An Efficient Direct Approach to Visual SLAM

Geraldo Silveira, Ezio Malis, *Associate Member, IEEE*, and Patrick Rives, *Member, IEEE*

***Abstract*—The majorityof visual simultaneous localization and mapping (SLAM) approaches consider feature correspondences as an input to the joint process of estimating the camera pose and the scene structure. In this paper, we propose a new approach for simultaneously obtaining the correspondences, the camera pose, the scene structure, and the illumination changes, all directly using image intensities as observations. Exploitation of all possible image information leads to more accurate estimates and avoids the inherent difficulties of reliably associating features. We also show here that, in this case, structural constraints can be enforced within the procedure as well (instead of *a posteriori*), namely the cheirality, the rigidity, and those related to the lighting variations. We formulate the visual SLAM problem as a nonlinear image alignment task. The proposed parameters to perform this task are optimally computed by an efficient second-order approximation method for fast processing and avoidance of irrelevant minima. Furthermore, a new solution to the visual SLAM initialization problem is described whereby no assumptions are made about either the scene or the camera motion. Experimental results are provided for a variety of scenes, including urban and outdoor ones, under general camera motion and different types of perturbations.**

***Index Terms*—Illumination changes, image registration, structure and motion, vision-based simultaneous localization and mapping (SLAM).**

I. Introduction

IN ORDER TO autonomously navigate in an unknown environment, a robot must be able to build a representation of the surrounding map and self-localize with respect to it. Even though it is possible to perform the latter without the former by computer vision using an appropriate tensor (e.g., the essential matrix), precision may be lost rapidly. This happens because important structural constraints, e.g., the scene rigidity, are not effectively exploited in a long run. Having understood that both estimation processes are intimately tied together, an appealing strategy is then to perform them simultaneously. This is generally referred to as simultaneous localization and mapping (SLAM) in the robotics community. This class of methods focuses on computationally tractable algorithms that incrementally (i.e., causally) integrate information. At the expense of usually accumulating drifts earlier, they are suitable to real-time operation required by robotic platforms. A slightly different class of methods, mainly developed by the computer vision community, refers to structure from motion (SFM) techniques. Noncausal schemes fall into this latter class. These algorithms, mostly aimed at high levels of accuracy, are allowed to run in a time-consuming batch process. This paper focuses on the former class. The reader may refer to, e.g., [1] and [2] for some well-established SFM methods.

Manuscript received December 15, 2007; revised July 4, 2008. First published September 26, 2008; current version published October 31, 2008. This work was supported in part by the Brazilian Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) Foundation under Grant 1886/03-7 and in part by the International Agreement Fundação de Amparo à Pesquisa do Estado de São Paulo-Institut National de Recherche en Informatique et en Automatique (FAPESP-INRIA) under Grant 04/13467-5. This paper was recommended for publication by Associate Editor J. Neira and Editor L. Parker upon evaluation of the reviewers' comments.

G. Silveira is with the Institut National de Recherche en Informatique et en Automatique (INRIA), Project Advanced Robotics and Autonomous System (ARobAS), Sophia-Antipolis 06902, France, and also with the Division of Robotics and Computational Vision (DRVC), Centro de Pesquisa Renato Archer (CenPRA) Research Center, Campinas 13069-901, Brazil (e-mail: geraldo.silveira@cenpra.gov.br).

E. Malis and P. Rives are with the Institut National de Recherche en Informatique et en Automatique (INRIA), Project Advanced Robotics and Autonomous System (ARobAS), Sophia-Antipolis 06902, France (e-mail: ezio.malis@sophia.inria.fr; patrick.rives@sophia.inria.fr).

This paper has supplementary downloadable multimedia material available at http://ieeexplore.ieee.org. provided by the author. *Contents:* The multimedia material is composed of four videos. Each one corresponds to a different image sequence, as described in the accompanying article. This material is 10.8 MB in size. *Description:* In all videos, the left frame shows the input images superimposed with the aligned regions, i.e., the exploited regions. The right frame shows both the 3-D camera pose and the scene structure being incrementally and causally recovered. Only the regions that are currently exploited by the technique are displayed in both the frames. *Player information:* The demos were encoded with MSMPEG4V2 codec (Microsoft MPEG-4 v2). They were tested both under Linux with MPlayer as well as under Windows with Microsoft Windows Media Player version 10.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TRO.2008.2004829

A. Related Work

The techniques that simultaneously and causally reconstruct the camera pose and the scene structure can be divided into two classes, which are briefly discussed next.

1) Feature-Based Methods to Visual SLAM: A standard scheme to visual SLAM consists of first extracting a sufficiently large set of features (e.g., points, lines), and robustly matching them between successive images. These corresponding features are the input to the joint process of estimating the camera pose and scene structure. The majority of visual SLAM methods fall into this class, e.g., [3]–[5], independently of the applied filtering technique, e.g., extended Kalman filter EKF-SLAM [6] and FastSLAM 2.0 [7]. This represents the discrete case. Another possibility consists of computing the needed correspondences in the form of optical flow (the velocity). This has been exploited in, e.g., [8] and [9]. In both cases, since the prior step of data association is error-prone, care must be taken in order to avoid propagating them to subsequent steps. On the other hand, these methods may handle large interframe displacements of the objects.

2) Direct Methods to Visual SLAM: In this class of methods, the intensity value of the pixels is directly exploited to obtain the required parameters. That is, there is no prior step of data association: this is simultaneously solved. An important strength of these methods concerns the level of accuracy that they can attain. This characteristic is mainly due to the exploitation of all possible image information, even from areas where gradient information is weak. The reader may refer to, e.g., [10] for a more profound discussion about this subject.

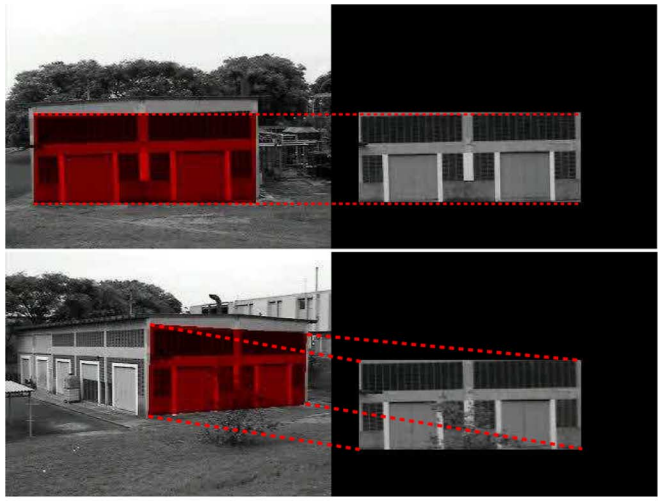
In this spirit, the technique proposed in [11] can be assigned to this class. However, it does not consider the strong coupling between motion and structure in their separated estimation processes from pixel intensities. Furthermore, it is sensitive to variable illumination. In that method, new information is initialized with a "best guess." The technique proposed in [12], though using a unified framework, relies on the linearity of image gradient. This limits the system to work under very small interframe displacements of the objects. This approach is relatively robust to lighting variations, but its model of illumination changes is overparameterized (which may lead, for example, to convergence problems). New information is initialized in a separate filter, and is inserted into the main filter after a probation period. Also, in a unified framework, central catadioptric cameras are adequately dealt with in [13]. The latter uses the same approximation method we use in this paper for obtaining the related optimal parameters. Nevertheless, its set of parameters is different from ours not only because illumination changes are handled here, but also due to the structural constraints we explicitly enforce. Moreover, initialization is not a concern in that work.

B. Overview of the Method

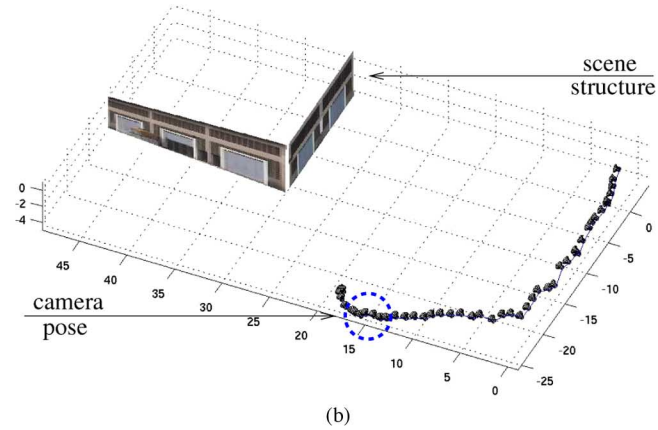
We formulate the visual SLAM problem as a nonlinear image registration task. In other words, we consider visual SLAM as the problem of estimating the appropriate parameters that optimally align a reference image with successive frames of a video sequence. A subset of the proposed parameters is naturally the camera pose and scene structure. Since the result of direct image alignments is such that each pixel intensity is matched as closely as possible across images, the technique in fact also returns a dense correspondence (see Fig. 1).

