Methods in Oceanography 15–16 (2016) 90–113

Contents lists available at ScienceDirect

Methods in Oceanography

journal homepage: www.elsevier.com/locate/mio

Full length article

Refractive 3D reconstruction on underwater images

Anne Jordt [a,b,*], Kevin Köser [a], Reinhard Koch [b]

[a] GEOMAR Helmholtz Centre for Ocean Research, Wischhofstr 1-3, 24148 Kiel, Germany
[b] Christian-Albrechts-University of Kiel, Hermann-Rodewald-Str. 3, 24118 Kiel, Germany

HIGHLIGHTS

- Complete 3D reconstruction system from images and videos.
- Refractive Structure from Motion for flat port underwater cameras.
- Eliminates systematic modeling error caused by using perspective camera model.

ARTICLE INFO

Article history:
Received 30 June 2015
Received in revised form
10 March 2016
Accepted 23 March 2016
Available online 26 April 2016

Keywords:
Underwater reconstruction
Refraction
Flat port
Structure-from-Motion
Plane sweep
3D model

ABSTRACT

Cameras can be considered measurement devices complementary to acoustic sensors when it comes to surveying marine environments. When calibrated and used correctly, these visual sensors are well-suited for automated detection, quantification, mapping, and monitoring applications and when aiming at high-accuracy 3D models or change detection. In underwater scenarios, cameras are often set up in pressure housings with a flat glass window, a flat port, which allows them to observe the environment. In this contribution, a geometric model for image formation is discussed that explicitly considers refraction at the interface under realistic assumptions like a slightly misaligned camera (w.r.t. the glass normal) and thick glass ports as common for deep sea applications. Then, starting from camera calibration, a complete, fully automated 3D reconstruction system is discussed that takes an image sequence and produces a 3D model. Newly derived refractive estimators for sparse two-view geometry, pose estimation, bundle

* Corresponding author at: GEOMAR Helmholtz Centre for Ocean Research, Wischhofstr 1-3, 24148 Kiel, Germany.
E-mail address: ajordt@geomar.de (A. Jordt).
URL: http://www.geomar.de (A. Jordt).

http://dx.doi.org/10.1016/j.mio.2016.03.001

adjustment, and dense depth estimation are discussed and evaluated in detail.

1. Introduction

In the last decades, cameras in oceanography have advanced from a pure qualitative inspection and manual evaluation tool to a visual sensor with great potential for automated survey, measuring, mapping and monitoring. Applications for underwater imaging include construction and inspection in the offshore industry, marine habitat monitoring (Costa et al., 2006; Williams et al., 2012), geological questions (Kwasnitschka et al., 2013), archeological documentation (Henderson et al., 2013; Bingham et al., 2010), exploration for deep sea mining, and many more. All of the mentioned applications have in common, that scientists benefit from a digital 3D model of the scene or object of interest in order to allow interactive inspection, measurements, documentation (over time), and classification and quantification. This work introduces a system for automated 3D reconstruction from images, also called Structure from Motion (SfM), adapted to the underwater imaging environment.

As with all sensors, the acquisition process for visual data has to be carefully inspected to avoid systematic errors or biases in the observed data. In particular, submerged cameras view the underwater world from within a pressure housing, often through a glass (or acrylic or sapphire) plate, called *flat port*. For deep sea applications at high water pressure, the thickness of these ports can be in the range of several centimeters (Fig. 2 on the right) and is therefore usually not negligible. When light rays enter such a housing at a non-zero incidence angle, they are refracted due to the different media densities according to Snell's law (Hecht, 1998). This causes imaging geometry to deviate from classical photogrammetry in air and the often used perspective (i.e. single-viewpoint) camera model to become invalid (Kotowski, 1988). Classical approaches for automated 3D reconstruction from images (Snavely et al., 2008; Farenzena et al., 2009; Johnson-Roberson et al., 2010) however do assume a single viewpoint model.

Fig. 1 shows the refraction effect on a temple model in a fish tank that is flooded and viewed from outside, fully analogous to a camera being submerged behind a flat port housing. It is clearly visible that the image is deformed, and, maybe less obvious, the deformation is distance dependent and cannot be modeled by simple 2D image operations. The effect can also be depicted geometrically (Fig. 2): the rays coming from the water are refracted at the water–glass interface and again at the glass–air interface.

When ignoring refraction (dashed lines), one can observe that the rays do not intersect in a common center of projection any more, hence the often used single-view-point camera model is invalid and a tailored solution is required.

In the next section, the evolution of underwater imaging models and previous work related to refractive imaging are discussed. Afterwards, we will derive a model for the imaging process of flat port cameras based on the underlying physical principles and define optimal estimation criteria under a Gaussian noise assumption for image observations. In Section 4, the first complete refractive visual reconstruction system is proposed, consisting of refractive single view and two view estimators for bootstrapping, refractive bundle adjustment, and finally dense refractive stereo estimation. The approach is evaluated in detail in Section 5. Here, we discuss also a potential benefit of refractive vision over 3D reconstruction in air, namely the potential to extract absolute scale from monocular image sequences. Finally, the system is demonstrated on several real-world sequences.

2. Previous work and novel contributions

In photogrammetry it is well known that refraction should be considered in order to obtain accurate measurements (Kotowski, 1988). Nevertheless, in the earlier days of underwater

Fig. 1. A model of the entrance to the Abu Simbel temple in Egypt is placed in a fish tank. A static camera views the scene through a flat glass. When increasing the water level in the fish tank, refractive effects are clearly visible. These do not only scale the image (in 2D) but allow to "look around" foreground objects (in 3D).

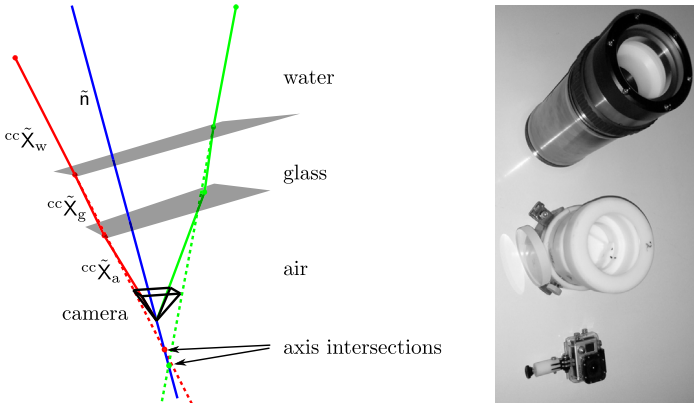
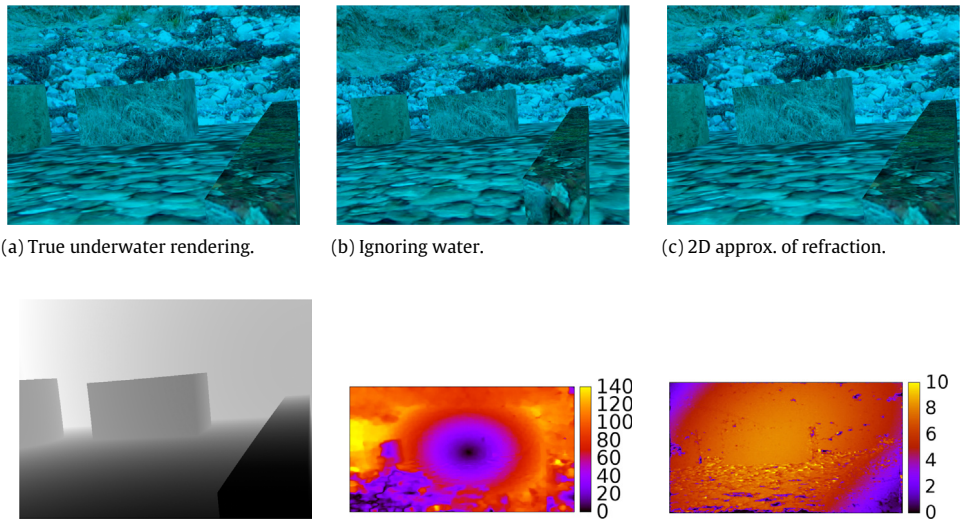


Fig. 2. Refractive camera with the perspective camera inside the housing in black. The inner and outer glass interface planes are depicted in gray. The blue line shows the interface normal intersecting the camera's center of projection. In red and green are the ray segments for two image points. Note how the dashed lines on the left do not intersect the camera's center of projection, but the blue line in different points. This shows that the perspective camera model is invalid. The right image shows sample camera housings for 6000 m (top, port ca. 2 cm sapphire), several hundred meters (center, port ca. 1 cm glass) and flat water (bottom, port ca. 1 mm acrylic glass). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

photogrammetry and computer vision, computational power was limited and tractable approximations were sought that described underwater image observations. Towards this end, Fryer and Fraser (1986), Harvey and Shortis (1998), and Lavest et al. (2000) postulated that the refractive effect can be approximated essentially by an image stretch from the center in radial direction. These approximations ignore the dependence of the image deformation on the 3D scene layout and model refraction as a pure 2D effect that can be absorbed into radial distortion, focal length, and potentially other intrinsic camera parameters. It can however be shown that the error depends on the distance between 3D point and camera (Sedlazeck and Koch, 2012) and 2D approximations can have large systematic errors. Fig. 3 shows a comparison between images rendered with the correct refractive camera model and with the perspective camera model. A refracted underwater image and a depth map are displayed on the left. The middle column shows the rendered image using the perspective camera model without any compensation for refraction. The right column shows the rendering results for the perspective camera model with an approximation of the refractive effect as in Harvey and Shortis (1998).

In this case, the pixel-wise difference between refractive and perspective image can be tens of pixels for high resolution cameras. Despite this error, a lot of works can be found in the literature,



(a) True underwater rendering. (b) Ignoring water. (c) 2D approx. of refraction.

(d) True depth map. (e) Error of (b)/pixel. (f) Error of (c)/pixel.

Fig. 3. Left column: (a) synthetic underwater image (considering refraction) and (d) corresponding depth map below. Middle column: (b) rendered image with perspective camera model ignoring refraction and (e) pixel movement compared to refractive image. Right column: (c) rendered image with perspective approximation of refraction and (f) pixel movement compared to refractive image.

where the perspective camera model is used in applications that utilize imaging geometry like mosaicking (Xu and Negahdaripour, 2001) or Structure from Motion (Johnson-Roberson et al., 2010; Inglis et al., 2012; Sedlazeck et al., 2009).

One solution to eliminate the systematic modeling error is to use a general camera model, where for each pixel, a ray or raxel (Grossberg and Nayar, 2005) is determined. Several works are concerned with reconstructing scenes based on such general camera models, e.g. Mouragnon et al. (2009) and Ramalingam et al. (2006). However, robustly obtaining a precise camera calibration for a camera model where for each pixel a ray is calibrated independently is often difficult. Therefore, it is also interesting to model refraction explicitly by parameterizing the interface housing. Calibration approaches for this can be found in Treibitz et al. (2008), Li et al. (1997), Agrawal et al. (2012) and Jordt-Sedlazeck and Koch (2012). In Treibitz et al. (2008) very thin glass and parallelism between glass and imaging sensor are assumed. The authors of Li et al. (1997) calibrate a stereo rig for a camera housing with thick glass and interface inclination. Agrawal et al. (2012) demonstrate how to calibrate a monocular camera, where the glass thickness can be modeled explicitly and a possible tilt between glass and imaging sensor is calibrated as well. Jordt-Sedlazeck and Koch (2012) use the results in Agrawal et al. (2012) as an initialization and apply an Analysis-by-Synthesis approach for optimization.

