Piecewise-planar 3D approximation from wide-baseline stereo

C. Verleysen and C. De Vleeschouwer
ICTEAM institute, Université catholique de Louvain (UCL), Louvain-la-Neuve, Belgium
{cedric.verleysen, christophe.devleeschouwer}@uclouvain.be

Abstract

This paper approximates the 3D geometry of a scene by a small number of 3D planes. The method is especially suited to man-made scenes, and only requires two calibrated wide-baseline views as inputs. It relies on the computation of a dense but noisy 3D point cloud, as for example obtained by matching DAISY descriptors [35] between the views. It then segments one of the two reference images, and adopts a multi-model fitting process to assign a 3D plane to each region, when the region is not detected as occluded. A pool of 3D plane hypotheses is first derived from the 3D point cloud, to include planes that reasonably approximate the part of the 3D point cloud observed from each reference view between randomly selected triplets of 3D points. The hypothesis-to-region assignment problem is then formulated as an energy-minimization problem, which simultaneously optimizes an original data-fidelity term, the assignment smoothness over neighboring regions, and the number of assigned planar proxies. The synthesis of intermediate viewpoints demonstrates the effectiveness of our 3D reconstruction, and thereby the relevance of our proposed data fidelity-metric.

1. Introduction

Estimating the 3D model of a scene from images captured by widely separated cameras offers two advantages compared to its estimation from small-baseline stereo. First, the 3D estimated from triangulation in wide-baseline setups is less impacted by unprecise correspondences or calibration inaccuracies than in small-baseline ones [17]. Second, 3D reconstruction from widely separated views results in 3D models that are consistent with a wider range of viewpoints, thereby enabling to synthesize a larger range of virtual views of the scene. However, the large occlusions and strong (projective) deformations affecting wide-baseline views make the determination of a dense matching much more challenging than in its small-baseline counterpart. For this reason, most of the state-of-the-art multi-view stereo (MVS) methods still rely on a dense network of small-baseline stereo pairs [1] [23] [36] to estimate the 3D, even if the two outermost cameras might form a wide-baseline stereo pair. Due to cost or practical deployment constraints, it is however not always possible to install many cameras around the scene. To address the reconstruction problem in sparse acquisition setups, our paper promotes the use of prior knowledge about the 3D geometry of the scene. Namely, it proposes a solution to reconstruct a scene from only two wide-baseline views, in cases for which the 3D scene exhibits a piecewise-planar geometry, as often encountered in man-made[1] scenes.

Our piecewise-planar reconstruction is formulated as a 3D planes assignment problem over the 2D regions that are obtained in one of the two reference images based on a color segmentation [40][2]. In contrast to most previous works dealing with wide-baseline setups [4] [1] [27], our method builds upon a dense 3D point cloud[3], instead of a sparse set of correspondences between keypoints. Although dense point clouds offer the advantage to provide 3D cues for challenging surfaces, *e.g.*, textureless or with repetitive patterns such as paved floors, they are generally much more corrupted by noise and 3D outliers than sparse ones. This noise makes it ineffective to directly fit planar models to the cloud. Therefore, our method first derives a set of planar hypotheses from the cloud, and then assigns them to the image regions. The assignment is done by optimizing an energy function that favors (i) assignment smoothness across neighboring regions, (ii) consistency between the assigned models and the dense point cloud, and (iii) sparsity of plane models. The success of our approach funda-

[1] A man-made scene is composed of manufactured 3D objects, which are observed by real cameras.

[2] In practice, the parameters of the segmentation are tuned to oversegment the image, so that it becomes unlikely that pixels that belong to the same region lie on distinct planar surfaces.

[3] We define a dense point cloud as a set of 3D points whose projection fully covers (at least one of) the reference images.

mentally depends on the capacity to derive accurate planar models hypotheses, and on the definition of a data fidelity metric that is able to deal with the noise inherent to the dense cloud. Overall, the main contributions of our proposed *plane hypotheses assignment* method are:

- A method to define, from a dense but noisy point cloud, a set of 3D plane hypotheses that includes most of the planar surfaces composing the 3D scene, while having a small cardinality (Section 3).
- A plane-to-region data-fidelity metric that accounts for the inaccuracy and ambiguity of the matching inherent to a dense 3D point cloud construction (Section 4).
- An energy-driven formulation of the plane-to-region assignment problem, which maximizes the data-fidelity and the smoothness of the plane assignment over the regions, while minimizing the number of assigned planes. This last term guarantees to approximate the 3D with a small number of planes, without having to fix this parameter *a priori* or having to merge many similar plane models *a posteriori*, as done in [4] (Section 5).

To the best of our knowledge, our work is the first one to approximate the 3D of a scene based on planar proxies that are estimated from a dense 3D point cloud in a way that explicitly balances the approximation error and the number of planar models covering the scene. Our validation demonstrates that it results in an accurate, low complexity 3D representation of the scene, perfectly adapted for light-weighted storage and transmission.

2. Related works

Many previous works have considered images to reconstruct the 3D of a scene. Their findings and observations have largely inspired and motivated our approach.

The most mature approaches are the ones estimating the 3D of a from small-baseline stereo. They have been extensively evaluated through the Middlebury challenge. Several of the top-ranked algorithms [25] [26] rely on image segmentation. Working at the region level has been proven to increase the robustness of the matching data-fidelity [18] [21] while effectively propagating depth information from textured to ambiguous regions [44]. We have thus adopted a region-based paradigm in our wide-baseline setup as well.

In contrast to small-baseline stereo, the reconstruction from wide-baseline images offers the advantage to generate 3D models that are consistent with a large range of view angles. It also benefits from more accurate triangulation, but suffers from severe occlusions , photometric and geometric deformations between the views. Therefore, the related previous art generally require many ($\gg$ 2) images to either derive a few reliable correspondences [38], or to fuse multiple depth-maps together [14] [43]. Moreover, most of those methods disambiguates the matching based on a strong regularization, which tends to over-smooth the depth [32] [31] [2], or even to propagate it to wrong pixels when the image gradient is not sufficient at the 3D structure's border [6]. As an alternative to depth-maps fusion, plane-sweeping methods investigate multiple depth hypotheses by sweeping a plane [7] through the 3D space, either orthogonally to one of the camera's axis [3] [16] or along a few principal directions [12]. Although their GPU-based implementations achieve real-time performances [42] [24] [15], plane-sweeping assumes the Manhattan world hypothesis, *i.e.*, that the 3D surfaces are orthogonal to the sweeping directions.

