2007 IEEE International Conference on
Robotics and Automation
Roma, Italy, 10-14 April 2007

An Efficient Direct Method for Improving visual SLAM

Geraldo Silveira, Ezio Malis, Patrick Rives

Abstract— **Traditionally in monocular SLAM, interest features are extracted and matched in successive images. Outliers are rejected a posteriori during a pose estimation process, and then the structure of the scene is reconstructed. In this paper, we propose a new approach for computing robustly and simultaneously the 3D camera displacement, the scene structure and the illumination changes *directly* from image intensity discrepancies. In this way, instead of depending on particular features, all possible image information is exploited. That problem is solved by using an efficient second-order optimization procedure and thus, high convergence rates and large domains of convergence are obtained. Furthermore, a new solution to the visual SLAM initialization problem is given whereby no assumptions are made either about the scene or the camera motion. The proposed approach is validated on experimental and simulated data. Comparisons with existing methods show significant performance improvements.**

I. INTRODUCTION

Visual SLAM consists in estimating the motion of a camera while simultaneously reconstructing the environment in which it navigates. In the computer vision community this problem is also called Structure From Motion [1]. This challenging task is traditionally divided into three major steps. First, carefully chosen, distinctive image features are extracted, e.g. by SIFT or Harris detector, and then tracked (or matched) between successive images. Once this data association problem has been solved, only pixel coordinates of the salient points will be available for further processing. However, the data association is never perfect and the outliers (aberrant measures) are usually rejected in a second step using a robust technique (e.g. RANSAC). The objective is to find a set of corresponding points free from mismatches that allows to estimate a tensor containing the camera displacement (e.g. the Essential matrix, the Trifocal tensor, etc.). Once the camera displacement has been extracted from the tensor, one can reconstruct the structure of the scene up to a scale factor. The reader may refer to the techniques proposed e.g. in [2], [3] among many others.

In this work, we depart from this paradigm and we propose a new approach to perform that core of monocular SLAM. The proposed technique computes simultaneously the 3D pose and the scene structure *directly* from image intensity discrepancies. In this way, instead of depending on particular features, all possible image information is exploited. In other words, motion and structure are directly used to align multiple reference image patches with the current image so that each pixel intensity is matched as closely as possible. Here, besides these global and local geometric parameters, global and local photometric ones are also included in the optimization process. This enables the system to work under illumination changes and to achieve more accurate alignments. In turn, the global variables related to motion *directly* enforce the rigidity constraint of the scene during the minimization. Hence, besides increasing accuracy, the technique becomes naturally robust to outliers.

The proposed technique is also different from the existing direct methods in many other aspects. The strategy [4], besides being sensitive to variable illumination, does not consider the strong coupling between motion and structure in their separated estimation processes. The method [5], though using a unified framework, relies on the linearity of image gradient which limits the system to very slow camera motions. The proposed unified approach is in fact based on an efficient second-order minimization procedure. Thus, higher convergence rates and larger domains of convergence are obtained. Furthermore, we propose a suitable structure parameterization which enforces, during the optimization, its positive depth (cheirality) constraint. Moreover, we advocate the parameterization of the visual tracking by the Lie algebra, which further improves its stability and accuracy. In addition, it is well-known that representing a scene as composed by planes leads to an improvement of computer vision algorithms in terms of accuracy, stability and rate of convergence [6]. For this reason, we suppose that any regular surface can be locally approximated by a set of planar patches. To respect real-time requirements, an appropriate selection of a subset is performed. Other contribution is concerning the initialization of the visual SLAM. This is not a trivial issue since, at the beginning of the task, *any* scene can be viewed as composed by a single plane: the plane at infinity [7]. The scene structure only appears when the translation of the camera becomes sufficiently large with respect to the depths. Given this ill conditioning, some systems e.g. [4] rely on a simple solution: one installs a reference pattern in the environment and uses it in the initial frame. Other systems e.g. [8] propose to recover the Essential matrix and to decompose it. However, if the scene is planar such a matrix is degenerate, and its decomposition provides an erroneous translation vector. In this article, a new solution for initializing the system is proposed whereby the environment is neither altered nor assumed that it is non-planar.

The experimental and simulation results also demonstrate that the image regions survive for larger camera motions and variations of illumination than by using traditional methods. Hence, by exploiting the same information in long periods of time, one avoids an early accumulation of drifts or even a total failure of the system.

Geraldo Silveira is with INRIA Sophia-Antipolis – Project ICARE, 2004 Route des Lucioles, BP 93, 06902 Sophia-Antipolis Cedex, France, and with the CenPRA Research Center – DRVC Division, Rod. Dom Pedro I, km 143,6, Amarais, CEP 13069-901, Campinas/SP, Brazil, Geraldo.Silveira@sophia.inria.fr

Ezio Malis and Patrick Rives are with INRIA Sophia-Antipolis – Project ICARE, 2004 Route des Lucioles, BP 93, 06902 Sophia-Antipolis Cedex, France, FirstName.LastName@sophia.inria.fr

II. THEORETICAL BACKGROUND

The gradient operator with respect to a variable $\mathbf{v}$ is here represented by $\nabla_{\mathbf{v}}(\cdot)$, while $\{v_i\}_{i=1}^{n}$ corresponds to the set $\{v_1, v_2, \ldots, v_n\}$, $(\mathbf{Q}^{-1})^{\top}=(\mathbf{Q}^{\top})^{-1}$ is abbreviated by $\mathbf{Q}^{-\top}$, and $\mathbf{0}$ is a matrix of zeros of appropriate dimensions. We also follow the usual notations $\widehat{\mathbf{v}}$, $\widetilde{\mathbf{v}}$, $\overline{\mathbf{v}}$, $\mathbf{v}'$ to represent respectively the estimate, an increment to be found, an augmented version and a modified one of $\mathbf{v}$.

