Photometric Gaussian Mixtures for Direct Virtual Visual Servoing of Omnidirectional Camera

Seif-Eddine Guerbas, Nathan Crombez, Guillaume Caron, El Mustapha Mouaddib

Photometric Gaussian Mixtures for Direct Virtual Visual Servoing of Omnidirectional Camera

Seif Eddine Guerbas[1], Nathan Crombez[2], Guillaume Caron[1,3] and El Mustapha Mouaddib[1]
[1]Université de Picardie Jules Verne, MIS laboratory, 33 rue Saint Leu, 80039 Amiens Cedex 1, France
[2]Université Bourgogne Franche-Comté, UTBM, CIAD Laboratory, 90010 Belfort, France
[3] CNRS/AIST JRL, Tsukuba Central 1, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8560, Japan
{seif-eddine.guerbas, guillaume.caron, mouaddib}@u-picardie.fr, nathan.crombez@utbm.fr

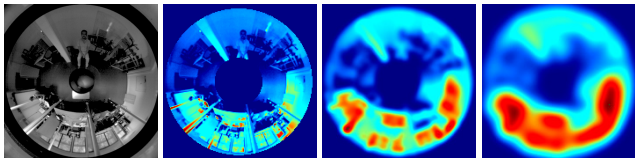
Abstract

This paper tackles direct 3D model-based pose tracking. It considers the Photometric Gaussian Mixtures (PGM) transform of omnidirectional images as direct features. The contributions include an adaptation of the pose optimization to omnidirectional cameras and a rethink of the initialization and optimization rules of the PGM extent. These enhancements produce a giant leap in the convergence domain width. Application to images acquired onboard a mobile robot within an urban environment described by a large 3D colored point cloud shows significant robustness to large inter-frame motion, compared to approaches that directly use pixel brightness as direct features.

1. Introduction

Direct image alignment [1] and direct visual servoing [2] (DVS) have significantly progressed during the last decade in their respective communities, namely computer vision and robotics. The best known direct approaches concern visual odometry [3] and visual Simultaneous Localization And Mapping [4] (SLAM). For a while, direct approaches were known to save time by avoiding features processing and to be of high accuracy whereas suffering of a narrow convergence domain [5]. While usually overcome by encapsulation in a pyramidal scheme [6], the latter narrowness was recently enlarged intrinsically by direct approaches relying on transforms of images: scale space [7], frequency domain [8], photometric moments [9] or Photometric Gaussian Mixtures [10] (PGM). The latter optimizes the Gaussian extent of PGMs (Fig. 1 shows its impact on PGM smoothness) together with camera pose degrees-of-freedom. This allows to significantly enlarge the convergence domain of DVS with conventional camera.

This paper investigates the application of PGM to omnidirectional (panoramic) vision. The motivation comes from the fact that the wider field of view of an omnidirectional camera compared to a conventional camera (perspective-like) allows more reliable 3D motion estimation [11] and higher localization rates [12]. This is due to the possibility to capture information better spread around the viewpoint. This paper not only considers the PGM of omnidirectional images but its use in a direct approach for camera and robot localization with respect to a 3D model of the environment. Interestingly, the bridge from Visual Servoing (VS) to the full scale alignment of an image on 3D model is well established as Virtual Visual Servoing [13]. It was applied to direct 3D model-based pose tracking in omnidirectional images from pixel brightness [14, 15], making full benefit of environment 3D point cloud with colors for robot localization, though sensitive to the amount of inter-frame motion.



(a) Image $\mathbf{I}$ (b) $\lambda = 0.1$ (c) $\lambda = 3$ (d) $\lambda = 7$

Figure 1: (a) An acquired image $\mathbf{I}$ and (b-d) PGMs $\mathbf{G}(\mathbf{I}, \lambda)$ for various extents λ.

