Proceedings of the 2006 IEEE/RSJ
International Conference on Intelligent Robots and Systems
October 9 - 15, 2006, Beijing, China

Real-time Robust Detection of Planar Regions in a Pair of Images

Geraldo Silveira [*,†], Ezio Malis [*], Patrick Rives [*]

[*] INRIA Sophia-Antipolis – Project ICARE
2004 Route des Lucioles, BP 93
06902 Sophia-Antipolis Cedex, France
FirstName.LastName@sophia.inria.fr

[†] CenPRA Research Center – DRVC Division
Rod. Dom Pedro I, km 143,6, Amarais
CEP 13069-901, Campinas/SP, Brazil
Geraldo.Silveira@cenpra.gov.br

***Abstract*— This work presents a method for segmenting image patches which correspond to planar regions in the scene. The method consists of an efficient and robust solution for detecting multiple planar regions in a global optimal sense. Moreover, in contrast with existing techniques which also work on intensity images, neither assumptions about the scene are made nor heuristic hypotheses are formulated. More specifically, the proposed method is based on a systematic, progressive voting procedure from the solution of a linear system, which exploits the two-view geometry. Hence, besides avoiding intermediary depth maps, the progressive mechanism together with such a convergence mapping drastically reduce the computational and storage complexities of the approach. Results from both synthetic and real-world scenes in different scenarios and under various kinds of strong noise confirm its effectiveness and robustness against large camera calibration errors and to the presence of outliers.**

I. Introduction

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 [1]. For this reason, this paper focus on the detection of image patches which correspond to the projections of these planar regions in the scene. The task is performed by using a pair of images, which are not necessarily captured by a stereo rig. That is, the two images could be acquired by a single moving camera. To solve the problem, the most of the techniques that have been proposed in the literature compute the depth map as a preliminary step [2], [3], [4] (or the disparity map as in [5]). On the contrary, the method we propose use directly the intensity images. Therefore, by working directly with the image we gain computational efficiency and we avoid error propagation. Similar strategies which also work on intensity images, make some assumptions about the scene. For example, in [6] and [7] the authors assume the presence of lines in the image, which is a valid assumption for many man-made structures although limiting their applicability. In fact, besides not requiring structured scenes, other scene constraints are also not assumed here, such as perpendicularity or verticality constraints [8] and symmetry on the imaged object [9], as well as multiple hypotheses formulation and testing [4] are also not performed. In the paper, we consider robotic applications of the algorithm, as for example plane-based template tracking and robot pose recovery. Thus, robustness aspects as well as real-time performance are of particular importance. In [10], the authors propose to use projective invariants defined by quintuples of assumed coplanar points. However, as remarked in [7], its main drawback is their sensitivity to errors in the localization of image points. Thus, it is not suited to our purposes given its lack of robustness. In addition, we do not assume a particular configuration of the camera with respect to the scene. For example, having a car-mounted camera always pointing towards the road plane can be a constraint that can help to reduce the complexity of the problem. Instead, a generic algorithm is developed here, which resorts to an efficient robust technique. The proposed approach makes use of the basic equation that links the projection of a same scene point onto a pair of images. By rewriting such equation in linear form and by using triplets of corresponding image features likely to be coplanar, a systematic, progressive voting procedure is proposed to partition the image into multiple highly reliable planar seed regions. Robust techniques to tackle multistructured data in a global optimal sense are generally designed by means of voting methods [11]. Within the guess-and-test paradigm of RANSAC, for example, one only searches for an inlier/outlier dichotomy. Within a voting framework, on the other hand, accumulation of evidence and management of the parameter space are performed globally. However, contrarily to the standard Hough transform, a pixel in the image is not mapped into all points on a hypersurface in the parameter space (divergence mapping). This mapping is the main source of computational and memory inefficiency of such a transform. Here, the solution of the carefully assembled linear system maps to a single point (convergence mapping). See Fig. 1 for an illustration in the Cartesian space, although the approach uses image features instead of 3D points. In this paper, it is then shown how to perform such convergence mapping by exploiting the plane-based two-view geometry. Various advantages of such a mapping is discussed here. Moreover, computational and storage complexities are shown to be further benefited from a progressive mechanism. On effect, a planar region is segmented as soon as the contents of the parameter space allow for such a decision. Indeed, this also contributes for operating over real-time systems since voting and plane detection are interleaved processes and thus, permitting the algorithm to be interrupted at any time and still providing useful information. Furthermore, besides the labeled features, all of them which also verify the plane-based projective equations are removed from input data, therefore considerably reducing time complexity as well. This also represents an attractive feature of the algorithm since the complexity of the proposed Planar Region Detector (PRD) is dependent on the complexity of the image. For example,

1-4244-0259-X/06/$20.00 ©2006 IEEE

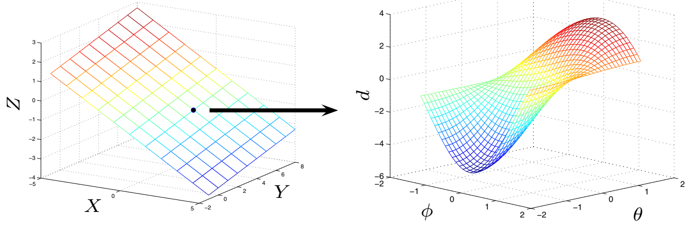

Fig. 1. Illustration of the divergence mapping performed by a standard Hough transform to detect planes. Instead of mapping a point to an hypersurface, if a convergence mapping is deployed then three points would map to a single point. Various advantages of this latter mapping are discussed in the text.

in ideal conditions and if the scene contains a single plane, the storage complexity is of order 1, instead of being cubic with respect to the discretization size. Results from both real images as well as from synthetic scenes under various kinds of strong noise confirm its efficiency and robustness against large camera calibration errors, and to the presence of noisy, mismatched features (outliers).