Roughly speaking, the optimal parameters are obtained as follows. Consider a parametric generative model that deforms (warps) an image. Using an estimate of the parameters, an image can be warped toward another one. The residual between the warped image and the second one is then used to iteratively refine the parameters of the models. In this paper, we focus on a deterministic optimal formulation of the visual SLAM. As for the uncertainty calculations, one can either directly cast the image registration as a stochastic optimization problem, or couple the approach with a standard filtering technique (e.g., EKF). The latter alternative is considered here, but the former is believed to represent a promising research direction.

Despite the impressive computing power to date, in a real-time setting, the entire image cannot, in general, be considered for processing. Therefore, an adequate selection of image regions is performed in this paper. Given that the selected regions may either leave the field-of-view or simply not fit the used models, the technique is able to both reject the latter and automatically insert new regions. Also, to improve computational efficiency [14], the scene is geometrically modeled as a collection of planar surfaces. This modeling is considered by all direct methods mentioned in Section I-A.2 as well.



(a)



(b)

Fig. 1. HANGAR sequence: 751-frame example of visual SLAM by aligning reference regions to successive images. All pixels within both regions are exploited, leading to a precise result. The recovered angle between walls is $89.7°$. The regions are defined relative to where they were first viewed and transfered to a common reference frame only for visualization purposes. (a) (Top) Reference region is selected. (Bottom) Using appropriate parameters, this region is automatically aligned (registered) to a different image. The image on the right is the warped region that is used to compute a residual. Other reference regions may be continuously selected and aligned if computing resources are available. (b) Subset of the parameters recovered by the proposed alignment algorithm is naturally the camera pose and the scene structure. Since monocular images are used, the scale factor is set arbitrarily.

C. Contributions

In this paper, a new approach to visual SLAM is proposed. We formulate it as a direct image registration problem. In order to solve it efficiently, consistently, and robustly, a new photogeometric generative model is presented, i.e., besides the global and local geometric parameters, global and local photometric ones are considered as optimization variables as well. This enables the system to work under generic illumination changes and achieve more accurate alignments. In turn, the global variables related to motion directly enforce the rigidity constraint of the scene within the minimization process. We remark that the proposed framework still preserves the advantages from motion parameterization using the Lie algebra. With regard to the last

but not the least structural constraint of the scene, the positive depth constraint (i.e., cheirality), a new structure parameterization is proposed to enforce it during the optimization also. Surprisingly, none of the existing direct approaches have exploited this constraint. The simultaneous enforcement within the optimization (instead of *a posteriori*) of all these structural constraints significantly contributes to improving robustness, stability, and accuracy.

Another contribution of this paper concerns the initialization of the visual SLAM. This is not a trivial issue since the scene structure becomes observable only when the amount of translation is sufficiently large with respect to its depths [15], [16]. Given this ill-conditioning, some systems, e.g., [11] rely on a simple solution: one installs a known target in the environment and uses it in the initial frame. Other systems recover and decompose the essential matrix. However, if the scene is planar, then such a matrix is degenerate, which leads to an erroneous translation vector. In this paper, a new solution for initializing the system is proposed whereby the environment is neither altered nor assumed to be nonplanar.

This paper is a revised and extended version of the visual SLAM approach that we have proposed in [17]. Besides, more thorough experiments are carried out, and a technique to automatically insert new regions is described.

II. Preliminaries

Besides the standard notations, in the sequel we adopt $\widetilde{\mathbf{v}}$, $\overline{\mathbf{v}}$, $\mathbf{v}'$, and $\|\mathbf{v}\|$ to, respectively, represent an increment to be found, an augmented version, a modified version, and the Euclidean norm of a variable $\mathbf{v}$. Here, a superscripted asterisk, e.g., $\mathbf{v}^*$ is used to represent a variable defined with respect to the reference frame, whereas a superscripted circle, e.g., $\mathbf{v}^\circ$ denotes its optimal value relative to a given cost function. Also, braces represent a set, e.g., $\{v_i\}_{i=1}^n = \{v_1, v_2, \ldots, v_n\}$, and $\mathbf{0}$ (respectively $\mathbf{1}$) is a matrix of zeros (respectively ones) of appropriate dimensions. Moreover, let $\mathbf{p} = [u, v, 1]^\top$ be the homogeneous vector containing the image coordinates of a pixel. Then, we denote as $\mathcal{I}(\mathbf{p})$ the image intensity of the pixel $\mathbf{p}$. Bilinear interpolation is used for subpixel coordinates. Consider an image $\mathcal{I}^*$ of 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 a homogeneous transformation matrix $\mathbf{T} \in \mathbb{SE}(3)$.

A. Lie Algebra $\mathfrak{se}(3)$ and the Lie Group $\mathbb{SE}(3)$

Let $\mathbf{A}_i$, $i = 1, 2, \ldots, 6$, be the canonical basis of the Lie algebra $\mathfrak{se}(3)$ [18]. Any $\mathbf{A} \in \mathfrak{se}(3)$ can thus be written as a linear combination of $\mathbf{A}_i$

$$\mathbf{A}(\mathbf{v}) = \sum_{i=1}^{6} \nu_i \mathbf{A}_i \in \mathfrak{se}(3) \tag{1}$$

where $\mathbf{v} = [\nu_1, \nu_2, \ldots, \nu_6]^\top \in \mathbb{R}^6$ represents its coordinates.

The Lie algebra $\mathfrak{se}(3)$ is related to its Lie group $\mathbb{SE}(3)$ via the exponential map

$$\exp\colon \mathfrak{se}(3) \to \mathbb{SE}(3); \qquad \mathbf{A}(\mathbf{v}) \mapsto \exp\big(\mathbf{A}(\mathbf{v})\big). \tag{2}$$

The mapping (2) is smooth and one-to-one onto, with a smooth inverse, within a very large neighborhood around the origin of $\mathfrak{se}(3)$ and the identity element of $\mathbb{SE}(3)$. The most important benefits of using this parameterization in our problem will be made clear when applying it in Sections II-C and III-D.

B. Plane-Based Two-View Epipolar Geometry

As previously stated, for efficiency reasons, we model the scene as a collection of planar regions. In this case, the coordinates of a pixel $\mathbf{p}^*$ in such a region of $\mathcal{I}^*$ are linked to its corresponding $\mathbf{p}$ in $\mathcal{I}$ by a projective homography $\mathbf{G}$ [15]

$$\mathbf{p} \propto \mathbf{G}\,\mathbf{p}^*. \tag{3}$$

The symbol "$\propto$" indicates proportionality up to a nonzero-scale factor. A warping operator $\mathbf{w}$ can then be defined as

$$\mathbf{p} = \mathbf{w}(\mathbf{G}, \mathbf{p}^*) \tag{4}$$

$$= \left[\frac{g_{11}u^* + g_{12}v^* + g_{13}}{g_{31}u^* + g_{32}v^* + g_{33}}, \frac{g_{21}u^* + g_{22}v^* + g_{23}}{g_{31}u^* + g_{32}v^* + g_{33}}, 1\right]^\top \tag{5}$$

where $\{g_{ij}\}$ denotes the elements of the matrix $\mathbf{G}$.

Consider the calibrated setting, where $\mathbf{K}$ denotes the upper triangular (3×3) matrix containing the camera's intrinsic parameters. Using the equation of the plane together with of the rigid-body motion, $\mathbf{G}$ can be written as a function of the camera displacement and the scene structure

$$\mathbf{G}(\mathbf{T}, \mathbf{n}_d^*) \propto \mathbf{K}\,(\mathbf{R} + \mathbf{t}\,\mathbf{n}_d^{*\top})\,\mathbf{K}^{-1} \tag{6}$$

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

C. Model-Based Image Alignment Parameterized in $\mathbb{SE}(3)$

Consider a textured planar surface, or that it can be locally approximated by a plane. For simplicity, let us suppose for the moment that the scaled normal vector $\mathbf{n}_d^*$ (i.e., the metric model) of this planar target is known. We will show in Section III-C how the image alignment (registration) problem can be adequately solved when this metric model is unknown.