A. The Lie Group and The Lie Algebra of $\mathbb{SE}(3)$

Consider an image $\mathcal{I}^*$ captured from a rigid scene. After displacing the camera by a rotation $\mathbf{R} \in \mathbb{SO}(3)$ and a translation $\mathbf{t} \in \mathbb{R}^3$, another image $\mathcal{I}$ of the same scene is acquired. This motion can be represented by the homogeneous transformation matrix $\mathbf{T} \in \mathbb{SE}(3)$.

Let $\mathbf{A}_i$, $i = 1, 2, \ldots, 6$, be the canonical basis of the Lie algebra $\mathfrak{se}(3)$ [9]. Any $\mathbf{A} \in \mathfrak{se}(3)$ can be written as a linear combination of the $\mathbf{A}_i$:

$$\mathbf{A}(\mathbf{x}) = \sum_{i=1}^{6} x_i \mathbf{A}_i \quad \in \mathfrak{se}(3), \tag{1}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_6]^{\top} \in \mathbb{R}^6$, and x_i is the i-th element of the base field. Such an algebra is transformed to the Lie group $\mathbb{SE}(3)$ via the exponential map:

$$\exp : \mathfrak{se}(3) \mapsto \mathbb{SE}(3); \quad \mathbf{A}(\mathbf{x}) \mapsto e^{(\mathbf{A}(\mathbf{x}))}. \tag{2}$$

The mapping (2) is smooth and one-to-one onto, with a smooth inverse, within a neighborhood of the identity element of $\mathfrak{se}(3)$ and the identity element of $\mathbb{SE}(3)$. However, these properties are valid within a very large region. Besides, due to this mapping the resulting matrix is always in the group, and no approximation is performed. Hence, the local parameterization (1) improves stability and accuracy, and is highly suitable to express incremental 3D displacements.

B. Visual Tracking Parameterized in the $\mathbb{SE}(3)$

Let $\mathbf{p} \in \mathbb{P}^2$ be the vector containing the image coordinates of a pixel. Then, we denote $\mathcal{I}(\mathbf{p})$ the image intensity of the pixel $\mathbf{p}$. Consider that an appropriate planar region $\mathcal{R}$ has been defined in $\mathcal{I}^*$ (see Section III-A). The coordinates of a pixel $\mathbf{p}^*$ defined in $\mathcal{R}^* \subset \mathcal{I}^*$ are linked to its corresponding $\mathbf{p}$ in $\mathcal{I}$ by a projective homography $\mathbf{G} \in \mathbb{SL}(3)$. Thus, a warping operator can be defined:

$$\mathbf{w}(\,\cdot\,;\mathbf{G}) : \mathbb{P}^2 \mapsto \mathbb{P}^2; \quad \mathbf{p}^* \mapsto \mathbf{p} = \mathbf{w}(\mathbf{p}^*;\mathbf{G}). \tag{3}$$

Let $\mathbf{K}$ be the upper triangular (3×3) matrix containing the camera intrinsic parameters. Given $\mathbf{K}$, then $\mathbf{G}$ can be written

$$\mathbf{G}(\mathbf{T}, \mathbf{n}^*) = \mathbf{K} \left(\frac{\mathbf{R} + \mathbf{t}\,\mathbf{n}^{*\top}}{\sqrt[3]{1 + \mathbf{t}^{\top}\mathbf{R}\,\mathbf{n}^*}} \right) \mathbf{K}^{-1}, \tag{4}$$

where $\mathbf{n}^* \in \mathbb{R}^3$ denotes the normal vector of the plane scaled by its distance to the reference camera frame.

For simplicity, let us suppose for the moment that the normal vector is known (in the Section III-C we will show how the problem can be solved if the normal vector is unknown). The problem of geometrical direct visual tracking can be formulated as a search for the optimal matrix $\mathbf{T}$ to warp all the pixels in the region $\mathcal{R}^* \subset \mathcal{I}^*$ so that their intensity values match as closely as possible to their corresponding ones in the current image $\mathcal{I}$. For that, a non-linear minimization procedure has to be derived since the pixel intensity $\mathcal{I}(\mathbf{p})$ are, in general, non-linear in $\mathbf{p}$. A standard technique to solve this problem consists in performing an expansion of the cost function in Taylor series and after applying a necessary condition of optimality. The solution of the obtained linear least squares problem iteratively updates an initial guess until convergence. Hence, given an $\widehat{\mathbf{T}}$ of $\mathbf{T}$, the problem is to find the optimal $\widetilde{\mathbf{T}} = \mathbf{T}(\widetilde{\mathbf{x}})$ through an iterative method which solves

$$\min_{\widetilde{\mathbf{x}} \in \mathbb{R}^6} \frac{1}{2} \sum_{\mathbf{p}_i^* \in \mathcal{R}^*} \left[\mathcal{I}\Big(\mathbf{w}\big(\mathbf{p}_i^*; \mathbf{G}\big(\widehat{\mathbf{T}}\mathbf{T}(\widetilde{\mathbf{x}})\big)\big)\Big) - \mathcal{I}^*(\mathbf{p}_i^*) \right]^2, \tag{5}$$

with an update of the transformation matrix as

$$\widehat{\mathbf{T}} \longleftarrow \widehat{\mathbf{T}}\mathbf{T}(\widetilde{\mathbf{x}}) = \widehat{\mathbf{T}}\, e^{(\mathbf{A}(\widetilde{\mathbf{x}}))}, \tag{6}$$

by using (2). The convergence may be established when the increments become arbitrarily small, i.e. $\|\widetilde{\mathbf{x}}\| < \epsilon$.

III. THE DIRECT VISUAL SLAM APPROACH

This section presents an unified approach where geometric and photometric models are included in a direct visual SLAM. Furthermore, it is also shown how to initialize and to obtain consistently and efficiently the optimal global and local parameters related to those models.

A. Selection of the Interest Regions

Due to real-time requirements and the fact that the entire image may not contain sufficient information for constraining all the parameters of the model, the whole image is not considered for processing. Indeed, a selection of image patches according to an appropriate score is needed. For direct methods, high scores should reflect strong image gradient along different directions.