To avoid multiplying the number of views or raising the Manhattan world assumption, many authors have proposed to constrain the 3D reconstruction based on geometric primitives. Typically, they first estimate a sparse (and hopefully less noisy) 3D points cloud from the matching of salient points in image pairs [1] [29] [30] [28], and then fit 3D primitives to those points. The fitting can be direct, *e.g.* based on a RANSAC-based approach(es) [10] [45], or indirect, *e.g.* based on the detection of line segments or vanishing directions [41] [37]. Those methods only achieve good reconstruction when either multiple small-baseline input views are available [41], or manual interactions are tolerated to specify high-level scene informations [20] [23] (*e.g.* ajacency, alignment, regularity, etc.). To alleviate those drawbacks, the so-called "propose-and-assign" approaches have been considered. Instead of directly fitting primitives to the data, they first derive a number of 3D primitive candidates, which are then assigned to the parts of the sparse 3D point cloud they best approximate. In this formulation, the assignment is handled globally over the whole scene, through an energy-minimization process. Under the piecewise-planarity assumption, the primitives correspond to 3D planes [27], and a Markov-Random-Field (MRF) formulation is considered to propagate the assignment to the pixels that are not represented in the sparse point cloud. For increased robustness, Bodis *et al.* [4] have recently proposed to lift-up the regularized assignment at the region level. In their approach, a plane candidate is assigned to each region,

and the number of proposed models is reduced *a posteriori* by merging the most similar ones. Their remarkable method strongly accelerates the reconstruction, from many minutes to a few seconds, due to the small amount of treated regions and their abstinence from using any expensive photoconsistency computation [22]. In practice, assigning planes to regions rather than to pixels however suffers from a main drawback: regions that are not represented in the sparse point cloud, and that do not have a MRF neighbor with similar planar structure, can not be modeled properly. This happens frequently in large and uniform regions presenting repetitive and non-discriminant patterns, like grass/floor planes. More generally, defining regions is an issue for methods that build on a sparse point cloud: too large regions violate the region planarity assumption [44], while too small regions might not have associated 3D points, meaning that their 3D can not be inferred. Our work overcomes this issue by adopting a dense point cloud as input 3D cues. A few previous works have also adopted a piecewise-planar assumption to fit multiple planes to a dense cloud. They however generally need a reliable dense 3D point cloud, which in turns requires many views: the impressive work in [11] and in [1] build respectively on 3 million and a few hundred thousands images. Far from those huge number of views, [13] uses the depths obtained from ten images (spread on approximatively 5 meters) to fit planar hypotheses on segmented regions, but relies on application-dependent priors, embedded in classifiers that are trained from manually labeled data. As illustrated in Section 1 of the supplementary material, estimating the 3D planes independently on each region without those application-dependent priors appears to be too sensitive to the strong noise inherent to a dense point cloud derived from a few wide-baseline pairs.

Our region-based plane assignment method offers thus a unique asset in that it requires only two wide-baseline views to determine an accurate piecewise-planar approximation of man-made scenes. It relies on dense point cloud estimation to properly deal with surfaces containing few discriminant salient points, but introduces an original data-fidelity metric and considers a multi-model fitting method to deal with the strong noise inherent to the dense nature of the cloud. It does not assume dominant directions, like in a Manhattan world hypothesis and does not require user interactions.

3. Planar models proposition

The 3D planes hypotheses to be considered during the plane-to-region assignment process (Section 5) are derived from a dense cloud of 3D points through a 3-steps procedure.

In the first step, a dense 3D point cloud is generated by determining, for each pixel $\mathbf{x}$ belonging to the first view $\mathcal{I}$, the corresponding pixel $\mathbf{x}'$ in the second view $\mathcal{I}'$, and triangulating [17] these correspondences. A correspondence $\mathbf{x}'$ is determined for each $\mathbf{x} \in \mathbf{\Omega}_{\mathcal{I}}$ (where $\mathbf{\Omega}_{\mathcal{I}}$ is the spatial domain of the image $\mathcal{I}$) based on a simple "Winner-Takes-All" (WTA) [26] method, restricted to the epipolar line $\mathbf{l}' = \mathbf{F} \cdot \widetilde{\mathbf{x}}$ associated to $\mathbf{x}$:

$$\mathbf{x}' = \underset{\mathbf{y}' \in \mathbf{F} \cdot \widetilde{\mathbf{x}}}{\operatorname{argmin}} \left\| \mathbf{d}\left(\mathbf{x}\right) - \mathbf{d}\left(\mathbf{y}'\right) \right\|_2^2, \tag{1}$$

where $\mathbf{F}$ is the fundamental matrix of the calibrated stereo pair, $\widetilde{\mathbf{x}}$ are the homogeneous coordinates [17] of $\mathbf{x}$, $\mathbf{d}\left(\mathbf{x}\right)$ is a descriptor associated to this pixel, and $\|\cdot\|_2$ is the ℓ_2 norm. In our validations, the Daisy descriptors [35] have been chosen for their robustness against wide-baseline geometric distortions, and their appropriateness for dense estimation [34].

In the second step, we derive M planar models from this noisy 3D point cloud. Therefore, we randomly (uniformly) select M triplets of (non-colinear) 3D points $\mathbf{X}_t$ (with $t = \{1,2,3\}$ defining the index of the 3D point in the Triplet) to generate M plane candidates $\boldsymbol{\pi}_m$ (with $1 \leq m \leq M$), each one parametrized as $\boldsymbol{\pi}_m = \begin{bmatrix} a_m & b_m & c_m & d_m \end{bmatrix}^\top$ to represent the plane $a_m x + b_m y + c_m z + d_m = 0$, or equivalently by $\boldsymbol{\pi}_m = \begin{bmatrix} a_m/d_m & b_m/d_m & c_m/d_m & 1 \end{bmatrix}^\top \triangleq \begin{bmatrix} \boldsymbol{\eta}_m^\top & 1 \end{bmatrix}^\top$.

In the last step, we derive from the M plane candidates, a small number of $K \ll M$ planes that are expected to capture most of the representative planar structures in the scene. Therefore, we first assign a quality value $q\left(\boldsymbol{\pi}_m\right)$ to each of the M plane candidates. This is done by considering the triangular patch $[\boldsymbol{\pi}_m]$ lying on the plane $\boldsymbol{\pi}_m$ and delimited by the triplet $\{\mathbf{X}_t\}_{t=\{1,2,3\}}$.

The 2D region representing this triangular patch $[\boldsymbol{\pi}_m]$ in the first (respectively second) reference view is denoted $\mathbf{\Delta}_m$ (respectively $\mathbf{\Delta}'_m$), and is defined by:

$$\mathbf{\Delta}_m = \left\{ \mathbf{x} \in \mathbf{\Omega}_{\mathcal{I}} \;\middle|\; \mathbf{x} \in \mathbf{P} \cdot [\boldsymbol{\pi}_m] \right\}$$
$$\mathbf{\Delta}'_m = \left\{ \mathbf{x}' \in \mathbf{\Omega}_{\mathcal{I}'} \;\middle|\; \mathbf{x}' \in \mathbf{P}' \cdot [\boldsymbol{\pi}_m] \right\},$$

with $\mathbf{P} \in \mathbb{R}^{3\times4}$ (respectively $\mathbf{P}' \in \mathbb{R}^{3\times4}$) the projection matrix of the first (respectively second) reference view.