This paper is organized as follows. Section II presents some modeling aspects whereas the proposed approach is formulated in Section III. The results are then shown and discussed in the Section IV. Finally, Section V summarizes the article and some references are afterward given.

II. Theoretical Background

Let $\mathcal{F}$ be the camera frame whose origin $\mathcal{O}$ coincides with the center of projection $\mathcal{C}$, and whose plane $(\vec{x}, \vec{y})$ is parallel to the plane of projection. Suppose that $\mathcal{F}$ is displaced with respect to another coordinate system $\mathcal{F}'$ in the Euclidean space by $\mathbf{R} \in SO(3)$ and $\mathbf{t} = [t_x, t_y, t_z]^\top \in \mathbb{R}^3$, respectively the rotation matrix and the translation vector. The notation $[\mathbf{a}]_\times$ represents the skew symmetric matrix associated to vector $\mathbf{a}$, whereas $\{a_i\}_{i=1}^k$ corresponds to the set $\{a_1, a_2, \ldots, a_k\}$. Also, $\mathcal{R}(\mathbf{A})$ and $\mathcal{N}(\mathbf{A})$ denote respectively the range and the null space of a matrix $\mathbf{A}$, $(\mathbf{A}^{-1})^\top = (\mathbf{A}^\top)^{-1}$ is abbreviated by $\mathbf{A}^{-\top}$, and $\mathbf{0}$ is a matrix of zeros of appropriate dimensions.

A. Camera Model

Consider the pinhole camera model. In this case, a 3D point with homogeneous coordinates $\mathbf{P}_i = \left[X_i, Y_i, Z_i, 1\right]^\top$ defined with respect to frame $\mathcal{F}$, $i = 1, 2, \ldots, N$, is projected onto the image space $\mathcal{I} \subset \mathbb{R}^2$ as a point with pixels homogeneous coordinates $\mathbf{p}_i \in \mathbb{P}^2$ through

$$\mathbf{p}_i = [u_i, v_i, 1]^\top \propto \mathbf{K} \left[\begin{array}{cc} \mathbf{I}_3 & \mathbf{0} \end{array}\right] \mathbf{P}_i, \tag{1}$$

where $\mathbf{K} \in \mathbb{R}^{3\times3}$ is an upper triangular matrix that gathers the camera intrinsic parameters

$$\mathbf{K} = \begin{bmatrix} f & fs & u_0 \\ 0 & fr & v_0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{2}$$

with focal lengths f in pixels, principal point $\mathbf{p}_0 = [u_0, v_0, 1]^\top$ in pixels, aspect ratio r and skew s. Correspondingly, the same point $\mathbf{P}_i \in \mathbb{P}^3$ is projected onto the image space $\mathcal{I}' \subset \mathbb{R}^2$ associated to $\mathcal{F}'$ as

$$\mathbf{p}'_i = [u'_i, v'_i, 1]^\top \propto \mathbf{K} \left[\begin{array}{cc} \mathbf{R} & \mathbf{t} \end{array}\right] \mathbf{P}_i. \tag{3}$$

Then, from the general rigid-body equation of motion along with (1) and (3), it is possible to obtain the fundamental relation that links the projection of $\mathbf{P}_i$ onto both images:

$$\mathbf{p}'_i \propto \mathbf{K}\,\mathbf{R}\,\mathbf{K}^{-1}\mathbf{p}_i + Z_i^{-1}\,\mathbf{K}\,\mathbf{t}. \tag{4}$$

B. Plane-based Two-view Geometry

Consider the normal vector description of a plane $\boldsymbol{\pi} = \left[\mathbf{n}^\top, -d\right]^\top \in \mathbb{R}^4 : \|\mathbf{n}\| = 1, d > 0$. Let $\boldsymbol{\pi}$ (resp. $\boldsymbol{\pi}'$) be defined with respect to frame $\mathcal{F}$ (resp. $\mathcal{F}'$). If a 3D point $\mathbf{P}_i$ lies on such planar surface then

$$\mathbf{n}^\top\,\mathbf{P}_i = \mathbf{n}^\top Z_i\,\mathbf{K}^{-1}\mathbf{p}_i = d, \tag{5}$$

and hence

$$Z_i^{-1} = d^{-1}\,\mathbf{n}^\top \mathbf{K}^{-1}\mathbf{p}_i. \tag{6}$$

By plugging (6) into (4), a projective mapping $\mathbf{H} : \mathbb{P}^2 \mapsto \mathbb{P}^2$ (also referred to as the projective homography) defined up to a non-zero scale factor is achieved:

$$\mathbf{p}'_i \propto \mathbf{H}\,\mathbf{p}_i \tag{7}$$

with $\mathbf{H} = \mathbf{K}\,\left(\mathbf{R} + d^{-1}\,\mathbf{t}\,\mathbf{n}^\top\right)\,\mathbf{K}^{-1}$.

