Refractive Structure-from-Motion on Underwater Images

Anne Jordt-Sedlazeck and Reinhard Koch
Institute of Compute Science, Kiel University, Germany
{sedlazeck,rk}@mip.informatik.uni-kiel.de

Abstract

In underwater environments, cameras need to be confined in an underwater housing, viewing the scene through a piece of glass. In case of flat port underwater housings, light rays entering the camera housing are refracted twice, due to different medium densities of water, glass, and air. This causes the usually linear rays of light to bend and the commonly used pinhole camera model to be invalid. When using the pinhole camera model without explicitly modeling refraction in Structure-from-Motion (SfM) methods, a systematic model error occurs. Therefore, in this paper, we propose a system for computing camera path and 3D points with explicit incorporation of refraction using new methods for pose estimation. Additionally, a new error function is introduced for non-linear optimization, especially bundle adjustment. The proposed method allows to increase reconstruction accuracy and is evaluated in a set of experiments, where the proposed method's performance is compared to SfM with the perspective camera model.

1. Introduction

In the last decade, many applications for images captured underwater arose. They include scientific exploration of geological or archaeological structures on the sea floor [2], maintenance of offshore oil rigs, inspection of ship hulls, and measurements of ships and other fisheries [6]. Due to the need of gaining measurements in the above described scenarios, the geometry of image formation is often utilized. However, cameras used in an underwater environment are usually confined in an underwater housing filled with air, viewing the scene through a piece of glass. In case of this glass being a flat port, the light rays entering the camera housing are refracted twice, once at the water-glass interface and again at the glass-air interface. Many of the above described applications require the camera to be lowered into the deep sea, sometimes to water depths of thousands of meters. Therefore, the underwater housing needs to be strong enough to withstand immense water pressures, requiring the glass interface to be several centimeters thick. The double refraction causes the usually straight rays of light to bend and change direction depending on the interface incidence angles. When following the ray in water in Figure 1 without refraction (dashed line), it does not intersect the camera center. In fact, Treibitz *et al*. [28] showed that the perspective camera model is invalid below water due to the rays not intersecting in one common center of projection. Despite that, the perspective camera model is often used for underwater images, approximating the refractive effect to some extent. For example Lavest *et al*. [18] showed that a camera calibrated below water approximates refraction with focal length and radial distortion and Sedlazeck and Koch [25] showed that principal point and camera pose absorb some of this model error in addition to focal length and radial distortion. Due to the perspective model being invalid, a systematic model error is introduced, when applying perspective algorithms utilizing imaging geometry like mosaicing or Structure-from-Motion (SfM) [9, 27] to underwater images. Even though, several works can be found in the literature, where the perspective camera model is used to reconstruct 3D scenes in underwater environments (*e.g*. [3, 12, 15]).

In contrast to using the perspective camera model in order to approximate refraction, refraction can also be modeled explicitly, where first a parametrization of the glass port of the housing needs to be found and calibrated. An early approach was introduced by Li *et al*. [19] coming from the area of photogrammetry where the housing of a stereo rig can be calibrated. The calibration routine proposed by Treibitz *et al*. [28] assumes a flat port interface with very thin glass and parallelism between glass and imaging sensor. More recently, Agrawal *et al*. [1] showed how a more general camera with thick glass and a possible inclination angle between glass interface and imaging sensor can be calibrated, and Jordt-Sedlazeck *et al*. [14] proposed a non-linear optimization based on an initialization with Agrawal's method. Building upon a valid calibration of an underwater camera, meaning the intrinsics and a housing parametrization are known, several approaches to refractive SfM exist. Chari *et al*. [5] derive a complete theoretical framework, however it has never been implemented. Chang

et al. [4] proposed a method for refractive SfM, where the camera views a scene at the bottom of a pool through the water surface and the camera's yaw and pitch with respect to the water surface are assumed to be known. The most recent work of Kang *et al*. [16] showed results for 3D reconstruction with relative pose between two images with explicit incorporation of refraction. They rely on outlier-free correspondences, which have to be selected manually and glass thickness is not modeled explicitly. The system cannot handle image sequences and because of the use of the reprojection error during bundle adjustment, it cannot be extended easily.

Our Contribution: In this paper, we propose a more general method for refractive SfM that can evaluate video sequences with more general patterns of movement compared to [4]. The main problem to overcome is that due to refraction, the computation of the refractive re-projection error is infeasible in large non-linear optimization problems like bundle adjustment [29]. Therefore we propose a new error function that can be computed efficiently and even enables the analytic derivation of the necessary Jacobian matrices of the error function. In addition, we propose new methods for relative and absolute pose computation. Finally, a refractive plane sweep proposed in [13] is used to estimate dense depth maps for each view, which are then used to create the final 3D model. Controlled experiments show that the proposed method performs better than a comparable perspective method, where the refractive effect is only approximated.

2. Refractive Camera Model and Non-linear Error Function

The camera model is the standard pinhole camera model with distortion [9, 27]. Hence the camera's intrinsics are defined in the camera matrix $\mathbf{K}$ containing focal length f, aspect ratio a, and a principal point (c_x, c_y), complemented by two coefficients for radial distortion r_1 and r_2 and two coefficients for tangential distortion t_1 and t_2. The camera's extrinsics are the rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{C}$, resulting in the projection matrix $\mathbf{P} = \mathbf{K}[\mathbf{R}^\mathrm{T}| - \mathbf{R}^\mathrm{T}\mathbf{C}]$. Those parameters allow to project 3D points onto 2D image points, but also to back-project 2D image points onto 3D rays. Refraction at the underwater housing is described by Snell's law [11] and depends on the different medias' indexes of refraction n_a for air, n_g for glass, and n_w for water. As seen in Figure 1, the rays coming from the water do not intersect in the camera's center of projection. However, [1] determined that a camera behind a flat port underwater housing is an axial camera, *i.e*. all rays coming from the water intersect a common axis defined by the camera center and the interface normal (blue line in Fig. 1). Moreover, all segments of the light ray $\mathbf{r}_\mathrm{a}$ in air, $\mathbf{r}_\mathrm{g}$ in glass, and $\mathbf{r}_\mathrm{w}$ in water, and the interface normal $\mathbf{n}$ lie in one common plane, the Plane of Refraction POR (pale blue plane in Fig. 1). In order to back-project a ray from a 2D image point, the ray in air $\mathbf{r}_\mathrm{a}$ is determined using the perspective parameters explained above. Then, the ray direction in glass $\mathbf{r}_\mathrm{g}$ is computed by [1]:

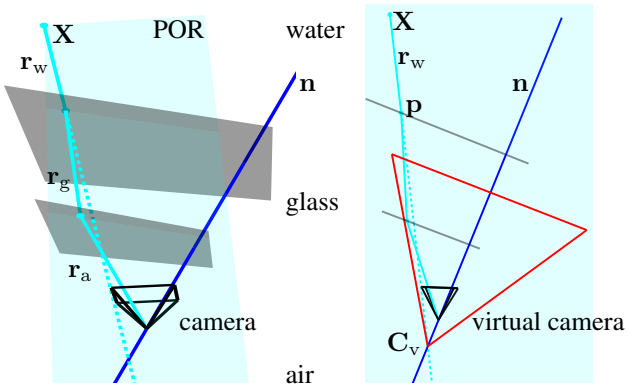

Figure 1. Left: refractive projection with ray segments in water $\mathbf{r}_\mathrm{w}$, glass $\mathbf{r}_\mathrm{g}$, and air $\mathbf{r}_\mathrm{a}$. All ray segments together with the interface normal lie in a common plane, the Plane of Refraction (POR). The blue line depicts the interface normal passing through the center of projection which is intersected by all rays $\mathbf{r}_\mathrm{w}$ [1] (dashed line). Right: projection onto the POR. The virtual camera's center $\mathbf{C}_\mathrm{v}$ is located at the intersection of the un-refracted ray (dashed line) and the interface normal (blue) and its focal length is $f_\mathrm{v} = d$.

$$\mathbf{r}_\mathrm{g} = \frac{n_\mathrm{a}}{n_\mathrm{g}}\mathbf{r}_\mathrm{a} + \left(-\frac{n_\mathrm{a}}{n_\mathrm{g}}\mathbf{r}_\mathrm{a}^\mathrm{T}\mathbf{n} + \sqrt{1 - \frac{n_\mathrm{a}}{n_\mathrm{g}}(1 - (\mathbf{r}_\mathrm{a}^\mathrm{T}\mathbf{n})^2)}\right)\mathbf{n}\,. \tag{1}$$

Using $\mathbf{r}_\mathrm{g}$, n_g, and n_w, the ray in water $\mathbf{r}_\mathrm{w}$ is computed respectively. Along with the interface distance d and the interface thickness d_g, $\mathbf{r}_\mathrm{a}$ and $\mathbf{r}_\mathrm{g}$ allow to determine a starting point $\mathbf{p}$ of the ray $\mathbf{r}_\mathrm{w}$ on the outer glass plane (Fig. 1):

$$\mathbf{p} = \frac{d}{\mathbf{n}^\mathrm{T}\mathbf{r}_\mathrm{a}}\mathbf{r}_\mathrm{a} + \frac{d_\mathrm{g}}{\mathbf{n}^\mathrm{T}\mathbf{r}_\mathrm{g}}\mathbf{r}_\mathrm{g}. \tag{2}$$

Hence for each pixel, a raxel [8] can be computed using the proposed parameters, instead of calibrating each raxel independently, which is often difficult. Using the proposed parameter set, [1] derived two constraints for the flat port underwater camera. The first one is called the Flat Refractive Constraint (FRC) and states that if a 3D point $\mathbf{X}$ has been transformed into the local camera coordinate system, its direction should be the same as the ray in water $\mathbf{r}_\mathrm{w}$, hence:

$$(\mathbf{R}^\mathrm{T}\mathbf{X} - \mathbf{R}^\mathrm{T}\mathbf{C} - \mathbf{p}) \times \mathbf{r}_\mathrm{w} = \mathbf{0} \quad \text{(FRC)}. \tag{3}$$

From the POR follows that:

$$(\mathbf{R}^\mathrm{T}\mathbf{X} - \mathbf{R}^\mathrm{T}\mathbf{C})^\mathrm{T}(\mathbf{n} \times \mathbf{r}_\mathrm{w}) = 0 \quad \text{(PORC)}. \tag{4}$$

2.1. Virtual Camera Error Function

When projecting a 3D point into a camera confined in an underwater housing with explicit refraction computation, Agrawal *et al*. [1] determined that a 12^{th} degree polynomial needs to be solved. While this insight allows solving the projection problem much more efficiently than previous approaches, where usually the projection was determined by an optimization using the back-projection function [17], it is still infeasible in classic SfM, especially bundle adjustment. Therefore, a new error function is introduced. It builds upon the idea in [24], where a virtual camera is defined for each 2D point into which the corresponding 3D point can be projected perspectively (Fig. 1 on the right). Note that a similar idea has been expressed in [23], however, the proposed method is adapted to the refractive case and is exact for each pixel. The virtual camera error is computed using the ray in water $\mathbf{r}_\mathrm{w}$ and its starting point $\mathbf{p}$ as described above to define a virtual perspective camera. However, in contrast to [24], we propose using the intersection of the ray in water with the line defined by the interface normal $\mathbf{n}$ and the camera coordinate system origin to define the virtual camera center $\mathbf{C}_\mathrm{v}$ (the camera's axis as in [1]), therefore solving for the scaling factor λ in $\mathbf{p} + \lambda\mathbf{r}_\mathrm{w} = \mathbf{n}$. The virtual rotation $\mathbf{R}_\mathrm{v}$ is defined through its rotation axis, which is the cross product between interface normal and optical axis and its rotation angle, which is the scalar product between interface normal and optical axis. The virtual focal length is $f_\mathrm{v} = d$. Thus, a 3D point $\mathbf{X}$ in the global coordinate system is first transformed into a point in the local camera coordinate system $\mathbf{X}_\mathrm{l}$ and then into the virtual camera coordinate system $\mathbf{X}_\mathrm{v}$ by:

$$\mathbf{X}_\mathrm{l} = \mathbf{R}^\mathrm{T}\mathbf{X} - \mathbf{R}^\mathrm{T}\mathbf{C} \tag{5}$$

$$\mathbf{X}_\mathrm{v} = \mathbf{R}_\mathrm{v}^\mathrm{T}\mathbf{X}_\mathrm{l} - \mathbf{R}_\mathrm{v}^\mathrm{T}\mathbf{C}_\mathrm{v}. \tag{6}$$

The starting point on the outer interface is also transformed into the virtual camera:

$$\mathbf{p}_\mathrm{v} = \mathbf{R}_\mathrm{v}^\mathrm{T}\mathbf{p} - \mathbf{R}_\mathrm{v}^\mathrm{T}\mathbf{C}_\mathrm{v}. \tag{7}$$

The error is then computed from the 2D projections of $\mathbf{X}_\mathrm{v}$ and $\mathbf{p}_\mathrm{v}$ onto the virtual image plane:

$$\mathbf{g}_\mathrm{v} = \begin{pmatrix} \frac{f_\mathrm{v}}{X_{\mathrm{v}_z}} X_{\mathrm{v}_x} - \frac{f_\mathrm{v}}{p_{\mathrm{v}_z}} p_{\mathrm{v}_x} \\ \frac{f_\mathrm{v}}{X_{\mathrm{v}_z}} X_{\mathrm{v}_y} - \frac{f_\mathrm{v}}{p_{\mathrm{v}_z}} p_{\mathrm{v}_y} \end{pmatrix}. \tag{8}$$

$\mathbf{g}_\mathrm{v}$ can be used as a non-linear error function for optimization with different parametrizations. For example considering one camera and a set of n 2D-3D correspondences, when only extrinsic parameters and 3D points are unknown, the known ray in water and the virtual camera center are used:

$$\epsilon_\mathrm{ext} = \sum_{i<n} \parallel \mathbf{g}_{\mathrm{v}i}(\mathbf{C}, \mathbf{R}, \mathbf{X}_i, \mathbf{p}_i, \mathbf{r}_{\mathrm{w}_i}, \mathbf{C}_{\mathrm{v}_i}) \parallel_2^2 . \tag{9}$$

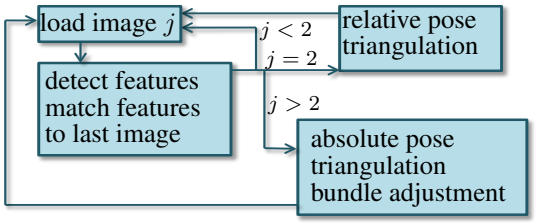

Figure 2. For both, perspective and refractive SfM, a basic SfM algorithm is used, where two images are used for initialization using relative pose and triangulation. Once 3D points exist and 2D-3D correspondences can be matched, absolute pose is determined and the whole scene is refined using bundle adjustment.

If extrinsics, 3D points, interface distance d, and interface normal are unknown, then the ray in air $\mathbf{r}_{\mathrm{a}_i}$ is used for error computation:

$$\epsilon_\mathrm{ext+housing} = \sum_{i<n} \parallel \mathbf{g}_\mathrm{v}(\mathbf{C}, \mathbf{R}, \mathbf{X}_i, d, \mathbf{n}, \mathbf{r}_{\mathrm{a}_i}) \parallel_2^2 . \tag{10}$$

In case the relative pose between two cameras is optimized, but no 3D points exist, 2D-2D correspondences between the two images are used to triangulate the 3D points [10] allowing the use of the following error function:

$$\epsilon_\mathrm{ext2View} = \sum_{i<n} \parallel \mathbf{g}_\mathrm{v}(\mathbf{C}, \mathbf{R}, \mathbf{p}_i, \mathbf{r}_{\mathrm{w}_i}, \mathbf{C}_{\mathrm{v}_i}, \mathbf{p'}_i, \mathbf{r'}_{\mathrm{w}_i}, \mathbf{C'}_{\mathrm{v}_i}) \parallel_2^2, \tag{11}$$

where $\mathbf{p}_i$, $\mathbf{r}_{\mathrm{w}_i}$, $\mathbf{C}_{\mathrm{v}_i}$ and $\mathbf{p'}_i$, $\mathbf{r'}_{\mathrm{w}_i}$, and $\mathbf{C'}_{\mathrm{v}_i}$ describe ray and virtual camera center in the first and second image for correspondence i respectively.

3. Refractive SfM

A typical reconstruction system is depicted in Fig. 2. In two images, features are detected and matched, then the relative pose of the second camera with respect to the first is computed. Next, the 2D-2D correspondences and camera poses are used for triangulation [10]. This allows to find 2D-3D correspondences for the next image, hence the absolute pose with respect to the 3D points is computed. After adding a new image and triangulating new points, nonlinear optimization is applied to the scene.

