10.1590/s1982-21702018000300023

*ARTICLE*

# A JOINT EFFORT OF SPEEDED-UP ROBUST FEATURES ALGORITHM AND A DISPARITY-BASED MODEL FOR 3D INDOOR MAPPING USING RGB-D DATA

## *Combinação do algoritmo SURF e de uma abordagem baseado em valores de disparidade para mapeamento indoor 3D usando dados RGB-D*

Marcos Aurélio Basso[1] - ORCID: 0003-1245-2509

Daniel Rodrigues dos Santos[2] - ORCID: 0000-0001-7977-7426

[1]Universidade Federal Rural do Rio de Janeiro, Engenharia, Seropédica - RJ, Brasil.
E-mail: marcosbasso@ufrrj.br

[2]Universidade Federal do Paraná, Geomática, Curitiba - PR, Brasil.
E-mail: danielsantos@ufpr.br

Received in May 5th, 2017

Accepted in March 20th, 2018

### *Abstract:*

In this paper, we present a method for 3D mapping of indoor environments using RGB-D data. The contribution of our proposed method is two-fold. First, our method exploits a joint effort of the speed-up robust features (SURF) algorithm and a disparity-to-plane model for a coarse-to-fine registration procedure. Once the coarse-to-fine registration task accumulates errors, the same features can appear in two different locations of the map. This is known as the loop closure problem. Then, the variance-covariance matrix that describes the uncertainty of transformation parameters (3D rotation and 3D translation) for view-based loop closure detection followed by a graph-based optimization are proposed to achieve a 3D consistent indoor map. To demonstrate and evaluate the effectiveness of the proposed method, experimental datasets obtained in three indoor environments with different levels of details are used. The experimental results shown that the proposed framework can create 3D indoor maps with an error of 11,97 cm into object space that corresponds to a positional imprecision around 1,5% at the distance of 9 m travelled by sensor.

**Keywords**: RGB-D data; SURF algorithm; Disparity-to-plane model; Loop closure; Graph optimization.

**How to cite this article:** Basso, M. A. and Santos, D.R. A joint effort of speeded-up robust features algorithm and a disparity-based model for 3D indoor mapping using RGB-D data. *Bulletin of Geodetic Sciences*, Vol. 24 (3), 351-366, Jul-Sept, 2018.

*Resumo:*

Neste trabalho é apresentado um método para mapeamento 3D de ambientes internos usando dados RGB-D. Há basicamente dois aspectos importantes a serem discutidos neste trabalho: 1) o método explora um esforço conjunto entre o algoritmo SURF e um modelo matemático baseado em uma abordagem disparidade-a-plano para registo de nuvens de pontos. Uma vez que a tarefa de registro das nuvens de pontos acumula erros, ocorrem erros de fechamento; 2) A matriz de variância-covariância que descreve a incerteza dos parâmetros de transformação (rotação 3D e translação 3D) é usada para a deteção de erro de fechamento na tarefa de otimização baseada em grafos. Para demonstrar e avaliar a eficácia do método proposto foram usados conjuntos de dados experimentais obtidos em três ambientes internos com diferentes níveis de detalhes. Os resultados experimentais mostram que o método proposto pode criar mapas 3D com um erro de 11,97 cm que corresponde a uma acurácia posicional em torno de 1,5% da distância de 9 m percorrida pelo sensor.

***Palavras-chave**:* Dados RGB-D; algoritmo SURF; Modelo de disparidade; Erro de fechamento; Otimização por grafos.

# 1. Introduction

3D mapping of indoor environments is an important task in many engineering applications, such as, simultaneous localization and mapping systems (SLAM), surveillance and emergency managements, navigation, positioning, robotics, forensics, virtual tours, crisis management, modeling, infrastructure inspections, urban design and others. Basically, the most important existing 3D mapping of indoor environments method using RGB-D data is based on three main steps (Henry et al, 2012): 1) pairwise registration; 2) loop closure detection; 3) global optimization. The pairwise registration task aims to estimate the transformation parameters between pairs of point clouds. The most popular approach to registering point clouds is the iterative closest point (ICP) algorithm (Besl and McKay, 1992). Pairwise registration task is prone to error due to the random error of individual points, leading to incorrectness in the 3D map, as described by Khoshelham et al. (2013). To handle with registration errors the loop closure detection should be used (Du et al., 2011). Loop closure corresponds to the global adjustemnt for simultaneous refinement of all the transformation parameters in a sequence. Once transformation parameters from registration task give us some constraints (poses between point clouds, adjacents or not), we can represent these constraints using a graph structure. Basically, there are three main graph optimization approaches proposed in the literature: tree-based network optimizer (TORO), idealized by Grisetti et al. (2007); general framework for graph-based optimization (G2O), proposed by Kümmerle et al. (2011); and sparse bundle adjustment (SBA), proposed by Lourakis and Argyros (2009). In this paper, the graph optimization of the complete data sequence is performed using the graph-based optimization to minimize the registration errors.

Nowadays, RGB-D sensors are quite useful solution to build colored 3D indoor maps because it can exploit both the visual and the depth information. Its advantages compared with LASER scanning sensors are the lightweight, the low cost and it is much more flexible (Dos Santos et al., 2016). In this paper we propose a method for 3D indoor mapping using RGB-D data. The contribution of our proposed method is two-fold. First, we propose a joint effort of speed-up robust features (Bay et al., 2008) and a disparity-based model (Dos Santos et al., 2016) to include additional constraints in the graph describing the coarse-to-fine registration between RGB-D data. Second, we investigate the variance-covariance matrix that describes the uncertainty of the transformation parameters (3D rotation and 3D translation) to weight the graph-based optimization.

The paper proceeds with the related work in the Section 2. In Section 3, the proposed method for 3D indoor mapping is described. Experimental evaluation of the method is presented in Section 4. The paper concludes with a discussion in Section 5.

## 2. Related Work

Basically, 3D indoor mapping method is based on two key problems: pairwise registration and global registration with loop closure.

