and ICP. At the end, the red square points overlap the corresponding black round points



(a) (b)

Figure 4. Final pairs estimated applying ICP directly to the 3D point clouds, re-projected onto the image plane (a) and in 3D space (b). In (b), black arrows indicate the positions reached by the red square points after ICP registration. Evidently, no good motion estimate is achieved

3. Experimental results

The method was tested using the Shrimp robot, equipped with a Videre Design stereo head (see Figure 5). The Shrimp is an off-road rover characterized by a passive non-hyperstatic structure which makes it able to adapt to a large range of obstacles. It has six motorized wheels and is composed of four main parts: the body, the articulated front fork and the two side bogies. This rover is able to overcome steps of twice its wheel diameter and can climb regular stairs. More details can be found in [9].

Several tests were performed, in order to verify the effectiveness of the method for different motion conditions and environments. Here, results of three different tests are presented.

In the first test, the robot was guided on a flat surface, in a typical office-like indoor environment (Figure 6.a). The other two tests were performed on a simulated rocky surface (Figure 6.b). In all the experiments, the robot was driven at 6 cm/s. 3D information is referred to a reference frame attached to the chassis of the robot, as shown in Figure 5. The algorithms were developed in C++ language code. A PC with 2.40 GHz processor and 256MB RAM was employed. Detailed results of the experiments are reported in the rest of this section.

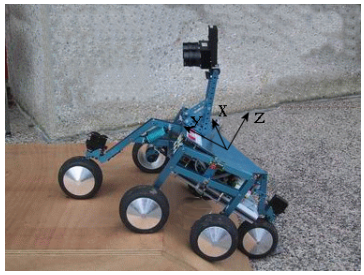


Figure 5. The Shrimp mobile robot

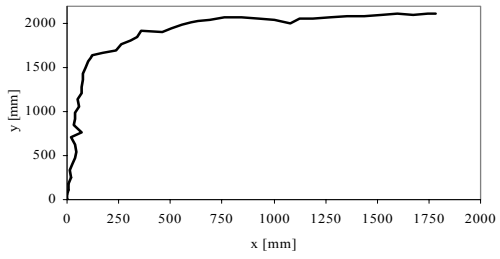
(a) (b)

Figure 6. Indoor test environment (a); Simulated rocky environment (b)

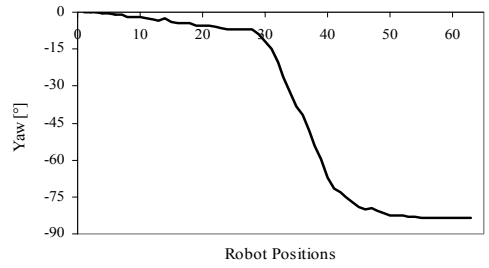
Flat surface. The first test was performed with the robot moving on a flat surface, in a typical office-like indoor environment. The ability of the system to reach a target position was evaluated, guiding the robot through an L-shaped path of 1780 (x) x 2200 (y) mm to a predefined location. Five runs were executed. The graphs in Figure 7.a and 7.b show, respectively, the estimated trajectory and the variation of the yaw angle during one run. The i-th percentage errors ($e^i_{px\%}$, $e^i_{py\%}$) are defined as:

$$e^i_{px\%} = 100 \cdot \left| \frac{p_{Tx} - p^i_{ex}}{p_{Tx}} \right| \qquad e^i_{py\%} = 100 \cdot \left| \frac{p_{Ty} - p^i_{ey}}{p_{Ty}} \right| \qquad (2)$$

where $[p_{Tx}, p_{Ty}]$, denotes the position of the target, and $[p^i_{ex}, p^i_{ey}]$ is the estimated final position of the robot at the i-th run. The computed mean percentage errors and corresponding standard deviations are of 1.9±2.2% along x and of 2.4±1.8% along y.



