Towards Photometric Stereo for Underwater Robots

Monika Roznere and Alberto Quattrini Li

***Abstract*—This paper discusses a photometric stereo framework to enable real-time scene 3D reconstruction of underwater structures with a non-stationary small low-cost underwater robot equipped with a monocular camera and fixed intensity-controllable lights. Previous approaches for underwater photometric stereo provided accurate scene reconstruction results, assuming that the robot is stationary at the bottom. This assumption limits the structures that can be reconstructed. In addition, lights are assumed to be either on or off and set at relatively large distances to the camera. To achieve photometric stereo on a small low-cost robot, such as the BlueROV2, there are two fundamental questions that we explore in this paper: 1) the use of images consisting light intensity changes, so that changes in light position are not necessary; 2) the relaxation of the camera/robot stationary assumption. After defining the photometric stereo model, we experimentally show that changes in light intensity result in saturated areas on the object, providing constraints to solve the photometric stereo problem. In addition, we discuss the use of binary features, such as ORB, to estimate small motion and incorporate the estimations in the photometric stereo objective function. The photometric stereo framework and insights discussed in this paper are the foundation for the full implementation of the photometric stereo model on a low-cost robot, which will then be used within an active perception pipeline to let the robot explore a structure and capture high-quality data.**

I. Introduction

In this paper, we present preliminary work for solving the photometric stereo model in the case of non-stationary underwater robots. Photometric stereo is a well known computer vision technique for reconstructing high resolution scenes or objects, typically out of water and considering stationary cameras.

Scene reconstruction is a common and important aspect in many underwater tasks, particularly for inspecting man-made structures (i.e., dams, oil rigs, ship hulls) [1], monitoring target biological locations [2], and exploring reefs [3] and archaeological sites [4]. Autonomous Underwater Vehicles (AUVs) are becoming more commonly dispatched to tackle these various tasks. Not only can AUVs stay longer underwater than a diver, but they are also typically set up with a modular sensory suit – at the very least with an IMU, monocular or stereo camera, single-beam echosounder, and lights [5], and more extensively (and more upscaled in price) with a multibeam sonar, side scan sonar, and guidance-based equipment (i.e., fiber-optic gyroscope (FOG) IMU, acoustic Doppler Velocity Log (DVL)) [6], [7]. We note that the new Water Linked DVLs may be considered as low-cost sensors (∼USD 6k), but will be omitted in this paper.

Department of Computer Science, Dartmouth College, USA {monika.roznere.gr, alberto.quattrini.li}@dartmouth.edu

Fig. 1: A photometric stereo framework for non-stationary underwater robots allows low-cost AUVs, like the BluerROV2, here to explore shipwrecks while computing scene reconstruction models, even under dynamic lighting conditions.

Multibeam and other sonars were shown to be extremely useful as sensory input for accurate underwater scene reconstruction [1], [8]. However, sonars lack the visual (i.e., color, texture) and resolution characteristics that cameras provide – which can be enriched by fusing sonar and camera(s) data. On its own, monocular camera-imagery input cannot provide accurate scene depth information [9]. IMU or DVL data can be integrated [10]–[12], but in our case with a low-cost underwater robot like the BlueROV2, we will uphold that the IMU is too unreliable. We have shown in previous work that a low-cost single-beam echosounder can improve monocular camera scene depth estimation [5].

Photometric Stereo (PS) relies solely on camera imagery *and* light sources (artificial or natural), so it does not require expensive multibeam or side scan sonars. It is originally [13] based on the observation that an object's surface normals can be estimated by observing changes in the surface points' reflected light intensities among different images, where light source(s) change position, but the camera's position *always* stays in place. For a review of different PS methods, we refer the reader to these works [14], [15].

An underwater environment is a difficult scenario for any camera-based scene reconstruction technique. Light attenuates (reduces in intensity) more extensively in water than in air; as light waves travel in water, they are scattered and absorbed by the different particles they collide with. Attenuation is the main reason why images taken underwater look as if they have lost color and contrast (e.g., murky, blurry). The amount and manner that the light attenuates are dependent on the oceanic properties in that location at that time and the

distance that the light must travel. Accurately calculating the attenuation parameters for a certain body of water requires a spectrophotometer and other specialty instruments [16]. Fortunately, there has been work on how to estimate the attenuation values, typically for the goal of color correcting or enhancing underwater images [17]–[19]. Many of these methods followed an image formation model, extended from the model used for in-air imagery, such that it includes the additional lighting effects that occur in water [20].

Underwater PS follows a similar image formation model [21]–[23]. It furthermore includes information on how incoming light attenuates during travel, how it reflects from each surface point of the scene object, and the distances of each surface point from the camera and light(s).

Another issue related to underwater scenarios is that the behavior of ambient light – the amount of sunlight directly above the surface of the water and how it enters the water – is very unpredictable and dynamic. The amount of ambient light entering the water depends on the weather, time of day, and presence of any obstacles (i.e., boats, swimmers). When the light rays enter the water, they are refracted, and with the presence of surface waves, the light rays are directed to multiple different directions over time, causing light ripple effects on the below underwater scene. Many works that utilized the underwater image formation model either did not consider how to thoroughly solve the issue of the dynamic ambient light [12], deployed AUVs at a sufficient depth where on-board lights can be assumed to be the main sources of illumination, or conducted experiments at night [3], [24], [25]. Unpredictable, dynamic light – not surprisingly – harmfully affects the results of underwater scene reconstruction models based on input from camera-imagery only [12].