The obtained refractive calibration can be utilized in a refractive reconstruction approach. Towards this end, Chari and Sturm (2009) derive a theory for refractive SfM for a camera looking through a water surface, however, the investigation remains theoretical. Chang and Chen (2011) is a system, where the camera views a scene on the bottom of a fish tank through the water surface. It is assumed that the camera's yaw and pitch are known. More general in terms of external sensors is the work of Kang et al. (2012). This work is targeted to reconstructions of two photos and requires manual elimination of outliers in the correspondences. A key problem of refractive vision is that forward projection of 3D points into cameras is difficult. For the simple case of parallelism between glass and image sensor and negligible glass thickness, Glaeser and Schröcker (2000) showed how 3D points can be projected using a 4th-degree polynomial. However, in the more general case of practical deep sea housings, forward projection can either be expensively approximated using back projection and inverse methods or, as recently shown by Agrawal et al. (2012) through the solution of a 12th-degree

polynomial, which is still way more complicated than the simple matrix multiplication for solving the same problem in air. Due to inefficient projections of 3D points into a refractive camera, it is infeasible to extend (Chang and Chen, 2011; Kang et al., 2012) to large scenes with many images.

Thus, no general system for refractive 3D reconstruction exists, which is the aim of this paper. Explicitly modeling refraction allows to eliminate the systematic model error introduced by applying methods designed for images captured in air at the cost of expensive forward projection. To obtain a computationally tractable procedure, an efficient error function is required in bundle adjustment, which classically relies on many forward projections. This allows to optimize reconstructions efficiently, and hence allows to apply the reconstruction algorithm to scenes with more than a few tens of images.

Contributions

This work is the first to propose, implement, and evaluate a complete scalable 3D reconstruction system that can be used for deep sea flat port cameras. It builds on and extends two preliminary conference publications (and for further aspects, the reader is also referred to Jordt, 2014): Jordt-Sedlazeck and Koch (2013), which introduce the refractive Structure-from-Motion routine and Jordt-Sedlazeck et al. (2013), which describe a refractive Plane Sweep algorithm. In this work, both methods are combined into a complete system for refractive reconstruction by improving the non-linear optimization to work on larger scenes. In addition, the system is compared against a method based on a general camera model in air (Mouragnon et al., 2009), which can be adapted to cover refraction in the general camera. Additionally, we investigate in which cases absolute scene scale is observable for monocular flat port cameras.

In the following, first the geometric image formation model will be introduced. This includes an efficient formulation for maximum likelihood estimation given Gaussian noise on the image observations despite the fact that refractive forward projection is expensive.

3. Refractive image formation

In order to describe the geometric image formation, when observing an object in the water using a camera inside a pressure housing, the interface of the housing is explicitly considered. For the camera inside the pressure housing we require a geometrical imaging model that allows computation of the ray in space for each position in the image, i.e. that we can backproject pixels onto 3D lines of sight in the air around the camera. Many different models exist, for instance ideal pinhole cameras, wide angle lenses with radial and tangential distortion (Heikkila and Silven, 1997), fisheyes (Scaramuzza et al., 2006), non-central cameras (Grossberg and Nayar, 2005), and so on. For simplicity of presentation, and since we will reason only about rays in space, in the remainder of this document we will assume a pinhole camera with a single center of projection (analogous considerations can be made for other camera models). The intrinsic parameters of such cameras can be calibrated with standard calibration toolboxes in air[1,2] (Schiller et al., 2008). Given an image or image coordinates for a camera with a certain set of intrinsics, it is straightforward to compute the image or coordinates for a different setting of these parameters (e.g. undistortion). Therefore, in the following, we assume without loss of generality that we have a *canonical camera* with focal length 1 and principal point 0 and no radial distortion (cf. also to Hartley and Zisserman, 2004 and Szeliski, 2011) that maps 3D points ${}^{wc}\mathbf{X} \in \mathbb{P}^3$ in world coordinates to homogeneous 2D points $\mathbf{x} \in \mathbb{P}^2$ in the image:

$$\mathbf{x} = (\mathbf{R}^T \mid -\mathbf{R}^T\mathsf{C})\,{}^{wc}\mathbf{X}. \tag{1}$$

Here, $\mathbf{R}$ and C represent the camera's orientation matrix and euclidean position vector respectively. In the remainder, 3D points will be annotated with ^{wc} or ^{cc} depending on whether they are in world or

[1] http://www.opencv.org.

[2] http://www.vision.caltech.edu/bouguetj/calib_doc/.

camera coordinates. The idea of the proposed approach lies in explicitly modeling the camera housing, which is assumed to contain a thick flat transparent window with parallel interfaces to air and water. For deep sea applications in several kilometers water depth, these ports can be several centimeters thick such that they cannot be considered infinitesimally small. The thickness depends on the material used, the pressure (i.e. the depth), and the diameter of the opening. Consequently, for the most general case, a light ray travels from an object through the water into the "glass" and then into the air that surrounds the lens, before finally reaching the lens. The refractive camera (a perspective camera inside an underwater housing) is then completed by a parametrization of refraction at the glass interface. These parameters include the distance d between center of projection and glass, the glass thickness d_g, and the glass normal $\tilde{n}$ in the camera coordinate system. In addition, the indexes of refraction for air n_a, glass n_g, and water n_w are required.

The *backprojection* of an image point into space now works as follows: after computing the (simply backprojected) normalized ray from the camera center inside the housing ${}^{cc}\tilde{X}_a = \mathbf{R}\mathbf{x}/\|\mathbf{R}\mathbf{x}\|$, it can be intersected with the glass and refracted using Snell's law to determine the ray direction in glass ${}^{cc}\tilde{X}_g$ (Agrawal et al., 2012):

$$ {}^{cc}X_g = \frac{n_a}{n_g}{}^{cc}\tilde{X}_a + \left(-\frac{n_a}{n_g}{}^{cc}\tilde{X}_a{}^T\tilde{n} + \sqrt{1-\frac{n_a}{n_g}(1-({}^{cc}\tilde{X}_a{}^T\tilde{n})^2)}\right)\tilde{n}. \tag{2} $$

After normalizing ${}^{cc}X_g$, it can be used to determine the ray in water ${}^{cc}\tilde{X}_w$ respectively. For each ray, its starting point in water on the outer glass plane can be determined by ${}^{cc}X_s = \frac{d}{{}^{cc}\tilde{X}_a^T\tilde{n}}{}^{cc}\tilde{X}_a + \frac{d_g}{{}^{cc}\tilde{X}_g^T\tilde{n}}{}^{cc}\tilde{X}_g$. The resulting ray in water in the local camera coordinate system can then be transformed into the world coordinate system by:

$$ {}^{wc}X_s = \mathbf{R}{}^{cc}X_s + C \qquad {}^{wc}\tilde{X}_w = \mathbf{R}{}^{cc}\tilde{X}_w. \tag{3} $$

Agrawal et al. (2012) determined several other interesting properties of refractive underwater cameras: one major insight was that all rays coming from the water intersect a common axis defined by the camera's center of projection and the interface normal (Fig. 2, left). In addition, Agrawal et al. (2012) determine that according to the second part of Snell's law, all ray segments and the interface normal lie in one common plane, the plane of refraction (POR), which allows to derive the POR constraint (Fig. 4 on the left):

$$ (\mathbf{R}^{Twc}X - \mathbf{R}^TC)^T(\tilde{n} \times {}^{cc}\tilde{X}_w) = 0, \tag{4} $$

which determines that a 3D point transformed into the camera coordinate system should lie on the POR as well. A second useful constraint derived by Agrawal et al. (2012) is the flat refractive constraint (FRC) that states that a 3D point in camera coordinates that is transferred onto the starting point ${}^{cc}X_s$ should have a zero angle with the ray in water ${}^{cc}\tilde{X}_w$ (Fig. 4 on the right):

$$ (\mathbf{R}^{Twc}X - \mathbf{R}^TC - {}^{cc}X_s) \times {}^{cc}\tilde{X}_w = 0. \tag{5} $$

Note that (5) allows to derive an angle error similar to the one used in Mouragnon et al. (2009) as can be used for general camera models.

The fact that all ray segments lie in the POR allowed to derive a projection of 3D points into the image plane by solving a 12th degree polynomial. Compared to solving the projection using iterative methods (similar to what is done in Mulsow (2010)) and backprojection according to Eq. (2), this was a huge improvement. Still, for reasonably large scenes, bundle adjustment and dense stereo require millions of projections and can become intractable when the number of data grows.

3.1. Lifting observations from image space to outer glass space

Projecting 3D points into an image is a task that is at least implicitly required in several estimators, either as part of an objective function to be optimized or as a means to classify correspondences into inliers and outliers according to a model. In both cases, a 2D–3D correspondence exists, i.e. a computed 3D point and the corresponding observation, the 2D point in the image. After projecting the 3D point

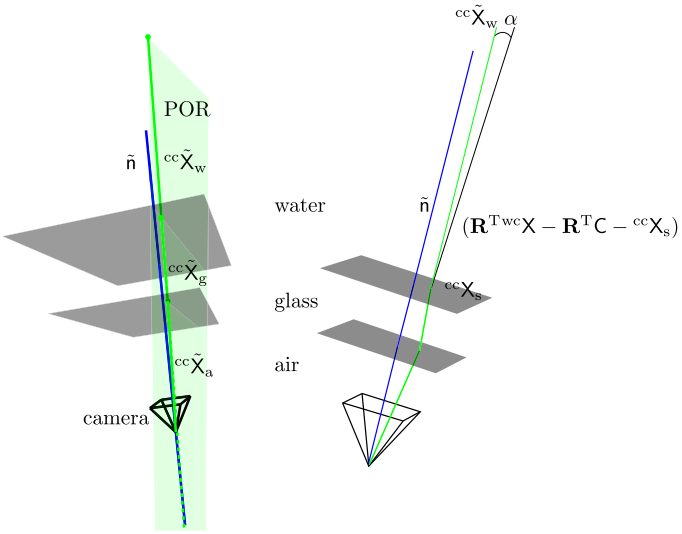


Fig. 4. Depicted in on the left is the plane of refraction (POR) for the green point shown in Fig. 2, containing all ray segments and the interface normal $\tilde{n}$. Depicted on the right is the Flat Refractive Constraint (FRC), where a ray segment in water is transformed into the local camera coordinates and compared to the local ray in water. The FRC states that the angle α should be zero. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

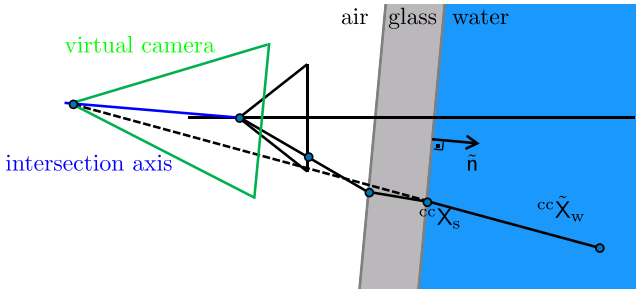


Fig. 5. For a 2D image position, the corresponding ray in water is intersected with the axis defined by center of projection and interface normal (blue line). This intersection defines the virtual camera center (green). The virtual camera's optical axis is defined by the interface normal. Consequently, for a 3D–2D correspondence, the 3D point can be projected into the virtual camera perspectively and the 2D point can be transformed into the virtual camera, thus defining a reprojection error that can be computed efficiently. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

into the image, the difference of the projected point to the observation is called *reprojection error* in the literature. Since refraction makes projecting 3D points directly into the image prohibitively expensive, the problem is reformulated such that the 2D observation is backprojected onto the outer glass interface. This removes the need for considering any further refractions in the objective function and simple linear projections can be used as in air.

Instead of using the reprojection error, we propose a different, non-linear error function that defines a *virtual camera* for each 2D observation, into which the corresponding 3D point can be projected perspectively (see Fig. 5). The basic idea is similar to the idea in Ramalingam et al. (2006), however, here the virtual camera is defined for each pixel and computes an exact error. An earlier version described in Sedlazeck and Koch (2011) required long computation times due to the need to compute a caustic point for each virtual camera. Here instead, we propose using the axis defined by center of projection and interface normal, which is intersected by each ray in water (Agrawal et al.,

2012) and serves as a virtual camera center C_v. The optical axis of the virtual camera is determined by the interface normal. Note that the virtual camera error can only be computed in case of existing 2D–3D correspondences, the 2D point determines the virtual camera and can be transformed onto the virtual image plane. The 3D point can be projected perspectively into the virtual camera, thus allowing to compute the reprojection error in this camera. The resulting virtual camera error can be computed efficiently and analytic derivatives exist.

