Dense Accurate Urban Mapping from Spherical RGB-D Images

Renato Martins, Eduardo Fernandez-Moral, Patrick Rives

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.

The authors are with Inria Sophia-Antipolis Méditerranée, France {name.surname}@inria.fr

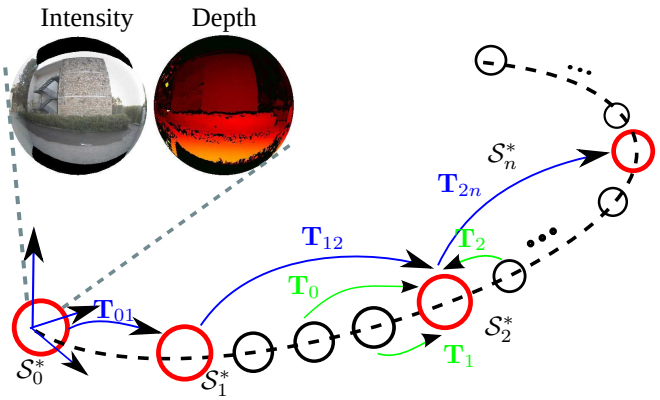

Fig. 1. Topometric graph: the nodes are RGB-D spherical images, which are relied by rigid transformations $\mathbf{T} \in \mathbb{SE}(3)$. The poses $\mathbf{T}_{ij}$ lay keyframes $\mathcal{S}^*$ (in red), while $\mathbf{T}_k$ relate near spheres (in black) to a particular keyframe. Only keyframes are kept in the final map model. In this work, each keyframe is built by exploiting the redundancy of nearby frames to reduce its noise and uncertainty.

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 photo-geometric 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

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[m]x8[m]x8[m]) 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 closeby 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 $\mathcal{S} = \{\mathcal{I}, \mathcal{D}\}$ is composed of $\mathcal{I} \in [0,1]^{m \times n}$ as pixel intensities and $\mathcal{D} \in \mathbb{R}^{m \times n}$ 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

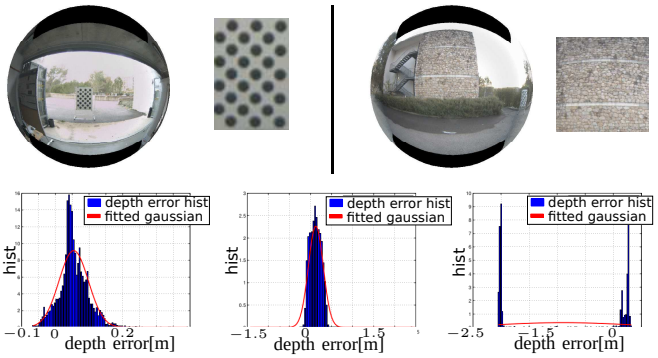

Fig. 2. Top row: reconstructed spheres in two different configurations within their respective segmented targets (a calibration grid and the wall of a building). Bottom row: depth error pdfs for: *i)* calibration grid at 5[m] using SGBM method; *ii)* building wall at 5[m] using SGBM; and *iii)* calibration grid at 5[m] using ELAS. A big region of the calibration target had erroneous disparity inference with ELAS, which causes a multi-modal distribution shape (as can be noticed in the bottom-right image).

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 they 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 $\mathcal{D}$, $s : \mathbb{R}^4 \mapsto \mathbb{R}, s(\mathbf{q}) = 0$ smooth and with a normal vector for each pixel. The normal vector is given by its gradient $\mathbf{n} = \nabla s(\mathbf{q})$, orthogonal to its tangent plane $\mathcal{P}(\mathbf{n}, d)$: $\mathbf{n}^T\mathbf{q} + d$ with $d = -\mathbf{n}^T\mathbf{q}_0, \forall \mathbf{q}_0 \in \mathcal{P}$. A small window neighbourhood is considered to estimate $\mathbf{n}$ 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) $\mathcal{C}_i$ with a set of points $\mathbf{q} \in \mathcal{C}_i$, and an extracted planar patch model $\mathcal{P}_i$, the orthogonal distance based metric ($\mathcal{L}_d$) and the normal consistency ($\mathcal{L}_n$) errors in this region are defined as

$$\begin{aligned} \mathcal{L}_d(\mathcal{C}_i, \mathcal{P}_i) &= \int\int_{\mathbf{q} \in \mathcal{C}_i} \|\mathbf{n}^T\mathbf{q} + d\|_2^2 dq \\ \mathcal{L}_n(\mathcal{C}_i, \mathcal{P}_i) &= \int\int_{\mathbf{q} \in \mathcal{C}_i} \|\mathbf{n}^T\mathbf{n}(\mathbf{q})\|_1 dq \end{aligned} \tag{1}$$

