Dense 3D Indoor Scene Reconstruction from Spherical Images

Thiago L. T. da Silveira*
Center of Computational Sciences
Federal University of Rio Grande
Rio Grande, RS, Brazil
Email: tltsilveira@furg.br

Cláudio R. Jung
Institute of Informatics
Federal Univeristy of Rio Grande do Sul
Porto Alegre, RS, Brazil
Email: crjung@inf.ufrgs.br

***Abstract*—Techniques for 3D reconstruction of scenes based on images are popular and support a number of secondary applications. Traditional approaches require several captures for covering whole environments due to the narrow field of view (FoV) of the pinhole-based/perspective cameras. This paper summarizes the main contributions of the homonym Ph.D. Thesis, which addresses the 3D scene reconstruction problem by considering omnidirectional (spherical or 360°) cameras that present a $360^\circ \times 180^\circ$ FoV. Although spherical imagery have the benefit of the full-FoV, they are also challenging due to the inherent distortions involved in the capture and representation of such images, which might compromise the use of many well-established algorithms for image processing and computer vision. The referred Ph.D. Thesis introduces novel methodologies for estimating dense depth maps from two or more uncalibrated and temporally unordered 360° images. It also presents a framework for inferring depth from a single spherical image. We validate our approaches using both synthetic data and computer-generated imagery, showing competitive results concerning other state-of-the-art methods.**

I. Introduction

Image-based 3D scene reconstruction approaches have been widely studied by the scientific community, having applications in archaeological [1] and architectural modeling [2], robot navigation [3], autonomous driving systems [4], and infrastructure inspection [5], just to mention a few. Most existing techniques deal with the traditional pinhole-based/perspective cameras, which present a narrow field of view (FoV), and, hence, require several captures to model large scenes. Classically, multi-view stereo (MVS) approaches rely on calibrated image sets and produce dense 3D reconstructions [6], whereas structure from motion (SfM) and visual simultaneous localization and mapping (V-SLAM) methods estimate both the camera poses and the 3D geometry from video sequences but in a much sparser feature level [7].

In the past few years, omnidirectional (spherical or 360°) images and videos started to become popular thanks to the release of easy-to-use consumer-grade devices for acquisition and visualization, and the growing number of novel artificial, mixed, and virtual reality (AR/MR/VR) applications [8]. Oppositely to pinhole-based media, 360° imagery are intrinsically defined over the surface of the unit sphere, and thus, present a full $360^\circ \times 180^\circ$ FoV [9]. Fig. 1(a) illustrates such an example[1]. From the application point of view, in theory, a dense 3D representation of a given scenario (excepting for disocclusions) might be generated from two views only, as occurs in the traditional stereo matching case, simplifying the capture and the methods' pipelines.

(a) (b)

Fig. 1. Example of a 360° image mapped (a) onto the sphere and (b) to the plane (in equirectangular format).

Nonetheless, 360° media contain intrinsic distortions associated to the camera model that become apparent when projecting the sphere on the plane [11]. The most commonly adopted sphere-to-plane mapping is the so-called equirectangular format, depicted in Fig. 1(b). As one can see, the distortions appear more intensely near to the image poles; but they are also prominent on objects that are close to the camera [12]. Therefore, most of the algorithms developed so far by the scientific community may not be capable of performing the tasks they were designed for when applied to omnidirectional images and videos [8], [11], [13], [14].

The main goal of the referred Ph.D. Thesis was to build a method for *dense* depth estimation from indoor scenes based on two or more non-calibrated and temporally unordered 360° captures. In fact, we explore the flexibility of SfM/V-SLAM approaches but requiring a dense output, as done by MVS methods, while attaining the particularities of the spherical domain. The Ph.D. Thesis also introduced a pipeline for inferring depth from a single 360° image, taking advantage of existing techniques for perspective single-view depth inference. Here,

*This article summarizes the main contributions of the Ph.D. Thesis entitled "Dense 3D Indoor Scene Reconstruction from Spherical Images" authored by Thiago L. T. da Silveira and advised by Cláudio R. Jung.

[1]The real images in this manuscript were obtained from the publicly available SUN360 database [10].

we focus on indoor scenarios since every imaged point can be associated to a (scaled) physical distance, unlike outdoors.

The rest of this article is organized as follows. Section II briefly exposes the related works for 3D reconstruction that rely on spherical imagery only. The four main contributions of the Ph.D. Thesis associated to this manuscript are highlighted throughout the Section III. Some final remarks and future investigations are drawn in Section IV. Finally, Section V lists the published works related to the discussed Ph.D. Thesis.

II. Related Work

Classically, techniques for 3D geometry estimation are classified according to the number of views required. The same rule applies to the omnidirectional context. There are methods that tackle the problem using just one view (single-image stereo); a pair of captures (stereo matching); or multiple spherical images (SfM/V-SLAM or MVS).

