Manhattan and Piecewise-Planar Constraints for Dense Monocular Mapping

Alejo Concha, Wajahat Hussain, Luis Montano and Javier Civera
Aragón Institute of Engineering Research (I3A)
Universidad de Zaragoza, Spain
{alejocb, hussain, montano, jcivera}@unizar.es

***Abstract*—This paper presents a variational formulation for real-time dense 3D mapping from a *RGB* monocular sequence that incorporates Manhattan and piecewise-planar constraints in indoor and outdoor man-made scenes.**

The state-of-the-art variational approaches are based on the minimization of an energy functional composed of two terms, the first one accounting for the photometric compatibility in multiple views, and the second one favoring smooth solutions. We show that the addition of a third energy term modelling Manhattan and piecewise-planar structures greatly improves the accuracy of the dense visual maps, particularly for low-textured man-made environments where the data term can be ambiguous.

We evaluate two different methods to provide such Manhattan and piecewise-planar constraints based on 1) multiview superpixel geometry and 2) multiview layout estimation and scene understanding. Our experiments include the largest map produced by variational methods from a *RGB* sequence and demonstrate a reduction in the median depth error up to a factor $5\times$.

I. Introduction

Real-time and fully dense –one point per pixel– 3D reconstruction from a monocular *RGB* sequence has been recently achieved thanks to the advances in the optimisation techniques and the availability of powerful graphical units. The so-called *DTAM*, standing for *Dense Tracking and Mapping* [17], and previous approaches [20, 15] estimate the depth for every pixel in a reference image by minimizing an energy functional composed of a data term and a regularization term. The data term, or photometric error, accounts for the pixel color difference between every pixel in the reference image and its correspondences in a number of short-baseline views. The regularization term favors smooth and low-gradient depth solutions.

DTAM shows an excellent performance when the data term is highly informative and the regularization term only plays a role in a small number of contiguous pixels. For *RGB* sequences, this is usually the case in highly textured scenes and large-parallax camera motions. But as the highly informative data becomes sparse the standard gradient-based regularizer might produce maps of low accuracy. In large textureless image regions there might be several depth solutions that hold the low-gradient regularization constraint but are of low accuracy. The lack of texture in that areas produces depth estimations mostly dominated by the image noise. *DTAM* can also be inaccurate for low-parallax camera motions and non-Lambertian effects –this latest limitation is mentioned in the original paper [17].

These three limitations described above –large textureless regions, low-parallax motions and non-Lambertian surfaces– are usually found in man-made scenes. The contribution of this paper is to model the Manhattan and piecewise-planar structures usually found in such scenes in order to improve the accuracy of the variational approach to dense mapping. Specifically, we model the Manhattan and piecewise-planar constraints as an extra term in the *DTAM* energy function. We evaluate two different methods to extract the parameters of the planarity constraints; the first one based on multiview superpixel geometry –piecewise-planar constraint– and the second one based on multiview layout estimation –Manhattan constraint.

Our experimental results show that our approach

improves by a large margin the accuracy of the standard *DTAM* in man-made scenes. We present the largest dense 3D map produced by a DTAM-like system from a monocular *RGB* camera, and demonstrate the key role that our formulation plays to achieve accurate results.

The rest of the paper is organised as follows. Section II describes the related work, section III relates how to extract the Manhattan and piecewise-planar constraints from multiple views, and section IV details the variational formulation including such constraints. Section V presents the experimental results and finally section VI concludes.

II. Related Work

The more recent research on dense and real-time mapping has focused on *RGB-D* cameras, e.g. [16, 23, 12]. Our approach addresses the more challenging case of *RGB* monocular sequences, being [20, 15, 17] the first works that used the total variational framework of [24] to achieve dense and real-time multiview 3D reconstructions.

There are several recent papers that use high-level understanding to improve geometry-only 3D reconstructions. [2] recognizes objects and estimates a sparse 3D map jointly, improving over the two tasks performed separately. [3] uses object category constraints to densely reconstruct 3D object models. [5] uses object shape constraints to improve 3D dense reconstructions. [18] recognizes RGB patches and searches for depth data in a RGB-D dataset to fill textureless gaps in Structure from Motion (SfM) maps. The Manhattan assumption has also been combined with traditional sparse 3D reconstruction in order to fill textureless gaps [8, 9, 22]. Our contribution is the addition of the Manhattan constraints in the variational formulation of *DTAM*, which has several advantages. First, the formulation allows a parallel computation and real-time performance using graphical units. Secondly, having a photometric and smoothness constraint *per pixel* instead of a sparse set of constraints improves the accuracy and robustness of the results.

Multiview *sparse* 3D reconstructions have made use of layout estimation and scene understanding in [7, 21]; and superpixels in [14, 4]. Again, our contribution over them is the use of such cues in a variational framework.

III. Manhattan and Piecewise-Planar constraints from Multiple Views

Our algorithm takes as the only input a *RGB* monocular image sequence $\mathcal{V}$. Dense mapping based on variational methods selects first a sparsely sampled set of keyframes $\{\mathbf{I}_1, \dots, \mathbf{I}_r, \dots, \mathbf{I}_j, \dots, \mathbf{I}_m\} \in \mathcal{V}$. The goal is to estimate an accurate inverse depth $\rho(\mathbf{u})$ for every pixel $\mathbf{u}$ in a reference image $\mathbf{I}_r$ using the information of a subset of close keyframes.