The problem of "metric model"-based direct image alignment can be formulated as a search for the optimal matrix $\mathbf{T} \in \mathbb{SE}(3)$ to warp all the pixels in a region $\mathcal{R}^* \subset \mathcal{I}^*$ so that their intensity values match as closely as possible to their corresponding ones in the current image $\mathcal{I}$ [19]. Since one seeks an optimal pose given a scene model, this technique can also be referred to as model-based visual odometry, or simply *localization*. To this end, a nonlinear minimization procedure has to be derived since the pixel intensity $\mathcal{I}(\mathbf{p})$ is, in general, nonlinear in $\mathbf{p}$. More formally, given an estimate $\widehat{\mathbf{T}}$ of $\mathbf{T}$, the problem is to find the optimal incremental $\widetilde{\mathbf{T}} = \mathbf{T}(\widetilde{\mathbf{v}})$ through an iterative method, e.g., [19] that solves

$$\min_{\widetilde{\mathbf{v}} \in \mathbb{R}^6} \frac{1}{2} \sum_{\mathbf{p}_i^* \in \mathcal{R}^*} \Big[\mathcal{I}\Big(\mathbf{w}\big(\mathbf{G}\big(\mathbf{T}(\widetilde{\mathbf{v}})\,\widehat{\mathbf{T}}\big), \mathbf{p}_i^*\big)\Big) - \mathcal{I}^*(\mathbf{p}_i^*)\Big]^2 \tag{7}$$

with an update of the transformation matrix as

$$\widehat{\mathbf{T}} \longleftarrow \mathbf{T}(\widetilde{\mathbf{v}})\,\widehat{\mathbf{T}} = \exp\big(\mathbf{A}(\widetilde{\mathbf{v}})\big)\,\widehat{\mathbf{T}} \tag{8}$$

by using the mapping (2). The arrow "$\longleftarrow$" denotes the update assignment within the iterations. The convergence may then be established when the increments become arbitrarily small, i.e., $\|\widetilde{\mathbf{v}}\| < \epsilon$. Due to the properties of this mapping, the resulting matrix $\widehat{\mathbf{T}}$ in (8) is always in the group, and hence, no approximation is performed. If this parameterization is not applied, the resulting $\widehat{\mathbf{T}}$ has to be projected onto its group manifold, clearly reducing its rate and domain of convergence. Therefore, the local parameterization (1) improves stability and accuracy, and thus, is highly suitable to express incremental displacements. Another important property will be exploited in Section III-D to solve optimization problems, e.g., (7) efficiently and with nice convergence properties.

III. Proposed Direct Visual SLAM Approach

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

A. Selection of Image Regions

In order to satisfy the real-time requirements, we select a set of nonoverlapping image patches according to an appropriate score. For direct methods, high scores should reflect strong image gradient along different directions.

Let the image region $\mathcal{R}^* \subset \mathcal{I}^*$ be a $(w \times w)$ matrix containing pixel intensities. Then, obtain a suitable gradient-based image $\mathcal{G}^*$ from $\mathcal{I}^*$. Given $\mathcal{G}^*$, a score image $\mathcal{S}^*$ can be defined as the sum of all values of $\mathcal{G}^*$ within a $(w \times w)$ block centered at every pixel. A second criterion to be considered, possibly with a different weight, is based on the quantity of local extrema of $\mathcal{G}^*$ (denoted $\mathcal{E}^*$) within each block. This may prevent the system from assigning high scores on single peaks, which would define patches with the same drawbacks as regions defined around standard interest points (e.g., Harris corners). The neighborhood of an isolated point may not contain enough information to constrain all degrees of freedom. Other criteria are also possible, e.g., the degree of spread of the regions, but these earlier two have shown to be sufficient.

All needed block operations are efficiently performed by a convolution (denoted by "$\otimes$") with the $(w \times w)$ kernel $\mathcal{K}_w = \mathbf{1}$

$$\mathcal{S}^* = \lambda\, \mathcal{G}^* \otimes \mathcal{K}_w + \eta\, \mathcal{E}^* \otimes \mathcal{K}_w \tag{9}$$

$$= (\lambda\, \mathcal{G}^* + \eta\, \mathcal{E}^*) \otimes \mathcal{K}_w . \tag{10}$$

Typical weights are $\lambda = \|\mathcal{G}^* \otimes \mathcal{K}_w\|^{-1}$ and $\eta = 1$. The resulting $\mathcal{S}^*$ contains the scores that are sorted, without any absolute thresholds on the strengths to be tuned. The amount of regions (defined around each score) considered for further processing depends only on the available computing resources.

B. Handling Generic Illumination Changes

An important issue to all vision-based methods is their robustness to variable lighting. A widely used technique to increase this robustness is to model the change in illumination as an affine transformation [20]. Despite the fact that improved results are obtained, only global changes are modeled.

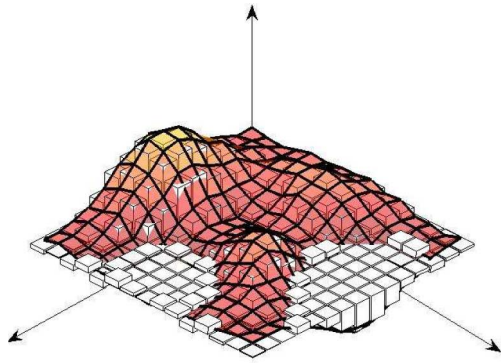


Fig. 2. (Boxes) Discretized surface for approximating (colored) the lighting changes.

Recently, we proposed in [21] a new model of illumination changes to cope with generic lighting variations. Illumination changes are viewed as a surface that can evolve with time. In that paper, we have successfully applied it to the direct visual tracking problem parameterized in the projective space. Here, we will show that this model can be straightforwardly applied to the efficient direct visual tracking problem parameterized in the Euclidean space. Indeed, for efficiency reasons, we use here the discretized realization of that generic model (see Fig. 2). Let the region have a sufficiently small size. Lighting variations are then explained by a *local* and a *global* term $\alpha, \beta \in \mathbb{R}$, respectively:

$$\mathcal{I}'(\alpha, \beta, \mathbf{p}_i) = \alpha\, \mathcal{I}(\mathbf{p}_i) + \beta. \tag{11}$$

This piecewise affine model (there is an α per region) can be interpreted as a photometric generative model for regulating the contrast of a particular region and the brightness of the entire image. This discretized model has been shown to be a good compromise between modeling error and computational complexity (it has few parameters and leads to a sparse Jacobian, as shown in Section IV). Nevertheless, it still does not require any prior knowledge about either the reflectance properties of the surface, which can be non-Lambertian, or the characteristics of the light sources, such as power, number, and their pose in space.

We remark that the model (11) is different from existing ones when applied to different parts of the same image. For example, the method proposed in [12] uses an affine model consisting of two local parameters per region. That is, it does not consider the global variations explicitly, which represent, e.g., the shift in the camera gain. In this latter overparameterized formulation, estimation of many more parameters are required. This may degrade frame-rate performance, and even worse, it may lead to convergence problems. Another important difference regards to how the related parameters are obtained. The global and local parameters related to our model are simultaneously obtained by an efficient second-order approximation method, yielding nicer convergence properties.

In fact, given that an iterative procedure is used and that the update rule for the illumination parameters can simply be

$$\begin{cases} \widehat{\alpha} \longleftarrow \widetilde{\alpha} + \widehat{\alpha} \\ \widehat{\beta} \longleftarrow \widetilde{\beta} + \widehat{\beta} \end{cases} \tag{12}$$

we can define the transformed pixel intensity as

$$\mathcal{I}'\big(\widetilde{\mathbf{v}}, \widetilde{\alpha}, \widetilde{\beta}, \mathbf{p}_i^*\big) = (\widetilde{\alpha} + \widehat{\alpha})\, \mathcal{I}\Big(\mathbf{w}\big(\mathbf{G}\big(\mathbf{T}(\widetilde{\mathbf{v}})\, \widehat{\mathbf{T}}\big), \mathbf{p}_i^*\big)\Big) + \widetilde{\beta} + \widehat{\beta}. \tag{13}$$

This can then be viewed as a photogeometric generative model. Therefore, by incorporating (13), the model-based visual tracking problem (7) becomes

$$\min_{\substack{\widetilde{\mathbf{v}} \in \mathbb{R}^6 \\ \widetilde{\alpha}, \widetilde{\beta} \in \mathbb{R}}} \frac{1}{2} \sum_{\mathbf{p}_i^* \in \mathcal{R}^*} \big[\, \mathcal{I}'\big(\widetilde{\mathbf{v}}, \widetilde{\alpha}, \widetilde{\beta}, \mathbf{p}_i^*\big) - \mathcal{I}^*(\mathbf{p}_i^*) \,\big]^2. \tag{14}$$

C. Full System

Since the metric model of the scene is unknown *a priori*, its structure parameters must be included in (14) as optimization variables as well. Indeed, the depth of some image points (not necessarily image features) together with a regularization function can be used as these variables. The latter function is needed in two-image direct reconstructions in order to avoid obtaining an underconstrained system (more unknowns than equations). As stated previously, we represent the scene here as a collection of planar regions. This, in fact, acts as our regularization function. This choice leads to a versatile and computationally efficient description of the scene (it has few parameters and leads to a sparse Jacobian, as will be shown).