3.1. Relative Pose

At the beginning of the reconstruction process, no 3D points are known, only a set of n 2D-2D correspondences between two images can be matched. Therefore, these correspondences are used to compute the relative pose of the second camera with respect to the first. The first camera is set into the world coordinate system origin. Let $\mathbf{p}_i$ and $\mathbf{r}_{\mathrm{w}_i}$ be the ray for the i^{th} 2D point in image one and $\mathbf{p'}_i$ and $\mathbf{r'}_{\mathrm{w}_i}$ be the corresponding ray for the 2D point in the second image. Then, two constraints can be used for determining the

unknown rotation $\mathbf{R}$ and translation $\mathbf{C}$:

$$\mathbf{p}_i + \lambda_i \mathbf{r}_{\mathrm{w}_i} = \mathbf{R}\mathbf{p'}_i + \mathbf{C} + \lambda'_i \mathbf{R}\mathbf{r'}_{\mathrm{w}_i} \quad (12)$$

$$(\mathbf{R}\mathbf{p'}_i + \mathbf{C} + \lambda'_i \mathbf{R}\mathbf{r'}_{\mathrm{w}_i} - \mathbf{p}_i) \times \mathbf{r}_{\mathrm{w}_i} = \mathbf{0}, \quad (13)$$

where the first constraint describes the triangulation of the corresponding but unknown 3D point and the second constraint is derived from the FRC (3). λ_i and λ'_i are the unknown lengths of the rays in water $\mathbf{r}_{\mathrm{w}_i}$ and $\mathbf{r'}_{\mathrm{w}_i}$. Hence, a set of equations for all n correspondences with both (12) and (13) is non-linear in the unknowns. Consequently, an alternating, iterative scheme is used, where the equation system is solved for $\mathbf{R}$ and $\mathbf{C}$ and the λ_i and λ'_i are updated by solving for λ_i and λ'_i in the PORC (4):

$$(\mathbf{R}\mathbf{p'}_i + \mathbf{C} + \lambda'_i \mathbf{R}\mathbf{r'}_{\mathrm{w}_i})^{\mathrm{T}} (\mathbf{n} \times \mathbf{r}_{\mathrm{w}_i}) = 0 \quad (14)$$

$$(\mathbf{R}^{\mathrm{T}}(\mathbf{p}_i + \lambda_i \mathbf{r}_{\mathrm{w}_i}) - \mathbf{R}^{\mathrm{T}}\mathbf{C})^{\mathrm{T}} (\mathbf{n} \times \mathbf{r'}_{\mathrm{w}_i}) = 0.$$

The initial solution gained by this iterative scheme is often still quite far from the true relative pose. However, we found that it is a good initial estimate for a Levenberg-Marquardt optimization using the virtual camera error $\epsilon_{\mathrm{ext2View}}$ (11). Both iterative and optimizer schemes are applied within a RANSAC framework [7] in order to be robust against outliers in the data.

Note that due to the rays starting on the glass interface being metric, in theory, relative pose as described here on underwater cameras can yield the absolute distance to the second camera as opposed to perspective relative pose, where the baseline to the second camera is usually scaled to one. This is due to the fact that common perspective scenes can be rescaled consistently by applying a scaling transform $\mathbf{T}$ to all 3D points and its inverse $\mathbf{T}^{-1}$ to the extrinsics of all projection matrices $\mathbf{P}$. However, in the refractive camera model interface distance and thickness would need to be scaled along with the 3D points and the camera's extrinsics, thereby changing the starting points and directions of the rays corresponding to image points. Consequently, relative pose in the refractive case is not invariant to scale changes in the camera translation. In case of synthetic correspondences with zero noise we found this to be true. However, in case of small amounts of noise added to the correspondences, we were not able to determine the scale of the translation, due to the noise superimposing the signal, *i.e.* the rays' starting points are only a few millimeters apart, but the camera movement and typical 3D point distances are in the order of centimeters and meters respectively. In Figure 3, the virtual camera error function is depicted for a pair of cameras, with a random set of correspondences. The curves depict the sensitivity of the error function to changes of the translation scale for different noise levels, while keeping everything else constant. In case of zero noise, a minimum at the correct scale is clearly visible. However, even in case of noise added to the 2D correspondences with $\sigma = 0.1$,

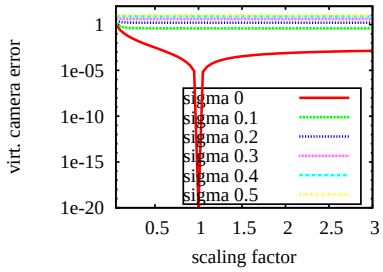

Figure 3. For two cameras random 2D-2D correspondences were used to compute the virtual camera error for different scales of the camera translation. Since the rays were not scaled, the error should increase for erroneous scales of the camera translation and have a clear minimum at 1, which is true for zero noise. However, as soon as noise is added to the correspondences, the error function does not have a minimum at the correct scale anymore. Hence, the retrieval of correct scale can only be achieved theoretically.

it is clear, that scale cannot be determined. The same was true for the re-projection error or the angle error proposed in [21].

