Piecewise-Planar StereoScan: Sequential Structure and Motion using Plane Primitives

Carolina Raposo, Michel Antunes, and João P. Barreto, *Member*

Abstract—The article describes a pipeline that receives as input a sequence of stereo images, and outputs the camera motion and a Piecewise-Planar Reconstruction (PPR) of the scene. The pipeline, named Piecewise-Planar StereoScan (PPSS), works as follows: the planes in the scene are detected for each stereo view using semi-dense depth estimation; the relative pose is computed by a new closed-form minimal algorithm that only uses point correspondences whenever plane detections do not fully constrain the motion; the camera motion and the PPR are jointly refined by alternating between discrete optimization and continuous bundle adjustment; and, finally, the detected 3D planes are segmented in images using a new framework that handles low texture and visibility issues. PPSS is extensively validated in indoor and outdoor datasets, and benchmarked against two popular point-based SfM pipelines. The experiments confirm that plane-based visual odometry is resilient to situations of small image overlap, poor texture, specularity, and perceptual aliasing where the fast LIBVISO2 [1] pipeline fails. The comparison against VisualSfM+CMVS/PMVS [2], [3] shows that, for a similar computational complexity, PPSS is more accurate and provides much more compelling and visually pleasant 3D models. These results strongly suggest that plane primitives are an advantageous alternative to point correspondences for applications of SfM and 3D reconstruction in man-made environments.

Index Terms—Structure and Motion, Piecewise-Planar Reconstruction, Stereo Image Sequences, MRF

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 [4]. 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 [4], [5], [6], [7], [8], [9], [10]. In addition, piecewise-planar 3D models are perceptually pleasing and geometrically simple, making the rendering, storage, and transmission substantially less complex when compared to point-cloud models [11], [12].

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 [13], [14]. Taguchi et al. [13] highlight that plane features are less numerous than point features, favouring fast correspondence and scalability, and that the global character of plane-primitives helps avoiding local minima issues. 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 [14].

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, and the final piecewise-planar model is obtained by simply concatenating the PPR results from consecutive frames.

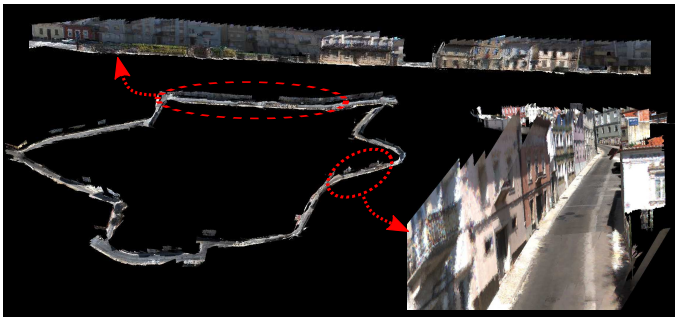


Fig. 1. Experiment in an urban sequence acquired by a moving vehicle with a forward-looking stereo pair. The sequence has 1370 frames covering 1100m. The errors in loop closing were 1.8% in translation and $0.5°$ in rotation with PPSS, which compares against 3.6% in translation and $26°$ in rotation with LIBVISO2 [1] and 2.0% in translation and $4.3°$ in rotation with VisualSFM [2], [3]. In addition, PPSS provides a visually pleasant PPR that can be viewed in the following link https://youtu.be/IhELZ3-wPU0.

The pipeline builds on the work of Antunes et al. [7], [15] in PPR from semi-dense depth estimation using the SymStereo framework, which proved to outperform competing methods for the case of two calibrated views [16]. We start by running a simplified version of Antunes' algorithm in each input stereo pair and use these initial plane detections

- *Carolina Raposo and João Barreto are with the Institute of Systems and Robotics, University of Coimbra, Portugal.*
- *Michel Antunes is with the Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg*

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 [17]. 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 [13]. The first issue is efficiently solved by matching triplets of planes using the angles between their normals. False correspondences are also pruned in [13], [17], [18] 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 [14] [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 adjustment [4]. Moreover, 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 [19], [20], [21]. Thus, we propose to carry the joint refinement of motion and structure using the PEARL framework [19] 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 [7], [15].

As a final step, the 3D plane surfaces detected in the scene are segmented in each stereo pair through dense labelling of the pixels. This can be accomplished using a standard MRF formulation, as proposed in [4], [7], with a data-term that quantifies label likelihood based on left-right photo consistency, and a smoothness term for regularization. We verified that these prior methods have difficulties in handling low-textured regions, where photo-consistency is ambiguous, and do not take into account coherence in visibility across successive frames. The article proposes a new MRF formulation, specific for sequential PPR, that largely solves the above mentioned issues.

The PPSS pipeline has been introduced in a conference publication that advanced the idea of using plane primitives for motion estimation [22]. The present article extends and consolidates this earlier work by introducing several improvements and providing a much more thorough experimental validation. The most notable improvement is the use of a new MRF formulation for labelling the image pixels using refined 3D plane hypotheses (Section 6). This new formulation enables to segment textureless plane regions and take into account visibility consistency across views, leading to 3D models that are complete, accurate and visually compelling. The article also provides a more thorough validation of PPSS in indoor and outdoor datasets, including a long and challenging urban sequence acquired by a moving vehicle. The experiments clearly illustrate the pros-and-cons of using plane primitives, as opposed to point correspondences, in Structure-from-Motion (SfM), visual odometry and SLAM pipelines. As advanced in [22], plane-based methods are resilient to small image overlap, poor texture, specularity, and perceptual aliasing, working in a number of practical situations where point-based methods usually fail. The present article goes one step further and compares our plane-based method with a sophisticated point-based pipeline with similar runtime, with experiments showing that the former outperforms the latter in accuracy and visual quality of reconstruction results. This is an important observation, because it suggests for the first time, that plane primitives should be faced as an alternative to point correspondences, and not as a mere complement to handle specific circumstances where the latter ones fail.

In summary, this article is an exploratory work in plane-based SfM, visual odometry or SLAM that conveys the message that plane primitives are an effective alternative to point correspondences. The benefits are resilience to a number of common situations where point-based methods fail, higher overall accuracy in motion estimation, and more complete, compact and visually pleasant 3D models. The paper starts by presenting the background work (Section 2) and an overview of the PPSS pipeline with a justification for the need of its different modules or steps (Section 3). The overall contributions can be summarised as follows, with the first two bullets referring to modules that were initially proposed in [22]:

1) A new closed-form minimal algorithm for determining the relative pose between stereo views from 3D planes that, in case the available planes are insufficient to fully constrain the motion, it automatically combines planes and points to assure operation in all circumstances (Section 4);
2) A PEARL formulation for simultaneously refining the camera motion and the 3D plane hypotheses that form the piecewise-planar model (Section 5);
3) A new method for labelling pixels into plane regions that handles low texture surfaces and assures coherence in visibility across views, enabling the computation of complete and visually compelling 3D models (Section 6);
4) A pipeline for camera motion estimation and PPR that is extensively tested with experiments showing that plane-based SfM is able to handle situations where popular fast methods fail and, comparing with point-based methods with similar computational complexity, it leads to more accurate results and visually pleasant 3D models (Section 7).

1.1 Related Work

Our work relates with previous methods for PPR [4], [5], [6], [8], [9], [10] that operate in a batch manner by first applying point-based SfM to estimate the relative pose between

1. *Image point correspondences* refer to inter-stereo correspondences

monocular views [12], and then reconstructing the planes from all images simultaneously. Unlike these methods, the algorithm herein described carries the 3D modelling in a sequential manner using a sliding window approach to combine the contributions of consecutive stereo pairs. This is an important difference that also enables applications in visual odometry and SLAM [23], [24], [25]. We ran comparative experiments against the broadly used stereo visual odometry algorithm LIBVISO2 [1] and the more sophisticated SfM system VisualSFM [2], [3] complemented with CMVS/PMVS [26], [27]. The experiments confirm the benefits of using plane-primitives by showing that our method outperforms point-based methods not only in terms of accuracy and robustness to situations of wide baseline, repetitive appearance, low texture and specularities, but also in the quality and completeness of 3D reconstructions.

Since one of the new additions to the pipeline with respect to [22] is the dense plane labelling of image pixels using an MRF formulation, it is worth reviewing previous formulations for the same purpose. In [6], Sinha et al. present an MRF where the data term includes not only photo-consistency cues, but also cues of geometric proximity between points and lines, and free space violation. Inspired by this idea, the modification to the data term of the MRF proposed in this article consists in using the previously computed semi-dense labelling information [15] to decrease the cost of certain pixel assignments, which avoids violation of free-space and enables the labelling of low texture regions where photo-consistency alone leads to ambiguous results. In [4], Gallup et al. use a multi-view plane linking step that enforces global consistency across overlapping views. However, and since in our case the MRF optimization is independently performed in each stereo pair for the sake of scalability, inconsistencies may occur with different labels being assigned to the same structure in different views. In this article, this issue is tackled using a post-processing step that modifies the dense labelling of each view independently. This is achieved by using the information from the dense labelling of another view that overlaps with the first.

2 Background

This section gives a brief review of background concepts that are useful for better understanding the proposed pipeline. We shortly discuss the PEaRL algorithm [19] for geometric multi-model fitting, where the fitting is formulated as an energy-based optimization problem. Then, an overview of the two-view PPR framework proposed in [7] is provided.

