Dense Accurate Urban Mapping from Spherical RGB-D Images
Renato Martins, Eduardo Fernandez-Moral, Patrick Rives

To cite this version:

HAL Id: hal-01237848
https://inria.hal.science/hal-01237848
Submitted on 3 Dec 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Dense Accurate Urban Mapping from Spherical RGB-D Images

Renato Martins, Eduardo Fernandez-Moral and Patrick Rives

Abstract—This paper presents a methodology to combine information from a sequence of RGB-D spherical views acquired by a home-made multi-stereo device in order to improve the computed depth images both in terms of accuracy and completeness. This methodology is embedded in a larger visual mapping framework aiming to produce accurate and dense topometric urban maps. Our method is based on two main filtering stages. Firstly, we perform a segmentation process considering both geometric and photometric image constraints, followed by a regularization step (spatial-integration). We then proceed to a fusion stage where the geometric information is further refined by considering the depth images of nearby frames (temporal integration). This methodology can be applied to other projective models, such as perspective stereo images. Our approach is evaluated within the frameworks of image registration, localization and mapping, demonstrating higher accuracy and larger convergence domains over different datasets.

I. INTRODUCTION

Producing dense, accurate and compact 3D maps from stereo sequences is relevant for a large number of applications, from localization and mapping [10] to scene rendering [8] and remains a challenging task. The main issues derive from the difficulty to take into account the visibility constraints and to represent the geometric discontinuities arising from real surfaces. In addition, errors coming from a sum of different sources directly affect the quality of photometric and geometric data acquired by the commodity imaging sensors.

In [10], we proposed a topometric representation composed of a graph of geolocalized RGB-D spherical images computed from a home-made multi-stereo device (fig. 1). The RGB-D spherical images are positioned in the scene accurately thanks to a robust dense visual odometry (VO) method. In order to limit drift inherent to VO and to process large urban scenes in constant time, only representative keyframes selected based on an information criterion are kept in the final representation.

However, the images that were not selected as keyframes were dropped out and not exploited to improve the scene representation. This leads to losing useful information which could be used to enhance the quality of the RGB-D keyframes. In this paper, we propose a novel scheme which aims to integrate all the information available in a lower number of RGB-D keyframes. The main objective is to build more accurate keyframes (depths maps) using the RGB-D acquisitions from different points-of-view. The overall approach is performed in two steps. First we perform a segmentation of the spherical image in isotropic planar patches (in the 3D space), which implicitly apply a coherent regularization prior to a fusion stage. The fusion relies on our previous work [5], where coherent regularized frames are merged in a single keyframe, taking into account the related uncertainties and their co-visibility. In summary, we exploit the redundancy of nearby frames to reduce noise and the uncertainty of the keyframe’s depth images. Note that our method can be also applied for building representations in indoor environments with Kinect/Asus sensors.

Experiments are presented in the context of image registration and mapping, showing that both precision and map consistency are greatly improved using this approach. The improved depth information represented a break for dense registration techniques in a set of experiments. We particularly show the improvement in the convergence domain of direct registration and on 3D plane segmentation, which has been previously applied to more compact scene modelling towards to stable life-long mapping [3].

II. RELATED WORK

A variety of methods have been proposed for RGB-D mapping and depth sensor fusion [9] [6] [10] [13], recently boosted by the release of commodity and popular Microsoft Kinect sensor. The goal of inferring pose and dense mapping is closely related to [2] and [9], who applies robust photometric cost functions for dealing with scene changes on indoor contexts. The works of [8] and [17] explores complementary aspects (quantization and surface smoothness) by explicitly accounting for discontinuities and outliers in

The authors are with Inria Sophia-Antipolis Méditerranée, France {name.surname}@inria.fr
photo-realistic scene rendering. Both aspects are considered in this work.