When running methods like bundle adjustment using a sum-of-squares error function, it is assumed that the measurement error on the correspondences is normally distributed, an assumption that can be made when no outliers are present in the data. We have empirically verified for practically relevant settings of distance and glass inclination that a normally distributed observation error in the image stays approximately Gaussian also in the virtual camera (Kolmogorov–Smirnov test Press et al., 2002 with 5% significance level). Normally distributed observation errors in the virtual cameras mean that minimizing the sum of squared virtual reprojection errors is a maximum likelihood estimator.

3.2. Calibration

The model described above depends on the knowledge of camera parameters, housing parameters and water parameters. In this section we will briefly discuss calibration techniques to obtain these parameters.[3] The optical properties of sea water and glass, i.e. their indexes of refraction, can be obtained from oceanographic models or material property sheets, respectively. The same holds for the thickness of the glass, which is usually known in practical systems.

The intrinsic camera parameters like focal length, but also the housing parameters need to be calibrated, which is achieved in two steps. First, the camera's intrinsics are calibrated by capturing checkerboard images in air. The method described in Schiller et al. (2008) uses an Analysis-by-Synthesis (AbyS) approach (Koch, 1993) and can calibrate perspective monocular cameras, but also stereo rigs with high accuracy. Second, the calibration approach for the housing parameters is also based on a set of checkerboard images, this time captured below the water surface (see Jordt-Sedlazeck and Koch, 2012 for details).

In cases where only approximate information is available for certain parameters, or in case checkerboard calibration is not feasible, parameters can also be optimized inside the Structure-from-Motion pipeline directly on the scientific image sequences, i.e. during bundle adjustment. This Structure-from-Motion pipeline will be described in the next section.

4. Automated refractive reconstruction

After having discussed refractive image formation and virtual camera errors, in this section we will discuss the actual automated 3D reconstruction approach, which follows a common sequential Structure-from-Motion (SfM, refer to Hartley and Zisserman, 2004 and Szeliski, 2011 for classic SfM methods): find image feature correspondences, robustly estimate single view or two-view relations to reject outliers and to locally extend the reconstruction, and perform bundle adjustment to obtain an optimal sparse reconstruction. Then, for each key frame compute dense distance information per pixel (a depth map) and fuse the depth maps to obtain the final 3D model. For all these steps different variations and strategies exist in the literature, depending on whether motion models apply or unordered image collections are reconstructed, depending on whether the goal is online or batch reconstruction and which additional sensors are available. It should be noted that the actual reconstruction strategy is not the main focus of this paper. Rather, we want to demonstrate feasibility of refractive reconstruction by presenting a complete system that relies as much on the visual information as possible in order to evaluate the difference refractive reconstruction makes.

[3] Calibration software, targets and instructions will be made available on http://www.geomar.de/go/cameracalibration-e.

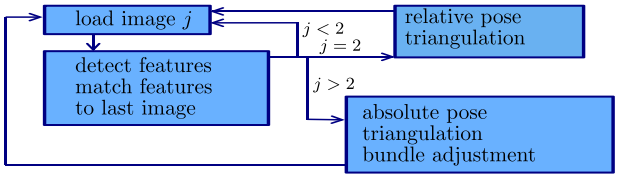


Fig. 6. Experiments are conducted on a classic, sequential SfM pipeline. Images are loaded and features are detected and matched to the last image. In case of the images being the first two, the relative pose of the second image with respect to the first is computed as in initialization. This allows to triangulate a set of 3D points for the 2D–2D image correspondences. Other images are added sequentially by computing the absolute pose of the image with respect to the existing 3D points. Additional 3D points are triangulated and bundle adjustment is applied in order to improve the reconstructed scene.

4.1. Structure from motion

The calibration of the camera housing is utilized in a refractive Structure-from-Motion (SfM) approach in order to determine the camera poses and 3D points of the scene. The algorithmic pipeline can be seen in Fig. 6. After loading a pair of consecutive images, SIFT-features are detected and matched using (Wu, 2007), yielding what we call 2D–2D correspondences. Then, the *relative pose* of the second image with respect to the first needs to be computed robustly. After that, a set of 3D points can be triangulated (Hartley and Sturm, 1997) and associated with the 2D feature points. A new image can be added to the reconstruction by matching its features to the features from existing frames, leading to a new set of 2D–2D correspondences. Some of those have an associated 3D point, which allows robustly computing *absolute pose* using those 2D–3D correspondences in a RANSAC-framework (Fischler and Bolles, 1981). After adding an image to the reconstruction, all camera poses and 3D points are optimized using bundle adjustment (McGlone, 2004; Triggs et al., 2000). Note that the same system can be used (a) in case of classic perspective reconstruction, (b) for a ray-based reconstruction described in Mouragnon et al. (2009), and (c) for the proposed refractive approach, thus allowing a direct comparison. In the following, algorithms for refractive relative and absolute pose, as well as for refractive bundle adjustment will be introduced.

4.1.1. Refractive relative pose

Based on a set of 2D–2D correspondences between two images, the relative pose of the second image with respect to the first can be computed. For ease of presentation, and without loss of generality, we assume that the first image has been taken at the origin of the world and with orientation identity matrix. Therefore, ${}^{\mathrm{cc}}\tilde{\mathrm{X}}_{\mathrm{w}} = {}^{\mathrm{wc}}\tilde{\mathrm{X}}_{\mathrm{w}}$ and ${}^{\mathrm{cc}}\mathrm{X}_{\mathrm{s}} = {}^{\mathrm{wc}}\mathrm{X}_{\mathrm{s}}$ for all rays of the first image. Thus we seek orientation $\mathbf{R}$ and position C of the second image. In the perspective case, relative pose relates to epipolar geometry (Hartley and Zisserman, 2004) and minimal methods like for example Nistér's five-point algorithm can be applied (Nistér, 2004). In case of refractive or other general camera models Pless (2003), Mouragnon et al. (2009), and Li et al. (2008) all examine a linear approach based on intersecting Plücker lines for the 2D–2D correspondences:

$$\begin{aligned} 0 &= {}^{\mathrm{wc}}\tilde{\mathrm{X}}_{\mathrm{w}}{}^{\mathrm{T}}(\mathbf{R}({}^{\mathrm{cc}}\tilde{\mathrm{X}}'_{\mathrm{w}} \times {}^{\mathrm{cc}}\mathrm{X}'_{\mathrm{s}}) - [\mathsf{C}]_{\times}\mathbf{R}{}^{\mathrm{cc}}\tilde{\mathrm{X}}'_{\mathrm{w}}) + ({}^{\mathrm{wc}}\tilde{\mathrm{X}}_{\mathrm{w}} \times {}^{\mathrm{wc}}\mathrm{X}_{\mathrm{s}})^{\mathrm{T}}(\mathbf{R}{}^{\mathrm{cc}}\tilde{\mathrm{X}}'_{\mathrm{w}}) \\ &= \begin{pmatrix} {}^{\mathrm{wc}}\tilde{\mathrm{X}}_{\mathrm{w}} \\ ({}^{\mathrm{wc}}\tilde{\mathrm{X}}_{\mathrm{w}} \times {}^{\mathrm{wc}}\mathrm{X}_{\mathrm{s}}) \end{pmatrix}^{\mathrm{T}} \underbrace{\begin{pmatrix} -[\mathsf{C}]_{\times}\mathbf{R} & \mathbf{R} \\ \mathbf{R} & \mathbf{0}_{3\times 3} \end{pmatrix}}_{\mathbf{E}_{\mathrm{GEC}}} \begin{pmatrix} {}^{\mathrm{cc}}\tilde{\mathrm{X}}'_{\mathrm{w}} \\ ({}^{\mathrm{cc}}\tilde{\mathrm{X}}'_{\mathrm{w}} \times {}^{\mathrm{cc}}\mathrm{X}'_{\mathrm{s}}) \end{pmatrix}. \end{aligned} \tag{6}$$

This is called Generalized Epipolar Constraint (GEC). 17 correspondences can be used to construct a system of equations linear in the entries of $\mathbf{E}_{\mathrm{GEC}}$ (Pless, 2003):

$$\mathbf{A}\ \mathrm{vec}(\mathbf{E}_{\mathrm{GEC}}) = 0. \tag{7}$$

Li et al. (2008) propose to split the matrix $\mathbf{A}$ into two parts $\mathbf{A}_{\mathrm{E}}$ for $-[\mathsf{C}]_{\times}\mathbf{R}$ and $\mathbf{A}_{\mathrm{R}}$ for $\mathbf{R}$ and solve the system $(\mathbf{A}_{\mathrm{R}}\mathbf{A}_{\mathrm{R}}^{+} - \mathbf{I})\mathbf{A}_{\mathrm{E}}(e_{11} \ldots e_{33})^{\mathrm{T}} = 0$. This method will be called *Li method* in the remainder. When using the general system of equations with 17 correspondences, the fact that underwater cameras are

axial cameras causes the system (7) to have two zero singular values, and hence a two-dimensional solution space $(e_{11} \dots e_{33})^{\mathrm{T}} = \mu \mathbf{e}_1 + \nu \mathbf{e}_2,\ \mu, \nu \in \mathbb{R}$. μ can be set to one due to the solution being determined up to scale only and a suitable ν can be found by utilizing a constraint on the rotation matrix. This method outperformed the Li method in our experiments and will be called *Pless method* in the evaluation.

A novel refractive relative pose solver. A different possibility for determining refractive relative pose can be developed by using three geometrical constraints. The first one is that two corresponding rays in water intersect in the same 3D point, which can be expressed by the triangulation constraint:

$$^{\mathrm{wc}}\mathsf{X}_{\mathrm{s}} + \kappa\, ^{\mathrm{wc}}\tilde{\mathsf{X}}_{\mathrm{w}} = \mathbf{R}\, ^{\mathrm{cc}}\mathsf{X}'_{\mathrm{s}} + \mathsf{C} + \kappa' \mathbf{R}\, ^{\mathrm{cc}}\tilde{\mathsf{X}}'_{\mathrm{w}}, \tag{8}$$

where again $\mathbf{R}$ and C represent the transformation of the second camera relative to the first, $^{\mathrm{wc}}\mathsf{X}_{\mathrm{s}}$, $^{\mathrm{wc}}\tilde{\mathsf{X}}_{\mathrm{w}}$ and $^{\mathrm{cc}}\mathsf{X}'_{\mathrm{s}}$, $^{\mathrm{cc}}\tilde{\mathsf{X}}'_{\mathrm{w}}$ are the rays in water for the two corresponding points, and κ and κ' are the scaling factors for the rays in water in order for the rays to intersecting a common 3D point. Based on the same entities, a constraint can be derived by transforming the ray of the second camera pose into the first and applying the FRC Eq. (5):

$$(\mathbf{R}\, ^{\mathrm{cc}}\mathsf{X}'_{\mathrm{s}} + \mathsf{C} + \kappa' \mathbf{R}\, ^{\mathrm{cc}}\tilde{\mathsf{X}}'_{\mathrm{w}} - ^{\mathrm{wc}}\mathsf{X}_{\mathrm{s}}) \times ^{\mathrm{wc}}\tilde{\mathsf{X}}_{\mathrm{w}} = 0. \tag{9}$$

Both constraints are non-linear in the unknowns $\mathbf{R}$, C, κ and κ' and we apply an iterative approach to solve for the unknowns. A set of K correspondences is used, where the κ_k and $\kappa'_k,\ k \in \{1, \dots, K\}$ are all initialized with 3 m, which corresponds to common underwater visibility conditions, i.e. the 3D Points are assumed to be 3 m away from the camera. Then, of the 6 equations gained from constraints (8) and (9), 3 are linearly independent and are used to stack a linear equations system to solve for $\mathbf{R}$ and C. Next, the updated rotation and translation are used to update all κ_k and κ'_k using the POR constraint (4):

$$\begin{aligned} (\mathbf{R}\, ^{\mathrm{cc}}\mathsf{X}'_{\mathrm{s}_k} + \mathsf{C} + \kappa'_k \mathbf{R}\, ^{\mathrm{cc}}\tilde{\mathsf{X}}'_{\mathrm{w}_k})^{\mathrm{T}} (\tilde{\mathsf{n}} \times ^{\mathrm{wc}}\tilde{\mathsf{X}}_{\mathrm{w}_k}) = 0 \\ (\mathbf{R}^{\mathrm{T}}\, ^{\mathrm{wc}}\mathsf{X}_{\mathrm{s}_k} - \tilde{\mathsf{R}}^{\mathrm{T}}\mathsf{C} + \kappa'_k \mathbf{R}^{\mathrm{T}}\, ^{\mathrm{wc}}\tilde{\mathsf{X}}_{\mathrm{w}_k})^{\mathrm{T}} (\tilde{\mathsf{n}} \times ^{\mathrm{cc}}\tilde{\mathsf{X}}'_{\mathrm{w}_k}) = 0. \end{aligned} \tag{10}$$

The resulting κ_k and κ'_k are then used to update rotation and translation.