3.2. Absolute Pose

Once 3D points exist, 2D-3D correspondences can be found for an underwater image. This allows computing the absolute pose with respect to the 3D points for which the following equation is utilized for each point i:

$$\mathbf{X}_i = \mathbf{R}\mathbf{p}_i + \mathbf{C} + \lambda_i \mathbf{R}\mathbf{r}_{\mathrm{w}_i}, \quad (15)$$

with $\mathbf{R}$ and $\mathbf{C}$ being the unknown pose and λ_i being the distance between outer interface point and 3D point, which is also unknown. Hence, (15) is non-linear in the unknowns. As in the relative pose case, the absolute pose problem is solved by an alternating iteration, where a linear system of equations with all 2D-3D correspondences is solved for $\mathbf{R}$ and $\mathbf{C}$ using (15). Then, all λ_i are updated by:

$$\lambda_i = (\mathbf{R}\mathbf{r}_{\mathrm{w}_i})^{\mathrm{T}} (\mathbf{X}_i - \mathbf{C} - \mathbf{R}\mathbf{p}_i). \quad (16)$$

The resulting initial solution is optimized using the Levenberg-Marquardt algorithm with the above described virtual camera error function ϵ_{ext} (9). Both iterative and non-linear optimization are used within a RANSAC framework. We found this method to compute absolute pose more robustly than the methods described by Nistér and Stewénius in [22] or Sturm *et al.* in [26], whose solutions work on a minimum number of 3 required 2D-3D correspondences. The reason is that underwater images suffer from degraded contrast. Consequently, the correspondences have a certain amount of noise, and hence a method that is more robust against noise but may require more than the minimal number of 3 correspondences can outperform [22] and [26] within a RANSAC framework.

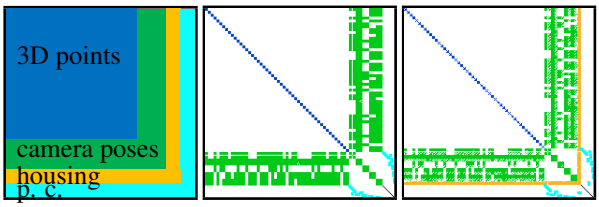

Figure 4. Sparse matrices for refractive bundle adjustment. Left: general block structure of sparse matrix N. Middle: optimization of 3D points and camera poses. Note that only colored parts are non-zero. Right: additional optimization of housing parameters. Colors: 3D points blue, poses green, housing parameters orange, constraints between parameters cyan.

3.3. Refractive Bundle Adjustment

When the scene consists of camera poses and 3D points, bundle adjustment [29] is applied after each newly added camera, optimizing scene geometry by minimizing a non-linear error function. In classical bundle adjustment, the re- projection error can be computed efficiently and is minimized in the classical iterative least squares solver. In order to achieve that, each 3D point is projected into each camera that saw this point numerous times. The Jacobian of the explicit error function is derived preferably analytically in the direction of all parameters. Due to these requirements, the use of the re-projection error in case of refractive bundle adjustment is infeasible, even if Agrawal's 12^{th} degree polynomial is used. However, in contrast to the refractive projection, the virtual camera error function can be computed efficiently and its analytic derivatives can also be computed[1]. With $\mathbf{t}$ containing all camera poses and all 3D points, hence all parameters and $\mathbf{l}$ containing all observations, in this case all rays with starting point, direction, and virtual camera center, the virtual camera error function $\mathbf{g}_{\mathrm{v}}(\mathbf{t}, \mathbf{l}) = \mathbf{0}$ is an implicit constraint for bundle adjustment. Consequently, the minimization problem is solved using the Gauss-Helmert model [20, 29], a generalization to the commonly known least squares solver. Note, that for implicit constraints the derivatives in parameter direction $\mathbf{t}$ and observation direction $\mathbf{l}$ need to be computed. The use of analytic Jacobi matrices allows fast and accurate computations in case of the virtual camera error. In addition, as in common perspective bundle adjustment, the parameter ordering in the sparse matrix can be arranged such that the Schur complement [29] can be used to efficiently solve the linear system of equations in each iteration (Fig. 4). Because of using the axis intersection instead of the caustic point, the analytic derivations of the error function, and the application of the Schur complement for solving the system of equations in each iteration, the proposed adjustment method can run in seconds rather than hours as mentioned in [24].

Note that the described approach to bundle adjustment with the virtual camera error function can be used to optimize camera poses and 3D points in case the observations are vectors with nine entries with ray starting point on outer interface $\mathbf{p}$, ray direction $\mathbf{r}_{\mathrm{w}}$, and axis intersection $\mathbf{C}_{\mathrm{v}}$. However, in case of the observations being the normalized rays in air $\mathbf{r}_{\mathrm{a}}$, the virtual camera error can also be computed, thus interface distance d and interface normal $\mathbf{n}$ can also be optimized efficiently (compare to Eq. (9) and (10)).

[1] *e.g.* using Maxima (`http://maxima.sourceforge.net/`)

4. Experiments

In addition to the refractive SfM presented here, we also implemented a classical perspective SfM, were a perspective calibration approximated the underwater conditions and we will compare both approaches. Initializations for all λ_i and λ'_i in relative pose estimation were set to 3000 mm, which is far enough for the distance dependent error to be initialized robustly. In case of absolute pose it was sufficient to set all initial λ_i to one.

4.1. Synthetic Images

In order to evaluate the performance of refractive SfM, sets of synthetically rendered images with varying glass port configurations were used. The image size was 800×600 pixels in all cases with a focal length of 800 pixels. The principal point was in the middle, and no radial distortion was set. The interface thickness was fixed to $d_g = 30\,\mathrm{mm}$, the interface distance was chosen from $d \in \{-5\,\mathrm{mm}, 0\,\mathrm{mm}, 5\,\mathrm{mm}, 10\,\mathrm{mm}, 20\,\mathrm{mm}, 50\,\mathrm{mm}, 100\,\mathrm{mm}\}$. The interface normal was determined by its two angles θ_1 and θ_2, where θ_2 is the angle between optical axis and normal and θ_1 is the angle by which the interface is turned around the normal. θ_1 was set fixed to $30°$, while $\theta_2 \in \{0°, 0.5°, 1°, 3°\}$. This gives 28 different configurations for which sets of images where rendered. Since the calibrated approximation of the perspective camera model is distance dependent, we rendered two different scenes, one within the calibration distance between 1000 mm and 4000 mm and one being further away. The top rows in in Figures 5 and 6 show exemplary input images and ground truth depth maps. In the second row are the average errors of the 3D points compared to ground truth, the third row shows the average camera translation error, and the last row shows the re-projection error. The first column shows the results for using the perspective camera on underwater images, while the second column depicts the results for using the proposed method. As can be seen, the error of the refractive method is lower than with the perspective camera method. With increasing interface distance d and increasing interface tilt θ, the perspective approximation of refraction becomes less accurate. Hence, the reconstruction error increases. In case of the proposed method, where the refractive effect is modeled explicitly, the reconstruction