Hence, define firstly a suitable gradient-based image $\mathcal{G}$ issued from $\mathcal{I}$. Also, let the image region $\mathcal{R}^*$ be a $(w \times w)$ matrix containing pixel intensities, whose size can be viewed as a compromise between robustness and accuracy. Given $\mathcal{G}$, a possible score image $\mathcal{S}$ may be obtained as the sum of all values of $\mathcal{G}$ within a $(w \times w)$ block centered at every pixel. This operation is well-known as a convolution of the kernel $\mathcal{K}_w$ composed by "ones" with $\mathcal{G}$

$$\mathcal{S} = \mathcal{G} \otimes \mathcal{K}_w, \tag{7}$$

which is performed extremely fast. Also, a second criterion to be added to (7), possibly with a different weight, is based on the quantity of local maxima of $\mathcal{G}$ within each block. This prevents from assigning high scores on single peaks. The resulting $\mathcal{S}$ contains the scores which are then sorted, without any absolute thresholds on the strengths to be tuned.

B. Improving the Robustness to Illumination Changes

An important issue to all vision-based methods which work directly with pixel intensities is its robustness to variable lighting. A naïve method to increase its robustness is by performing a photometric normalization. For example, the

image region may be normalized by using the mean and the standard deviation. Instead of this remediation, illumination parameters are explicitly taken into account as follows. A warped image region changed due to an illumination variation is here explained by two terms $\alpha, \beta \in \mathbb{R}$:

$$\mathcal{I}'\Big(\mathbf{w}\big(\mathbf{p}_i^*;\mathbf{G}(\mathbf{T})\big)\Big) = \alpha\,\mathcal{I}\Big(\mathbf{w}\big(\mathbf{p}_i^*;\mathbf{G}(\mathbf{T})\big)\Big) + \beta, \quad (8)$$

$\forall \mathbf{p}_i^* \in \mathcal{R}^*$. Since (8) comprises local and global parameters, this piecewise linear model can be interpreted as a model for regulating the contrast of a particular region and the brightness of the entire image. This has been shown to be a good compromise between modeling error (especially if each $\mathcal{R}^*$ has a small size) and computational complexity (few parameters together with a sparse Jacobian, as shown in Subsection III-E). In addition, it does not require any a priori knowledge about either the reflectance properties of the surface or the characteristics of the light sources. Thus, given that an iterative procedure has to be used and that the update rule for the illumination parameters is simply

$$\begin{cases} \widehat{\alpha} \longleftarrow \widehat{\alpha} + \widetilde{\alpha} \\ \widehat{\beta} \longleftarrow \widehat{\beta} + \widetilde{\beta}, \end{cases} \quad (9)$$

we can define the warped and illumination compensated pixel intensity as

$$\mathcal{I}'_i = (\widehat{\alpha} + \widetilde{\alpha})\,\mathcal{I}\Big(\mathbf{w}\big(\mathbf{p}_i^*;\mathbf{G}\big(\widehat{\mathbf{T}}\mathbf{T}(\widetilde{\mathbf{x}})\big)\big)\Big) + \widehat{\beta} + \widetilde{\beta}. \quad (10)$$

Therefore, by incorporating (10), the model-based visual tracking problem (5) becomes

$$\min_{\substack{\widetilde{\mathbf{x}}\in\mathbb{R}^6 \\ \widetilde{\alpha},\widetilde{\beta}\in\mathbb{R}}} \frac{1}{2} \sum_{\mathbf{p}_i^*\in\mathcal{R}^*} \big[\,\mathcal{I}'_i - \mathcal{I}^*(\mathbf{p}_i^*)\,\big]^2. \quad (11)$$

As a remark, the proposed model of lighting variation (8) is different from existing ones when applied to different parts of the same image. For example, the method [5], though the model is also affine, uses two local parameters for each region. By not considering the global contribution explicitly, which represents the variation of the ambient reflection, estimation of many more parameters are required. This may degrade frame-rate performance and, even worse, it may lead to convergence problems. Second, the parameters related to our model are obtained by performing an efficient second-order approximation of the cost function. Hence, nicer convergence properties are obtained without never computing the Hessians explicitly (see Section III-D). Furthermore, some approaches modify the reference $\mathcal{R}^* \subset \mathcal{I}^*$ to compensate severe changes of its appearance in the current $\mathcal{I}$. Instead, we propose to consider also the last warped and illumination compensated $(k-1)$ image of the sequence in the optimization problem. Thus, we transform (11) into

$$\min_{\substack{\widetilde{\mathbf{x}}\in\mathbb{R}^6 \\ \widetilde{\alpha},\widetilde{\beta}\in\mathbb{R}}} \frac{1}{2} \sum_{\mathbf{p}_i^*\in\mathcal{R}^*} \Bigg\{ \big[\,\mathcal{I}'_i - \mathcal{I}^*(\mathbf{p}_i^*)\,\big]^2 + \Big[\mathcal{I}'_i - \alpha^{(k-1)}\,\mathcal{I}\Big(\mathbf{w}\big(\mathbf{p}_i^*;\mathbf{G}\big(\mathbf{T}^{(k-1)}\big)\big)\Big) + \beta^{(k-1)}\Big]^2 \Bigg\}. \quad (12)$$

The application of this latter modification is optional (since it increases the number of equations to be solved) and depends on the complexity of the scenario.