Both, the iterative and the Pless method are embedded in a RANSAC (Fischler and Bolles, 1981) framework in order to deal with outliers. Good solutions are optimized using the virtual camera error of all inliers within the RANSAC framework.

Scene scale. An interesting result of modeling refraction explicitly is that scene scale can theoretically be determined from a two-view setting. This is in contrast to the perspective case. Perspective reconstructions are determined up to a similarity transform (Hartley and Zisserman, 2004), i.e. absolute position, orientation, and a scale factor are not observable from image data alone. Algebraically, a similarity transformation $\mathbf{T} \in \mathbb{P}^{4\times4}$ can be applied to perspective projection matrices and points by $\mathbf{x} = \mathbf{P}\mathbf{T}\mathbf{T}^{-1\,\mathrm{wc}}\mathbf{X}$ without changing image observations. However, in the refractive case, such a transform would not scale interface distance and thickness, and thereby not change the starting points and directions, which are given in a fixed unit. Consequently, relative pose in the refractive camera case is not invariant against changes in scale.

4.1.2. Refractive absolute pose

Once the relative pose between the first two images is known, 3D points are triangulated using refractive triangulation through the interface, choosing the 3D point with the smallest sum of squared distances to the projection rays (Kanatani, 1996).

This allows to add more images to the reconstruction by computing the camera's absolute pose with respect to the 3D points. In case of perspective cameras, a lot of methods for absolute pose exist, e.g. Dementhon and Davis (1995) which will be used for comparing perspective against refractive reconstruction. In case of general camera models, Sturm et al. (2006) and Nistér and Stewénius (2007) both proposed methods that work on the minimal set of three 2D–3D correspondences. However, these methods are strongly sensitive to noise in the correspondences. Usually, 2D–3D correspondences are classified into inliers and outliers, where the inliers are assumed to have some

measurement noise, which is assumed to be normally distributed, while the outliers do not follow this distribution and can have very large errors e.g. due to mismatches during correspondence search. When running the methods within a RANSAC framework (Fischler and Bolles, 1981), a lot of trials fail due to the correspondences being noisy, even though they should all be classified as inliers. This can happen especially when working on underwater images, where contrast and visibility are often diminished, making feature detection difficult due to an increased noise level. Therefore, rather than solving directly for the absolute pose using a minimal solution, we propose an iterative scheme for computing absolute pose based on $c > 3$ (we use $c = 7$) correspondences. For each point correspondence the following constraint is used:

$$^{wc}\mathsf{X} = \mathbf{R}^{cc}\mathsf{X}_s + \mathsf{C} + \kappa\mathbf{R}^{cc}\tilde{\mathsf{X}}_w, \tag{11}$$

which is non-linear in the unknowns $\mathbf{R}$, C, and κ. Initializing κ for each correspondence with 1 allows to use a similar iterative scheme as in the relative pose case, by solving for $\mathbf{R}$ and C using (11) first, and updating all κ afterwards:

$$\kappa = (\mathbf{R}^{cc}\tilde{\mathsf{X}}_w)^T(^{wc}\mathsf{X} - \mathsf{C} - \mathbf{R}^{cc}\mathsf{X}_s). \tag{12}$$

4.1.3. Refractive bundle adjustment

The techniques for relative pose, absolute pose and triangulation described so far consider only a subset of the data and not all observations in all images at once. Although their results can serve as approximations for scene geometry and camera motion, the estimates are not optimal, in particular in presence of noisy observations. This is the goal of bundle adjustment (Triggs et al., 2000), where all camera poses and 3D points are optimized to jointly best match the observed feature position, i.e. bundle adjustment is a maximum likelihood estimator in case of normally distributed observations. A good introduction to bundle adjustment can be found in Triggs et al. (2000), McGlone (2004) and for increased readability we use the same symbols and letters for parameters (p), observations (l), covariances ($\mathbf{C}_{ll}$), constraints ($\mathsf{g}(\mathsf{p}, \mathsf{l})$ and $\mathsf{h}(\mathsf{p})$) and Jacobian matrices ($\mathbf{A}_g$ and $\mathbf{B}_g$) as the latter.

The objective function to be minimized in bundle adjustment contains the sum of squared distances between projected 3D points and measured 2D points in the images. This leads to an explicit functional dependence of the expected observations from the parameters $\mathsf{f}(\mathsf{p}) = \mathsf{l}$, where p contains the parameters for all images and points and the vector l contains all observations, i.e. all measured 2D points in all images. Depending on the number of camera images and the number of 3D points, the number of 3D point projections into the images during optimization in standard bundle adjustment can be in the order of tens of thousands or even millions in case of large scale reconstructions.

However, as already seen in Section 3, projecting a point into a refractive camera requires solving a 12th degree polynomial, hence causing refractive bundle adjustment to be infeasible. This problem is solved by using the virtual camera error function introduced in Section 3, where 3D points can be projected perspectively. Additionally, it is possible to compute analytic derivatives, thus allowing efficient computation of the parameter Jacobian $\mathbf{A}_g = \frac{\partial g}{\partial p}$. However, using the proposed virtual camera error function causes the constraint $\mathsf{g}(\mathsf{p}, \mathsf{l}) = 0$ to be implicit. This is the reason why the Gauss–Helmert model (McGlone, 2004; Triggs et al., 2000) needs to be used for optimization. Compared to the Gauss–Markov model, which is commonly used in classic bundle adjustment, the Gauss–Helmert model is a more general formulation of the optimization problem, allowing to deal with implicit constraints. Due to the dependence of g on the observations, the observation Jacobian $\mathbf{B}_g = \frac{\partial g}{\partial l}$ needs to be computed in addition to the parameter Jacobian, yielding the linearized system of equations:

$$\underbrace{\begin{bmatrix} \mathbf{A}_g{}^T(\mathbf{B}_g\mathbf{C}_{ll}\mathbf{B}_g{}^T)^{-1}\mathbf{A}_g & \mathbf{H}_h{}^T \\ \mathbf{H}_h & \mathbf{0} \end{bmatrix}}_{\mathbf{N}} \begin{bmatrix} \Delta\mathsf{p} \\ \mathsf{k}_h \end{bmatrix} = \begin{bmatrix} -\mathbf{A}_g{}^T(\mathbf{B}_g\mathbf{C}_{ll}\mathbf{B}_g{}^T)^{-1}\mathsf{g}(\mathsf{p}, \mathsf{l}) \\ -\mathsf{h}(\mathsf{p}) \end{bmatrix}, \tag{13}$$

where $\mathbf{C}_{ll}$ is the observation covariance, h are constraints between parameters like quaternion unit length, with $\mathbf{H}_h = \frac{\partial h}{\partial p}$ being the Jacobian of the parameter constraints, and k being a set of Lagrange Multipliers.

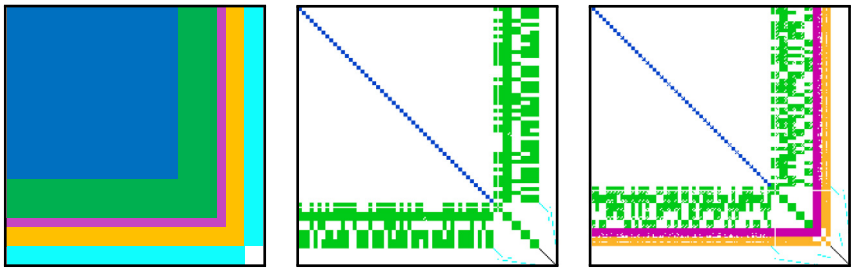


Fig. 7. Visualizations of bundle adjustment matrix **N**. Left: the blue part contains parameters for the 3D points, green is for the camera extrinsics, magenta for the local rig transform, orange for the housing parameters, and cyan for constraints between parameters like quaternion unit length or unit length of the camera housing. Middle: matrix for monocular adjustment without optimization of housing parameters, only colored parts are non-zero. Left: adjustment matrix for stereo optimization with optimization of rig transform and housing parameters of both cameras. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

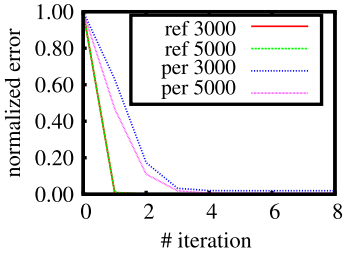


Fig. 8. Bundle adjustment convergence. Depicted are four different runs on random data, where camera poses and 3D points were optimized. The red and green trial are runs on underwater data with the virtual camera error with 3000 and 5000 3D points respectively. The blue and magenta curves depict runs on perspective data. Due to the absolute errors not being comparable, all errors depicted have been normalized. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Solving (13) then leads to successive updates $\Delta \mathbf{p}$ on the parameter vector. Note that in case of using the Gauss–Helmert model, a parameter ordering can be found that causes the sparse matrix **N** to have a large block diagonal part with the 3D point parameters that allows applying the Schur complement for solving Eq. (13) efficiently (see also Fig. 7). Convergence and run-time of the proposed algorithm is comparable to the perspective counterpart as can be observed in Fig. 8. In summary, using the virtual camera error with the analytically computed derivatives allows to run bundle adjustment in seconds rather than hours, allowing to apply refractive SfM to large image sequences.

4.1.4. Stereo sequences

If a stereo camera system is used instead of a monocular camera system, the refractive image formation theory and estimators can be applied in a very similar way. The main differences for an SfM system lie in the different steps of the reconstruction, i.e. the system is calibrated beforehand using the method in Sedlazeck and Koch (2011). This provides a fixed generalized essential matrix for 2D–2D inlier/outlier determination within a stereo pair. Inliers within a stereo pair can be triangulated to obtain 3D points and subsequent stereo pairs can be added to the reconstruction using absolute pose estimation using 2D–3D correspondences of both cameras at the same time (no relative pose initialization is required). Bundle adjustment is parametrized in a way that for each stereo-pair only one pose is computed and the relative transformation between the cameras of the stereo system is the same for all image pairs. Fixing the distance between the two stereo cameras also fixes the scale of the 3D reconstruction.