Remark 2.1: Such an homographic mapping is obtained, independently if the object is planar or not, if $\mathbf{t} = \mathbf{0}$ (i.e. the camera undergoes a pure rotation motion). In this case, $\mathbf{H} = \mathbf{K}\,\mathbf{R}\,\mathbf{K}^{-1}$ and depth information is completely lost.

III. The Proposed Planar Region Detector

The proposed Planar Region Detector (PRD) is based on a progressive voting procedure which performs a convergence mapping by using the solution of a linear system. This section presents how such a linear system is assembled and also derives the conditions to perform plane detection by using a pair of images. In addition, further details on constraining the search space are discussed in the next subsections.

A. How a Vote is Performed

The vote is the solution of a linear system, which is derived from the following. Equation (4) together with (6) allows for rewriting the fundamental equation that links the projection of the same 3D point $\mathbf{P}_i$ in a pair of images as

$$\mathbf{p}'_i \propto \mathbf{H}_\infty\,\mathbf{p}_i + \mathbf{e}\,\mathbf{n}_d^\top \mathbf{K}^{-1}\,\mathbf{p}_i, \tag{8}$$

with the projective homography of the plane at infinity $\mathbf{H}_\infty = \mathbf{K}\,\mathbf{R}\,\mathbf{K}^{-1}$, the epipole $\mathbf{e} = \mathbf{K}\,\mathbf{t}$, and the normal vector scaled by the distance to $\mathcal{F}$, $\mathbf{n}_d = \mathbf{n}/d$. Next, define the normalized version of the latter:

$$\mathbf{x} \triangleq \mathbf{K}^{-\top}\mathbf{n}_d. \tag{9}$$

By pre-multiplying both members of (8) by $[\mathbf{p}'_i]_\times$, and knowing that $\mathbf{x}^\top\mathbf{p}_i = \mathbf{p}_i^\top\mathbf{x}$, the following linear system is achieved:

$$\mathbf{A}_i\,\mathbf{x} = \mathbf{b}_i, \tag{10}$$

with

$$\begin{cases} \mathbf{A}_i = [\mathbf{p}'_i]_\times \mathbf{e}\,\mathbf{p}_i^\top \\ \mathbf{b}_i = -[\mathbf{p}'_i]_\times \mathbf{H}_\infty\,\mathbf{p}_i. \end{cases} \tag{11}$$

Notice also that such a system of equations is fully defined in the image space. However, matrix $\mathbf{A}_i \in \mathbb{R}^{3\times3}$ has maximum rank 1 since it can be seen as a product of two 3-vectors, i.e. as $\mathbf{A}_i = \mathbf{c}_i\mathbf{p}_i^\top$, where $\mathbf{c}_i = [\mathbf{p}'_i]_\times\mathbf{e}$. This is an obvious statement from a geometric point of view since at least 3 points are needed to constraint the 3 dofs of a plane (2 dofs for $\mathbf{n}$ and 1 dof for d). Hence, the parameters related to a plane is recovered by stacking three Eqs. (10), one for each pair of corresponding points $\mathbf{p}'_i \leftrightarrow \mathbf{p}_i$, which gives

$$\bar{\mathbf{A}}\,\mathbf{x} = \bar{\mathbf{b}} \tag{12}$$

with $\bar{\mathbf{A}} = \left[\{\mathbf{A}_i\}_{i=1}^3\right] \in \mathbb{R}^{9\times3}$ and $\bar{\mathbf{b}} = \left[\{\mathbf{b}_i\}_{i=1}^3\right] \in \mathbb{R}^9$. The solution of such a rectangular linear system is obtained in the least-squares sense by solving its normal equations

$$\bar{\mathbf{A}}^\top\bar{\mathbf{A}}\,\mathbf{x} = \bar{\mathbf{A}}^\top\,\bar{\mathbf{b}}, \tag{13}$$

which is performed extremely fast given its low dimensionality. Furthermore, if noise is not too large then those 9 equations can be reduced to 6 by using only the first 2 equations of each $\mathbf{A}_i$. The linearly independent equation is either the first or the second one. Now, it is important to study in which conditions the solution (the vote) of such a system is unique.

Proposition 3.1 (Existence and uniqueness of solution): The assembled linear system (13) from 3 pairs $\mathbf{p}'_i \leftrightarrow \mathbf{p}_i$ is consistent and has a unique solution if:

- $\mathbf{t} \neq \mathbf{0}$;
- the 3 points are non-collinear.

