ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume V-2-2022
XXIV ISPRS Congress (2022 edition), 6–11 June 2022, Nice, France

ROTATION ONLY BUNDLE-ADJUSTMENT FOR THE GENERALIZED CAMERA MODEL AND ITS APPLICATION FOR LARGE-SCALE UNDERWATER IMAGE-SETS

Bashar Elnashef *, Sagi Filin

Mapping and Geo-Information Engineering, Technion – Israel Institute of Technology, Haifa, Israel
(basharnashef, filin)@technion.ac.il

Commission IIWG 9

KEY WORDS: Underwater imaging, Flat-refractive, Bundle adjustment, Rotation averaging, Generalized camera model (GCM), SfM

ABSTRACT:

The generalized camera model allows handling a broad range of imaging systems for which the common central perspective model no longer applies. While offering greater modeling flexibility, designated orientation procedures need to support its application. In that respect, models that facilitate the computation of the bundle-adjustment solution are imperative as a means to solve the motion and structure of a large set of images. Being non-linear, the bundle adjustment solution requires good initial estimates and the filtering of outlying matches. To facilitate such a solution, we propose in this paper a method for estimating the global rotations of a set of views independent of their translation and the scene structure. This method is then used to improve the robustness, efficiency, and convergence of the solution when used as a refinement to the global rotations. The proposed pipeline is evaluated by experiments on synthetic and real-world data of axial flat-refractive cameras. It shows that the proposed method produces more accurate and optimal initial estimates of the global rotations than the state-of-the-art rotation averaging method.

1. INTRODUCTION

The use of the generalized camera model (GCM) has seen a growing interest in recent years with the advancement of imaging systems and multicamera installations (Miraldo and Cardoso, 2020). Notable examples include fish-eye (Castanheiro et al., 2021), central or noncentral catadioptric (Filin et al., 2020), rolling shutter (Fan and Dai, 2021), panoramic (Ji et al., 2020), and underwater cameras (Telem and Filin, 2010), as well as multi-view rigs (Miraldo and Cardoso, 2020). The generalized form facilitates the modeling of a broad range of imaging systems but it also requires suitable pose estimation procedures to support it. Because of the deviation from the central perspective form, integration into structure-from-motion (SfM) and bundle adjustment (BA) solutions requires also robust procedures that generate good initial estimates for the pose parameters and the filtering of outlying matches. The predominant strategy of recent years initializes such solutions by estimating the global motion of each view via an averaging procedure whose aim is to recover global motions from a set of relative ones. Such a scheme, commonly termed motion averaging, provides a fast and accurate method for the cameras' motion estimation with a wide range of applications, including view registration, robotic path estimation, super-resolution, etc. (Chatterjee and Govindu, 2013; Eriksson et al., 2018).

Motion averaging can be partitioned into rotation averaging (RA) followed by translation averaging (TA) procedures (Hartley et al., 2013; Chatterjee and Govindu, 2017; Schonberger and Frahm, 2016; Eriksson et al., 2019; Chen et al., 2021). RA consists of estimating global camera orientations (in reference to a predefined datum) that best agree with the complete set of plausible pairwise relative orientations within the image set. These global orientations are estimated such that the disagreement is minimized and the errors are distributed over the entire set of pairwise constraints. Differing from the BA solution, in RA only the rotations are estimated as a function of the relative pairwise rotations but not so that the image-related reprojection errors are minimized (Chen et al., 2021). As the set of unknown parameters is small, compared to the full BA framework (involving structure + motion), the computation is faster and the procedure is simpler to perform. Being an indirect minimization form, not in reference to the image measurements, and as all relative rotations are treated equally, even when a different number of points is utilized for the individual estimations, this procedure is not optimal. Recently, a method that considers both pair-wise rotation estimates and image measurements was proposed for central cameras (Lee and Civera, 2021). There, given the RA estimates, a rotation-only bundle adjustment (ROBA) solution, which optimizes the rotations over all image measurements, was performed. For that, the authors extended a pair-wise relative orientation model that allows the rotation to be recovered independently of the translation into a multiple-view framework. While offering better initial estimates, their solution only fits central cameras and cannot handle non-conventional imaging systems.

Considering the growing utilization of the generalized camera models, we study in this paper the computation of global rotations as an optimization form and use the image rays direction as the direct input. This solution can be integrated into the SfM pipeline to refine the initial absolute rotations by RA methods before the BA solution commences. While our solution form is general, we demonstrate its application on underwater flat refractive imaging configurations. It has been identified that the flat refractive configuration is axial by nature (cf., Telem and Filin, 2010), also demonstrating that the nature of this system translates to depth dependence in terms of the image related corrections (Telem and Filin, 2010; Nocerino et al., 2021). The literature shows that the predominant method to handle the axial nature of this system is by approximating it as a

*Corresponding author

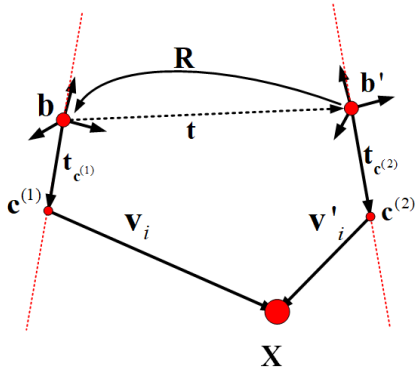

Figure 1: Geometry of the generalized relative pose problem for multi-camera systems. The unknowns are the transformation parameters between the two viewpoints $\mathbf{b}$ and $\mathbf{b}'$, given by $\mathbf{t}$ and $\mathbf{R}$. The observation vectors $\mathbf{v}_i$ and $\mathbf{v}'_i$ and the position of the camera centers $\mathbf{c}^{(1)}$ and $\mathbf{c}^{(2)}$ with respect to $\mathbf{b}$ and $\mathbf{b}'$, given by $\mathbf{t}_{\mathbf{c}^{(1)}}$ and $\mathbf{t}_{\mathbf{c}^{(2)}}$, respectively.

perspective one with distortions (Chadebecq et al., 2020). This produces aberrations and inaccuracies in the reconstruction because of the depth-dependent 3-D error in the system. To alleviate such errors, reweighted BA procedures have been recently proposed, where the weights assigned to the image measurements are changed according to the error introduced by their depth (Nocerino et al., 2021). Aspiring for a physically exact solution to the flat-refractive camera model, we utilize the GCM and solve the global orientations as initial estimates for the BA solution. The advantages of such a representation are the direct form by which the measurements are introduced into the model and the physically exact modeling of the system. We extend a pairwise rotation-only relative orientation solution to handle multiple views by aggregating a set of two-view costs and minimizing them through nonlinear optimization. We test the proposed generalized rotation only bundle BA through a set of simulated tests and on real-world data. As the results show, under typical configurations, the obtained parameters are sufficiently close to the actual ones and may facilitate a variety of applications such as coarse reconstruction or coarse localization. Finally, we compare our results to the state-of-the-art RA method (Chatterjee and Govindu, 2017) and demonstrate consistent and significant gains in accuracy. The organization of this paper is as follows: Sec. (2) develops the generalized two-view rotation-only solution to the flat-refractive setup. Sec. (3) describes the generalized rotation-only solution. In Sec. (4) we present the experimental results, and Sec. (5) presents the discussion and conclusions.

