Piecewise-Planar StereoScan: Structure and Motion from Plane Primitives

Carolina Raposo, Michel Antunes, and Joao P. Barreto

Institute of Systems and Robotics
University of Coimbra, 3030 Coimbra, Portugal

Abstract. This article describes a pipeline that receives as input a sequence of images acquired by a calibrated stereo rig and outputs the camera motion and a Piecewise-Planar Reconstruction (PPR) of the scene. It firstly detects the 3D planes viewed by each stereo pair from semi-dense depth estimation. This is followed by estimating the pose between consecutive views using a new closed-form minimal algorithm that relies in point correspondences only when plane correspondences are insufficient to fully constrain the motion. Finally, the camera motion and the PPR are jointly refined, alternating between discrete optimization for generating plane hypotheses and continuous bundle adjustment. The approach differs from previous works in PPR by determining the poses from plane-primitives, by jointly estimating motion and piecewise-planar structure, and by operating sequentially, being suitable for applications of SLAM and visual odometry. Experiments are carried in challenging wide-baseline datasets where conventional point-based SfM usually fails.

Keywords: Structure and Motion, Piecewise-Planar Reconstruction.

1 Introduction

Although multi-view stereo has been an intensive field of research in the last few decades, current methods still have difficulty in handling situations of weak or repetitive texture, variable illumination, non-lambertian reflection, and high surface slant [11]. In this context, it makes sense to explore the fact that man-made environments are usually dominated by large plane surfaces to improve the accuracy and robustness of 3D reconstruction. This is the key idea behind the so-called Piecewise-Planar Reconstruction (PPR) methods that use the strong planarity assumption as a prior to overcome the above mentioned issues [11,9,22,2,26,10,18]. In addition, piecewise-planar 3D models are perceptually pleasing and geometrically simple, and thus their rendering, storage, and transmission is substantially less complex when compared to conventional point-cloud models [1,23]. The usefulness of plane primitives is not limited to multi-view stereo reconstruction as shown by recent works in SLAM for RGB-D cameras that estimate the motion from plane correspondences [24,21]. Taguchi et al. highlight that plane features are much less numerous than point features, favoring fast correspondence and scalability, and that the global character of

D. Fleet et al. (Eds.): ECCV 2014, Part II, LNCS 8690, pp. 48–63, 2014.

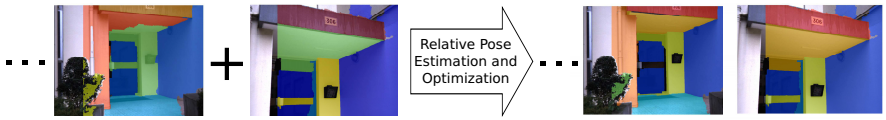


Fig. 1. Back-propagation of planes across stereo pairs: a closer view of the top horizontal plane allows its correct detection and propagation to previous stereo pairs. Note that the overlaid planes in the output images are identified by different colors.

plane-primitives helps avoiding local minima issues [24]. Also, man-made environments are often dominated by large size planes that enable correspondence across wide baseline images and, since plane-primitives are mostly in the static background, the motion estimation is specially resilient to dynamic foreground [21].

This article describes a pipeline for passive stereo that combines the benefits of PPR and plane-based odometry by recovering both structure and motion from plane-primitives. The algorithm receives as input an image sequence acquired by a calibrated stereo rig and outputs the camera motion and 3D planes in the scene. These planes are segmented in each stereo pair using a standard Markov-Random Field (MRF) labeling [11,22,4], and the final piecewise-planar model is obtained by simply concatenating the PPR results from consecutive frames.

The pipeline builds on the work of Antunes et al. [2] in PPR from semi-dense depth estimation using symmetry energy, which proved to outperform competing methods for the case of two calibrated views [4]. We start by running a simplified version of Antunes' algorithm in each input stereo pair and use these initial plane detections to compute the relative pose between consecutive frames. It is well known that the registration of two sets of 3D planes can be carried in closed-form from a minimum of 3 plane correspondences [13]. In our case, the estimation of the relative pose from plane-primitives raises two issues: establishing plane correspondences across stereo pairs, and determining the motion whenever the available planes do not fully constrain the problem [24]. The first issue is efficiently solved by matching triplets of planes using the angles between their normals. False correspondences are also pruned in [13,20,24] using this angular metric. Concerning the second issue it is shown that the undetermined situations can be overcome by either using 2 planes and 1 image point correspondence, or 1 plane and 3 image point correspondences [21] [1]. We derive closed-form minimal solutions for these cases and apply them in a hierarchical RANSAC that estimates the relative pose using point matches only when strictly necessary.

The next step is the joint refinement of camera motion and initial plane detections to obtain a coherent piecewise-planar model of the scene. In general, independent stereo detections of the same 3D plane are slightly different and must be merged into a single hypothesis before proceeding to bundle adjust-

[1] In this paper *image point correspondences* refer to inter-stereo point correspondences meaning point matches between the images of two different stereo pairs.

ment [11]. Moreover, and as shown in Fig. 1, it often happens that the same plane is wrongly reconstructed in a faraway view and correctly detected in a closer view, which means that the first plane hypothesis must be discarded and replaced by the second. We show that linking, fusing, and back-propagating plane hypotheses across stereo pairs can be conveniently formulated as a multi-model fitting problem that is efficiently solved using global energy minimization [15,17,6]. Thus, we propose to carry the joint refinement of motion and structure using a PEaRL framework [15] that alternates between a discrete optimization step, whose objective is to re-assign plane hypotheses to stereo pairs, and a continuous bundle adjustment step that refines the reconstruction results using the symmetry-energies arising from the initial semi-dense depth estimations [2,4].