We then extract, from the point cloud, the set of 3D points $\mathbf{X}$ projecting in $\mathbf{\Delta}_m$ or in $\mathbf{\Delta}'_m$. For the sake of simplicity, we slightly abuse the notation in the rest of the paper and write $\mathbf{X} \in \mathbf{\Delta}_m$ when the projection $\mathbf{P} \cdot \widetilde{\mathbf{X}}$ of the 3D point $\mathbf{X}$ falls into the 2D triangle $\mathbf{\Delta}_m$. We write analogously $\mathbf{X} \in \mathbf{\Delta}'_m$.

Given those definitions, the proposed quality value $q\left(\boldsymbol{\pi}_m\right)$ quantifies how close is the plane candidate $\boldsymbol{\pi}_m$

from the 3D points $\mathbf{X}_j \in \{\mathbf{\Delta}_m \cup \mathbf{\Delta}'_m\}$, with $j \leq \mathcal{J}$, $\mathcal{J}$ being the number of 3D points projecting onto $\mathbf{\Delta}_m$ or $\mathbf{\Delta}'_m$. This is done by counting the fraction of 3D points $\mathbf{X}_j \in \{\mathbf{\Delta}_m \cup \mathbf{\Delta}'_m\}$ that are closer from the 3D plane $\boldsymbol{\pi}_m$ than a predefined threshold $T_d \in \mathbb{R}^+$:

$$q\left(\boldsymbol{\pi}_m\right) = \frac{1}{\mathcal{J}} \sum_{j=1}^{\mathcal{J}} \left(d(\boldsymbol{\pi}_m, \mathbf{X}_j) \leq T_d\right),$$

in which

$$d\left(\boldsymbol{\pi}_m, \mathbf{X}_j\right) = \frac{|\boldsymbol{\pi}_m^\top \cdot \widetilde{\mathbf{X}}_j|}{\|\boldsymbol{\eta}_m\|_2}$$

is the orthogonal distance between a 3D plane $\boldsymbol{\pi}_m$ and the 3D point $\mathbf{X}_j$.

The relevance of the quality value $q\left(\boldsymbol{\pi}_m\right)$ is assessed, in Section 2 of the supplementary material, by showing that the distributions of this metric largely differs for ground-truth and random planes.

Based on this plane quality value $q\left(\boldsymbol{\pi}_m\right)$, we select, from the M plane candidates $\boldsymbol{\pi}_m$, the $K \ll M$ most representative ones by applying a weighted k-means [8] on the $\boldsymbol{\eta}_m \in \mathbb{R}^3$ vectors. The weight associated to the plane candidate $\boldsymbol{\pi}_m$ in the weighted k-means is chosen to be its quality value $q\left(\boldsymbol{\pi}_m\right)$.

In summary, although we initially generate a tremendeous amount of M plane candidates to guarantee that this random selection includes the 3D ground-truth, our plane-to-region assignment method avoids to compute $M \cdot N$ plane/region association metrics (N being the number of regions), and reduces it to $K \cdot N$, with $K \ll M$.

4. Cost of assigning a 3D plane to a 2D region

This section proposes a novel data-fidelity metric to quantify how well a given 3D plane $\boldsymbol{\pi}$ approximates the 3D surface associated to a region $\mathcal{R}$ in image $\mathcal{I}$. Fundamentally, our data fidelity measures the proximity between the investigated (plane) model $\boldsymbol{\pi}$ and the 3D points that project into the 2D region $\mathcal{R}$ or its counterpart $\mathcal{R}_{\boldsymbol{\pi}}$, obtained in $\mathcal{I}'$ using the homography $\mathbf{H}_{\boldsymbol{\pi}}$ induced by the 3D plane $\boldsymbol{\pi} = [a \;\; b \;\; c \;\; d]^\top$ [17], *i.e.*, $\mathcal{R}_{\boldsymbol{\pi}} = \left\{\mathbf{H}_{\boldsymbol{\pi}} \cdot \widetilde{\mathbf{x}}_j : \mathbf{x}_j \in \mathcal{R}\right\}$. To modulate our data-fidelity metric according to the discriminativeness of the textures observed in the 2D views, we propose to account for the inaccuracy and the ambiguity of the 2D descriptors associations that support the 3D points definition.

Indeed, a matching between a pair of 2D points $\mathbf{x} \in \mathbf{\Omega}_{\mathcal{I}}$ and $\mathbf{x}' \in \mathbf{\Omega}_{\mathcal{I}'}$ is expected to be reliable when the 2D point descriptors $\mathbf{d}\left(\mathbf{x}\right)$ and $\mathbf{d}\left(\mathbf{x}'\right)$ are (1) very similar, and (2) quite discriminant, which means they are different from most of the alternative matches along the epipolar line, *i.e.*, $\mathbf{d}\left(\mathbf{x}\right)$ different from $\mathbf{d}\left(\mathbf{y}'\right)$ with $\mathbf{y}' \in \mathbf{F} \cdot \widetilde{\mathbf{x}}$ (see Equation (1)). For a 3D point $\mathbf{X}$ associated to the triangulation of two matched pixels $\mathbf{x}$ and $\mathbf{x}'$, we introduce:

- the matching inaccuracy, denoted by $m_i\left(\mathbf{X}\right)$, to measure how dissimilar are the descriptors $\mathbf{d}\left(\mathbf{x}\right)$ and $\mathbf{d}\left(\mathbf{x}'\right)$ of the two corresponding 2D points $\mathbf{x} \leftrightarrow \mathbf{x}'$ associated to $\mathbf{X}$. We define it by:

  $$m_i\left(\mathbf{X}\right) = \frac{1}{\mathcal{D}} \left\|\mathbf{d}\left(\mathbf{P} \cdot \widetilde{\mathbf{X}}\right) - \mathbf{d}\left(\mathbf{P}' \cdot \widetilde{\mathbf{X}}\right)\right\|_2,$$

  where $\mathcal{D}$ is the size of the descriptor used during the matching phase.