2. GENERALIZED CAMERA MODEL

Image-related measurements for the GCM are usually expressed by Plücker coordinates, a 6-D vector of which the first three elements correspond to the ray direction, and the latter three are given by the cross-product between a point on the line and its direction. For a multi-camera system whose center does not coincide with a specific camera center, we denote its positions at two different epochs by $\mathbf{b}$ and $\mathbf{b}'$, and relate them by a 3-D rigid body transformation, where $\mathbf{t}$ and $\mathbf{R}$ are the respective translation and rotation (Fig. 1). We also consider two different cameras centers, $\mathbf{c}^{(1)}$ and $\mathbf{c}^{(2)}$, each viewing at a different epoch the same object-space point by the two respective rays, $\mathbf{v}_i$ and $\mathbf{v}'_i$, in reference to $\mathbf{b}$ and $\mathbf{b}'$ (Fig. 1). In that form, the Plücker line coordinates of the two observations are given

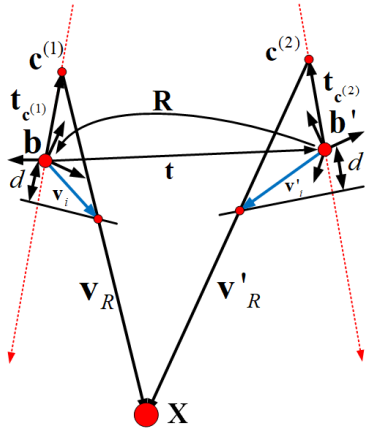

Figure 2: Geometry of the axial camera relative pose problem as a multi-camera system. The unknowns are the transformation parameters between the two viewpoints $\mathbf{b}$ and $\mathbf{b}'$, given by $\mathbf{t}$ and $\mathbf{R}$. The observation vectors $\mathbf{v}_i$ and $\mathbf{v}'_i$ and the position of the camera centers $\mathbf{c}^{(1)}$ and $\mathbf{c}^{(2)}$ with respect to $\mathbf{b}$ and $\mathbf{b}'$, given by $\mathbf{t}_{\mathbf{c}^{(1)}} = k\mathbf{n} = [0, 0, -k]^T$ and $\mathbf{t}_{\mathbf{c}^{(2)}} = k'\mathbf{n} = [0, 0, -k']^T$, respectively. The axial nature dictates that all camera centers to be on the same system axis, n (dashed red line) for each one of the viewpoints. Note that, $\mathbf{b}$ and $\mathbf{b}'$ must also lie on the axes of the systems.

by:

$$\mathbf{L}_i = \begin{pmatrix} \mathbf{v}_i \\ \mathbf{t}_{\mathbf{c}^{(1)}} \times \mathbf{v}_i \end{pmatrix} \qquad \mathbf{L}'_i = \begin{pmatrix} \mathbf{v}'_i \\ \mathbf{t}_{\mathbf{c}^{(2)}} \times \mathbf{v}'_i \end{pmatrix} \tag{1}$$

Integration of the Plücker line transformation and the intersection-constraint (cf. Förstner and Wrobel, 2016, for more details) leads to the generalized epipolar constraint (GEC):

$$\mathbf{L}_i^T \begin{pmatrix} [\mathbf{t}]_\times \mathbf{R} & \mathbf{R} \\ \mathbf{R} & \mathbf{0} \end{pmatrix} \mathbf{L}'_i = 0 \tag{2}$$

where $[\mathbf{t}]_\times$ represents the skew-symmetric form of $\mathbf{t}$. Substituting Eq. (1) into Eq. (2), we obtain the generalized epipolar constraint (GEC):

$$\mathbf{v}_i^T [\mathbf{t}]_\times \mathbf{R} \mathbf{v}'_i + \mathbf{v}_i^T \left(\mathbf{R} [\mathbf{t}_{\mathbf{c}^{(1)}}]_\times - [\mathbf{t}_{\mathbf{c}^{(2)}}]_\times \mathbf{R} \right) \mathbf{v}'_i = 0 \tag{3}$$

Similar to the central case, this formulation allows solving linearly for the relative pose. However, the linear solution has a large redundant parametrization and requires 17 correspondences for solving only 6 DoF (Kim et al., 2009).

2.1 Application to flat-refractive geometry of underwater cameras

When applied to underwater cameras that image through a flat interface, one needs to account for refraction at the interface that bends the incident ray direction and introduces a non-linear trajectory (Fig. 2). We can maintain the collinearity of the incident ray by offsetting the camera center along an axis whose direction is the normal to the interface by a magnitude $\tilde{k}f$, such that the modified ray direction under such a formulation becomes,

$$\mathbf{v}_{Ri} = \left(x_i, y_i, f \left(1 + \tilde{k}_i \right) \right)^T \tag{4}$$

where x_i, y_i, are the image plane coordinates given in the calibrated camera frame, and $\tilde{k}$ is a correction factor to the principal distance, f, whose magnitude depends on the image point co-

ordinates (cf. Appendix A). Next, defining k_i as the correction to the principal distance, where $\mathbf{t}_{\mathbf{c}^{(i)}} = k_i\mathbf{n} = [0, 0, -k_i]^T$ and $\mathbf{n} = [0, 0, 1]^T$, the image ray can be expressed in the Plücker line representation by,

$$\begin{aligned}\mathbf{L}_i &= \begin{pmatrix} \mathbf{v}_{Ri} \\ -\left(k_i[\mathbf{n}]_\times \mathbf{v}_{Ri}\right) \end{pmatrix} \\ &= \begin{pmatrix} x_i & y_i & f(1+\tilde{k}_i) & -k_i y_i & k_i x_i & 0 \end{pmatrix}^T \end{aligned} \tag{5}$$

such that $k_i\mathbf{n}$ is the modified position of the camera center and $\mathbf{v}_{Ri}$ is the ray direction. Note that Eq. (5) defines a linear form of the ray trajectory through refraction which is expressed by image-related quantities only. Substituting the Plücker line-coordinates from Eq. (5) into Eq. (2) and with further derivations we obtain:

$$\left(\mathbf{v}_{Ri} \times \mathbf{R}\mathbf{v}'_{Ri}\right)\mathbf{t} + \mathbf{v}_{Ri}^T\left(k'_i\mathbf{R}[\mathbf{n}]_\times - k_i[\mathbf{n}]_\times\mathbf{R}\right)\mathbf{v}'_{Ri} = 0 \tag{6}$$

equivalent to Eq. (3), thereby establishing the analogy of the flat-refractive camera model to the GCM and GEC forms.