2.1 Energy-based Multi-Model Fitting

The authors of [19] discussed that methods that greedily search for models with most inliers while ignoring the overall classification of data are inappropriate for multi-model fitting, and formulating the fitting as a labelling problem with a global energy function is preferable. Following this, they propose the PEaRL algorithm that consists in 3 steps:

1) Propose an initial set of models (labels) $\mathcal{P}$ from the observations.
2) Expand the label set for estimating the spatial support (inlier classification).
3) Re-estimate the inlier models by minimizing some error function.

Given an initial model set $\mathcal{P}$, the multi-model fitting is cast as a global optimization where each model in $\mathcal{P}$ is interpreted as a label l. Consider that $d \in \mathcal{D}$ is a data point and that l_d is a label in $\mathcal{P}$ assigned to d. The objective is to compute the global labelling $\mathbf{l} = \{l_d | d \in \mathcal{D}\}$ such that the following energy is minimized:

$$E(\mathbf{l}) = \underbrace{\sum_{d \in \mathcal{D}} D_d(l_d)}_{\text{data term}} + \lambda_S \underbrace{\sum_{d,e \in \mathcal{N}} V_{d,e}(l_d, l_e)}_{\text{smoothness term}} + \underbrace{\lambda_L \cdot |\mathcal{F}_l|}_{\text{label term}}, \quad (1)$$

where $\mathcal{N}$ is the neighbourhood system considered for d, $D_d(l_d)$ is some error that measures the likelihood of point d belonging to l_d, and $V_{d,e}$ is the spatial smoothness term that encourages piecewise smooth labelling by penalizing configurations $\mathbf{l}$ that assign to neighbouring nodes d and e different labels. The label term is used for describing the data points using as few models as possible, with $\mathcal{F}_l$ being the subset of different models assigned to the nodes d by the labelling $\mathbf{l}$ (refer to [19]). In order to handle outlier data points in $\mathcal{D}$, the outlier label $l_\emptyset$ is used.

Energies containing only data and smoothness terms can be minimized using α-expansion [28]. In case all the terms of Equation 1 are taken into account, then the energy can be optimized using an extension of α-expansion proposed in [19]. On the other hand, if the smoothness term is not considered, the problem becomes an Uncapacitated Facility Location (UFL) instance, which can be solved very efficiently using the message passing inference algorithm [20].

The third step of PEaRL consists in re-estimating the model labels l in $\mathcal{P}$, given the non-empty set of inliers. The new set of labels is then used in a new expand step, and the algorithm iterates between discrete labelling and model refinement until the energy of Equation 1 stops decreasing.

2.2 Pixel-wise Plane Labelling

Given a finite set of plane hypotheses contained in the scene, the final step of many existing PPR algorithms (refer to Section 1.1) is to assign one of these planes to each pixel of the input images. For this purpose, a standard MRF formulation involving only the data and smoothness terms in Equation 1 is employed. The nodes $\mathbf{p}$ are the image pixels, and the labels $l \in \mathcal{P}$ are the plane hypotheses, where the label set $\mathcal{P}$ contains the scene planes $\mathcal{P}_0$ and the infinite plane Π_∞. The data term is defined as

$$D_{\mathbf{p}}(l) = \begin{cases} \min(\rho(l), \rho_m) & \text{if } l \in \mathcal{P} \\ \alpha \rho_m & \text{if } l = l_\emptyset \end{cases}, \quad (2)$$

where $\rho(l)$ is the photo-consistency between the pixels in the two views put into correspondence by the plane associated to label l, ρ_m truncates the cost and $\alpha < 1$. For measuring the photo-consistency the matching cost Zero-mean Normalized Cross-correlation (ZNCC) is used. The smoothness term is defined as in [4].

2.3 Two-View Semi-Dense PPR

The first step of the proposed pipeline consists in obtaining a semi-dense PPR of the scene for each stereo pair. The method described in [7] was chosen for this purpose, mainly due to two characteristics: the fact that it was specifically designed for using two views, while other existing methods (refer to Section 1.1) receive multiple images as input; and because superior results in terms of accuracy were reported when compared to other PPR methods [4], [6]. The 3 major steps of the pipeline presented in [7] are as follows.

Step 1 Stereo-Rangefinding along virtual cut planes: Antunes et al. [16] have introduced a new stereo cost function, dubbed SymStereo, that is particularly well suited for estimating depth along a virtual cut plane passing in-between cameras. The approach uses a symmetry-based metric for obtaining an energy function that encodes the likelihood of each point in the virtual plane being a 3D point of the scene. In other words, the contour where the cut plane meets the scene should be a ridge of maxima in the energy function.

Step 2 Detection of plane hypotheses: Knowing that the intersections of the virtual planes with the planes in the scene are lines, the energy computed in Step 1 is used as input to a Hough transform for extracting line segments. Each pair of lines provides a plausible plane hypothesis.

Step 3 Discrete-Continuous optimization: Let us assume a pencil of virtual cut planes $\mathbf{\Phi}_j$ intersecting the baseline in its midpoint. This can be thought of as an image created by a virtual camera that is located between the cameras (cyclopean eye), where each pixel is originated from the back-projection ray $\mathbf{d}_{j,r}$, corresponding to the intersection between the epipolar plane $\mathbf{\Psi}_r$ and the virtual plane $\mathbf{\Phi}_j$. In [7], the multi-plane fitting problem consisting in assigning a plane label to each pixel of the cyclopean eye is formulated using the PEARL algorithm, with an energy formulation as the one in Equation 1. Since we use their method in an initialization stage, we downsized the energy formulation by ignoring the smoothness term. In this case, the problem is reduced to an UFL instance and the solver of [20] can be used. This modification provides a less accurate but sufficiently good semi-dense PPR of the scene.

3 Overview of the Problem and Proposed Solution

We propose a structure and motion framework that is able to automatically recover the camera positions and orientations along with a PPR of the scene from a stereo sequence. The explanation is given for a two-stereo pair sequence, being extended to longer sequences in a straightforward manner.

The starting point is a semi-dense PPR obtained for each stereo pair using a simplified and computationally more efficient version of the method proposed in [7], as summarized in Section 2.3. Due to its simplicity, problems such as spurious plane detections, inaccuracies due to e.g. low texture and slant, may occur. The reconstruction in Fig. 2a presents some of these issues: the red plane in view 1 is inaccurately recovered due to its long distance to the camera, there is over-segmentation of the frontal plane, and the pink plane in view 2 is incorrectly segmented due to low texture and poor illumination. The output of this step is a set of plane hypotheses, with each reconstructed line contour assigned to one of these hypotheses.

The relative motion between consecutive stereo pairs is determined by registering plane hypotheses. This requires an algorithm for associating and registering planes, as well as strategies to identify situations where plane information is insufficient and must be complemented with point correspondences. The method for carrying this step is presented in Sec. 4. Given the relative motion between frames, the plane hypotheses arising from each pair can be represented in a common reference frame and dense pixel labelling can be carried using a standard MRF formulation as defined in Sec. 2.2. The problem is that a 3D plane in the scene can give rise to multiple labels due to inaccuracies in the plane and relative motion estimation. Fig. 2b illustrates this situation, where it can be seen that there is over segmentation in the labelling, and the reconstruction contains spurious planes.

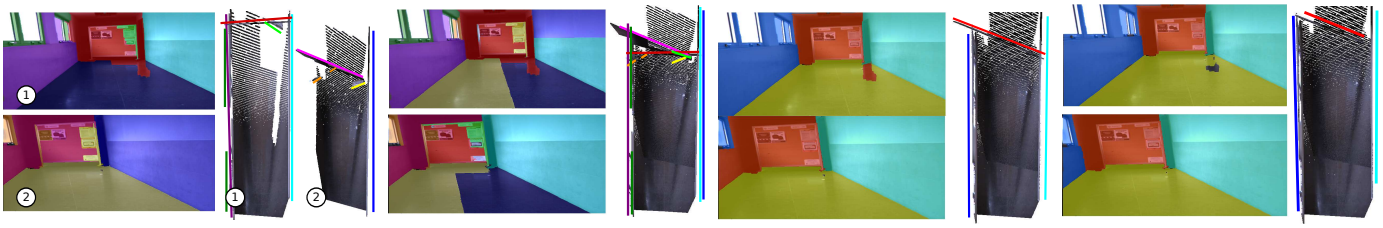
As a consequence, a mechanism for merging plane hypotheses and back-propagating information across views, while simultaneously refining the camera motion and plane parameters is required. This is achieved using the PEARL algorithm reviewed in Sec. 2.1. It consists of a discrete optimization - planes detected in cameras C_i and C_{i+1} are assigned to pixels of the cyclopean eye, by minimizing an energy function in the format of Equation 1 - followed by the joint continuous optimization of the chosen planes and the relative pose. Since the discrete optimization makes use of the energies provided by SymStereo, the first step of this pipeline must be the described PPR. Further details are given in Sec. 5.