The next subsections describe the two algorithms that extract the rough planar structure of a man-made scene. III-A details how to extract the rough layout of a Manhattan-like indoor scene using high-level multiview scene understanding. III-B explains how we estimate a sparse piecewise-planar reconstruction by reconstructing a set of superpixels from multiple views. Both algorithms will provide depth priors in planar textureless areas that we incorporate to the variational formulation as detailed in section IV.

In both cases we need two preprocessing steps. We extract first a set of salient points $\mathbf{u}^* \in \mathbf{u}$ in every keyframe, compute correspondences and estimate the salient points' 3D positions $\mathbf{p} = \left(\mathbf{p}_1^\top \ \dots \ \mathbf{p}_i^\top \ \dots \ \mathbf{p}_n^\top\right)^\top$ and camera poses $\mathbf{c} = \left(\mathbf{c}_1^\top \ \dots \ \mathbf{c}_r^\top \ \dots \ \mathbf{c}_j^\top \ \dots \ \mathbf{c}_m^\top\right)^\top$ using a standard Bundle Adjustment optimization [19].

In the second preprocessing step, we segment every keyframe $\mathbf{I}_r$ into a set of superpixels $\mathcal{S}_r = \{s_1^r, \dots, s_l^r, \dots, s_t^r\}$ using the algorithm by Felzenszwalb et al. [6].

A. Manhattan constraints from Multiview Layout Estimation

One of the goals of indoor scene understanding is the estimation of the rough geometry of a room –its layout– and the classification of every image pixel $\mathbf{u}$ into the wall, floor, ceiling or clutter classes. In this paper we basically use the algorithm of [10] and extend it to a multiview case.

The main assumption is that we are in a cuboid room. The geometric model of the room layout $\mathcal{L}$ will be composed of six planes $\mathcal{L} =$

$\{\pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6\}$. Every plane π_k will be parametrized by its plane normal $\mathbf{n}_k$ and distance to the origin d_k. We first estimate the plane normals $\mathbf{n}_k$ by extracting the vanishing points $\mathbf{v}_k^r$ from the dominant directions of the room in every keyframe $\mathbf{I}_r$ [13]. We backproject them to the 3D world $\mathbf{V}_k^r = \mathbf{K}_r^{-1}\mathbf{v}_k^r$, group them into three clusters, and take their centroids.

We use the sparse point cloud estimation $\mathbf{p}$ to compute the plane distances d_k. Specifically, we hypothesize several distances d_k^h and build histograms for the distance between the point cloud $\mathbf{p}$ and such plane distances d_k^h. We assume that in a room most of the salient features will be close to its geometric boundaries and we take as initial seed the hypothesis with the minimum median.

Finally, we label every superpixel from the segmentation $\mathcal{S}_r = \{s_1^r, \ldots, s_l^r, \ldots, s_t^r\}$ into 4 different classes $\{W, F, C, Cl\}$ –wall, floor, ceiling and clutter respectively. See [11] for details on the superpixel features and the classification algorithm used. We will not constraint the depth of the pixels $\mathbf{u} \in Cl$ that are labeled as clutter. For the rest of the pixels $\mathbf{u} \in \{W, F, C\}$ we will compute an *a priori* inverse depth $\rho_\pi(\mathbf{u})$ from the intersection between the backprojected ray $\mathbf{K}_r^{-1}\mathbf{u}$ and the layout plane $\pi_k \in \mathcal{L}$ where it has been classified

$$\rho_\pi(\mathbf{u}) = || - \frac{\mathbf{u}\mathbf{K}_r^{-1}\mathbf{R}_r\mathbf{n}_k}{d_k\mathbf{K}_r^{-1}\mathbf{u}}|| \ . \tag{1}$$

B. Piecewise-planar constraints from Multiview Superpixel Geometry

We assume that the superpixels $\mathcal{S}_r = \{s_1^r, \ldots, s_l^r, \ldots, s_t^r\}$ correspond to approximately planar areas in the scene. We will estimate their 3D parameters using [4], which we will summarize here for completeness.

We can estimate the geometric parameters $\mathbf{\Pi} = \left(\pi_1^\top \ \ldots \ \pi_k^\top \ \ldots \ \pi_q^\top\right)^\top$ for the q planar superpixels $\{s_1, \ldots, s_k, \ldots, s_q\}$ that were matched in two or more keyframes with the following optimization

$$\hat{\mathbf{\Pi}} = \arg\min_{\mathbf{\Pi}} \sum_{r=1}^{m}\sum_{k=1}^{q} F(\epsilon_{s_k}^r) \ . \tag{2}$$

$\epsilon_{s_k}^r = \mathbf{u}_{s_k}^r - \mathbf{h}(\pi_k, \mathbf{c}_r)$ stands for the reprojection error of the superpixel s_k *contours* in the keyframe $\mathbf{I}_r$. As we are approximating the superpixels by planar surfaces, $\mathbf{h}$ stands for a homography model. We use a robust function of the error $F(*)$ to avoid the influence of outliers. As before, every superpixel π_k is parametrized by its plane normal $\mathbf{n}_k$ and distance to the origin d_k.