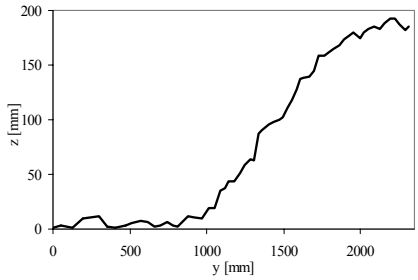
(a)



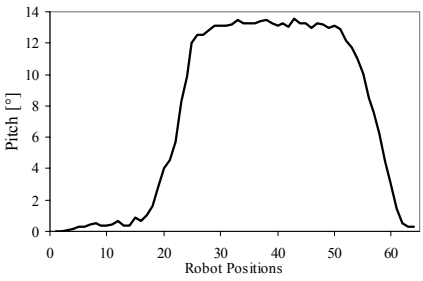
(b)

Figure 7. Estimated robot L-shaped trajectory (a) and corresponding yaw angle variation (b)

Ramp trajectory. In this test, the robot moves on a rocky surface. After a forward displacement, it climbs a ramp of about 12° of inclination to reach a target position located at a distance of 2200 mm along y, at a quote of 200 mm over the initial position of the robot. The test was repeated five times. Figure 8.a and 8.b show, respectively, the trajectory in the (y-z) plane and the pitch angle variation during one run. Mean percentage errors and standard deviations of 2.4±1.9% along y and of 6.0±4.7% along z were computed, using the definitions in (2) for the (y-z) plane.



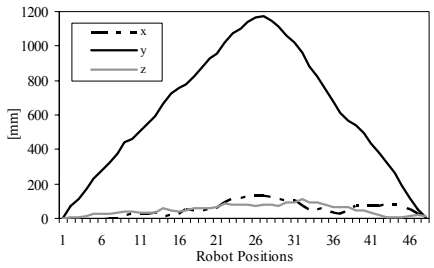
(a)



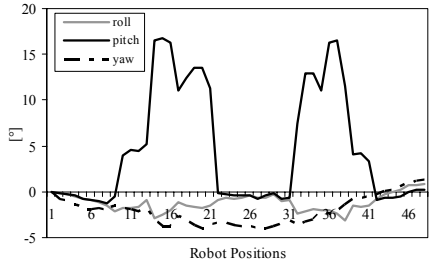
(b)

Figure 8. Estimated robot ramp trajectory (a) and corresponding pitch angle variation (b)

Step trajectory. In this test, the robot was guided to overcome two consecutive steps of 50 mm and 100 mm, at first moving forward for 1100 mm and then backward to the start position. Here, the variations of all the six degrees-of-freedom of the vehicle can be clearly observed, as shown in Figure 9.a and 9.b, representing respectively, the estimated 3D positions and the Euler angles during one test. Ten runs were executed. In each run, the robot started at a marked location and was driven back to the same location. The discrepancy between the actual robot position and the estimated position is the so-called Return Position Error (RPE) [16]. The following mean absolute RPE and corresponding standard deviation was computed at each spatial direction: 2.6±3.3 cm (x), 3.1±2.8 cm (y), 6.2±3.8 cm (z).



(a)



(b)

Figure 9. Estimated 3D positions (a) and Euler angles (b) during the step test

4. Conclusions

In this paper, a stereovision algorithm for real-time 6DoF ego-motion estimation was presented. The method integrates image intensity and 3D stereo information in the well-known Iterative Closest Point scheme, and allows overcoming two basic problems of standard ICP, i.e. its failure in presence of gross outliers and its inability to segment the input data points.

First of all, the proposed approach was introduced. Then, a sample case was examined to illustrate how the various steps of the method work. Finally, experimental tests with an all-terrain rover were presented, performed on both a flat surface and a rock-like soil, proving the algorithm to be accurate and effective for visual odometry.

5. References

[1] P.J. Besl and N.D. McKay, "A Method for Registration of 3-D Shapes", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 14, No. 2, February 1992, pp. 239-256.

[2] J. Borenstein, B. Everett, and L. Feng, *Navigating Mobile Robots: Systems and Techniques*, A. K. Peters, Ltd., Wellesley, MA, ISBN 1-56881-058-X, 1996.

[3] P.I. Corke, D. Strelow, and S. Singh, "Omnidirectional Visual Odometry for a Planetary Rover", *Proceedings of IROS 2004*, Japan, 2004.

[4] A.J. Davison, "Real-Time Simultaneous Localization and Mapping with a Single Camera", *IEEE Int. Conf. on Computer Vision*, Nice, 2003, pp. 1403-1410.

[5] J. Diebel, K. Reuterswärd, S. Thrun, J. Davis, and R. Gupta, "Simultaneous Localization and Mapping with Active Stereo Vision", *IEEE/RSJ Conf. on Intelligent Robots and Systems (IROS)*, Japan, 2004.

[6] M. Dunbabin, K. Usher, and P. Corke, "Visual motion estimation for an autonomous underwater reef monitoring robot", *Field and Service Robotics Conference (FSR 2005)*, Port Douglas, Qld., 2005, pp. 57-68.