From the geometrical point of view, at least two views from the same scene are required for estimating the 3D position of correspondent points [15]. However, deep learning approaches have allowed solutions for depth inference from a single image. In this context, we highlight pioneer studies like [16]–[18], in which the former is one of our contributions. Other recent works focus on identifying and reconstructing 3D layouts from a single image [19], [20], where only information about the joints of two or more planes are actually estimated. It is worth mentioning, however, that learning-based approaches that infer depth from a single image do not incorporate any real geometric constraint from multiple views. Also, they are strongly dependent on the training datasets, being or not accurate in general, unseen, contexts.

There are only few works [21]–[23] that estimate depth from pairs of views. The studies [21], [22] (and their prior analyses) estimate the relative pose between the cameras using traditional A-KAZE features [24] and the eight-point algorithm (8-PA) [25]; derotate and estimate dense features from the views; and refine the pose using an iterative non-linear approach. Lai and colleagues [23] present an encoder-decoder model that deals with stereo-rectfied image pairs with a small, fixed baseline. Their convolutional neural network, although not adapted to the spherical distortions, encourages the depth estimates from left and right boundaries to connect. Although it is possible to extract dense depth from two views only, stereo-based methods are much more prone to noise in feature matching and pose estimates than methods considering more captures. Thus, most methods work with many views.

The techniques proposed in [2], [26] work with stereo-rectified image pairs in an MVS context. The former work [26] introduces a time-consuming approach for disparity estimation based on hierarchical partial differential equations. From the disparities, a 3D mesh is constructed and aligned using an adaptation of the iterative closest point algorithm [27]. The latter study [2] considers that the scene can be modeled as a collection of blocks (Manhattan world). Their method segment planes, which are adjusted and registered together. Afterward, the resultant planes are refined and used in the cuboid representation of the scene. Although these methods deal with multiple views, they require stereo-rectified image pairs, which may jeopardize practical applications.

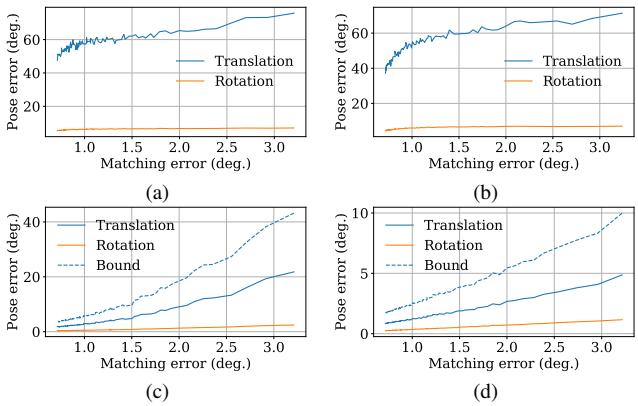


Fig. 2. 5-DoF pose error and bounds for different FoVs and noise levels. The graphics relate to (a)(b) two regular narrow-FoV cameras, (c) a camera with wide-angle fisheye lenses, and (d) a full-FoV 360° camera [30], [31].

Another possible limitation is the need of video sequences for estimating the 3D geometry. The methods from [1], [8], [28], [29] use traditional approaches – 8-PA and direct linear transform (DLT) [15] – for initial pose and 3D geometry estimation based on sparse keypoint matching. After linearly extracting the extrinsic parameters of the cameras – using either the 8-PA or the "spherical n-point problem" (SnP) [29] – it is common to apply a non-linear refinement technique to the pose estimates [1], [28], [29]. Furthermore, an iterative joint non-linear refinement of both the pose and the 3D geometry – bundle adjustment (BA) – is often considered [8], [28], [29]. Although effective, it is well known that BA approaches do not scale well with the number of cameras and imaged points [15]. As a final step, depending on the target application, a "densification" approach can be applied on the *sparse* depth map or 3D point cloud [1], [28]. Each method [1], [8], [28], [29] presents a slightly different approach from the others, and contributes to the literature of SfM/V-SLAM, extending the basic concepts of the classical pinhole imaging [15] to the spherical context.

Our main contributions, which are detailed throughout the next section, address some of the issues raised in this section, especially in the context of multiple views.