Since it is hard to find more recent works of omnidirectional image direct alignment with a 3D model of an environment, [14, 15] are considered as baselines. Despite the variety of contributions in the field of neural networks, only conventional or rectified images are considered as input of pose detection [16] approaches. One could generate conventional images from omnidirectional ones [17] to feed the latter methods but, in this paper, we focus on using acquired images directly without geometric pre-transformation. This way, one can think about considering large scale direct visual SLAM with omnidirectional images [18] but none implement yet localization within a pre-built map. The latter functionality is handled by handcrafted feature-based ap-

proaches. But even the state-of-the-art ORB-based visual SLAM [19] fails in localizing acquired omnidirectional images in a map that has been pre-built from another camera, thus making hard to share the map and requiring additional sensors to estimate the full scale 3D pose of the camera.

The contributions of this paper benefit from both the use of a pre-built 3D model of an environment and the properties of omnidirectional vision and are summarized as:

- a 3D model-based visual tracking approach robust to very large inter-frame motion;
- a new solution of initialization and optimization of the PGM extent, enlarging the convergence domain;
- PGM adaptation to the omnidirectional camera.

The rest of the paper shortly describes the PGM-based omnidirectional visual servoing, focusing on the contribution regarding the new rules of initialization and optimization of the Gaussian extent of PGMs. Then, Section 3 reports their evaluation in a virtual environment made of 3D scans of streets. Finally, Section 4 presents preliminary results of direct 3D model-based tracking in omnidirectional images transformed as PGMs, before conclusion (Sec. 5).

2. PGM for Omnidirectional Visual Servoing

2.1. Photometric Gaussian Mixture

From an omnidirectional image $\mathbf{I}$ with $M \in \mathbb{N}$ pixels, we express its PGM in the exact same way as [10] did for conventional images, *i.e.*, as a mixture of M Gaussians, sharing a unique Gaussian extent parameter $\lambda \in \mathbb{R}^*_+$, weighted by brightness $I(\mathbf{u})$, for the M pixels of coordinates $\mathbf{u} = (u, v) \in \Omega \subset \mathbb{N}^2$ of the image. To distinguish image coordinates from PGM coordinates, we write the latter $\mathbf{u_g} = (u_g, v_g) \in \Omega$, leading to express a PGM sample as:

$$G(\mathbf{u_g}, \mathbf{I}, \lambda) = \sum_{\mathbf{u}} I(\mathbf{u}) \exp\left(-\frac{(u_g - u)^2 + (v_g - v)^2}{2\lambda^2}\right). \quad (1)$$

The PGM of an image $\mathbf{I}$ is written $G(\mathbf{I}, \lambda)$, for compactness. Figure 1 shows the impact of λ on the PGM.

2.2. PGM-based omnidirectional visual servoing

Visual servoing is similar to a Gauss-Newton optimization that computes camera pose increments $\mathbf{v} \in \mathbb{R}^6$, minimizing the error between a reference (desired) image $\mathbf{I}^*$ and the one to align, namely the current image $\mathbf{I}$. $\mathbf{I}$ is acquired at pose $\mathbf{r} = (t_X, t_Y, t_Z, \theta w_X, \theta w_Y, \theta w_Z) \in \mathbb{R}^6$, with $||[w_X, w_Y, w_Z]|| = 1$ and $\theta \in [-\pi, \pi]$ representing the rotation as axis-angle. $\mathbf{I}^*$ is acquired at pose $\mathbf{r}^* \in \mathbb{R}^6$. Then, highlighting the dependence of PGMs to the camera pose as $G(\mathbf{r}, \mathbf{I}, \lambda)$ for $\mathbf{I}$ and $G(\mathbf{r}^*, \mathbf{I}^*, \lambda^*)$ for $\mathbf{I}^*$, and by stacking all their samples (1) as, respectively, $\mathbf{G}(\mathbf{r}, \lambda) \in \mathbb{R}^M$ and $\mathbf{G}^* \in \mathbb{R}^M$, we express the error vector to regulate to zero:

$$\mathbf{e}(\mathbf{r}, \lambda) = \mathbf{G}(\mathbf{r}, \lambda) - \mathbf{G}^* \in \mathbb{R}^M. \quad (2)$$