[7] M.A. Garcia and A. Solanas, "3D Simultaneous Localization and Modeling from Stereo Vision", *Proceedings of the 2004 IEEE International Conference on Robotics & Automation*, New Orleans, LA, 2004, pp. 847-853.

[8] K. Konolige, "Small Vision Systems: Hardware and Implementation", *8th International Symposium on Robotics Research*, Japan, 1997.

[9] P. Lamon and R. Siegwart, "Inertial and 3D-odometry fusion in rough terrain – Towards real 3D navigation", *International Conference on Intelligent Robots and Systems*, Japan, 2004.

[10] A. Mallet, S. Lacroix, and L. Gallo, "Position Estimation in Outdoor Environments using Pixel Tracking and Stereovision", *IEEE Int. Conf. on Robotics and Automation*, San Francisco, CA, USA, 2000, pp. 3519-3524.

[11] L.H. Matthies, *Dynamic Stereo Vision*, PhD thesis, Carnegie Mellon University, 1989.

[12] I.A.D. Nesnas, M. Bajaracharya, R. Madison, E. Bandiari, C. Kunz, M. Deans, and M. Bualat, "Visual Target Tracking for Rover-based Planetary Exploration", *Proceedings of the 2004 IEEE Aerospace Conference*, Big Sky, Montana, 2004.

[13] D. Nistér, "Preemptive RANSAC for Live Structure and Motion Estimation", *IEEE International Conference on Computer Vision*, Nice, 2003, pp. 199-206.

[14] D. Nistér, O. Naroditsky, and J. Bergen, "Visual Odometry", *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Volume 1, 2004, pp. 652-659.

[15] A. Nüchter, K. Lingemann, J. Hertzberg, and H. Surmann, "Heuristic-Based Laser Scan Matching for Outdoor 6D SLAM", *in KI 2005: Advances in Artificial*

Intelligence, 28th Annual German Conference on AI, Proceedings Springer LNAI, Koblenz, Germany, 2005.

[16] L. Ojeda, G. Reina, and J. Borenstein, "Experimental Results from FLEXnav: An Expert Rule-based Dead-reckoning System for Mars Rovers", *IEEE Aerospace Conference*, Big Sky, MT, USA, 2004.

[17] C.F. Olson, L.H. Matthies, M. Schoppers, and M.W. Maimone, "Rover Navigation Using Stereo Ego-Motion", *Robotics and Autonomous Systems*, 43, 2003, pp. 215-229.

[18] S. I. Roumeliotis, A.E. Johnson, and J.F. Montgomery, "Augmenting Inertial Navigation with Image-Based Motion Estimation", *Proceedings of the 2002 IEEE International Conference on Robotics & Automation*, Washington, 2002, pp. 4326-4333.

[19] S. Rusinkiewicz and M. Levoy, "Efficient Variants of the ICP Algorithm", *Proceedings of IEEE 3DIM*, Canada, 2001, pp. 145-152.

[20] J.M. Saéz and F. Escolano, "A global 3D map-building approach using stereo vision", *Proceedings of the 2004 IEEE International Conference on Robotics and Automation,* 2004, pp. 1197-1202.

[21] S. Se, D. Lowe, and J. Little, "Vision-Based Mobile Robot Localization and Mapping Using Scale-Invariant Features", *ICRA 2001*, 2001, Korea, pp. 2051-2058.

[22] J. Shi and C. Tomasi, "Good Features to Track", *IEEE Conference of Computer Vision and Pattern Recognition*, CA, 1994, pp. 593-600.

[23] H. Surmann, A. Nüchter, and J. Hertzberg, "An Autonomous Mobile Robot with a 3D Laser Range Finder for 3D Exploration and Digitalization of Indoor Environments", *Journal Robotics and Autonomous Systems*, Vol. 45, 2003, pp. 181-198.

[24] M.W. Walker, L. Shao, and R.A. Volz, "Estimating 3-D Location Parameters using Dual Number Quaternions", *CVGIP: Image Understanding*, 54, 1991, pp. 358-367.

[25] S. Weik, "Registration of 3-D Partial Surface Models using Luminance and Depth Information", *Proceedings of the First International Conference on Recent Advances in 3-D Digital Imaging and Modeling*, 1997.

[26] Z. Zhang, "Iterative Point Matching for Registration of Free-Form Curves", *IRA Rapports de Recherche N° 1658 Programme 4 Robotique, Image et Vision*, 1992.