We include the structure parameters as follows. First, we perform a parameterization of the scaled normal vector $\mathbf{n}_d^* \in \mathbb{R}^3$ by using the depth $z_i^* > 0$ of any (noncollinear) three image points $\mathbf{p}_i^*, i = 1, 2, 3$, within the region $\mathcal{R}^*$ (e.g., its corners). For a 3-D point that lies on the plane $\mathbf{n}_d^*$ and the equation of perspective projection, we have

$$\mathbf{n}_d^{*\top}\, \mathbf{K}^{-1} \mathbf{p}_i^* = \frac{1}{z_i^*}. \tag{15}$$

Using these three points, define the vector of inverse depths

$$\mathbf{z}^* = \left[\frac{1}{z_1^*}, \frac{1}{z_2^*}, \frac{1}{z_3^*}\right]^\top \tag{16}$$

which is the natural value to be computed. The relation between both representations is then

$$\mathbf{n}_d^* = \mathbf{M}\, \mathbf{z}^* \qquad \text{with } \mathbf{M} = \mathbf{K}^\top [\, \mathbf{p}_1^*, \mathbf{p}_2^*, \mathbf{p}_3^* \,]^{-\top} \in \mathbb{R}^{3\times 3}. \tag{17}$$

Next, given that the depths must be strictly positive scalars and that an iterative procedure has to be devised, we propose to parameterize them as

$$\mathbf{z}^* \;=\; \mathbf{z}^*(\mathbf{y}) = \exp(\mathbf{y}) > 0, \qquad \mathbf{y} \in \mathbb{R}^3. \tag{18}$$

This provides the update rule

$$\widehat{\mathbf{z}}^* \longleftarrow \mathbf{z}^*(\widetilde{\mathbf{y}}) \cdot \widehat{\mathbf{z}}^* \;=\; \exp(\widetilde{\mathbf{y}}) \cdot \widehat{\mathbf{z}}^* \tag{19}$$

where "$\cdot$" denotes element-wise multiplication.

Remark III.1 (Cheirality Constraint). By using the proposed efficient parameterization of the structure $\mathbf{n}_d^*\big(\mathbf{z}^*(\mathbf{y})\big) \in \mathbb{R}^3$, we enforce, within the optimization procedure, that the scene is always in front of the camera. That is, $z_i^* > 0\ \forall i$.

Accordingly, the photogeometric generative model expressed in (13) has to be changed into

$$\begin{aligned} &\mathcal{I}''\big(\widetilde{\mathbf{v}}, \widetilde{\alpha}, \widetilde{\beta}, \widetilde{\mathbf{y}}, \mathbf{p}_i^*\big) \\ &\quad = (\widehat{\alpha}+\widetilde{\alpha})\, \mathcal{I}\Big(\mathbf{w}\big(\mathbf{G}\big(\mathbf{T}(\widetilde{\mathbf{v}})\, \widehat{\mathbf{T}},\ \mathbf{n}_d^*(\mathbf{z}^*(\widetilde{\mathbf{y}}) \cdot \widehat{\mathbf{z}}^*)\big), \mathbf{p}_i^*\big)\Big) + \widehat{\beta} + \widetilde{\beta}. \end{aligned} \tag{20}$$

Incorporating this modification into all regions $\mathcal{R}_j^*,\ j = 1, 2, \ldots, n$, our problem becomes

$$\min_{\mathbf{x} \in \mathbb{R}^{6+4n}} \frac{1}{2} \sum_j \sum_{\mathbf{p}_{ij}^* \in \mathcal{R}_j^*} \big[\, \underbrace{\mathcal{I}''(\mathbf{x}, \mathbf{p}_{ij}^*) - \mathcal{I}^*(\mathbf{p}_{ij}^*)}_{d_{ij}(\mathbf{x})} \,\big]^2 \tag{21}$$

and where $\mathbf{x} = \big[\, \widetilde{\mathbf{v}}^\top,\ \widetilde{\beta},\ \{\widetilde{\alpha}_j,\ \widetilde{\mathbf{y}}_j^\top\}_{j=1}^n \,\big]^\top$ has $7 + 4n - 1$ parameters, since the scale factor cannot be recovered from monocular images only. Thus, one has to fix it (to a strictly positive value) to obtain a consistent solution to the problem. It can be noted that the set $\mathbf{x}$ comprises both global geometric and photometric parameters $\big(\widetilde{\mathbf{v}}^\top, \widetilde{\beta}\big)$, as well as local geometric and photometric ones $\big(\{\widetilde{\alpha}_j,\ \widetilde{\mathbf{y}}_j^\top\}_{j=1}^n\big)$.

Remark III.2 (Rigidity Constraint). Observe that in formulation (21), the regions are not independently tracked. In fact, the rigidity constraint of the scene is explicitly enforced, within the optimization procedure also, since all regions share the same incremental motion parameters.

D. Optimization Procedure

Concisely, our system (21) can then be interpreted as seeking the optimal value

$$\mathbf{x}^\circ = \underset{\mathbf{x} \in \mathbb{R}^{6+4n}}{\arg\min}\ \frac{1}{2} \big\|\, \mathbf{d}(\mathbf{x}) \,\big\|^2 \tag{22}$$

such that the norm of the vector of intensity discrepancies $\mathbf{d}(\mathbf{x}) = \{d_{ij}(\mathbf{x})\}$ is minimized. In order to iteratively solve this nonlinear optimization problem, an expansion in Taylor series is first performed. To this end, another key technique to achieve nice convergence properties is to perform an efficient second-order approximation of $\mathbf{d}(\mathbf{x})$ [22]. Indeed, it can be shown that, neglecting the third-order remainder, a second-order approximation of $\mathbf{d}(\mathbf{x})$ around $\mathbf{x} = \mathbf{0}$ is

$$\mathbf{d}(\mathbf{x}) = \mathbf{d}(\mathbf{0}) + \frac{1}{2}\big(\mathbf{J}(\mathbf{0}) + \mathbf{J}(\mathbf{x})\big)\, \mathbf{x}. \tag{23}$$

In our case, the current Jacobian $\mathbf{J}(\mathbf{0})$ is divided into the Jacobian relative to the motion parameters, the illumination parameters, and the structure parameters

$$\mathbf{J}(\mathbf{0}) = \big[\, \mathbf{J}_{\mathbf{v}}(\mathbf{0}),\ \mathbf{J}_{\alpha\beta}(\mathbf{0}),\ \mathbf{J}_{\mathbf{z}^*}(\mathbf{0}) \,\big] \tag{24}$$

where

$$\begin{cases} \mathbf{J}_{\mathbf{v}}(\mathbf{0}) = \widehat{\alpha}\, \mathbf{J}_{\mathcal{I}} \mathbf{J}_{\mathbf{w}} \mathbf{J}_{\hat{\mathbf{T}}} \mathbf{J}_{\mathbf{V}}(\mathbf{0}) \\ \mathbf{J}_{\alpha\beta}(\mathbf{0}) = \big[\, \nabla_{\hat{\beta}}\, \mathcal{I}''(\mathbf{0}),\ \nabla_{\hat{\alpha}}\, \mathcal{I}''(\mathbf{0}) \,\big] = \big[\, 1,\ \mathcal{I} \,\big] \\ \mathbf{J}_{\mathbf{z}^*}(\mathbf{0}) = \widehat{\alpha}\, \mathbf{J}_{\mathcal{I}} \mathbf{J}_{\mathbf{w}} \mathbf{J}_{\hat{\mathbf{n}}^*} \mathbf{M}\, \mathbf{z}^*(\mathbf{0}) \end{cases}$$

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

$$\mathbf{J}(\mathbf{x}) = \left[\, \mathbf{J}_{\mathbf{v}}(\mathbf{x}),\ \mathbf{J}_{\alpha\beta}(\mathbf{x}),\ \mathbf{J}_{\mathbf{z}^*}(\mathbf{x}) \,\right] \tag{25}$$

where

$$\begin{cases} \mathbf{J}_{\mathbf{v}}(\mathbf{x}) = \alpha\, \mathbf{J}_{\mathcal{I}^*} \mathbf{J}_{\mathbf{w}} \mathbf{J}_{\mathbf{T}} \mathbf{J}_{\mathbf{V}}(\mathbf{x}) \\ \mathbf{J}_{\alpha\beta}(\mathbf{x}) = \left[\, 1,\ \mathcal{I}^* \,\right] \\ \mathbf{J}_{\mathbf{z}^*}(\mathbf{x}) = \alpha\, \mathbf{J}_{\mathcal{I}^*} \mathbf{J}_{\mathbf{w}} \mathbf{J}_{\mathbf{n}^*} \mathbf{M}\, \mathbf{z}^*(\mathbf{x}). \end{cases}$$

By applying a necessary condition for $\mathbf{x} = \mathbf{x}^\circ$ to be an extremum of our cost function in (22) gives

$$\nabla_{\mathbf{x}} \left(\frac{1}{2}\, \mathbf{d}(\mathbf{x})^\top \mathbf{d}(\mathbf{x}) \right) \Big|_{\mathbf{x}=\mathbf{x}^\circ} = \nabla_{\mathbf{x}} \big(\mathbf{d}(\mathbf{x})\big)^\top \Big|_{\mathbf{x}=\mathbf{x}^\circ} \mathbf{d}(\mathbf{x}^\circ) = \mathbf{0}. \tag{26}$$

Provided that $\mathbf{J}(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}^\circ}$ is full rank (see Section IV) and using (23) around $\mathbf{x} = \mathbf{x}^\circ$, one has from (26)

$$\frac{1}{2} \big(\mathbf{J}(\mathbf{0}) + \mathbf{J}(\mathbf{x})\big)\, \mathbf{x}^\circ = -\mathbf{d}(\mathbf{0}). \tag{27}$$