III. Contributions of the Ph.D. Thesis

We discuss the main contributions of the referred Ph.D. Thesis in the following. Section III-A presents a perturbation analysis of the traditional algorithm for relative pose estimation on wide-FoV imagery. Then, we use these results for proposing an approach that provides a dense 3D geometry estimate of indoor scenes based on multiple uncalibrated and unordered spherical images, described in Section III-B. Section III-C refers to a novel algorithm for 360° image oversegmentation and presents its applications to pose and depth estimation. Finally, in Section III-D, we introduce a framework for inferring depth from a single omnidirectional

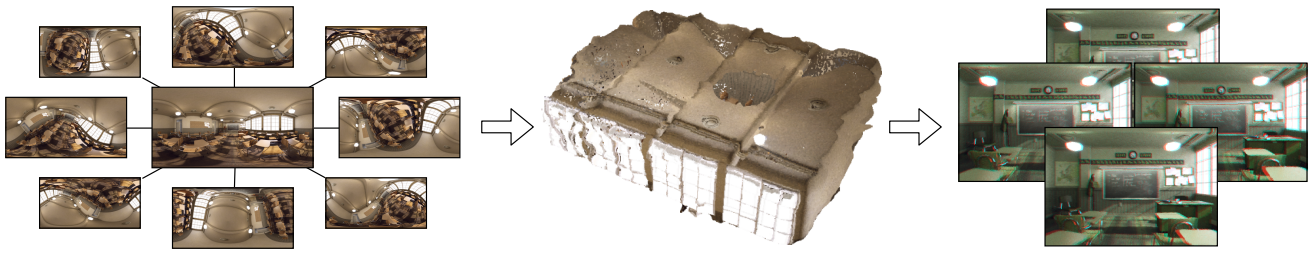


Fig. 3. An overview of the proposed multi-view method. Input reference and supporting images allow for dense 3D geometry estimation and, then, narrow-FoV view synthesis for 3-DoF+ VR applications.

image, which can be coupled to existing approaches for monocular depth prediction of planar images. We refer the reader to the complete text [31] for more details.

A. Perturbation Analysis for the Eight-Point Algorithm

This contribution presents both theoretical and experimental analyses for the estimate of Epipolar (Fundamental/Essential) matrices under noisy conditions using the popular 8-PA [25]. The 8-PA is a linear solution for estimating the relative pose between two cameras – the five-degrees of freedom (5-DoF) pose – through the epipolar constraint that relates a set of eight or more key-point correspondences. We rewrite the original 8-PA formulation to work with normalized homogeneous coordinates, making it applicable to every central projection camera, including both the pinhole and spherical camera models.

Then, our approach explores existing bounds for singular subspaces – namely Wedin's [32] and Merikoski, Sarria and Tarazagas' [33] bounds – and relates them to the 8-PA. We do not assume any error distribution for the matched features and do not require the noiseless measurement matrix to be known, unlike other recent approaches [34]. Specifically, we obtain that the sine error bound is inversely proportional to the second least singular value of the observation matrix, which is strongly affected by the spatial distribution of the correspondences.

Our experimental validation indicates that the bounds and effective errors tend to decrease as the camera FoV increases. In particular, the features extracted when using narrow-FoV images are spatially concentrated, leading to larger bounds, and, according to our experiments, also larger errors in the estimate of the Epipolar matrix. On the other hand, images with wider FoV (in the limit case, spherical images) tend to present a much better spatial distribution of features, leading to smaller bounds and also smaller effective errors in the estimated matrix.

Additionally, we present bounds for the unit translation vector extracted from the Essential matrix based on singular subspace analysis. We experimentally show that the rotation error associated to the Essential matrix is much smaller than the translation one, regardless of the FoV. The latter is typically proportional to the Epipolar matrix error, and dominates the camera pose errors. Fig. 2 illustrates the actual translation and rotation errors as a function of the matching error (in degrees) depending on the camera FoV. Translation error bounds are provided for wide-FoV camera cases. Also, we experimentally show that non-linear approaches for pose refinement tend to be much more effective when the matched features are concentrated, resembling the narrow-FoV scenario.

An empirical analysis of the impact of using either traditional planar or spherical feature extractors on pairs of 360° images in the context of pose estimation was documented in [35]. The theoretical analysis and the results associated with the contribution revisited in this section were published in [30].

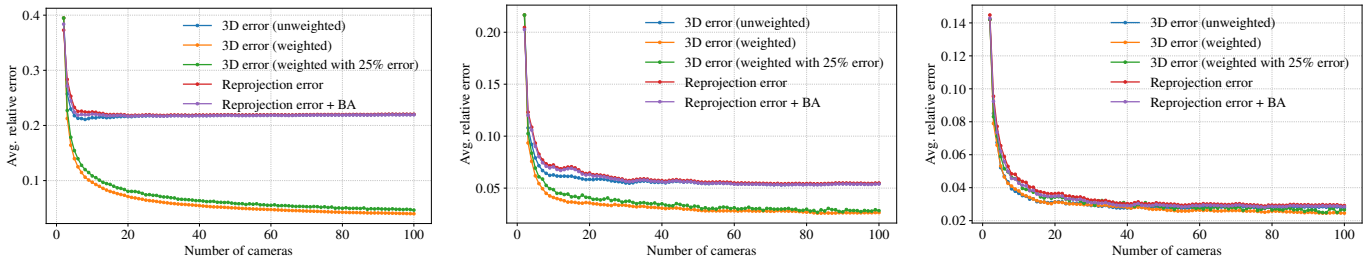
B. Dense 3D Reconstruction from Multiple Spherical Images

We use the results from Section III-A to introduce a linear approach for estimating a dense 3D representation of indoor scenes from multiple uncalibrated and temporally unordered spherical images. The method estimates a depth value for each pixel (in equirectangular format) of a reference view, relying on one or more additional views. Fig. 3[2] illustrates the inputs (reference and supporting views) and possible outputs provided by our technique.