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

cluster is employed to build isometric patches (in the 3D space). This avoids undesirable effects as aliasing details of far objects and the over-sampling of structures close to the sensor, as shown in fig. 3. The number of allowed points per patch depends on the desired area A and the spatial density point's distribution $d(\rho_1) = \frac{N}{4\pi\rho_1^2}$, with N being the total number of points. Given the number of points $n_1 = d(\rho_1)A$ at range ρ_1, a same area patch at range ρ_2 has $n_2 = n_1 \left(\frac{\rho_1}{\rho_2}\right)^2$ points. Then it is sufficient to consider an interval around the value of n_2 as minimum and max points in the region growing procedure, as presented in the algorithm III-B. This segmentation also greatly reduces scene complexity and simultaneously applies surface regularization.

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 $\mathcal{I}$ of the raw spheres by employing a superpixel based segmentation. The chosen method in this work is the single linear iterative clustering (SLIC) algorithm [1], which encodes nice properties as strong adherence to boundaries and compactness, besides of a simple tuning parameter being the approximate size of the desired superpixel cells. Afterwards, a mean patch $(d_s, \mathbf{n}_s)$ related to each superpixel region $\mathcal{M}_i$ is extracted considering all the patches that are mostly englobed by that superpixel.

As a consequence of the isotropic area segmentation (described in the previous section), the intensity based regularization is just the combination of all $\mathcal{P}_i$ in $\mathcal{M}_i$ that are bellow a threshold from the computed mean patch

Algorithm III-B.1 : Patch geometric scene segmentation

Input1: ρ_1 and A: $n_1 = d(\rho_1)A$
Input2: max errors ϵ_n and ϵ_d
for all pixels $\mathbf{p}^*$ in $\mathcal{D}$ **do**
  $\rho_2 = \mathcal{D}(\mathbf{p}^*)$ and $n_2 = n_1 \left(\frac{\rho_1}{\rho_2}\right)^2$
  $n_{min} = \max(n_2 - 2, 3)$ and $n_{max} = n_2$
  while $size(\mathcal{C}_i(k)) < n_{max}$ & $\delta_n(\mathbf{p}) < \epsilon_n$& $\delta_d(\mathbf{p}, \mathcal{C}_i) < \epsilon_d$ **do**
    $\mathcal{C}_i(k+1) = [\mathcal{C}_i(k)\ g(\mathbf{p})]$
    Reevaluate plane equation on $C_i(k+1)$
  end while
end for

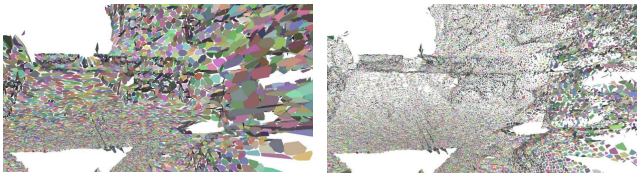

Fig. 3. Isotropic growing region. Keeping the area of patches constant avoids flatting geometry details when far from the sensor (the distribution of points is asymptotically decreasing with the range). In the left a structure sampling of 20 points per patch. The right figure has the same assumption that we aim to construct patches of around 20 points when from 5[m] from the sensor, which gives an square of 10[cm] edge. This area is then propagated to all depth levels, resulting in isotropic patches.

superpixel $(d_s, \mathbf{n}_s)$. All $\mathcal{P}_i$ verifying $\|d_i\mathbf{n}_i - d_s\mathbf{n}_s\|_2 < \epsilon_1$ and $\|\mathbf{n}_i^T\mathbf{n}_s\|_1 < \epsilon_2$ are combined in $\mathcal{P}_f(d_f, \mathbf{n}_f)$ by:

$$\begin{cases} \mathbf{n}_{f(k+1)} = (k\mathbf{n}_{f(k)} + \mathbf{n}_j)/(k+1) \text{ and} \\ d_{f(k+1)} = \|(kd_{f(k)}\mathbf{n}_{f(k)} + d_j\mathbf{n}_j)\|_2/(k+1) \end{cases} \quad (2)$$

The set of pixels $\mathbf{p} = \mathbf{p}_i \cup \mathbf{p}_j$, such as depth $\mathcal{D}(\mathbf{p})$ belongs to either $\mathcal{P}_i, \mathcal{P}_j$ planes, are then fulfilled by employing the resulting final plane parameters in (2): $\mathcal{D}_f(\mathbf{p}) = \left\| \frac{d_f}{(\mathbf{n}_f^T \mathbf{q}_S(\mathbf{p}))} \right\|_1$. One would ask why not performing the regularization directly over the superpixels regions $\mathcal{M}_i$ instead of using the scene geometry. The reason is that outliers in the depth measurements can roughly affect the estimates of the final regularized patch.