The superpixel correspondences between several views are computed as follows. We first search for pairwise correspondences between two keyframes $\mathbf{I}_r$ and $\mathbf{I}_j$ using a Monte Carlo approach. For every superpixel $\mathbf{s}_k$ in $\mathbf{I}_r$ we create several plane hypotheses π_k^h. The plane hypothesis are ranked according to the reprojection error of the superpixel contours in image $\mathbf{I}_j$

$$\epsilon_{s_k^h} \quad = \quad ||\mathbf{u}_{s_k^h}^j - \mathbf{h}\left(\mathbf{u}_{s_k^h}^j, \pi_k^h, \mathbf{c}_r, \mathbf{c}_j\right)|| \tag{3}$$

The planar superpixel hypotheses π_k^h with the smallest error $\epsilon_{s_k^h}$ are taken as the initial seed for the optimization of equation 2.

The Manhattan inverse depth prior $\rho_\pi(\mathbf{u})$ for each pixel $\mathbf{u} \in s_k$ is computed as the intersection of its backprojected ray and the plane π_k (equation 1).

IV. A Variational Formulation for Dense Mapping with Manhattan and Piecewise-Planar Constraints

The variational approaches to mapping aim to estimate the inverse depth $\rho(\mathbf{u})$ for every pixel $\mathbf{u}$ of a reference image $\mathbf{I_r}$. In order to do that we minimize a global energy function E_ρ; which is the weighted sum of a photometric error data term $\mathbf{C}(\mathbf{u}, \rho(\mathbf{u}))$, a robust spatial regularization term $\mathbf{G}(\mathbf{u}, \rho(\mathbf{u}))$ and our newly proposed Manhattan or piecewise-planar term $\mathbf{M}(\mathbf{u}, \rho(\mathbf{u}), \rho_\pi(\mathbf{u}))$

$$\begin{aligned} E_\rho \quad &= \quad \int (\lambda_1 \mathbf{C}(\mathbf{u}, \rho(\mathbf{u})) + \mathbf{G}(\mathbf{u}, \rho(\mathbf{u})) \\ &+ \quad \frac{\lambda_2}{2}\mathbf{M}(\mathbf{u}, \rho(\mathbf{u}), \rho_\pi(\mathbf{u}))\partial\mathbf{u} \end{aligned} \tag{4}$$

λ_1 and λ_2 are the weighting factor that account for the relative importance of the photometric, Manhattan/piecewise-planar and smoothness costs.

The photometric term. As in [17], our photometric error is based on color difference between the reference image and the set of short-baseline images. Every pixel $\mathbf{u}$ of the reference image $\mathbf{I}_r$ is first backprojected at an inverse distance ρ and projected again in every close image $\mathbf{I}_j$.

$$\mathbf{u}^j = \mathbf{T}_{rj}(\mathbf{u}, \rho) = \mathbf{K}\mathbf{R}^\top \left(\begin{pmatrix} \frac{\mathbf{K}^{-1}\mathbf{u}}{||\mathbf{K}^{-1}\mathbf{u}||} \\ \rho \end{pmatrix} - \mathbf{t} \right) \tag{5}$$

The photometric error is the summation of the color error between every pixel in the reference image and its corresponding in every other image at an hypothesized inverse distance ρ

$$\mathbf{C}(\mathbf{u}, \rho(\mathbf{u})) = \frac{1}{|I_s|} \sum_{j=1, j\neq r}^{m} ||\epsilon(\mathbf{I}_j, \mathbf{I}_r, \mathbf{u}, \rho)||_1 \tag{6}$$

$$\epsilon(\mathbf{I}_j, \mathbf{I}_r, \mathbf{u}, \rho) = \mathbf{I}_r(\mathbf{u}) - \mathbf{I}_j(\mathbf{T}_{rj}(\mathbf{u}, \rho)) \tag{7}$$

The gradient regularizer. The gradient regularizer is the Huber norm of the weighted gradient of the inverse depth map $||\nabla\rho(\mathbf{u})||_\epsilon$

$$\mathbf{G}(\mathbf{u^r}, \rho(\mathbf{u})) = \mathbf{g}(\mathbf{u^r})||\nabla\rho(\mathbf{u})||_\epsilon \tag{8}$$

Depth discontinuities often coincides with contours. $g(\mathbf{u})$ is a per-pixel weight that decreases the regularization strength for high-gradient pixels.

$$g(\mathbf{u}) = \mathbf{e}^{-\alpha||\nabla \mathbf{I_r}(\mathbf{u})||_\mathbf{2}} \tag{9}$$

The Manhattan or piecewise-planar constraint. The third term is the Manhattan or piecewise-planar constraint. It measures how far is every point from the estimated planar prior ρ_π detailed in section III:

$$\mathbf{M}(\mathbf{u}, \rho(\mathbf{u}), \rho_\pi(\mathbf{u})) = ||\rho(\mathbf{u}) - \rho_\pi(\mathbf{u})||_\mathbf{2}^\mathbf{2} \tag{10}$$

In the areas of the image where we do not have a planar constraint (highly textured or classified as clutter) we set $\lambda_2 = 0$.