C. The Full System

Since the 3D model of the scene is unknown a priori, its structure parameters must be included as optimization variables as well. For that, we perform a parameterization of the scaled normal vector $\mathbf{n}^* \in \mathbb{R}^3$ by using the depth $z_i > 0$ of any 3 image points $\mathbf{p}_i^*$ within $\mathcal{R}^*$ (e.g. its corners) as follows. Define the vector $\mathbf{z} \triangleq \big[z_1^{-1}, z_2^{-1}, z_3^{-1}\big]^\top$. Then,

$$\mathbf{n}^* = \mathbf{K}^\top \big[\ \mathbf{p}_1\ \mathbf{p}_2\ \mathbf{p}_3\ \big]^{-\top} \mathbf{z} \triangleq \mathbf{M}\,\mathbf{z}, \quad \mathbf{M} \in \mathbb{R}^{3\times 3}, \quad (13)$$

from the equations of the perspective projection of 3D points which lies on the plane. Next, given that an iterative procedure has to be devised and that the depths must be strictly positive scalars, we also parameterize them as $\mathbf{z} = \mathbf{z}(\mathbf{y}) = e^{\mathbf{y}}$, $\mathbf{y} \in \mathbb{R}^3$. This provides the update rule

$$\widehat{\mathbf{z}} \longleftarrow \widehat{\mathbf{z}} \cdot \mathbf{z}(\widetilde{\mathbf{y}}) = \widehat{\mathbf{z}} \cdot e^{\widetilde{\mathbf{y}}} = \mathrm{diag}(\widehat{\mathbf{z}})\, e^{\widetilde{\mathbf{y}}}, \quad (14)$$

where "$\cdot$" denotes elementwise multiplication. Therefore, by using this parameterization $\mathbf{n}^* = \mathbf{n}^*(\mathbf{z}(\mathbf{y}))$ we enforce, during the iterative optimization procedure, that the scene region is always in front of the camera.

Accordingly, the warped and illumination compensated pixel intensity expressed in (10) has to be changed into

$$\mathcal{I}''_i = (\widehat{\alpha}+\widetilde{\alpha})\,\mathcal{I}\Big(\mathbf{w}\big(\mathbf{p}_i^*;\mathbf{G}\big(\widehat{\mathbf{T}}\mathbf{T}(\widetilde{\mathbf{x}}),\mathbf{n}^*(\widehat{\mathbf{z}}\cdot\mathbf{z}(\widetilde{\mathbf{y}}))\big)\big)\Big) + \widehat{\beta} + \widetilde{\beta}.$$

In order to incorporate this latter modification into all regions $\mathcal{R}_j^*,\ j = 1, 2, \ldots, n$, Eq. (12) is changed into

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{7+4n-1}} \frac{1}{2} \sum_j \sum_{\mathbf{p}_{ij}^*\in\mathcal{R}_j^*} \Bigg\{ \underbrace{\big[\,\mathcal{I}''_{ij} - \mathcal{I}^*(\mathbf{p}_{ij}^*)\,\big]}_{d'_{ij}}{}^2 + \underbrace{\Big[\mathcal{I}''_{ij} - \alpha_j^{(k-1)}\,\mathcal{I}\Big(\mathbf{w}\big(\mathbf{p}_{ij}^*;\mathbf{G}\big(\mathbf{T}^{(k-1)},\mathbf{n}_j^{*(k-1)}\big)\big)\Big) + \beta^{(k-1)}\Big]}_{d''_{ij}}{}^2 \Bigg\} \quad (15)$$

where $\boldsymbol{\theta} = \big[\,\widetilde{\mathbf{x}}^\top,\ \widetilde{\beta},\ \{\widetilde{\alpha}_j,\ \widetilde{\mathbf{y}}_j\}_{j=1}^n\,\big]^\top$. Remark that in this case the regions are not tracked independently. In fact, the rigidity constraint of the scene is also explicitly enforced since all the regions share the same motion parameters. Concisely, our system can then be rewritten so as to find the optimal value

$$\boldsymbol{\theta}^\circ = \underset{\boldsymbol{\theta}\in\mathbb{R}^{6+4n}}{\arg\min}\ \frac{1}{2}\big\|\,\mathbf{d}(\boldsymbol{\theta})\,\big\|^2 \quad (16)$$

such that norm of the vector of intensity discrepancies

$$\mathbf{d}(\boldsymbol{\theta}) = \Big[\{d'_{i1}\}_i, \ldots, \{d'_{in}\}_i, \{d''_{i1}\}_i, \ldots, \{d''_{in}\}_i\Big]^\top, \quad (17)$$

whose elements are defined in (15), is minimized. In the next subsection, an efficient algorithm to solve it is developed.

In all case, regions which violate the models (e.g. independently moving ones) must be detected and discarded by the algorithm. For that, two meaningful metrics are used to evaluate the j-th template: a photometric measure as well as a geometric one. The former is here defined as

$$\mathrm{RMS}_j^2 \triangleq \frac{1}{2\,\mathrm{card}(\mathcal{R}_j^*)} \sum_{\mathbf{p}_{ij}^*\in\mathcal{R}_j^*} \big(d'^2_{ij} + d''^2_{ij}\big)\,, \quad (18)$$

where $\mathrm{card}(\cdot)$ denotes the cardinality of the set. Notice that the illumination variation has already been compensated for

in this measure. The geometric one is naturally the side ratio between the current and the previously warped region. That is, if a template shrinks or elongates significantly in at least one direction, this may mean insufficient content for constraining all the parameters (and can thus be discarded).

D. The Optimization Procedure

This subsection extends the efficient second-order minimization [10] in order to iteratively solve the problem (16). Indeed, it can be shown that, neglecting the third-order remainder, an efficient second-order approximation of $\mathbf{d}(\boldsymbol{\theta})$ around $\boldsymbol{\theta} = \mathbf{0}$ is

$$\mathbf{d}(\boldsymbol{\theta}) = \mathbf{d}(\mathbf{0}) + \tfrac{1}{2}\big(\mathbf{J}(\mathbf{0}) + \mathbf{J}(\boldsymbol{\theta})\big)\,\boldsymbol{\theta}. \tag{19}$$