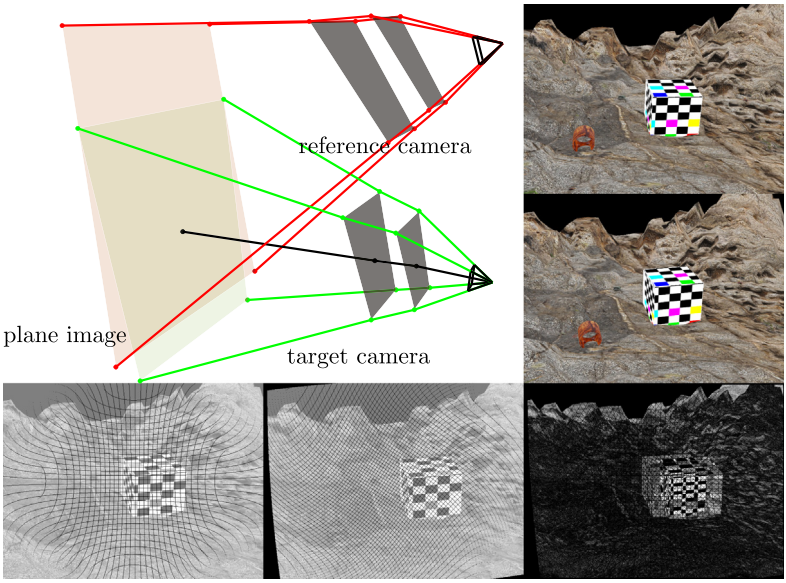


Fig. 9. Refractive plane sweep. Top left: camera configuration with plane image. The green camera is the target camera (image), the red the reference camera. Top right: the corresponding input images. The bottom row shows typical artifacts by forward mapping as they would appear without GPU interpolation, and which would hinder depth estimation. From left to right: plane image of target camera, plane image of reference camera, and difference image of the two plane images. The interpolated images on the GPU (not displayed) do not contain this distortion. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.2. Dense depth estimation using refractive plane sweep

The camera path and sparse 3D point cloud are utilized for dense depth computation, in order to gain a detailed 3D model. In case of refractive cameras, the dense depth algorithm needs to fulfill several design constraints:

- 3D–2D projections are expensive and should not be used
- rectification approaches known from pinhole cameras require single center of projection and are not applicable
- homographies are invalid (as single-view-point camera model is invalid).

When considering dense depth estimation for an image pair, the image for which the depth is to be computed will be called target image, the second image reference image. The proposed algorithm that meets all three constraints has a simple basic idea. It is in essence a plane sweep algorithm (Gallup et al., 2007), however, instead of warping entire images using homographies, a plane image is defined on the current hypothesis plane (see also Fig. 9). For the target image, hypothesis planes are swept through the space in front of the camera. For each hypothesis plane, the four image corners of the target image are back-projected using (2) and intersect the hypothesis plane defining the corners of the plane image. Both target and reference image are then warped onto the hypothesis plane, which poses a forward mapping, causing the plane image to be incomplete. However, by implementing the method on the GPU, holes in both plane images can be filled efficiently by interpolation. As common in plane sweep, also more than two images can now be utilized to increase robustness (we use three views in our experiments). Any appropriate local (dis-)similarity measures can be chosen to evaluate the agreement of the images on the plane hypothesis, such as normalized cross correlation (NCC) or sum of absolute differences (SAD). Our sample implementation uses two sweeps with different measures. During the first sweep, NCC is used for patch comparison yielding a preliminary depth map. In the second sweep, SAD with a shiftable window approach is used for patch comparison and good NCC results are used as weights. Note that the algorithm can be applied to all camera models for which

a ray with starting point and direction can be computed, thus is also applicable to a wide range of other general cameras including the perspective camera.

5. Evaluation

The system described in the previous sections was evaluated on numerical data, synthetically rendered images, and real underwater data. The experiments on synthetic data allow a comparison of the results against ground truth and also against other methods. On real data, the system using the classical perspective camera model is compared against the newly proposed refractive method. The images in this case were captured in a controlled lab environment in a fish tank. Finally, the new method is applied to real deep sea data captured on different scientific cruises.

5.1. Numerical experiments on refractive relative pose and scene scale

Fig. 11 shows the results of running the Pless method and the iterative method on 100 synthetic data sets with normal distributed noise with increasing variance σ on a scene of size 4 m in each direction. The iterative method slightly outperforms the linear method, both are comparable in accuracy to standard perspective methods like the 8-point algorithm (Hartley and Zisserman, 2004) on perspective data. Note, that the methods were run within a RANSAC framework with newly created data sets for each σ, which explains the non-monotonicity of the curves.

In Fig. 12 we investigate an interesting property of the refractive formulation: while in classical perspective monocular reconstructions in air, the absolute scale of the scene is not observable (Hartley and Zisserman, 2004), the fixed glass thickness and interface distance (known in mm by calibration) carry information about the absolute scale of a scene when observed underwater. For different camera types ultra wide angle action camera, wide angle deep sea system and an extreme setting with a wide angle lens camera and glass distance $d = 200$ mm and glass distance $d_g = 100$ mm, we simulate two images with image size 800×600 pixels, with a camera baseline of 0.5 m. The 3D points were between 0.5 and 4 m from the cameras. In the wide angle case, the focal length was 100 pixels, while in the wide angle deep sea case, the focal length was 700 pixels. After adding noise to the correspondences, the scale of the baseline between the two cameras is changed without scaling interface distance and glass thickness and the reprojection error is computed. It can be seen that in all cameras, there is a clear minimum for the resulting average reprojection error when no noise is present. However, for realistic noise levels the correct scale of the scene generates only a visible minimum in the error function for the extreme camera setting on the right and therefore this insight does not seem to be practically usable for standard underwater cameras. In the next section, the perspective and refractive methods are applied to synthetic image sequences.

5.2. Synthetic image sequences

The synthetic images are rendered using the simulator described in Sedlazeck and Koch (2011). The experimental setup was as follows:

- create synthetic scene
- create camera housing configurations with different interface distances chosen from $d \in [-10, 150$ mm$]$ and interface tilts defined by $\theta_1 = 30^\circ$ and $\theta_2 \in [0^\circ, 3^\circ]$ (see Fig. 13)
- for each housing configuration, render a set of refractive, underwater checkerboard images
- calibrate perspectively based on the checkerboard images in order to approximate refractive effect
- for each housing configuration, render synthetic scene with constant camera trajectory
- compute 3D point cloud and camera path using different methods
- evaluate reconstruction errors compared to known ground truth.

Two such scenes were rendered, the first one with a monocular camera and the second one with a stereo camera rig. Both scenes and camera trajectories are depicted in Fig. 10. In case of the monocular camera, the proposed method is compared against two other methods, one using the approximative,

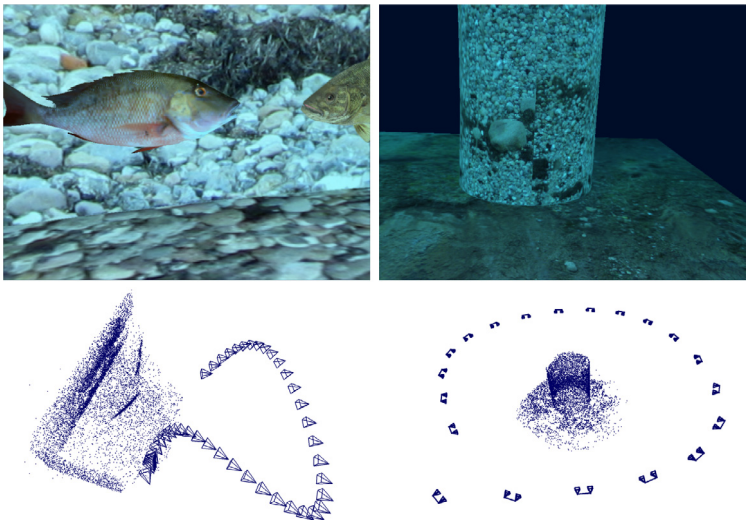


Fig. 10. Top row: exemplary images from the rendered scenes. Bottom row: scene structure and camera trajectory. Note that the scenes differ not only in structure and camera path, but also in camera–object distance, i.e. in the first scene the closest views have camera–object distances between 550 and 1900 mm and the furthest have 1300 –2300 mm. In the second (stereo) scene, the camera was moved in an approximate orbit around the scene, hence the camera–object distances were almost constant for all views (3000 –6000 mm).

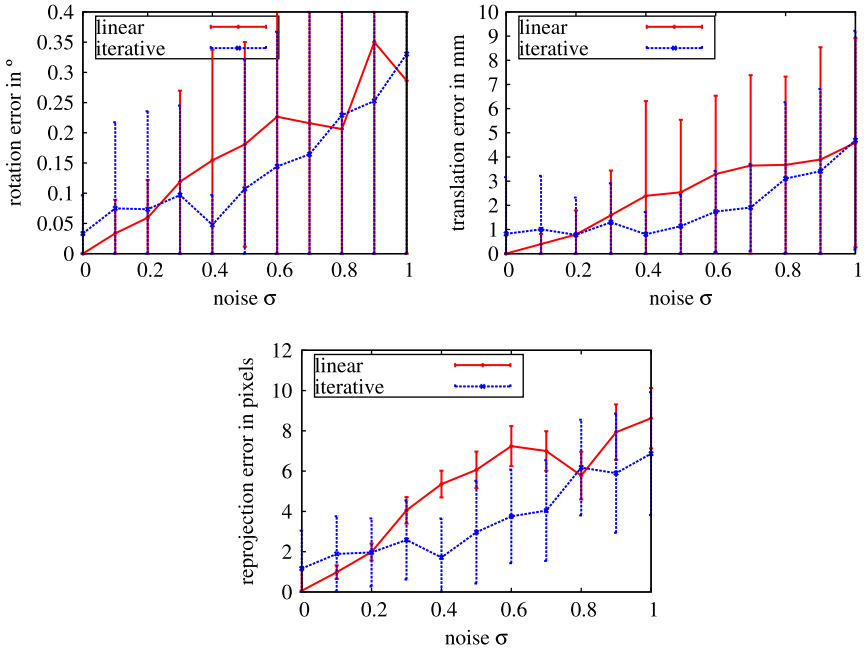


Fig. 11. Results of relative pose estimation. Left: orientation error for linear and iterative method, right: translation error. Bottom: reprojection error for both methods.

perspective camera, and another based on the angle error described for general camera models in air in Mouragnon et al. (2009). Although not designed as an underwater system its ray-based reasoning

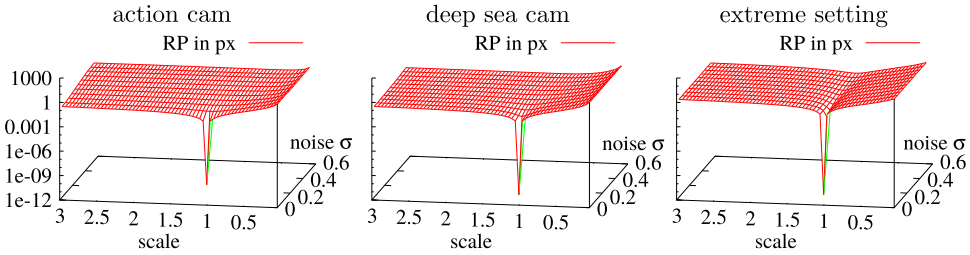


Fig. 12. Observability of absolute scale in two-view relative pose setting under different levels of noise. Left: ultra wide angle camera with 1 mm distance to port, 1 mm thickness. Center: wide angle camera with 25 mm from port, 10 mm thickness. Right: theoretical setup with ultra wide angle, distance 200 mm, glass thickness 100 mm. We plot the residual error in pixels when changing the scale of the scene and the distance of the housings (but keeping the known interface thickness and distance) from the correct value of 1. Absolute scale becomes more clearly observable with thicker glass, larger distance and wider field of view of the camera and can be identified in noiseless data easily (peak in front rows). As the noise level increases it becomes quickly infeasible to find the exact minimum reliably (back rows of plots correspond to only 0.6 pixels noise level).