Firstly, our approach computes and matches sparse spherical ORB (SPHORB) features [36] that relate the reference view to the others, and estimates the 5-DoF camera poses using the 8-PA supported by outlier removal [37]. We found in [35] that SPHORB performs well and scatters the features on the image pairs, occupying a large region, regardless of the distortions. Our method benefits from the small rotation errors associated with wide-FoV imagery (confer Section III-A) to derotate – i.e., remove the relative rotation of – all supporting images.

Then, the proposed approach uses an optical flow algorithm to obtain dense matches, attaining for the circularity property of equirectangular images. We choose to use the DeepFlow [38] algorithm, which performs well under large-displacement and (affine) distortion conditions. Although DeepFlow was not designed for dealing with spherical imagery, we can somehow identify good and bad correspondences and treat them accordingly. More precisely, we propose to explore a joint photometric-geometric confidence metric that uses cross-checking and epipolar geometry consistency to detect and remove the contribution of inconsistent flow vectors. One may note that dense correspondences can be naturally converted to dense depth maps, as required.

[2]The "Classroom" environment is fully available under license CC0 at https://www.blender.org. Color images and depth maps are rendered using Blender.



(a) (b) (c)

Fig. 4. Relative error associated to different approaches for calibrated reconstruction as the number of views increases. The graphics relate to different angular noise levels on feature matching: (a) 7.25°, (b) 3.20°, and (c) 2.27° [31].

A subset of the best dense matches is used to accurately estimate the full 6-DoF pose (SnP problem) of each supporting view using a simplified linear approach that optimizes only the remaining translation-related 3-DoF (recalling that the images are derotated). Then, we estimate the 3D scene geometry from the (now) *calibrated* cameras and dense matches by minimizing a weighted error in the 3D space. The proposed weighting scheme benefits from an initial (unweighted) estimate of the geometry, and performs as good as the reprojection error (even when refining with BA) under low feature matching error (around 2.27° [13]). Our calibrated reconstruction algorithm, although, performs the best in higher noise conditions, as depicted in Fig. 4. Finally, we adapt the domain transform [39], an image-guided filter, to the spherical domain for imposing edge-aware spatial smoothness, taking advantage of the fact that the depth map and reference image are registered. Different depth estimates for the image set shown in Fig. 3 (left) are illustrated in Fig. 5. The point cloud associated to the final depth map from Fig. 5 is shown in Fig. 3 (middle).

We further explore the output of our method to implement a simplified depth-image-based rendering (DIBR) technique for synthesizing coherent binocular stereo pairs and small head motion parallax. This is the application context for 3-DoF+ VR immersive exploration using head-mounted displays (HMDs), which is implementable with the aligned color plus depth representation. More precisely, we project the synthesized views to the user's HMD viewport and fill the small holes that come from the occlusions or disocclusions using the fast hierarchical hole filling (HHF) [40]. Some visual results are shown in cyan-red anaglyphs in Fig. 3 (right).

The results from this contribution showed that it is possible to propose a method that naturally produces a dense depth map even in uncalibrated and unordered 360° camera setups. For acquiring dense correspondences, we can indeed use a traditional large-displacement optical-flow algorithm provided that they are assessed for quality and properly weighted. Also, our results showed that we can rely on linear approaches only for 5-/6-DoF pose estimation and for (weighted) calibrated 3D reconstruction, still having competing results with state-of-the-art methods that use traditional, but expensive, non-linear approaches based on BA. The methodology and results reviewed in this section were published in [41]. An exploratory study about stereoscopy and DIBR techniques was published in [42].