- the matching ambiguity, denoted by $m_a\left(\mathbf{X}\right)$, to measure the percentage of pixel candidates $\mathbf{y}' \in \mathbf{F} \cdot \widetilde{\mathbf{x}}$ satisfying $\frac{1}{\mathcal{D}} \left\|\mathbf{d}\left(\mathbf{x}\right) - \mathbf{d}\left(\mathbf{y}'\right)\right\|_2 \leq \frac{m}{\mathcal{D}} \cdot \left\|\mathbf{d}\left(\mathbf{x}\right) - \mathbf{d}\left(\mathbf{x}'\right)\right\|_2 + b$, among the pixels $\mathbf{y}'$ lying on the epipolar line associated to $\mathbf{x}$. In this definition, m and b are respectively set to 1.5 and 0.002. Our experiments have revealed that these parameters do not strongly affect the performance of our method.

To evaluate the relevance of those metrics, Figure 1 plots their distributions for two classes of 3D points that project in a region for which a planar ground-truth plane $\boldsymbol{\pi}^\star$ model (notated GT model) has been manually defined: (1) the green plot considers the "inliers" to the manual ground-truth plane $\boldsymbol{\pi}^\star$ associated to the region (*i.e.*, the $\mathbf{X}$ satisfying $d\left(\boldsymbol{\pi}^\star, \mathbf{X}\right) \leq 0.1$ [m]), while (2) the red plot refers to the outliers (with distance $d\left(\boldsymbol{\pi}^\star, \mathbf{X}\right) > 1$ [m]) compared to this ground-truth plane.

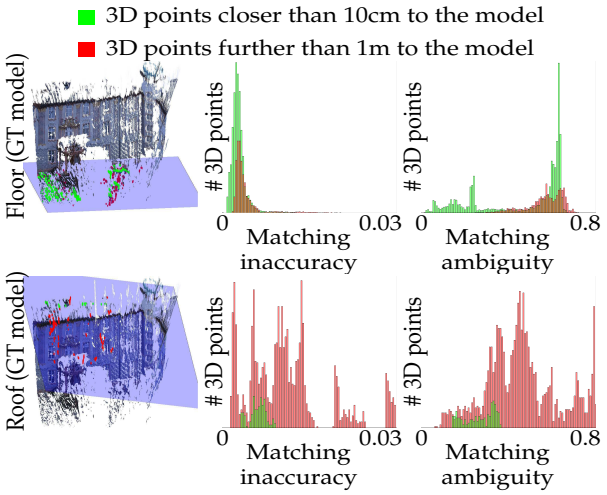


Figure 1: Distribution of the inaccuracy and ambiguity of the 3D points associated to two ground-truth 3D planar regions. The floor is textureless, while the roof is only composed of repetitive textures.

Figure 1 reveals that, whilst being different, the inliers and outliers distributions largely overlap each others. This prevents the accurate classification of the 3D points into an inlier and outlier class based on those two metrics.

Since it is not possible to identify the inliers directly from the 3D points inaccuracy and ambiguity, we have adopted an indirect statistical approach to estimate whether a 3D plane correctly fits the 3D point cloud associated to an image region. In short, we analyze whether the points that are sufficiently (as defined below) accurate and unambiguous lie close to the plane model. Formally, let $\mathcal{C}^{\boldsymbol{\tau}}_{\mathcal{R},\boldsymbol{\pi}}$ denote the set of 3D points $\mathbf{X}$ satisfying the three following criteria:

$$\begin{cases} \mathbf{X} \in \{\mathcal{R} \cup \mathcal{R}_{\boldsymbol{\pi}}\} \\ m_i(\mathbf{X}) \leq \tau_i \\ m_a(\mathbf{X}) \leq \tau_a \end{cases}$$

where $\boldsymbol{\tau} = \{\tau_i, \tau_a\}$ and $\tau_i \in \mathbb{R}^+$ and $\tau_a \in \mathbb{R}^+$ are thresholds on the matching inaccuracy and ambiguity. As in Section 3, we abuse the notation, and write $\mathbf{X} \in \{\mathcal{R} \cup \mathcal{R}_{\boldsymbol{\pi}}\}$ to indicate that the 3D point $\mathbf{X}$ projects onto the 2D region $\mathcal{R}$, or its counterpart $\mathcal{R}_{\boldsymbol{\pi}}$ in $\mathcal{I}'$.

Given a pair $\boldsymbol{\tau} = \{\tau_i, \tau_a\}$, we analyze how the 3D points in $\mathcal{C}^{\boldsymbol{\tau}}_{\mathcal{R},\boldsymbol{\pi}}$ scatter away from the investigated plane $\boldsymbol{\pi}$, by introducing the scattering function $f_{\mathcal{C}^{\boldsymbol{\tau}}_{\mathcal{R},\boldsymbol{\pi}}}(l, \boldsymbol{\pi})$ to define the fraction of 3D points in $\mathcal{C}^{\boldsymbol{\tau}}_{\mathcal{R},\boldsymbol{\pi}}$ whose distance to $\boldsymbol{\pi}$ is smaller than $l \in \mathbb{R}^+$, given a pair $\boldsymbol{\tau} = \{\tau_i, \tau_a\}$.

Examples of scattering functions are presented in Section 4 of the supplementary material. They indicate that the area under curve (AuC) of the scattering function is a good indicator of plane model relevance. We introduce $A(\boldsymbol{\tau}, \boldsymbol{\pi}) \in [0; 1]$ to denote the area under curve of the scattering function $f_{\mathcal{C}^{\boldsymbol{\tau}}_{\mathcal{R},\boldsymbol{\pi}}}(l, \boldsymbol{\pi})$:

$$A(\boldsymbol{\tau}, \boldsymbol{\pi}) = \int_0^{l_{\text{lim}}} f_{\mathcal{C}^{\boldsymbol{\tau}}_{\mathcal{R},\boldsymbol{\pi}}}(l, \boldsymbol{\pi}) \; dl.$$

Roughly speaking, this area reflects the likelihood that the 3D points $\mathbf{X} \in \mathcal{C}^{\boldsymbol{\tau}}_{\mathcal{R},\boldsymbol{\pi}}$, spread on the interval $[0; l_{\text{lim}}]$ around the investigated plane $\boldsymbol{\pi}$, are "close" from this plane. The choice of $\{\tau_a, \tau_i\}$ is however important. It should keep the subset $\mathcal{C}^{\boldsymbol{\tau}}_{\mathcal{R},\boldsymbol{\pi}}$ sufficiently large, while making sure that the most reliable points have the largest impact. To avoid the tricky/delicate tuning of the parameters $\{\tau_a, \tau_i\}$, we consider the scattering function for several subsets of 3D points, each subset corresponding to an increasing level of accuracy/unambiguity. The AuC are then merged based on a geometric mean, to decide whether a plane model is valid or not. Formally, the data-fidelity $c(\mathcal{R}, \boldsymbol{\pi}) \in [0; 1]$ of assigning a plane $\boldsymbol{\pi}$ to a region $\mathcal{R}$ is thus defined, based on the geometric mean of a sequence of T tests $\boldsymbol{\tau}^{(t)} = \{\tau_i^{(t)}, \tau_a^{(t)}\}$ on the accuracy/unambiguity of the 3D points $\mathbf{X} \in \{\mathcal{R}, \mathcal{R}_{\boldsymbol{\pi}}\}$. In practice, the sequence of tests is defined as:

$$\boldsymbol{\tau}^{(t)} = \boldsymbol{\tau}^{(1)} - \frac{t-1}{T-1} \cdot \left(\boldsymbol{\tau}^{(1)} - \boldsymbol{\tau}^{(T)}\right) \quad \forall t \in \{1, \cdots, T\},$$

with $\boldsymbol{\tau}^{(1)}$ (respectively $\boldsymbol{\tau}^{(T)}$) the set of maximum (respectively minimum) investigated thresholds and the data-fidelity is measured as:

$$c(\mathcal{R}, \boldsymbol{\pi}) = \begin{cases} 0 & \text{if } \mathcal{R}_{\boldsymbol{\pi}} \cap \boldsymbol{\Omega}_{\mathcal{I}'} = \emptyset \\ \exp\left(\frac{1}{T}\sum_{t=1}^{T} \log\left(A\left(\boldsymbol{\tau}^{(t)}, \boldsymbol{\pi}\right)\right)\right) & \text{otherwise.} \end{cases}$$

Finally, it is worth noting that we do not consider, in the computation of the data-fidelity, the area $A\left(\boldsymbol{\tau}^{(t)}, \boldsymbol{\pi}\right)$ which are computed on less than a certain number, set to 10 in practice, of 3D points $\mathbf{X} \in \mathcal{C}^{\boldsymbol{\tau}^{(t)}}_{\mathcal{R},\boldsymbol{\pi}}$.

Computing a geometric mean and ignoring too small subsets makes the decision relatively independent from the actual subsets definition, as long as a sufficiently fine sequence of $\{\tau_a, \tau_i\}$ thresholds is considered. Ignoring small subsets avoids that statistically non-representative subsets impact the mean. Using a fine sequence of thresholds ensures that accurate and unambiguous points largely impact the mean, since they are part of much more subsets than unreliable points.

5. Sparse piecewise-planar approximation

The assignment of a planar model to each of the N regions is formulated as a multi-model fitting problem, using the state-of-the-art *Propose, Expand and Re-Learn* (PEARL) algorithm [19]. As its name indicates, the PEARL inference optimization is composed of three steps: the proposition of a set of models ("propose stage"), the label inference ("expand stage") and the models reestimation ("re-learn stage"). We now explain how the PEARL framework is adapted to fit our problem.

In the "propose" stage, while the original paper requires to generate several thousands of models candidates, we rely on Section 3 to limit ourselves to a few hundreds (only 200 candidates are used in our validations). This enables us to strongly accelerate the optimization, while keeping the same accuracy (as shown in Section 5 of the supplementary material).

In the "expand" stage, one planar model is assigned to each image region. The inference problem is expressed as an energy-driven minimization [19] (solved by α-expansion [5]). In our case, it minimizes:

$$E(\mathbf{L}) = \sum_{n=1}^{N} (1 - c(\mathcal{R}_n, \boldsymbol{\pi}(L_n))) + \lambda \sum_{(p,q)\in\mathcal{N}} \omega_{pq}\, \delta(L_p \neq L_q) + \beta |\mathcal{L}_{\mathbf{L}}|$$

where $\mathbf{L} = \{L_1, L_2, \cdots, L_N\}$ are the labels assigned to the N regions. Each label L_n refers either to one of the K models, or to an occlusion label $L_\emptyset$, which allows to explicitly model the occluded regions. $c(\mathcal{R}_n, \boldsymbol{\pi}(L_n)) \in [0;1]$ is the cost of assigning the $\boldsymbol{\pi}(L_n)$ model to the n^{th} region (see Section 4), $\delta(.)$ is the indicator function, $|\mathcal{L}_{\mathbf{L}}|$ is the number of assigned models[4] and ω_{pq} is a weight associated to a pair of neighboring regions that encourages spatial coherence, defined as:

$$\omega_{pq} = \begin{cases} 1 - \mathbb{E}\left[|\nabla \mathcal{I}(\mathbf{x})|\right]_{\mathbf{x}\in\mathcal{B}} & \text{if } \mathcal{R}_p \text{ and } \mathcal{R}_q \text{ have a common border } \mathcal{B} \\ 0 & \text{otherwise} \end{cases}$$

where the gradient amplitude $|\nabla \mathcal{I}(\mathbf{x})|$ is rescaled to $[0;1]$, by applying contrast stretching over the entire gradient image, and $\mathbb{E}[.]$ represents the mean operator.

Eventually, in the "re-learn" stage, PEARL extracts, for each assigned label $L_n \neq L_\emptyset$, the set $\mathcal{P}_{L_n}$ of region assigned to this label, and reestimates the associated model. This reestimation is done by selecting the set of 3D points that project into one of the regions of $\mathcal{P}_{L_n}$ and applying RANSAC [10] (with inlier threshold τ) to robustly fit a new plane model to these 3D points, based on the inlier score proposed in [4].

The PEARL algorithm iterates sequentially between the three stages, until $E(\mathbf{L})$ reaches a minimum. In contrast to [19], our implementation also iterates over the regions that are initially assigned to the $L_\emptyset$ label.

6. Experiments

This section considers various man-made scenes, and demonstrates that our method is able to locate their main 3D planes, as well as to detect their occluded regions. The accuracy of the 3D model is then validated by generating free-viewpoints around the piecewise-planar reconstructed scenes. To complement those results, Section 5 of the supplementary material presents additional results showing that our plane proposition phase is effective at generating a small set of 3D plane hypotheses that includes the 3D ground-truth of the scene.

[4]This term encourages parsimony, to describe the scene with as few plane models as possible.

We consider 10 well-known and calibrated sequences representing street-level captures of (man-made) building scenes (indoor and outdoor). While these datasets provide multiple different views of each scene, we have arbitrarily selected two distant views among the available ones to define a set of wide-baseline stereo pairs.

To segment the left view, we rely on [40] [39] to learn the dominant colors in the image[5]. Given this set of C dominant colors, the segmentation problem is defined as the assignment of each pixel to one of the C classes. To impose the smoothness among neighboring pixels, this assignment problem is solved by graph-cut optimization [9], in which the data-fidelity term is defined as the ℓ_2 distance between the C dominant colors and the pixel color, and the smoothness term is proportional to the inverse of the amplitude of the gradient of two neighboring pixels. This method results into a set of N regions.