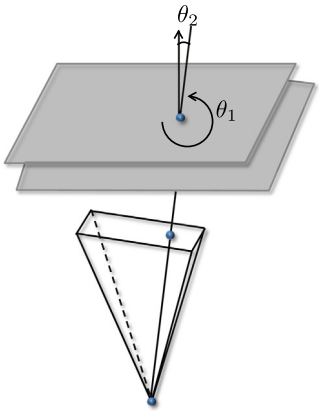


Fig. 13. Angles θ_1 and θ_2 describing the interface tilt.

can be used to avoid the systematic error present in the other underwater systems. The authors of Mouragnon et al. (2009) use the linear relative pose method described in Section 4.1.1 and run a maximum-likelihood estimation (optimization based on the angle error between rays), which is equivalent to the FRC (5) for underwater cameras. For absolute pose computation, they use a non-linear optimization minimizing the angle error within a RANSAC framework. Bundle adjustment is based on the angle error as well. In our tests, the initialization with the linear method often failed when using image data, and we used the iterative method as a fall back. In terms of implementation, the bundle adjustment using the angle error parametrizes rotations with incremental Euler angles, while the refractive bundle adjustment used quaternions. Note that all three methods were configured to follow the outline shown in Fig. 6 in order to be comparable.

The results are summarized in Fig. 14. From left to right the columns show the results using the perspective approximation as a camera model, the angle error, and the newly proposed refractive method. The *x*- and *y*-axis show the interface distance and tilt respectively. The first row shows the 3D error, i.e. the error of the 3D points compared to their ground truth points. The second row shows resulting errors in camera translation, and the last row shows the reprojection error. The results concerning the estimation of baseline scaling showed that in the refractive case, scale can be computed as long as there is no noise on the correspondences. In the perspective case, scene scale cannot be estimated. Therefore, the results in Fig. 14 assume a known scene scale between the first two view points, which is kept constant throughout the reconstruction. This assumption can often be met in real oceanic applications by utilizing the platform's navigation data or adding a pair of lasers with

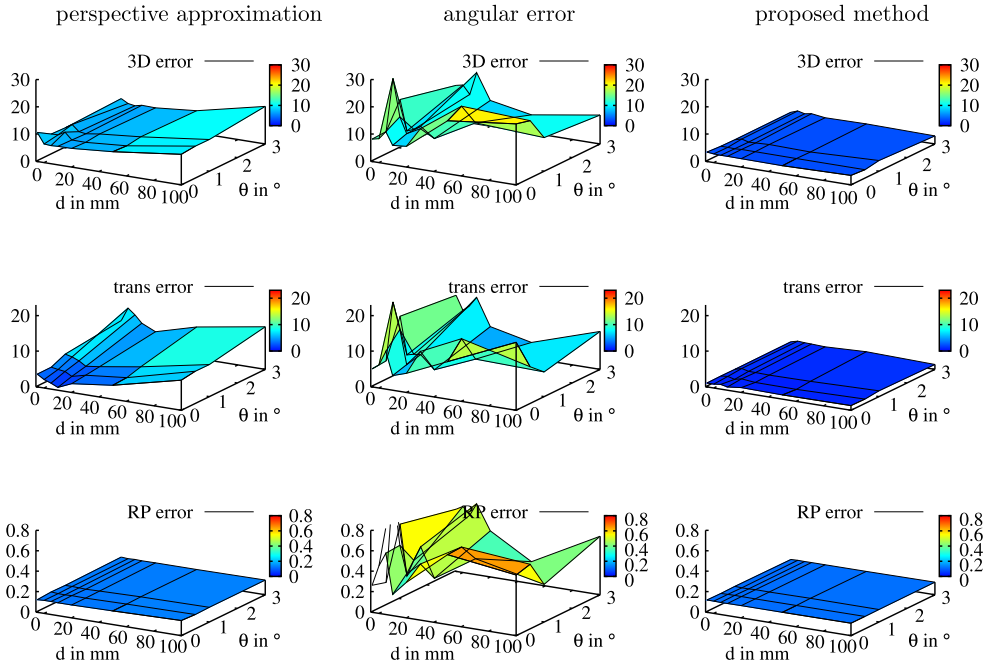


Fig. 14. Results of monocular SfM on fish sequence. Left column: perspective camera model on underwater images. Middle column: results of the method described in Mouragnon et al. (2009) with the iterative relative pose method as a fall back. Right column: refractive camera model on underwater images. From top to bottom: 3D error in mm, camera translation error in mm, and reprojection error in pixels. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

known distance to the system. It is clearly visible that using the perspective camera model causes the reconstruction to have a systematic modeling error, which depends on the interface distance and especially the interface tilt. The method based on the angle error does not have such a systematic modeling error, but a fairly large error altogether. When using the refractive camera model, the systematic modeling error is eliminated completely and the results are much more accurate than in the angle error case. The reason for this is that minimizing the angular error as proposed in Mouragnon et al. (2009) is not a maximum likelihood estimator for observation with Gaussian noise in perspective images, while the novel system is. Even though the resulting errors do not seem to be very large in this example, one has to keep in mind, that the reconstructed camera trajectories were fairly short. In real image sequences, hundreds or several thousands of images are common and the systematic error will accumulate over time.

Fig. 15 shows reconstruction results on the second, orbit-like sequence using a stereo camera rig. As in case of the first scene, the systematic modeling error in case of using the perspective camera model is clearly visible and removed completely in case of using the refractive camera model.

As the next evaluation after the so far sparse reconstructions, the described refractive plane sweep algorithm is applied to the image in order to obtain dense depth maps. Exemplary result images can be seen in Fig. 16. The left column shows a ground truth depth map and input image, the second column shows the resulting depth map using the perspective camera model and an error image compared to ground truth. The third column shows results in case of using the proposed refractive method. Note, that the error images are inverted for better visibility, i.e. darker colors mean a high error. It is clearly visible that the refractive result has less errors than the perspective result. In addition, the error in case of the perspective depth map is depending on the distance between the camera and the 3D objects, which can be observed on the floor and back wall, where the error reaches the order of 20 cm. The average error across all images and pixels is depicted in Fig. 17. As in the case of the SfM results,

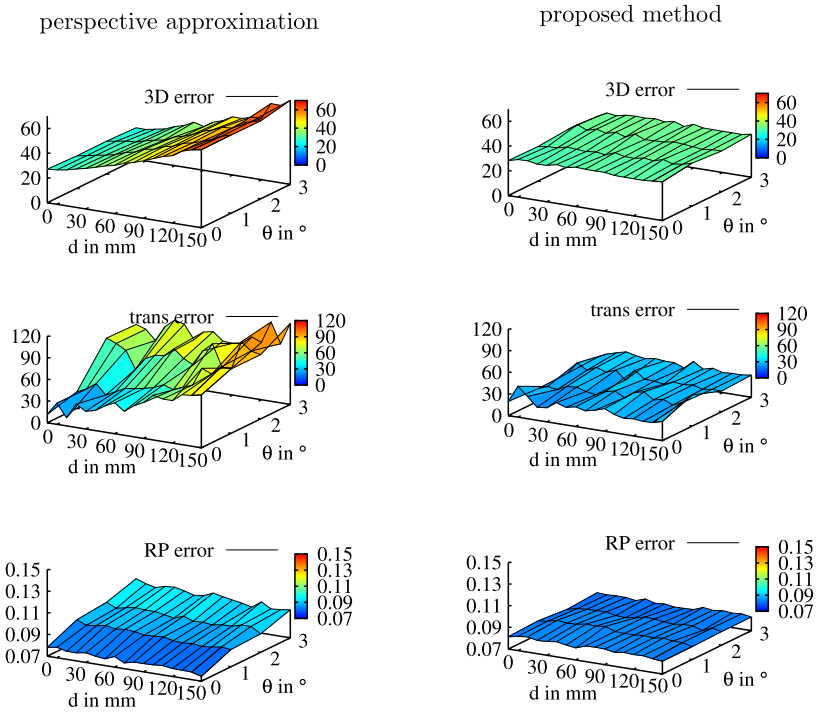


Fig. 15. Results of stereo SfM on orbit-like sequence. Left column: perspective camera model on underwater images. Right column: refractive camera model on underwater images. From top to bottom: 3D error in mm, camera translation error in mm, and reprojection error in pixels. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the systematic modeling error in case of using the perspective camera model on underwater data (Fig. 17, left) is clearly visible. When using the refractive camera model, it is eliminated completely, the remaining error is in the order of the quantization error induced by the discrete depth hypothesis planes of the Plane Sweep algorithm. In summary, the results clearly demonstrate that it is important to model refraction correctly, when reconstructing from underwater data captured with flat port underwater camera housings and that the proposed method is superior to the angle error method, which was designed for more general camera models.

5.3. Real data

5.3.1. Abu Simbel sequence

In order to test the approach on real data, an image sequence was recorded in a controlled lab environment. The use of a camera housing with different configurations was simulated by placing the cameras in front of a fish tank at different distances and camera tilts (see Fig. 18, top). Note that the image does not depict the experiment setup with correct scales. The camera was placed at distances between 7 and 149 mm from the glass. The fish tank itself was of the size 500 mm × 1000 mm × 500 mm and was filled with water. Inside, a model of the entrance to the Abu Simbel temple in Egypt was roughly rotated around its vertical axis, while capturing images. As can be seen in Fig. 18 in the second row, the model is reflected at the bottom, but also on the sides of the lab tank. In addition, the tank itself, but also small gas bubbles on the glass violate the rigid scene assumption when rotating the model. Therefore, each image was roughly segmented prior to reconstruction. Reconstructions (a)–(g) in Fig. 18 show reconstruction results using the refractive (blue) and the perspective camera model. The number of input images, glass configuration, and the

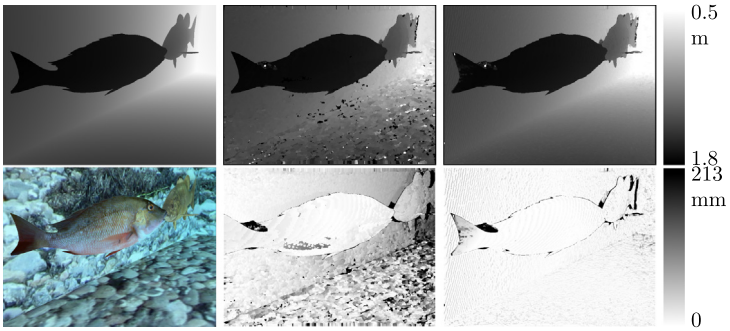


Fig. 16. Exemplary result of refractive plane sweep with housing configuration $d = 100$ mm and $\theta_2 = 3°$. Top row: ground truth depth map, resulting depth map using the perspective model, and resulting depth map using the refractive model. Bottom row: input image, pixel-wise difference to ground truth for perspective result, and pixel-wise difference to ground truth for refractive result.

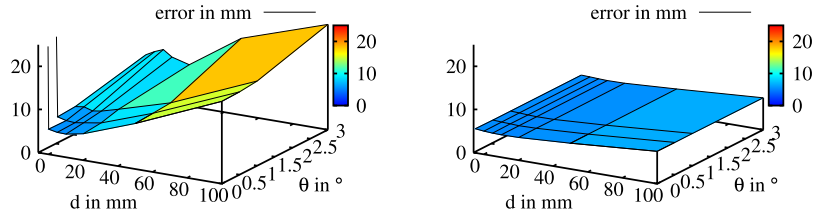


Fig. 17. Refractive plane sweep results. Results for a close scene with distances up to 2300 mm. Left: results of perspective model on perspective images. Right: results of perspective camera on underwater images, and results of refractive camera on underwater images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
The two rightmost columns show the average distance μ_Δ and standard deviation σ_Δ between estimated classical perspective and novel refractive camera positions for seven different camera-glass configurations. Note the large differences in trials (a) and (f), which are cases, where the classical perspective model failed (compare to Fig. 18).

Trial	#images	d/mm	θ/°	μ_Δ/mm	σ_Δ/mm
a	46	7.88	0.34	350.879	312.876
b	52	10.60	0.25	24.791	4.423
c	67	51.95	0.29	26.4426	14.4046
d	65	61.47	7.36	186.571	82.5256
e	76	76.96	29.29	115.596	31.4714
f	87	95.45	0.12	609.384	194.478
g	79	149.39	0.12	79.5105	37.9085

resulting average distance and standard deviation between perspective and refractive camera poses are summarized in Table 1. Although there is no ground truth available in this case, it is clear, that the proposed refractive method outperforms the perspective method. Additionally, the increasing differences between the perspective and refractive results with increasing interface distance and tilt indicate the same systematic modeling error as observed in the synthetic case.