In summary, the contributions of these article are threefold: (i) a method for estimating the relative pose between two stereo cameras that preferentially uses plane-primitives. This method differs from the algorithm for RGB-D cameras [24] because it uses image correspondences instead of 3D points for handling the undetermined cases; (ii) a PEaRL formulation for simultaneously refining camera motion and piecewise-planar model of the scene; and (iii) a complete stereo pipeline for PPR and motion estimation that is validated in challenging wide-baseline sequences for which conventional point-based SfM fails.

1.1 Related Work

Our work relates with previous methods for PPR [9,22,11,26,10,18] that operate in a batch manner by first applying point-based SfM to estimate the relative pose between monocular views [23], and then reconstructing the plane surfaces from all images in simultaneous. Unlike these methods, the algorithm herein described carries the 3D modeling in a sequential manner using a sliding window approach to concatenate the contributions of consecutive stereo pairs. This is an important difference that enables applications in visual odometry and SLAM. Since the article also proposes a method for estimating relative camera pose, it relates with prior works in visual odometry for stereo cameras [19,12,16,7,25]. We ran comparative experiments against the broadly used LIBVISO2 algorithm [12] that confirm the benefits of using plane-primitives, as opposed to image point matches, to recover the camera motion. In particular our method outperforms LIBVISO2 in the case of little overlap between stereo pairs.

2 Background

This section gives a brief review of background concepts that are useful for better understanding the proposed pipeline. It uses energy-based methods for solving two multi-model fitting problems, for which the theoretical basis is presented. Moreover, it builds on top of the PPR framework proposed in [2], whose main aspects are introduced in section 2.2.

2.1 Energy-Based Multi-Model Fitting

Several PPR methods start by obtaining a sparse 3D reconstruction of the scene, and solve a multi-model fitting problem for generating likely plane hypotheses. It has been recently stated in [15] that formulating the multi-model fitting as an optimal labeling problem with a global energy function is usually preferable than using RANSAC-based [11] or histogram-based [22] methods, mostly because they tend to ignore the overall classification of the input data.

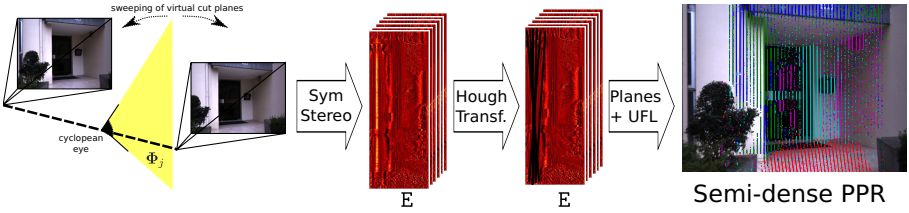
The optimization problem that arises from the multi-model fitting can be cast as an Uncapacitated Facility Location (UFL) problem, whenever the relationships between data nodes is not taken into account. The objective is to assign a label to each data point by minimizing a global energy function, $E = D + L$, that consists of data, D, and label costs, L. UFL problems can be efficiently solved using a message passing inference algorithm [17].

Whenever the dependencies between the data points are taken into account, a smoothness term S must be added to the previous energy function. In this case, the multi-model fitting can be formulated as an optimization problem using the PEARL algorithm [15]. The objective is also to assign a label to each data point, but this time by minimizing an energy function in the form $E = D + S + L$, which is efficiently achieved using α-expansion [15].

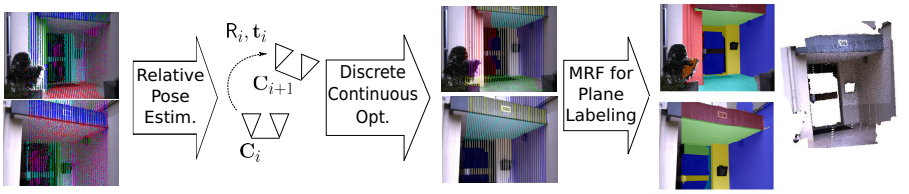
2.2 Semi-dense Piecewise Planar Stereo Reconstruction

Our method starts by obtaining a semi-dense PPR of the scene for each stereo pair using the method proposed in [2]. This framework was chosen as our starting point since it reported superior results when compared to other PPR methods [22,11] in stereo reconstruction, both in terms of accuracy and computational time. The method starts by employing a sparse set of M virtual cut planes $\mathbf{\Phi}_j$ intersecting the baseline in its midpoint for obtaining the energy $\mathtt{E}$ for each virtual plane using the SymStereo framework [3] (refer to Fig. 2(a)). This can be thought of as an image created by a virtual camera that is located between the cameras (cyclopean eye), where each epipolar plane $\mathbf{\Psi}_r$ projects onto one row and each virtual plane $\mathbf{\Phi}_j$ projects onto one column of the cyclopean image. Each pixel of the cyclopean eye is originated from the back-projection ray $\mathbf{d}_{j,r}$. For a particular virtual cut plane, each pixel in $\mathtt{E}$ provides the matching likelihood of a certain pair of pixels in the stereo views. The energy $\mathtt{E}$ is used as input to a Hough transform for extracting a set of line segments, which are the intersections of the virtual planes with the scene planes, and then each set of two lines provides a plane hypothesis. This is illustrated in the third step of the scheme in Fig. 2(a).

PPR is a *chicken-and-egg* problem since the accuracy of the plane hypotheses is inevitably limited by the accuracy of the initial 3D reconstruction that significantly depends on taking into account the fact of the scene being dominated by planar surfaces. Methods for PPR such as [22,11] that treat stereo matching and plane detection in a sequential and independent manner are affected by this problem. In [2], the multi-model plane fitting is formulated in a simultaneous and integrated manner as an optimization problem using the PEARL algorithm,



(a) Single Stereo Pair Semi-dense PPR



(b) Dense PPR from a Stereo Sequence

Fig. 2. Different steps of the proposed pipeline. (a) For each stereo pair, a semi-dense PPR is computed as described in section 3.1. The inlier set of planes for each view, along with the corresponding energies, is the input to the pipeline in (b), for which an overview is given in section 3.2. After the optimization step, colors identify planes. Note that a plane was randomly assigned to the black areas of the door due to its very weak texture, and the reconstructed points were removed.

overcoming this issue. Our pipeline follows this idea while fusing several stereo pairs. The objective in the formulation is to assign to each back-projection ray of the cyclopean eye a plane label of the initial plane set. They showed that the symmetry energy can be improved by repositioning the virtual cut plane according to the surface slant [4]. We use this idea in the continuous optimization step for improved performance. As a final step, a MRF formulation for labeling pixels is proposed. Our pipeline also contains this post processing step to obtain individual stereo reconstructions that are subsequently merged.