Having accurate estimations of planes and camera motion should allow using a standard MRF to obtain a correct dense labelling. However, and as observed in Fig. 2c, this is not always the case as can be noticed in the labelling of the left wall and the floor. These problems have two main reasons: low-textured surfaces may cause photo-consistency to fail, and since the optimization is carried individually for each stereo pair, it might occur that the labelling becomes inconsistent across frames, leading to visibility issues. They are tackled in a final dense labelling step that includes an MRF segmentation followed by a novel post-processing step. Details on this new method are given in Sec. 6. As shown in Fig. 2d, by concatenating the individual reconstructions, it is possible to obtain a dense PPR for the complete sequence and overcome occlusion issues.

As depicted in Fig. 3, the solutions proposed for tackling the aforementioned problems are concatenated in a pipeline that takes as input the individual semi-dense labellings of stereo pairs and a set of plane hypotheses, and outputs a global dense labelling for the stereo sequence.

4 Relative Pose Estimation

Consider two consecutive stereo pairs C_i and C_{i+1}. Let $\mathbf{\Pi}_k^{(i)}$ and $\mathbf{\Pi}_k^{(i+1)}$, with $k{=}1 \ldots K$ be putative plane correspondences across the two pairs. Our objective is to use these plane correspondences to estimate the relative pose $(\mathrm{R}_i, \mathbf{t}_i)$ between the stereo cameras. In [17], it was shown that two minimal sets of 3 corresponding 3D planes can be registered in a closed-form manner if their normals span the entire 3D space. More recently, Taguchi et al. [13] used this registration



(a) Starting point (b) Motion estim. + standard MRF (c) PEARL + standard MRF (d) PEARL + proposed labelling

Fig. 2. Segmentation and reconstruction results obtained in different scenarios: (a) Planes detected in each view independently, where a standard MRF is used only for visualization purposes. (b) Result obtained when applying a regular MRF after the motion initialization. (c) PEARL refinement followed by a regular MRF. (d) Result obtained when applying the proposed pipeline with the new MRF formulation. In all cases, lines with colors corresponding to the detected vertical and nearly vertical planes are shown in the top view of the 3D models.

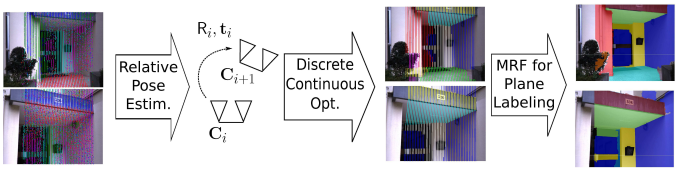
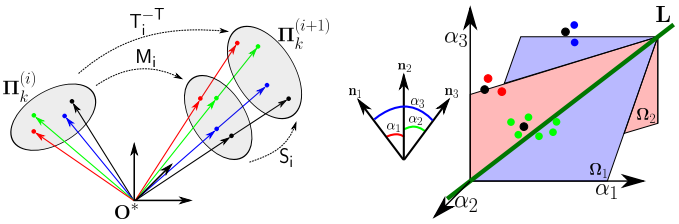


Fig. 3. Different stages of the pipeline. The planes detected in each view are used for pose estimation. The pose and the plane hypotheses are refined in a discrete-continuous optimization step and a final MRF is used for dense labelling. Scene planes are identified with colors.



(a) Registration (b) Triplet matching

Fig. 4. (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-neighbours 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 $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ and line L.

algorithm as a starting point for their plane-based SLAM for RGB-D cameras. They studied the singular configurations and showed how to use reconstructed 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

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

for $k{=}1,2,3$ verifies, where I_3 is the $3{\times}3$ identity matrix, and the 3D planes have the homogeneous representation $\mathbf{\Pi}_k^{(i)}{\sim}[\mathbf{n}_k^\mathsf{T}\ 1]^\mathsf{T}$ and $\mathbf{\Pi}_k^{(i+1)}{\sim}[\mathbf{s}_k^\mathsf{T}\ 1]^\mathsf{T}$. 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 3 can be seen as a projective transformation in $\mathcal{P}^{3*}$ that maps points $\mathbf{\Pi}_k^{(i)}$ into $\mathbf{\Pi}_k^{(i+1)}$ through a rotation M_i followed by a projective scaling S_i, as illustrated in Fig. 4a. R_i is firstly computed by applying the algorithm from [29] that provides a unique solution for aligning two sets of unitary vectors. By replacing R_i in Equation 3, it can be shown after some algebraic manipulation that $\mathbf{t}_i$ is computed by solving the following linear system of equations

$$\underbrace{\begin{bmatrix} \mathbf{s}_1^\mathsf{T}\mathbf{s}_1 & 0 & 0 \\ 0 & \mathbf{s}_2^\mathsf{T}\mathbf{s}_2 & 0 \\ 0 & 0 & \mathbf{s}_3^\mathsf{T}\mathbf{s}_3 \end{bmatrix} \begin{bmatrix} \mathbf{n}_1^\mathsf{T} \\ \mathbf{n}_2^\mathsf{T} \\ \mathbf{n}_3^\mathsf{T} \end{bmatrix}}_{\mathsf{N}_i} \mathsf{R}_i^\mathsf{T}\mathbf{t}_i = \begin{bmatrix} \mathbf{s}_1^\mathsf{T}\mathbf{s}_1 - \mathbf{s}_1^\mathsf{T}\mathsf{R}_i\mathbf{n}_1 \\ \mathbf{s}_2^\mathsf{T}\mathbf{s}_2 - \mathbf{s}_2^\mathsf{T}\mathsf{R}_i\mathbf{n}_2 \\ \mathbf{s}_3^\mathsf{T}\mathbf{s}_3 - \mathbf{s}_3^\mathsf{T}\mathsf{R}_i\mathbf{n}_3 \end{bmatrix} \quad (4)$$

It comes in a straightforward manner that if the 3 normals do not span the 3D space, then N_i is rank deficient and the computation of 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 [29] 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 4.

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 $\mathbf{x}_k^{(i)}, \mathbf{x}_k^{(i+1)}, k{=}1\ldots3$. Related to this problem is the work described in [30], 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.

The 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 $\mathbf{t}_u{=}[t_x\ t_y\ t_z]^\mathsf{T}$, where t_z can be computed as follows. In the rotated configuration, we have

$$[0\ 0\ z_2\ 1]^\mathsf{T} \sim \begin{bmatrix} \mathsf{R}_u & \mathbf{0} \\ -[t_x\ t_y\ t_z]\mathsf{R}_u & 1 \end{bmatrix} [0\ 0\ z_1\ 1]^\mathsf{T}. \qquad (5)$$

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 [30], allowing the epipolar constraint to be written as $\mathsf{A}[t_x\ t_y\ 1]^\mathsf{T} = \mathbf{0}$, where the $3{\times}3$ 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

The relative pose estimation method uses a hierarchical RANSAC scheme that works by considering the maximum number of planes in the image pair, 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 (2 planes and 1 point, and 1 plane and 3 points).

The method starts by building a descriptor (refer to Fig. 4b)) 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 neighbours 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. 4b 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 method in Sec. 4.1. The semi-dense PPR generates a set of line cuts in each frame associated to each 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.

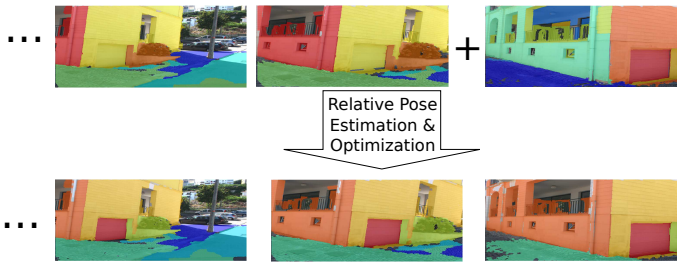


Fig. 5. Back-propagation across stereo pairs: a closer view of the door plane allows its correct detection and propagation to previous pairs.

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 low score, 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-neighbours approach. Since there is only one angle, each match gives rise to two hypotheses. A local feature detector (SURF [31]) is used for extracting point features and solutions are computed in a RANSAC framework from two planes and one point correspondences (Sec. 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 Sec. 4.3. In theory, the scoring metric might fail if the planes surfaces lack texture. An hybrid score metric that mixes planes and points raises other 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 PPR. The previous single stereo PPR and relative pose estimation steps yield two sets of planes defined in the reference frames of cameras C_i and C_{i+1}, $\mathbf{\Pi}_k^{(i)}, k{=}1\ldots K_i$ and $\mathbf{\Pi}_k^{(i+1)}, k{=}1\ldots K_{i+1}$, respectively, and an estimate for the relative pose R_i, $\mathbf{t}_i$ between the cameras. The optimization is achieved using the PEaRL algorithm (Sec. 2.1). The initial set of plane models $\mathcal{P}_0$ for PEaRL is the union of the $(K_i{+}K_{i+1})$ planes. Then, 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 (Sec. 2.3). The objective is to estimate the point on $\mathbf{d}_{j,r}^{(i)}$ that most likely belongs to a planar surface. This problem is cast as a labelling 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 $l_{\mathbf{d}_{j,r}^{(i)}}$. The set of possible labels is $\mathcal{F} = \{\mathcal{P}_0, l_\emptyset\}$, where $l_\emptyset$ is the discard label and is used for identifying non-planar

structures. This labelling problem is solved by minimizing an energy function E in the form of Equation 1, where the data and smoothness terms are modified such that they sum over the whole stereo sequence, becoming

$$\underbrace{\sum_{i}\sum_{\mathbf{d}_{j,r}^{(i)}\in\mathcal{D}} D_{d_{j,r}^{(i)}}(l_{\mathbf{d}_{j,r}^{(i)}})}_{\text{Data Term}} \text{ and } \underbrace{\sum_{i}\sum_{\mathbf{d}_{j,r}^{(i)},\mathbf{e}_{j,r}^{(i)}\in\mathcal{N}} V_{\mathbf{d}_{j,r}^{(i)},\mathbf{e}_{j,r}^{(i)}}(l_{\mathbf{d}_{j,r}^{(i)}}, l_{\mathbf{e}_{j,r}^{(i)}})}_{\text{Smoothness Term}},$$

respectively, where $\mathcal{N}$ is the 4×4 neighbourhood of $\mathbf{d}_{j,r}^{(i)}$ and V is the spacial smoothness term. The data term $D_{\mathbf{d}_{j,r}^{(i)}}$ for the back-projection ray $\mathbf{d}_{j,r}^{(i)}$ is defined as