5.3.2. Deep sea data

Apart from testing on real images captured in a controlled lab environment, the refractive reconstruction was also applied to deep ocean data. In this case, the camera was an HDTV video camera enclosed in an underwater housing rated for water depths of 6000 m equipped with a flat glass port.

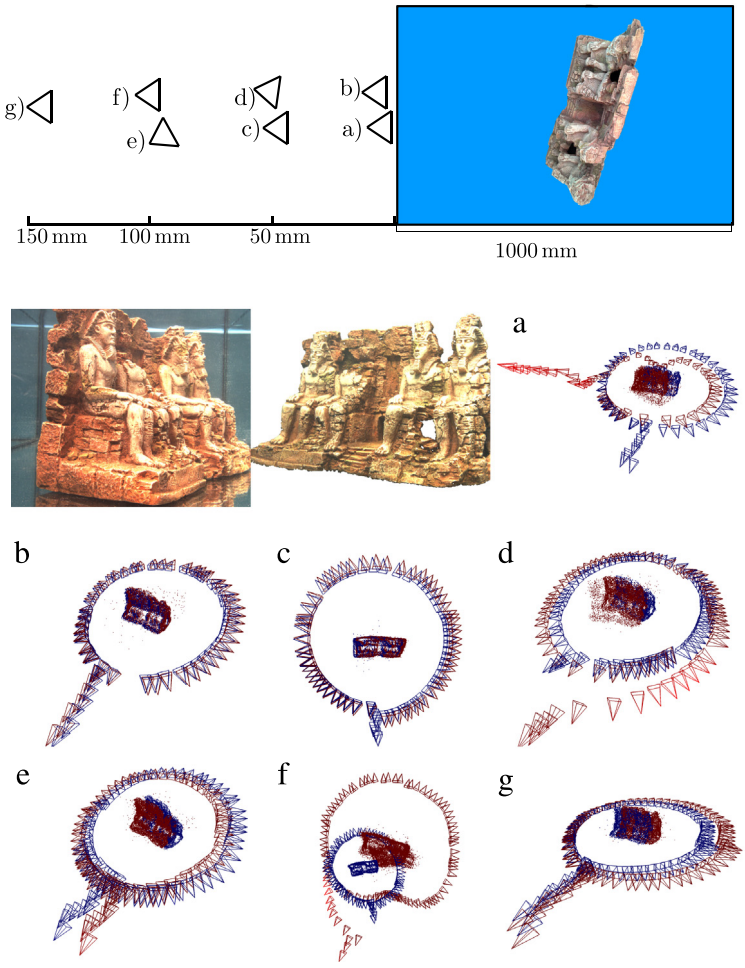


Fig. 18. The first image shows the experimental setup. The camera was placed at different configurations in front of a lab tank. Note that for better visibility, the tank is not up to scale. The second row shows an exemplary input image. Due to the mirrored scene in the tank bottom and other features like small air bubbles on the tank walls, all input images have been roughly segmented (second image in second row). Images (a)–(g) show reconstruction results for the seven different camera-glass configurations. Blue is the camera trajectory and point cloud from the refractive reconstruction, red is from perspective reconstruction (refer to Table 1 for differences in mm between perspective and refractive results). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The camera was attached to the ROV Kiel 6000.[4] It is used to explore deep sea structures as can be seen in Fig. 19. The top row shows input images of a black smoker, a hydrothermal vent found at the Middle Atlantic ridge.

The formation of black smokers, their composition, growth rates and also the faunal communities that populate the habitats around them are important research topics in several ocean science disciplines. The results shown here can be used for volumetric and surface measurements and they can serve as a frame for visualizing biological data, for biomass assessment and many other scenarios. If the same black smoker is visited again, it will be possible to determine changes in volume over time

[4] http://www.geomar.de/en/centre/central-facilities/tlz/rovkiel6000/.

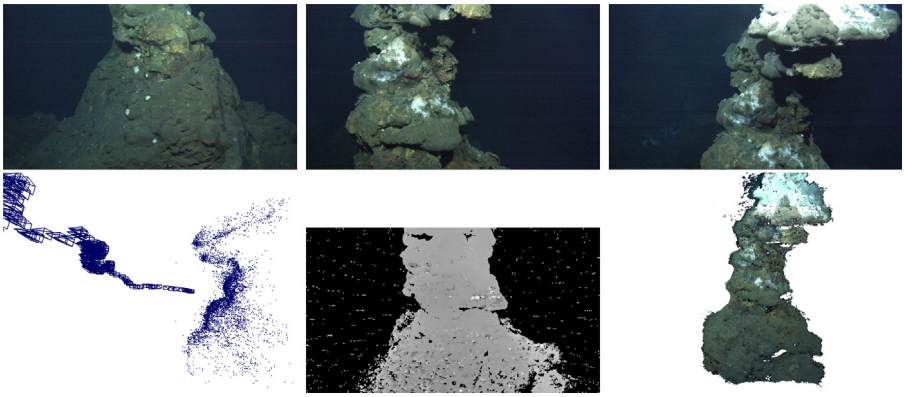


Fig. 19. Results of refractive 3D reconstruction of a black smoker at the middle Atlantic ridge. Top row: Sample images from video sequence. Bottom row: 3D point cloud with camera path (left), sample depth map (center) and textured 3D model (right).

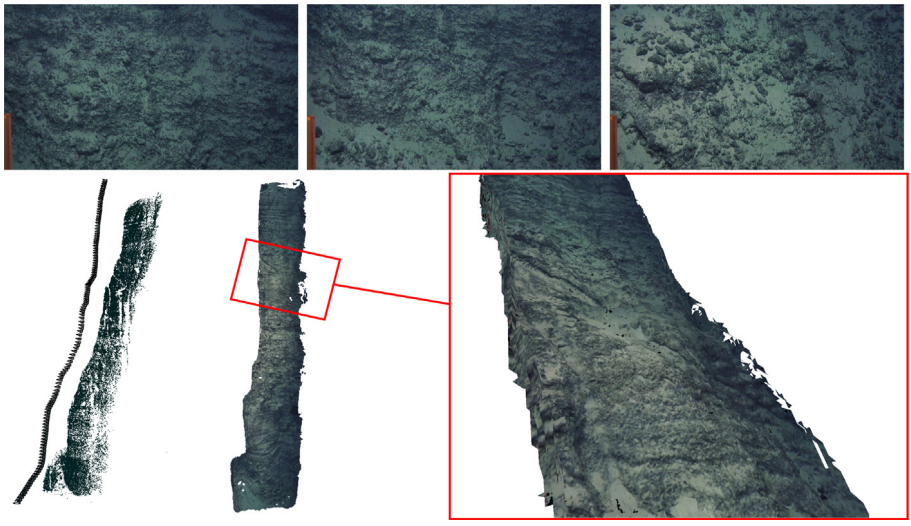


Fig. 20. Results on underwater volcano (Cape Verdes, 3500 m water depth). Top row: sample images from video. Bottom row: reconstructed point cloud with camera path (left), reconstructed 3D model with exemplary detected geological feature (center) and a joint (right, marked in red rectangle) according to Kwasnitschka et al. (2013). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and therefore to monitor its development. The second row in Fig. 19 shows the reconstruction result with sparse 3D point cloud and camera path, exemplary depth map, and textured 3D model.

Fig. 20 shows input images and reconstruction of a part of the inner flank of an underwater volcano found near the Cape Verdes at approximately 3000 m water depth. Here, a geologic research question lies in the actual formation process of this particular volcano. The typical field work flow requires studying the flanks and searching for fault and joint lines, which is however very difficult from small field of view image data and closeup video. However, the digital 3D model resulting from our method allows to *interactively* view and investigate the entire volcano flank after the dive either on a 2D or 3D screen or even in a 3D viewing arena. This is much closer to how geologists usually do their field on land (compare also Kwasnitschka et al., 2013). Finally, Fig. 21 shows input images and reconstruction of the inside wall of an underwater lava lake. The structure was formed by sinking lava levels in the lake causing the edges to tear horizontally. By the resulting 3D model the structures are documented and reconstructed on a mm-scale (depending on the camera's viewing distance).

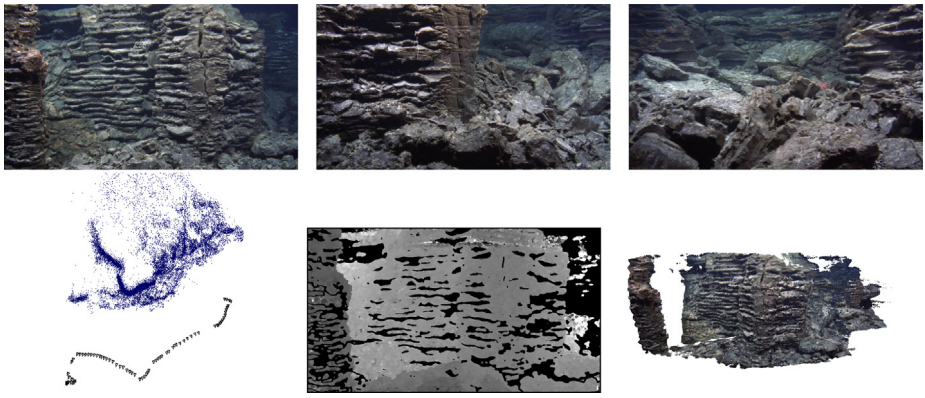


Fig. 21. Reconstruction of an inside wall of a lava lake as the middle Atlantic ridge (4°48′S, 12°22.5′W). Top row: sample images from video. Bottom row: sparse point cloud and cameras (left), sample depth map (center), 3D model (right).

6. Conclusion and future work

When capturing visual oceanographic data for measuring, mapping, and reconstruction, cameras are often mounted in a pressure housing with a flat port. For deep sea housings, the glass becomes significantly thick leading to two refractions per ray. On top, high resolution cameras nowadays provide an angular resolution (e.g. 5000 pixels covering 50° field of view) that makes it difficult to align the optical axis of the camera perfectly (to less than a pixel) with the normal of the port. The approach taken in this paper models the physical parameters of the housing explicitly and integrates them into a refractive image formation model for practical 3D reconstruction of deep sea sequences.

While this refractive projection would be computationally infeasible for large scale vision when applied in a brute force manner, the key idea proposed here is to lift the feature observations from the image onto the outer glass interface of the port. Optimizing a virtual camera error is still a maximum likelihood estimator for the original problem when the image observations contain Gaussian noise. In bundle adjustment this required switching from estimation in the Gauss–Markov model (as for traditional bundle adjustment in air) to the more general Gauss–Helmert model that supports implicit constraints, but which runs at comparable computational costs as perspective bundle adjustment.

The different steps of the refractive reconstruction pipeline have been evaluated in detail on several image sequences and showed similar performance as their counterparts without refraction. The estimation of absolute scale from monocular image sequences, which is not possible without refraction, is feasible in noise-free settings. However, the signal-to-noise ratio for standard action cameras in underwater housings or our deep sea housings at reasonable working distances to the scene is poor. In any case, the reconstruction usually needs to be geo-referenced using absolute navigation data, which are then also used to determine the scale of the reconstruction. The resulting 3D models, as shown for the geology applications, proved the usefulness of the system.

Limitations and failure cases. Refractive direct solvers, as needed for reconstructing unordered image collections are still less reliable, as e.g. the generalized essential matrix requires using many correspondences and minimal solvers did not cope well with the high noise often present in real oceanographic data. Also, when purely relying on vision data, reconstructions are often disrupted when no visual structures are present or when sediment is dispersed into the water column. This study should therefore be seen as one component of a bigger system that should detect vision failure and integrates and weights also other sensors such as sonar or any kind of navigation.