Proof: First of all, associated systems of normal equations are always consistent since $\bar{\mathbf{A}}^\top\bar{\mathbf{b}} \in \mathcal{R}(\bar{\mathbf{A}}^\top) = \mathcal{R}(\bar{\mathbf{A}}^\top\bar{\mathbf{A}})$. Thus, we only need to proof the uniqueness of solution for such a system under those conditions. The proof consists in demonstrating that $\mathcal{N}(\bar{\mathbf{A}}^\top\bar{\mathbf{A}}) = \mathcal{N}(\bar{\mathbf{A}}) = \mathbf{0}$ or, equivalently, that $\bar{\mathbf{A}}$ is a full rank matrix if those conditions are verified. We start by observing that $\mathbf{t} \neq \mathbf{0}$ is a necessary and sufficient condition to avoid a null coefficient matrix $\bar{\mathbf{A}}$. This can be seen directly from its submatrices in (10). In fact, as pointed in Remark 2.1, if $\mathbf{t} = \mathbf{0}$ then the entire image corresponds to the plane at infinity $\boldsymbol{\pi}_\infty$, since there exist a solution such that $\lim_{\|\mathbf{x}\|\to0^+} d = 1/\|\mathbf{K}^\top\mathbf{x}\| = \infty$, using (9). However, this is a necessary but not a sufficient condition to guarantee that $\mathrm{rank}(\bar{\mathbf{A}}) = 3$. In fact, $\exists\mathbf{y} \neq \mathbf{0} : \bar{\mathbf{A}}\,\mathbf{y} = \mathbf{0}$ when the third image point is a linear combination of the first two, i.e. $\mathbf{p}_3 = \alpha\mathbf{p}_1 + \beta\mathbf{p}_2,\ \alpha,\beta \neq 0$. In this case, $\mathbf{y} = \gamma[\mathbf{p}_1]_\times\mathbf{p}_2,\ \forall\gamma \neq 0,$ is such a vector. Hence, if an image point is collinear with the others, then $\bar{\mathbf{A}}$ is rank-deficient. ■

Therefore, after verifying the conditions stated in the Proposition 3.1, the normal vector of the plane and its distance to $\mathcal{F}$ described by a certain triplet of points are obtained from the solution of (13), $\mathbf{x} = (\bar{\mathbf{A}}^\top\bar{\mathbf{A}})^{-1}\,\bar{\mathbf{A}}^\top\,\bar{\mathbf{b}}$, and Eq. (9) as

$$\begin{cases} d = \|\mathbf{n}_d\|^{-1} \\ \mathbf{n} = \mathbf{n}_d\,d. \end{cases} \tag{14}$$

This is then used to perform the convergence mapping, i.e. to perform a single vote instead of voting the whole parameter space (see Fig. 1). In fact, a transformation from Cartesian to orthogonal-axis coordinate system is used: the unit normal vector is written as a function of the tilt and slant angles, i.e. $\mathbf{n} = \mathbf{n}(\phi,\theta) \in [-\pi/2,\pi/2] \times [-\pi/2,\pi/2]$. Indeed, the solution (14) of the k-th triplet of points may be stored e.g. as a member of a linear list together with its number of votes s (the score), i.e. as a member of $\mathcal{S} = \left\{[\mathbf{n}_k^\top, d_k, s_k]\right\}$. If a particular solution already exists according to a given resolution, then its corresponding score is simply incremented: $s_k \leftarrow s_k + 1$; otherwise a new member is appended to the set of solutions $\mathcal{S}$ with $s_k = 1$. This convergence mapping presents several other advantages. First of all, a huge parameter space does not need hence to be allocated. The storage is performed dynamically as systems are solved. A second advantage of such mapping is that the parameter space does not need to be defined a priori. Two votes are regarded as the same according to an error criterion without boundaries on the parameter space, which means an infinite range for such a space. Furthermore, by using the plane-based two-view geometry within a progressive mechanism, as demonstrated in the next subsection, an enormous reduction of the computational complexity is yielded.

In addition, given the well-known robustness characteristics of voting procedures, even if the set of camera parameters are uncalibrated (instead of determining $\mathbf{e}$ and $\mathbf{H}_\infty$ in the image), i.e. only a coarse estimate $\left\{\widehat{\mathbf{K}}, \widehat{\mathbf{R}}, \widehat{\mathbf{t}}\right\}$ is provided and there exist outliers, it is still possible to cluster planar regions in the image provided that the conditions stated above are verified. See both the simulation and experimental results in the Section IV. This robustness property is an attractive characteristic of the approach since it is able to tolerate large errors in its inputs.

B. A Progressive Procedure for Fast Plane Segmentation

This section shows why a progressive procedure further contributes for reducing the time complexity, which becomes in fact dependent on the image. As it will be demonstrated, all possible combinations of 3 points do not need necessarily to be voted. On effect, a plane is clustered as soon as the contents of the accumulator permit such a decision, which involves checking if the score is large enough together with a plane verification step (this latter is described in next subsection).