Solution. The energy is composed of two convex terms $g(\mathbf{u})||\nabla\rho(\mathbf{u})||_\epsilon + \frac{\lambda_\mathbf{2}}{\mathbf{2}}||\rho(\mathbf{u}) - \rho_\pi(\mathbf{u})||_\mathbf{2}^\mathbf{2}$ and a non-convex term $\lambda_1\mathbf{C}(\mathbf{u}, \rho(\mathbf{u}))$. The convex terms and the non-convex term are optimized differently. An auxiliary variable $\mathbf{a}$ is used to make these two terms independent from each other:

$$E_{\rho,\mathbf{a}} = \int \Big(\lambda_1 \mathbf{C}(\mathbf{u}, \mathbf{a}(\mathbf{u})) + g(\mathbf{u})||\nabla\rho(\mathbf{u})||_\epsilon + \\ + \frac{\lambda_2}{2}||\rho(\mathbf{u}) - \rho_\pi(\mathbf{u})||_2^2 + \frac{1}{2\theta}(\rho(\mathbf{u}) - \mathbf{a}(\mathbf{u}))^\mathbf{2} \Big) \partial\mathbf{u} \tag{11}$$

The coupling term $\frac{1}{2\theta}(\rho(\mathbf{u}) - \mathbf{a}(\mathbf{u}))^\mathbf{2}$ will enforce ρ and $\mathbf{a}$ to become the same as θ is drived to 0 iteratively. Therefore, equation 11 will result in the original energy 4.

The non-convex term will be optimised by sampling and the convex terms will be efficiently optimised using a primal-dual approach.

The convex terms are converted to their primal-dual formulation using the Legendre-Fenchel transformation (details and proofs in [1]). The energy in the equation 11 is then minimized as follows

$$\underset{\mathbf{q}, ||\mathbf{q}||_\mathbf{2} \leq \mathbf{1}}{\arg\max} \left\{ \underset{\rho, a}{\arg\min}\, \mathbf{E}(\rho, \mathbf{a}, \mathbf{q}) \right\} \tag{12}$$

$$\mathbf{E}(\rho, \mathbf{a}, \mathbf{q}) = \Big\{ \langle \mathbf{g}\mathbf{A}\rho, \mathbf{q} \rangle - \delta_\mathbf{q}(\mathbf{q}) - \frac{\epsilon}{\mathbf{2}}||\mathbf{q}||_\mathbf{2}^\mathbf{2} \\ + \frac{\lambda_2}{2}||\rho - \rho_\pi||_2^2 + \frac{1}{2\theta}(\rho - \mathbf{a})^\mathbf{2} + \lambda_\mathbf{1}\mathbf{C}(\mathbf{a}) \Big\} \tag{13}$$

Where $\mathbf{q}$ is the dual variable and $\mathbf{A}\rho$ stands for the gradient of ρ.

For the dual variable $\mathbf{q}$ the energy has to be maximized, therefore a gradient ascent step is computed:

$$\frac{\partial \mathbf{E}(\rho, \mathbf{a}, \mathbf{q})}{\partial \mathbf{q}} = \mathbf{g}\mathbf{A}\rho - \epsilon\mathbf{q} \tag{14}$$

Discretizing and rearranging terms:

$$\mathbf{q^{n+1}} = (\mathbf{q^n} + \sigma_\mathbf{q}\mathbf{g}\mathbf{A}\rho^\mathbf{n}) \,/\, (\mathbf{1} + \sigma_\mathbf{q}\epsilon) \tag{15}$$

$$\mathbf{q^{n+1}} = \mathbf{q^{n+1}} / \max(\mathbf{1}, ||\mathbf{q^{n+1}}||_\mathbf{1}) \tag{16}$$

In the case of the variable ρ, the energy is minimized, therefore a gradient descent step is computed. Using the divergence theorem $\frac{\partial \langle A\rho, q \rangle}{\partial \rho} =$

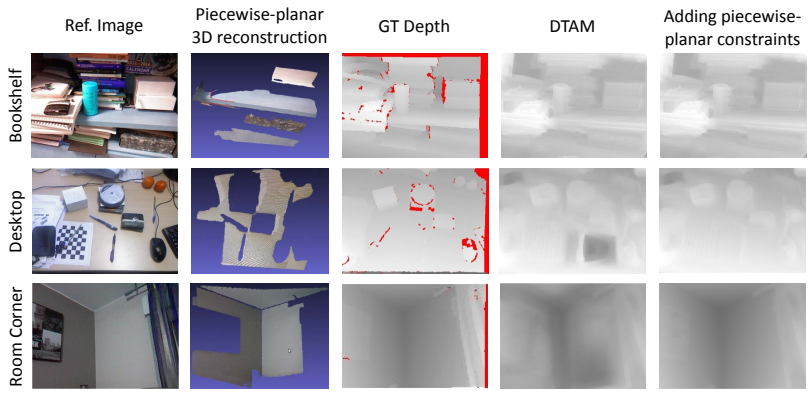


Fig. 1. Summary of the small-scale experiments. 1^{st} column: Reference image. 2^{nd} column: Piecewise-planar reconstruction using superpixels. 3^{rd} column: ground truth depth –red stands for no-depth-data. 4^{th} column: *DTAM* depth. 5^{th} column: Depth using the piecewise-planar constraints. Notice how the latest column is visually closer to the ground truth than the *DTAM* one.