This is not a linear system in $\mathbf{x}^\circ$ because of $\mathbf{J}(\mathbf{x})$. However, due to the suitable parameterization of the alignment (see Section II-C), we exploit the left-invariance property of the vector fields on Lie groups [18]. In fact, given that the space of the parameters $\mathbf{x}$ is homeomorphic to a Lie group defined over $\mathbb{SE}(3) \times \mathbb{R}^{4n}$, this property means that $\mathbf{J}_{\mathbf{V}}(\mathbf{x})\, \mathbf{x}^\circ = \mathbf{J}_{\mathbf{V}}(\mathbf{0})\, \mathbf{x}^\circ$. Then, provided that $\mathbf{J}_{\mathbf{T}} \approx \mathbf{J}_{\hat{\mathbf{T}}}$ and $\mathbf{J}_{\mathbf{n}^*} \approx \mathbf{J}_{\hat{\mathbf{n}}^*}$, the left-hand side of (27) can be written as

$$\begin{aligned} &\frac{1}{2} \big(\mathbf{J}(\mathbf{0}) + \mathbf{J}(\mathbf{x})\big)\, \mathbf{x}^\circ = \mathbf{J}'\, \mathbf{x}^\circ = \left[\, \mathbf{J}'_{\mathbf{v}},\ \mathbf{J}'_{\alpha\beta},\ \mathbf{J}'_{\mathbf{z}^*} \right] \mathbf{x}^\circ \\ &= \frac{1}{2} \left[\widehat{\alpha}\, (\mathbf{J}_{\mathcal{I}} + \mathbf{J}_{\mathcal{I}^*}) \mathbf{J}_{\mathbf{w}} \mathbf{J}''_{\mathbf{v}},\ \left[2,\ (\mathcal{I} + \mathcal{I}^*)\right],\ \widehat{\alpha}\, (\mathbf{J}_{\mathcal{I}} + \mathbf{J}_{\mathcal{I}^*}) \mathbf{J}_{\mathbf{w}} \mathbf{J}''_{\mathbf{z}^*} \right] \mathbf{x}^\circ \end{aligned} \tag{28}$$

with $\mathbf{J}''_{\mathbf{v}} = \mathbf{J}_{\hat{\mathbf{T}}} \mathbf{J}_{\mathbf{V}}(\mathbf{0})$ and $\mathbf{J}''_{\mathbf{z}^*} = \mathbf{J}_{\hat{\mathbf{n}}^*} \mathbf{M}\, \mathbf{z}^*(\mathbf{0})$.

By appropriately stacking each $\mathbf{J}'$ above to take into consideration all regions $j = 1, 2, \ldots, n$, i.e.,

$$\begin{aligned} \overline{\mathbf{J}}' &= \left[\begin{array}{c|ccccc|cccc} \mathbf{J}'_{1\mathbf{v}} & \mathbf{1} & \mathbf{J}'_{1\alpha} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{J}'_{1\mathbf{z}^*} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{J}'_{2\mathbf{v}} & \mathbf{1} & \mathbf{0} & \mathbf{J}'_{2\alpha} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{J}'_{2\mathbf{z}^*} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{J}'_{n\mathbf{v}} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{J}'_{n\alpha} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{J}'_{n\mathbf{z}^*} \end{array}\right] \\ &= \left[\, \overline{\mathbf{J}}'_{\mathbf{v}},\ \overline{\mathbf{J}}'_{\alpha\beta},\ \overline{\mathbf{J}}'_{\mathbf{z}^*} \,\right] \end{aligned} \tag{29}$$

a rectangular linear system is hence finally achieved

$$\overline{\mathbf{J}}'\, \mathbf{x}^\circ = -\mathbf{d}(\mathbf{0}) \tag{30}$$

whose solution $\mathbf{x}^\circ$ is obtained in the least-squares sense by solving its normal equations. The optimal solution is found by iteratively updating the parameters according to (8), (12), and (19) until the displacements become arbitrarily small.

Therefore, we provide a second-order approximation method that leads to a computationally efficient optimization procedure because only first-order derivatives are involved. In other words, differently from second-order minimization techniques (e.g., Newton), the Hessians are never computed explicitly. This also contributes to obtain nicer convergence properties. Furthermore, the proposed model of illumination changes together with the used representation of the scene yield sparse (diagonal) Jacobians, respectively, $\overline{\mathbf{J}}'_{\alpha\beta}$ and $\overline{\mathbf{J}}'_{\mathbf{z}^*}$, as shown in (29). Efficiency is then further improved.

IV. Initialization of the System

In this section, a method to initialize the proposed visual SLAM formulation is described. Essentially, the technique consists of a hierarchical framework in the sense of the number of parameters to explain the image motion.

A. Hierarchical Formulation

At the beginning of the task, the amount of translation may be small relative to the distance to the scene. If this occurs, the augmented Jacobian of the structure $\overline{\mathbf{J}}'_{\mathbf{z}^*}$ [see (29)] is ill-conditioned, which means that the structure parameters are not yet observable. In this situation, the motion parameters together with the illumination ones can explain most of the image differences. The latter reasoning also applies once the optimal structure parameters (i.e., the map) have already been obtained. In this case, there is no reason to maintain them as optimization variables. Besides that their values may be perturbed, e.g., when the image resolution decreases, less parameters in the minimization mean more available computing resources. Once again, motion parameters and illumination ones can explain most of the image discrepancies. As a matter of fact, in this case, the proposed visual SLAM approach effectively runs in a robust localization mode.

Therefore, for every new image, we initially attempt to align the regions by using only a subset of parameters from (30)

$$\left[\, \overline{\mathbf{J}}'_{\mathbf{v}},\ \overline{\mathbf{J}}'_{\alpha\beta} \,\right] \left[\, \widetilde{\mathbf{v}}^{\circ\top},\ \widetilde{\beta}^\circ,\ \left\{\widetilde{\alpha}^\circ_j\right\}_{j=1}^n \,\right]^\top = -\mathbf{d}(\mathbf{0}) \tag{31}$$

whose solution $\left[\, \widetilde{\mathbf{v}}^{\circ\top},\ \widetilde{\beta}^\circ,\ \left\{\widetilde{\alpha}^\circ_j\right\}_{j=1}^n \,\right]^\top$ is also obtained in the least-square sense, and then it iteratively updates (8) and (12). The structure parameters are only simultaneously used as optimization variables, i.e., by solving (30), whenever the difference between the resulting cost value by using (31) and the resulting one from previous (image) optimization exceeds the image noise. We remark that in any case, the structure (plus motion and illumination) parameters are required to compute the discrepancies $\mathbf{d}(\mathbf{0})$. These parameters can either be the optimal ones from preceding image registrations or an initial value. In fact, this shows how all past observations have contributed to incrementally building and maintaining a coherent description of the map (and locations).

B. Augmenting the Domain and the Rate of Convergence

A limitation of the visual SLAM approach proposed in Section III regards its domain of convergence. Although the parameters are obtained by a second-order approximation method with nice convergence properties, it does not ensure that the global minimum will be reached. Global optimization methods such as simulated annealing are too time-consuming to be considered in a real-time setting.

However, a possible solution to avoid getting wedged in local minima consists of using, e.g., feature-based techniques as a bootstrap to our method. We remark that even though a recovered set of parameters can represent a local minimum, it may be close to the global one. Hence, the regions may still have been effectively aligned in the image. A standard pose recovery technique can then be used with all these registered (i.e., corresponding) pixels. Afterward, the scene can be reconstructed by triangulating them [15]. In addition to augmenting the domain of convergence, this approach may also augment the rate of convergence. If these estimated motion and/or structure are closer to the true ones than those by using the proposed approach, they will act in this case as a prediction for aligning a new image.

Other predictors can additionally be tested to improve convergence properties. In fact, the coupling between the deterministic image registration proposed in Section III with a probabilistic filtering technique can be performed at this stage. Here, we use a variable-order Kalman filter to provide both another estimate of the optimization variables and the covariances. The input (i.e., observations) to the filtering are the recovered parameters from the optimization process. In order to initialize the system (i.e., when a new image is available), the best set of parameters among all predictors is simply chosen by comparing their resulting cost value.

V. Region Rejection and Insertion

A. Outliers Rejection

Within direct methods, outliers correspond to regions that do not fit the models, e.g., regions related to independently moving objects. Surface discontinuities and occluding boundaries can also be viewed as outliers. Hence, they must be detected and discarded by the algorithm. For this, two meaningful metrics are used to evaluate the jth template: a photometric measure as well as a geometric one. The photometric measure is defined directly from our cost function in (21) as

$$\varepsilon_j^2(\mathbf{x}^\circ) = \frac{1}{\mathrm{card}(\mathcal{R}_j^*)} \sum_{\mathbf{p}_{ij}^* \in \mathcal{R}_j^*} d_{ij}^2(\mathbf{x}^\circ) \tag{32}$$

where $\mathrm{card}(\cdot)$ denotes the cardinality of the set. Notice that the illumination variations have already been compensated here. The geometric measure is the side ratio between the current and the previously warped region. That is, if a template significantly shrinks or elongates in at least one direction, this may signify insufficient content for constraining all parameters (and can thus be discarded). We remark that while (32) is evaluated after obtaining the optimal solution, the geometric measure can be evaluated within the iterations, provided that the region has been adequately initialized (see next section). This may prevent such regions from perturbing the solution.