2.2 GEC rotation-only solution

Denoting $\bar{\mathbf{t}} = [\mathbf{t}, 1]^T$, Kneip and Li (2014) expressed Eq. (6) as follows:

$$\mathbf{g}_i^T\bar{\mathbf{t}} = \begin{pmatrix} \mathbf{v}_{Ri} \times \mathbf{R}\mathbf{v}'_{Ri} \\ \mathbf{v}_{Ri}^T\left(k'_i\mathbf{R}[\mathbf{n}]_\times - k_i[\mathbf{n}]_\times\mathbf{R}\right)\mathbf{v}'_{Ri} \end{pmatrix}^T \bar{\mathbf{t}} = 0 \tag{7}$$

where $\mathbf{g}_i$ is termed the generalized epipolar plane normal vector, and $\bar{\mathbf{t}}$ the homogeneous translation vector, which has an arbitrary scale. Having n generalized normal vectors, the following constraint can be generated:

$$\mathbf{G}_i^T\bar{\mathbf{t}} = \left(\mathbf{g}_1 \dots \mathbf{g}_n\right)^T\bar{\mathbf{t}} = 0 \tag{8}$$

This expression constrains $\bar{\mathbf{t}}$ by a $n \times 4$ matrix that depends only on the rotation parameters. As the trivial solution is not allowed, the rank of $\mathbf{G}$ has to be 3. Hence, given an arbitrary number of correspondences, n, a rank minimization of $\mathbf{G}$ over $\mathbf{R}$ can be reached by minimizing the smallest eigenvalue of

$$\mathbf{H} = \mathbf{G}\mathbf{G}^T = \sum_n^{i=1} \mathbf{g}\mathbf{g}^T \tag{9}$$

Solving $\mathbf{R}$ through $\mathbf{H}$ can be done by the minimization of

$$\mathbf{R}^* = \underset{\mathbf{R}}{\operatorname{argmin}}\, \lambda_{\mathbf{H}}\left(\mathbf{R}\right) \tag{10}$$

where $\lambda_{\mathbf{H}}\left(\mathbf{R}\right)$ is the smallest eigenvalue of $\mathbf{H}$, which is a function of $\mathbf{R}$. Let $a\lambda^4 + b\lambda^3 + c\lambda^2 + d\lambda + e = 0$ be the fourth degree polynomial whose roots are the eigenvalues of $\mathbf{H}$. The coefficients $\{a, b, c, d, e\}$ can be derived from $\det\left(\mathbf{H} - \lambda\mathbf{I}_{4\times4}\right)$. The smallest root can be obtained in closed form by applying Ferrari's solution:

$$\alpha = -{}^{3b^2}/_8 + c; \qquad \beta = {}^{b^3}/_8 - {}^{bc}/_2 + d \tag{11}$$

$$\gamma = -{}^{3b^4}/_{256} + {}^{b^2c}/_{16} - {}^{bd}/_4 + e; \quad p = -{}^{\alpha^2}/_{12} - \gamma \tag{12}$$

$$q = -{}^{\alpha^3}/_{108} + {}^{\alpha\gamma}/_3 - {}^{\beta^2}/_8; \qquad h = -{}^{p^3}/_{27} \tag{13}$$

$$\theta_1 = h^{1/6} cos\left({}^1/_3 \arccos\left(-{}^q/_{2\sqrt{h}}\right)\right); \qquad \theta_2 = h^{1/3} \tag{14}$$

$$y = -\frac{5\alpha}{6} - \frac{p\theta_1}{3\theta_2} + \theta_1; \qquad w = \sqrt{\alpha + 2y} \tag{15}$$

and,

$$\lambda_{\mathbf{H},min} = -\frac{b}{4} - {}^w/_2 - {}^1/_2\sqrt{-3\alpha - 2y + {}^{2\beta}/_w} \tag{16}$$

Hence, $\mathbf{R}^*$ that minimizes the GEC (Eq. 6) can be obtained directly by nonlinearly solving Eq. (15). The GEC rotation-only solution allows computing pair-wise orientations between all image pairs with a sufficient number of correspondences.

3. GENERALIZED ROTATION ONLY BUNDLE ADJUSTMENT (G-ROBA)

Given the set of plausible relative orientations, it is possible to reconstruct a view-graph $\mathcal{G}$ that encodes all connections between the pair of views. We define $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ such that $\|\mathcal{V}\| = N$ and $\|\mathcal{E}\| = M$, where $\mathcal{V}$ is the set of N cameras and $\mathcal{E}$ is the set of M edges representing the relative orientation between individual cameras. In this representation the edge $ij \in \mathcal{E}$ represents the computed relative rotation $\mathbf{R}_{ij}$ between the cameras i and j, such that $[i, j] \in \mathcal{V}$. We denote the collection of all relative rotations by $\mathbf{R}_\mathcal{E}$. The set of all 3-D rotations $\mathbf{R}_\mathcal{V} = \{\mathbf{R}_1, \mathbf{R}_1, \dots, \mathbf{R}_N\}$ completely specifies the global rotation of all the cameras with respect to a given reference frame. If $\mathcal{E}$ spans the entire view graph, the global rotations of all the cameras can be solved by using the pairwise relative rotations. As the cameras i and j have global rotations of $\mathbf{R}_i$ and $\mathbf{R}_j$ respectively in a given reference frame, the relative rotation between them should obey the relationship, $\mathbf{R}_{ij} = \mathbf{R}_j\mathbf{R}_i^T$, $\forall\, ij \in \mathcal{E}$.

The problem of relative rotation averaging can be stated as follows: given a sufficient number of relative rotations $\mathbf{R}_{ij} \in \mathbf{R}_\mathcal{E}$, we seek an estimate of the global camera rotations, $\mathbf{R}_\mathcal{V}$. In practice, we always have a larger number of edges than what is required to span the view-graph, i.e., $M > N - 1$, implying that we have a redundant set of observations. We also note that due to the presence of noise or outliers, the set of relative rotations is inconsistent, i.e., we cannot find a solution $\mathbf{R}_\mathcal{V} = \{\mathbf{R}_1, \mathbf{R}_1, \dots, \mathbf{R}_N\}$ that exactly satisfies all constraints $\{\mathbf{R}_{ij} = \mathbf{R}_j\mathbf{R}_i^T | \forall ij \in \mathcal{E}\}$.