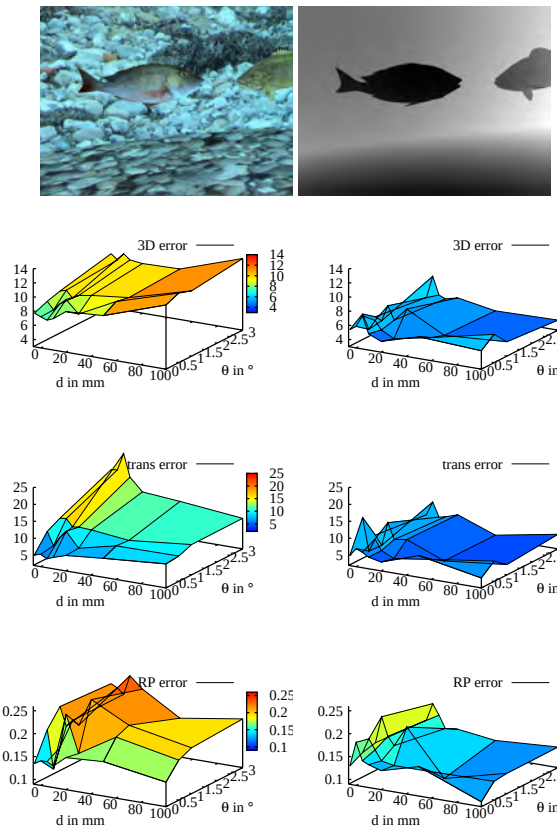

Figure 5. Top row left: exemplary input image. Top row right: ground truth depth map, scene-camera distance was between 1000 mm and 4000 mm. Rows 2-4: In the left column are the results of perspective SfM on underwater images, the right column depicts the results from refractive SfM. In the second row is the average error of the 3D points compared to ground truth. The third row depicts the translation error of the resulting camera poses, while the bottom row shows the average re-projection error. All experiments were conducted on the same camera movement, only the interface distance d and the interface tilt θ changed.

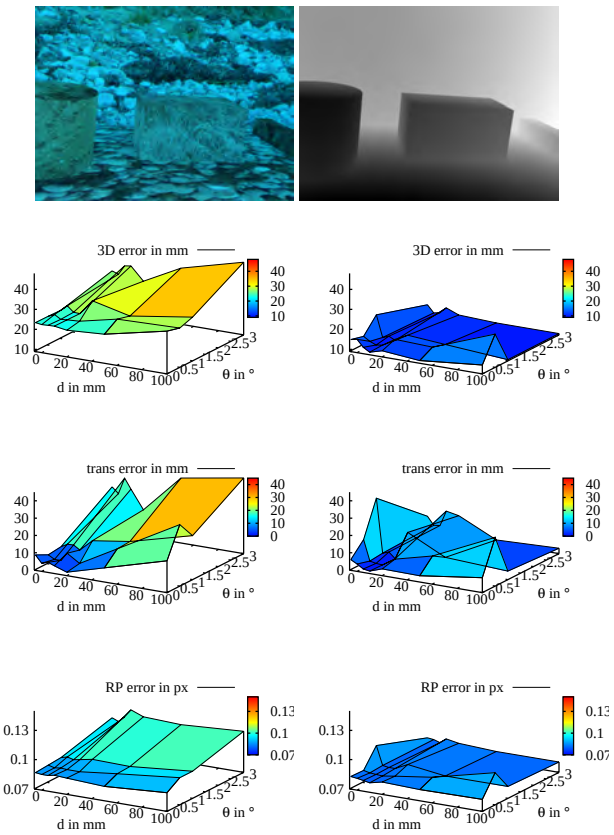

Figure 6. In the top row are exemplary input image and ground truth depth map of the scene. The scene-camera distance was between 4000 mm and 9000 mm. Rows 2-4 show the same analysis as in Figure 5.

errors do not increase with increasing interface distance or tilt.

4.2. Real Images

In order to test the described system on real images in a controlled environment, a fish tank of the size $500\,\mathrm{mm} \times 500\,\mathrm{mm} \times 1000\,\mathrm{mm}$ was filled with water and a camera was placed outside of it. The captured scene shows a scaled model of the entrance to the Abu Simbel temple in Egypt with a size of about $380\,\mathrm{mm} \times 280\,\mathrm{mm} \times 180\,\mathrm{mm}$. Since the camera was not allowed to move with respect to the glass while capturing the images, the Abu Simbel model was moved inside the water at a distance range between 300 mm and 750 mm from the camera. Consequently, the back ground of the tank violated the rigid scene constraint, which is why all images were roughly segmented in order to eliminate errors caused by features on the background or the mirrored object at the bottom or the sides of the tank (see Fig. 7 for exemplary original and segmented images). Four data sets with different camera-glass configurations (d, $\mathbf{n}$) were captured and used for reconstruction with the refractive and the perspective method. Calibrations for both methods were achieved using checkerboard images. As can be seen in Figure 8 and Table 1 the distance between perspectively and refractively reconstructed camera poses increases with increasing interface distance and interface tilt, indicating the influence of the systematic model error of the perspective approximation. Note that the gaps in the camera path really occurred because the model was moved manually in the water.