Consider the symmetric transfer error derived from the projective mapping expressed in (7)

$$e_i^2(\mathbf{p}_i,\mathbf{p}'_i,\mathbf{H}) = \left\|\mathbf{p}'_i - \mathbf{H}\,\mathbf{p}_i\right\|^2 + \left\|\mathbf{p}_i - \mathbf{H}^{-1}\,\mathbf{p}'_i\right\|^2. \tag{15}$$

The upper bound on the number of votes, which would achieve a time complexity of $O(N^3)$ from the binomial coefficient $\binom{N}{3}$, is significantly reduced by using a progressive mechanism since:

- A sliding spatial subdivision of the image is performed. That is, instead of a prior division of the image, a local grouping is performed by a function $\varphi : \mathbb{P}^2 \to \mathbb{R}$ (e.g. geometric- or photometric-based) so that only an open disk $\mathcal{D} \subset \mathcal{I}$ of radius $r, R > 0$ centered at $\mathbf{p}_i$, i.e.

$$\mathcal{D}_{r,R}(\mathbf{p}_i) = \left\{\forall\mathbf{p} \in \mathcal{I} : r < \|\varphi(\mathbf{p}) - \varphi(\mathbf{p}_i)\| < R\right\}, \tag{16}$$

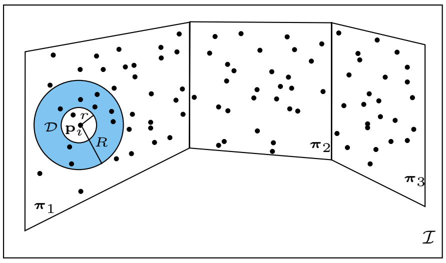

Fig. 2. Illustration of a geometric-based local grouping in the image by using a disk $\mathcal{D}$ of radius r, $R > 0$ centered at $\mathbf{p}_i$. In the case of a photometric measure, the intensity of the pixel is who plays the role in devising the region.

is considered at a time. See Fig. 2 for an illustration. Notice that this procedure, besides grouping points likely to be coplanar and reducing complexity, it also contributes to avoid clustering dominant (virtual) planes;

- very importantly, after a bin reaches a sufficiently large score, i.e. $s_k > \varepsilon$, and $\mathbf{H}$ is afterward calculated, all the points which also verify the plane-based two-view geometry are removed from input data. Hence, an enormous reduction of the computational complexity is yielded. See Fig. 3 for a simple example. In other words, a region growing is performed by using (15) in order to segment the plane $\boldsymbol{\pi}_j$, $j = 1, 2, \ldots, M$, i.e.

$$\text{proj}(\boldsymbol{\pi}_j) \subseteq \left\{ \forall \mathbf{p}_i \in \mathcal{I} : e_i^2(\mathbf{p}_i, \mathbf{p}_i', \mathbf{H}) < \epsilon^2 \right\}, \qquad (17)$$

where the operator proj$(\cdot)$ represents the perspective projection of the entity. For the χ^2 distribution with a probability $p = 0.95$ and standard deviation σ, the threshold may be given as $\epsilon^2 = 5.99\sigma^2$. Such a step also guarantees that, from a local clustering, a global segmentation is achieved;

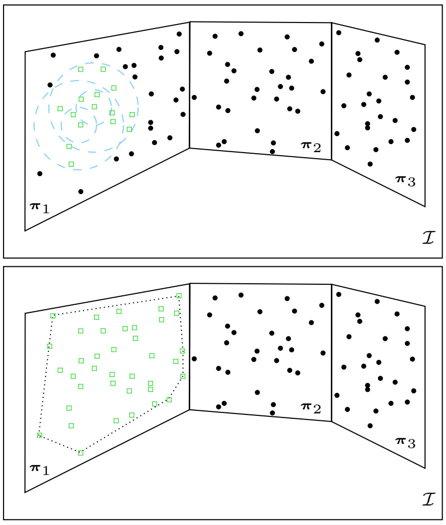

Fig. 3. After detecting a planar seed region (first image), a region growing is performed to remove all points which also belong to the plane (bottom image). The top image also depicts the movement of the disk to another image point.

- and finally, given the number of remaining image features $a \leq N$ at a given iteration, the terminating condition $\binom{a}{3} = \frac{a(a-1)(a-2)}{6} \leq \varepsilon$ can be imposed since a minimum number of image points is necessary to perform the detection. That is, if $a < a_i \in \mathbb{R}_+$ then $\nexists \boldsymbol{\pi}_{j+1}$, and the algorithm has finished. The $\{a_i\}_{i=1}^3$ are then the roots of the cubic polynomial $a^3 - 3a^2 + 2a - 6\varepsilon = 0$.

C. Plane Verification

Given the objective of partitioning highly reliable planar regions, a step of plane verification is needed. Indeed, after obtaining robustly which image features belong to a certain plane, whose number is described by $\ell \in \left[\lceil a_i \in \mathbb{R}_+ \rceil, 3\varepsilon\right]$, the smallest convex set containing them (i.e. the convex hull)

$$\mathcal{H} \equiv \left\{ \sum_{i=1}^{\ell} \mu_i \mathbf{p}_i : \mu_i \geq 0, \ \forall i, \ \text{and} \ \sum_{i=1}^{\ell} \mu_i = 1 \right\} \qquad (18)$$

may be used to form the templates in both views. This is achieved with complexity $O(\ell \log \ell)$ since specialized algorithms exist for the 2-dimensional case. The $\lceil \cdot \rceil$ denotes the ceiling function, which gives $\forall x \in \mathbb{R}$ the smallest integer $\geq x$. Plane verification can thus be performed: one template is warped into the other frame in order to have them compared. In this work, the zero-mean normalized cross-correlation score is used as the measure of similarity of the corresponding templates. A candidate plane also fails to pass the verification if its area is too small to be considered. In addition, there exist other advantages of using the convex hull as the plane boundaries. Firstly, one may allow those points to belong to several planes, what would represent their intersection line. Also, an hybrid strategy is hence deployed: image features and image templates are combined in the detector, providing higher reliability and meaningful information.