**Table 1:** Summarized description of the advantages and disadvantages of both the pairwise registration and graph optimization methods.

| Method | Advantages | Disadvantages |
|---|---|---|
| Point cloud registration | This approach uses salient feature surfaces, such as: points; planes; lines or free-form surfaces. | The method requires an accurate initial transformation and high overlap area between the point clouds and in the most situations it is time-consuming. |
| Combined depth and intensity values | This method integrates both the depth and the intensity information in 3D registration. | Low resolution of intensity images delivered by both LASER scanning. Range devices provide inaccurate feature location and it is highly time-consuming. |
| Global registration with loop closure | Loop closure ensures that new constraints can be inserted in the graph optimization. The global registration minimizes the registration errors. | In most cases the uncertainty of the transformation parameters is not employed to weight the optimization process. Usually, the refinement of the pairwise registration not always held. |

We will cover the main related work about the techniques in our study by grouping them into three main different categories: (1) **point cloud registration:** One of the most popular method is the ICP (iterative closest point) algorithm for registration of 3D shapes developed by Besl and McKay (1992). According Henry et al. (2012) "ICP has been shown to be effective when the two point clouds are already nearly aligned". (2) **Combined depth and intensity values:** There are many combined depth and intensity extraction algorithms proposed in recent years. Al-Manasir and Fraser (2006) presented a method that uses bundle adjustment of images taken by a digital camera to register the LASER point cloud data rather than ICP algorithm. Another interesting method was given in Barnea and Filin (2008) which suggested the integration of both the depth values and the intensity information in pairwise registration process. (3) **Global registration with loop closure:** Loop closure is a technique used to recognize when the sensor has returned to a previously visited location. It is useful to insert constraints into graph optimization task. Henry et al. (2012) used a subset of the registered frames to keep the graph relatively sparse. The loop closure is detected based on the number of visual features stablished. In other words, a loop closure is detected if sufficient amount of correspondences is correctly determined. This is done, by means of a joint effort between RGB-D data and the ICP algorithm. Once loop closure is detected, the tree-based network optimizer is activated. In Chow et al. (2014) the ICP algorithm is carried out in a model-to-scene fashion to handle the loop closure problem. The authors used the sparse bundle adjustment to obtain a consistent 3D indoor map. A summarized description of the advantages and disadvantages of the before mentioned methods is presented in Table 1. Note that our objective is both the pairwise registration and the global registration tasks. We first propose a coarse-to-fine registration task using a joint optimization algorithm combining visual features to carry out a coarse registration and a disparity-based model to realize the fine registration step, instead of only apply a coarse registration step as done by Henry et al. (2012). Second, we use the variance-covariance matrix that describes the uncertainty of the transformation parameters to weight the graph-based optimization task. In our method, the variance-covariance matrix is obtained by minimizing disparity-to-plane distances using iterative least-squares estimator, instead of accept it as identity matrix or using the Cholesky solver such as introduced by Kümmerle et al. (2011).

## 3. Method

This section explains the proposed method for 3D mapping of indoor environments using RGB-D data. In particular it combines a speed-up robust features algorithm with a disparity-to-plane model to find refined transformation parameters. Then, new constraints are inserted using the loop closure procedure and the registration errors are minimized using a graph-based optimizer. The result of this framework is a globaly consistent 3D indoor map. The Figure 1 presents the generic structure of our proposed method.

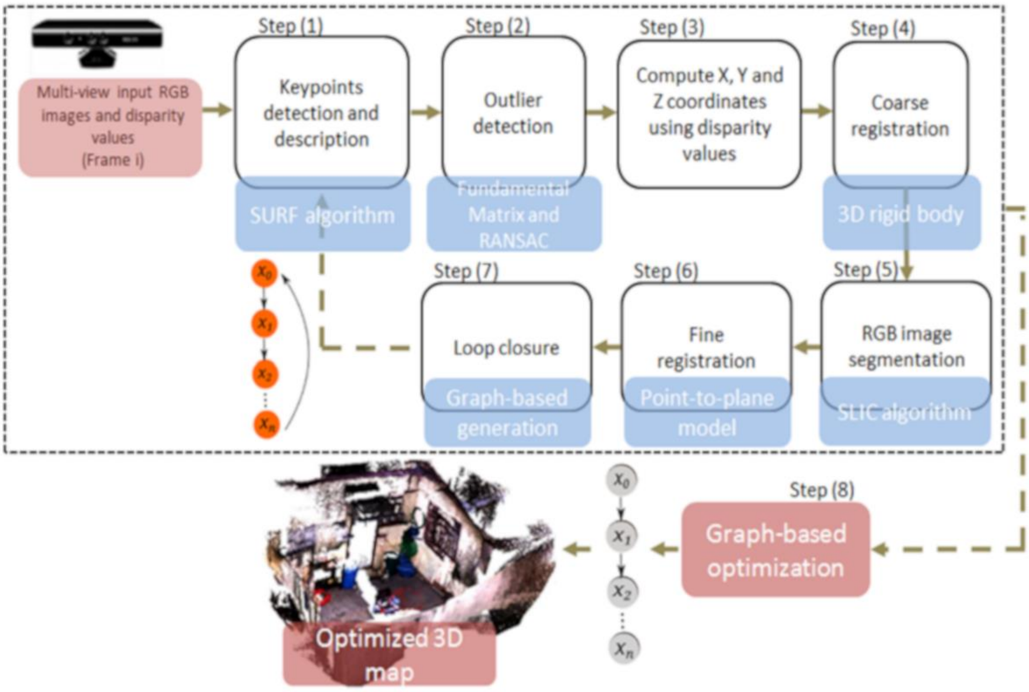

Figure 1: Generic structure of the proposed method.

Figure 1 shows the proposed method for 3D mapping of indoor environments using RGB-D data. Our method is divided in three main parts: 1) pairwise registration; 2) loop closure detection; 3) global registration. Given RGB-D data is obtained with Kinect sensor. The before mentioned tasks are described, as follows.

## 3.1 Pairwise registration

In this work, pairwise registration problem aims to estimate the transformation parameters between pairs of RGB-D frames. As before mentioned, a coarse-to-fine registration task using a joint optimization algorithm combining visual features to compute initial approximation of the transformation parameters (3D rotation and 3D translation) and a disparity-based model to realize the fine registration step is proposed. Herein, the main idea is performing the pairwise registration using RGB-D data. This task is separated in six steps, as shows Figure 1. First, keypoints (visual features) should be detected and their correspondences obtained. The scale invariant feature transform (SIFT) algorithm, developed by Lowe (2004) is one of the most used feature point detection and feature descriptor algorithm. However, is relatively time consuming in terms of reliability and robustness (Wu et al., 2013). It is divided in three stages: a) keypoints extraction and detection; b) description; and 3) matching. The general framework of SURF algorithm can be found in Bay et al. (2008). Basically, the focus is producing a feature descriptor that allows quick and highly discriminatory assessments with other features. Second, the outliers are detected using the fundamental matrix ($F$) and RANSAC algorithm. The RANSAC computes $F$ using the selected the normalized eight-point algorithm (Hartley and Zisserman, 2003). Then, the fitness for all points putatively matched is computed. If the fitness is better than initial $F$, replace $F$ with fitness and the number of random trials ($N$) is updated for every iteration in the RANSAC algorithm (Fischler and Bolles, 1981). The results it is a set of matched keypoints. Third, in order to associate the keypoints with their corresponding depth values we used the method proposed by Dos Santos et al. (2016). Firstly, depth values are computed using the formulation introduced in Khoshelham and Elberink (2012), as follows:

$$z_i = \frac{1}{c_0 + c_1 d} \tag{1}$$

intercept of the line, respectively and $d$ the denormalized disparity value. After, the correct associated 3D point is search using the epipolar geometry concept. Assuming that the shift of the depth values is sufficient to align the depth image with the RGB frame, as described in Henry et al. (2012), $x_i$ and $y_i$ coordinates using RGB-D data are computed as follows (Khoshelham and Elberink, 2012):

$$x_i = \frac{z_i}{f} x_i^{IR} + x_0 + \delta_x \tag{2}$$

$$y_i = \frac{z_i}{f} y_i^{IR} + y_0 + \delta_y \tag{3}$$

where $x_i$, $y_i$ denotes coordinates for each point on surface, $x_0$, $y_0$ represents the principal point coordinates, $\delta_x$, $\delta_y$ denotes the lens distortion, $x_i^{IR}$, $y_i^{IR}$ are the 2D coordinates for each pixel in current infrared image (IR image) and $f$ denotes the calibrated focal distance of infrared camera (IR camera). The before mentioned task should be realized for each pair of RGB-D frame. The result is a set of corresponding 3D points.