D. Spherical Dense Registration

A direct registration procedure is applied for retrieving the relative pose $\mathbf{x} \in \mathbb{R}^6$ between two spheres by minimizing a cost function which accounts for the differences in intensity and depth of all pixels between the reference and target frames (see e.g. [2]). Given an initial estimate $\widehat{\mathbf{T}}$, the cost function $\mathfrak{F}_S = \frac{1}{2}\|e_\mathcal{I}\|_\mathcal{I}^2 + \frac{\lambda^2}{2}\|e_\rho\|_\mathcal{D}^2$ can be written explicitly as:

$$\mathfrak{F}_S = \tfrac{1}{2}\sum_{\mathbf{p}^*} \mathbf{W}^I(\mathbf{p}^*) \left\| \mathcal{I}\left(w(\mathbf{p}^*, \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}))\right) - \mathcal{I}^*(\mathbf{p}^*) \right\|^2 + \tfrac{\lambda^2}{2}\sum_{\mathbf{p}^*} \mathbf{W}^D(\mathbf{p}^*) \left\| \mathbf{n}^T \left(g(w(\mathbf{p}^*, \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}))) - \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}) g^*(\mathbf{p}^*)\right) \right\|^2, \quad (3)$$

with $w(\bullet)$ being the warping function depending on the pose to be estimated, λ is a tuning parameter to effectively scale both error terms (intensity and depth consistency), $\mathbf{W}^I$ and $\mathbf{W}^D$ are weights related to the confidence of each measure (with $\mathbf{W}^I$ computed from a Huber robust estimator as in [9]). The normal vector $\mathbf{n}$ is computed at the reference sphere within the neighbourhoods of each 3D point $g(\mathbf{p}^*)$, where g^{-1} corresponds to the projection model, e.g., perspective or spherical.

In addition to this standard dense VO formulation, we combine a set of implementation strategies for both avoiding local minima in the optimization and for speeding-up the computation. Thus, the registration is achieved from a coarse to fine scheme using different pyramid resolutions. The subpixel registration is then only applied on the larger resolution, while the registration on other resolutions is done with nearest neighbour approximation. Also, we store a depth-buffer of the projected pixels to take into account possible occlusions [11]. This direct registration of spherical images is applied in the sequence for both fusion-regularization of the keyframe mapping and for odometry.

E. Merging Close Spheres

Finally, based on a distance criterion between spheres that encapsulates the similarity of both geometry and intensity, a filtering scheme can be set up for combining "nearby" augmented regularized spheres in order to reduce the error in the retained keyframe. For this, we follow the approach of [5] which is briefly summarized for the sake of completeness.

Assuming that a set of n near spheres share enough coherent information, their combination into a keyframe is

performed in two stages. First, each neighbouring depth image is warped on the spherical keyframe reference sphere as $\mathcal{D}_w(\mathbf{p}^*) = \mathcal{D}_t(w(\mathbf{p}^*, \mathbf{T}))$, where $w(\bullet)$ is the warping function and

$$\begin{aligned} \mathcal{D}_t(\mathbf{p}) &= \sqrt{\mathbf{q}_w(\mathbf{p}, \mathbf{T})^T \mathbf{q}_w(\mathbf{p}, \mathbf{T})} \\ \mathbf{q}_w(\mathbf{p}, \mathbf{T}) &= \left([\mathbf{I}\ \mathbf{0}] \mathbf{T} \begin{bmatrix} g(\mathbf{p}) \\ 1 \end{bmatrix} \right) \end{aligned} \tag{4}$$

The second stage corresponds to finding the keyframe model that minimizes the square error between the previously warped segmented spheres $\mathcal{D}^*(\mathbf{p}) = \text{argmin}_{\mathbf{D}} \sum_{i=1}^{n} \Pi_D(\mathbf{p})(\mathcal{D}_{w_i}(\mathbf{p}) - \mathcal{D}(\mathbf{p}))^2$, which can be stated sequentially as a simple weighted average as follows:

$$\begin{cases} \mathcal{D}^*_{k+1}(\mathbf{p}) = \dfrac{\mathbf{W}^D_k(\mathbf{p})\mathcal{D}^*_k(\mathbf{p}) + \Pi_D(\mathbf{p})\mathcal{D}_{w_k}(\mathbf{p})}{\mathbf{W}^D_k(\mathbf{p}) + \Pi_D(\mathbf{p})} \\ \mathbf{W}^D_{k+1} = \mathbf{W}^D_k + \Pi_D \end{cases} \tag{5}$$

where, $\Pi_D(\mathbf{p}) = 1/\sigma^2_{\mathcal{D}_w}(\mathbf{p})$ is the confidence (the inverse of the uncertainty) resulting from the blending of both pose and structure errors; and $\mathbf{W}^D_0$, 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 show in fig. 6, 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

Fig. 4. Globeye stereo sensor and acquisition system.