$-div(\mathbf{q}) = \mathbf{A^T q}$, where $\mathbf{A^T}$ forms the negative divergence operator:

$$\frac{\partial \mathbf{E}(\rho, \mathbf{a}, \mathbf{q})}{\partial \rho} = \mathbf{g A^T q} + \frac{\mathbf{1}}{\theta}(\rho - \mathbf{a}) + \lambda_\mathbf{2}(\rho - \rho_\pi) \quad (17)$$

Discretizing and rearranging terms:

$$\rho^{n+1} = \frac{\left(\rho^n + \sigma_\rho \left(-\mathbf{g A^T q^{n+1}} + \frac{\mathbf{a^n}}{\theta^\mathbf{n}} + \lambda_\mathbf{2}\rho_\pi\right)\right)}{(1 + \frac{\sigma_\rho}{\theta^n} + \lambda_2 \sigma_\rho)} \quad (18)$$

The remaining non-convex function is minimised using a point-wise search for each $\mathbf{a}$ in the range $\mathbf{a} = [\rho_\mathbf{min}, \rho_\mathbf{max}]$:

$$\arg\min_a \mathbf{E^{aux}}(\rho, \mathbf{a}) \quad (19)$$

$$\mathbf{E^{aux}}(\rho, \mathbf{a}) = \frac{\mathbf{1}}{\mathbf{2}\theta}(\rho - \mathbf{a})^\mathbf{2} + \lambda_\mathbf{1}\mathbf{C}(\mathbf{a})) \quad (20)$$

Equations 15, 16, 18, 19 are performed iteratively until $\theta^{n+1} = \theta^n(1 - 0.001 * n)$ is below a certain threshold. Variables are initialized as follows: $\mathbf{q^0} = \mathbf{0}$ and $\rho^0 = \mathbf{a^0} = \arg\min_\mathbf{a} \mathbf{C}(\mathbf{u}, \mathbf{a})$

Finally, we use the sub-sample accuracy method recommended in [17]:

$$\mathbf{a^{n+1}} = \mathbf{a^{n+1}} - \frac{\nabla \mathbf{E^{aux}}}{\nabla^\mathbf{2} \mathbf{E^{aux}}} \quad (21)$$

V. Experimental Results

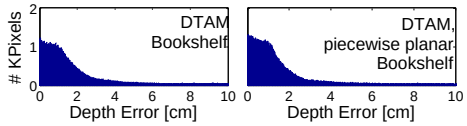
We have evaluated our proposal using indoor sequences in sections V-A and V-B. We recorded the sequences with a *RGB-D* camera, took the depth channel *D* as the ground truth depth and used the *RGB* data as the input for the standard *DTAM* algorithm and our proposal incorporating the planarity priors. We also present results with outdoor sequences in section V-C. Only qualitative results are shown there due to the limitations of *RGB-D* sensors outdoors.

A. Small-scale experiments

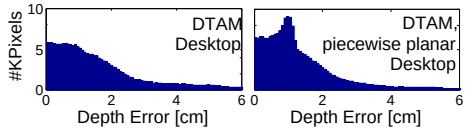
Figure 1 summarizes the three small-scale experiments we have performed –*Bookshelf*, *Desktop* and *Room Corner*. For each one of them we show the reference image in the first column, the piecewise-planar reconstruction of the largest superpixels in the second row, and the ground truth depth, the standard *DTAM* depth and the depth using our proposal in the third, fourth and fifth columns respectively.

Figure 2 shows the histogram of the depth errors for the three experiments and both cases; standard *DTAM* and our approach –adding piecewise-planar constraints. Notice first how for the highly textured *Bookshelf* image, both *DTAM* and our approach performances are very similar and close to the ground truth. In any case, thanks to the extra

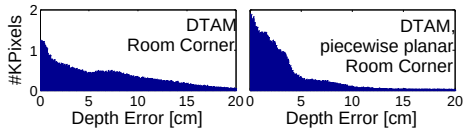
constraints, the median error is reduced 8%, from $1.3cm$ to $1.2cm$.



(a) Bookshelf seq. Median error reduction: 11%.



(b) Desktop seq. Median error reduction: 26%.



(c) Room corner seq. Median error reduction: 60%.

Fig. 2. Depth error histograms for the standard variational mapping approach (*DTAM*) and our proposal using piecewise-planar constraints.

For the *Desktop* and *Room Corner* experiments, where significantly large textureless areas exist, our proposal shows a noticeable improvement. *DTAM* depth maps present some defects that piecewise-planar constraints are able to reduce. We believe the errors in the *Desktop* experiment were caused by light reflections. The *Room Corner* sequence, being recorded in a larger scene, also has the challenge of lower-parallax camera motions and hence the addition of constraints makes the improvement even larger.

