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×.

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, 13, 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 [2, 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 \( \{I_1, \ldots, I_r, \ldots, I_j, \ldots, I_m\} \in \mathcal{V} \). The goal is to estimate an accurate inverse depth \( \rho(u) \) for every pixel \( u \) in a reference image \( I_r \) 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 \( u^* \in u \) in every keyframe, compute correspondences and estimate the salient points’ 3D positions \( p = (p_1^\top \ldots p_i^\top \ldots p_n^\top)\top \) and camera poses \( c = (c_1^\top \ldots c_r^\top \ldots c_j^\top \ldots c_m^\top)\top \) using a standard Bundle Adjustment optimization [19].

In the second preprocessing step, we segment every keyframe \( I_r \) into a set of superpixels \( S_r = \{s_1^r, \ldots, s_{r_1}^r \} \) 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 \( u \) 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 \(\pi_k\) 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 \(V'_k\) from the dominant directions of the room in every keyframe \(I_r\) [13]. We backproject them to the 3D world \(V'_k = K_r^{-1}V'_k\), group them into three clusters, and take their centroids.

We use the sparse point cloud estimation \(p\) 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 \(p\) 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 \(S_r = \{s'_1, \ldots, s'_k, \ldots, s'_q\}\) 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 \(u \in Cl\) that are labeled as clutter. For the rest of the pixels \(u \in \{W, F, C\}\) we will compute an \(a\ priori\) inverse depth \(\rho^h(u)\) from the intersection between the backprojected ray \(K_r^{-1}u\) and the layout plane \(\pi_k \in L\) where it has been classified

\[
\rho^h(u) = \| - \frac{uK_r^{-1}R_n}{d}K_r^{-1}u \| . \tag{1}
\]

**B. Piecewise-planar constraints from Multiview Superpixel Geometry**

We assume that the superpixels \(S_r = \{s'_1, \ldots, s'_k, \ldots, s'_q\}\) 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 \(\Pi = \{\pi_1^\top \ldots \pi_k^\top \ldots \pi_q^\top\}\) 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

\[
\hat{\Pi} = \arg \min_{\Pi} \sum_{r=1}^{m} \sum_{k=1}^{q} F(\epsilon_{r_k}^h). \tag{2}
\]

\(\epsilon_{r_k}^h = u_{r_k}^h - h(\pi_k, c_r)\) stands for the reprojection error of the superpixel \(s_k\) contours in the keyframe \(I_r\). As we are approximating the superpixels by planar surfaces, \(h\) stands for a homography model. We use a robust function of the error \(F(\ast)\) to avoid the influence of outliers. As before, every superpixel \(\pi_k\) 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 \(I_r\) and \(I_j\) using a Monte Carlo approach. For every superpixel \(s_k\) in \(I_r\), we create several plane hypotheses \(\pi_k^h\). The plane hypothesis are ranked according to the reprojection error of the superpixel contours in image \(I_j\)

\[
\epsilon_{s_k}^h = \|u_{s_k}^h - h\left(u_{s_k}^h, \pi_k^h, c_r, c_j\right)\| . \tag{3}
\]

The planar superpixel hypotheses \(\pi_k^h\) 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 \(\rho^h(u)\) for each pixel \(u \in s_k\) is computed as the intersection of its backprojected ray and the plane \(\pi_k\) (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 \(\rho(u)\) for every pixel \(u\) of a reference image \(I_1\). In order to do that we minimize a global energy function \(E_\rho\), which is the weighted sum of a photometric error data term \(C(u, \rho(u))\), a robust spatial regularization term \(G(u, \rho(u))\) and our newly proposed Manhattan or piecewise-planar term \(M(u, \rho(u), \rho^h(u))\)

\[
E_\rho = \int \left(\lambda_1 C(u, \rho(u)) + G(u, \rho(u))\right) + \frac{\lambda_2}{2} M(u, \rho(u), \rho^h(u)) \right) du \tag{4}
\]

\(\lambda_1\) and \(\lambda_2\) are the weighting factor that account for the relative importance of the photometric, Manhattan/piecewise-planar and smoothness costs.
The photometric term. As in [7], our photometric error is based on color difference between the reference image and the set of short-baseline images. Every pixel $u$ of the reference image $I_r$ is first backprojected at an inverse distance $\rho$ and projected again in every close image $I_j$.

$$u^t = T_{rj}(u, \rho) = KR^T\left(\frac{K^{-1}u - t}{\rho}\right)$$

(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 $\rho$

$$C(u, \rho(u)) = \frac{1}{|I_s|} \sum_{j=1, j\neq r}^m ||\epsilon(I_j, I_r, u, \rho)||_1$$

(6)

$$\epsilon(I_j, I_r, u, \rho) = I_r(u) - I_j(T_{rj}(u, \rho))$$

(7)