D. Discussion on the Complexities of the Algorithm

The standard Hough transform is neither computational nor memory efficient given its respective complexities $O(N\,N_a^2)$ and $O(N_a^3)$ for a system with 3 dofs, where N_a is the size of the accumulator. To a large extent, this is due to its divergence mapping. Let N_{min} be the number of features describing the smallest plane in the image. With respect to the computational complexity of the algorithm, by using the results in [12] it can be shown that the worst case is then $O(M\,(\varepsilon N^3/N_{\text{min}}^3)^2)$ even if a simple linear list is used. If hash tables are used then the complexity drops to $O(\varepsilon\,M\,N^3/N_{\text{min}}^3)$. In both cases, one can observe that those complexities are usually considerably smaller than $O(N\,N_a^2)$ and are dependent on the complexity of the data. For example, if a single plane is present in the image and the input data is noiseless, the time complexity is of $10 \sim 15$ instead of being exponential. With respect to the memory complexity, if a simple list is used as storage, it can be shown that it has an upper bound of $O(\varepsilon N^3/N_{\text{min}}^3)$. Such complexity is usually considerably lower than $O(N_a^3)$ too. Using again that simple example, i.e. in ideal conditions and if the scene contains a single plane, only 1 bin is needed. In case of using hash tables as storage, memory complexity drops to $3N_h$, i.e. to $O(N_h)$, where N_h is the length of the tables.

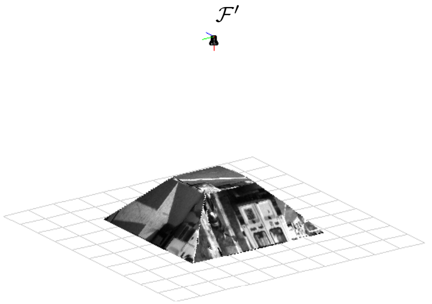

Fig. 4. The textured synthetic scene designed for the systematic tests.

IV. Results and Discussion

In order to assess the performance of the algorithm, we have tested it against a large data set of both simulated and real images. In all cases, the resolution for the normal vector was set to 5° (for ϕ and θ) and to 0.05m for the distance d. With respect to their boundaries, as already stated, they do not need to be defined a priori. Also, the disk parameters were set to $r = 5$ and $R = 50$ pixels. The corresponding interest points can be furnished by e.g. the Harris detector together with a standard correlation-based matching algorithm. In addition, in accordance with probabilistic Hough-like transforms, where as low as 2% of the number of points was used (~ 300 here), the threshold on the minimum number of votes was then set to $\varepsilon = 15$ from $\left(\frac{0.02*300}{3}\right)$. Those parameters remained constant for all the experiments.

A synthetic 3D scene was constructed in order to have a ground truth for a large range of variation for each input variable. However, real textured images were used to simulate realistic situations as closely as possible. The artificially created scene is composed by four planes disposed in pyramidal form, but cut by another plane on its top. Onto each one of the five planes, a different texture was applied (see Fig. 4). The reference camera frame $\mathcal{F}'$ is positioned at the center of the pyramid pointing downwards and whose distance to the farthest plane (the top plane) is of $d = 1$m. This distance does not represent a restricting fact given that is the amount of scaled translation $\|\mathbf{t}\|/d$ between the frames (along with the focal length) which plays an important role for scene reconstruction from a pair of images. This would represent the baseline with respect to depth in case of stereoscopic images. We have then conducted more than 10,000 simulations to investigate the performance of the PRD algorithm. For every simulation, a normally distributed, independent noise η_i with mean 0 and standard deviation $1/6$ is added to every input camera parameter: $\widehat{a}_i = a_i(1 + \frac{1}{6}\eta_i)$. This means that such an input has an error of up to 50% in 99.7% of the cases. From $\mathcal{F}'$, random directions of translation as well as random rotations were used to displace the camera by a varying amount of $\|\mathbf{t}\|/d \in [0.01, 0.5]$. Also, the image may contain fewer planes for large displacements (large baselines), since we do not enforce that all planes are always in the image. The median number of corresponding points $\mathbf{p}'_i \leftrightarrow \mathbf{p}_i$, along with the interquartile range, and of the percentage of outliers in the data are shown in the Fig. 5. A $\mathbf{p}'_i \leftrightarrow \mathbf{p}_i$ is said to be an outlier here if the known warping (ground truth) of the extracted point in the first image and the extracted point in the second view gives an error over 5σ pixels. It was considered that the point detector has a standard deviation of $\sigma = 1$ pixel. Indeed, from such a large, noisy input data set, we have computed two measures for assessing the performance of the PRD. In the Fig. 6, the median number of detected planar regions as well as of the rate of false positives are shown. Firstly, as predicted in the Section III-A, if $\|\mathbf{t}\|$ is too small then *any* scene may be viewed as composed by a single plane (the plane at infinity). That explains the high rate of false positives for $\|\mathbf{t}\|/d = 0.01$. However, for all the other cases, a median of zero false positive planes was achieved. Moreover, notice that this happens even if a large number of outliers is present in the process as well (compare Figs. 5 and Fig. 6 for large displacements), although reducing the number of planes to be detected. Such a result confirms the robustness of the PRD against large errors in the camera parameters and to the presence of outliers.