B. Insertion of New Regions

Given that regions may leave the field-of-view or eventually be rejected from the optimization, the system must be able to insert new regions whenever computing resources are available. The initialization of new regions follows the natural way of specialization: we start by the most generic stratum to the most specialized one. In other words, we first characterize each new region in the projective space. Using this knowledge and of the recovered interframe displacement, we can obtain its best possible Euclidean structure until that moment.

This algorithm is detailed as follows. Let the current image be indexed by "τ." New regions can be selected in this image according to the procedure described in Section III-A. Denote this image by $\mathcal{I}_\tau^*$ since it contains the reference template of these particular regions. Then, we have the following steps.

1) When a new image is available, obtain the projective homography that best aligns each jth selected region

$$\{\mathbf{G}_j^\circ, \alpha_j^\circ, \beta_j^\circ\} = \underset{\substack{\mathbf{G}_j \in \mathbb{SL}(3) \\ \alpha_j, \beta_j \in \mathbb{R}}}{\arg\min} \frac{1}{2} \times \sum_{\mathbf{p}_{ij}^* \in \mathcal{R}_j^*} \left[\alpha_j \mathcal{I}\big(\mathbf{w}(\mathbf{G}_j, \mathbf{p}_{ij}^*)\big) + \beta_j - \mathcal{I}_\tau^*(\mathbf{p}_{ij}^*)\right]^2 \tag{33}$$

as described in [21]. Since each region is treated independently, we have $8+2$ parameters to be recovered per region. Optionally, this procedure may be initialized by, e.g., a correlation measure.

2) Determine the scaled normal vector relative to the frame where the region was first viewed (i.e., corresponding to $\mathcal{I}_\tau^*$) using the closed-form solution described in [23]

$$\widehat{\mathbf{n}}_{d\,j}^* = \frac{\left(\mu\,\mathbf{K}^{-1}\,\mathbf{G}_j^\circ\,\mathbf{K} - \mathbf{R}_\tau^\circ\right)^\top \mathbf{t}_\tau^\circ}{\|\mathbf{t}_\tau^\circ\|^2} \tag{34}$$

with the obtained $\mathbf{G}_j^\circ$ in step 1 and the local displacement from the visual SLAM result (30) or (31). The factor $\mu \in \mathbb{R}$ is given from the median singular value of $\mathbf{K}^{-1}\,\mathbf{G}_j^\circ\,\mathbf{K}$. Of course, one must have $\mathbf{t}_\tau^\circ \neq \mathbf{0}$.

3) An iterative refinement may then be conducted using the same procedure as described in Section III-D, but using only the structure as optimization variable, i.e., with only three parameters to be recovered per region.

If the jth new region is not declared as an outlier, it is ready to be exploited from the next image. To this end, the photogenerative model (20) can adequately incorporate each new relative reference frame by multiplying the global $\widehat{\mathbf{T}} = \widehat{\mathbf{T}}_0$ by the inverse of the relative ${}^\tau\mathbf{T}_0$.

This insertion algorithm is intrinsically different to existing direct ones. For example, besides being sensitive to variable lighting, the method in [11] does not take into account all available knowledge to initialize $\widehat{\mathbf{n}}_{d\,j}^*$ (it uses a "best guess"). This may lead to convergence problems. Furthermore, differently from [12] where new regions are backprojected to the global reference frame, we avoid altering the original information by adequately incorporating them in (20). This possibility is also an attractive characteristic of the proposed SLAM formulation.

VI. Experimental Results

In order to validate the algorithm and assess its performance, we have tested it with both synthetic and real-world images. All

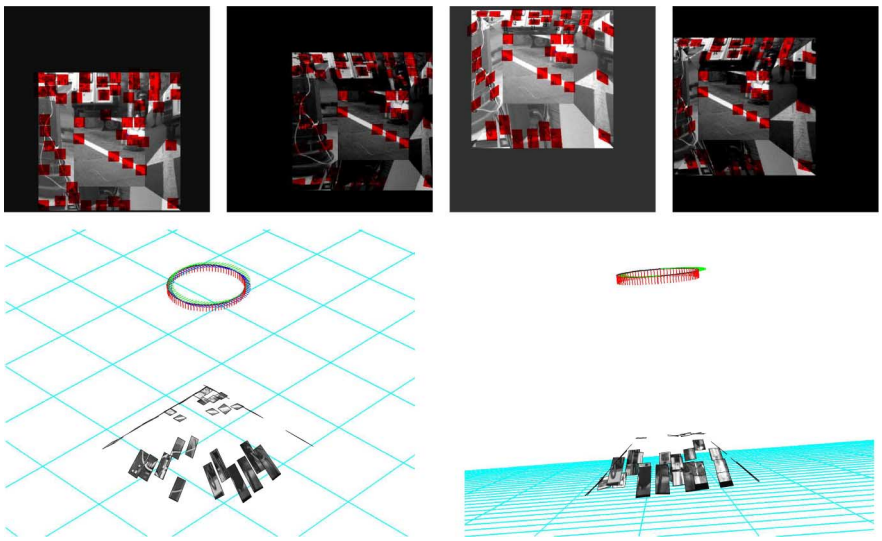


Fig. 3. (Top) Excerpts from the 81-frame PYRAMID sequence superimposed with the regions aligned (in red) by using the proposed approach. Observe the successful rejection of regions that do not fit the models (notably in the junctions of planes). (Bottom) Reconstructed structure and motion (represented by three-color frames) seen from different viewpoints. Final pose drift is of less than 0.001% of the total amount of translation and 0.091° for the rotation.

results can be found as multimedia material published in IEEE Xplore with this paper. In all cases, trivial initial conditions are used: $\widehat{\mathbf{T}}^{(0)} = \mathbf{I}_4, \widehat{\alpha}_j^{(0)} = 1, \widehat{\beta}_j^{(0)} = 0, \widehat{\mathbf{n}}_{d\,j}^{*(0)} = [0,0,1]^\top \forall j$. The photometric error is here measured by its rms (32). The jth region is declared as an outlier if either $\varepsilon_j > 20$ or if its geometric error is over 50%. The rms of the image noise is considered to be of 0.6 level of grayscale. Moreover, we emphasize that no other sensory device than a single camera is used.

A. PYRAMID *Sequence*

A synthetic scene was constructed so that a ground truth is available. It is composed of 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. Then, a sequence of images was generated by displacing the camera while varying the illumination conditions. With respect to the trajectory, the camera performs a circular motion. The objective is twofold. First, returning the camera to the starting pose offers an important benchmark for SLAM algorithms. Second, this aims to show that past observations *de facto* contribute, within the proposed incremental technique, to build and maintain a coherent description of the structure and motion. With respect to the lighting variations, they are created by applying an $\alpha^{(k)}$ that linearly changes the image intensities up to 50% of its original value, and a $\beta^{(k)}$ that varies sinusoidally with amplitude of 50 levels of grayscale.

We have then compared our approach (see some SLAM results in Fig. 3), which started with 50 regions of size 21×21 pixels, with traditional methods as well as with a direct method. With regard to standard methods, we used SIFT keypoints (1025 matches were initially found), and the subpixel Harris detector along with a zero-mean normalized cross-correlation with mutual consistency check for matching these latter points (235 were initially matched). Other than the initial ones, no features or regions are initialized here. Moreover, there is a relevant difference about how feature correspondences are established along the sequence. While keypoints are matched between the first (reference) and the current images, the latter had to be made between successive images (i.e., had to be tracked). In all cases, corresponding features were fed into a random sample access (RANSAC) procedure (typically 300 trials) with the state-of-the-art five-point algorithm [24] for robustly recovering the pose. This corresponds to a standard feature-based framework where a two-image reconstruction is considered and a nonplanar scene is assumed (because of the five-point algorithm). The comparisons are depicted in Fig. 4, where those strategies are respectively referred to as 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, i.e., $\arccos(\mathbf{t}^\top \hat{\mathbf{t}}/(\|\mathbf{t}\|\|\hat{\mathbf{t}}\|))$. Notice that, despite exploiting many more features, the standard techniques obtain relatively larger errors, especially for large displacements (i.e., middle of the loop) and significant lighting changes. In addition, the results show an increasing percentage of outliers and a rapidly decreasing number of corresponding features. Therefore, to avoid an early failure, these methods certainly require a more frequent replacement of features. As a remark, despite their relative inferior accuracy, feature-based methods can have a larger domain of convergence, and thus, may be used as a bootstrap to our technique (as discussed in Section IV-B). For the requested accuracy, the proposed approach performed along the sequence of a median of seven iterations returned a median photometric error of 9.84 levels of grayscale, and used a median of 10.4% of each (500×500) image. For this sequence where perfect camera's intrinsic parameters are available, the proposed method realized a drift between the original and final pose (since a closed loop is performed) of less than 0.001% of the total amount of translation and 0.091° for the rotation. This shows that precise results with minimal drift are obtained.