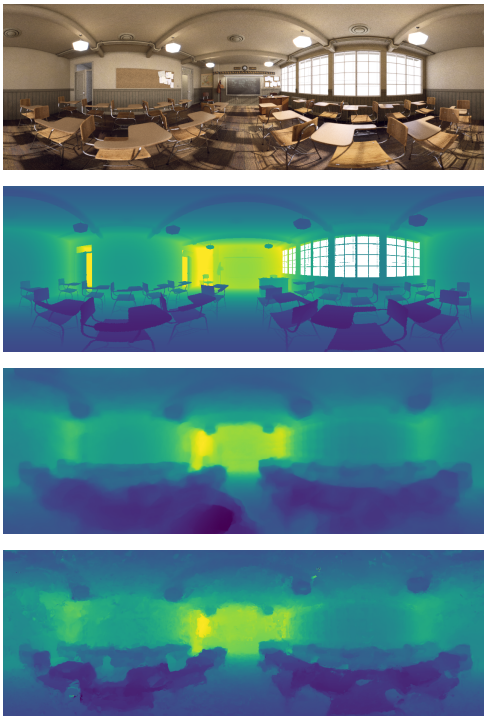


Fig. 5. Example of depth estimation based on the views in Fig 3. From top to bottom: reference view, ground-truth and estimated depth maps using the unweighted and weighted approaches [31], [41].

C. Spherical Superpixels and Applications to 6-DoF Pose and Multi-View Depth Estimation

The contributions of this section are three-folded. Firstly, we present a novel superpixel algorithm suited for omnidirectional images that encourages the segments to adhere to borders and keep a regular size on the sphere. Because this novel algorithm extends the simple non-iterative clustering (SNIC) [43] to the spherical domain, we name it as spherical SNIC (SSNIC). Fig. 6 illustrates the application of both algorithms – SNIC and SSNIC – to a 360° image. Similarly to the spherical SLIC [44], spherical SNIC is applicable to any image oversegmentation problem defined in the spherical domain. Our algorithm, although, performs faster than spherical SLIC, and generates

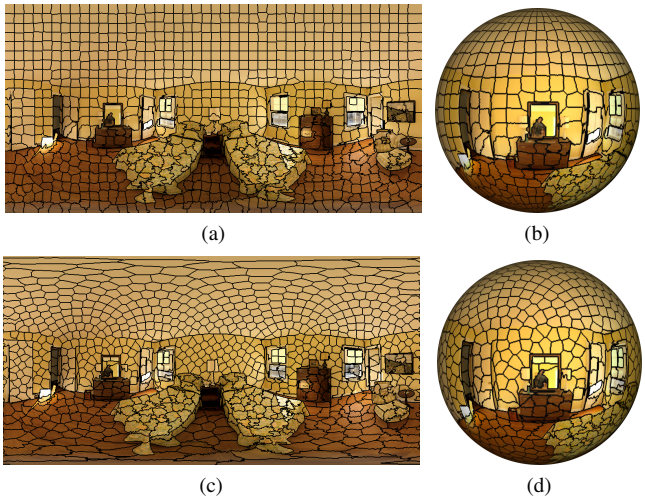


Fig. 6. Example of a 360° image segmented using (a)(b) SNIC and (c)(d) SSNIC. Images shown in (a)(c) equirectangular and (b)(d) spherical formats.

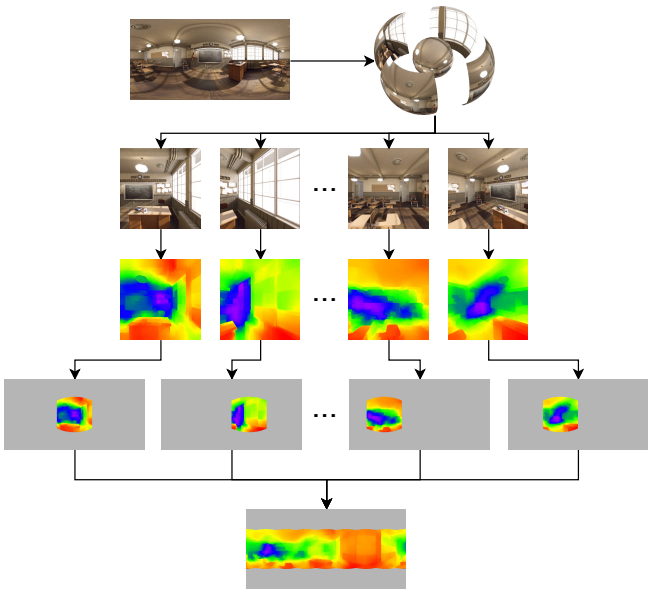


Fig. 7. Pipeline of the proposed approach for inferring depth from a single spherical image.

nearly uniformly distributed points on the sphere. SSNIC's software registration is available at [45].

Based on the results from Section III-A and SSNIC properties, we conduct an experimental analysis for determining if the linear solutions for the SnP problem are affected by the distribution of the features. Synthetic feature matching experiments confirm our hypothesis, indicating that scattered correspondences allow for better 6-DoF pose estimates than concentrated ones. Thus, we propose to select representative matches from the SSNIC labels using the joint confidence metric from Section III-B, so that we enforce nearly-uniform spatial distribution. However, in our experiments using real feature matching, this approach did not improve the results obtained by the unconstrained correspondence selection from Section III-B. The reason for this may be that most of the "good matches" are along the entire image equator line, which spans a large horizontal FoV.