For a quantitative evaluation, see the histograms in figures 2(b) and 2(c) and table I. The reduction of the median error is 26% in the *Bookshelf* sequence and 60% in the *Room Corner* sequence.

B. Large-scale Experiment

The sequence for this experiment was recorded in our *Lab* with a hand-held camera, covering approximately half of the room. As the scene is larger than the previous ones, we estimated the depth for 5 references images with the information of around 50 close keyframes for each one. Each

Sequence	**Median Depth Error [cm]**	
	DTAM	**Our approach**
Bookshelf	1.3	1.2
Desktop	1.5	1.1
Room Corner	5.6	2.3
Lab (Layout)	27.0	13.4
Lab (Superpixels)		5.2

TABLE I
MEDIAN OF THE ESTIMATED DEPTH ERROR FOR THE STANDARD *DTAM* AND OUR APPROACH.

one of the *DTAM* challenges that we mentioned in the introduction –large textureless areas, low-parallax motions and non-Lambertian effects– appears in this sequence. Up to our knowledge, it is also the largest scene ever mapped using variational methods. More details on the experiment can be seen in the video accompanying the paper.

As the sequence images a box-like room, we evaluate the performance of the two algorithms to extract the planar priors described in sections III-A and III-B. Figure 3 shows a qualitative summary of the experiment. Each row shows the results for the depth of each reference image. In columns we show, respectively: The reference *RGB* image, the multiview superpixels, the multiview layout, the ground truth depth, the results of standard *DTAM*, and finally the results of our approach both using superpixel and layout constraints. Notice that the depth of our approach is always closer to the ground truth depth image. The better accuracy of our approach can also be seen in figure 4, that shows top and side views of the complete map. The top view of the *DTAM* map, figure 4(a), shows a more inaccurate reconstruction than the one of our approach in figure 4(b).

Specifically, for this experiment *DTAM*'s median depth error is $27.0cm$, our approach's one using the constraint coming from the layout is $13.4cm$ and our approach's one using the constraint coming from the superpixels is $5.2cm$. The distribution of the depth errors can be better appreciated in the histograms of figure 5. The difference between the two constraints is due to the lower maturity of the scene understanding techniques compared to multiview geometry. In any case, we believe that

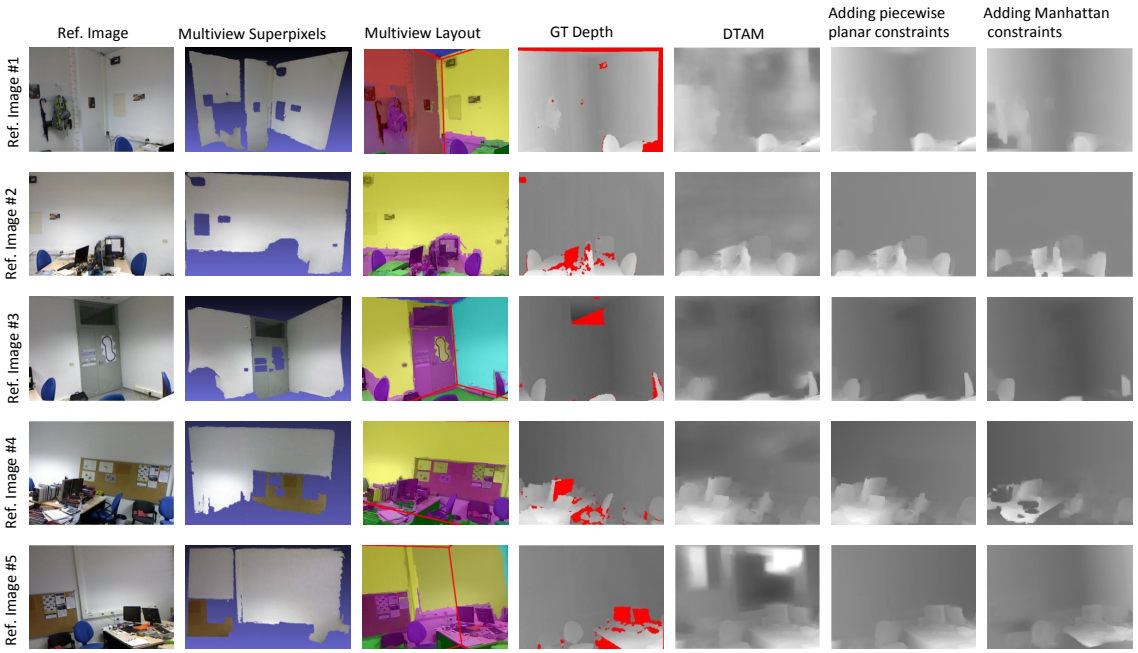


Fig. 3. Large-scale experiment. Each row shows the results for a reference image. 1^{st} column: *RGB* image. 2^{nd} column: multiview superpixel reconstruction. 3^{rd} column: layout constraint. Red lines stand for the projected box. Magenta stands for clutter, green for floor and dark blue for ceiling. Other colors stand for walls. 4^{th} column: ground truth depth –red stands for no-depth-data. 5^{th}: column *DTAM* depth. 6^{th} column: *DTAM* results using piecewise-planar constraints. 7^{th} column: *DTAM* results using Manhattan constraints. The improvement of the depth maps of *DTAM* with planarity constraints against the standard *DTAM* is visually noticeable.

the semantic information that the former provides could be of great interest for robotics, and hence could be an interesting line for further research.