Some recent methods as [15] and [13] employ non-linear energy minimization operators for dealing with the inferred depth measurements errors and outliers. These errors are mainly from three different sources: i) wrong pixel matching assignments (particularly in low textured regions); ii) occlusions; and iii) violation of the lambertian surface assumption (i.e. variant spectral reflections, mirrors and windows/glass structures). A limitation of these previous works is the reduced size of the models (a workspace smaller than \(8\times8\times8\)) imposed by the volumetric voxel-grid sampling, which grows unbounded along the scene scale since the representation quality is directly tight to sampling resolution (although the efforts made on voxel-adaptive grids as in [14]).

As stated previously, the presented methodology is directly related to [10] and [5]. The keyframe based framework is presented in [10], but the discarded frames (highly redundant) are not exploited to improve the graph information. This leads to the need of a higher number of frames to represent the same scene and a smaller convergence domain for registration, reducing the region of exploration using the learned model. By last, we extend the method proposed in [5] by exploiting the rigidity of neighbourhoods through a joint depth – color segmentation, which has a clear improvement in the maps produced, specially when considering the more challenging data coming from stereo outdoor sequences.

III. PROPOSED APPROACH

Our algorithm is mainly composed of two steps: i) segmentation and regularization of the geometric component of the reconstructed spherical images; and ii) registering and merging closely-regularized frames in a common reference model, through a direct registration procedure. In the first step, the point cloud built is segmented in planar patches employing a region growing approach, whilst enforcing an isotropic spatial distribution (equal area). Next, a regularization of the planar patches is made considering the photometric component of the data. These stages allow one to be more robust to the noise in the computed stereo depth maps, coming from wrong disparity block matching assignments.

The second step consists of a fusion technique, where the geometric information of a keyframe is further refined by the regularized depth images of nearby frames. The similarity criterion to combine a segmented sphere to a keyframe is given while performing a dense registration on both intensity and geometry components [9]. We derive next these steps for building more accurate depth maps, which also allows a sparser keyframe based ego-centric representation.

A. Notation and Preliminaries

A spherical image \(S = \{I, D\}\) is composed of \(I \in [0,1]^{m \times n}\) as pixel intensities and \(D \in \mathbb{R}^{m \times n}\) as the depth. The depth information is obtained from stereo correspondence using the Semi-Global Block Matching algorithm (SGBM) [7]. The disparity image was also computed using the state of the art algorithm ELAS [4], but the conclusions we get were similar to [16]: ELAS overregularize the depth maps producing border blending artifacts, where posterior filtering regularization might have marginal influence. On the other hand, while SGBM gives less smoothed maps (gaps with no depth information), the errors are more local and thus can be treated following the steps presented in the next sections. An example of this phenomenon is given in the experiment showed in fig. 2.

Let us consider a surface in \(D\), \(s : \mathbb{R}^4 \rightarrow \mathbb{R}, s(q) = 0\) smooth and with a normal vector for each pixel. The normal vector is given by its gradient \(n = \nabla s(q)\), orthogonal to its tangent plane \(P(n,d)\): \(n^Tq + d = 0\), \(\forall q \in P\). A small window neighbourhood is considered to estimate \(n\) when \(s\) has sharp features, which increases noise influence on the estimate (see the work of [12] which analyses the effects of noise, curvature, and sampling density and focused on how automatically select the appropriate window size). Given a region (or a cluster representative plane) \(C_i\) with a set of points \(q \in C_i\), and an extracted planar patch model \(P_i\), the orthogonal distance based metric \(L_d\) and the normal consistency \(L_n\) errors in this region are defined as

\[
\begin{align*}
L_d(C_i, P_i) &= \int_{q \in C_i} ||n^Tq + d||_2 dq \\
L_n(C_i, P_i) &= \int_{q \in C_i} ||n^Tn(q)||_1 dq
\end{align*}
\]

In the following sections, we introduce the segmentation/regularization and posterior registration procedures assuming these concepts and metrics.