The Jacobians can be found in [10], where this procedure was used for solving the problem (5). In our case, the current Jacobian $\mathbf{J}(\mathbf{0})$ within the j-th region is divided into the Jacobian with respect to the motion parameters, to the illumination parameters, and to the structure parameters:

$$\mathbf{J}(\mathbf{0}) = \big[\ \mathbf{J}_{\mathbf{x}}(\mathbf{0}),\ \mathbf{J}_{\alpha\beta}(\mathbf{0}),\ \mathbf{J}_{\mathbf{z}}(\mathbf{0})\ \big], \tag{20}$$

where

$$\begin{cases} \mathbf{J}_{\mathbf{x}}(\mathbf{0}) = \widehat{\alpha}\,\mathbf{J}_{\mathcal{I}}\mathbf{J}_{\mathbf{w}}\mathbf{J}_{\mathbf{K}}\mathbf{J}_{\widehat{\mathbf{T}}}\mathbf{J}_{\mathbf{X}}(\mathbf{0}) \\ \mathbf{J}_{\alpha\beta}(\mathbf{0}) = \big[\ \nabla_{\widehat{\beta}}\mathcal{I}'',\ \nabla_{\widehat{\alpha}}\mathcal{I}''\ \big] = \big[\ 1,\ \mathcal{I}\ \big] \\ \mathbf{J}_{\mathbf{z}}(\mathbf{0}) = \widehat{\alpha}\,\mathbf{J}_{\mathcal{I}}\mathbf{J}_{\mathbf{w}}\mathbf{J}_{\mathbf{K}}\mathbf{J}_{\widehat{\mathbf{n}}}\mathbf{M}\,\mathbf{z}, \end{cases}$$

by applying the chain rule. Correspondingly, the Jacobian $\mathbf{J}(\boldsymbol{\theta})$ is divided into

$$\mathbf{J}(\boldsymbol{\theta}) = \big[\ \mathbf{J}_{\mathbf{x}}(\boldsymbol{\theta}),\ \mathbf{J}_{\alpha\beta}(\boldsymbol{\theta}),\ \mathbf{J}_{\mathbf{z}}(\boldsymbol{\theta})\ \big], \tag{21}$$

where

$$\begin{cases} \mathbf{J}_{\mathbf{x}}(\boldsymbol{\theta}) = \widehat{\alpha}\,\mathbf{J}_{\mathcal{I}^*}\mathbf{J}_{\mathbf{w}}\mathbf{J}_{\mathbf{K}}\mathbf{J}_{\mathbf{T}}\mathbf{J}_{\mathbf{X}}(\boldsymbol{\theta}) \\ \mathbf{J}_{\alpha\beta}(\boldsymbol{\theta}) = \big[\ 1,\ \mathcal{I}^*\ \big] \\ \mathbf{J}_{\mathbf{z}}(\boldsymbol{\theta}) = \widehat{\alpha}\,\mathbf{J}_{\mathcal{I}^*}\mathbf{J}_{\mathbf{w}}\mathbf{J}_{\mathbf{K}}\mathbf{J}_{\mathbf{n}}\mathbf{M}\,\mathbf{z}. \end{cases}$$

By applying a necessary condition for $\boldsymbol{\theta} = \boldsymbol{\theta}^\circ$ to be an extremum of our cost function in (16) gives

$$\nabla_{\boldsymbol{\theta}}\big(\tfrac{1}{2}\,\mathbf{d}(\boldsymbol{\theta})^\top\mathbf{d}(\boldsymbol{\theta})\big)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\circ} = \nabla_{\boldsymbol{\theta}}\big(\mathbf{d}(\boldsymbol{\theta})\big)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\circ}^{\top}\mathbf{d}(\boldsymbol{\theta}^\circ) = \mathbf{0}. \tag{22}$$

Provided that $\mathbf{J}(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\circ}$ is full rank (see Section III-E) and using (19) around $\boldsymbol{\theta} = \boldsymbol{\theta}^\circ$, Eq. (22) yields

$$\tfrac{1}{2}\big(\mathbf{J}(\mathbf{0}) + \mathbf{J}(\boldsymbol{\theta}^\circ)\big)\ \boldsymbol{\theta}^\circ = -\mathbf{d}(\mathbf{0}). \tag{23}$$

This is not a linear system in $\boldsymbol{\theta}^\circ$ because of $\mathbf{J}(\boldsymbol{\theta}^\circ)$. However, due to the suitable parameterization of the tracking (see Section II-B), we exploit the left-invariance property of the vector fields on $\mathbb{SE}(3)$ [9]: $\mathbf{J}_{\mathbf{X}}(\boldsymbol{\theta}^\circ)\ \boldsymbol{\theta}^\circ = \mathbf{J}_{\mathbf{X}}(\mathbf{0})\ \boldsymbol{\theta}^\circ$. Moreover, provided that $\mathbf{J}_{\mathbf{T}} \approx \mathbf{J}_{\widehat{\mathbf{T}}}$ and $\mathbf{J}_{\mathbf{n}} \approx \mathbf{J}_{\widehat{\mathbf{n}}}$, we can write the left hand side of (23) as

$$\mathbf{J}'\,\boldsymbol{\theta}^\circ \triangleq \frac{1}{2}\Big[\widehat{\alpha}\,(\mathbf{J}_{\mathcal{I}}+\mathbf{J}_{\mathcal{I}^*})\mathbf{J}''_{\mathbf{x}}, \big[2, (\mathcal{I}+\mathcal{I}^*)\big], \widehat{\alpha}\,(\mathbf{J}_{\mathcal{I}}+\mathbf{J}_{\mathcal{I}^*})\mathbf{J}''_{\mathbf{z}}\Big]\,\boldsymbol{\theta}^\circ, \tag{24}$$

where $\mathbf{J}''_{\mathbf{x}} = \mathbf{J}_{\mathbf{w}}\mathbf{J}_{\mathbf{K}}\mathbf{J}_{\widehat{\mathbf{T}}}\mathbf{J}_{\mathbf{X}}(\mathbf{0})$ and $\mathbf{J}''_{\mathbf{z}} = \mathbf{J}_{\mathbf{w}}\mathbf{J}_{\mathbf{K}}\mathbf{J}_{\widehat{\mathbf{n}}}\mathbf{M}\,\mathbf{z}$. Then, by stacking appropriately each $\mathbf{J}'_j = \mathbf{J}'$ given in (24) to take into consideration all templates $j = 1, 2, \ldots, n$ (see Section III-E), as well as all the Jacobians associated to $\{d''_j\}_{j=1}^n$, the following rectangular linear system is achieved:

$$\overline{\mathbf{J}}'\,\boldsymbol{\theta}^\circ = -\mathbf{d}(\mathbf{0}), \tag{25}$$

whose solution[1] $\boldsymbol{\theta}^\circ$ iteratively updates the minimization parameters according to (6), (9) and (14) until it becomes arbitrarily small or until the cost value is arbitrarily close to stability. Therefore, we provide a second-order minimization procedure which is computationally efficient because it only involves first-order derivatives. In other words, differently to other second-order techniques (e.g. Gauss-Newton), the Hessians are never computed explicitly. This in turn contributes to obtain nicer convergence properties.