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

where the coordinate x_l is the column defined by the hypothesis l, corresponding to the intersection of $\mathbf{d}_{j,r}^{(i)}$ with the plane indexed by l. 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 neighbouring 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 $\mathbf{l}$. For each camera i, the smoothness term V is defined as in [7], 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 l, the corresponding energies $\mathrm{E}^{(i)}$ that come from SymStereo can be recomputed to enhance the likelihood measure with respect to a particular range of slant values [7]. These energies are used in the third step of PEARL. Let $\mathbf{\Pi}_l$ be the plane associated to l to which has been assigned a non-empty set of inliers $\mathbf{D}(l) = \{\mathbf{d}\in\mathcal{D} | l_\mathbf{d} = l\}$. All the inlier planes $\{\mathbf{\Pi}_{l_k}\}$ and the relative pose R_i, $\mathbf{t}_i$ are refined simultaneously by minimizing the error function:

$$\{\mathrm{R}_i^*, \mathbf{t}_i^*, \{\mathbf{\Pi}_{l_k}^*\}\} = \min_{\mathrm{R}_i,\mathbf{t}_i,\{\mathbf{\Pi}_{l_k}\}} \sum_i \sum_k \sum_{\mathbf{d}_{j,r}^{(i)}\in\mathbf{D}(l)} \left(1-\mathrm{E}_j^{(i)}(r,x_{\mathbf{\Pi}_{l_k}})\right) + \delta e_{ph}, \quad (6)$$

where $x_{\mathbf{\Pi}_{l_k}}$ is the point defined by the intersection of $\mathbf{d}_{j,r}^{(i)}$ with $\mathbf{\Pi}_{l_k}$, 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 $\mathcal{P}_1 = \{\mathbf{\Pi}_{l_k}^*\}$ is then used in a new expand step, and we iterate between discrete labelling and plane refinement until the α-expansion optimization does not decrease the energy E.

A sliding window approach is applied where at most one relative pose is refined. The exchange of planes between cameras, illustrated in Fig. 5, has an important role in the 3D modelling 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 have plane propagation across distant views, a final PEARL step using a significantly larger sliding window is applied to the whole sequence.

6 Dense Labelling Formulation

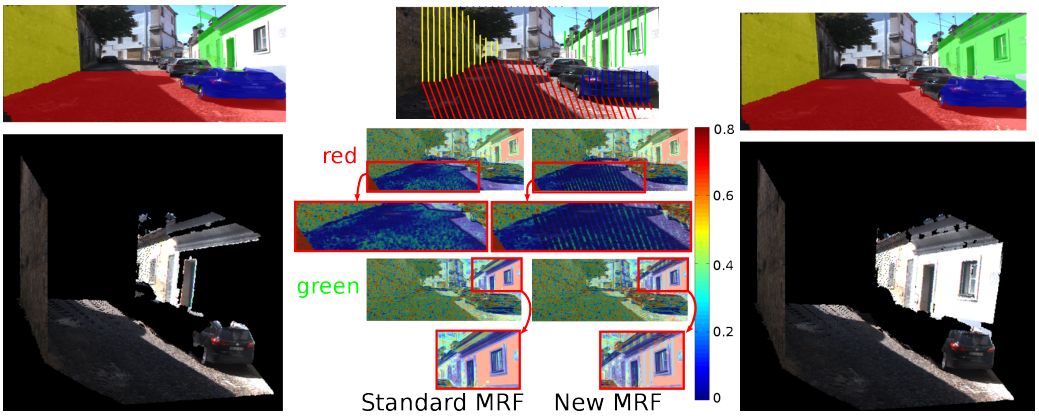
In our previous publication [22], the planes are densely segmented in each stereo pair using a standard MRF labelling as described in Sec. 2.2. Since its data cost solely relies on photo-consistency cues, this formulation tends to provide inaccurate labellings in cases of lack of texture or presence of non-planar surfaces at long distances. The result was not only the reconstruction of non-planar objects such as trees and pedestrians, but also the existence of occlusions and areas that failed to be reconstructed. Two new improvements to the described dense labelling framework are proposed for overcoming these issues. The first one consists in changing the data cost in the MRF formulation, whereas the second is a post-processing step that ensures coherence across views.

6.1 Updated Data Term

The standard MRF formulation described in Sec. 2.2 does not handle cases of textureless regions because the photo-consistency measure (ZNCC) becomes infinity in those regions due to the null variance. This often leads to pixels in textureless regions being discarded because they are assigned the highest cost ρ_m, although the correct plane hypothesis is in the label set. Fig. 6a shows such an example where a large part of the white wall is not reconstructed. This is an important issue as it occurs frequently in outdoor scenarios. One possible solution to the problem would be to increase the penalty for plane change in the smoothness term, which will tend to extend neighbour planes to the textureless region. However, this has many disadvantages including the possibility of reconstructing non-planar objects and not recovering small planes.

Since the SymStereo framework [16] is able to handle low texture, the estimated planes and the semi-dense labelling are usually accurate (Fig. 6b). Although this information has not been used in the dense labelling step, it could be relevant for partially overcoming some issues. Thus, it is proposed to incorporate the semi-dense labelling information in the MRF formulation by changing the data cost with the intent of enforcing coherence between semi-dense and dense labellings. The smoothness term then plays the role of extending this labelling information to neighbouring pixels. The outcome is a robust framework that combines photo-consistency and, indirectly, symmetry cues from SymStereo [16].

The data cost in Equation 2 is modified by decreasing the cost of pixel $\mathbf{p}$ and its neighbours being assigned label f in case $\mathbf{p}$ obtained the same label during the semi-dense labelling. In other words, let us first consider the set of all pixels $\mathbf{p}_i$ that were assigned label f in the semi-dense labelling. Define the $n\times n$ neighbourhood of $\mathbf{p}_i$ as $\mathcal{N}_{\mathbf{p}_i}^f$ and the union of all these neighbourhoods as $\mathcal{U}^f$. If a given pixel $\mathbf{p}$ belongs to $\mathcal{U}^f$, the cost of that pixel being assigned label $l = f$ is decreased by a constant value λ. Also, if a pixel was discarded in the semi-dense labelling, i.e. if $f = l_\emptyset$, it and



(a) Standard MRF (b) Semi-dense labelling (c) New MRF

Fig. 6. (a) Due to the lack of texture, the white wall is not fully reconstructed when using the standard MRF formulation. (b) Data cost of each pixel, for two different labels corresponding to a scene plane, obtained with a standard MRF (left column) and new the MRF formulation (right column). The color corresponding to each plane in the semi-dense labelling (top image) is indicated next to the data cost matrices. The color bar corresponds to the matching cost. (c) The new MRF assigns the correct label to the textureless pixels.

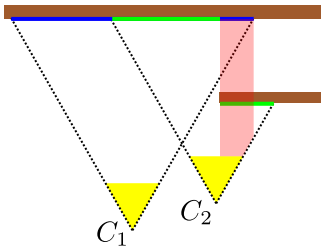


Fig. 7. Example of a case where a legitimate occlusion originates label inconsistency. The proposed approach detects that this occlusion is a legitimate one and does not change the labels of the corresponding pixels.

its neighbours have a constant data cost for all plane labels. The new data cost is then defined as

$$D_{\mathbf{p}}(l)=\begin{cases}\max\left(\min(\rho(l),\rho_m)-\lambda,0\right) & \text{if } l\in\mathcal{P}\wedge\mathbf{p}\in\mathcal{U}^l\\ \alpha\rho_m & \text{if } l=l_\emptyset\vee\mathbf{p}\in\mathcal{U}^{l_\emptyset}\\ \min(\rho(l),\rho_m) & \text{otherwise}\end{cases}, \tag{7}$$

where λ is the parameter that controls the decrease in the cost and $\alpha < 1$. Another parameter to be controlled is the size n of the neighbourhood $\mathcal{N}^f_{\mathbf{p}_i}$. This parameter must be tuned by taking into account the density of virtual cut planes (smaller n for higher density). The remaining parameters are defined as in Sec. 2.2.