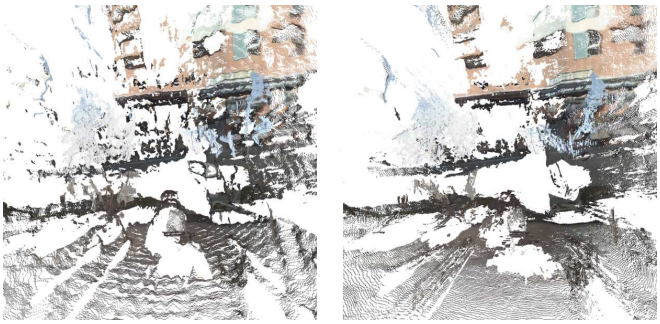

Fig. 5. Bird's-eye of view rendered images from one of the keyframe nodes (the related RGB image is showed in the bottom-right of fig. 4) with and without fusion.

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 [1]. 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. 7). 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 III-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)

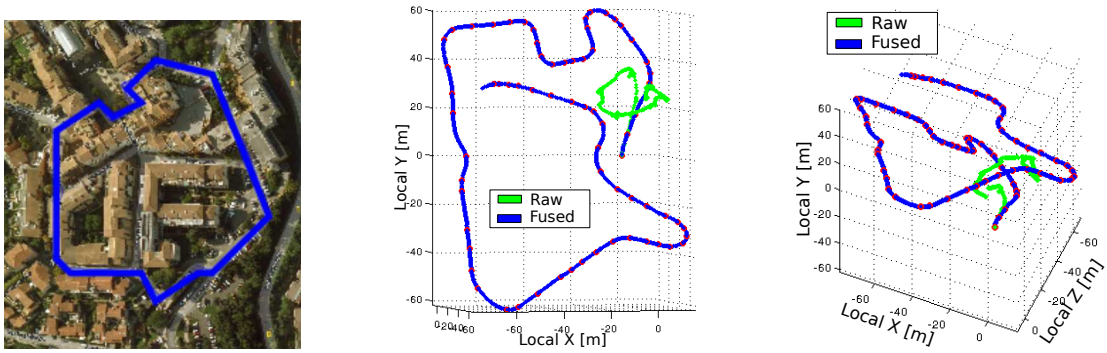

Fig. 6. Trajectories and final topometric graph for the Garbejaire sequence (seq1): the bird's-eye of view of the urban 0.7[km] travelled path (left in blue) and the final topometric model (the middle and right images) using both raw (green) and fused data (blue). It is noticeable the graph consistency improvement in the resulting model after the fusion procedure and only about 20% of the original spheres were retained.

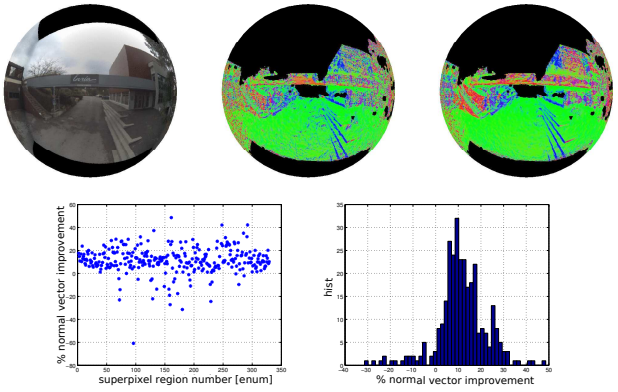

Fig. 7. Normal vector consistency for the raw data (middle) and filtered (right) in the seq2 dataset. The normal vectors are color encoded for visualization, then smooth planar surfaces in the scene should have the same color. The improvement of the filtering procedure varied between 10 to 30% using the metrics of normal consistency in (1) (bottom right and left images).

TABLE I

AVERAGE RESULTS FOR PLANE SEGMENTATION.