We seek to find an estimate of $\mathbf{R}_\mathcal{V}$ that is most consistent with the observed relative rotations. This can be obtained by minimizing a cost function that penalizes the discrepancy between the observed relative rotations $\mathbf{R}_{ij}$ and the one suggested by the estimate $\mathbf{R}_j\mathbf{R}_i^T$, i.e.,

$$\mathbf{R}_\mathcal{V} = \underset{\{\mathbf{R}_1, \mathbf{R}_1, \dots, \mathbf{R}_N\}}{\operatorname{argmin}} \sum_{(i,j)\in\mathcal{E}} \rho\left(d\left(\mathbf{R}_{ij}, \mathbf{R}_j\mathbf{R}_i^T\right)\right) \tag{17}$$

where $d(.)$ is a distance measure between two rotations in $SO(3)$ and $\rho(.)$ is a loss function defined over this distance measure. We follow Chatterjee and Govindu (2017) that apply a two-step approach in which first an L_1-iterative reweighted least-squares (L_1-IRLS) is used for initialization and is then switched to an $L_{1/2}$-IRLS for additional refinement.

Next, we extend the idea of utilizing the two-view rotation-only solution for a rotation-only BA (ROBA, Lee and Civera, 2021) to the case of generalized cameras. Given the set of all edges $\mathcal{E}$, the global rotations of the N cameras are computed using the RA. This provides the set of global rotations denoted by $\{\mathbf{R}_1, \dots, \mathbf{R}_N\}$. For each edge in $\mathcal{E}$ a constraint on two rotations from the set of N global rotations can be generated using

Eq. (7). This form allows minimizing the repreojection as it is directly related to the image measurements. However, in its form, it can only relate two views. To extend the relative orientation problem to handle multiple views, we define a cost given by $\lambda_{\mathbf{H},min}$ which can be considered as a measure of how good is the estimate and amounts to zero when no noise exists in the system. For each edge, the value of $\lambda_{\mathbf{H},min}$ is computed while using the global rotation estimates as inputs and the sum of all costs is then minimized in a single optimization. Lee and Civera (2021) showed that the cost $\sqrt{\lambda_{\mathbf{H},min}}$ performed better and improved the convergence rate. Hence, the optimization problem can be formulated as

$$\{\mathbf{R}_1^*, \dots, \mathbf{R}_N^*\} = \underset{\mathbf{R}_1,\dots,\mathbf{R}_N}{\operatorname{argmin}} \mathcal{C}\left(\mathbf{R}_1, \dots, \mathbf{R}_N\right) \tag{18}$$

with,

$$\mathcal{C}\left(\mathbf{R}_1, \dots, \mathbf{R}_N\right) = \sum_{(j,k)\in\mathcal{E}} \sqrt{\lambda_{\mathbf{H}}\left(\mathbf{R}_{jk}\right)} \tag{19}$$

To solve Eq. (18) iteratively, the first-order gradient-based optimization algorithm for stochastic objective functions, ADAM (Kingma and Ba, 2014) was used. The hyper-parameters, β_1 and β_2, were set to the default values (i.e., $\beta_1 = 0.9$, $\beta_2 = 0.999$). The step size was set to $\alpha = 0.01$ at the beginning from which was then switched to $\alpha = 0.001$ permanently once the cost increased in five successive iterations.

4. EXPERIMENTS

To test the performance of the proposed global rotation estimation solution, experiments were carried out via simulations and validated using real-world data. For the simulated experiments, we use a 640×480 pixels frame camera with a 525 pixel focal length. The distance to the interface, d was set to 30 mm, and the indices of refraction to $\mu_0 = 1$ and $\mu_1 = 1.33$. We generated 3-D points at random distances $D \sim \mathcal{U}\left(1, 5\right)$ m from the xy-plane while ensuring that each view observed at least 10 corresponding points with neighboring views. The image points were formed by forward-projecting the generated 3-D points using ray-tracing (Kunz and Singh, 2008). They were then perturbed by Gaussian noise characterized by $\mathcal{N}\left(0, \sigma^2\right)$. Image pairs with more than n_{in} corresponding points formed an edge in $\mathcal{E}$. Our base setup consisted of $n_{views} = 100$, and we set $n_{in} = 50$ and $\sigma = 1$ pixel. We then evaluated also the influence of the following alterations to the base setup: *i*) an increase in the number of minimal corresponding points to $n_{in} = 100$, thereby limiting the number of permissible pairwise connections; *ii*) fewer views covering the same scene to $n_{views} = 30$, thereby reducing the overlap; *iii*) more views, to $n_{views} = 300$, thereby increasing the overlap; *iv*) decreased noise level, to $\sigma = 0.5$ pixel; and *v*) increased noise level, to $\sigma = 2$ pixels. Each setup was simulated by 200 different configurations of randomly sampled camera rotations and a 3-D set of points. We tested these setups on two imaging scenarios: *i*) of a closed-loop where n cameras were uniformly distributed one unit apart from one another, while their optical axis direction was uniformly perturbed by an $\theta \sim \mathcal{U}\left(0, 20^\circ\right)$ with respect to the z-axis; and *ii*) of an image block made of n_{strips} strips, with n_{views} images per strip. The rotations were uniformly perturbed by $\theta \sim \mathcal{U}\left(0, 5^\circ\right)$ off the nadir direction.

To simulate as realistic as a possible scenario, the relative rotation estimates were computed using the method from Sec. (2.2), The rotation, $\mathbf{R}_{ij}$ between images i and j initialized the RA solution. Finally, the output of the RA was introduced to the G-ROBA solution. To compute the RA solution, the state-of-the-art model by Chatterjee and Govindu (2017) was used. Its implementation was based on the code publicly shared by the authors. For performance evaluation, we note that our method was implemented in Python and tested on an Intel i7-3770 CPU, 3.40GHz PC 16GB RAM.

As the simulated parameters are given in absolute terms and our solutions is in reference to an arbitrary datum, the global rotation estimates $(\hat{\mathbf{R}}_1, \dots, \hat{\mathbf{R}}_N)$ do not share the same reference frame as their ground-truth counterparts $(\mathbf{R}_1^{gt}, \dots, \mathbf{R}_N^{gt})$. Therefore, they must first be aligned with the ground-truth to evaluate the accuracy. To do so, it is customary to estimate a rotation that transforms the estimated global rotations to the ground-truth system. Being a nonlinear single rotation-averaging problem, it is solved iteratively (Hartley et al., 2013). We compute this by minimizing the L_1 and L_2 norms, yielding two such rotations, $\mathbf{R}_{L_1}$ and $\mathbf{R}_{L_2}$, respectively that are estimated by:

$$\mathbf{R}_{L_1} = \underset{\mathbf{R}_{L_1}}{\operatorname{argmin}} \sum_{j=1}^{N} d\left(\mathbf{R}_{L_1}, \mathbf{R}_j^T \mathbf{R}_j^{gt}\right) \tag{20}$$

and,

$$\mathbf{R}_{L_2} = \underset{\mathbf{R}_{L_2}}{\operatorname{argmin}} \sum_{j=1}^{N} d\left(\mathbf{R}_{L_2}, \mathbf{R}_j^T \mathbf{R}_j^{gt}\right)^2 \tag{21}$$

where $d(\cdot, \cdot)$ denotes the geodesic distance between the two rotations, i.e., $d(\mathbf{R}_1, \mathbf{R}_2) = \arccos\left(\left(tr\left(\mathbf{R}_1 \mathbf{R}_2^T\right) - 1\right)/2\right)$. Next, all the estimated global rotations are rotated by $\mathbf{R}_{L_1}$ and $\mathbf{R}_{L_2}$ to produce two different alignments corresponding to the L_1 and L_2 minimization norms. The mean and median angular errors using these two alignment methods are presented and analyzed.