In (2), the Gaussian extent λ is variable while λ^* of $\mathbf{G}^*$ is not [10], thus they are possibly different. λ is not constant because it is optimized in addition to $\mathbf{r}$ by a Gauss-Newton method computing iteratively increments as [10]:

$$[\mathbf{v}^T, \dot{\lambda}]^T = -\mu\, [\mathbf{L_G} \;\; \mathbf{J}_\lambda]^+ \mathbf{e}(\mathbf{r}, \lambda), \quad (3)$$

where $[\,]^+$ is the pseudo-inverse operator, $\mathbf{L_G} \in \mathbb{R}^{M \times 6}$ is the interaction matrix related to $G(\mathbf{r}, \mathbf{I}, \lambda)$ at pose $\mathbf{r}$ and $\mathbf{J}_\lambda \in \mathbb{R}^{M \times 1}$ is the Jacobian of $G(\mathbf{r}, \mathbf{I}, \lambda)$ with respect to λ. $\mathbf{J}_\lambda$ is the same as in [10]. However, $\mathbf{L_G}$ is now expressed for the unified central camera projection model (UCM) [20] instead of the perspective one used in [10]. Considering intrinsic parameters $\alpha_u \in \mathbb{R}^*$, $\alpha_v \in \mathbb{R}^*$ as the generalized focal length, $u_0 \in \mathbb{R}$, $v_0 \in \mathbb{R}$ as the principal point coordinates and $\xi \in \mathbb{R}$ as the mirror shape parameter, the UCM relates 3D points $\mathbf{X} = [X, Y, Z]^T \in \mathbb{R}^3$ to digital image points $\mathbf{u_g}$ as:

$$u_g = \alpha_u x_g + u_0 \text{ and } v_g = \alpha_v y_g + v_0, \quad (4)$$

with $x_g = X/(Z + \xi\rho)$, $y_g = Y/(Z + \xi\rho)$ and $\rho = \sqrt{X^2 + Y^2 + Z^2}$. Then, each line $\mathbf{L}_G$ of $\mathbf{L_G}$ is expressed as:

$$\mathbf{L}_G = \left(\sum_{\mathbf{u}} \nabla_{\mathbf{u_g}} G\right) \begin{bmatrix} \alpha_u & 0 \\ 0 & \alpha_v \end{bmatrix} \mathbf{L}_{\mathbf{x_g}}, \quad (5)$$

where:

$$\mathbf{L}_{\mathbf{x_g}} = \begin{bmatrix} -\frac{1+x_g^2(1-\xi(\gamma+\xi))+y_g^2}{\rho(\gamma+\xi)} & \frac{\xi x_g y_g}{\rho} \\ \frac{\xi x_g y_g}{\rho} & -\frac{1+y_g^2(1-\xi(\gamma+\xi))+x_g^2}{\rho(\gamma+\xi)} \\ \frac{\gamma x_g}{\rho} & \frac{\gamma y_g}{\rho} \\ x_g y_g & \frac{(1+y_g^2)\gamma - \xi x_g^2}{\gamma+\xi} \\ -\frac{(1+x_g^2)\gamma - \xi y_g^2}{\gamma+\xi} & -x_g y_g \\ y_g & -x_g \end{bmatrix}^T, \quad (6)$$

with $\gamma = \sqrt{1 + (1-\xi^2)(x_g^2 + y_g^2)}$ [14]. This is the key difference in the computation of $\mathbf{L_G}$ compared to [10].

2.3. The two stages strategy: new rules

In (3), λ is optimized for the ideal behavior of PGM VS [10], *i.e.*, a large convergence domain (large λ) and a high precision at convergence (λ tends to λ^*, set small). To achieve this ideal behavior, [10] reports a sequence of two PGM VS, that we name here *Rule0*: Step 1 with a large λ^* (exact value depends on experiments) and $\lambda = \alpha\lambda^*$, with $\alpha = 2$, at initialization; Step 2 with constant $\lambda = \lambda^* = 1$.