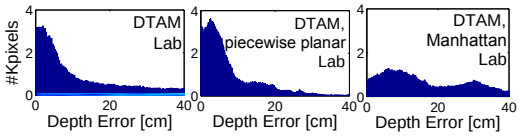
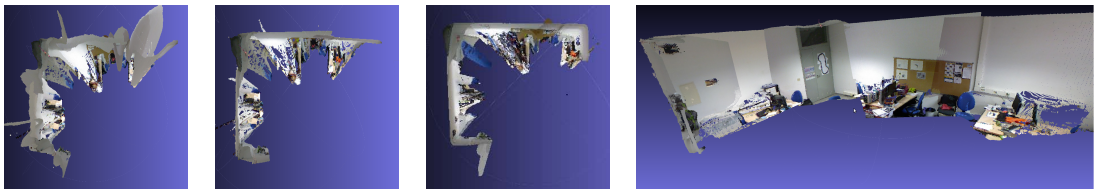


Fig. 5. Depth error histograms for the standard variational mapping approach (*DTAM*), and our proposed additions of piecewise-planar and Manhattan constraints. Median error is reduced in a factor $5\times$ in the best case (piecewise-planar constraints), from $27.0cm$ to $5.2cm$

C. Outdoor Scenes

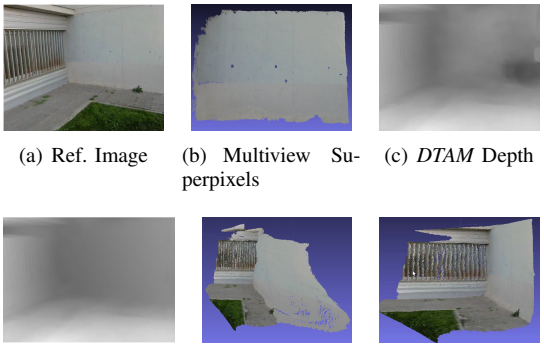
Figures 6 and 7 summarize the two outdoor experiments that we performed, in a building corner and a building façade respectively. Though ground truth depth is not available, the accuracy improvement is noticeable from the figures. Notice the defect of the *DTAM* depth image in the right wall of the building corner –figure 6(c)– and the planar depth that our approach estimated in the same area –figure 6(d). *DTAM* produces a distorted 3D map –see figure 6(e)–, mostly noticeable in the right wall; while the 3D map of our approach is accurate in this area –figure 6(f).

For the façade experiment observe how the *DTAM* depth image of figure 7(c) does not correspond to a mostly constant depth, as it should. This depth errors result in the erroneous wavy 3D reconstruction of figure 7(e). Notice how the depth image of our approach in 7(d) reflects the constant depth of the façade, and how the 3D map of figure 7(f) is a more accurate reconstruction of the scene.



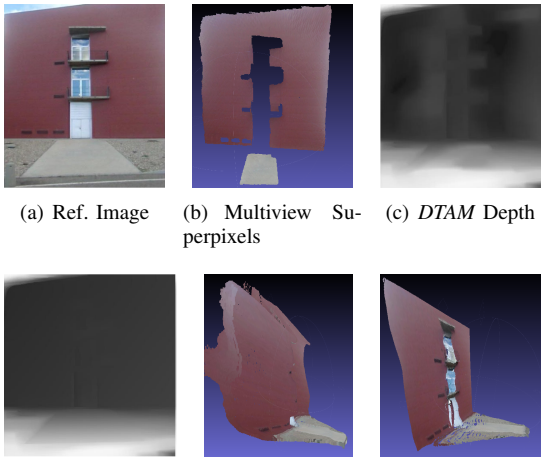
(a) Standard *DTAM*. Top view (b) *DTAM* + piecewise planar. Top view (c) *DTAM* + Manhattan. Top view (d) *DTAM* + piecewise-planar. Side view

Fig. 4. 3D maps for the *Lab* experiment. Notice the large *DTAM* errors in the top view of figure (a) and the more accurate reconstruction of figures (b) –adding piecewise-planar constraints– and (c) –adding the Manhattan constraint. Notice the different errors of the algorithms: (b) shows small misalignment errors, while (c) is globally consistent but with large errors in the objects and final parts of two walls due to inaccuracies in the layout and labels. (d) shows a side view of *DTAM* using piecewise-planar constraints, the one with most accurate results. Quantitative results are in figure 5 and table I



(a) Ref. Image (b) Multiview Superpixels (c) *DTAM* Depth (d) Piecewise-planar *DTAM* Depth (e) *DTAM* Map (f) Piecewise-planar *DTAM* Map

Fig. 6. Outdoor Corner seq. The higher accuracy of the piecewise-planar *DTAM* depth is qualitatively noticeable.