3 Overview of the Approach

We propose a structure and motion framework that is able to automatically recover the camera positions and orientations along with a piecewise planar reconstruction (PPR) of the scene from a stereo sequence. For each stereo pair, a semi-dense reconstruction is obtained using a simplified version of the algorithm described in section 2.2. The motion between consecutive frames is initialized in a RANSAC-like framework, where plane primitives are favoured over point correspondences. A sliding window approach is then used in an optimization step where the energy-based multi-model fitting algorithm PEaRL [15] is applied.

3.1 Semi-dense PPR from a Single Stereo Pair

For each stereo pair, a semi-dense piecewise planar reconstruction of the 3D scene is obtained (Fig. 2(a)). This is done by using a modified version of the method proposed in [2] and briefly reviewed in section 2.2. The original work formulates the multi-model plane fitting as an optimization problem using the PEARL algorithm. However, this problem can be cast as a UFL problem whenever no smoothness term is considered. This provides a less accurate but sufficiently good semi-dense PPR of the scene, being much faster than the original method.

3.2 PPR from a Stereo Sequence

Our algorithm takes as input the semi-dense labeling computed individually for each stereo pair i and a set of plane hypotheses $\mathbf{\Pi}_k^i, k = 1, \dots K$, and outputs the semi-dense labeling of a sequence of stereo pairs in conjunction with the relative pose between the consecutive pairs in the sequence, as illustrated in Fig. 2(b).

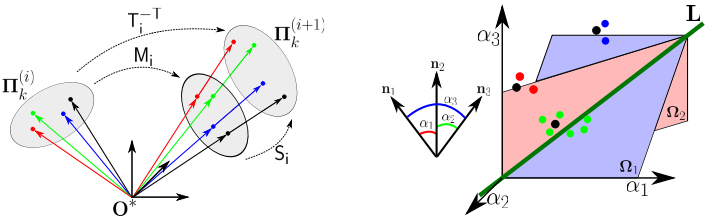
Although the explanation is given for a sequence of only two stereo pairs, it is extended to longer sequences in a straightforward manner. Our method consists of two main steps, which are an initialization of the relative pose R_i, $\mathbf{t}_i$ between cameras C_i and C_{i+1}, and a subsequent bundle adjustment step that alternates between discrete and continuous optimization for refining pose and structure.

The relative pose estimation is carried out using the planes from stereo pairs i and $i+1$ in a hierarchical scheme in the sense that it is obtained using the highest possible number of corresponding planes. A detailed explanation of this step is given in section 4. The energy-based multi-model fitting algorithm PEARL is applied in the optimization step. It consists of a discrete optimization step, where planes detected in cameras C_i and C_{i+1} are assigned to pixels of the cyclopean eye of those cameras, by minimizing an energy function with data, smoothness and label terms. Next, the chosen planes and the relative pose are jointly optimized in the continuous step. Further details are given in section 5.

For visualization purposes, a dense labeling for each stereo pair is generated in a MRF approach. By concatenating the individual reconstructions, it is possible to obtain a dense piecewise planar reconstruction for the complete sequence.

4 Relative Pose Estimation

Consider two consecutive stereo pairs C_i and C_{i+1} and two sets of plane detections. Let $\mathbf{\Pi}_k^{(i)}$ and $\mathbf{\Pi}_k^{(i+1)}$, with $k = 1 \dots K$ be putative plane correspondences across the two pairs. Our objective is to use these plane correspondences to estimate the relative pose $(\mathsf{R}_i, \mathbf{t}_i)$ between the stereo cameras. In [13], it was first shown that two sets of 3D planes can be registered in a closed-form manner from a minimum of 3 correspondences as long as their normals span the entire 3D space. More recently, Taguchi et al. [24] used this registration algorithm as a starting point for their plane-based SLAM method for RGB-D cameras. They studied the singular configurations and showed how to use reconstructed



(a) Registration (b) Triplet matching

Fig. 3. (a) The relative pose estimation can be cast as a point registration problem in the dual projective space $\mathcal{P}^{3*}$. (b) A descriptor is computed for the plane triplets and used in a nearest-neighbors approach for finding putative matches between the planes. Similarities between angles in the descriptor give rise to different hypotheses, depicted by the points near planes $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ and line L.

3D points to disambiguate motion whenever the information provided by planes was insufficient. We revisit this registration problem and show how to disambiguate the motion by directly using inter-stereo image point correspondences, in order to avoid having to reconstruct points from passive stereo.