Our method depends only on two types of parameters. First, the RANSAC inliers/outliers parameter τ and the parameter l_{lim} (representing the investigated orthogonal distance around the proposed 3D plane) are fixed, based on rough prior human knowledge about the depth variability in the scene. In all our experiments, l_{lim} has been chosen between 30cm and 1.5m, while τ has been fixed to $\tau = l_{\text{lim}}/5$. Second, for the parameters of the PEARL optimization, we have set the pairwise term to $\lambda = 0.1$ and the occlusion data-fidelity to $c(\mathcal{R}_n, \boldsymbol{\pi}(L_\emptyset)) = 0.5$ in all our experiments. This last parameter is a good trade-off between accepting plane assumptions on regions associated to noisy 3D points and discarding bad planes. The labeling weight β is chosen between $[0.1; 0.5]$ to lead to a visual reasonable trade-off between number of planes and accuracy of representation.

Figure 2 illustrates the results of the different steps of our algorithm, as well as the projection of the first view onto the second one via the piecewise-planar approximated model. Occlusions are highlighted in black. From top to down, the used datasets are: CastleP19/FountainP11 [33] and Model-house/Wadham/MertonIII [41]. Similar validations on other well-known wide-baseline datasets, such as HerzJesuP25/Oxford Corridor/Library/MertonI and MertonII, are presented in the supplementary material (Section 6).

[5]The required color dissimilarity threshold for learning the dominant colors consituting the image has been set to 20 in all the experiments. This parameter influences the number of obtained regions. Our experiments have revealed that it does not strongly affect the performance of our piecewise-planar 3D approximation, as long as the image is over-segmented.

First, we note that most of the 3D planes are correctly estimated (columns (d) and (e) in Figure 2) , despite the presence of noise in the dense 3D point cloud. This performance is due to the high robustness of the proposed data-fidelity metric $c(\mathcal{R}, \boldsymbol{\pi})$, which simultaneously considers the 3D points of $\mathcal{R}$ and $\mathcal{R}_{\boldsymbol{\pi}}$.

Second, the failure cases of our approach are not frequent, and can be divided into three classes: wrong 3D plane model assignment, assignment of visible regions to the occlusion label, and wrong 3D plane estimation. The first class of errors appears either when the 3D points projecting in $\mathcal{R}$ and $\mathcal{R}_{\boldsymbol{\pi}}$ are strongly contaminated by 3D outliers, and the high value of β pushes towards the propagation of a wrong model from an adjacent region (*e.g.*, in the sky regions in the 4^{th} row of Figure 2), or when the initial image segmentation defines regions that cover two distinct planar models (*e.g.*, on the gutter at the middle of the left wall of Merton II, presented in the 5^{th} row of Figure 7 in the supplementary material). The second class of errors (assignment of a visible region to the occlusion label) appears in the absence of 3D points in the interval $[0; l_{\text{lim}}]$ around the 3D ground-truth. This behavior can be observed on the tower of the Library dataset (4^{th} row of Figure 2, column (d)). The third class of errors (wrong model estimation) can affect either the large and challenging 2D regions, or the smallest ones. In the first case, the inaccuracy of the plane estimation originates from the planar re-estimation of PEARL, using RANSAC on a region that is contaminated by more than 50% of 3D "outliers". This behavior can be observed on the Oxford corridor image (supplementary material, fourth row, third column), or on the terrace of the Model-house sequence (supplementary material, Figure 7, second row, third column). The second case appears in very small regions for which the spatial concentration of their associated 3D points makes the (RANSAC-based) fitted plane more prone to errors than if the 3D points were spatially spread (large region). This problem affects the roof of the windows of the Merton I dataset, as attested by the fact that those regions are projected on the grass in the other view, as illustrated in the fourth row in Figure 7 of the supplementary material, column (e).

To complete our visual experimental results, Figure 3 (as well as Figure 8 and many videos, in the supplementary material) demonstrates the effectiveness of our dense, piecewise-planar 3D approximation method, by projecting the textured 3D piecewise-planar models on virtual intermediate views.

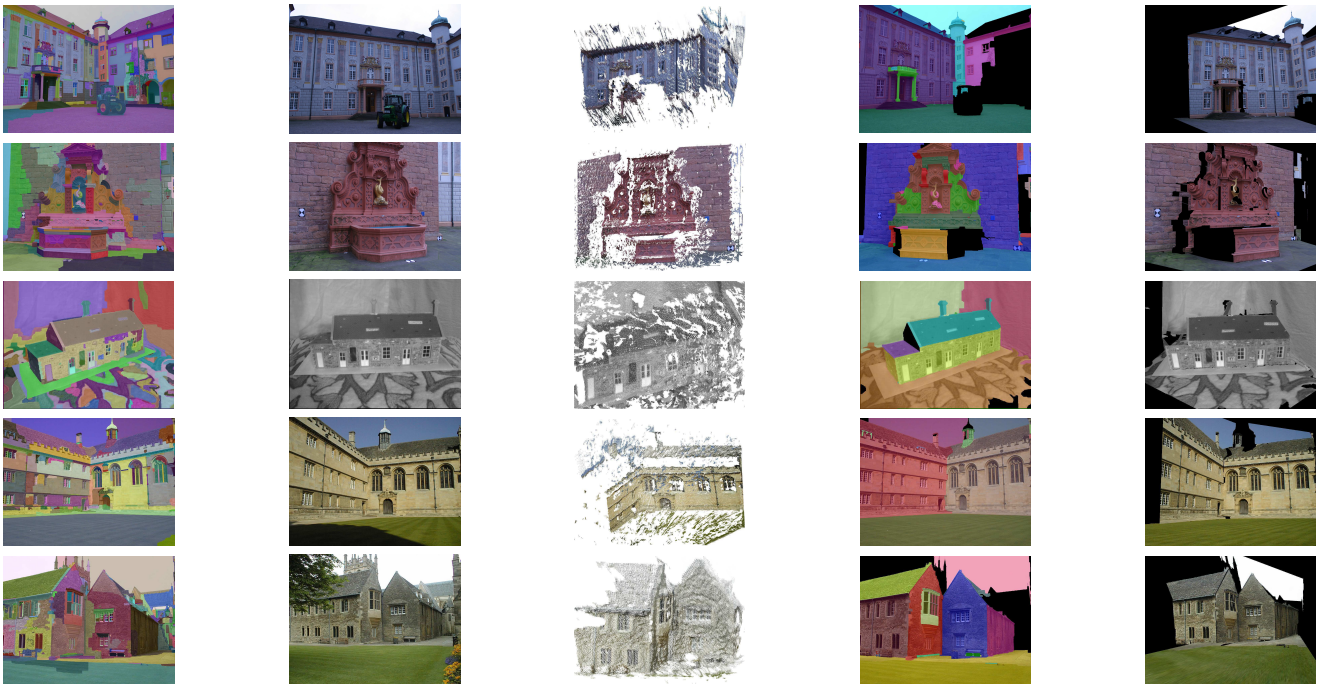
Regarding complexity, our method reconstructs each 3D scenes in a few minutes (from 4 to approximatively 20 minutes, according to the resolution of the reference images, Matlab implementation on a 2.4GHz Intel I5 CPU, 8Gb RAM machine), which are divided into three parts: approximatively 60% of the running time is dedicated to the dense point cloud generation (using currently a non-parallel implementation of the WTA method), 25% on the plane proposition phase (in which the location of the pixels in $\boldsymbol{\Delta}$ and $\boldsymbol{\Delta}'$ takes the most of the running time), and 15% for the rest.