(a) Ref. Image (b) Multiview Superpixels (c) *DTAM* Depth (d) Piecewise-planar *DTAM* Depth (e) *DTAM* Map (f) Piecewise-planar *DTAM* Map

Fig. 7. Façade seq. The higher accuracy of the piecewise-planar *DTAM* depth is qualitatively noticeable.

VI. Conclusion

In this paper we have presented an algorithm that integrates Manhattan and piecewise-planar constraints into a variational formulation for real-time dense 3D mapping from a *RGB* camera. In our experiments we have shown that our proposal improves the accuracy of the state-of-the-art approaches in indoor and outdoor man-made scenes. We achieve a reduction factor up to $5\times$ in the median depth error of the reconstruction. The planarity constraints that we add are particularly relevant in image sequences where the data term is ambiguous or noisy; which is the case for low-texture scenes, low-parallax camera motions or distortions of the projection model –e.g., non-Lambertian surfaces.

We have evaluated two different methods to provide our approach with planarity constraints. In our experiments, multiview superpixel geometry produces 3D reconstructions of higher accuracy than the ones based on layout and understanding.

Acknowledgments

This research was funded by the Spanish government with the projects IPT-2012-1309-430000 and DPI2012-32168.

REFERENCES

[1] A. Angeli A. Handa, R. Newcombe and A. Davison. Applications of Legendre-Fenchel transformation to computer vision problems. In *Technical Report DTR11-7, Imperial College*, 2011.

[2] Sid Yingze Bao and Silvio Savarese. Semantic structure from motion. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2025–2032. IEEE, 2011.

[3] Y Bao, Manmohan Chandraker, Yuanqing Lin, and Silvio Savarese. Dense object reconstruction with semantic priors. In *26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[4] Alejo Concha and Javier Civera. Using superpixels in monocular SLAM. In *IEEE International Conference on Robotics and Automation*, Hong Kong, June 2014.

[5] Amaury Dame, Victor A Prisacariu, Carl Y Ren, and Ian Reid. Dense reconstruction using 3D object shape priors. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1288–1295, 2013.

[6] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[7] Alex Flint, David Murray, and Ian Reid. Manhattan scene understanding using monocular, stereo, and 3D features. In *2011 IEEE International Conference on Computer Vision (ICCV)*, pages 2228–2235, 2011.

[8] Y. Furukawa, B. Curless, S.M. Seitz, and R. Szeliski. Reconstructing building interiors from images. In *Proc. Int. Conf. on Computer Vision*, pages 80–87, 2009.

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

[10] Varsha Hedau, Derek Hoiem, and David Forsyth. Recovering the spatial layout of cluttered rooms. In *Computer vision, 2009 IEEE 12th international conference on*, pages 1849–1856. IEEE, 2009.

[11] D. Hoiem, A.A. Efros, and M. Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1):151–172, 2007.

[12] Christian Kerl, Jurgen Sturm, and Daniel Cremers. Dense visual SLAM for RGB-D cameras. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 2100–2106. IEEE, 2013.

[13] Jana Košecká and Wei Zhang. Video compass. In *Computer VisionECCV 2002*, pages 476–490. Springer, 2006.

[14] Branislav Mičušík and Jana Košecká. Multi-view superpixel stereo in urban environments. *International journal of computer vision*, 89(1):106–119, 2010.

[15] R.A. Newcombe and A.J. Davison. Live dense reconstruction with a single moving camera. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1498–1505. IEEE, 2010.

[16] Richard A Newcombe, Andrew J Davison, Shahram Izadi, Pushmeet Kohli, Otmar Hilliges, Jamie Shotton, David Molyneaux, Steve Hodges, David Kim, and Andrew Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *10th IEEE international symposium on Mixed and augmented reality (ISMAR)*, pages 127–136, 2011.

[17] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. DTAM: Dense tracking and mapping in real-time. In *2011 IEEE International Conference on Computer Vision (ICCV)*, pages 2320–2327, 2011.

[18] Andrew Owens, Jianxiong Xiao, Antonio Torralba, and William Freeman. Shape anchors for data-driven multi-view reconstruction. In *2013 IEEE International Conference on Computer Vision (ICCV)*, Sydney, Australia, December 2013.

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

[20] Jan Stühmer, Stefan Gumhold, and Daniel Cremers. Real-time dense geometry from a handheld camera. In *Pattern Recognition*, pages 11–20. Springer, 2010.

[21] Grace Tsai, Changhai Xu, Jingen Liu, and Benjamin Kuipers. Real-time indoor scene understanding using bayesian filtering with motion cues. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 121–128. IEEE, 2011.

[22] Carlos A Vanegas, Daniel G Aliaga, and Bedrich Benes. Building reconstruction using manhattan-world grammars. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 358–365. IEEE, 2010.

[23] Thomas Whelan, Michael Kaess, Maurice Fallon, Hordur Johannsson, John Leonard, and John McDonald. Kintinuous: Spatially extended KinectFusion. 2012.

[24] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime TV-L 1 optical flow. In *Pattern Recognition, Lecture Notes in Computer Science*, volume 4713, pages 214–223. Springer, 2007.