E. Initialization of the System

In this subsection, a new method to initialize the visual SLAM is presented. The technique consists in exploiting the conditioning of the Jacobians of the proposed minimization algorithm. For that, let us first of all rewrite the Jacobian $\mathbf{J}'$ defined in (24) as

$$\mathbf{J}' = \big[\ \mathbf{J}'_{\mathbf{x}},\ \mathbf{J}'_{\alpha\beta},\ \mathbf{J}'_{\mathbf{z}}\ \big]. \tag{26}$$

Next, let us expand the augmented Jacobian $\overline{\mathbf{J}}'$ in (25):

$$\overline{\mathbf{J}}' = \left[\begin{array}{c|ccc|ccc} \mathbf{J}'_{\mathbf{x}1} & \mathbf{1} & \mathbf{J}'_{\alpha 1} & \mathbf{0} & \mathbf{0} & \mathbf{J}'_{\mathbf{z}1} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{J}'_{\mathbf{x}n} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{J}'_{\alpha n} & \mathbf{0} & \mathbf{0} & \mathbf{J}'_{\mathbf{z}n} \end{array}\right] \tag{27}$$

$$= \big[\ \overline{\mathbf{J}}'_{\mathbf{x}},\ \overline{\mathbf{J}}'_{\alpha\beta},\ \overline{\mathbf{J}}'_{\mathbf{z}}\ \big]. \tag{28}$$

At the beginning of the task the translation may be small relative to the distance to the scene. If this occurs, the augmented Jacobian of the structure $\overline{\mathbf{J}}'_{\mathbf{z}}$ is ill-conditioned, which means that the structure parameters are not observable yet. In this situation, the motion parameters together with the illumination ones can explain most of the image differences. Thus, the set of linear equations (25) is initially changed into

$$\big[\ \overline{\mathbf{J}}'_{\mathbf{x}},\ \overline{\mathbf{J}}'_{\alpha\beta}\ \big]\ \big[\widetilde{\mathbf{x}}^{\circ\top},\ \widetilde{\beta}^\circ,\ \{\widetilde{\alpha}^\circ_j\}_{j=1}^n\big]^\top = -\mathbf{d}(\mathbf{0}), \tag{29}$$

whose solution $\big[\widetilde{\mathbf{x}}^{\circ\top},\ \widetilde{\beta}^\circ,\ \{\widetilde{\alpha}^\circ_j\}_{j=1}^n\big]^\top$ is also obtained in the least-square sense, and then it iteratively updates (6) and (9). The structure parameters are only used jointly to explain the image discrepancies, i.e. by solving (25), whenever the difference between the resulting cost value by using (29) and the resulting one from previous (image) optimization exceeds the image noise. This minimal parameterization presents many strengths. First, by not including unobservable parameters in the process, the pose ones are not indirectly perturbed. Second, there is no delayed initialization: all templates are always directly exploited to compute the motion. Furthermore, once the optimal structure parameters for a given set of regions are obtained, there is no reason to maintain them in the optimization (although they are back whenever that difference is exceeded). Besides that their values may be perturbed e.g. the image resolution decreases (or whenever a partial occlusion is present), less parameters in the minimization mean more available computing resources. In this case, another set of regions can be selected. Moreover, a variable-order (the regions may drop in and out) Kalman filter was used to speed up the system by providing an estimate of the minimization parameters for the next image.

[1] obtained in the least-squares sense by solving its normal equations $\overline{\mathbf{J}}'^{\top}\overline{\mathbf{J}}'\,\boldsymbol{\theta}^\circ = -\overline{\mathbf{J}}'^{\top}\mathbf{d}(\mathbf{0})$.

IV. RESULTS

In order to validate the algorithm and to assess its performance, we have tested it with both simulated and real-world scenes. Due to paper length restrictions, we report here only one sequence for each scenario. In all cases, the trivial initial conditions were used: $\widehat{\mathbf{T}}_0 = \mathbf{I}_4, \widehat{\alpha}_{j0} = 1, \widehat{\beta}_{j0} = 0, \widehat{\mathbf{n}}^*_{j0} = [0,0,1]^\top, j = 1,2,\ldots,n$. We also emphasize that: (i) bundle adjustment is not performed in any case; (ii) no off-line training phase is carried out; and (iii) a priori knowledge is not imposed anywhere.

A synthetic 3D scene was constructed so that a ground truth is available. It is composed by four planes disposed in pyramidal form, and cut by another plane on its top. In order to simulate realistic situations as closely as possible, textured images were mapped onto the planes (see Fig. 1). Afterward, a sequence of images were generated by displacing the viewpoint while varying the illumination conditions significantly: we apply an $\alpha^{(k)}$ which changes the image intensities up to 50% of its original value, and a $\beta^{(k)}$ which varies sinusoidally with amplitude of 50 pixels. We have then compared our approach (using 30 regions of size 21×21) with traditional methods as well as with a direct method. The results obtained by the proposed technique are shown in the Fig. 1, which successfully tracks the image regions by performing robustly and simultaneously the reconstruction of the pose and the scene. With regard to standard methods, we tested the SIFT keypoints, and the sub-pixel Harris detector along with a Zero-mean Normalized Cross-Correlation with mutual consistency check for matching those latter points. Afterward, 100 of those matched salient features were fed into a RANSAC procedure (typically 300 trials) with the state-of-the-art 5-point algorithm [11] for robustly recovering the pose. The comparisons are shown in the Fig. 2, where those strategies are called respectively by S+R+5P and H+ZNCC+R+5P. Since the scale factor is supposed to be unknown, the translation error is measured by the angle between the actual and the recovered translation directions. Notice that, despite the fact that more features were used, larger errors were obtained by applying those techniques to that challenging sequence, especially at the initialization step and for large displacements. Observe that the proposed initialization procedure performs well. In addition, the results show a rapidly decreasing number of tracked features, and an increasing percentage of outliers. Therefore, to avoid an early failure, a more frequent replacement of features is certainly required by those methods. As a remark, despite their relative inferior accuracy, feature-based methods can be global and thus could be used as an input to our technique. With respect to the direct methods, we have made a comparison with [5], whose results can be seen in the Fig. 3. Given that the displacements (motion and illumination) were not very small, what violate their assumptions, that algorithm failed in the seventh image of the sequence. Due to the efficient second-order approximation we propose, larger inter-frame displacements are allowed. The method proposed in [4] could not be applied since the scene is supposed to be unknown, and that it is not possible to alter the environment (it needs a reference pattern for the initialization).