In our experimental convergence study (Sec. 3), we observed that setting $\lambda = 2\lambda^*$ at the initialization of Step 1 may lead to unexpected divergence. As such setting can lead to a current and a desired PGM of very different orders of magnitude, we assume it is the cause of the problem.

To simplify, we propose to remove factor α and set $\lambda = \lambda^*$ large at the initialization of Step 1. We name this

new rule *Rule1*. Then, we define *Rule2* that keeps Step 2 with $\lambda = \lambda^* = 1$ (or smaller) but we relax the constancy of λ for more coherence with respect to Step 1, more freedom and to use the exact same control law (3). The following Section 3 validates the adaptation of the PGM to omnidirectional cameras and evaluates the latter introduced rules.

3. Evaluation in a virtual environment

The virtual environment is a point cloud of four streets in the city of Amiens, France. The point cloud is a registration of thirteen 3D scans with Red-Green-Blue photographic colors, all acquired with a Lidar scanner Faro Focus 3D. The virtual camera simulates the UCM (Sec. 2.2) implemented with a vertex shader in the Unity 3D software (`http://unity.com`) bridged to our C++ implementation. Camera intrinsic parameters match those of a real catadioptric camera (Sec. 4) calibrated classically by observing known chessboards [21].

This evaluation compares the PGM omnidirectional VS (PGMoVS) with *Rule0*, *Rule1* and *Rule2*, and the seminal Photometric omnidirectional VS [14] (PoVS). Only virtual images are considered for fair quantitative comparison as previous works of visual odometry evaluations did [11].

3.1. Protocol

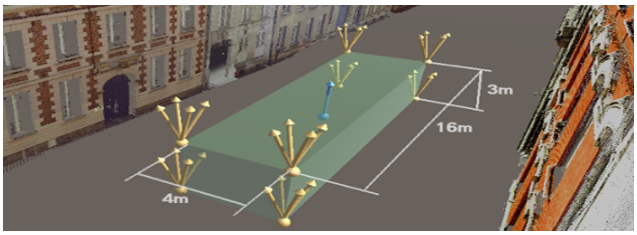
In order to evaluate the convergence domain, 64 initial poses $\mathbf{r}$ are generated around various desired poses $\mathbf{r}^*$ with combinations of transformations $t_X = \{-8\text{m}, 8\text{m}\}$, $t_Y = \{-2\text{m}, 2\text{m}\}$, $t_Z = \{-1.5\text{m}, 1.5\text{m}\}$, $\theta w_X = \{-10°, 10°\}$, $\theta w_Y = \{-10°, 10°\}$ and $\theta w_Z = \{-15°, 15°\}$. The initial positions form together a volume included in streets of about 12m width (Fig. 2a). For variety, we consider 7 desired poses $\mathbf{r}^*$ spread within the 3D model (Fig. 2b-2h).

For PGMoVS, we limit Step 1 to 120 iterations, 250 in total for both steps. For all three rules, $\lambda = 15$ in order to have sufficient overlapping area between $\mathbf{G}$ and $\mathbf{G}^*$. The control gain μ in (3) is set to 0.2 for every VS.

3.2. Results

Figure 3 shows the cumulative distribution of the 448 final position errors for each of the four VS compared. Setting a convergence threshold equal to 2cm, *Rule2* achieves a 97% success rate while *Rule1* and *Rule0* respectively achieve 78% and 70%. For *Rule1* and *Rule0*, Step 2 may converge to a local minimum, whereas for *Rule2* most of the local minima are removed, allowing to reach the global minimum. In contrast, all the PoVS diverge, which is not surprising as initial errors allowing PoVS to converge are known to be below 1.3m [14] whereas initial errors are greater than 8m in this evaluation.