4.1 Relative Pose from 3 Plane Correspondences

The registration problem between stereo pairs i and $i+1$ is the one of estimating R_i and $\mathbf{t}_i$ such that

$$\boldsymbol{\Pi}_k^{(i+1)} \sim \underbrace{\begin{bmatrix} \mathsf{R}_i & \mathbf{0} \\ -\mathbf{t}_i^\mathsf{T}\mathsf{R}_i & 1 \end{bmatrix}}_{\mathsf{T}_i^{-\mathsf{T}}} \boldsymbol{\Pi}_k^{(i)} \sim \underbrace{\begin{bmatrix} \mathsf{I}_3 & \mathbf{0} \\ -\mathbf{t}_i^\mathsf{T} & 1 \end{bmatrix}}_{\mathsf{S}_i} \underbrace{\begin{bmatrix} \mathsf{R}_i & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}}_{\mathsf{M}_i} \boldsymbol{\Pi}_k^{(i)}, k = 1, 2, 3 \tag{1}$$

verifies, where I_3 is the 3×3 identity matrix, and the 3D planes have the homogeneous representation $\boldsymbol{\Pi}_k^{(i)} \sim [\mathbf{n}_k^\mathsf{T} \quad 1]^\mathsf{T}$ and $\boldsymbol{\Pi}_k^{(i+1)} \sim [\mathbf{m}_k^\mathsf{T} \quad 1]^\mathsf{T}$. Knowing that points and planes are dual entities in 3D - a plane in the projective space $\mathcal{P}^3$ is represented as a point in the dual space $\mathcal{P}^{3*}$, and vice-versa - equation (1) can be seen as a projective transformation in $\mathcal{P}^{3*}$ that maps points $\boldsymbol{\Pi}_k^{(i)}$ into points $\boldsymbol{\Pi}_k^{(i+1)}$ through a rotation transformation M_i followed by a projective scaling S_i, as is illustrated in Fig. 3(a). R_i is firstly computed by applying the algorithm from [14] that provides a unique solution for aligning two sets of unitary vectors.

By replacing R_i in equation (1), it can be shown after some algebraic manipulation that $\mathbf{t}_i$ is computed by solving the following linear system of equations

$$\begin{bmatrix} \mathbf{m}_1^\mathsf{T}\mathbf{m}_1 & 0 & 0 \\ 0 & \mathbf{m}_2^\mathsf{T}\mathbf{m}_2 & 0 \\ 0 & 0 & \mathbf{m}_3^\mathsf{T}\mathbf{m}_3 \end{bmatrix} \underbrace{\begin{bmatrix} \mathbf{n}_1^\mathsf{T} \\ \mathbf{n}_2^\mathsf{T} \\ \mathbf{n}_3^\mathsf{T} \end{bmatrix}}_{\mathsf{N}_i} \mathsf{R}_i^\mathsf{T}\mathbf{t}_i = \begin{bmatrix} \mathbf{m}_1^\mathsf{T}\mathbf{m}_1 - \mathbf{m}_1^\mathsf{T}\mathsf{R}_i\mathbf{n}_1 \\ \mathbf{m}_2^\mathsf{T}\mathbf{m}_2 - \mathbf{m}_2^\mathsf{T}\mathsf{R}_i\mathbf{n}_2 \\ \mathbf{m}_3^\mathsf{T}\mathbf{m}_3 - \mathbf{m}_3^\mathsf{T}\mathsf{R}_i\mathbf{n}_3 \end{bmatrix}. \tag{2}$$

From this equation, it comes in a straightforward manner that if the three normals do not span the entire 3D space, then N_i is rank deficient and the problem of determining the translation becomes underdetermined.

4.2 Relative Pose Estimation in Case N_i Has Rank 2

The matrix of the normal vectors N_i can have rank 2 whenever there are only two corresponding planes available or the three planes have a configuration such that their normals are co-planar. An example of this situation happens when at least two planes are parallel. The rotation R_i is estimated using Horn's algorithm [14] since two corresponding planes suffice. However, there is a 2D space for translation, and thus there is one remaining DOF to be estimated. Given an image point correspondence $\mathbf{x}^{(i)}, \mathbf{x}^{(i+1)}$ between the reference views of the two stereo pairs C_i and C_{i+1}, the translation $\mathbf{t}_i$ can be fully determined by stacking the epipolar constraint ${\mathbf{x}^{(i+1)}}^\mathsf{T} \mathsf{E}_i \mathbf{x}^{(i)} = 0$, where $\mathsf{E}_i = [\mathbf{t}_i]_\times \mathsf{R}_i$ is the essential matrix, to the two linear constraints in equation 2.

4.3 Relative Pose Estimation in Case N_i Has Rank 1

Whenever there is a single plane correspondence or the putative plane correspondences are all parallel, the registration leads to the computation of 2 DOF for the rotation. In this case N_i has rank 1, and thus 1 DOF for the translation can be estimated. We show for the first time that in this case the relative pose can be determined from a minimum of 3 additional image point correspondences $\mathbf{x}_k^{(i)}, \mathbf{x}_k^{(i+1)}, k = 1 \ldots 3$. Related to this problem is the work described in [8], where a minimal solution for the case of two known orientation angles is given. Our problem differs from it because we have an extra constraint for the translation.

Our reasoning is explained in the 3D space instead of the dual space. Both stereo cameras C_i and C_{i+1} are independently rotated so that the z axes of their reference views are aligned with the plane normal, through transformations P_i and P_{i+1}. This implies that the rotated cameras become related by an unknown rotation around the z axis, $\mathsf{R}_u(\theta)$, and a translation $\mathbf{t}_u = [t_x \quad t_y \quad t_z]^\mathsf{T}$, where t_z can be computed as follows. In the rotated configuration, equation 1 becomes

$$\begin{bmatrix} 0 \\ 0 \\ z_2 \\ 1 \end{bmatrix} \sim \begin{bmatrix} & \mathsf{R}_u & & \mathbf{0} \\ -[t_x & t_y & t_z]\mathsf{R}_u & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ z_1 \\ 1 \end{bmatrix}. \tag{3}$$

Thus, t_z can be determined by $t_z = -\frac{z_1/z_2 - 1}{z_1}$. The remaining 3 DOF ($\theta$, t_x and t_y) can then be determined from 3 point correspondences using the epipolar constraint. The essential matrix E_i has a simplified form as in [8], allowing the epipolar constraint to be written as $\mathsf{A}[t_x \quad t_y \quad 1]^\mathsf{T} = \mathbf{0}$, where the 3×3-matrix A depends on θ, which can be computed using the hidden variable method. This originates up to 4 solutions for the motion in the rotated configuration, T_u. The real motion T_i can then be retrieved by simply computing $\mathsf{T}_i = \mathsf{P}_{i+1}^{-1} \mathsf{T}_u \mathsf{P}_i$.