Remark that if a pixel of the cyclopean eye is assigned label $f \neq l_\emptyset$, its location in image I is computed from the corresponding plane equation. On the other hand, pixels assigned the discard label $l_\emptyset$ will be reconstructed using the output of SymStereo. For each pixel of the cyclopean eye assigned $l_\emptyset$, the maximum value of the symmetry based matching cost is used for the reconstruction [7], [16].

Fig. 6b shows the semi-dense labelling and the data costs computed with the standard MRF (left column) and the new MRF (right column) for the detected planes. Note that the data costs obtained with the new formulation are usually lower in low textured regions when compared to the standard approach, which proves the robustness of the proposed formulation. This is particularly observable for the green plane (last row), where the reconstruction of the whole surface is now possible (refer to Fig. 6c).

6.2 Label Consistency across Views

A coherent labelling of an image sequence is the one in which corresponding areas in different views have the same label, i.e. represent the same plane. The discrete optimization step partially solves this problem by selecting an appropriate subset of planes. However, as shown in Fig. 2c, a standard MRF for computing the dense labelling may still originate inconsistencies in small areas of the images, leading to problems such as occlusions and reconstruction of non-planar objects. This section proposes a post-processing

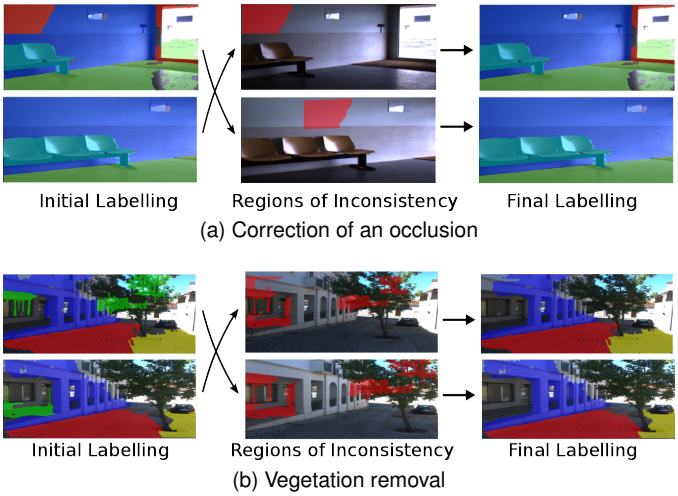


Fig. 8. New post-processing step applied after the MRF labelling to ensure label consistency across frames. The initial dense labelling of each of the two frames (images on the left) is modified by finding the areas of disagreement between labels (images in the middle) and changing them so that corresponding areas in the scene have the same label (images on the right). This allows to correct problems of (a) occlusion and (b) the reconstruction of non-planar objects such as vegetation.

step that enforces consistency across two consecutive views. We refer to the image of camera C_i by I_i, and to its dense labelling by D_i. The procedure is described for correcting D_1, being applied similarly for D_2. It is as follows:

1) Reconstruct the points in image I_1 from the assigned plane labels, representing the corresponding labelling as D_1^2;
2) Find areas of inconsistency between D_1^2 and D_2 by detecting the pixel locations where the assigned labels are different. In Figs. 8a and 8b these areas are shown in red (middle images);
3) The labels of pixels that belong to these areas must be modified to achieve coherence across views. The proposed approach treats the discard and non-discard labels differently. Two labels are considered

(a) No post-processing (b) $l_\emptyset$ as absorbing element (c) $l_\emptyset$ as neutral element

Fig. 9. Reconstruction results obtained when (a) not using the proposed post-processing step, (b) using the post-processing step as it is proposed, and (c) using the post-processing step with the alternative of considering the discard label as a neutral element.

for each of these pixels, l_1^2 from D_1^2 and l_2 from D_2. If $l_2 = l_\emptyset$, then l_1 is also set to $l_\emptyset$, where l_1 is the label corresponding to l_1^2 in image I_1. Thus, the discard label $l_\emptyset$ works like an absorbing element;

4) Otherwise, if $l_2 \neq l_1^2 \wedge l_2 \neq l_\emptyset \wedge l_1^2 \neq l_\emptyset$ the point that is being analysed is reconstructed using both l_2 and l_1^2. The distances of both 3D points to camera C_2 are computed and l_1 is set to the label that yields the shortest distance. The reasoning is that, in general, closer surfaces are reconstructed more accurately than farther ones.

Note that there are cases in which having label inconsistency between views is correct. Fig. 7 illustrates such an example. Two cameras observe a scene that contains two walls. While camera C_1 only observes the farther wall, as shown by the blue line, camera C_2 views both of them since they are inside its Field-of-View (FOV), depicted by the green line. The area in red shows the region where there is overlap of the cameras' FOVs, originating label inconsistencies. However, the points that belong to the closer wall viewed by camera C_2, when projected on the image plane of camera C_1, fall outside its FOV. Thus, our post-processing step does not modify the labellings in this case, being able to distinguish between legitimate and non-legitimate occlusions.

The described procedure allows to achieve consistency in all pixels of two views, enabling the correction of possible reconstruction errors. Fig. 8 depicts two different common errors that can occur: occlusions and reconstruction of non-planar objects. In Fig. 8a, although all label assignments are correct for the first view, some pixels that belong to the wall were incorrectly assigned to the door plane in the second view. This generates an inconsistency, since the area that is incorrectly reconstructed cannot be observed by any of the cameras as it is occluded by the wall. Correcting the labelling so that it becomes coherent across frames yields an accurate reconstruction without occlusions. Fig. 8b depicts an example where a non-planar object (vegetation) is assigned to an existing plane when it is observed from a long distance, and only correctly discarded in a closer view. By ensuring consistency across frames, all of the pixels belonging to the tree become discarded, originating a reconstruction without non-planar objects. Remark that this only happens because the discard label is being considered as an absorbing element. It could instead work as a neutral element, and discarded pixels in one view that originated inconsistencies would be assigned the other label. In Fig. 9, concatenated reconstructions are shown for three different scenarios. It can be seen in Fig. 9b that using the proposed post-processing step - discard label as an absorbing element - originates the complete removal of pedestrians, but has the disadvantage of also removing the surrounding parts, caused by the fact that the pedestrians are moving. If the discard label was considered as a neutral element (Fig. 9c), the outcome would be the reconstruction of the same pedestrian in both views, becoming visible in different positions. Although this can be an interesting application in certain cases, considering $l_\emptyset$ as a neutral element would cause the vegetation in Fig 8b to be reconstructed, which is a very poor approximation. Since this is a situation that happens very often in urban scenes, the choice of considering the discard label as an absorbing element is more appropriate.

7 Experimental Results

This section reports several experiments that validate the PPSS framework, show that using plane primitives effectively improves the accuracy of the pipeline and compare its performance against two conventional point-based methods with very different complexities.

7.1 Competing method, datasets and metrics

PPSS is compared against the two competing point-based methods LIBVISO2 [1] and VisualSFM + CMVS/PMVS [2], [3], [26], [27]. LIBVISO2 is a fast, ready to use stereo method broadly employed by the community. However, since it can be argued that LIBVISO2 sacrifices accuracy by low computational time, we also tested with VisualSFM + CMVS/PMVS, which is a sophisticated pipeline that can arguably do the best possible using point primitives. Since VisualSFM is not specifically tailored to work with stereo sequences, we limited the matching to the six different pairs of frames in every two consecutive stereo pairs, and performed a constrained BA where the extrinsic calibration of the stereo camera was fixed between the left and right images of each stereo pair and kept unchanged.

In the absence of a suitable public dataset, we collected new image sequences using three distinct camera setups whose characteristics are summarized in Table Fig. 10a. Setup S1 is a Bumblebee camera from PointGrey that was used to collect indoor sequences. The other two setups consist of a pair of synchronized cameras mounted on the

Setup	S_1	S_2	S_3
Brief Description	PtGrey Bumblebee	Vehicle Forward	Vehicle Lateral
Resolution	1024×384	1142×410	1187×436
Baseline	$24cm$	$80cm$	$80cm$
FPS	-	7.5	7.5
Max. Range	10.3m	30.3m	31.6m

(a) Specifications of the acquisition setups

(b) Cameras mounted on a car and respective acquired stereo pair

Fig. 10. (a) Information about image resolution, stereo camera baseline, acquisition rate and maximum range for accurate depth estimation for the 3 acquisition setups. The maximum range is computed by considering a depth error of 2% and a disparity error of 0.6 pixels.(b) Vehicle setups S_2 and S_3 with a corresponding acquired stereo pair enclosed in a black and red box, respectively.

roof of a vehicle whose images were calibrated and rectified using standard methods [32]. In Setup S2 the cameras were mounted in a forward looking position, while in S3 the cameras were pointing to the right side of the vehicle (Fig. 10b).

Due to the absence of ground truth in the acquired datasets, the quality of the motion estimation in loop closing sequences is measured quantitatively by computing the loop closing error in the following way.