B. Segmentation of Planar Patches

The raw sphere is segmented to produce planar patches. The clustering nearby points in the image algorithm is based on the choice of a seed (the point and its related normal vector) and its neighbouring points are then tested by considering the spatial metrics as defined in (1). If these two criteria are fulfilled within pre-defined thresholds, then the point is included in the cluster.

Since the surface resolution decreases within the distance to the sensor, an adaptive number of allowed points at each
C. Combining Intensity Coherent Patches

Instead of warping the geometric raw patches directly, we enforce the surface regularization taking into account the photometric image component $I$ of the raw spheres by employing a superpixel based segmentation. The chosen method in this work is the single linear iterative clustering (SLIC) algorithm \cite{A2}, which encodes nice properties as strong adherence to boundaries and compactness, besides of the \cite{A1} algorithm, which encodes nice properties as strong adherence to boundaries and compactness, besides of the \cite{A1} algorithm, which encodes nice properties as strong adherence to boundaries and compactness, besides of

This segmentation also greatly reduces scene complexity and simultaneously applies surface regularization.

Algorithm III-B.1 : Patch geometric scene segmentation

1. Input1: $\rho_1$ and $A$: $n_1 = d(\rho_1)A$
2. Input2: max errors $\epsilon_n$ and $\epsilon_d$
3. for all pixels $p^*$ in $D$ do
4. \quad $\rho_2 = D(p^*)$ and $n_2 = n_1 \left(\frac{2}{3}n\right)^2$
5. \quad $\min_n = \max(n_2 - 2, 3)$ and $n_{\max} = n_2$
6. \quad while size($C_i(k)$) < $n_{\max}$ & $\delta_n(p) < \epsilon_n$ & $\delta_d(p, C_i) < \epsilon_d$ do
7. \quad $C_i(k + 1) = [C_i(k) \ g(p)]$
8. \quad Reevaluate plane equation on $C_i(k + 1)$
9. \quad end while
10. end for

Fig. 3. Isotropic growing region. Keeping the area of patches constant avoids flattening geometry details when far from the sensor (the distribution of points is asymptotically decreasing with the range). In the left a structure sampling of 20 points per patch. The right figure has the same assumption that we aim to construct patches of around 20 points when from 5[m] from the sensor, which gives an area of 10[cm]^2. This area is then propagated to all depth levels, resulting in isotropic patches.
performed in two stages. First, each neighbouring depth image is warped on the spherical keyframe reference sphere as $D_w(p^*) = D_k(w(p^*, T))$, where $w(\bullet)$ is the warping function

$$q_w(p, T) = \begin{bmatrix} T \end{bmatrix}$$

The second stage corresponds to finding the keyframe model that minimizes the square error between the previously warped segmented spheres

$$D^*(p) = \arg\min_{\Pi_D} \sum_{i=1}^{n} \Pi_D(p)(D_w(p) - D(p))^2$$

which can be stated sequentially as a simple weighted average as follows:

$$\begin{cases} D_{k+1}^*(p) = W_k^D(p)D_k^*(p) + \Pi_D(p)D_w(p) \\ W_{k+1}^D = W_k^D + \Pi_D(p) \end{cases}$$

where $\Pi_D(p) = 1/\sigma^2_{D_k}(p)$ is the confidence (the inverse of the uncertainty) resulting from the blending of both pose and structure errors; and $W_k^D$, being the uncertainty of the keyframe model initialization.

IV. EXPERIMENTS

In this section we evaluate the results of direct registration on a sequence of urban images with and without fusion (for short, we employ the word fusion to refer to the described procedure of segmentation/regularization plus combination of near spheres). We show that the fusion improves the accuracy of dense registration and its robustness allowing a larger region of convergence. Moreover, it also results in a higher efficiency for localization in a previously created map.