4.4 Robust Algorithm for Computing the Relative Pose

Our relative pose estimation algorithm uses an hierarchical RANSAC scheme that works by considering the maximum number of planes present in the image pair, and only using point correspondences when strictly necessary. It first attempts to compute the pose from 3 plane correspondences, using subsequently less plane correspondences in case of failure, meaning that it tries to carry the registration with 2 planes and 1 point, and if this fails, with 1 plane and 3 points.

The method starts by building a descriptor (refer to Fig. 3(b))) for matching triplets of planes, which consists of the 3 angles between the plane normals sorted by increasing value, in both stereo pairs. Putative matches are established using a nearest neighbors approach. Remark that the descriptor implicitly establishes plane correspondences between elements in the triplet and that typically there is a relatively small number of triplets for each view. In case the angles in the descriptor are sufficiently different from each other, the descriptor establishes plane correspondences directly. However, if two of the angles are similar, two possible sets of element-wise correspondences are considered. This is the case in Fig. 3(b) where the point in the descriptor space is close to plane $\mathbf{\Omega}_2$ that defines $\alpha_1 = \alpha_2$ (and identical for plane $\mathbf{\Omega}_1$ that defines $\alpha_2 = \alpha_3$). Similarly, if all three angles are close, six possible hypotheses for matches must be considered. This is the case when the point is close to the line L that defines $\alpha_1 = \alpha_2 = \alpha_3$.

For each triple correspondence, a solution is computed using the procedure in subsection 4.1. The semi-dense PPR step generates a set of line cuts in each frame associated to each reconstructed scene plane. A patch containing the pixels around the projection of each line cut in the left image of camera C_i is selected and projected onto the left image of camera C_{i+1}, using the homography induced by the respective plane. Line cuts that have a photo-geometric error below a predefined threshold are considered for computing a score ϵ.

The pose estimation is performed in a RANSAC framework. If there are no matching triplets of planes or the number of inlier line cuts for the computed solutions originates a score too low, the algorithm attempts to use 2 plane correspondences. A descriptor consisting of the angle between the 2 plane normals is considered for both stereo pairs and matches are established using a nearest-neighbors approach. Since there is only one angle, each match gives rise to two hypotheses. A local feature detector (SURF [5]) is used for extracting point features and solutions are computed in a RANSAC framework from two planes and one point correspondences (subsection 4.2). The models' inliers are computed as in the previous stage. Similarly, if there are no acceptable corresponding pairs of planes, the motion is estimated using one plane and three point-correspondences, as described in subsection 4.3. Note that in theory the scoring metric might fail if the planes surfaces lack texture. An hybrid score metric that mixes planes and points raises other type of issues, such as normalization. The metric used in this work always provided acceptable results, and thus it was kept unaltered.

5 Discrete-Continuous Bundle Adjustment

This section describes the optimization step that is carried for jointly refining the motion and the piecewise planar structure. From the previous single stereo PPR and relative pose estimation steps come two sets of planes defined in the reference frames of cameras C_i and C_{i+1}, $\mathbf{\Pi}_k^{(i)}, k = 1 \dots K_i$ and $\mathbf{\Pi}_k^{(i+1)}, k = 1 \dots K_{i+1}$, respectively, and an initialization for the relative pose R_i, $\mathbf{t}_i$ between the cameras. The optimization is achieved using the PEARL algorithm that consists in three steps: (i) propose an initial set of plausible models (labels) from the data, (ii) expand the label set for estimating its spatial support (inlier classification), and (iii) re-estimate the inlier models by minimizing some error function.

The initial set of plane models $\mathcal{P}_0$ for PEARL is the union of the $(K_i + K_{i+1})$ planes detected in each stereo pair separately. Then, the objective is to expand the models and estimate their spatial support. Consider the cyclopean eye relative to camera i, whose back-projection rays are denoted by $\mathbf{d}_{j,r}^{(i)}$, where r indexes a particular epipolar plane (refer to section 2.2). The objective is to estimate the point on $\mathbf{d}_{j,r}^{(i)}$ that most likely belongs to a planar surface. As stated previously, this problem can be cast as a labeling problem, in which the nodes of the graph are the back-projection rays $\mathbf{d}_{j,r}^{(i)} \in \mathcal{D}$, and to which we want to assign a plane label $f_{\mathbf{d}_{j,r}^{(i)}}$. The set of possible labels is $\mathcal{F} = \{\mathcal{P}_0, f_\emptyset\}$, where $f_\emptyset$ is the discard label and is mostly used for identifying non-planar structures. This labeling problem is solved by minimizing an energy function E defined by

$$E(\mathbf{f}) = \underbrace{\sum_i \sum_{\mathbf{d}_{j,r}^{(i)} \in \mathcal{D}} D_{d_{j,r}^{(i)}}(f_{\mathbf{d}_{j,r}^{(i)}})}_{\text{data term}} + \lambda_S \underbrace{\sum_i \sum_{\mathbf{d}_{j,r}^{(i)}, \mathbf{e}_{j,r}^{(i)} \in \mathcal{N}} V_{\mathbf{d}_{j,r}^{(i)}, \mathbf{e}_{j,r}^{(i)}}(f_{\mathbf{d}_{j,r}^{(i)}}, f_{\mathbf{e}_{j,r}^{(i)}})}_{\text{smoothness term}} + \underbrace{\lambda_L \cdot |\mathcal{F}_f|}_{\text{label term}}, \tag{4}$$