7. Conclusion

We express the 3D reconstruction as a generalized plane assignment problem over 2D image regions, in which the occluded regions are explicitly modeled. We rely on a dense, and thus inherently highly corrupted, 3D point cloud to allow the approximation of challenging (*e.g.*, textureless or repetitively patterned) 2D regions, *e.g.*, grass floors. Therefore, we adopt a multi-model fitting framework. It relies on a limited number (*e.g.*, $\approx$ 200) of candidate plane models, and formulates the plane assignment problem as an energy-driven formulation, which simultaneously optimizes a data-fidelity term, the smoothness of the plane assignment over the regions and the number of used models. Our main contributions have to do with the computation of a small set of relevant candidate models, and the derivation of a data-fidelity metric that measures the fitting error while considering the inaccuracy and the ambiguity associated to the 2D matches used to defined the 3D points . Also, to the best of our knowledge, by simultaneously optimizing the data-fidelity, the smoothness and the number of assigned models, our light-weight method is the first one to densely approximate a 3D scene while simultaneously targeting a minimal number of models. We have demonstrated the accuracy of the approximated 3D models by interpolating virtual views around a variety of man-made scene, on which traditional MVS methods fail [4].

Acknowledgments

This work has been funded by the Belgian NSF (under F.N.R.S and F.R.I.A grants) and by the Walloon Region projects AOC and PTZ-PILOT.



(a) (b) (c) (d) (e)

Figure 2: (Best viewed in color). Based on the segmentation (a) of one of the two wide-baseline views ((a) and (b)) and on their associated dense point cloud (c), our method approximates the 3D surface by the minimum set of of 3D planes. In (d), regions assigned to the same 3D plane are illustrated with a same color. The reprojection of the optimal piecewise-planar reconstruction, textured based on the first view (a) and projected in the second view (b), is represented in (e).

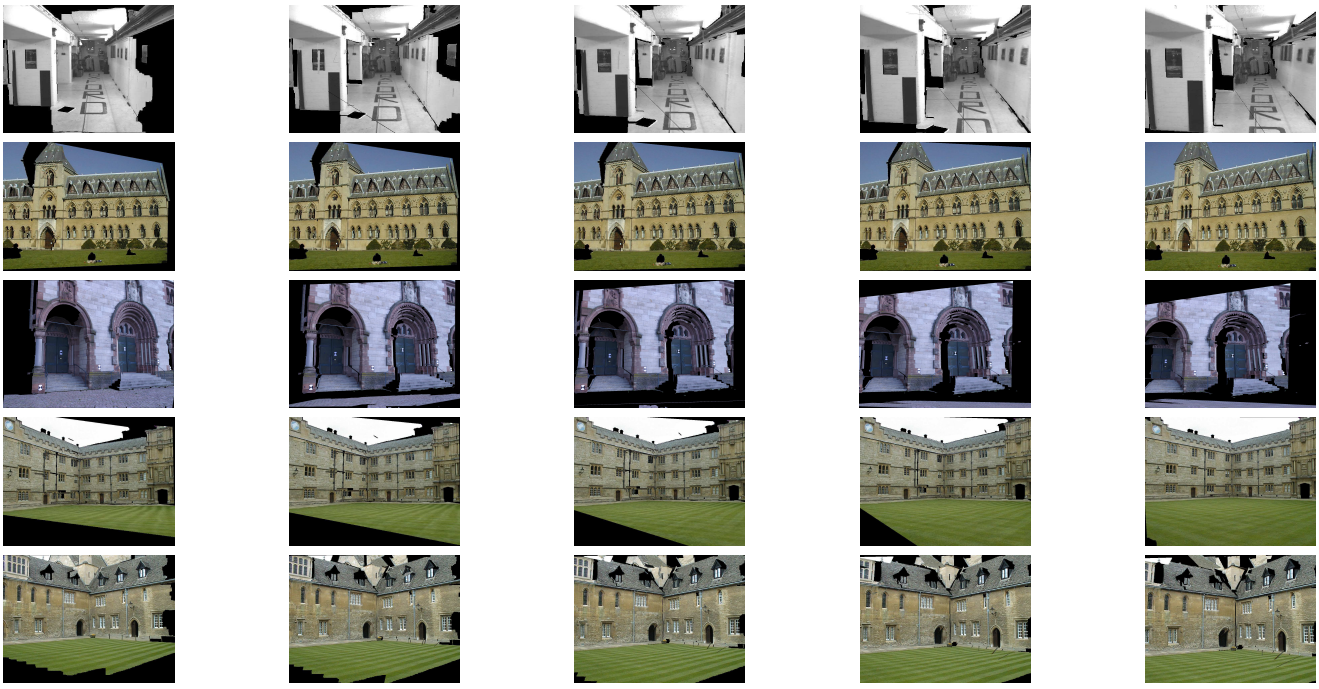


Figure 3: Projection of the textured piecewise-planar approximation of the scene's 3D on virtual views in-between the two cameras of the wide-baseline stereo pair.

References

[1] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski. Building Rome in a day. In *International Conference on Computer Vision (ICCV)*, volume 1, pages 105–112. IEEE, 2011. 1, 2, 3

[2] L. Bagnato, P. Frossard, and P. Vandergheynst. Optical flow and depth from motion for omnidirectional images using a tv-l1 variational framework on graphs. In *International Conference on Image Processing (ICIP)*, pages 1469–1472. IEEE, 2009. 2

[3] L. Ballan, G. J. Brostow, J. Puwein, and M. Pollefeys. Unstructured video-based rendering: Interactive exploration of casually captured videos. *Transactions on Graphics (TOG)*, 29(4):87, 2010. 2

[4] A. Bodis-Szomoru, H. Riemenschneider, and L. Van Gool. Fast, approximate piecewise-planar modeling based on sparse structure-from-motion and superpixels. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 469–476. IEEE, 2014. 1, 2, 6, 7