We compare the performance of the application of the RA and G-ROBA in terms of the mean angular error following the L_1 and L_2 alignment. For evaluation, we define the metrics mn1 and mn2 as the mean angular error (in degrees) following the L_1/L_2 alignment, respectively. Similarly, we also denote the median angular error for the L_1/L_2 alignments as md1 and md2, respectively. Figures 4, 5, 6, and 7 illustrate the results in which on each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles (Q1 and Q3), respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted as blobs, individually.

Closed-loop configuration – Figs. 4 & 5 plot the results of the closed-loop simulation. In all settings considered, the G-ROBA solution demonstrated an improvement over the RA. For example, for the baseline case and using the L_1 alignment (Fig. 4), the mean angular error dropped from $2.41^\circ \pm 0.63^\circ$ when using the RA to $1.38^\circ \pm 0.57^\circ$ using the G-ROBA. For the L_2 alignment, the mean angular error dropped from $2.39^\circ \pm 0.67^\circ$ to $1.41^\circ \pm 0.59^\circ$ (Fig. 5). The addition of more points improved the results, reaching a mean as low as 1.05° when using our method. Furthermore, using fewer views improved the results when compared to the baseline and more views cases with 0.29° compared to 1.38° and 3.89°, respectively. Finally, the noise test showed that our method scales well with the increase of noise, reaching a mean of $1.22^\circ, 1.38^\circ$, and 1.62 for 0.5, 1, and 2 pixels, respectively. This demonstrates the model's capa-

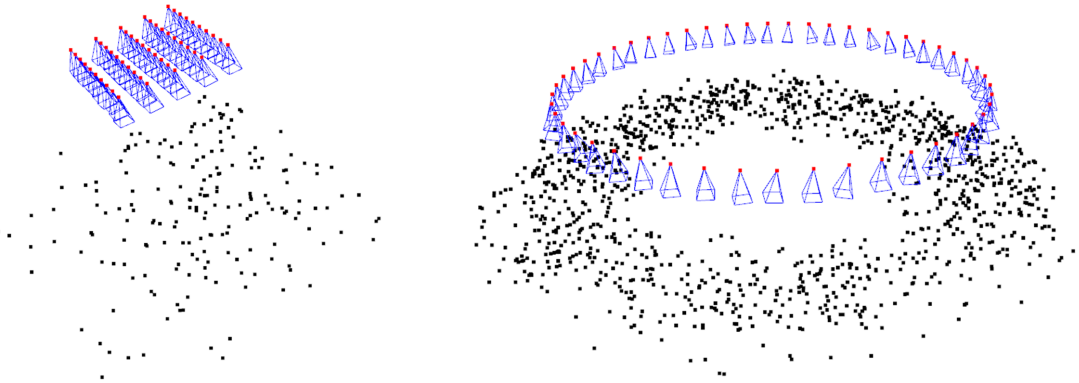

Figure 3: Simulated imaging configurations. (left) We simulate a block structure with n_{strips} as the number of strips, each containing n_{views} number of views. The rotations are perturbed by random angles $\theta \sim \mathcal{U}(0, 5°)$; (right) we uniformly distribute n cameras on a circle on the xy-plane such that the neighbors are evenly spaced along the circle perimeter. After aligning their optical axes with the z-axis, we perturb the rotations by random angles $\theta \sim \mathcal{U}(0, 20°)$.

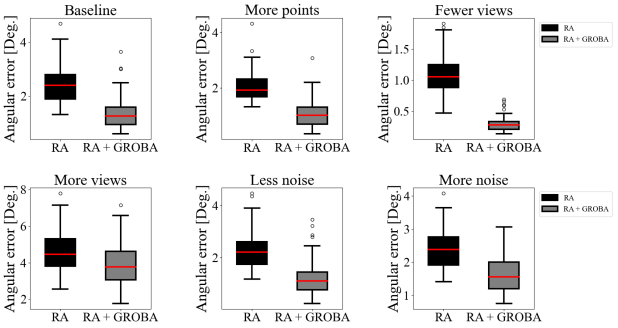

Figure 4: Results of the synthetic data test settings. Comparison between RA and G-ROBA (initialized by RA) in terms of the mean angular error after the L_1 alignment.

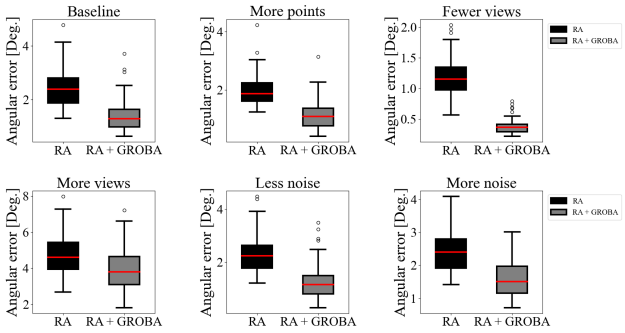

Figure 5: Results of the synthetic data test settings. Comparison between RA and G-ROBA (initialized by RA) in terms of the mean angular error after the L_2 alignment.

bility to handle large quantities of noise while also maintaining a linear trend.

Block configuration – For the base case, and using the L_1 alignment, the mean angular error was $1.18° \pm 0.29°$ when using the RA and $0.51° \pm 0.18°$ for the G-ROBA solution. Using the L_2 alignment, the respective mean angular errors were $1.42° \pm 0.36°$ and $0.62° \pm 0.24°$ (Fig. 5). Additional points improved the results with a mean error of $0.39°$ and fewer views improved the results even further, reaching a $0.17°$ mean error. These results show, similar to the closed-loop simulation, that both methods performed better for the cases where more points were added and lesser views of the same scene were observed. Comparing these settings to the baseline and more views cases show an improvement by a factor with $0.5°$ and $1.15°$ for the base and more views cases, respectively. Hence, for all settings considered, the G-ROBA solution demonstrated improved results compared to the RA. Here also the angular error for all six settings was lower than that of the closed-loop configuration. We attribute this to the number of edges gained by the side overlap between adjacent strips.