Fourth, to estimate the initial approximation of the transformation parameters for the fine registration task we propose a coarse point-based alignment procedure. Then, since exists a set of three or more corresponding 3D points, a 3D rigid transformation is used to carried out the coarse registration task. The single-cost function that represents the 3D rigid transformation, which minimize the point-to-point error ($\mathbf{e}$) is expressed as:

$$\mathbf{e} = \sum_{i=1}^{n} \left\| \mathbf{X}_{i,Ref} - \mathbf{R} \cdot \mathbf{X}_{i,Pesq} - \mathbf{t} \right\| \tag{4}$$

where $\mathbf{X}_{i,Ref}, \mathbf{X}_{i,Pesq}$ denotes the 3D coordinates of point $i$ in reference RGB-D frame ($Ref$) and a search RGB-D frame ($Pesq$), respectively, $n$ denotes the number of corresponding 3D points used for each pair of RGB-D frame, $\mathbf{R}$ denotes the 3D rotation matrix and $\mathbf{t}$ is the 3D translation vector. The least-squares solution for $\mathbf{R}$ and $\mathbf{t}$ can be obtained: $\hat{\boldsymbol{x}} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{w}$, where A represents the Jacobian of the equation with respect to the observations, **w** is obtained by the observation and the approximate values of the unknowns. This arrangement gives to the system an initial approximation for the fine registration task. In the fifth step of the proposed pairwise registration method, we segmented pairs of RGB images to associate the centroid of the segmented regions to the depth values extracted from Depth images. Figure 2 illustrates the main steps to associate the centroid of the segmented regions in the RGB images to the depth values extracted from Depth images.

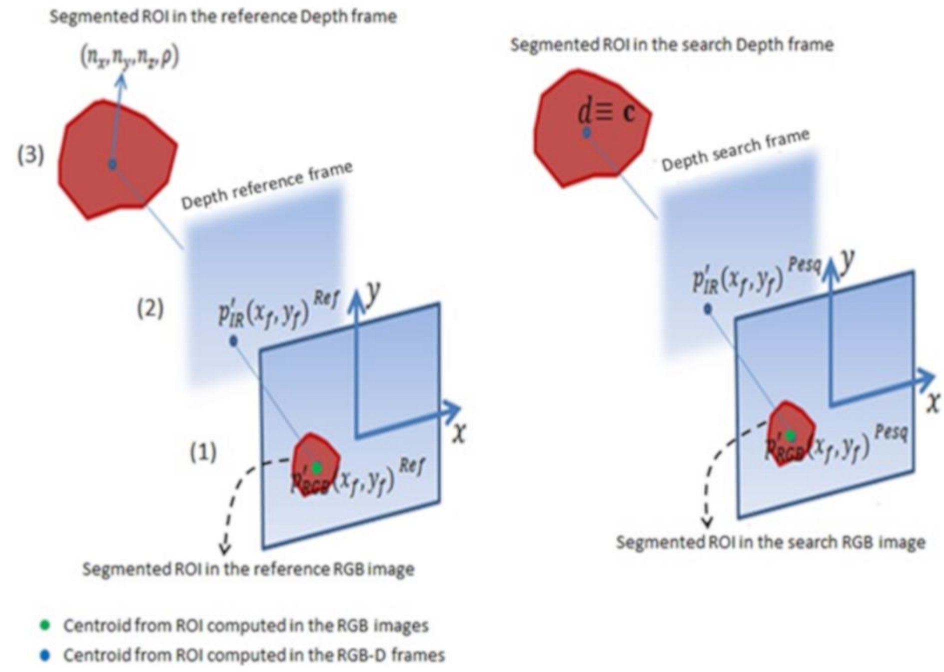
Segmented ROI in the reference Depth frame
$(n_x, n_y, n_z, \rho)$
(3)
Depth reference frame
(2)
$p'_{IR}(x_f, y_f)^{Ref}$
y
x
(1)
$p'_{RGB}(x_f, y_f)^{Ref}$
Segmented ROI in the reference RGB image
Segmented ROI in the search Depth frame
$d \equiv c$
Depth search frame
$p'_{IR}(x_f, y_f)^{Pesq}$
y
x
$p'_{RGB}(x_f, y_f)^{Pesq}$
Segmented ROI in the search RGB image
● Centroid from ROI computed in the RGB images
● Centroid from ROI computed in the RGB-D frames

**Figure 2:** Description of the tasks to associate depth values with the centroid of the segmented region in the pair of RGB images.

According Khoshelham et al. (2013) the Kinect sensor captures Depth and colour (or RGB) 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 (or frame). In Figure 2, the regions of interest (ROI) are segmented in a pair of RGB images using the simple linear iterative clustering (SLIC) algorithm proposed by Achanta et al. (2012). Basically, SLIC performs a local clustering of pixels in 5-D space defined by values of the nonlinear transformation of RGB frame. It samples $K$ regularly spaced cluster centers and moves them to seed locations corresponding to the lowest gradient position in a 3×3 neighbourhood. Each pixel in the image is associated with the nearest cluster centroid whose search area overlaps this pixel. For all the pixels associated with the nearest cluster centroid, a new centroid is computed as the average vector of all the pixels belonging to the cluster. At the end of this process, a few stray labels may remain. For each pair of RGB images the centroids $p'(x_f, y_f)^{Ref}$ and $p'(x_f, y_f)^{Pesq}$ are computed by means of the average position of all the point in the segmented ROI for matching process. Thus, the correspondences are obtained using the approximate nearest neighbours (ANN) algorithm proposed by Muja and Lowe (2009). After, the 3D coordinates $X_i = [x_i \; y_i \; z_i]^T$ for each segmented ROI is computed using Equations (1)-(3). Third, the segmented ROI from reference Depth frame ($Ref$) are fitted to obtain the unit normal vector ($\mathbf{n}$). Once the plane is represented as a 3D point $X_i$ and $\mathbf{n}$ (normal vector), and the distance from a point $\boldsymbol{p}_q \in \mathrm{ROI}$ to the plane is defined as $\rho_i = (\boldsymbol{p}_q - X_i)\mathbf{n}$. The solution for $\mathbf{n}$ is given by analysing the eigenvalues and eigenvectors of the matrix $C \in \mathbb{R}^{3x3}$ of $\mathrm{ROI}$, as follows:

$$C = \sum_{i=1}^{m} (\boldsymbol{p}_i - X)\,(\boldsymbol{p}_i - X)^T, \mathbf{C}v_j = \lambda_j v_j, j \in \{0,1,2\} \quad (5)$$