With respect to existing direct methods, we have made a comparison with [12]. Given that the displacements (motion and

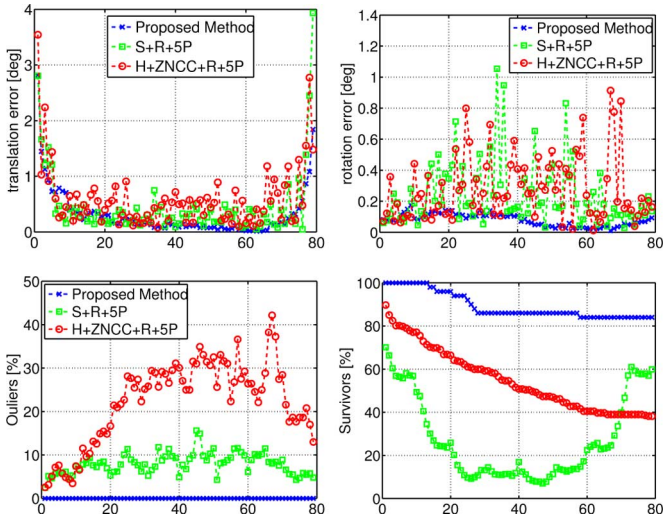


Fig. 4. Results obtained from the proposed approach and traditional methods for the PYRAMID sequence. (Top) Errors in the recovered motion. Relatively larger errors were obtained from traditional methods for large displacements and illumination changes. (Bottom) Percentage concerning the exploited regions and features. The notion of an outlier is made uniform here by using the same threshold for both features and any pixel of a region.

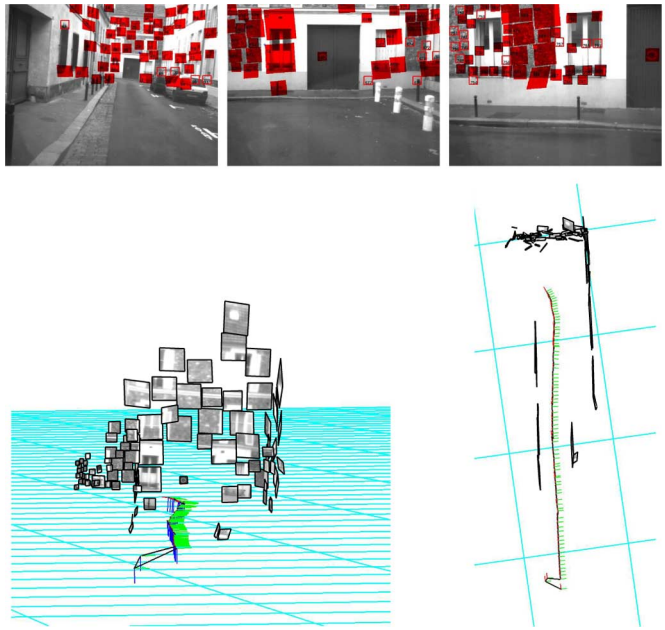


Fig. 5. (Top) Excerpts from the 81-frame CANYON sequence superimposed with the regions registered (in red) by using the proposed approach. Observe the significant change in scale between first and last image. (Bottom) Reconstructed structure and motion seen from different viewpoints. Recovered poses are represented by three-color frames, and only the most stable regions are shown. See the parallelism and/or perpendicularity between most of them.

illumination) were not very small, which violate their assumptions, that algorithm failed at the beginning of the sequence. Our solution is able to deal with larger interframe displacements. The method proposed in [11] could not be applied since the scene is supposed to be unknown, and it is not possible to alter the environment (it needs a known target for the initialization).

B. HANGAR *Sequence*

The application of the proposed technique to this outdoor sequence (see Fig. 1) also has a twofold objective. First, it aims at offering a didactic overview of the method, especially concerning the insertion of new information (the second region). Second, it shows its degree of robustness to different kinds of noise, e.g., shaking motion, image blur, etc. Very importantly, although we model the scene as a collection of planar regions, some occluding nonplanar objects have appeared throughout the sequence, e.g., see the tree in Fig. 1(a). These disturbances have not significantly perturbed the estimation process since they carry substantially less information compared to other parts of the patches. For the requested accuracy, the approach performed along the sequence a median of five iterations, and returned a median photometric error of 13.37 levels of grayscale. The recovered angle between the two walls is of 89.7°, using a median of 22.59% of each (320×240) image. This geometric measure is also an important benchmark for evaluating the technique (considering that these walls are truly perpendicular), since pose and structure are intimately tied together. The total displacement of the camera is of approximately 50 m, and the images were captured by a hand-held camcorder at 25 Hz.

C. CANYON *Sequence*

We also run the proposed algorithm on a representative urban sequence, captured at approximately 12 Hz. It is also a challenging sequence in the sense that large interframe displacements are carried out, the objects are disposed at very different distances from the camera, and because there exists a significant change in scale. Furthermore, it corresponds to a typical urban scenario where cameras can be of particular importance for localization: narrow streets. In this case, positions from GPS may not be available or not sufficiently reliable. The obtained results are shown in Fig. 5, where the visual SLAM is successfully performed. The starting image was chosen such that the dominant plane is further away from the initial camera pose, compared to [17]. This choice aims to show the limitation of the optimization approach, which is local by nature. Notice that in the beginning of the task, despite the fact that the regions are effectively aligned in the images, the recovered motion and structure are not coherent with the true ones (see first camera poses in Fig. 5). This means that the algorithm got wedged in a local minimum. Thanks to the solution proposed in Section IV-B, this minimum is adequately treated and the correct parameters are subsequently obtained. For the requested accuracy, the approach performed along the sequence a median of 12 iterations, returned a median photometric error of 10.77 levels of grayscale, used a median of 34 image regions of size 31×31 pixels (at the time they are selected), and exploited a median of 17.01% of each (760×578) image. The total displacement of the camera is of approximately 60 m.

D. ROUND-ABOUT *Sequence*

This sequence is also illustrative since other different types of noise are present, e.g., pedestrians and moving vehicles.

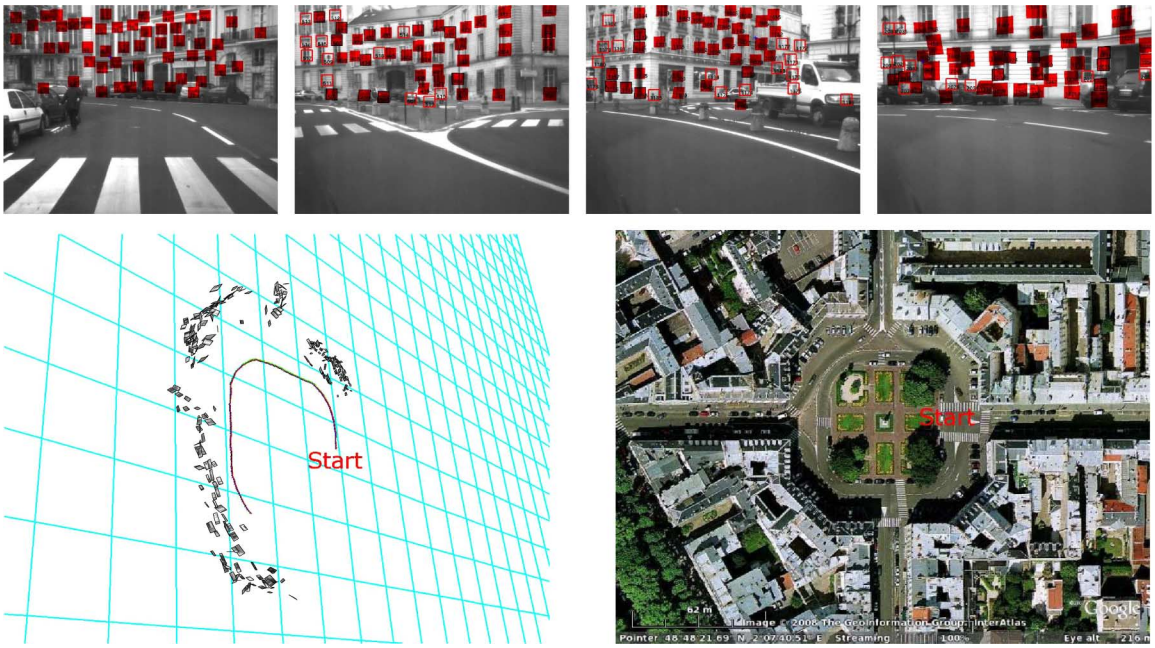


Fig. 6. (Top) Excerpts from the 230-frame ROUND-ABOUT sequence superimposed with the regions aligned (in red) by using the proposed approach. Observe the presence of a pedestrian in the first image and a moving car in the third image. (Bottom left) Reconstructed structure and motion. Recovered poses are represented by very small frames. (Bottom right) Satellite image of the scenario. The path length is of approximately 150 m.