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) and the Cluster of Excellence "The Future Ocean" through grant "Opti-Acoustic Sensor Fusion for Highly Detailed and Accurate 3D Modeling".

References

Agrawal, A., Ramalingam, S., Taguchi, Y., Chari, V., 2012. A theory of multi-layer flat refractive geometry. In: CVPR.
Bingham, B., Foley, B., Singh, H., Camilli, R., Delaporta, K., Eustice, R., Mallios, A., Mindell, D., Roman, C., Sakellariou, D., 2010. Robotic tools for deep water archaeology: Surveying an ancient shipwreck with an autonomous underwater vehicle. J. Field Robot. 27, 702–717.
Chang, Y.-J., Chen, T., 2011. Multi-view 3D reconstruction for scenes under the refractive plane with known vertical direction. In: IEEE International Conference on Computer Vision, ICCV.
Chari, V., Sturm, P., 2009. Multiple-view geometry of the refractive plane. In: Proceedings of the 20th British Machine Vision Conference, London, UK.
Costa, C., Loy, A., Cataudella, S., Davis, D., Scardi, M., 2006. Extracting fish size using dual underwater cameras. Aquac. Eng. 35 (3), 218–227.
Dementhon, D.F., Davis, L.S., 1995. Model-based object pose in 25 lines of code. Int. J. Comput. Vis. 15, 123–141. http://dx.doi.org/10.1007/BF01450852.
Farenzena, M., Fusiello, A., Gherardi, R., 2009. Structure-and-motion pipeline on a hierarchical cluster tree. In: 3DIM09, pp. 1489–1496.
Fischler, M., Bolles, R., 1981. Random sampling consensus: a paradigm for model fitting with application to image analysis and automated cartography. Commun. ACM 24 (6), 381–395. http://dx.doi.org/10.1145/358669.358692.
Fryer, J.G., Fraser, C.S., 1986. On the calibration of underwater cameras. Photogramm. Rec. 12, 73–85.
Gallup, D., Frahm, J.-M., Mordohai, P., Qingxiong, Y., Pollefeys, M., 2007. Real-time plane-sweeping stereo with multiple sweeping directions. In: IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07, pp. 1–8.
Glaeser, G., Schröcker, H.-P., 2000. Reflections on refractions. J. Geom. Graph. (JGG) 4, 1–18.
Grossberg, M.D., Nayar, S.K., 2005. The raxel imaging model and ray-based calibration. Int. J. Comput. Vis. 61 (2), 119–137.
Hartley, R.I., Sturm, P.F., 1997. Triangulation. Comput. Vis. Image Underst. 68 (2), 146–157.
Hartley, R., Zisserman, A., 2004. Multiple View Geometry in Computer Vision (Second Edition), second ed. Cambridge University Press.
Harvey, E.S., Shortis, M.R., 1998. Calibration stability of an underwater stereo-video system: Implications for measurement accuracy and precision. Mar. Technol. Soc. J. 32, 3–17.
Hecht, E., 1998. Optics, fourth ed. Addison-Wesley.
Heikkila, J., Silven, O., 1997. A four-step camera calibration procedure with implicit image correction. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 0, pp. 1106–1112 http://dx.doi.org/10.1109/CVPR.1997.609468.
Henderson, J., Pizarro, O., Johnson-Roberson, M., Mahon, I., 2013. Mapping submerged archaeological sites using stereo-vision photogrammetry. Int. J. Naut. Archaeol. 42 (2), 243–256. http://dx.doi.org/10.1111/1095-9270.12016.
Inglis, G., Smart, C., Vaughn, I., Roman, C., 2012. A pipeline for structured light bathymetric mapping. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, pp. 4425–4432. http://dx.doi.org/10.1109/IROS.2012.6386038.
Johnson-Roberson, M., Pizarro, O., Williams, S.B., Mahon, I.J., 2010. Generation and visualization of large-scale three-dimensional reconstructions from underwater robotic surveys. J. Field Robot., 27.
Jordt, A., 2014. Underwater 3D reconstruction based on physical models for refraction and underwater light propagation (Ph.D. thesis), Christian-Albrechts-Universtät zu Kiel, Germany.
Jordt-Sedlazeck, A., Jung, D., Koch, R., 2013. Refractive plane sweep for underwater images. In: Weickert, J., Hein, M., Schiele, B. (Eds.), Pattern Recognition. In: Lecture Notes in Computer Science, vol. 8142. Springer, Berlin, Heidelberg, pp. 333–342. http://dx.doi.org/10.1007/978-3-642-40602-7_36.
Jordt-Sedlazeck, A., Koch, R., 2012. Refractive calibration of underwater cameras. In: Fitzgibbon, A., Lazebnik, S., Pietro, P., Sato, Y., Schmid, C. (Eds.), Computer Vision—ECCV 2012. In: Lecture Notes in Computer Science, vol. 7576. Springer, Berlin, Heidelberg, pp. 846–859.
Jordt-Sedlazeck, A., Koch, R., 2013. Refractive structure-from-motion on underwater images. In: 2011 IEEE International Conference on Computer Vision, ICCV, pp. 57–64. http://dx.doi.org/10.1109/ICCV.2013.14.
Kanatani, K., 1996. Statistical Optimization for Geometric Computation: Theory and Practice. Elsevier.
Kang, L., Wu, L., Yang, Y.-H., 2012. Two-view underwater structure and motion for cameras under flat refractive interfaces. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (Eds.), Computer Vision—ECCV 2012. In: Lecture Notes in Computer Science, vol. 7575. Springer, Berlin, Heidelberg, pp. 303–316.
Koch, R., 1993. Dynamic 3-d scene analysis through synthesis feedback control. IEEE Trans. Pattern Anal. Mach. Intell. 15 (6), 556–568.
Kotowski, R., 1988. Phototriangulation in multi-media photogrammetry. Intl Archives of Photogrammetry and Remote Sensing XXVII.
Kwasnitschka, T., Hansteen, T.H., Devey, C.W., Kutterolf, S., 2013. Doing fieldwork on the seafloor: Photogrammetric techniques to yield 3D visual models from ROV video. Comput. Geosci. 52, 218–226.
Lavest, J.-M., Rives, G., Lapresté, J.-T., 2000. Underwater camera calibration. In: ECCV'00: Proceedings of the 6th European Conference on Computer Vision-Part II, pp. 654–668.
Li, H., Hartley, R., Kim, J.-H., 2008. A linear approach to motion estimation using generalized camera models. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008, pp. 1–8. http://dx.doi.org/10.1109/CVPR.2008.4587545.

Li, R., Li, H., Zou, W., Smith, R.G., Curran, T.A., 1997. Quantitative photogrammetric analysis of digital underwater video imagery. IEEE J. Ocean. Eng. 22 (2), 364–375. http://dx.doi.org/10.1109/48.585955.
McGlone, J.C. (Ed.), 2004. Manual of Photogrammetry, fifth ed. ASPRS.
Mouragnon, E., Lhuillier, M., Dhome, M., Dekeyser, F., Sayd, P., 2009. Generic and real-time structure from motion using local bundle adjustment. Image Vis. Comput. 27, 1178–1193. http://dx.doi.org/10.1016/j.imavis.2008.11.006.
Mulsow, C., 2010. A flexible multi-media bundle approach. In: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. 38. pp. 472–477.
Nistér, D., 2004. An efficient solution to the five-point relative pose problem. TPAMI 26, 756–777.
Nistér, D., Stewénius, H., 2007. A minimal solution to the generalised 3-Point pose problem. J. Math. Imaging Vision 27, 67–79.
Pless, R., 2003. Using many cameras as one. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings, Vol. 2, pp. 587–593. http://dx.doi.org/10.1109/CVPR.2003.1211520.
Press, W.H., Vetterling, W.T., Teukolsky, S.A., Saul, A., Flannery, B.P., 2002. Numerical Recipes in C++: The Art of Scientific Computing, second ed. Cambridge University Press, New York, NY, USA.
Ramalingam, S., Lodha, S.K., Sturm, P., 2006. A generic structure-from-motion framework. Comput. Vis. Image Underst. 103 (3), 218–228.
Scaramuzza, D., Martinelli, A., Siegwart, R., 2006. A flexible technique for accurate omnidirectional camera calibration and structure from motion. In: IEEE International Conference on Computer Vision Systems, 2006 ICVS'06, pp. 45–45. http://dx.doi.org/10.1109/ICVS.2006.3.
Schiller, I., Beder, C., Koch, R., 2008. Calibration of a PMD camera using a planar calibration object together with a multi-camera setup. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXVII, Part B3a, Beijing, China, pp. 297–302, xXI. ISPRS Congress.
Sedlazeck, A., Koch, R., 2011. Calibration of housing parameters for underwater stereo-camera rigs. In: Hoey, J., McKenna, S., Trucco, E. (Eds.), Proceedings of the British Machine Vision Conference. BMVA Press, ISBN: 1-901725-43-X, pp. 118.1–118.11. http://dx.doi.org/10.5244/C.25.118.
Sedlazeck, A., Koch, R., 2011. Simulating deep sea underwater images using physical models for light attenuation, scattering, and refraction. In: Eisert, P., Hornegger, J., Polthier, K. (Eds.), VMV 2011: Vision, Modeling & Visualization. Eurographics Association, Berlin, Germany, pp. 49–56. no. 978-3-905673-85-2.
Sedlazeck, A., Koch, R., 2012. Perspective and non-perspective camera models in underwater imaging–overview and error analysis. In: Dellaert, F., Frahm, J.-M., Pollefeys, M., Leal-Taixé, L., Rosenhahn, B. (Eds.), Outdoor and Large-Scale Real-World Scene Analysis. In: Lecture Notes in Computer Science, vol. 7474. Springer, Berlin, Heidelberg, pp. 212–242.
Sedlazeck, A., Köser, K., Koch, R., 2009. 3D reconstruction based on underwater video from ROV kiel 6000 considering underwater imaging conditions. In: Proc. OCEANS'09. OCEANS 2009-EUROPE, pp. 1–10. http://dx.doi.org/10.1109/OCEANSE.2009.5278305.
Snavely, N., Seitz, S., Szeliski, R., 2008. Modeling the world from Internet photo collections. Int. J. Comput. Vis. 80 (2), 189–210. http://dx.doi.org/10.1007/s11263-007-0107-3.
Sturm, P., Ramalingam, S., Lodha, S., 2006. On calibration, structure from motion and multi-view geometry for generic camera models. In: Daniilidis, K., Klette, R. (Eds.), Imaging Beyond the Pinhole Camera. In: Computational Imaging and Vision, vol. 33. Springer.
Szeliski, R., 2011. Computer Vision: Algorithms and Applications. Springer-Verlag.
Treibitz, T., Schechner, Y.Y., Singh, H., 2008. Flat refractive geometry. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition CVPR 2008, pp. 1–8.
Triggs, B., McLauchlan, P., Hartley, R., Fitzgibbon, A., 2000. Bundle adjustment—A modern synthesis. In: Triggs, B., Zisserman, A., Szeliski, R. (Eds.), Vision Algorithms: Theory and Practice. In: Lecture Notes in Computer Science, vol. 1883. Springer-Verlag, pp. 298–372.
Williams, S.B., Pizarro, O.R., Jakuba, M.V., Johnson, C.R., Barrett, N.S., Babcock, R.C., Kendrick, G.A., Steinberg, P.D., Heyward, A.J., Doherty, P.J., Mahon, I., Johnson-Roberson, M., Steinberg, D., Friedman, A., 2012. Monitoring of benthic reference sites: Using an autonomous underwater vehicle. IEEE Robot. Autom. Mag. 19 (1), 73–84. http://dx.doi.org/10.1109/MRA.2011.2181772.
Wu, C., 2007. SiftGPU: A GPU implementation of scale invariant feature transform (SIFT). http://cs.unc.edu/~ccwu/siftgpu.
Xu, X., Negahdaripour, S., 2001. Application of extended covariance intersection principle for mosaic-based optical positioning and navigation of underwatervehicle. In: ICRA'01, pp. 2759–2766.