These results show that the new *Rule2* significantly outperforms the previous *Rule0* and is hence, considered for the next experiment of 3D tracking (Sec. 4).



(a)

(b) (c) (d) (e)

(f) (g) (h)

Figure 2: Convergence domain evaluation: (a) Visualization of a parallelepiped formed by 64 initial camera poses (gold arrows) around a desired one (blue arrow); (b-h) Images rendered at the seven desired poses.

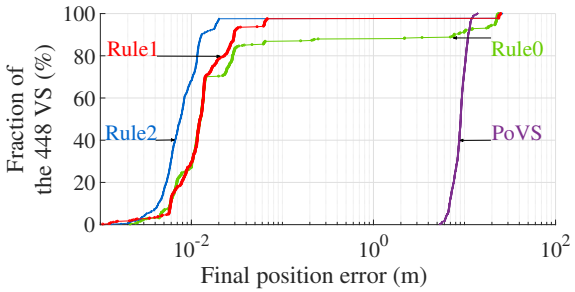
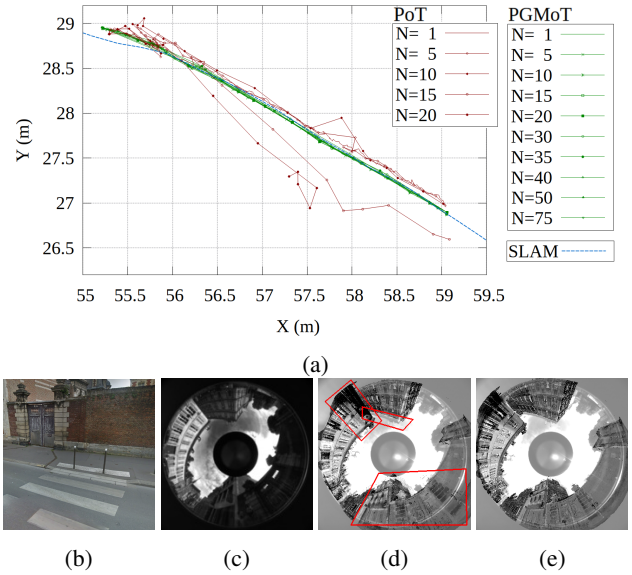


Figure 3: Cumulative distribution of position errors.

4. Experiment: tracking by alignments

To perform direct 3D model-based tracking using omnidirectional images transformed as PGMs, we compute the desired PGM ($\mathbf{G}^*$ in (3)) from an image captured by an actual camera. The current PGM, $\mathbf{G}$, is computed from an omnidirectional image rendered as in Section (3). Brightness of both images are centered and normalized to improve their consistency [15]. Tracking in a sequence of acquired omnidirectional images is done by successive executions of the virtual control law (3) with *Rule2*. The camera pose for the current acquired image is initialized with the optimal pose of the previous one in the sequence. The initial pose for the first image of the sequence is set manually.

The sequence of omnidirectional images is acquired by



(a)

(b) (c) (d) (e)

Figure 4: Direct 3D model-based tracking with PGMoT and PoT using 1 frame every N frame(s): (a) Estimated trajectories on (b) challenging uneven ground where images as (c) are acquired; (d-e) Superimposition of the last acquired and optimal virtual images ($N = 20$): (d) PoT, (e) PGMoT.

an IDS UI-1545LE-M-GL camera (30 images per second) equipped with a RemoteReality omnidirectional catadioptric optics (double mirror). The camera was attached to a Mobile Robots Pioneer 3AT, manually piloted at walking speed (3 km/h, on average) in streets[1]. The robot embeds a SICK LMS-200 Lidar (single horizontal measurement plane) used by the robot software for SLAM, considered for qualitative (not synchronized) comparisons (Fig. 4a).