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

$$G(u^r, \rho(u)) = g(u^r)||\nabla \rho(u)||_\epsilon$$

(8)

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

$$g(u) = e^{-\alpha ||\nabla \rho(u)||_2^2}$$

(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 $\rho_\pi$ detailed in section III

$$M(u, \rho(u), \rho_\pi(u)) = ||\rho(u) - \rho_\pi(u)||_2^2$$

(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 $g(u)||\nabla \rho(u)||_\epsilon + \frac{\lambda_2}{2} ||\rho(u) - \rho_\pi(u)||_2^2$ and a non-convex term $\lambda_1 C(u, \rho(u))$. The convex terms and the non-convex term are optimized differently. An auxiliary variable $a$ is used to make these two terms independent from each other:

$$E_{\rho,a} = \int \left(\lambda_1 C(u, a(u)) + g(u)||\nabla \rho(u)||_\epsilon + \frac{\lambda_2}{2} ||\rho(u) - \rho_\pi(u)||_2^2 + \frac{1}{2\theta}(\rho(u) - a(u))^2\right)du$$

(11)

The coupling term $\frac{1}{2\theta}(\rho(u) - a(u))^2$ will enforce $\rho$ and $a$ to become the same as $\theta$ is drived to 0 iteratively. Therefore, equation [11] will result in the original energy [3].

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

$$\arg \max_{q,||q||_2\leq 1} \left\{ \arg \min_{\rho,a} E(\rho, a, q) \right\}$$

(12)

$$E(\rho, a, q) = \left\{ \langle gA\rho, q \rangle - \delta_q(q) - \frac{\epsilon}{2}||q||_2^2 + \frac{\lambda_2}{2}||\rho - \rho_\pi||_2^2 + \frac{1}{2\theta}(\rho - a)^2 + \lambda_1 C(a) \right\}$$

(13)

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

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

$$\frac{\partial E(\rho, a, q)}{\partial q} = gA\rho - \epsilon q$$

(14)

Discretizing and rearranging terms:

$$q^{n+1} = (q^n + \sigma_q gA\rho^n) / (1 + \sigma_q \epsilon)$$

(15)

$$q^{n+1} = q^n / \max(1, ||q^n||_1)$$

(16)

In the case of the variable $\rho$, the energy is minimized, therefore a gradient descent step is computed. Using the divergence theorem $\frac{\partial (A\rho, q)}{\partial \rho} = $
Fig. 1. Summary of the small-scale experiments. 1st column: Reference image. 2nd column: Piecewise-planar reconstruction using superpixels. 3rd column: ground truth depth –red stands for no-depth-data. 4th column: DTAM depth. 5th column: Depth using the piecewise-planar constraints. Notice how the latest column is visually closer to the ground truth than the DTAM one.

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

\[
\frac{\partial \mathbf{E}(\rho, a, \mathbf{q})}{\partial \rho} = \mathbf{gA}^T \mathbf{q} + \frac{1}{\theta}(\rho-a) + \lambda_2(\rho - \rho_\pi) \quad (17)
\]

Discretizing and rearranging terms:

\[
\rho^{n+1} = \rho^n + \sigma_\rho \left( -\mathbf{gA}^T \mathbf{q}^{n+1} + \frac{n^n}{\theta_\pi} + \lambda_2 \rho_\pi \right) \quad (18)
\]

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

\[
\arg \min_a \mathbf{E}_{\text{aux}}(\rho, a) \quad (19)
\]

\[
\mathbf{E}_{\text{aux}}(\rho, a) = \frac{1}{2\theta}(\rho-a)^2 + \lambda_1 C(a) \quad (20)
\]

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

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

\[
a^{n+1} = a^{n+1} - \frac{\mathbf{V} E_{\text{aux}}}{\mathbf{V}^2 E_{\text{aux}}} \quad (21)
\]

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.

![Histograms](image)

(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 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

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Median Depth Error [cm]</th>
<th>DTAM</th>
<th>Our approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bookshelf</td>
<td>1.3</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>Desktop</td>
<td>1.5</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Room Corner</td>
<td>5.6</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>Lab (Layout)</td>
<td>27.0</td>
<td>13.4</td>
<td></td>
</tr>
<tr>
<td>Lab (Superpixels)</td>
<td>5.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE I

MEDIAN OF THE ESTIMATED DEPTH ERROR FOR THE STANDARD DTAM AND OUR APPROACH.
Fig. 3. Large-scale experiment. Each row shows the results for a reference image. 1\textsuperscript{st} column: RGB image. 2\textsuperscript{nd} column: multiview superpixel reconstruction. 3\textsuperscript{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\textsuperscript{th} column: ground truth depth –red stands for no-depth-data. 5\textsuperscript{th}: column DTAM depth. 6\textsuperscript{th} column: DTAM results using piecewise-planar constraints. 7\textsuperscript{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.

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× 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.
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 \textit{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