Finally, we propose to enforce spatial consistency *a priori*, i.e., during depth estimation, using SSNIC segments. Unlike the approach from Section III-B that estimates all depth values separately and applies an image guided filter in a post-processing step, here, we propose to enforce depth consistency within semantically grouped SSNIC regions. Our experimental results indicate that this approach can be effective for removing incoherent matches within superpixels, and it presented smaller average errors when estimating the 3D scene geometry. We also still obtain small gains when combining both *a priori* and *a posteriori* solutions.

D. Dense 3D Reconstruction from a Single Spherical Image

This additional contribution provides a framework for inferring depth from a single spherical image, so that it can be coupled to any existing single-view depth prediction method suited for perspective images [46], [47]. Single-image depth estimation is clearly an ill-posed problem, so that the baseline methods inevitably rely on machine learning for inference. This framework is intended to be used only in *extreme* cases, i.e., when it is not possible to capture more than one image. We depict the overview of the method in Fig. 7.

Our framework starts by extracting multiple *overlapping* tangent planar projections with smaller FoVs from the spherical image. A larger FoV implies in more contextual information, but also heavy distortions. This is a trade-off that impacts the rest of the pipeline. In our experiments, a diagonal FoV of 120° produced the best results. After extracting the narrow-FoV images from the spherical image, we apply a monocular planar depth estimation algorithm – as a black box – to each of them. Once it is done, we back-project the associated depth maps to the adequate locations on the sphere. In a third moment, we minimize the depth discrepancies along the pairwise intersections on the sphere and finally perform alpha-blending to obtain the final spherical depth map.

We perform tests by plugging three, to the time, state-of-the-art baseline algorithms, two of them from [47] and the other from [46], and compared the results on both synthetic and real spherical imagery. The results indicated that our approach outperforms two common strategies for adapting planar methods to the spherical domain: the application of a planar method (i) directly to equirectangular images or (ii) to multiple disjoint planar sections, mapped back to the sphere. Fig. 8 exemplifies (i), (ii) and our approach, using the method from [46] as the module for narrow-FoV single view depth estimation. The methodology and results discussed in this section were previously published in [16].

IV. Final Considerations

This article highlighted the main contributions of the Ph.D. Thesis entitled "Dense 3D Indoor Scene Reconstruction from

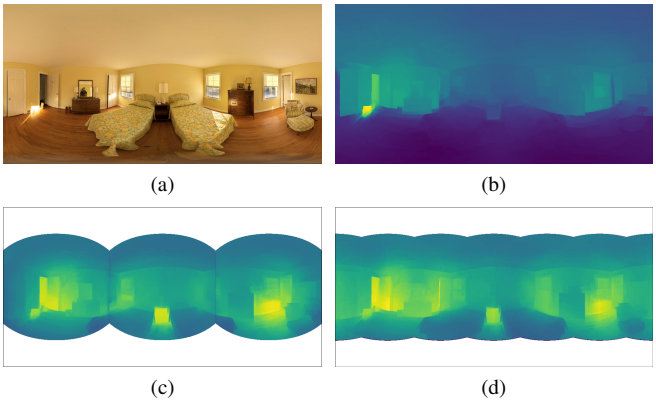


Fig. 8. Example of single spherical image depth estimation. Depth estimates obtained from (a) after applying [46] to (b) the full equirectangular image, (c) disjoint spherical sections and (d) overlapping spherical sections (and the post-processing described in the text) [16], [31].

Spherical Images", authored by Thiago L. T. da Silveira and advised by Cláudio R. Jung, which is fully accessible in [31].

We do believe that this work has advanced important theoretical and practical contributions to the image processing and computer vision fields. More precisely, we have addressed the pose and depth estimation problems using a single view, stereoscopic captures, and multiple uncalibrated and temporally unordered spherical images. Last but not least, we have advanced in related areas such as key-point matching, image oversegmentation, edge-aware filtering, depth-image-based rendering, etc.

We hope that the referred Ph.D. Thesis could be recognized as a solid starting point for future researches on dense 3D reconstruction based on spherical imagery, besides novel AR/MR/VR-related applications. In the future, we intend to better explore the use of spherical regions for spatially-consistent depth estimation, and incorporate monocular layout or depth estimates to the multi-view pipeline.

V. Intellectual Production

The contributions of this Ph.D. Thesis, published between 2016 and 2019, are highlighted in the following.