where λ_S and λ_L are weighting constants, $\mathbf{f}$ is the labeling being analyzed, $\mathcal{N}$ is the neighborhood of $\mathbf{d}_{j,r}^{(i)}$ and V is the spacial smoothness term. The label term forces the algorithm to use as few plane surfaces as possible. The data term $D_{\mathbf{d}_{j,r}^{(i)}}$ for the back-projection ray $\mathbf{d}_{j,r}^{(i)}$ is defined as

$$D_{\mathbf{d}_{j,r}^{(i)}}(f) = \begin{cases} \min(1 - \mathtt{E}_j^{(i)}(r, x_f), \tau) & \text{if } f \in \mathcal{P}_0 \\ \tau & \text{if } f = f_\emptyset \end{cases}$$

where the coordinate x_f is the column defined by the hypothesis f, corresponding to the intersection of $\mathbf{d}_{j,r}^{(i)}$ with the plane indexed by f. Using the camera pose, we can transform the planes detected in the stereo rig $i + 1$ to the stereo rig i, and vice versa. This allows us to use all the structure information available simultaneously and reconstruct planes in a particular view even if they were detected by a different camera. The smoothness term V is used to describe the relationships between nodes. No penalization is assigned to neighboring nodes receiving the same plane label, while in the case of one node obtaining the discard label, a non-zero cost is added to the plane configuration $\mathbf{f}$. For each camera i,

the smoothness term V is defined as in [2], which encourages label transitions near crease or occlusions edges . For further details refer to that work.

The output of this step is a set of planes shared by cameras C_i and C_{i+1}. Given the inliers of a particular plane label f, the corresponding energies $\mathtt{E}^{(i)}$ can be recomputed to enhance the likelihood measure with respect to a particular range of slant values [4]. These energies are used in the third step of PEARL.

Let $\mathbf{\Pi}_f$ be the plane associated to f to which has been assigned a non-empty set of inliers $\mathbf{D}(f) = \{\mathbf{d} \in \mathcal{D} | f_{\mathbf{d}} = f\}$. All the inlier planes $\{\mathbf{\Pi}_{f_k}\}$ and the relative pose R_i, $\mathbf{t}_i$ are refined simultaneously by minimizing the error function:

$$\{\mathsf{R}_i^*, \mathbf{t}_i^*, \{\mathbf{\Pi}_{f_k}^*\}\} = \min_{\mathsf{R}_i, \mathbf{t}_i, \{\mathbf{\Pi}_{f_k}\}} \sum_i \sum_k \sum_{\mathbf{d}_{j,r}^{(i)} \in \mathbf{D}(f)} \left(1 - \mathtt{E}_j^{(i)}(r, x_{\mathbf{\Pi}_{f_k}})\right) + \delta e_{ph}, \quad (5)$$

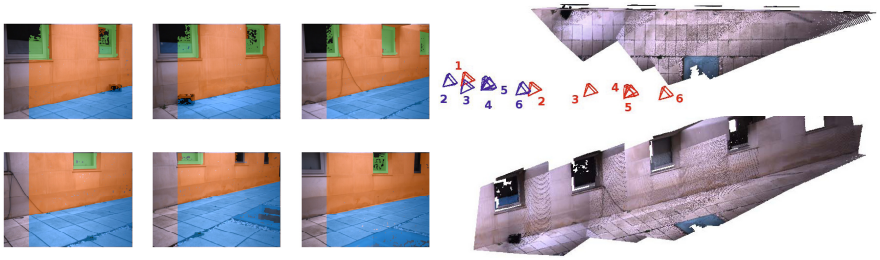
where $x_{\mathbf{\Pi}_{f_k}}$ is the column defined by the intersection of $\mathbf{d}_{j,r}^{(i)}$ with $\mathbf{\Pi}_{f_k}$, δ is a parameter that is zero whenever the optimization is carried out using 3 shared planes that span the 3D space and larger than zero otherwise, and e_{ph} is the photo-consistency error computed in a planar patch. The new set of plane labels $\mathcal{P}_1 = \left\{\mathbf{\Pi}_{f_k}^*\right\}$ is then used in a new expand step, and we iterate between discrete labeling and plane refinement until the α-expansion optimization does not decrease the energy of Equation 4.

A sliding window approach is applied where at most one relative pose is refined. The exchange of planes between cameras, described previously, has an important role in the 3D modeling process since it allows planar surfaces that are only properly detected in subsequent frames to be back-propagated and accurately reconstructed in previous images. Remark that plane information is only exchanged between different cameras inside the sliding window. In order to overcome this issue, a connected list is maintained containing the plane linking information across views and is updated whenever a new plane is back-propagated inside the considered window of cameras.

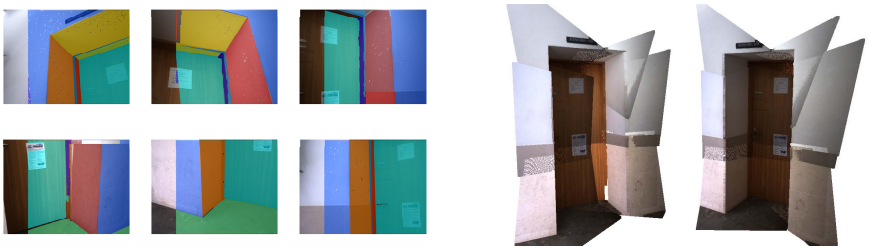
6 Experimental Results

In this section several experiments are shown in order to highlight the different advantages of the proposed method. The datasets were acquired using a stereo camera with a 24 cm baseline and a resolution of 1024×768 pixels. Experiments on short sequences of 3 to 6 images are presented, and the motion estimation is compared to the result obtained with the point-based method LIBVISO2 [12] (Fig. 4). LIBVISO2 only leads to plausible results in some of the experiments, in which cases the images of the 3D reconstructions include camera symbols in red and blue, if they were computed using our algorithm or LIBVISO2, respectively. For every experiment, the left images of all the stereo pairs that were used are shown with the overlaid MRF labeling, where each color identifies one plane.