4.1 Real-world experiments

To validate our method in real-world conditions we compared the computation of our global rotations to a publicly shared dataset by Bender et al. (2013) of the O'Hara Reef, Tasman Peninsula, Tasmania. The data consists of the raw onboard sensor information of the Sirius autonomous underwater vehicle (AUV), a modified version of the SeaBED AUV (Bender et

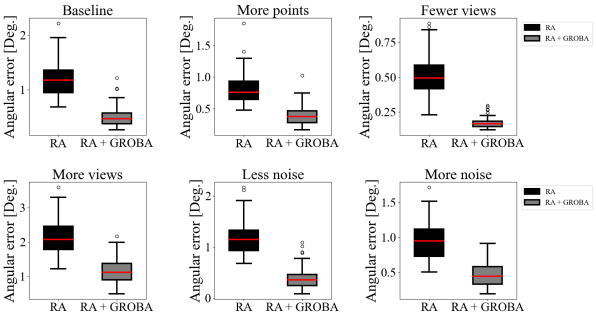

Figure 6: Results of the synthetic data test settings. Comparison between RA and G-ROBA (initialized by RA) in terms of the mean angular error after the L_1 alignment.

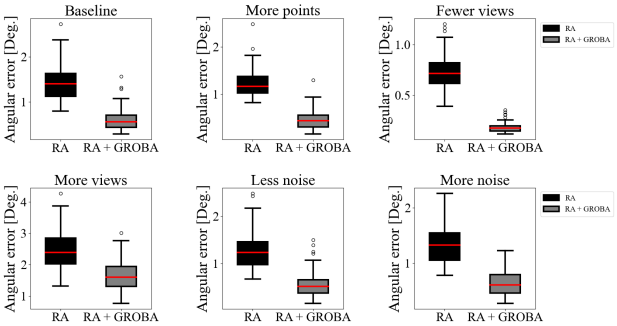

Figure 7: Results of the synthetic data test settings. Comparison between RA and G-ROBA (initialized by RA) in terms of the mean angular error after the L_2 alignment.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-2-2022-325-2022 | © Author(s) 2022. CC BY 4.0 License.

Settings	RA						RA + G-ROBA					
	mn1	**md1**	σ_1	**mn2**	**md2**	σ_2	**mn1**	**md1**	σ_1	**mn2**	**md2**	σ_2
Baseline	2.40	2.40	0.63	2.39	2.38	0.64	1.38	1.26	0.57	1.41	1.28	0.57
More points	2.04	1.92	0.50	1.97	1.86	0.50	1.05	1.01	0.47	1.12	1.08	0.47
Fewer views	1.08	1.05	0.31	1.18	1.15	0.31	0.29	0.28	0.11	0.37	0.36	0.11
More views	4.59	4.45	1.03	4.74	4.60	1.04	3.89	3.77	1.12	3.93	3.81	1.12
Less noise	2.25	2.21	0.63	2.28	2.24	0.64	1.22	1.11	0.62	1.25	1.14	0.62
More noise	2.40	2.39	0.56	2.42	2.40	0.57	1.62	1.55	0.50	1.58	1.51	0.50

Table 1: Results of the closed-loop simulation. The metrics, mn/md/σ1/2: mean/median angular error (in degrees) after the L_1/L_2 alignment, respectively. For all settings, the RA + G-ROBA improves the results of the RA.

Settings	RA						RA + G-ROBA					
	mn1	**md1**	σ_1	**mn2**	**md2**	σ_2	**mn1**	**md1**	σ_1	**mn2**	**md2**	σ_2
Baseline	1.18	1.17	0.29	1.40	1.40	0.35	0.50	0.47	0.18	0.61	0.55	0.24
More points	0.81	0.76	0.23	1.23	1.16	0.28	0.39	0.38	0.15	0.45	0.44	0.20
Fewer views	0.51	0.49	0.14	0.73	0.71	0.17	0.17	0.17	0.03	0.18	0.17	0.05
More views	2.14	2.08	0.47	2.47	2.39	0.58	1.15	1.12	0.35	1.65	1.60	0.47
Less noise	1.16	1.14	0.29	1.26	1.24	0.35	0.39	0.35	0.19	0.56	0.52	0.26
More noise	0.95	0.95	0.26	1.34	1.33	0.32	0.47	0.44	0.16	0.64	0.60	0.21

Table 2: Results of the block simulation. The metrics, mn/md/σ1/2: mean/median angular error (in degrees) after the L_1/L_2 alignment, respectively. For all settings, the RA + G-ROBA improves the results of the RA.

al., 2012). The vehicle was designed for high-resolution, geo-referenced imaging (Bryson et al., 2013). It includes 11,278 stereo image pairs and an onboard Imagenex DeltaT 260kHz Multibeam sensor. Integration of both data streams was used in a SLAM solution that is utilized here as a baseline for validation.[1] The set of images covers a seabed strip, more than 4 km long. The trajectory consists of a 4 km long straight transect, and a zig-zag path that traverses this transect, crossing it five times (Fig. 8). In our analysis we studied the RA and G-ROBA performance on *i*) the 4 km long open-end strip and the zig-zag path, *ii*) on both setups using only the left and then right cameras and *iii*) on the whole dataset, whose overall length was 8 km, and facilitates a loop closure.

Data processing: We used SIFT (Lowe, 2004) to extract feature points and the FLANN (Muja and Lowe, 2014) for the matching. To improve the efficiency of the matching stage and as the image order was known, we limited the possible matches to the five neighboring images in both directions. Considering the fact that the images were acquired in a strip-like campaign, this simplification had hardly any effect on the edges in the view graph. The underwater system parameters were calibrated using the recently proposed method of Elnashef and Filin (2022), and then the relative rotation of each edge we estimated according to Sec. (2.2). As a means to remove bad matches, we computed the relative orientation over all stereo-pairs and removed outlying matches using the random sample consensus (RanSaC) algorithm. Next, we computed the RA solutions and refined them using the proposed G-ROBA solution.