Let T_i be the estimated camera motion between positions i and $i+1$. For a sequence of F frames, the loop closing error is computed as the relative rotation and translation between the 4×4 identity matrix I_4 and the transformation

$$\mathsf{T}_e = \left(\prod_{i=1}^{F-1} \mathsf{T}_i \right) \mathsf{T}_F, \tag{8}$$

where T_F is the pose between position F and position 1. Since matrix multiplication does not have the commutative property, different errors will be obtained if we consider different starting points. Note that varying the starting point translates into a cycling permutation of the product defined in Equation 8. Although the norm of the translation component in T_e varies with the starting point, the angle of rotation θ associated with the rotation component R_e does not. From Rodrigues' rotation formula, θ is computed by

$$\theta = \cos^{-1}\left(\frac{1}{2} \left(tr(\mathsf{R}_e) - 1 \right) \right), \tag{9}$$

which, as can be seen, only depends on the trace tr of matrix R_e. Since the trace has the cyclic property, meaning that it is invariant under cyclic permutations, the trace of transformation T_e, and consequently of its rotation component R_e, does not vary with the starting point. This explains why the rotation error θ is invariant under different starting points. Thus, it suffices to present information about the translation error as a function of the starting position.

7.2 Impact of Plane-Primitives in Performance

This experiment serves to assess how the accuracy of the motion estimation varies with the number of planes. For this, we acquired a 20-frame loop-closing sequence of stereo images with setup S_1 that has small displacement between consecutive frames. Also, the acquisition was performed to guarantee that the sequence has at least three shared planes between consecutive views, allowing the camera motion to be estimated with 3 planes, 2 planes and 1 point, 1 plane and 3 points. Sample images of this sequence are shown in Figure 11a.

The motion errors shown in Fig. 11c evince the fact that using planes for motion estimation is beneficial since a significant increase in accuracy can be achieved with an increasing number of planes. For the case of 3 planes, median errors as low as $0.44°$ and $1.2cm$ are obtained, significantly outperforming both point-based methods. Considering that the scene walls and the floor are perpendicular (Fig. 11b), we provide a quantitative evaluation of the reconstruction result by measuring the deviation from perpendicularity of the estimated planes and reporting the average errors in the caption of Fig. 11. A comparison with the point-based methods is given by finding the 3D planes that give the best fit to the reconstructed 3D points corresponding to the walls and floor. Considering all stereo pairs of the sequence, an average error of $0.48°$ was obtained for PPSS, while VisualSFM and LIBVISO2 yielded errors of $1.03°$ and $1.27°$, respectively, demonstrating the high accuracy of the PPR obtained with our method.

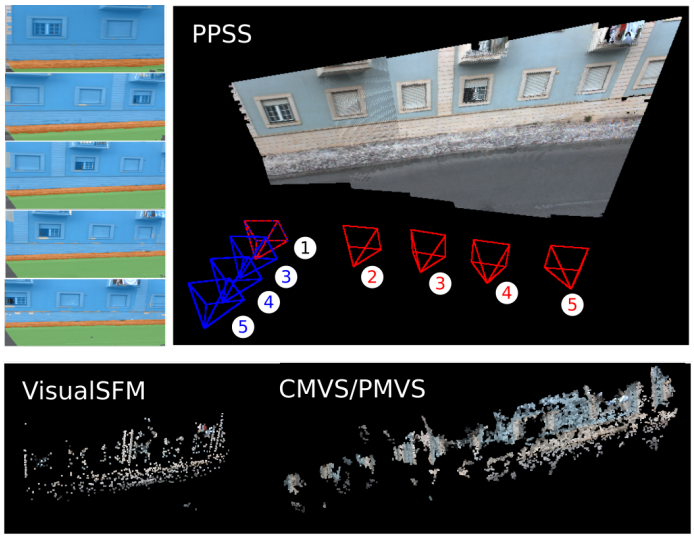
7.3 Benefits of PPSS

As already observed in [22], PPSS is able to handle cases of wide-baseline and low or repetitive texture (Fig. 12), as well as situations of reflection and dynamic foreground where the LIBVISO2 [1] fails. However, and since PPSS takes in average $45s$ to process each stereo frame, while LIBVISO2 does the same in only $50ms$, it is important to also compare against a more complex point-based method in order to be conclusive about benefits of using plane primitives for SfM. This section presents experiments in short indoor sequences where PPSS is confronted not only with LIBVISO2, but also with the more sophisticated VisualSFM [2], [3] pipeline that is complemented with CMVS/PMVS [26], [27] to obtain dense reconstructions. The VisualSFM + CMVS/PMVS takes in average $40s$ to process each stereo pair, having a computational complexity similar to PPSS.

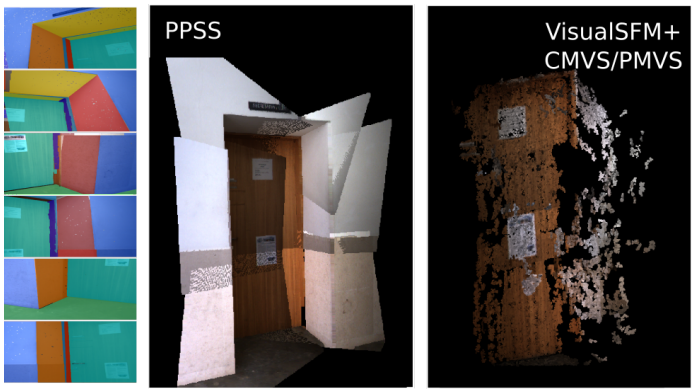
Two short stereo sequences were selected with the intent to show how using planes in motion estimation and reconstruction is advantageous over using solely points. In Fig. 12a, camera symbols are shown in red and blue, if they were computed using PPSS or LIBVISO2, respectively. Since the motion results provided by VisualSFM are similar to ours, and due to the absence of ground truth, camera symbols for VisualSFM are not included. The left images of the stereo pairs are shown with the overlaid plane labelling, where each color identifies one plane and the same color across images corresponds to the same plane. The sequence of images is sorted from top to bottom, and the cameras are numbered accordingly.

(a) Loop-closing sequence (b) Reconstruction by PPSS (c) Translation and rotation errors

Fig. 11. Experiment performed on a (a) 20-frame loop-closing sequence to evaluate how the use of planes affects the motion and reconstruction accuracies. (b) depicts the reconstruction obtained with our method and (c) shows the rotation and translation errors. The average reconstruction errors obtained when using 1, 2 and 3 planes are 1.07°, 0.59° and 0.48°, respectively. For the two point-based competing methods, the reconstruction errors were higher, being 1.27° for LIBVISO2 and 1.03° for VisualSfM.



(a) Perceptual aliasing



(b) Small overlap and low texture

Fig. 12. Example cases of the presence of (a) perceptual aliasing and (b) small overlap and low texture where the proposed plane-based method performs accurately, while the point-based method LIBVISO2 fails. Red and blue camera symbols show the relative pose estimated using the proposed method and LIBVISO2, respectively. Dense reconstructions obtained with VisualSFM+CMVS/PMVS are also shown.

A 5-stereo pair sequence of a scene with the presence of perceptual aliasing was acquired with stereo camera S_3. Results in Fig. 12a show that our method was able to provide an accurate reconstruction of the scene, properly distinguishing between the road (green) and side walk (orange) planes. On the contrary, LIBVISO2 performed poorly on this sequence, not being able to estimate the first relative pose and providing inaccurate results for the remaining ones. This is a consequence of the perceptual aliasing, as most of the extracted point matches are incorrect. Despite this difficulty, VisualSFM was able to correctly estimate the camera poses, providing a plausible sparse reconstruction. However, due to the small number of good point matches, the densification of the sparse model obtained with CMVS/PMVS is very poor, especially in the area corresponding to the first two stereo pairs.

PPSS yields a detailed reconstruction of a door and surrounding walls from a sequence of only six stereo pairs with minimum overlap (Fig. 12b). It can be seen that the white low-textured walls and the small interior planes were accurately recovered. LIBVISO2 failed to find sufficient point correspondences for estimating the camera motion and thus camera symbols are not shown. Our approach computed the camera motion using correspondences of two planes and one point, as there are no triplet correspondences in consecutive stereo pairs. In this challenging low-textured sequence, VisualSFM still managed to properly estimate the camera motion. However, since the majority of the point matches belong to the door, CMVS/PMVS is unable to recover the surrounding walls and all details that our approach provides.

These experiments demonstrate that, in fact, VisualSFM managed to overcome the situations where LIBVISO2 failed. However, due to the very small number of point matches, the reconstruction results obtained with CMVS/PMVS have incomparably much lower quality than the ones accomplished with PPSS. Thus, there are important benefits that prevail when comparing PPSS against sophisticated point-based pipelines with computational efforts of the same order of magnitude.