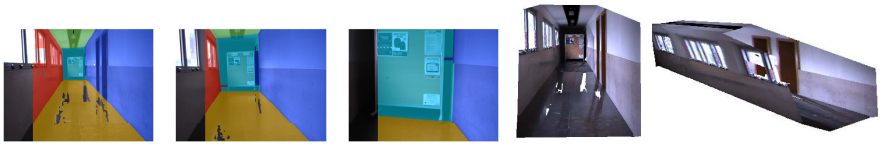
Example 1



Example 2



Example 3



Example 4

Example 5

Fig. 4. Structure and motion results, different colors identify different planes. Red/blue cameras represent the motion computed using our approach/LIBVISO2.

The sequence of images is sorted from left to right and top to bottom and the cameras are numbered accordingly. A 140-meter loop-closing experiment using a sequence of 60 frames acquired in an outdoor scene is also shown (Fig. 5).

Example 1. The 5-frame stereo sequence was acquired with significant overlap in order to illustrate the exchange of planes between frames. It can be seen that in the first stereo pairs, the top plane of the entrance has very small image support, and thus cannot be recovered, as shown in Fig. 1. Moreover, the back plane (containing the door) is poorly estimated since it is only observed from a long distance. These two planes are correctly reconstructed in the last frame, and back propagated to the initial frames, providing an accurate reconstruction of the whole scene. Our method and LIBVISO2 provided very similar results.

Example 2. This is an outdoor example where the scene is dominated by two plane directions. The relative pose between the consecutive cameras was obtained using two plane and one point correspondences. It can be seen that all the planes were correctly assigned across the different views, and an accurate reconstruction was obtained, evinced by the correct alignment of the floor lines and the detection of the windows. Also, this scene contains a significant amount of perceptual aliasing since consecutive views have only slight differences. Due to this fact, LIBVISO2 was unable to provide an acceptable result when computing the camera motion between the first 3 positions. However, it provided estimations very similar to ours for the last camera positions.

Example 3. A sequence of six stereo pairs with minimum overlap was acquired, originating a detailed reconstruction of a door. It can be seen that the white walls and the small interior planes were accurately recovered. LIBVISO2 failed to find sufficient point correspondences for estimating the camera motion. Our approach computed the camera motion using correspondences of two planes and one point, as there are no triplet correspondences in consecutive stereo pairs.

Example 4. This example illustrates the behavior of our method in a challenging situation of low textured surfaces, high slant and image specularities. A 14-meter corridor is accurately reconstructed using a sequence of only 3 stereo pairs. Our method does not rely on the Manhattan assumption, as shown by this example where the board is not perpendicular to the walls. LIBVISO2 was not able to compute the camera motion due to the small overlap between views.

Example 5. This outdoor example shows that our method is able to correctly distinguish between planar and non-planar objects, which can be used to automatically remove trees and vegetation from the final 3D model. The fact that the vegetation occupies a large part of the camera's field of view leads to a large percentage of incorrect point matches. Thus, LIBVISO2 provided very poor results for the estimation of the relative pose. As an example, camera 3 appears to be in an impossible position since it is in the vegetation's location.

Final Example. The reconstruction of an outdoor stereo sequence of 60 frames is shown in Fig. 5. The camera traveled 136.6 meters in loop, and the final loop closing error was 2.17% in translation and 2.67% in rotation. For validation and

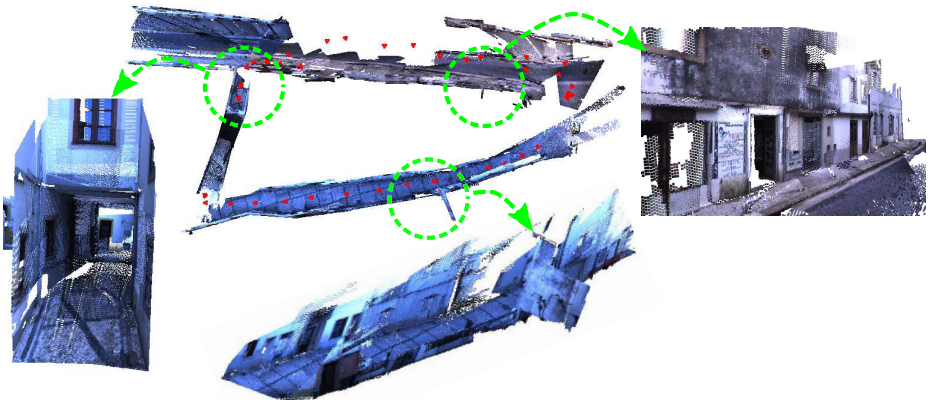


Fig. 5. Final example of a 137-meter loop closing experiment. Despite the challenging conditions, a good 3D reconstruction of the whole scene was obtained. A video is attached as supplementary material where some results can be seen in more detail.

computational time purposes, the selected size of the optimization window is two. The trajectory can be seen in the top view of the reconstructed scene, depicted by red camera symbols, showing that there is a large displacement between most of the positions, which translates into small overlap between consecutive stereo pairs. Moreover, the camera usually pointed forward, meaning that the algorithm dealt with strong surface slant. Under these circumstances, it was able to provide an accurate reconstruction of the scene. Images in different viewpoints are included to better illustrate the obtained results, where good alignment can be observed. Due to this acquisition conditions, LIBVISO2 was unable to find sufficient point matches to provide a plausible motion estimation.