Nevertheless, the technique automatically coped with such outliers. Excerpts from this sequence and the obtained SLAM results can be seen in Fig. 6. We can observe that coherent motion and structure are recovered. For the requested accuracy, the approach performed along the sequence a median of ten iterations, returned a median photometric error of 11.37 levels of grayscale, used a median of 37 image regions of size 31×31 pixels (at the time they are selected), and exploited a median of 10.84% of each (760×578) image. This sequence was captured at approximately 12 Hz by a camera-mounted car, where the path length measured by Google Earth is of approximately 150 m.

VII. CONCLUSION AND PERSPECTIVES

In this paper, we have proposed a different formulation of the vision-based SLAM problem. The technique is based on image alignment (i.e., image registration) using appropriate motion, structure, and illumination parameters, without first having to find feature correspondences. The major advantages and limitations of this approach are described here. Namely, the strengths concern its high accuracy and absence of feature extraction process. Additionally, we have proved that standard methods need to add more frequently new features to track, especially under either significant lighting variations or lengthy camera displacements. Hence, the proposed method reduces the drift by maintaining for longer the estimation of the displacement with respect to the same reference frame. On the other hand, in order to be tractable in real time, we use a local optimization procedure to obtain the related parameters. Alternatives to avoid getting trapped in local minima are then discussed in the paper. Another important research topic regards loop closure, which was not the objective of this paper. Nevertheless, we believe that the proposed direct technique is promising since existing ones (which have a smaller convergence domain) have already performed this task. Other future works may also focus on merging/growing regions with similar structure, which may lead to more stable and faster estimates.

REFERENCES

[1] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: A factorization method," *Int. J. Comput. Vis.*, vol. 9, no. 2, pp. 137–154, 1992.
[2] P. H. S. Torr and A. Zisserman, "Feature based methods for structure and motion estimation," in *Proc. Workshop Vis. Algorithms: Theory Pract.*, 1999, pp. 278–294.
[3] T. J. Broida, S. Chandrashekhar, and R. Chepalla, "Recursive 3-D motion estimation from a monocular image sequence," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 26, no. 4, pp. 639–656, Jul. 1990.
[4] A. Davison, "Real-time simultaneous localization and mapping with a single camera," in *Proc. Int. Conf. Comput. Vis.*, 2003, pp. 1403–1410.
[5] E. Eade and T. Drummond, "Scalable monocular SLAM," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun.17–22, 2006, vol. 1, pp. 469–476.
[6] R. C. Smith and P. Cheeseman, "On the representation and estimation of spatial uncertainty," *Int. J. Robot. Res.*, vol. 5, no. 4, pp. 56–68, 1986.
[7] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, "FastSLAM 2.0: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges," in *Proc. Int. Joint Conf. Artif. Intell.*, Acapulco, Mexico, Aug. 2003, pp. 1151–1156.
[8] A. R. Bruss and B. K. P. Horn, "Passive navigation," *Comput. Vis., Graph., Image Process.*, vol. 21, no. 1, pp. 3–20, 1983.
[9] R. Hummel and V. Sundareswaran, "Motion parameter estimation from global flow field data," *IEEE Pattern Anal. Mach. Intell.*, vol. 15, no. 5, pp. 459–476, May 1993.
[10] M. Irani and P. Anandan, "About direct methods," in *Proc. Workshop Vis. Algorithms: Theory Pract.*, Corfu, Greece, Sep.1999, pp. 267–277.
[11] N. D. Molton, A. J. Davison, and I. D. Reid, "Locally planar patch features for real-time structure from motion," presented at the Br. Mach. Vis. Conf. (BMVC), Kingston, U.K., Sep. 2004.
[12] H. Jin, P. Favaro, and S. Soatto, "A semidirect approach to structure from motion," *Vis. Comput.*, vol. 6, pp. 377–394, 2003.
[13] C. Mei, S. Benhimane, E. Malis, and P. Rives, "Constrained multiple planar template tracking for central catadioptric cameras," in *Proc. Br. Mach. Vis. Conf.*, Edinburgh, U.K., Sep. 2006, pp. 4–7.

[14] R. Szeliski and P. H. S. Torr, "Geometrically constrained structure from motion: Points on planes," in *Proc. Eur. Workshop 3-D Struct. Mult. Images Large-Scale Environ.*, 1998, pp. 171–186.
[15] O. Faugeras, *Three-Dimensional Computer Vision—A Geometric Viewpoint*. Cambridge, MA: MIT Press, 1993.
[16] G. Silveira, E. Malis, and P. Rives, "Real-time robust detection of planar regions in a pair of images," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Beijing, China, Oct. 2006, pp. 49–54.
[17] G. Silveira, E. Malis, and P. Rives, "An efficient direct method for improving visual SLAM," in *Proc. IEEE Int. Conf. Robot. Autom.*, Rome, Italy, Apr. 10–14, 2007, pp. 4090–4095.
[18] F. W. Warner, *Foundations of Differential Manifolds and Lie Groups*. New York: Springer-Verlag, 1987.
[19] S. Benhimane and E. Malis, "Integration of Euclidean constraints in template based visual tracking of piecewise-planar scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Beijing, China, Oct. 2006, pp. 1218–1223.
[20] S. Baker, R. Gross, and I. Matthews, "Lucas–Kanade 20 years on: A unifying framework: Part 3," Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-RI-TR-03-35, 2003.
[21] G. Silveira and E. Malis, "Real-time visual tracking under arbitrary illumination changes," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Minneapolis, MN, Jun. 17–22, 2007, pp. 1–6.
[22] E. Malis, "Improving vision-based control using efficient second-order minimization techniques," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 26/May 1, 2004, vol. 2, pp. 1843–1848.
[23] G. Silveira, E. Malis, and P. Rives, "The efficient E-3D visual servoing," *Int. J. Optomechatron.*, vol. 2, no. 3, pp. 166–184, Jul. 2008.
[24] D. Nister, "An efficient solution to the five-point relative pose problem," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun.18–20, 2003, vol. 2, pp. II-195–II-202.

Geraldo Silveira received the B.Sc. (Hons.) degree from the State University of Campinas (UNICAMP), Sao Paulo, Brazil, and the M.Sc. degree from the Federal University of Rio Grande do Norte (UFRN), Rio Grande do Norte, Brazil, in 2000 and 2002, respectively, both in electrical engineering. He is currently working toward the Ph.D. degree at the Ecole des Mines de Paris (ENSMP) and the Institut National de Recherche en Informatique et en Automatique (INRIA), Sophia-Antipolis, France.

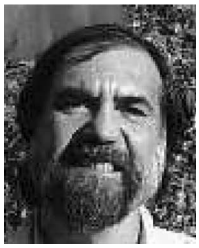
In 2002, he joined the Centro de Pesquisas Renato Archer (CenPRA), Sao Paulo, as a Research Engineer. His current research interests include computer vision, vision-based control, and robotics.

Mr. Silveira was ranked in Top 10 of 2002 by the Brazilian Computer Society for the M.Sc. thesis. In 2004, he received the Best Master's Thesis of 2001–2003 Award endowed by SIEMENS.

Ezio Malis (A'03) received the Graduate degrees in electronics and automatics from the University Politecnico di Milano, Milan, Italy, and the Ecole Supérieure d'Electricité (Supélec), Gif-Sur-Yvette, Paris, France, both in 1995 and the Ph.D. degree in computer vision and robot control from the University of Rennes, Rennes, France, in 1998.

In 2000, he joined the Institut National de Recherche en Informatique et en Automatique (INRIA), Sophia-Antipolis, France, as a Research Scientist. Prior to this, he spent two years as a Research Associate at the University of Cambridge, Cambridge, U.K. His current research interests include automatics, robotics, computer vision, and particular vision-based control.

Dr. Malis was the receipient of the IEEE King-Sun Fu Memorial Best Transactions Paper Award and the IEEE Wegbreit Best Vision Paper Award in 2002.



Patrick Rives (M'04) received the Doctorat de 3ième cycle degree in robotics from the Université des Sciences et Techniques du Languedoc, Montpellier, France, in 1981 and the Habilitation à diriger les recherches degree from the Université de Nice, Nice, France, in 1991.

He was a Research Fellow with the Institut National de la Recherche Scientifique (INRS) Laboratory, Montreal, QC, Canada, for one year. In 1982, he joined the Institut National de Recherche en Informatique et en Automatique (INRIA), Rennes, France. He is the Research Director at INRIA Sophia Antipolis-Méditérranée, Rennes, and the Head of the project team Advanced Robotics and Autonomous Systems (ARobAS). His main research interests include sensor-based control applied to the navigation and the control of mobile robots with a particular emphasis on sensor-based control techniques. He has also addressed the problems of autonomous navigation and SLAM for aerial, underwater, and urban vehicles.