For $C$ we determine its eigenvalues $\lambda_j \in \mathbb{R}$ and their corresponding eigenvectors $v_j$. For $0 \leq \lambda_0 \leq \lambda_1 \leq \lambda_2$, the eigenvector $v_0$ corresponding to the eigenvalue $\lambda_0$ that represents an approximation of $+\mathbf{n} = \{n_x, n_y, n_z\}$ or $-\mathbf{n}$. Then, the centroid $\mathbf{c} = [x_i{}^k \quad y_i{}^k \quad z_i{}^k]^{\mathrm{T}}$ is obtained by means of the average position of all the points in each corresponding segmented ROI from search Depth frame ($Pesq$). Finally, $\mathbf{c}$ is associated to the disparity value extracted from their corresponding depth value in the Depth search image. Note that the red surfaces denote segmented ROIs in the RGB images, green points denote the computed centroid from ROI in the RGB images and blue points represent the computed centroid from ROI in the Depth frames, which 3D points are determined using Equations (1)-(3). The depth values are extracted from Depth images and there is a centroid associated to them. Finally, we refine the transformation parameters ($\mathbf{R}, \mathbf{t}$) using the disparity-to-plane model introduced by Dos Santos et al. (2016). Let $\mathbf{c}_{Ref}$ be the set of centroids from segmented ROI in the reference Depth frame and $\mathbf{c}_{Pesq}$ the set of centroids from segmented ROI in the search Depth frame, the 3D rigid transformation ($\mathbf{T}$) between them consists of a 3D rotation and a 3D translation (transformation parameters). Conveniently, these transformation parameters are combined in a transformation matrix $\mathbf{T}$ of homogenous coordinates:

$$\mathbf{c}_{Pesq} = \mathbf{T}\mathbf{c}_{Ref} = \begin{bmatrix} \mathbf{R}^* & \mathbf{t}^* \\ 0 & 1 \end{bmatrix} \mathbf{c}_{Ref} \quad (6)$$

where $\mathbf{R}^*$ and $\mathbf{t}^*$ denotes the refined 3D rotation and 3D translation, respectively, and $\mathbf{c}' = [x_i{}^k \quad y_i{}^k \quad z_i{}^k \ \ 1]^T$ is the homogeneous representation of the centroid in 3D space.

For a set of centroids that lie on a segmented ROI, the disparity-to-plane model used to refine $\mathbf{R}$ and $\mathbf{t}$ can be written as:

$$\pi^T \mathbf{c}' = 0 \tag{7}$$

where $\pi = (\mathbf{n}^{\mathrm{T}}, -\rho)^{\mathrm{T}}$ is the homogeneous representation of the plane defined by a normal vector $\mathbf{n}$ unit length and its perpendicular distance $\rho$ from origin. Substituting Equation (6) in (7) give us:

$$[n_x \quad n_y \quad n_z \quad -\rho]^T \begin{bmatrix} \mathbf{R}^* & \mathbf{t}^* \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_i{}^k \\ y_i{}^k \\ z_i{}^k \\ 1 \end{bmatrix} = 0 \tag{8}$$

where $n_x, n_y, n_z$ denote the unit normal vector.

Substituting equations (1)-(3) in (8), the disparity-to-plane model can be expressed as (Dos Santos et al., 2016):

$$[n_x \quad n_y \quad n_z - \rho]^T \begin{bmatrix} \mathbf{R}^* & \boldsymbol{t}^* \\ 0_{1x3} & 1 \end{bmatrix} \begin{bmatrix} \frac{z_i^k}{f} x_i^{k\prime} \\ \frac{z_i^k}{f} y_i^{k\prime} \\ \frac{1}{c_0 + c_1 d} \\ 1 \end{bmatrix} = 0 \tag{9}$$

where $x_i^{k\prime}$, $y_i^{k\prime}$ represents the centroid coordinates for each segmented current images. Developing Equation (9) algebraically:

$$\Big( \big[n_x r_{11} + n_y r_{21} + n_z r_{31}\big] \frac{z_i^k}{f} x_i^{k\prime} + \big[n_x r_{12} + n_y r_{22} + n_z r_{32}\big] \frac{z_i^k}{f} y_i^{k\prime} + \big[n_x r_{13} + n_y r_{23} + n_z r_{33}\big] f + \big[n_x t_x + n_y t_y + n_z t_z - \rho\big] f (c_0 + c_1 d) \Big) = 0 \tag{10}$$

where $r_{11} \dots r_{33}$ represent the elements of $\mathbf{R}^*$and $t_x, t_y, t_z$ are the elements of $\mathbf{t}^*$. For a set of three or more non-parallel segmented ROI corresponding to the centroid $\mathbf{c}'$ previously established, the solution for $\mathbf{R}^*$and $\mathbf{t}^*$ can be obtained by means of least-squares criteria: $\hat{\boldsymbol{x}} = (\mathbf{A}^{\mathbf{T}}\mathbf{M}^{-\mathbf{1}}\mathbf{A})^{-1}\mathbf{A}^{\mathbf{T}}\mathbf{M}^{-\mathbf{1}}\boldsymbol{w}$, where $\mathbf{M} = \mathbf{B}\mathbf{B}^T$, A represents the Jacobian of the condition equation with respect to the unknown parameters, B represents the Jacobian of the condition equation with respect to the observations, **w** is obtained by evaluating the condition equation with the observation and the approximate values of the unknowns. Note that $\mathbf{R}$ and $\mathbf{t}$ are used as initial approximation to estimate $\mathbf{R}^*$ and $\mathbf{t}^*$ into fine registration task. Once the coarse-to-fine registration step accumulates errors, loop closure task should be actived. This task corresponds to the step seven of the proposed method, as shows Figure 1.

## 3.2 Loop closure detection

This task is formalized with the help of a graph-based optimizer. A graph is an ordered pair $G(l,U)$ comprising a set $l$ of nodes together with a set $U$ of edges. In this work, nodes are parameterized by 3D rotation and 3D translation components, which corresponds to the refined transformation parameters $l_i(\mathbf{R}_i^*, \mathbf{t}_i^*)$, for all $i \in [1,k]$, and edges represents constraints between the nodes. Loop closure ensures that new constraints can be inserted in the graph optimization. As described, loop closure corresponds to the global adjustemnt for simultaneous refinement of all the transformation parameters in a sequence. In this work, we used the loop closure method proposed by Hogman (2012). According Hogman (2012), the loop closure aims to detect revisited locations checking if the key frames matches with some of the previous one. Then, a transformation can be estimated between the current frame and key frame. Consequently, a new constraint can be inserted from it. Finally, the cumulated error of $G(l,U)$ can be minimized. However, if the time required for check the similarities between the key frame and current frame is not fast, the loop closure detection task can be highly time consuming. In order to improve the computational loop closure detection cost it was used a slide windows strategy proposed by Hogman (2012).