7.4 Camera Motion Estimation on the KITTI Dataset

Three sequences of the KITTI dataset [33] containing buildings were selected for testing PPSS, LIBVISO2 and VisualSFM. It was observed that using the original sequences, both point-based methods performed more accurately than

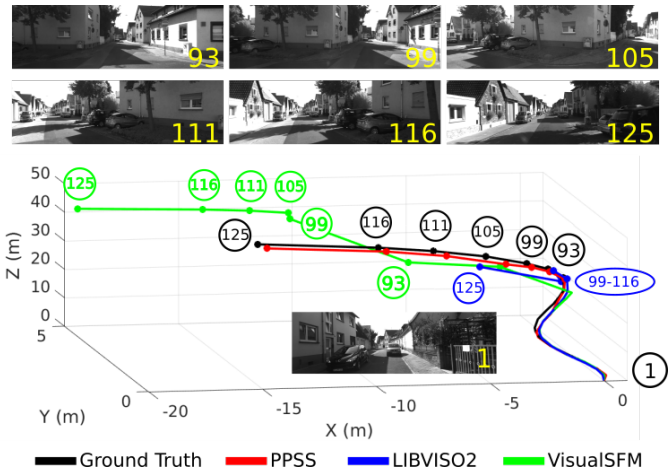


Fig. 13. Results on a sequence of the raw data KITTI dataset [33] (2011_09_26_*drive*_0046), using a sampling of frames that originates an average distance between consecutive frames of about 4 meters. The point-based methods LIBVISO2 and VisualSFM outperform our method when there is significant overlap between frames (as a consequence of straight movement) but diverge otherwise, while our method still provides acceptable results.

our method, originating average errors of approximately 2% and 5% in translation, and $0.1°$ and $0.15°$ in rotation, for LIBVISO2 and VisualSFM, respectively, as opposed to 6% and $0.5°$ for PPSS. The reason for this is that the KITTI sequences are not well suited to evaluate methods that rely in plane primitives. Although KITTI comprises sequences acquired in urban environment, the baseline between stereo cameras is small and only enables accurate depth estimation up to 13 meters [34]. Thus, many building façades that provide the plane surfaces to be used by PPSS are in general too far away to be properly reconstructed. The short baseline favours point-based methods because it provides large image overlap and prevents changes in perspective that can hamper matching.

In order to assess the performance with smaller amounts of data, and create difficulties to the point-based methods, we sampled the sequences by considering only every fifth frame (corresponding to about 4 meters between consecutive frames). We observed that both LIBVISO2 and VisualSFM have good performance when the combination of camera motion and scene structure results in images with much overlap, yielding similar errors as the ones obtained with the complete sequences. In cases of smaller overlap, they easily diverge, as shown in Fig. 13 that they were not able to accurately estimate the camera motion near the curve. As expected, reducing the number of frames did not influence the performance of PPSS as it properly handles situations of wide baseline. Thus, the average errors in the motion estimation were nearly the same as the ones obtained for the complete sequences.

7.5 3D Urban Modelling from Street Viewpoint

This experiment consists in the camera motion estimation and reconstruction of a 1370-frame 1100-meter loop-closing sequence acquired with stereo camera S_2. This is a challenging sequence of a curvy path in a hilly area of Coimbra, as can be seen in Figs 1 and 14c. The presence of slanted building façades, vegetation, moving vehicles and pedestrians further hampers the camera motion estimation and reconstruction processes.

Fig. 14b shows the translation error distributions and rotation errors obtained with the methods of the top table in Fig. 14a. It can be seen that even without the final optimization step, PPSS outperforms LIBVISO2 in the estimation of the rotation component. The optimization step, which was performed with a sliding window of 6 frames, significantly improved the results, yielding a small rotation error and thus translation errors with a small standard deviation. The bottom table in Fig. 14a shows the percentage of frames in which PPSS used 3, 2, 1 or no planes to compute the camera motion. Remark that despite providing a smaller minimum translation error, LIBVISO2 performs worse than our method as the median error is considerably larger. Since the translation value is affected by errors in the rotation, a large rotation error may lead to small translation errors, depending on the starting position, not meaning that the overall result is better. The more sophisticated VisualSFM performs considerably better than LIBVISO2 and PPSS without the final optimization step. However, despite its complexity, it is still not able to outperform PPSS with the optimization step. The superior results of our method can be explained by the fact that this sequence is challenging, presenting many curves and significant variations in height. Due to the low acquisition frame rate, consecutive stereo pairs sometimes present small overlap, preventing point-based methods from extracting enough point correspondences to accurately estimate the camera motion.

The lack of sufficient point matches is also evinced by the reconstruction result shown in Fig. 14c, which includes a top view of the full 3D model, with some areas in a different perspective. Each area is shown for both VisualSFM + CMVS/PMVS and PPSS + Opt., enclosed in a light green and cyan ellipse, respectively. These detailed views show that while our method provides a visually pleasant and complete 3D reconstruction, CMVS/PMVS provides very poor dense reconstructions in areas with faraway planes and sharp turns. This example demonstrates the clear superiority of PPSS in providing accurate and complete 3D models when compared to point-based approaches.

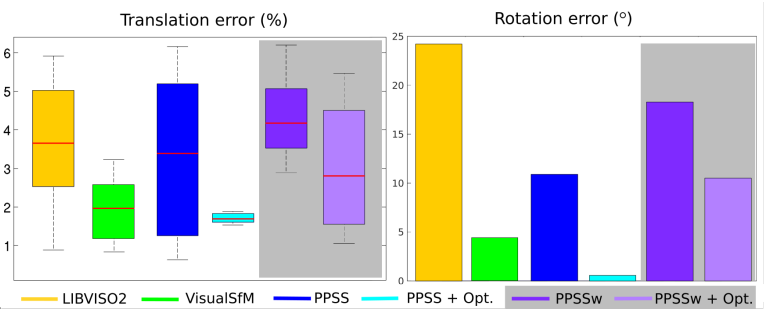
In order to assess the performance of PPSS in a much wider baseline dataset, the sequence was sampled by considering only every fourth frame and is denoted with a lower-case w (PPSSw and PPSSw + Opt. in the top table in Fig. 14a). A rotation error of approximately $10°$ and a median translation error of 3% were obtained, representing a degradation in overall accuracy. The reconstruction results obtained with the sampled sequence were slightly inferior not only due to the decreased quality of the camera motion estimation but also because less details are recovered. This can be observed in Fig. 14c in the reconstructed areas inside purple ellipses. However, such a small dataset still provided good results, being useful in applications that require working with small amounts of data. A video of the full reconstruction can be accessed at https://youtu.be/IhELZ3-wPU0.

The execution times of each modality of methods for the complete sequence are shown in the top table of Fig. 14a. Although LIBVISO2 runs in less than 5 minutes, its perfor-

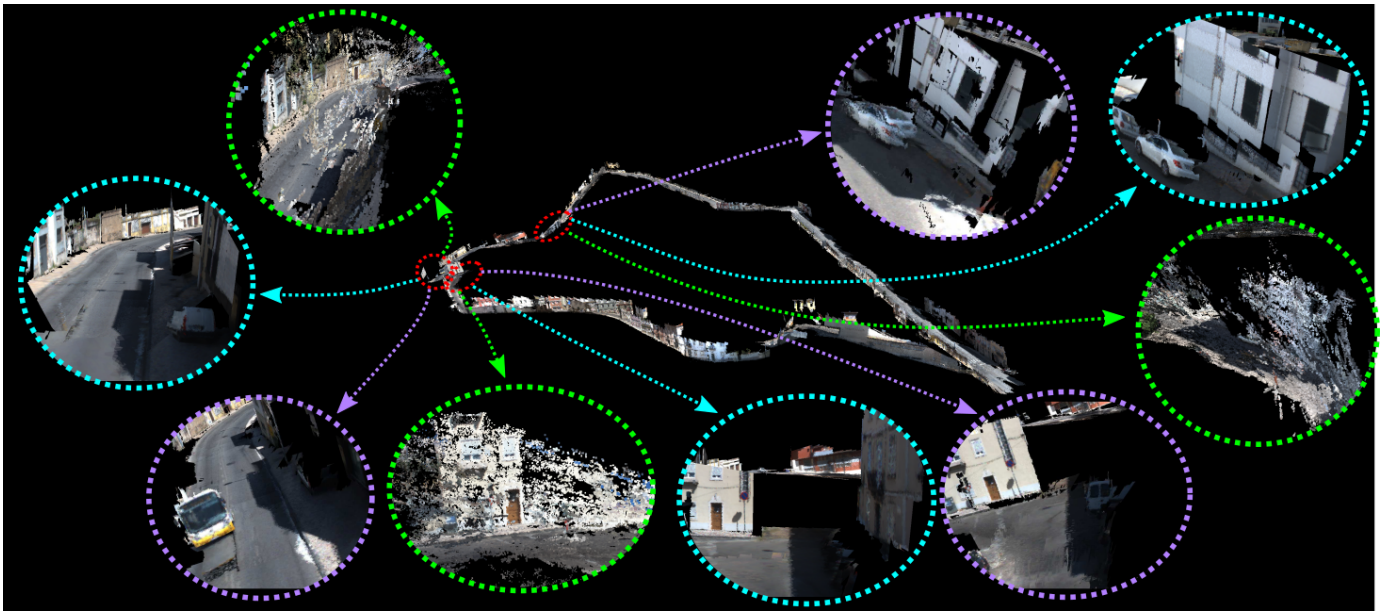
Modality	Method	No. Ims	Time
LIBVISO2	Point based	1370	0.08
VisualSFM	Point based	1370	15
PPSS	Plane based	1370	15.2
PPSS + Opt.	PPSS + final PEARL	1370	17
PPSSw	Plane based	**380**	4.2
PPSSw + Opt.	PPSS + final PEARL	**380**	4.8

Number of planes	3	2	1	0
%	15.4	59.9	24.3	0.2

(a) Top: different methods and their execution times (hours). Bottom: statistics of the number of planes used by PPSS.