Validations: Results of the three scenarios and test sets are listed in Table (4.1). For all scenarios and subsets, our solution outperformed RA error-wise. In scenario #1, we computed the angular error for the two subsets, namely, a zig-zag path with 6278 stereo images, and a straight path with 5000 stereo images with respect to their baseline. We observe that in the straight path set, the RA error was large, reaching a mean error of 6.22°, these results were improved by our solution reducing the error to 4.89°. Also, we listed the run-time for each set and show that the RA method is more efficient by at least an order of magnitude compared to our solution (Table 4.1). The application of both methods in scenario #2 shows that the error reached are slightly better than those from scenario #1. This we believe

[1] Tasmania O'Hara 7 - http://marine.acfr.usyd.edu.au/datasets/#home

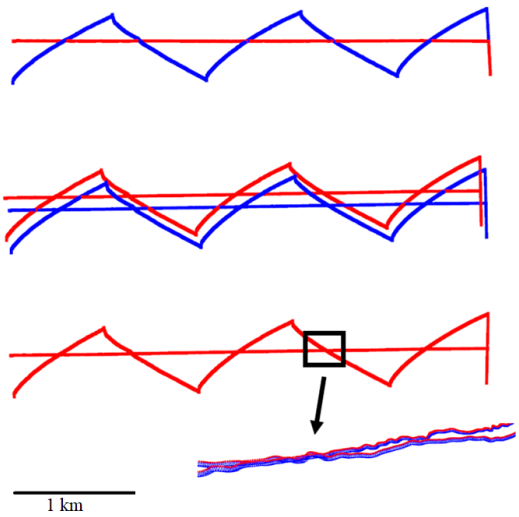

Figure 8: Validations scenarios in real-world experiments. (Top) The data are partitioned into two patches, a zig-zag path, and a straight path; (Middle) The data are partitioned into left and right images of the stereo-pair. Note that, the offset between the two lines is enlarged for a better illustration; (Bottom) Full dataset with a side view magnification at the intersection between the zig-zag and straight paths.

is a consequence of adding more edges at the loop closure positions, namely, the intersection between the zig-zag and straight paths (Fig. 8). To measure the consistency of the solution, we computed the angular error of the alignments between the left (11278 images) and right (11278 images) sets (Table 4.1). The mean errors were as low as 0.16° when using G-ROBA. Applying both methods over the entire dataset (scenario #3), yielded errors as low as 3.52° in comparison to mn1=2.27° for the RA and G-ROBA, respectively. Demonstrating once again that our solution provides an improvement over the RA and reduces the overall error.

5. CONCLUSIONS

This paper proposed a generalized version of the rotation-only bundle adjustment for the generalized camera model. It presented a complete pipeline adapted for that purpose that begins with a relative orientation of this axial camera form, through

Scenario	Data	RA					RA + G-ROBA				
		mn1	md1	mn2	md2	t [sec]	mn1	md1	mn2	md2	t [sec]
I	**Zig-Zag**	4.01	2.52	4.28	2.78	26	1.82	0.33	1.61	0.59	317
	Straight line	6.22	1.52	8.26	4.06	15	4.89	0.45	7.22	3.02	278
II	**Right**	5.59	4.19	5.87	3.92	166	2.89	0.88	3.06	1.11	611
	Left	5.42	4.22	5.70	3.88	166	2.76	0.95	3.15	1.02	611
	Relative	0.28	0.10	0.31	0.14	2	0.16	0.05	0.19	0.10	10
III	**Full dataset**	3.52	2.30	3.62	2.17	422	2.27	1.92	2.38	1.89	1288

Table 3: Comparison of the real-world results against the baseline dataset (Bryson et al., 2013) divided into three scenarios. (1) The data are partitioned into two sets, a zig-zag path (6278 stereo images), and a straight path (5000 stereo images), and solved independently from one another. Both sets are compared to their baseline counterpart, respectively. (2) The data are partitioned into left (11278 images) and right (11278 images) images. The two sets are then compared to their baseline counterpart and to one another (relative rotation between the left and right trajectories). (3) Full dataset (11278 stereo images). In all experiments, the RA + G-ROBA outperformed the RA, error-wise.

the establishment of a view-graph for the image set, the initialization by applying a rotation averaging procedure, and the generalized refinement of the global rotations. The sequential process poses little computational demand on the complete bundle adjustment solution while improving the estimated parameters that are introduced into it and the ability to filter outlying matches. Our experiments demonstrate that the proposed pipeline is general, and performs well even when no loop closure is enforced on the image block. With the introduction of the loop closure, the estimated rotations further improve, in some typical constellations providing accurate results sufficient for some coarse mapping tasks. Our evaluations also demonstrated that in all setups the G-ROBA solution outperformed the standard RA. Future research would study the integration of these solutions, into a global bundle adjustment and SfM procedures, as a means to obtain a suited underwater global orientation solution.

6. ACKNOWLEDGMENT

The authors would like to acknowledge the Australian Center for Field Robotics' marine robotics group for providing the data. Funding was provided in part by the Neubauer family foundation.

REFERENCES

Bender, A., Williams, S. B., Pizarro, O., 2012. Classification with probabilistic targets. *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 1780–1786.

Bender, A., Williams, S. B., Pizarro, O., 2013. Autonomous exploration of large-scale benthic environments. *2013 IEEE International Conference on Robotics and Automation*, IEEE, 390–396.

Bryson, M., Johnson-Roberson, M., Pizarro, O., Williams, S. B., 2013. Colour-consistent structure-from-motion models using underwater imagery. *Robotics: Science and Systems VIII*, 33.

Castanheiro, L. F., Tommaselli, A. M. G., Berveglieri, A., Campos, M. B., Junior, J. M., 2021. Modeling Hyperhemispherical Points and Calibrating a Dual-Fish-Eye System for Close-Range Applications. *Photogrammetric Engineering & Remote Sensing*, 87(5), 375–384.

Chadebecq, F., Vasconcelos, F., Lacher, R., Maneas, E., Desjardins, A., Ourselin, S., Vercauteren, T., Stoyanov, D., 2020. Refractive two-view reconstruction for underwater 3d vision. *International Journal of Computer Vision*, 128(5), 1101–1117.

Chatterjee, A., Govindu, V. M., 2013. Efficient and robust large-scale rotation averaging. *Proceedings of the IEEE International Conference on Computer Vision*, 521–528.

Chatterjee, A., Govindu, V. M., 2017. Robust relative rotation averaging. *IEEE transactions on pattern analysis and machine intelligence*, 40(4), 958–972.

Chen, Y., Zhao, J., Kneip, L., 2021. Hybrid rotation averaging: A fast and robust rotation averaging approach. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10358–10367.

Elnashef, B., Filin, S., 2020. DIRECT ESTIMATION OF THE RELATIVE ORIENTATION IN UNDERWATER ENVIRONMENT. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 5(2).

Elnashef, B., Filin, S., 2022. Target-free calibration of flat refractive imaging systems using two-view geometry. *Optics and Lasers in Engineering*, 150, 106856.

Eriksson, A., Olsson, C., Kahl, F., Chin, T.-J., 2018. Rotation averaging and strong duality. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 127–135.

Eriksson, A., Olsson, C., Kahl, F., Chin, T.-J., 2019. Rotation averaging with the chordal distance: Global minimizers and strong duality. *IEEE transactions on pattern analysis and machine intelligence*, 43(1), 256–268.

Fan, B., Dai, Y., 2021. Inverting a rolling shutter camera: bring rolling shutter images to high framerate global shutter video. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4228–4237.

Filin, S., Ilizirov, G., Elnashef, B., 2020. Robust pose estimation and calibration of catadioptric cameras with spherical mirrors. *Photogrammetric Engineering & Remote Sensing*, 86(1), 33–44.