7 Conclusions

We describe the first pipeline for sequential piecewise-planar reconstruction from images acquired by a moving stereo rig. The relative pose between consecutive frames is preferentially estimated using plane-primitives, and motion and structure are jointly refined within a PEARL framework [15] that alternates between discrete optimization to enforce coherent PPR across stereo frames, and continuous bundle adjustment to improve the accuracy of results. The experiments show that the use of plane-primitives to recover camera motion enables to handle sequences with little overlap that are very challenging for conventional SfM approaches. The approach proved to successfully handle situations of weak texture, high surface slant, repetitive structure, and non-lambertian reflection, being able to render detailed piecewise-planar models of the scene in cases of minimum visual coverage. We are using a straightforward MATLAB implementation for validation purposes. For getting an idea about the current runtime, our pipeline took around 2 hours for computing the motion and reconstructing the scene depicted in Fig. 5. As future work, we intend to develop a parallel version of the

pipeline to be ran in the GPU (note that the initial PPR is computed for each stereo rig independently) in order to decrease computational time and use larger optimization windows for further improving accuracy.

Acknowledgments. The authors thank Google, Inc for the support through a Faculty Research Award. Carolina Raposo acknowledges the Portuguese Science Foundation (FCT) for funding her PhD under grant SFRH/BD/88446/2012. The work was also partially supported by FCT and COMPETE program under Grant AMS-HMI12: RECI/EEI-AUT/0181/2012.

References

1. P.F. Alcantarilla, C. Beall, F. Dellaert.: Large-scale dense 3D reconstruction from stereo imagery. In: 5th Workshop on Planning, Perception and Navigation for Intelligent Vehicles (PPNIV13) (2013)
2. Antunes, M., Barreto, J.P.: Semi-dense piecewise planar stereo reconstruction using symstereo and pearl. In: 3DimPVT (2012)
3. Antunes, M., Barreto, J.P., Zabulis, X.: Plane surface detection and reconstruction using induced stereo symmetry. In: BMVC (2011)
4. Antunes, M.: Phd thesis: Stereo reconstruction using induced symmetry and 3D scene priors (2014), http://www2.isr.uc.pt/~michel/files/final.pdf
5. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (surf). Comput. Vis. Image Underst. 110(3), 346–359 (2008)
6. Delong, A., Osokin, A., Isack, H., Boykov, Y.: Fast approximate energy minimization with label costs. International Journal of Computer Vision 96(1), 1–27 (2012), http://dx.doi.org/10.1007/s11263-011-0437-z
7. Dunn, E., Clipp, B., Frahm, J.M.: A geometric solver for calibrated stereo egomotion. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 1187–1194 (November 2011)
8. Fraundorfer, F., Tanskanen, P., Pollefeys, M.: A minimal case solution to the calibrated relative pose problem for the case of two known orientation angles. In: Danilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 269–282. Springer, Heidelberg (2010)
9. Furukawa, Y., Curless, B., Seitz, S., Szeliski, R.: Manhattan-world stereo. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 1422–1429 (June 2009)
10. Furukawa, Y., Curless, B., Seitz, S., Szeliski, R.: Reconstructing building interiors from images. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 80–87 (September 2009)
11. Gallup, D., Frahm, J.M., Pollefeys, M.: Piecewise planar and non-planar stereo for urban scene reconstruction. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1418–1425 (June 2010)
12. Geiger, A., Ziegler, J., Stiller, C.: Stereoscan: Dense 3D reconstruction in real-time. In: Intelligent Vehicles Symposium (IV) (2011)
13. Grimson, W., Lozano-Pérez, T.: Model-based recognition and localization from sparse range or tactile data. International Journal of Robotics Research 3(3), 3–35 (1984), http://lis.csail.mit.edu/pubs/tlp/AIM-738.pdf

14. Horn, B.K.P.: Closed-form solution of absolute orientation using unit quaternions. J. Opt. Soc. Am. A 4(4), 629–642 (1987)
15. Isack, H., Boykov, Y.: Energy-based geometric multi-model fitting. IJCV (2012)
16. Kazik, T., Kneip, L., Nikolic, J., Pollefeys, M., Siegwart, R.: Real-time 6D stereo visual odometry with non-overlapping fields of view. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1529–1536 (June 2012)
17. Lazic, N., Frey, B.J., Aarabi, P.: Solving the uncapacitated facility location problem using message passing algorithms. Journal of Machine Learning Research (2010)
18. Micusik, B., Kosecka, J.: Piecewise planar city 3d modeling from street view panoramic sequences. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 2906–2912 (June 2009)
19. Nister, D., Naroditsky, O., Bergen, J.: Visual odometry. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, vol. 1, pp. I-652–I-659 (June 2004)
20. Pathak, K., Birk, A., Vaskevicius, N., Poppinga., J.: Fast registration based on noisy planes with unknown correspondences for 3D mapping. IEEE Transactions on Robotics 26(3), 424–441 (2010)
21. Raposo, C., Lourenco, M., Antunes, M., Barreto, J.P.: Plane-based odometry using an rgb-d camera, pp. 1–11 (September 2013)
22. Sinha, S., Steedly, D., Szeliski, R.: Piecewise planar stereo for image-based rendering. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1881–1888 (September)
23. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the world from internet photo collections. Int. J. Comput. Vision 80(2), 189–210 (2008), http://dx.doi.org/10.1007/s11263-007-0107-3
24. Taguchi, Y., Jian, Y.D., Ramalingam, S., Feng, C.: Slam using both points and planes for hand-held 3d sensors. In: Proceedings of the 2012 IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2012, pp. 321–322. IEEE Computer Society, Washington, DC (2012), http://dx.doi.org/10.1109/ISMAR.2012.6402594
25. Vasconcelos, F., Barreto, J., Nunes, U.: A minimal solution for the extrinsic calibration of a camera and a laser-rangefinder. IEEE Transactions on Pattern Analysis and Machine Intelligence PP(99), 1 (2012)
26. Werner, T., Zisserman, A.: New techniques for automated architectural reconstruction from photographs. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part II. LNCS, vol. 2351, pp. 541–555. Springer, Heidelberg (2002), http://dl.acm.org/citation.cfm?id=645316.649194