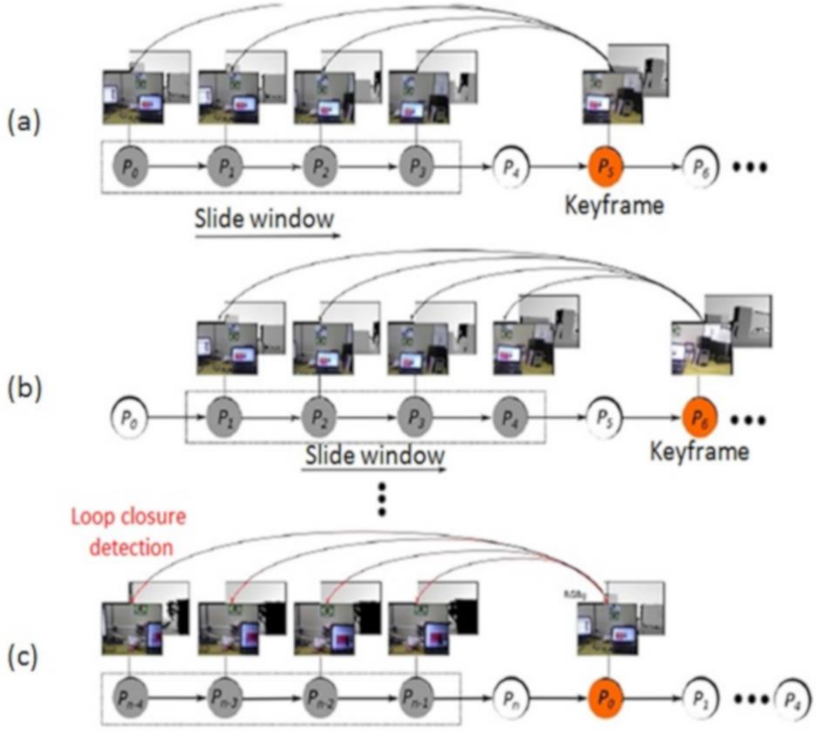

**Figure 3:** Slide window strategy proposed by Hogman (2012). (a) Loop closure detection between current frame and the old four frames. (b) Window slides step forward representation to check frames. (c) Loop closure detection process until no more key frames exist.

In Figure 3a, the detection of the loop closure is done by comparing the most recent frames ($P_5$) with the previous four frames ($P_0, \ldots, P_3$) belonging to the slide window, where the last frame ($P_4$) is ignored to improve the computational cost. The slide window is defined with a fixed size of four frames in the past. In other words, for the current frame $P_5$, the frames $P_0 - P_3$ are checked. Then, the loop closure is inserted between $P_5$ and the old frame with greater similarity. The similarity value is determined by a probability density function, as described in Hogman (2012). In this case, an edge ($U$) should be added to the graph. Thus, the graph is optimized using G2O graph-based method. For the next frame $P_6$, the window slides one step forward and the frames $P_1 - P_4$ are checked, as illustrated in Figure 3b. These tasks are successively executed until no more key frames exist, as depicted in Figure 3c. Then, whole pose graph can be optimized using a pose graph network optimizer as proposed by Grisseti et al. (2007). The graph to be optimized is written as:

$$l_i = \prod_{i=n}^{1} \mathbf{E}_{i-1}^{i} \, l_j^i \tag{11}$$

where $l_j^i$ represents the nodes formed by consecutive transformations and $\mathbf{E}_{i-1}^{i}$ denotes the edges created by relative transformations between the nodes.

## 3.3 Graph optimization

In this work, G2O graph-based optimizer is used for global adjustemnt of all the transformation parameters in a sequence. In G2O method, the directed edges are added in $G$ based on number of correct corresponding points and $l_i(\mathbf{R}_i^*, \mathbf{t}_i^*)$, such as described in Grisetti et al. (2007). To verify the consistency of the optimization problem an analyse of error function $e(l_i, l_j)$ should be executed, once an edge is characterized for both $e(l_i, l_j)$ and its variance-covariance matrix. According Khoshelham and Elberink (2012) the accuracy of depth image is not stable and also depends on the structure of the environment. For example, on some materials the reflection affects the measurement depth. Thus, is essential includes a metric to ensure that the constraint to be comprised in the loop closure detection task represents the real uncertainty of the transformation parameters. In Kümmerle et al. (2011) the variance-covariance matrix that describes the uncertainty of the transformation parameters to weight the graph-based optimization task is used as identity matrix or, usually, computed by Cholesky solver. The Cholesky solver only carried out the inverse of $\mathbf{A}^{\mathbf{T}}\mathbf{A}$, being $\mathbf{A}$ the Jacobian of the condition equation with respect to the unknown parameters. Note that Cholesky solver ignores the posteriori variance factor ($\hat{\sigma}_o^2$) and the number of redundant observations in the system (*r*). To compensate this deficiency, we introduced the variance-covariance matrix $\widehat{\mathbf{C}}_x$ of the refined transformation parameters $(\mathbf{R}_i^*, \mathbf{t}_i^*)$ in the graph optimization process. In this work, $\widehat{\mathbf{C}}_x$ is obtained jointly with the fine registration task, as follows:

$$\widehat{\mathbf{C}}_x = \hat{\sigma}_0^2 (\mathbf{A}^{\mathbf{T}} \mathbf{M}^{-1} \mathbf{A})^{-1} \tag{12}$$

where $\hat{\sigma}_0^2 = \frac{\mathbf{v}^T\mathbf{v}}{r}$, A represents the Jacobian of the condition equation with respect to the unknown parameters, $\mathbf{M} = \mathbf{B}\mathbf{B}^T$ with B denoting the Jacobian of the condition equation with respect to the observations and $\mathbf{v}$ represents the residual vector. Note that, we introduced $\hat{\sigma}_0^2$ computed with respect to the residual vector and the redundancy, as well as the Jacobian of the condition equation with respect to the observations.

## 4. Experiments and Results

To demonstrate and evaluate the effectiveness of our proposed method three experiments were conducted. Firstly, we calibrate the Kinect device using the Matlab camera calibration toolbox (Bouguet, 2004). The Bouguet's method improved the technique originally proposed by Zhang (2000). Basically, the well-known photogrammetric bundle adjustment with self-calibration method is used to estimate the interior orientation parameters (IOPs), the relative translation and rotation between the RGB and IR sensors ($\Delta x$,$\Delta y$, $\Delta z$, $\Delta\omega$, $\Delta\varphi$, $\Delta\kappa$). The lens distortions are modeled using the Brown model (Brown, 1971). Table 2 presents the estimated IOPs and the relative translation and rotation between the IR and RGB sensors. Second, a sequence of handled data was captured in order to test the suitability of the devised method for 3D indoor mapping applications. A person carried the Kinect sensor by hand leaving a certain place, turned around and then returning to the same position. A total of three indoor scenarios were captured and processed. Table 3 summarizes the trajectory of the sensor obtained after the graph optimization, the number of nodes ($l_i$) corresponding to the refined transformation parameters $(\mathbf{R}_i^*, \mathbf{t}_i^*)$ and the number of edges ($U$) representing the constraints added in the graph by loop closure detection task.