Förstner, W., Wrobel, B. P., 2016. *Photogrammetric computer vision*. Springer.

Hartley, R., Trumpf, J., Dai, Y., Li, H., 2013. Rotation averaging. *International journal of computer vision*, 103(3), 267–305.

Hecht, E., 2002. *Optics*. Pearson education, fourth edn, Addison-Wesley.

Ji, S., Qin, Z., Shan, J., Lu, M., 2020. Panoramic SLAM from a multiple fisheye camera rig. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159, 169–183.

Kim, J.-H., Li, H., Hartley, R., 2009. Motion estimation for nonoverlapping multicamera rigs: Linear algebraic and l geometric solutions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6), 1044–1059.

Kingma, D. P., Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kneip, L., Li, H., 2014. Efficient computation of relative pose for multi-camera systems. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 446–453.

Kunz, C., Singh, H., 2008. Hemispherical refraction and camera calibration in underwater vision. *OCEANS 2008*, IEEE, 1–7.

Lee, S. H., Civera, J., 2021. Rotation-only bundle adjustment. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 424–433.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), 91–110.

Miraldo, P., Cardoso, J. R., 2020. On the Generalized Essential Matrix Correction: An efficient solution to the problem and its applications. *Journal of Mathematical Imaging and Vision*, 62, 1107–1120.

Muja, M., Lowe, D. G., 2014. Scalable nearest neighbor algorithms for high dimensional data. *IEEE transactions on pattern analysis and machine intelligence*, 36(11), 2227–2240.

Nocerino, E., Menna, F., Grün, A., 2021. Bundle adjustment with polynomial point-to-camera distance dependent corrections for underwater photogrammetry. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences-ISPRS Archives*, 43(B2-2021), 673–679.

Schonberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Telem, G., Filin, S., 2010. Photogrammetric modeling of underwater environments. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(5), pp. 433–444.

A. AXIAL FLAT-REFRACTIVE UNDERWATER CAMERA MODEL

Flat-refractive imaging systems consist of a camera observing reference points through an interface (Fig. 2). The system is characterized by the refractive interface parameters, including the surface normal, $\mathbf{n} = [0, 0, -1]^T$, and the distance from the camera center to the interface, d; and by the index of refraction of the imaged medium, μ. Without loss of generality, we also consider the index of refraction of the medium in which the camera is stored to equal one. The image ray in the camera frame is defined by $\mathbf{v}_0 = [x, y, f]^T$, where x, y, are the image plane coordinates given in the calibrated camera frame, and f is the principal distance. The ray sequence $[\mathbf{v}_0, \mathbf{v}_1]$, from the perspective center to an object-space point, describes the light traversal within the plane of refraction. We use the vector form of Snell's law of refraction where the direction of the incident ray, $\mathbf{v}_1$, is defined as a function of the emergent ray, $\mathbf{v}_0$, and $\mathbf{n}$ (Hecht, 2002):

$$\mathbf{v}_1 = \xi \mathbf{v}_0 + \delta \mathbf{n} \tag{22}$$

where $\xi = 1/\mu$, and:

$$\delta = -\xi \mathbf{v}_0^T \mathbf{n} - \sqrt{\xi^2 \left(\mathbf{v}_0^T \mathbf{n}\right)^2 + \left(1 - \xi^2\right) \mathbf{v}_0^T \mathbf{v}_0} \tag{23}$$

Rather than tracing the ray trajectory (Eq. 22), a principal distance correction whose aim is to reestablish the collinearity between, $\mathbf{v}_1$ the direction of the incident ray, and, $\mathbf{v}_R$ the direction of the modified image-ray (Fig. 2) is introduced. Such correction is valid as the incident and emergent rays and the optical axis lie on the same plane of refraction. Defining $\mathbf{t}_{\mathbf{bc}_i} = k_i \mathbf{n}$, as the vector form of this offset, where k is a scalar correction to the principal distance, we express the renewed collinearity by the cross-product between the vectors $\mathbf{v}_1$ and $\mathbf{q}_1 - \mathbf{t}_{\mathbf{bc}_i}$, where $\mathbf{q}_1 = \frac{-d}{\mathbf{v}_0^T \mathbf{n}} \mathbf{v}_0$ is the point of refraction at the interface (Fig. 2). Using Eq. (22), we write:

$$\mathbf{v}_1 \times (\mathbf{q}_1 + \Delta \mathbf{c}) = (\xi \mathbf{v}_0 + \delta \mathbf{n}) \times \left(\frac{-d}{\mathbf{v}_0^T \mathbf{n}} \mathbf{v}_\eta - k \mathbf{n} \right) = \mathbf{0} \tag{24}$$

from which the following expressions for k and $\mathbf{v}_R$ can be derived,

$$k = d \left(\frac{1}{\xi} \sqrt{\frac{\left(\mathbf{v}_0^T \mathbf{n}\right)^2 + \left(1 - \xi^2\right) \left\| \mathbf{v}_0 \times \mathbf{n} \right\|^2}{\left(\mathbf{v}_0^T \mathbf{n}\right)^2}} - 1 \right) \tag{25}$$

and,

$$\mathbf{v}_R = \left(x, y, f \left(1 + \tilde{k} \right) \right)^T \tag{26}$$

where $\tilde{k} = k/d$. Next by setting, $t_{\mathbf{bc}_i} = k_i \mathbf{n}$ as the camera position along the system axis and $\mathbf{v}_{R,i}$ as the ray direction, leads to (Elnashef and Filin, 2020):

$$\begin{aligned} \mathbf{L} &= \begin{pmatrix} \mathbf{v}_R^T & -\left(k [\mathbf{n}]_\times \mathbf{v}_R\right)^T \end{pmatrix}^T \\ &= \begin{pmatrix} x & y & f(1+\tilde{k}) & -ky & kx & 0 \end{pmatrix}^T \end{aligned} \tag{27}$$

Note that, Eq. (27) defines a linear form of the ray trajectory through refraction which is expressed by image-related quantities only. Substituting the Plücker line-coordinates from Eq. (27) in Eq. (22) leads to:

$$\left(\mathbf{v}_i \times \mathbf{R} \mathbf{v}_i'\right) \mathbf{t} + \mathbf{v}_i^T \left(k' \mathbf{R} [\mathbf{n}]_\times - k [\mathbf{n}]_\times \mathbf{R}\right) \mathbf{v}_i' = 0 \tag{28}$$

with, the axial epipolar plane normal vector defined as follows:

$$\mathbf{g}_{Ri} = \begin{pmatrix} \mathbf{v}_i \times \mathbf{R} \mathbf{v}_i' \\ \mathbf{v}_i^T \left(k' \mathbf{R} [\mathbf{n}]_\times - k [\mathbf{n}]_\times \mathbf{R}\right) \mathbf{v}_i' \end{pmatrix} \tag{29}$$