[5] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(11):1222–1239, 2001. 5

[6] J. Braux-Zin, R. Dupont, and A. Bartoli. A general dense image matching framework combining direct and feature-based costs. In *International Conference on Computer Vision (ICCV)*, volume 1, pages 185–192. IEEE, 2013. 2

[7] R. T. Collins. A space-sweep approach to true multi-image matching. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 358–363. IEEE, 1996. 2

[8] R. C. De Amorim and B. Mirkin. Minkowski metric, feature weighting and anomalous cluster initializing in K-means clustering. *Pattern Recognition*, 45(3):1061–1075, 2013. 4

[9] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov. Fast approximate energy minimization with label costs. *International Journal of Computer Vision (IJCV)*, 96(1):1–27, 2012. 6

[10] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2, 6, 11

[11] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, et al. Building rome on a cloudless day. In *European Conference on Computer Vision (ECCV)*, pages 368–381. Springer, 2010. 3

[12] D. Gallup, J.-M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 1–8. IEEE, 2007. 2

[13] D. Gallup, J.-M. Frahm, and M. Pollefeys. Piecewise planar and non-planar stereo for urban scene reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 1418–1425. IEEE, 2010. 3

[14] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz. Multi-view stereo for community photo collections. In *International Conference on Computer Vision (ICCV)*, volume 1, pages 1–8. IEEE, 2007. 2

[15] P. Goorts, C. Ancuti, M. Dumont, S. Rogmans, and P. Bekaert. Real-time video-based view interpolation of soccer events using depth-selective plane sweeping. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, 2013. 2

[16] P. Goorts, M. Dumont, S. Rogmans, and P. Bekaert. An end-to-end system for free viewpoint video for smooth camera transitions. In *International Conference on 3D Imaging*, pages 1–7. IEEE, 2012. 2

[17] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2003. 1, 3, 4

[18] H. Hirschmüller and D. Scharstein. Evaluation of cost functions for stereo matching. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 1–8. IEEE, 2007. 2

[19] H. Isack and Y. Boykov. Energy-based geometric multi-model fitting. *International Journal of Computer Vision (IJCV)*, 97(2):123–147, 2012. 5, 6

[20] F. Lafarge and C. Mallet. Building large urban environments from unstructured point data. In *International Conference on Computer Vision (ICCV)*, volume 1, pages 1068–1075. IEEE, 2011. 2

[21] X. Mei, X. Sun, M. Zhou, H. Wang, X. Zhang, et al. On building an accurate stereo matching system on graphics hardware. In *Computer Vision Workshops (in conjunction with ICCV)*, volume 1, pages 467–474. IEEE, 2011. 2

[22] B. Mičušík and J. Košecká. Multi-view superpixel stereo in urban environments. *International Journal of Computer Vision (IJCV)*, 89(1):106–119, 2010. 3

[23] P. Musialski, P. Wonka, D. G. Aliaga, M. Wimmer, L. van Gool, and W. Purgathofer. A survey of urban reconstruction. *Computer Graphics Forum*, 32(6):146–177, 7 2013. 1, 2

[24] M. Pollefeys, D. Nistér, J.-M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S.-J. Kim, P. Merrell, et al. Detailed real-time urban 3D reconstruction from video. *International Journal of Computer Vision (IJCV)*, 78(2-3):143–167, 2008. 2

[25] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition*, pages 31–42. Springer, 2014. 2

[26] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision (IJCV)*, 47(1-3):7–42, 2002. 2, 3

[27] S. N. Sinha, D. Steedly, and R. Szeliski. Piecewise planar stereo for image-based rendering. In *International Conference on Computer Vision (ICCV)*, volume 1, pages 1881–1888. IEEE, 2009. 1, 2

[28] N. Snavely, R. Garg, S. M. Seitz, and R. Szeliski. Finding paths through the world's photos. *Transactions on Graphics (TOG)*, 27(3):15, 2008. 2

[29] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3D. *Transactions on graphics (TOG)*, 25(3):835–846, 2006. 2

[30] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *International Journal of Computer Vision (IJCV)*, 80(2):189–210, 2008. 2

[31] C. Strecha, R. Fransens, and L. Van Gool. Wide-baseline stereo from multiple views: a probabilistic account. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 552–560. IEEE, 2004. 2

[32] C. Strecha, T. Tuytelaars, and L. Van Gool. Dense matching of multiple wide-baseline views. In *International Conference on Computer Vision (ICCV)*, volume 1, pages 1194–1201. IEEE, 2003. 2

[33] C. Strecha, W. Von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 1–8. IEEE, 2008. 6, 11, 12, 16

[34] E. Tola. *DAISY: A Fast Descriptor for Dense Wide Baseline Stereo and Multiview Reconstruction*. PhD thesis, EPFL, 2010. 3

[35] E. Tola, V. Lepetit, and P. Fua. DAISY: An efficient dense descriptor applied to wide-baseline stereo. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(5):815–830, 2010. 1, 3

[36] E. Tola, C. Strecha, and P. Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 23(5):903–920, 2012. 1

[37] R. Toldo and A. Fusiello. Robust multiple structures estimation with j-linkage. In *European Conference on Computer Vision (ECCV)*, pages 537–547. Springer, 2008. 2

[38] M. Vergauwen and L. Van Gool. Web-based 3D reconstruction service. *Machine Vision and Applications*, 17(6):411–426, 2006. 2

[39] C. Verleysen and C. De Vleeschouwer. Recognition of sport players' numbers using fast-color segmentation. In *IS&T/SPIE Electronic Imaging*, pages 80350–80360. International Society for Optics and Photonics, 2012. 6

[40] C. Verleysen and C. De Vleeschouwer. Learning and propagation of dominant colors for fast video segmentation. In *Advanced Concepts for Intelligent Vision Systems*, pages 657–668. Springer, 2013. 1, 6

[41] T. Werner and A. Zisserman. New techniques for automated architectural reconstruction from photographs. In *European Conference on Computer Vision (ECCV)*, pages 541–555. Springer, 2002. 2, 6, 16

[42] R. Yang and M. Pollefeys. Multi-resolution real-time stereo on commodity graphics hardware. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–211. IEEE, 2003. 2

[43] C. Zach, T. Pock, and H. Bischof. A globally optimal algorithm for robust tv-l 1 range image integration. In *International Conference on Computer Vision (ICCV)*, volume 1, pages 1–8. IEEE, 2007. 2

[44] C. L. Zitnick and S. B. Kang. Stereo for image-based rendering using image over-segmentation. *International Journal of Computer Vision (IJCV)*, 75(1):49–65, 2007. 2, 3

[45] M. Zuliani, C. S. Kenney, and B. Manjunath. The multiransac algorithm and its application to detect planar homographies. In *International Conference on Image Processing (ICIP)*, volume 3, pages III–153. IEEE, 2005. 2