This section shows an extract of a 350m sequence, focusing on a part of about 4.35m when the robot leaves a sidewalk to cross a street (Fig. 4b). Although short, the sequence is of 210 images. In the firsts, the uneven ground leads the robot to shake while moving straight forward. Although the images are sharp (Fig. 4c), some inter-frame motion are large due to the shaking movements of the robot.

Tracking by direct alignment from PGMs (PGMoT), and from brightness (PoT) for comparison, have been conducted on this sequence using 1 frame every N frame(s). Various N values were used in order to simulate different robot speeds (e.g., $N = 20$ simulates an average displacement speed of about $45km/h$). Figure 4a shows all the estimated trajectories and the manually registered Lidar SLAM one, in order to evaluate the tracking. Due to the motion between consecutive images PoT struggles to precisely track the robot displacements and thus provides noisy trajectories. PoT is actually strongly influenced by the value of N. Indeed, the larger N, the more the drift. For $N = 20$, PoT diverges due to too important inter-frame motion for its limited convergence domain. By contrast, PGMoT remains consistent and succeeds to track the robot displacement even for higher N values. The poses estimated with PGMoT are more reliable, precise and thus produce smooth and accurate trajectories qualitatively on par with the Lidar SLAM.

Figure 4d shows the superimposition of the last real and virtual images of the sequence, rendered at the optimal pose computed with PoT for $N = 20$. The poor alignment is particularly visible in the areas highlighted in red whereas PGMoT leads to a much more precise alignment (Fig. 4e).

We also compare these results with ORB-based visual SLAM [19] (OpenVSLAM), extended to the unified camera model (UCM, Sec. 2.2). In short, while a sparse map and an erratic trajectory can be estimated from the sequence of acquired omnidirectional images (Fig. 4c), they are up to a scale factor, furthermore variable as there is no loop closure in the sequence of images. No way was found to automatically fix the scale despite the several investigations made. First, the sparse map is too sparse to be registered[2] with the dense point cloud of the 3D model. Second, a sparse map could be obtained from rendered images of the 3D model but not any acquired image could be localized in that map, even when built from images rendered at optimal poses obtained with PGMoT. So there is no way to compare quantitatively OpenVSLAM to PGMoT that reliably succeeds in the full scale estimation of camera, hence robot, poses.

5. Conclusion

This paper expresses omnidirectional direct visual servoing, representing images as Photometric Gaussian Mixtures. Evaluation in virtual scenes of photographic appearance shows a significant increase of the convergence domain compared to the previous state-of-the-art photometric omnidirectional direct visual servoing. The new rule of Gaussian extent initialization and optimization also shows a significant improvement over the state-of-the-art rules.

Experiments of direct 3D model-based tracking of the 3D model of a city in omnidirectional images acquired within streets of the same city show the new tracking succeeds where large inter-frame motion prevents the success of the former state-of-the-art one.

6. Acknowledgement

This work is carried out as part of the Interreg VA FCE ADAPT project “Assistive Devices for empowering disAbled People through robotic Technologies” (adapt-project.com). The Interreg FCE Programme is a European Territorial Cooperation programme that aims to fund high quality cooperation projects in the Channel border region between France and England. The Programme is funded by the European Regional Development Fund (ERDF).

[1] http://mis.u-picardie.fr/~g-caron/videos/PGMomni.mp4.

[2] with geometric algorithms of https://www.cloudcompare.org

References

[1] P.-C. Wu, H.-Y. Tseng, M.-H. Yang, and S.-Y. Chien, "Direct pose estimation for planar objects," *Computer Vision and Image Understanding*, vol. 172, pp. 50 – 66, 2018. 1

[2] C. Collewet and E. Marchand, "Photometric visual servoing," *IEEE Trans. on Robotics*, vol. 27, no. 4, pp. 828–834, 2011. 1

[3] A. Lakshmi, F. AGJ, and D. Deodhare, "Robust direct visual odometry estimation for a monocular camera under rotations," *IEEE Robotics and Autom. Letters*, vol. 3, no. 1, pp. 367–372, 2018. 1