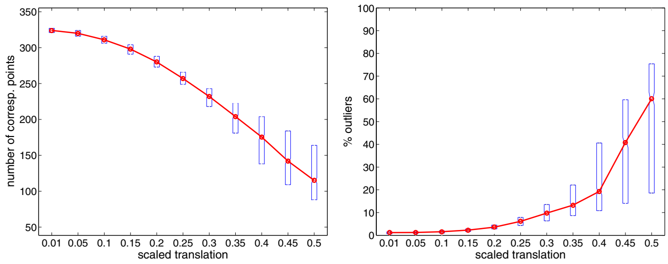

Fig. 5. Median number of corresponding points along with the interquartile range, as well as the median percentage of outliers present in the simulated data, as the amount of the scaled translation is varied.

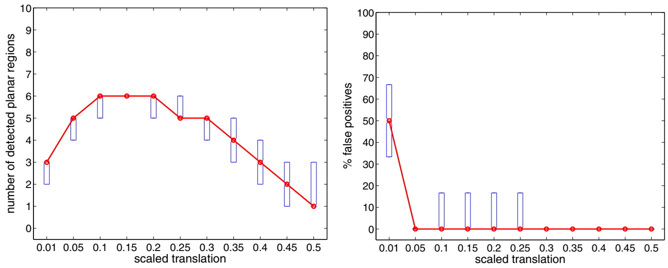

Fig. 6. Median number of detected planar regions and of the rate of false positives obtained from such a large, noisy input data set. The planes are not enforced to be always in the image anywhere in the simulations.

With respect to experimental results, three different scenarios were considered: an indoor scene, an urban scene, as well as an outdoor one. Due to paper length restrictions, only one example for each scenario is provided here. The corresponding detected planar regions by using the algorithm are shown in the first row of the Figs. 7, 8 and 9. For all pairs of images tested, $\widehat{\alpha}_u = \widehat{\alpha}_v = 500$ pixels were used with principal point as the middle of the image, zero skew, as well as $\widehat{\mathbf{R}} = \mathbf{I}_3$ and $\widehat{\mathbf{t}} = [-0.1, 0, -1]^\top$m for the rotation and translation motions, respectively. Albeit these parameters are obviously not the true ones, actual planes were detected, which confirms once again the robustness properties of the approach. Since the approach aims to cluster planar regions *in the image*, large errors on the camera parameters are tolerated. The effect of erroneous camera parameters appears on the values of the $\mathbf{n}_d$ in the Cartesian space. In addition, if precise camera parameters are provided, then accurate, stable, fast scene reconstruction can also be achieved by enforcing the rigidity of the scene as follows. From (7), one obtains that $\mathbf{t}\,\mathbf{n}_d^\top = \alpha\,\mathbf{K}^{-1}\,\mathbf{H}\,\mathbf{K}-\mathbf{R}$. By pre-multiplying both members by the transpose of the translation vector, each segmented planar region, $j = 1, 2, \ldots, M$, is described by:

$$\mathbf{n}_{dj} = \left(\alpha_j\,\mathbf{K}^{-1}\,\mathbf{H}_j\,\mathbf{K} - \mathbf{R}\right)^\top \mathbf{t}/\|\mathbf{t}\|^2, \tag{19}$$

where the factor $\alpha_j \in \mathbb{R}$ is given from the median singular value of $\mathbf{K}^{-1}\,\mathbf{H}_j\,\mathbf{K}$. The bottom images of the Figs. 7, 8 and 9 show the reconstructed scenes from disparate viewpoints, after performing a partial region growing. One can observe that even if no assumptions about the scene were made, perpendicularity and parallelism of the planes were achieved.

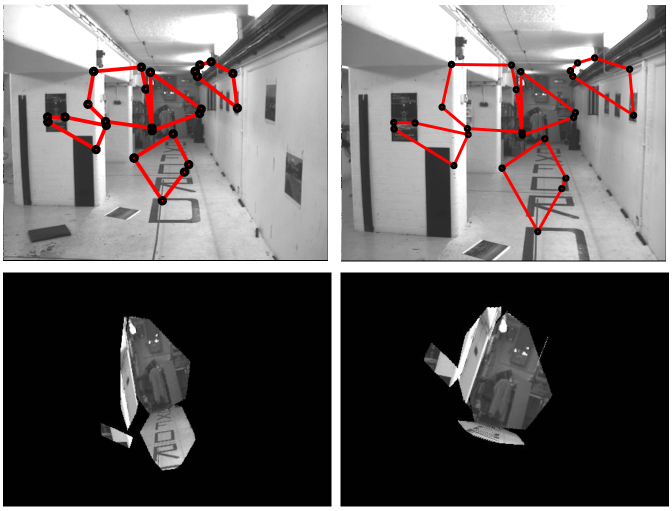

Fig. 7. First row: the Oxford images superposed by the detected planar regions (in red). Bottom images: the reconstructed scene after performing a partial region growing, and rendered from different viewpoints.