The main areas of application for our system are not small scale fish tanks, but cameras used in deep sea scenarios, where often very thick glass needs to be placed in front of the camera in order to withstand the immense water pressure. Therefore, we present some preliminary reconstruction results for images captured of an underwater volcano near the Cape Verdes at water depths of about 3000 m.

Figure 7. Abu Simbel original image and segmented image. Note the mirrored scene at the tank bottom. In the background, erroneous correspondences were detected and matched as well, making a rough segmentation as seen on the right necessary.

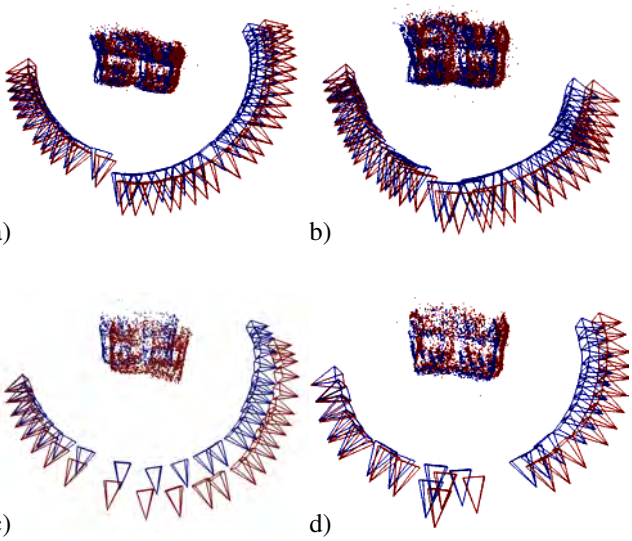

Figure 8. Abu Simbel results. Blue: refractive reconstruction. Red: perspective reconstruction. At the beginning, both camera paths are nearly identical. With an increasing number of cameras added to the reconstruction, the paths diverge.

case	d in mm	θ in $^\circ$	$\varnothing$ in mm	sd in mm
a	68	1.35	36.9476	19.9553
b	81	0.87	34.1247	17.908
c	50	8.96	102.804	58.0293
d	114	0.36	48.5285	27.9537

Table 1. Abu Simbel reconstruction results of four different camera-glass configurations. d is the calibrated interface distance, θ the angle between optical axis and interface normal $\mathbf{n}$. $\varnothing$ in mm and sd in mm are the average distance and standard deviation between perspectively and refractively reconstructed camera translation respectively.

In this case, we only had an intrinsic calibration of the camera made months after the expedition and no calibration of the interface or perspective calibration. Therefore, we auto-calibrate the interface normal and distance using bundle adjustment and present the refractive underwater reconstructions of four short sequences in Figure 9. Table 2 summarizes results of the auto-calibration of the camera housing. Although there is no means of determining the calibration error, the results of the first three runs are very close to each other, indicating a successful calibration.

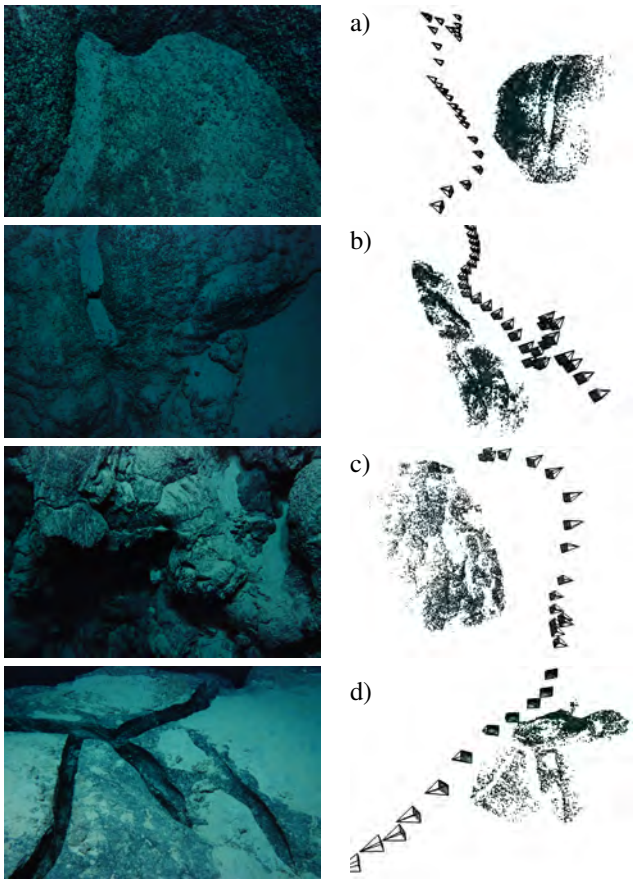

Figure 9. Deep sea volcano crater reconstructions of underwater volcano near the Cape Verdes at approximately 3000 m water depth. Left: exemplary input image (by Geomar Helmholtz Centre for Ocean Research), right: SfM result.

case	d in mm	θ in $^\circ$
a	48	1.23
b	20	1.24
c	38	1.29
d	47	3.15

Table 2. Auto-calibration results for the underwater housing for the four volcano sequences. d is the calibrated interface distance, which should roughly be the same in all four cases. θ, the angle between optical axis and interface normal $\mathbf{n}$, can vary to some extend.

5. Conclusion and Future Work

We presented a system for reconstruction with explicit incorporation of refraction at an underwater housing allowing to accurately reconstruct underwater scenes captured through a flat port even in deep sea scenarios, where the glass can be several centimeters thick. In order to avoid having to project 3D points, we introduced an error function using a virtual camera that allows to efficiently compute bun-

dle adjustment. Our system is also capable to auto-calibrate the glass interface using bundle adjustment, if the initial estimation is not too far off. It would be interesting to further investigate accuracy and limitations of auto-calibrating the housing parameters, therefore eliminating the need to capture checkerboard images below water, which is at best impractical in oceanographic applications.