**Table 2:** Interior orientation parameters, boresight and lever arm misalignment.

| IOPs | IR sensor | RGB sensor |
|---|---|---|
| $f_x$ | 5·398 mm $\pm$3·5e$^{-4}$ mm | 4·883 mm $\pm$3· 7e$^{-4}$ mm |
| $x_0$ | 0·055 mm $\pm$7·0e$^{-5}$ mm | -0·033 mm $\pm$7·0e$^{-5}$mm |
| $y_0$ | 0·115 mm $\pm$7·0e$^{-5}$ mm | -0·158 mm $\pm$7·0e$^{-5}$ mm |
| $k_1$ | -3·679e$^{-4}$ $\pm$8·0e$^{-6}$ $\mathrm{mm}^{-2}$ | -9·571e$^{-3}$ $\pm$8·5e$^{-6}$ $\mathrm{mm}^{-2}$ |
| $k_2$ | 4·265e$^{-4}$ $\pm$7·5e$^{-7}$ $\mathrm{mm}^{-4}$ | -6·111e$^{-3}$ $\pm$8·0e$^{-7}$ $\mathrm{mm}^{-4}$ |
| $k_3$ | 0·0 | 0·0 |
| $P_1$ | -1·122e$^{-3}$ $\pm$8·0e$^{-6}$ $\mathrm{mm}^{-1}$ | -5·923$e^{-4}$ $\pm$8·0e$^{-7}$ $\mathrm{mm}^{-1}$ |
| $P_2$ | 1·155e$^{-3}$ $\pm$8·0e$^{-6}$ $\mathrm{mm}^{-1}$ | 0·0 |

| $\Delta x$ (mm) | $\Delta y$ (mm) | $\Delta z$ (mm) | $\Delta\omega$ (º) | $\Delta\varphi$ (º) | $\Delta\kappa$ (º) |
|---|---|---|---|---|---|
| 26·1448 | 0·3433 | -2·1776 | 0·16713 | 0·30770 | 0·0966 |

The initial estimative of 2,6 cm for the baseline distances between the IR camera and RGB camera are close to the initial approximations obtained manually with micrometer equipment. In Table 2, note that the magnitude of radial distortions is larger in the RGB images.

**Table 3:** The number of nodes, edges and the trajectory of the sensor measured by the proposed method.

| Scenario | nodes ($\boldsymbol{l}_i(\mathbf{R}_i^*, \mathbf{t}_i^*)$) | edges ($\boldsymbol{U}$) | Trajectory of the sensor (m) |
|---|---|---|---|
| A | 62 | 65 | 9.73 |
| B | 33 | 52 | 3.99 |
| C | 101 | 131 | 9.77 |

The graph optimization aims to minimize the coarse-to-fine registration errors. To verify the consistency of the optimization problem the error function $e(l_i, l_j)$ is analysed. According Grisseti et al. (2007) $e$ is a function that computes the difference between the relative position of the two nodes ($l_i$ and $l_j$), that represents $l_j$ seen in the frame of $l_i$, and the ground observation ($g_i$) manually measured using a tape. In practice, $e(l_i, l_j)$ is computed through the difference between the obtained trajectory of the sensor and $g_i$. Figure 4 shows $e(l_i, l_j)$ for scenarios A-C. In Figure 4, $e(l_i, l_j)$ values closest to zero mean that the nodes $l_i$ and $l_j$ correspond to the ground observation. Note that, the largest magnitude of $e(l_i, l_j)$ was 15,9 cm found in the scenario B and the smallest magnitude of $e(l_i, l_j)$ was 2,7 cm for the scenario C. Once an edge is characterized for both $e(l_i, l_j)$ and the proposed $\hat{\mathbf{C}}_x$ matrix, when a loop closure is detected a constraint represented by $\hat{\mathbf{C}}_x$ between the nodes $l_i$ and $l_j$ is added to weight the graph optimization task. As expected, when a joint effort of SURF algorithm and disparity-based model is executed and the loop closure is detected small $e(l_i, l_j)$ values are achieved, as showed the scenario B. Usually, with textureless regions (scenarios A and C) the disparity-based model does not work appropriately and the $\hat{\mathbf{C}}_x$ cannot be computed. Then, the algorithm automatically attribute to $\hat{\mathbf{C}}_x$ the identity matrix. Consequently, are achieved high $e(l_i, l_j)$ values, as show the peaks in Figure 4.

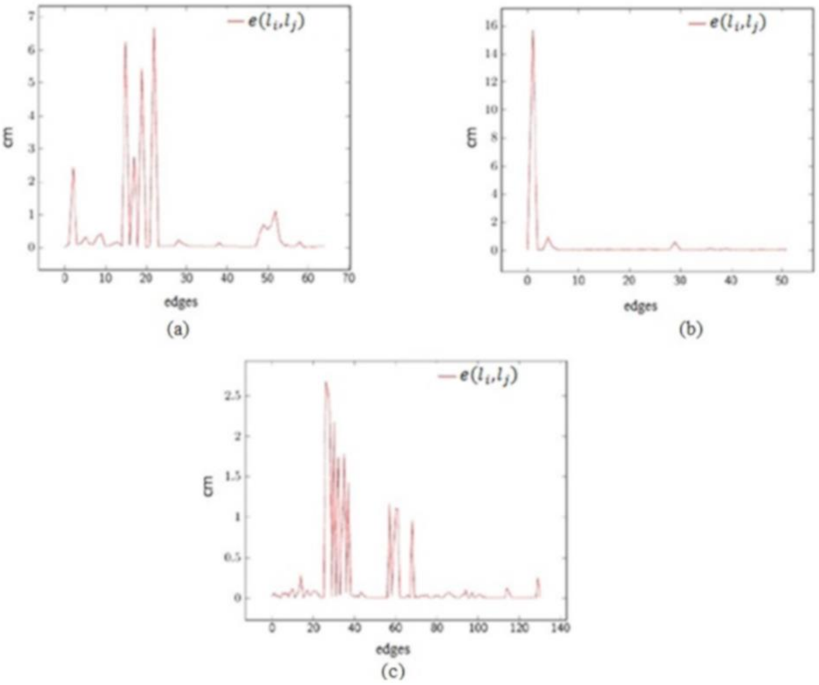

**Figure 4:** The $e(l_i, l_j)$ for scenario A (a), scenario B (b) and scenario C (c).

Third, in order to evaluate the global error of the graph optimization the root mean square (RMS) defined by $\sqrt{\frac{global\ error}{number\ of\ edges}}$ is computed. The $global\ error$ is defined as the absolute difference between a horizontal ground distance and the corresponding distance measured in 3D map after the graph optimization. The horizontal ground distance is measured with a tape and the corresponding distance is manually measured in the 3D indoor map obtained after graph optimization. The highest error was found for scenario A. Note that the error of the graph optimized corresponds to 1,5% over the distance travelled by the sensor in the experiments A and C. For example, an error of 11,97 cm in the graph corresponds to a positional imprecision around 1,5% at the distance of 9 m travelled by sensor. The 3D indoor map generated with the proposed method for scenarios A and B are displayed in Figure 5 and Figure 6.