In our tests we employ two different sequences of spherical RGB-D images recorded at the urban area of Garbejaire in Sophia-Antipolis (seq1) as shown in fig. 5 and in the Inria Sophia-Antipolis research center (seq2). The camera rig is fixed to a car (see fig. 4), which is driven at a variable speed depending on the road/traffic conditions, with an average of 30 km/h and a maximum speed of 80 km/h. The sequences are recorded at a frame rate of 20 Hz, where the six global shutter cameras of the stereo system are synchronized, producing spherical images with a resolution of 2048x665 (see fig. 4). Such sequences are fused offline to obtain maps that can be used later for localization (like in [10]) or for scene rendering (in a similar fashion to Google Street View) as we show in the accompanying video. Though real-time performance is out of the scope of this paper, we would like to remark that an optimized implementation of this work exploiting parallel computing and/or GPU might be used to achieve real-time performance.

A. Scene Rendering and Segmentation

The accuracy obtained after fusion is clearly apparent by visual inspection of the reconstructed point clouds. In figure 5, we can see on the left an image reconstructed from raw data, and the same frame on the right after fusion. Notice how the artefacts and the waves on the floor are removed in the view after fusion. Inspection of the depth images and the normal vector images also confirm this, where we see that flat and/or smooth surfaces like the building façades or the road are more regular in the fused images, whilst keeping sharp details of the surface (please refer to the accompanying video for more details of the point cloud reconstruction from different points-of-view).

We tested the segmentation of planar surfaces from the fused and non-fused depth maps. Since our sequences are recorded in structured scenarios, the detection of the planar surfaces in them provides qualitative information about how good are the obtained depth maps (fig. 5). Table I shows the average number of planar surfaces segmented, and the average number of point inliers per image on the corresponding sequence.

B. Direct RGB-D Registration

As described in section II-D direct registration of RGB-D images consists of retrieving the relative pose of two frames (6 DoF) by minimizing intensity and depth differences between the reference and the warped images. In order to evaluate the improvement on the accuracy of dense

---

1 video url:(www-sop.inria.fr/members/Renato-Jose.Martins/iros15.html)
registration (minimising the photometric error and both the photometric and the geometric errors) after our fusion stage, we compute the trajectory using dense registration on a set of 4000 subsequent fused spheres imposing the requirement to fulfil a smooth motion model (almost constant velocity) in 3D (6 DoF). This trajectory is used as groundtruth to compare the results of pair-wise registration of nearby frames in the sequence (which are not subsequent). Table II shows the average errors obtained by each registration method with and without fusion, confirming the error reduction of the improved data after fusion. Notice that the photo-consistency method is also more accurate when the filtered depth maps are used.

To further confirm the higher accuracy of dense registration after fusion we carry out another experiment in which the spherical image is divided into two halves (left and right), so that each one has 180 degrees FOV on the horizontal plane. We perform dense registration with the above methods in each half separately and compute their difference (deviation) in rotation and translation. The poses estimated by registering both should be the same. The average deviations for registration of 100 images in the Inria sequence are shown in table III, supporting the same findings as in the previous experiment.

**TABLE I
AVERAGE RESULTS FOR PLANE SEGMENTATION.**

<table>
<thead>
<tr>
<th></th>
<th>Av. num planes</th>
<th>Av. inliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq1 Regular</td>
<td>4.67</td>
<td>119 K</td>
</tr>
<tr>
<td>Seq1 Fusion</td>
<td>6.41</td>
<td>148 K</td>
</tr>
<tr>
<td>Seq2 Regular</td>
<td>4.23</td>
<td>124 K</td>
</tr>
<tr>
<td>Seq2 Fusion</td>
<td>6.15</td>
<td>159 K</td>
</tr>
</tbody>
</table>

**TABLE II
AVERAGE ERROR OF THE DIFFERENT REGISTRATION METHODS.**

<table>
<thead>
<tr>
<th></th>
<th>Av. Rot. Error (deg)</th>
<th>Av. Trans. Error (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raw</td>
<td>Fusion</td>
</tr>
<tr>
<td>Dense RGB-D</td>
<td>0.51</td>
<td>0.12</td>
</tr>
<tr>
<td>Photo-consistency</td>
<td>0.47</td>
<td>0.12</td>
</tr>
</tbody>
</table>