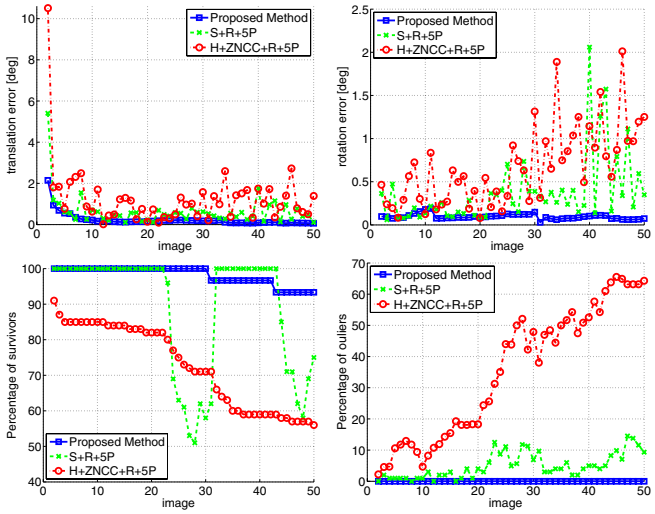

Fig. 2. Results from traditional methods and from the proposed approach. Top: errors in the recovered motion. Larger errors were obtained from traditional methods especially at the initialization step and for large displacements. Observe also that the proposed initialization step performs well.

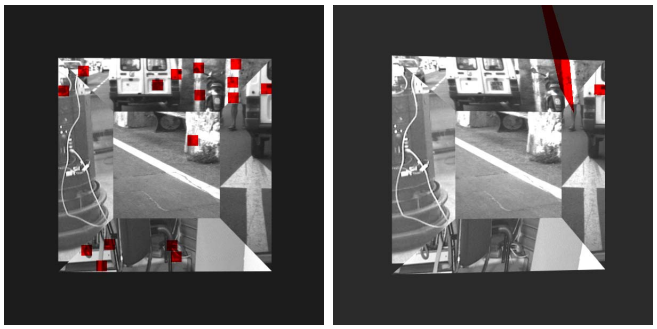

Fig. 3. Results of the tracking (in red) by using [5] at the 2nd and 7th image of the sequence, where it fails. The regions were the same as shown in Fig. 1. Observe that even at the 2nd image, many regions has already been discarded. This method also failed for the sequence presented in Fig. 4.

The results of the proposed method over a real-world sequence are shown in Fig. 4. At the beginning, the scene can be seen as the plane at infinity (see first frame). As the camera progresses, more accurate results are obtained. Note once again that the initialization step performs well. The regions were selected using [7] to show that much denser mapping can also be achieved with our technique.

V. CONCLUSIONS

In this paper, we have given various contributions for improving vision-based SLAM. First of all, we have handled the observability problems in the initialization step. Then, we have provided an efficient and robust method that directly computes the scene structure, the illumination variations and the camera displacement with respect to a reference frame. We have proved that standard methods need to add new features to track more frequently. Hence, the proposed method allows to reduce the drift by maintaining longer the estimation of the displacement with respect to the same reference frame. Future work will be devoted to the implementation of a complete visual SLAM method with small drift over large distances.