	Av. num planes	*Av. inliers*
Seq1 Regular	4.67	119 K
Seq1 Fusion	6.41	148 K
Seq2 Regular	4.23	124 K
Seq2 Fusion	6.15	159 K

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

TABLE II

AVERAGE ERROR OF THE DIFFERENT REGISTRATION METHODS.

	Av. Rot. Error (deg)		Av. Trans. Error (mm)	
	Raw	*Fusion*	*Raw*	*Fusion*
Dense RGB-D	0.51	**0.12**	3.4	**1.1**
Photo-consistency	0.47	**0.12**	2.9	**1.3**

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 III

AVERAGE DEVIATIONS OF HALF-SPHERE REGISTRATION.

	Rotation (deg)		Translation (mm)	
	Raw	*Fusion*	*Raw*	*Fusion*
Dense RGB-D	0.87	**0.16**	2.3	**0.89**
Photo-consistency	0.55	**0.18**	1.8	**0.88**

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 8). 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-

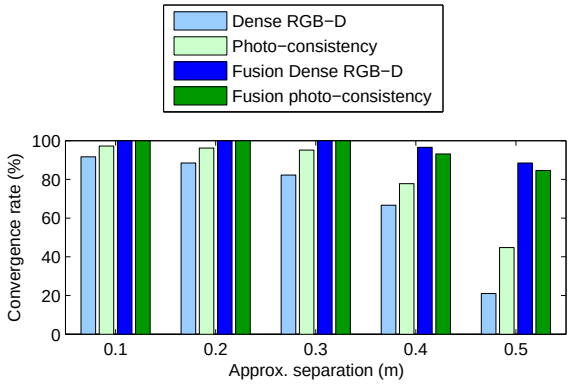

Fig. 8. Direct registration convergence for different image separations.

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).

TABLE IV

AVERAGE TIMINGS FOR DENSE PAIRWISE REGISTRATION.

Processing time	Raw	Fusion
Photo-consistency	0.91 s	0.65 s
Dense RGB-D	1.09 s	0.76 s

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

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC superpixels compared to state-of-the-art superpixels methods. *IEEE Trans. PAMI*, 34(11), 2012.
[2] C. Audras, A. Comport, M. Meilland, and P. Rives. Real-time dense appearance-based SLAM for RGB-D sensors. In *ACRA*, 2011.
[3] E. Fernandez-Moral, W. Mayol-Cuevas, V. Arevalo, and J. Gonzalez-Jimenez. Fast place recognition with plane-based maps. In *IEEE ICRA*, 2013.
[4] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In *ACCV*, 2010.
[5] T. Gokhool, R. Martins, P. Rives, and N. Despre. A compact spherical RGBD keyframe-based representation. In *IEEE ICRA*, 2015.
[6] T. Gokhool, M. Meilland, P. Rives, and E. Fernandez-Moral. A dense map building approach from spherical RGBD images. In *VISAPP*, 2014.
[7] H. Hirschmulle. Stereo processing by semi-global matching and mutual information. *IEEE Trans. PAMI*, 30(2), 2008.
[8] B. Huhle, T. Schairer, P. Jenke, and W. Straßer. Fusion of range and color images for denoising and resolution enhancement with a non-local filter. *CVIU*, 114, 2010.
[9] C. Kerl, J. Sturm, and D. Cremers. Dense visual SLAM for RGB-D cameras. In *IEEE IROS*, 2013.
[10] M. Meilland, A. Comport, and P. Rives. Dense omnidirectional RGB-D mapping of large-scale outdoor environments for real-time localization and autonomous navigation. *JFR*, 32(4), 2015.
[11] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J. Frahm, R. Yang, D. Nister, and M. Pollefeys. Real-time visibility-based fusion of depth maps. In *IEEE ICCV*, 2007.
[12] N. Mitra and A. Nguyen. Estimating surface normals in noisy point cloud data. In *ACM SCG*, 2003.
[13] R. Newcombe, S. Lovegrove, and A. Davison. DTAM: dense tracking and mapping in real-time. In *IEEE ICCV*, 2011.
[14] J. Ryde and M. Brunig. Non-cubic occupied voxel lists for robot maps. In *IEEE IROS*, 2009.
[15] C. Vogel, K. Schindler, and S. Roth. Piecewise rigid scene flow. In *IEEE ICCV*, 2013.
[16] Q. Wang, Z. Yu, C. Rasmussen, and J. Yu. Stereo vision based depth of field rendering on a mobile device. In *JEI*, 2014.
[17] Q. Zhang, M. Ye, R. Yang, Y. Matsushita, and B. Wilburn. Edge-preserving photometric stereo via depth fusion. In *IEEE CVPR*, 2012.