**TABLE III
AVERAGE DEVIATIONS OF HALF-SPHERE REGISTRATION.**

<table>
<thead>
<tr>
<th></th>
<th>Rotation (deg)</th>
<th>Translation (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raw</td>
<td>Fusion</td>
</tr>
<tr>
<td>Dense RGB-D</td>
<td>0.87</td>
<td>0.16</td>
</tr>
<tr>
<td>Photo-consistency</td>
<td>0.55</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Another benefit of the fused spheres is the higher robustness of direct registration, since the region of convergence enlarges as a consequence of the higher accuracy of the estimated depth maps. To illustrate this, we show the ratio of convergence of the methods evaluated above for registering sets of spheres with different separations, from 10[cm] to 50[cm] (see figure S). The convergence condition requires that the difference in the translation is less than 1[cm] and that the difference in the rotation is less than 1 degree with respect to our ground truth trajectory. The approximate distances shown in the figure correspond to that of our ground truth within a range of ±3[cm]. As expected, we can see that both techniques for direct registration (RGB-D and photo-consistency) present larger convergence domains than for the case where raw depth information is used. We can also conclude that when such raw depth is used, the photo-
consistency method has a larger convergence rate than the RGB-D method. This indicates that the photometric term in (3) is less influenced by wrong depth measurements than its geometric counterpart. On the other hand, the convergence is higher for RGB-D when employing the fused spheres as a consequence of the improved data. This means that fusion allows to handle larger separation between frames, and thus, lower camera frame rates and/or higher acquisition speed can be attained. This has also a direct impact on the hardware resources, reducing the computation and memory requirements as less frames need to be processed for odometry/mapping purposes.

Besides the advantages of higher accuracy and robustness, we remark that there is also an advantage on the computational cost of dense registration. Convergence is reached with a reduced number of iterations, and thus, the time required to register a pair of spheres is 44% shorter in average with respect to the same optimization using the raw depth images. Table IV presents the average computation times for both cases using a single CPU with a processor of 2.6 GHz. This result is only relevant for localization applications in which there is already a build representation like in [10], since the fusion and regularization steps introduce an additional processing stage whose cost is around $N$ times larger than the dense registration cost, being $N$ the size of the sliding window (the number of frames required to optimize one keyframe image).

![Fig. 8. Direct registration convergence for different image separations.](image)

**TABLE IV**

<table>
<thead>
<tr>
<th>Processing time</th>
<th>Raw</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photo-consistency</td>
<td>0.91 s</td>
<td>0.65 s</td>
</tr>
<tr>
<td>Dense RGB-D</td>
<td>1.09 s</td>
<td>0.76 s</td>
</tr>
</tbody>
</table>

V. CONCLUSIONS AND PERSPECTIVES

A method for fusion-regularization of stereo depth maps has been described. This method is applied to our particular case of spherical vision, though it can be generally applied to other contexts evolving RGB-D data (as showed in the experiments section [IV]). The improved accuracy of the keyframe geometric images applying this approach is useful for different tasks. Concretely, we show that it has several advantages for direct registration, increasing its accuracy and robustness. This is a key aspect for generating accurate maps to be used later on for robot localization or simple scene rendering, like e.g. Google Street View. The larger region of convergence for registration also contributes to store less frames in the representation.

One of the limitations of the experimental section is the lack of an accurate ground truth of the trajectory obtained with a method different than direct registration. This is planned to be overcomed by applying loop closure restrictions in our sequence. Also, a comparison with other techniques as those based in measurement stability [11] are left as future work.

ACKNOWLEDGEMENTS

This work was supported by CNPq of Brazil under grant number 216026/2013-0. The authors thank Florent Lafarge for discussions on the patch algorithm segmentation, and Tawsif Gokhool and the reviewers for their critical insights and suggestions for improving the material of this paper.

REFERENCES