[4] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *European C. on Computer Vision*, 2014. 1

[5] S. Park, T. Schöps, and M. Pollefeys, "Illumination change robustness in direct visual slam," in *IEEE Int. C. on Robotics and Autom.*, 2017, pp. 4523–4530. 1

[6] J.-Y. Bouguet, "Pyramidal implementation of the Lucas Kanade feature tracker," *Intel, MRL*, 2000. 1

[7] Y. Ahmine, G. Caron, E. Mouaddib, and F. Chouireb, "Adaptive Lucas-Kanade tracking," *Image and Vision Computing*, vol. 88, Aug. 2019. 1

[8] E. Marchand, "Direct visual servoing in the frequency domain," *IEEE Rob. and Autom. Letters*, vol. 5, no. 2, pp. 620–627, 2020. 1

[9] M. Bakthavatchalam, O. Tahri, and F. Chaumette, "A direct dense visual servoing approach using photometric moments," *IEEE Trans. on Robotics*, vol. 34, no. 5, pp. 1226–1239, 2018. 1

[10] N. Crombez, E. Mouaddib, G. Caron, and F. Chaumette, "Visual servoing with photometric gaussian mixtures as dense features," *IEEE Trans. on Robotics*, vol. 35, no. 1, pp. 49–63, 2019. 1, 2

[11] Z. Zhang, H. Rebecq, C. Forster, and D. Scaramuzza, "Benefit of large field-of-view cameras for visual odometry," in *IEEE Int. C. on Robotics and Autom.*, 2016, pp. 801–808. 1, 3

[12] K. Chappellet, G. Caron, F. Kanehiro, K. Sakurada, and A. Kheddar, "Benchmarking Cameras for OpenVSLAM Indoors," in *Int. C. on Pattern Recognition*, Milan, Italy, Jan. 2021. 1

[13] E. Marchand and François Chaumette, "Virtual visual servoing: A framework for real-time augmented reality," in *EUROGRAPHICS C.*, Saarebrün, Germany, 2002, vol. 21(3), pp. 289–298. 1

[14] G. Caron, E. Marchand, and E. Mouaddib, "Photometric visual servoing for omnidirectional cameras," *Autonomous Robots*, vol. 35, no. 2-3, pp. 177–193, Oct. 2013. 1, 2, 3

[15] N. Crombez, G. Caron, and E. Mouaddib, "Using dense point clouds as environment model for visual localization of mobile robot," in *IEEE Int. C. on Ubiquit. Robots and Ambient Intell.*, 2015, pp. 40–45. 1, 3

[16] U. Nadeem, M. Jalwana, M. Bennamoun, R. Togneri, and F. Sohel, "Direct Image to Point Cloud Descriptors Matching for 6-DOF Camera Localization in Dense 3D Point Clouds," in *Int. C. on Neural Information Processing*, 2019, pp. 222–234. 1

[17] M. Eder, M. Shvets, J. Lim, and J.-M. Frahm, "Tangent images for mitigating spherical distortion," in *IEEE/CVF C. on Computer Vision and Pattern Recognition*, June 2020. 1

[18] D. Caruso, J. Engel, and D. Cremers, "Large-scale direct slam for omnidirectional cameras," in *IEEE/RSJ Int. C. on Intell. Robots & Syst.*, 2015, pp. 141–148. 1

[19] S. Sumikura, M. Shibuya, and K. Sakurada, "OpenVSLAM: A Versatile Visual SLAM Framework," in *ACM Int. C. on Multimedia*, 2019, pp. 2292–2295. 2, 4

[20] J.P. Barreto, F. Martin, and R. Horaud, "Visual servoing/tracking using central catadioptric images," in *Experimental Robotics VIII*, 2003, pp. 245–254. 2

[21] G. Caron and D. Eynard, "Multiple camera types simultaneous stereo calibration," in *IEEE Int. C. on Robotics and Autom.*, 2011, pp. 2933–2938. 3