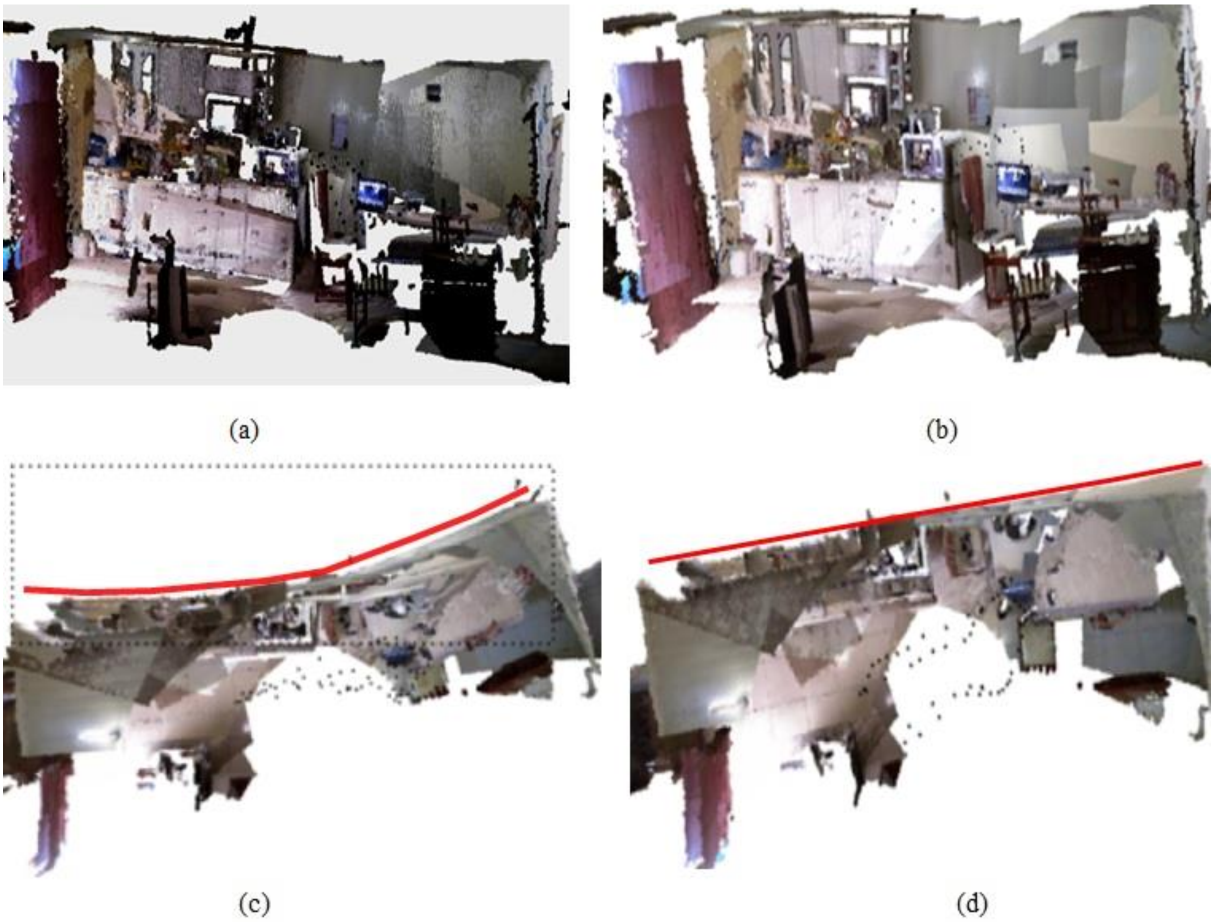

**Figure 5:** 3D indoor map for scenario A. (a) Scenario A before graph optimization, (b) Scenario A after graph optimization, (c) top view of the Scenario A before graph optimization and (d) top view of the Scenario A after graph optimization.

From Figure 5(c) a misalignment in the 3D indoor scenario before graph optimization can be observed (see the red curve projected in the kitchen wall). This effect is likely a result of the small local registration errors that are accumulated during the coarse-to-fine registration process. However, the misalignment and accumulated mapping errors are smoothed after the graph optimization task, as can be observed in Figure 5(d).

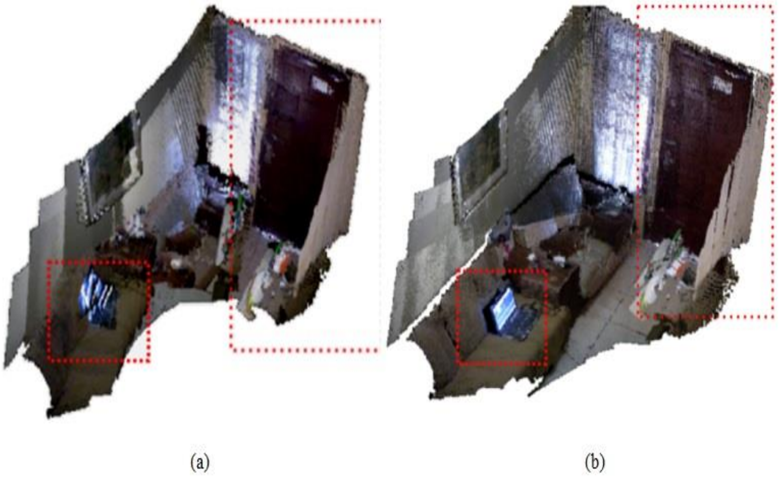

Figure 6: 3D indoor map for scenario B. (a) Scenario B before graph optimization, (b) Scenario B after graph optimization.

As can be observed in Figure 6a, a laptop in the down left corner indicated by a red square box and all the stuffs in the living room are not adequately modeled before graph optimization. After graph optimization the scenario B is consistently optimized and the before mentioned stuffs can be clearly visualized, as presented in Figure 6b. The evaluated scenarios A and B are mainly composed of doors, kitchen and living room stuffs. However, we evaluated the proposed method for a rather extreme case, such as, the scenario C that is essentially composed of flat walls, as depicted in Figure 7. Note that after the graph optimization the accumulated errors (see black narrow region in Figure 7a) are practically attenuated achieving a consistent 3D indoor map, as visually verified in Figure 7b and Figure 7c.