- Software registration **[45]** granted by ***Instituto Nacional de Propriedade Industrial (INPI).***
- Article **[30]** published in ***IEEE/CVF CVPR*** (Qualis A1). According to Google Scholar, CVPR holds the higher h5-index (240) among all conferences and journals in computer science[3]. In 2019, CVPR's acceptance rate was of 25.2%.
- Article **[41]** published in ***IEEE VR*** (Qualis A1). VR is the premier international conference on virtual reality and 3D user interfaces. In 2019, VR's acceptance rate was of 21.5%.
- Article **[42]** published in ***IEEE ICASSP*** (Qualis A1). ICASSP is the premier international conference on signal processing.

[3]Confer https://scholar.google.com/citations?view_op=top_venues\&hl=en

- Article **[16]** published in ***IEEE ICIP*** (Qualis A1). ICIP is the premier international conference on image processing.
- Article **[35]** published in ***SIBGRAPI Conference*** (Qualis B1). SIBGRAPI is the annual brazilian conference on graphics, patterns and images.

Acknowledgment

The first author thanks to the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001 - and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the scholarships received.

References

[1] A. Pagani, C. Gava, Y. Cui, B. Krolla, J.-M. Hengen, and D. Stricker, "Dense 3D Point Cloud Generation from Multiple High-resolution Spherical Images," *International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST)*, pp. 1–8, 2011.

[2] H. Kim and A. Hilton, "Block world reconstruction from spherical stereo image pairs," *Computer Vision and Image Understanding*, 2015.

[3] J. Moreau, S. Ambellouis, and Y. Ruiche, "3D reconstruction of urban environments based on fisheye stereovision," in *8th International Conference on Signal Image Technology and Internet Based Systems, SITIS 2012r*, 2012.

[4] J. Tong and X. Ning, "Depth measurement by omni-directional camera," in *2013 IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, IEEE-CYBER 2013*, 2013.

[5] S. Pathak, A. Moro, A. Yamashita, and H. Asama, "Dense 3D reconstruction from two spherical images via optical flow-based equirectangular epipolar rectification," in *IEEE International Conference on Imaging Systems and Techniques (IST)*, 2016, pp. 140–145.

[6] "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR'06)*, vol. 1. IEEE, 2006, pp. 519–528.

[7] O. Özyeşil, V. Voroninski, R. Basri, and A. Singer, "A survey of structure from motion." *Acta Numerica*, vol. 26, pp. 305–364, 2017.

[8] S. Im, H. Ha, F. Rameau, H.-G. Jeon, G. Choe, and I. S. Kweon, *All-Around Depth from Small Motion with a Spherical Panoramic Camera*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham, 2016, vol. 9907.

[9] T. Akihiko, I. Atsushi, and N. Ohnishi, "Two-and three-view geometry for spherical cameras," *Proc. of the Sixth Workshop on Omnidirectional Vision, Camera Networks and Non- classical Cameras*, vol. 105, pp. 29–34, 2005.

[10] J. Xiao, K. A. Ehinger, A. Oliva, and A. Torralba, "Recognizing scene viewpoint using panoramic place representation," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2012, pp. 2695–2702.

[11] Y.-C. Su and K. Grauman, "Learning Spherical Convolution for Fast Features from 360 Imagery," in *Conference on Neural Information Processing Systems*, 2017, pp. 529–539.

[12] J. Cruz-Mota, I. Bogdanova, B. Paquier, M. Bierlaire, and J. P. Thiran, "Scale invariant feature transform on the sphere: Theory and applications," *International Journal of Computer Vision*, vol. 98, no. 2, pp. 217–241, 2012.

[13] H. Guan and W. A. P. Smith, "BRISKS: Binary Features for Spherical Images on a Geodesic Grid," in *Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1–9.

[14] R. G. d. A. Azevedo, N. Birkbeck, F. De Simone, I. Janatra, B. Adsumilli, and P. Frossard, "Visual Distortions in 360-degree Videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. c, pp. 1–1, 2019.

[15] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, 2003.

[16] T. L. T. da Silveira, L. Dal'aqua, and C. R. Jung, "Indoor Depth Estimation From Single Spherical Images," in *IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 2935–2939.

[17] N. Zioulis, A. Karakottas, D. Zarpalas, and P. Daras, "OmniDepth: Dense Depth Estimation for Indoors Spherical Panoramas," 2018, pp. 453–471.

[18] M. Eder and J.-M. Frahm, "Convolutions on Spherical Images," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, may 2019.

[19] C. Zou, A. Colburn, Q. Shan, and D. Hoiem, "LayoutNet: Reconstructing the 3D Room Layout from a Single RGB Image," 2018.

[20] S.-T. Yang, F.-E. Wang, C.-H. Peng, P. Wonka, M. Sun, and H.-K. Chu, "DuLa-Net: A Dual-Projection Network for Estimating Room Layouts from a Single RGB Panorama," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3363–3372.