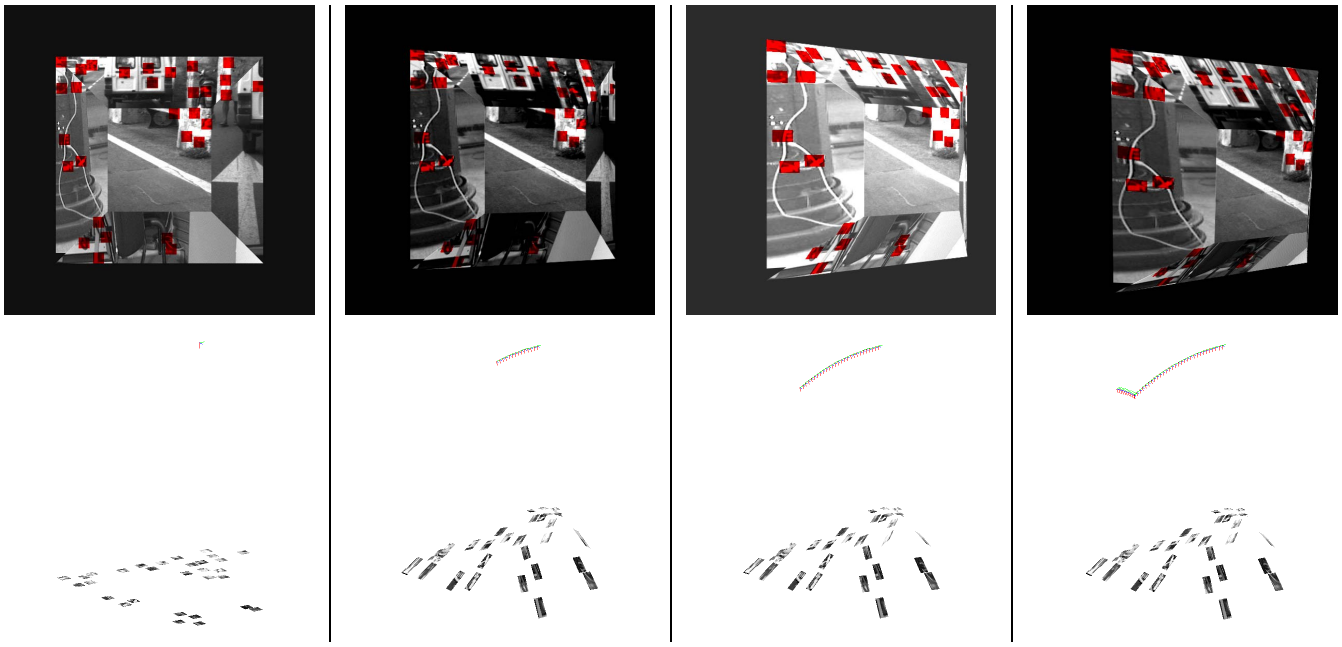

Fig. 1. Top: some frames of the Pyramid sequence superposed with the regions tracked (in red) by using the proposed approach. Observe the substantial illumination changes. Bottom: both pose and scene being incrementally reconstructed. At the beginning, the entire scene corresponds to the plane at infinity.

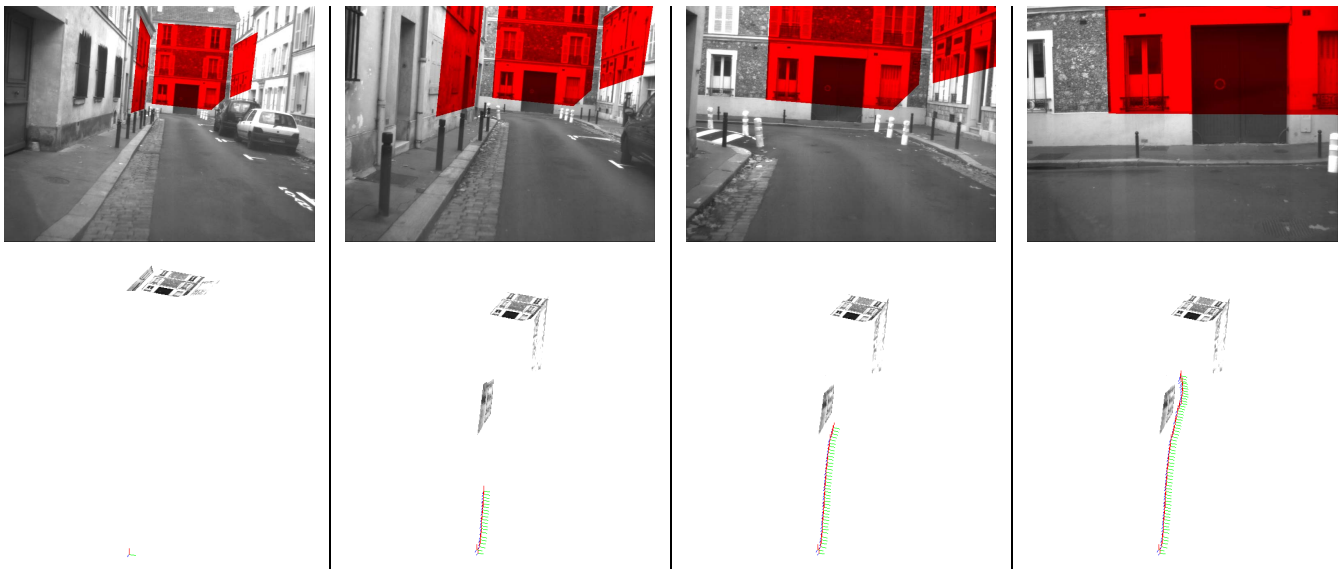

Fig. 4. Top: some frames of the Versailles sequence superposed with the regions tracked (in red) by using the proposed approach. Bottom: both pose and scene being incrementally reconstructed. Observe that, at the beginning of the task, the entire scene is viewed as the plane at infinity.

ACKNOWLEDGMENTS

This work is also partially supported by the CAPES Foundation under grant no. 1886/03-7, and by the international agreement FAPESP-INRIA under grant no. 04/13467-5.

REFERENCES

[1] O. Faugeras, *Three-Dimensional Computer Vision – A geometric viewpoint*. Cambridge: The MIT Press, 1993.

[2] P. H. S. Torr and A. Zisserman, "Feature based methods for structure and motion estimation," in *Workshop on Vision Algorithms: Theory and Practice*, 1999, pp. 278–294.

[3] S. Se, D. Lowe, and J. Little, "Vision-based global localization and mapping for mobile robots," *IEEE Transactions on Robotics and Automation*, vol. 21, no. 3, pp. 364–375, 2005.

[4] N. D. Molton, A. J. Davison, and I. D. Reid, "Locally planar patch features for real-time structure from motion," in *Proc. BMVC*, 2004.

[5] H. Jin, P. Favaro, and S. Soatto, "A semi-direct approach to structure from motion," *The Visual Computer*, vol. 6, pp. 377–394, 2003.

[6] R. Szeliski and P. H. S. Torr, "Geometrically constrained structure from motion: points on planes," in *Proc. Eur. Workshop on 3D Struct. from Mult. Images of Large-Scale Environments*, 1998, pp. 171 – 186.

[7] G. Silveira, E. Malis, and P. Rives, "Real-time robust detection of planar regions in a pair of images," in *Proc. of the IEEE/RSJ IROS*, China, 2006, pp. 49–54.

[8] D. Burschka and G. D. Hager, "V-GPS(SLAM): vision-based inertial system for mobile robots," in *Proc. of the IEEE ICRA*, USA, 2004.

[9] F. W. Warner, *Foundations of differential manifolds and Lie groups*. Springer Verlag, 1987.

[10] S. Benhimane and E. Malis, "Integration of Euclidean constraints in template based visual tracking of piecewise-planar scenes," in *Proc. of the IEEE/RSJ IROS*, China, 2006, pp. 1218–1223.

[11] D. Nistér, "An efficient solution to the five-point relative pose problem," in *Proc. of the IEEE CVPR*, vol. 2, 2003, pp. 195–202.