Acknowledgments This work has been supported by the German Science Foundation (KO 2044/6-1/2: 3D Modeling of Seafloor Structures from ROV-based Video Sequences).

References

[1] A. Agrawal, S. Ramalingam, T. Y., and V. Chari. A theory of multi-layer flat refractive geometry. In *CVPR*, 2012. 1, 2, 3

[2] B. Bingham, B. Foley, H. Singh, R. Camilli, D. K., R. Eustice, A. Mallios, D. Mindell, C. Roman, and D. Sakellariou. Robotic tools for deep water archaeology: Surveying an ancient shipwreck with an autonomous underwater vehicle. *J. Field Robotics*, 27:702–717, 2010. 1

[3] V. Brandou, A. Allais, M. Perrier, E. Malis, P. Rives, J. Sarrazin, and P. Sarradin. 3d reconstruction of natural underwater scenes using the stereovision system iris. In *Proc. OCEANS 2007 - Europe*, pages 1–6, 2007. 1

[4] Y.-J. Chang and T. Chen. Multi-view 3d reconstruction for scenes under the refractive plane with known vertical direction. In *IEEE International Conference on Computer Vision (ICCV)*, 2011. 2

[5] V. Chari and P. Sturm. Multiple-view geometry of the refractive plane. In *Proceedings of the 20th British Machine Vision Conference, London, UK*, sep 2009. 1

[6] C. Costa, A. Loy, S. Cataudella, D. Davis, and M. Scardi. Extracting fish size using dual underwater cameras. *Aquacultural Engineering*, 35(3):218 – 227, 2006. 1

[7] M. Fischler and R. Bolles. RANdom SAmpling Consensus: a paradigm for model fitting with application to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, June 1981. 4

[8] M. Grossberg and S. Nayar. The raxel imaging model and ray-based calibration. *International Journal of Computer Vision*, 61(2):119–137, 2005. 2

[9] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision (Second Edition)*. Cambridge University Press, second edition, 2004. 1, 2

[10] R. I. Hartley and P. F. Sturm. Triangulation. *Computer Vision and Image Understanding*, 68(2):146–157, 1997. 3

[11] E. Hecht. *Optics*. Addison-Wesley, 4th edition, 1998. 2

[12] M. Johnson-Roberson, O. Pizarro, S. Williams, and I. Mahon. Generation and visualization of large-scale three-dimensional reconstructions from underwater robotic surveys. *Journal of Field Robotics*, 27, 2010. 1

[13] A. Jordt-Sedlazeck, D. Jung, and R. Koch. Refractive plane sweep for underwater images. In *Pattern Recognition*, volume 8142 of *LNCS*, pages 333–342. 2013. 2

[14] A. Jordt-Sedlazeck and R. Koch. Refractive calibration of underwater cameras. In *Computer Vision, ECCV 2012*, volume 7576 of *LNCS*, pages 846–859. 2012. 1

[15] L. Kang, L. Wu, and Y.-H. Yang. Experimental study of the influence of refraction on underwater three-dimensional reconstruction using the svp camera model. *Applied Optics*, 51(31):7591–7603, Nov 2012. 1

[16] L. Kang, L. Wu, and Y.-H. Yang. Two-view underwater structure and motion for cameras under flat refractive interfaces. In *European Conference on Computer Vision (ECCV)*, volume 7575 of *LNCS*, pages 303–316. 2012. 2

[17] C. Kunz and H. Singh. Hemispherical refraction and camera calibration in underwater vision. In *OCEANS 2008*, pages 1 –7, 15-18 2008. 3

[18] J.-M. Lavest, G. Rives, and J.-T. Lapresté. Underwater camera calibration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 654–668, 2000. 1

[19] R. Li, H. Li, W. Zou, R. Smith, and T. Curran. Quantitative photogrammetric analysis of digital underwater video imagery. *Oceanic Engineering, IEEE Journal of*, 22(2):364 –375, apr 1997. 1

[20] J. C. McGlone, editor. *Manual of Photogrammetry*. ASPRS, 5th edition, 2004. 5

[21] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd. Generic and real-time structure from motion using local bundle adjustment. *Image and Vision Computing*, 27:1178–1193, 2009. 4

[22] D. Nistér and H. Stewénius. A minimal solution to the generalised 3-point pose problem. *Journal of Mathematical Imaging and Vision*, 27:67–79, 2007. 4

[23] S. Ramalingam, S. K. Lodha, and P. Sturm. A generic structure-from-motion framework. *Comput. Vis. Image Underst.*, 103(3):218–228, 2006. 3

[24] A. Sedlazeck and R. Koch. Calibration of housing parameters for underwater stereo-camera rigs. In *Proceedings of the British Machine Vision Conference*, pages 118.1–118.11. BMVA Press, 2011. 3, 5

[25] A. Sedlazeck and R. Koch. Perspective and non-perspective camera models in underwater imaging overview and error analysis. In *Outdoor and Large-Scale Real-World Scene Analysis*, volume 7474 of *LNCS*, pages 212–242. 2012. 1

[26] P. Sturm, S. Ramalingam, and S. Lodha. On calibration, structure from motion and multi-view geometry for generic camera models. In *Imaging Beyond the Pinhole Camera*, volume 33 of *Computational Imaging and Vision*. Springer, aug 2006. 4

[27] R. Szeliski. *Computer Vision Algorithms and Applications*. Springer-Verlag, 2011. 1, 2

[28] T. Treibitz, Y. Schechner, and H. Singh. Flat refractive geometry. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition CVPR 2008*, pages 1–8, 2008. 1

[29] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment – a modern synthesis. In B. Triggs, A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practice*, volume 1883 of *LNCS*, pages 298–372, 2000. 2, 5