(b) Translation and rotation errors



(c) 3D model

Fig. 14. (a) Different modalities of methods and datasets and statistics of the number of planes used, and (b) corresponding rotation and translation loop closing errors. Since the translation error depends on the rotation, the median value in the translation errors distribution (red mark) should be considered when assessing the quality of the motion estimation. (c) Full 3D reconstruction using PPSS + Opt.. Selected areas are shown in greater detail with cyan ellipses. A video of the 3D reconstruction can be accessed in https://youtu.be/wrBaV7O1Q7Q. For the same selected areas, the dense reconstructions obtained with VisualSFM + CMVS/PMVS and PPSSw + Opt. are shown inside green and purple ellipses, respectively.

mance is poor when compared to the other more time consuming approaches. Our method, despite being the slowest, only takes approximately 13% longer than VisualSfM and is the top performer by a large margin. This experiment clearly demonstrates the usefulness of plane primitives in odometry and PPR.

8 Conclusion

We described the PPSS pipeline for sequential PPR from stereo images, where the relative pose between consecutive frames is preferentially estimated using plane-primitives, and the motion and structure are jointly refined using a PEARL framework [19] that alternates between discrete optimization to enforce coherent PPR across stereo frames, and continuous bundle adjustment to improve the accuracy of the results. The rendering of complete and visually pleasant 3D models of the scene is possible thanks to a final dense labelling step, which assigns 3D planes hypotheses to image pixels in a manner that is robust to textureless regions and takes into account visibility constraints.

PPSS was tested in several datasets, including a challenging sequence collected by a stereo camera mounted on the roof of a car (Fig. 1 and 14). 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. It was also observed that, in general, the use of planes improves the overall accuracy of visual odometry, which suggests that plane primitives are an alternative to point correspondences that must be considered when developing SfM pipelines.

The current implementation of PPSS takes in average $45s$ to process each stereo pair, which, although inline with the times observed for VisualSFM+CMVS/PMVS, is significantly more than the execution time of LIBVISO2 in the same computer. We believe that there is substantial margin of improvement, namely by exploring the parallel nature of some of the PPSS components. A first effort in developing a GPU implementation of the two view PPR (Sec. 2.3) led to a speedup of $30\times$. This module corresponds to about 30% of

the current execution time of the PPSS pipeline. We intend to pursue this effort further with the goal of reaching 1 fps, while maintaining the benefits of plane-based SfM in terms of robustness, accuracy and quality of outputted 3D models.

ACKNOWLEDGMENTS

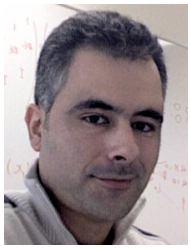
The authors thank Google, Inc for the support through a Faculty Research Award. Carolina Raposo acknowledges the Portuguese Science Foundation (FCT) for funding her PhD under grant SFRH/BD/88446/2012. The work was also partially supported by FCT and COMPETE program under Grant AMS-HMI12: RECI/EEI-AUT/0181/2012. The authors would also like to thank João Marcos for his contribution in the acquisition of datasets S_2 and S_3.

REFERENCES

[1] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3d reconstruction in real-time," in *Intelligent Vehicles Symposium (IV)*, 2011.
[2] C. Wu, "Towards linear-time incremental structure from motion," in *3DV*, 2013.
[3] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz, "Multicore bundle adjustment," in *CVPR*, 2011.
[4] D. Gallup, J.-M. Frahm, and M. Pollefeys, "Piecewise planar and non-planar stereo for urban scene reconstruction," in *CVPR*, 2010.
[5] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski, "Manhattan-world stereo," in *CVPR*, 2009.
[6] S. Sinha, D. Steedly, and R. Szeliski, "Piecewise planar stereo for image-based rendering," in *ICCV*, 2009.
[7] M. Antunes, J. P. Barreto, and U. Nunes, "Piecewise-planar reconstruction using two views," *Image and Vision Computing*, 2016.
[8] T. Werner and A. Zisserman, "New techniques for automated architectural reconstruction from photographs," in *ECCV*, 2002.
[9] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski, "Reconstructing building interiors from images," in *ICCV*, 2009.
[10] B. Micusik and J. Kosecka, "Piecewise planar city 3d modeling from street view panoramic sequences," in *CVPR*, 2009.
[11] P. Alcantarilla, C. Beall, and F. Dellaert, "Large-scale dense 3d reconstruction from stereo imagery," in *5th Workshop on Planning, Perception and Navigation for Intelligent Vehicles (PPNIV13)*, 2013.
[12] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the world from internet photo collections," *IJCV*, 2008.
[13] Y. Taguchi, Y.-D. Jian, S. Ramalingam, and C. Feng, "Slam using both points and planes for hand-held 3d sensors," in *ISMAR*, 2012.
[14] C. Raposo, M. Lourenco, M. Antunes, and J. P. Barreto, "Plane-based odometry using an rgb-d camera," 2013.
[15] M. Antunes and J. P. Barreto, "Semi-dense piecewise planar stereo reconstruction using symstereo and pearl," in *3DimPVT*, 2012.
[16] ——, "Symstereo: Stereo matching using induced symmetry," in *IJCV*, 2014.
[17] W. Grimson and T. Lozano-Pérez, "Model-based recognition and localization from sparse range or tactile data," *IJRR*, 1984.
[18] K. Pathak, A. Birk, N. Vaskevicius, and J. Poppinga, "Fast registration based on noisy planes with unknown correspondences for 3-d mapping," *T-RO*, 2010.
[19] H. Isack and Y. Boykov, "Energy-based geometric multi-model fitting," *IJCV*, 2012.
[20] N. Lazic, B. J. Frey, and P. Aarabi, "Solving the uncapacitated facility location problem using message passing algorithms." *Journal of Machine Learning Research*, 2010.
[21] A. Delong, A. Osokin, H. Isack, and Y. Boykov, "Fast approximate energy minimization with label costs," *IJCV*, 2012.
[22] C. Raposo, M. Antunes, and J. P. Barreto, "Piecewise-planar stereoscan:structure and motion from plane primitives," in *ECCV*, 2014.
[23] D. Nister, O. Naroditsky, and J. Bergen, "Visual odometry," in *CVPR*, 2004.
[24] T. Kazik, L. Kneip, J. Nikolic, M. Pollefeys, and R. Siegwart, "Real-time 6d stereo visual odometry with non-overlapping fields of view," in *CVPR*, 2012.
[25] E. Dunn, B. Clipp, and J.-M. Frahm, "A geometric solver for calibrated stereo egomotion," in *ICCV*, 2011.
[26] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Towards internet-scale multi-view stereo," in *CVPR*, 2010.
[27] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multi-view stereopsis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1362–1376, 2010.
[28] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE TPAMI*, 2001.
[29] B. K. P. Horn, "Closed-form solution of absolute orientation using unit quaternions," *J. Opt. Soc. Am. A*, 1987.
[30] F. Fraundorfer, P. Tanskanen, and M. Pollefeys, "A minimal case solution to the calibrated relative pose problem for the case of two known orientation angles," in *ECCV*, 2010.
[31] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *CVIU*, 2008.
[32] J.-Y. Bouguet, "Camera calibration toolbox for matlab," http://www.vision.caltech.edu/bouguetj/calib_doc/, 2008.
[33] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*, 2012.
[34] D. Gallup, J.-M. Frahm, P. Mordohai, and M. Pollefeys, "Variable baseline/resolution stereo," in *CVPR*, 2008.

Carolina Raposo received the Integrated Master's degree (BSc+MSc) in Electrical and Computer Engineering from the University of Coimbra, Portugal, in 2012. Since 2012 she is a computer vision researcher at the Institute for Systems and Robotics - Coimbra. She is currently a PhD student at the University of Coimbra, Portugal. Her main research interests lie on geometric computer vision, 3D reconstruction and Structure-from-Motion.

Michel Antunes received the Integrated M.Sc. degree in Biomedical Engineering and the Ph.D. degree in Electrical Engineering from the University of Coimbra, Portugal in 2008 and 2014, respectively. He has worked at the Institute of Systems and Robotics (ISR), Coimbra, Portugal, and at the Mitsubishi Electrical Research Laboratories (MERL), Boston, USA. His research interests include several topics on computer vision, such as stereo matching, 3D reconstruction and modelling, 3D geometry estimation, motion estimation and activity recognition. Since April 2014, Michel is a Research Associate at the Interdisciplinary Centre for Security, Reliability and Trust (SnT), Luxemburg.



João P. Barreto (M'99) received the "Licenciatura" and Ph.D. degrees from the University of Coimbra, Coimbra, Portugal, in 1997 and 2004, respectively. From 2003 to 2004, he was a Post-doctoral Researcher with the University of Pennsylvania, Philadelphia. He has been a Professor with the University of Coimbra, since 2004, where he is also a Senior Researcher with the Institute for Systems and Robotics. His current research interests include different topics in 3D computer vision, with a special emphasis in robotics and medicine. He is the author of more than 70 peer-reviewed publications and recipient of several distinctions and awards including a Google Faculty Research award and 5 Outstanding Reviewer Awards. He is Associate Editor for Computer Vision and Image Understanding, Image and Vision Computing and Journal of Mathematical Imaging and Vision.