[21] S. Pathak, A. Moro, H. Fujii, A. Yamashita, and H. Asama, "Virtual Reality with Motion Parallax by Dense Optical Flow-Based Depth Generation from Two Spherical Images," pp. 1–6, 2017.

[22] S. Pathak, A. Moro, A. Yamashita, and H. Asama, "Optical Flow-Based Epipolar Estimation of Spherical Image Pairs for 3D Reconstruction," *SICE Journal of Control, Measurement, and System Integration*, vol. 10, no. 5, pp. 476–485, 2017.

[23] P. K. Lai, S. Xie, J. Lang, and R. Laqaruere, "Real-Time Panoramic Depth Maps from Omni-directional Stereo Images for 6 DoF Videos in Virtual Reality," *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 405–412, 2019.

[24] P. Alcantarilla, J. Nuevo, and A. Bartoli, "Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces," *Procedings of the British Machine Vision Conference 2013*, pp. 13.1–13.11, 2013.

[25] H. C. Longuet-Higgins, "A Computer Algorithm for Reconstructing a Scene from Two Projections," in *Readings in computer vision: issues, problems, principles, and paradigms*, 1987, pp. 61–62.

[26] H. Kim and A. Hilton, "3D scene reconstruction from multiple spherical stereo pairs," *International Journal of Computer Vision*, 2013.

[27] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Transactions Pattern Analasys and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992. [Online]. Available: http://dx.doi.org/10.1109/34.121791

[28] J. Huang, Z. Chen, D. Ceylan, and H. Jin, "6-DOF VR videos with a single 360-camera," in *IEEE Virtual Reality (VR)*, 2017, pp. 37–44.

[29] H. Guan and W. A. P. Smith, "Structure-From-Motion in Spherical Video Using the von Mises-Fisher Distribution," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 711–723, feb 2017.

[30] T. L. T. da Silveira and C. R. Jung, "Perturbation Analysis of the 8-Point Algorithm: a Case Study for Wide FoV Cameras," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 757–11 766.

[31] T. L. T. da Silveira, "Dense 3d indoor scene reconstruction from spherical images," Ph.D. dissertation, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2019, available at https://lume.ufrgs.br/handle/10183/202142.

[32] P.-Å. Wedin, "Perturbation bounds in connection with singular value decomposition," *BIT Numerical Mathematics*, vol. 12, no. 1, pp. 99–111, 1972.

[33] J. K. Merikoski, H. Sarria, and P. Tarazaga, "Bounds for singular values using traces," *Linear Algebra and its Applications*, vol. 210, pp. 227–254, 1994.

[34] T. T. Cai, A. Zhang *et al.*, "Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics," *The Annals of Statistics*, vol. 46, no. 1, pp. 60–89, 2018.

[35] T. L. T. da Silveira and C. R. Jung, "Evaluation of Keypoint Extraction and Matching for Pose Estimation Using Pairs of Spherical Images," *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 374–381, 2017.

[36] Q. Zhao, W. Feng, L. Wan, and J. Zhang, "SPHORB: A Fast and Robust Binary Feature on the Sphere," *International Journal of Computer Vision*, vol. 113, no. 2, pp. 143–159, 2014.

[37] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[38] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "DeepFlow: Large displacement optical flow with deep matching," *Proceedings of the IEEE International Conference on Computer Vision*, no. Section 2, pp. 1385–1392, 2013.

[39] E. S. L. Gastal and M. M. Oliveira, "Domain transform for edge-aware image and video processing," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 69:1–69:12, 2011.

[40] M. Solh and G. AlRegib, "Hierarchical hole-filling for depth-based view synthesis in ftv and 3d video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 5, pp. 495–504, 2012.

[41] T. L. T. da Silveira and C. R. Jung, "Dense 3D Scene Reconstruction from Multiple Spherical Images for 3-DoF+ VR Applications," in *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2019, pp. 9–18.

[42] A. Q. de Oliveira, T. L. T. da Silveira, M. Walter, and C. R. Jung, "On the performance of DIBR methods when using depth maps from state-of-the-art stereo matching algorithms," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 2272–2276.

[43] R. Achanta and S. Süsstrunk, "Superpixels and polygons using simple non-iterative clustering," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017.

[44] Q. Zhao, F. Dai, Y. Ma, L. Wan, J. Zhang, and Y. Zhang, "Spherical Superpixel Segmentation," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1406–1417, 2018.

[45] T. L. T. da Silveira and C. R. Jung, "Snics: um método para supersegmentação de imagens esféricas," 2019, instituto Nacional de Propriedade Industrial (INPI). Número do registro: BR5120190029348.

[46] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2024–2039, oct 2016.

[47] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised Monocular Depth Estimation with Left-Right Consistency," in *Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279.