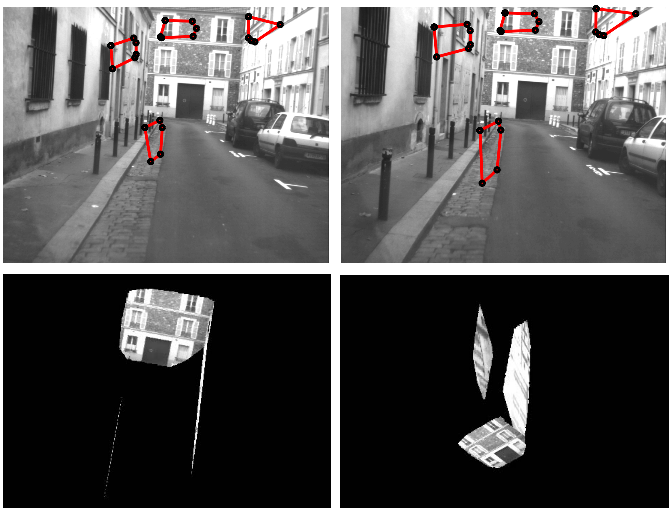

Fig. 8. The Versailles images superposed by the detected planar regions (in red) are shown in the first row. At the bottom, the reconstructed scene, after performing a partial region growing, as seen from different viewpoints.

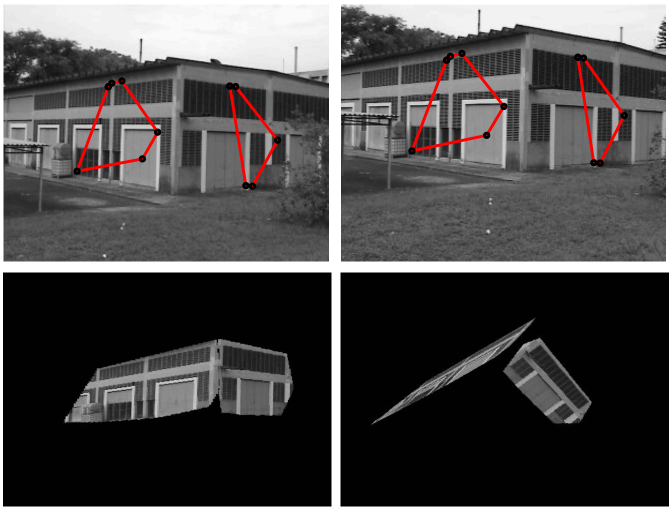

Fig. 9. The Hangar images superposed by the detected planar regions (in red) are shown in the first row. The reconstructed scene is shown at the bottom.

V. Conclusions

A new planar region detector is proposed in this work. In contrast with traditional methods, error propagation is avoided and no assumptions about the scene are made. The approach consists of an efficient robust technique which optimally clusters multiple highly reliable planar seed regions by exploiting the two-view geometry. It features fast speed, small storage, infinite range, high resolution, and very importantly, it is robust to very large errors in its inputs. Experimental results as well as by using synthetic data in different scenarios and under various types of strong noise are shown and discussed. Possible extensions may encompass a random sampling of the image features in order to even further speed up computations.

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

[2] K. Okada *et al.*, "Plane segment finder: Algorithm, implementation and applications," in *Proc. of the IEEE ICRA*, 2001, pp. 2120–2125.

[3] P. J. Besl and R. C. Jain, "Segmentation through variable-order surface fitting," *IEEE Transactions on PAMI*, vol. 10, no. 2, pp. 167–192, 1988.

[4] A. Bartoli, "Piecewise planar segmentation for automatic scene modeling," in *Proc. IEEE Int. Conf. on CVPR*, USA, 2001, pp. 283–289.

[5] E. Trucco, F. Isgrò, and F. Bracchi, "Plane detection in disparity space," in *Proc. of the IEE Int. Conf. on Visual Inf. Eng.*, UK, 2003, pp. 73–76.

[6] C. Baillard and A. Zisserman, "Automatic reconstruction of piecewise planar models from multiple views," in *Proc. IEEE Conf. on Comp. Vision and Patt. Recognition*, 1999, pp. 559–565.

[7] M. Lourakis, A. Argyros, and S. Orphanoudakis, "Plane detection in an uncalibrated image pair," in *Proc. BMVC*, 2002, pp. 587–596.

[8] A. Dick, P. Torr, and R. Cipolla, "Automatic 3D modelling of architecture," in *Proc. BMVC*, 2000, pp. 372–381.

[9] A. Y. Yang *et al.*, "Geometric segmentation of perspective images based on symmetry groups," in *Proc. Int. Conf. on Comp. Vision*, 2003.

[10] D. Sinclair and A. Blake, "Quantitative planar region detection," *International Journal of Computer Vision*, vol. 18, no. 1, pp. 77–91, 1996.

[11] P. Meer, *Emerging Topics in Computer Vision.* Prentice Hall, 2004, ch. Robust techniques for computer vision.

[12] L. Xu and E. Oja, "Randomized Hough Transform (RHT): Basic mechanisms, algorithms, and computational complexities," *CVGIP: Image Understanding*, vol. 57, no. 2, pp. 131–154, 1993.