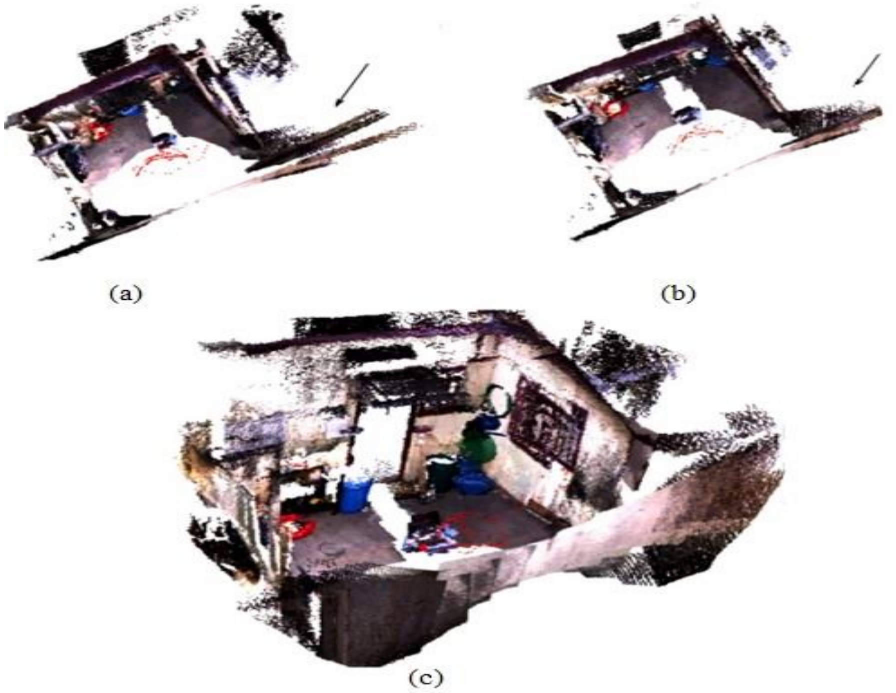

Figure 7: 3D indoor environment for scenario C. (a) top view of 3D scenario before graph optimization, (b) top view of 3D scenario after graph optimization and (c) a perspective view of 3D indoor map.

## 5. Conclusions

In this paper, a method for 3D mapping of indoor environments using RGB-D data is presented. For that purpose, a joint effort of the speed-up robust features algorithm and a disparity-to-plane model for a coarse-to-fine registration procedure is exploited. This strategy avoids the solution of difficult convergence, once errors in RGB-D data association negatively affect the performance of the registration procedure. We also detect loop closures using a variance-covariance matrix based on residual vector of the observations, redundancies and the Jacobian of the condition equation with respect to the observations. Thus, since loop closures are detected the uncertainty of the transformation parameters is carefully dealt in our method. These contributions have shown quite effective as can be observed when small $e(l_i, l_j)$ values are achieved. The coarse-to-fine registration and the weighting of the graph optimization using the mentioned strategy are the two main reasons that the method perform well in this work. Compared with point-based registration methods, plane-based approaches are more robust to outliers. In other words, outliers cannot have considerable impact on the correspondence model. In our method we joint a point-plane effort to appear sufficient. The graph is optimized using good approximations derived from proposed coarse-to-fine registration task combined with full MVC, as show the estimated global RMS. Limitations of the proposed method rely on high computational loop closure detection task and in the fact that the coarse-to-fine registration procedure cannot be used at indoor environments that contain only flat scenes. Their main disadvantages compared with LASER scanning sensors are lowest accuracy and it is not able to map outdoors environments. More sophisticated constraints to make it feasible to consider the full contribution will be the focus of the authors' future work.

## Acknowledgement

This work was supported in part by the CNPq (Conselho Nacional de Pesquisa e desenvolvimento) under Grant no. 303432/2016-0.

## REFERENCES

Achanta, R. et al. 2012. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *Pattern Analysis and Machine Intelligence*. *IEEE Transactions on*, 34(11), pp.2274–2282.

Bay, H., Tuytelaars, T. and Gool, L. V. 2006. Surf: Speeded up robust features. *In: Computer vision–ECCV*. [S.l.]: Springer, pp. 404–417.

Barnea, S. and Filin, S. 2008. Keypoint based autonomous registration of terrestrial LASER point-clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 63(1), pp.19–35. doi: 10.1016/j.isprsjprs.2007.05.005

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

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

Chow, J. et al. 2014. IMU and multiple RGB-D camera fusion for assisting indoor stop-and-go 3D terrestrial LASER scanning. *Robotics: Robot Vision*, 3(3), pp.247–280. doi: 10.3390/robotics3030247

Dos Santos, D. R. et al. 2016. Mapping indoor spaces by adaptive coarse-to-fine registration of RGB-D data. *IEEE Geoscience and Remote Sensing Letters*, 13(2), pp.262–266. doi: 10.1109/LGRS.2015.2508880

Du, H. et al. 2011. Interactive 3D modeling of indoor environments with a consumer depth camera. *Proceedings of the 13th International Conference on Ubiquitous Computing*, Beijing, China, ACM. doi: 10.1145/2030112.2030123

Fischler, M. A. and 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), pp.381-395. doi: 10.1145/358669.358692

Grisetti, G. Stachniss, C. Grzonka, S. and Burgard, W. 2007. A tree parameterization for efficiently computing maximum likelihood maps using gradient descent. *Proceedings of Robotics: Science and Systems* (RSS). doi: 10.15607/RSS.2007.III.009

Hartley, R. and Zisserman, A. 2003. *Multiple View Geometry in Computer Vision*. Cambridge University Press.

Henry, P., Krainin, M., Herbst, E., Ren, X. and Fox, D. 2012. RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments. *The International Journal of Robotics Research*, 31(5), pp.647–663. doi: 10.1177/0278364911434148

Khoshelham, K. and Elberink, S. O. 2012. Accuracy and resolution of Kinect depth data for indoor mapping applications. *Sensors (Basel)*, 12(2), pp.1437–1454. doi: 10.3390/s120201437

Khoshelham, K., dos Santos, D.R. and Vosselman, G. 2013. Generation and weighting of 3D point correspondences for improved registration of RGB-D data, *in: ISPRS Annals Volume II-5/W2: ISPRS Workshop LASER scanning.* (ed by M. Scaioni et al. Antalya, pp. 127-132). 11-13 November 2013, Antalya, Turkey. Kümmerle, R. et al. 2011. g2o: A general framework for graph optimization. *In: Proceedings - IEEE Conference on Robotics and Automation* (ICRA).

Lourakis, M. I. A. and Argyros, A. A. 2009. SBA: A software package for generic sparse bundle adjustment. *ACM Transactions on Mathematical Software*, 36(1), pp.1–30. doi: 10.1145/1486525.1486527

Lowe, D. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2), pp.91–110. doi: 10.1023/B:VISI.0000029664.99615.94

Muja, M. and Lowe, D.G. 2009. Fast approximate nearest neighbors with automatic algorithm configuration. *In: Proceedings International Conference on Computer Vision Theory and Applications* (VISAPP'09), 2009.

Wu, J., et al. 2013. A comparative study of sift and its variants. *Measurement Science Review,* 13(3), pp.122–131. doi: 10.2478/msr-2013-0021