Work on artificial light intensity control [26], [27] and camera relative to scene positioning [3] were shown to provide invaluable results when constructing 3D models of objects. However, many of these approaches assumed ambient light was negligible or could be handled with statistical methods [12]. By including repetitive ambient light estimation and collecting camera images under various artificial lighting for each camera-to-scene pose, we are hopeful to mitigate scene reconstruction error due to dynamic ambient lighting.

Accordingly, PS methodologies already require multiple images under different lighting conditions to optimize the scene's orientation and 3D surface. Thus, we believe that a PS approach can be applicable to low-cost AUVs that are exploring a shipwreck, under dynamic lighting conditions, and collecting images to be used for scene reconstruction – killing two birds with one stone.

Photometric stereo for non-stationary underwater robots is to the best of our knowledge a novel, unsolved problem. Previous PS works [24], [25] that tested with underwater robots required that the robots stay settled on the bottom to ensure that the camera does not move. This is undesirable as (1) target objects (i.e., corals, parts of shipwrecks) may be located meters above the bottom, (2) most underwater robots are setup to be neutrally or positively buoyant, thus requiring motor usage to stay on the bottom and that may cause sediment stirring and image hazing, and (3) it is common underwater practice to not touch the habitat in order to avoid accidental interference with sensitive organisms and artifacts.

PS for non-stationary underwater robots introduces new issues and research questions:

- **R1:** *Can images taken with the light position in place, but whose light intensity changes, be as informative as images taken with the light position changing?* Installing four individually controlled lights, each in a different position, is a waste of cable port usage (i.e., it would take up 4 of the 14 penetrators of a typical end cap on a BlueROV2). To minimize the number of lights required, we will look into a PS model that considers how different light intensities may change the scale of brightness in the scene, but, more interestingly, it may also identify surface patches whose brightness is at saturation – the reflected light (brightness) will not increase with a stronger light intensity.
- **R2:** *Can an object be well reconstructed using a PS model, even if the images were taken while the robot was moving?* The underwater robot cannot always stay in place while it is suspended in water; it will slightly move due to external water forces or motor usage. We will look into the following: how robot movement can be approximately estimated by short-term visual feature tracking, how much robot movement corresponds to reconstruction error, and how robot movement needs to be incorporated into the PS optimization scheme.

The ultimate goal is to implement an underwater 3D scene reconstruction algorithm that is embedded into an active perception pipeline for real-time visual data collection and mapping performed by low-cost underwater robots.

The rest of the paper is organized as follows: Section II explains the underwater image formation model, Section III illustrates the differences in light models, and Section IV describes the assumptions that we make about the scene. Section V finally combines all of the information from the previous sections to define the model used for solving the PS problem. Afterwards, in Section VI we will explain in more detail our research questions **R1** and **R2** and the preliminary work we have accomplished. Finally, the paper will wrap up with a discussion on future steps and a conclusion of our current observations.

II. Underwater Image Formation Model

In an underwater scene, an image I captured by the camera's image sensor follows the simplified image formation model [17], [18], [20]:

$$I = D + B \tag{1}$$

where it is composed of the direct signal D and backscatter B, as illustrated in Fig. 2. For the rest of the description of the image formation model, we will refer to grayscale imagery, as color RGB-based model is more complex with additional unknowns. In addition, explanations below will be simplified

and will describe a singular pixel point x that corresponds to a unique surface point, such that $I = I_x$.

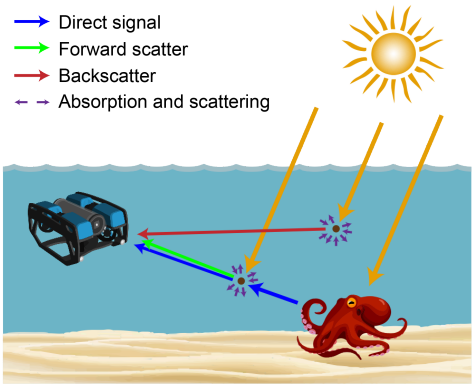

Fig. 2: Illustration of how direct signal and backscatter attenuation components arise. Direct signal consists of the information coming from the camera-visible scene, and backscatter consists of the light that was scattered to the camera before reaching the scene. Forward scatter is included, but is usually negligible or can be estimated as part of the camera properties.

A. Direct Signal

The direct signal D component corresponds to the amount of light that has traveled from the light source, reflected from the visible scene, and reached the camera's image sensor pixel. During the light beam's travel, it is attenuated based on the water medium's characteristics, including beam attenuation, absorption, and scattering coefficients. These attenuation properties are packaged and approximated as a singular coefficient value β^D, as described in:

$$D = J\frac{e^{-\beta^D(|PS_i|+|OP|)}}{(|PS_i|+|OP|)^2} \tag{2}$$

In color correction terminology, J is the unattenuated (corrected) image. In other words, D is the distorted image of J, attenuated by β_c^D and magnified by the distance from the light source S_i to the surface point P, $|PS_i|$, and the distance from the surface point to the camera O, $|OP|$.

More specifically, J consists of the light reflected from the surface point L_R, the surface point's albedo a, its unit normal vector $\hat{\mathbf{n}}$, and the incident light direction from the source $\hat{\mathbf{l}}_{\mathbf{PS_i}}$,

$$J = a\, L_R\, \hat{\mathbf{l}}_{\mathbf{PS_i}} \cdot \hat{\mathbf{n}} \tag{3}$$

For simplicity, $\hat{\mathbf{n}}$ and a can be combined as $\mathbf{n}$.

B. Backscatter

The backscatter B component, on the other hand, corresponds to the amount of light that *never* reached the visible scene, but was reflected by particles in the water and arrived at the camera's image sensor pixel. It is also affected by the water medium's attenuation properties, but unlike direct signal's attenuation coefficient, it is affected by a different proportion of the beam attenuation, absorption, and scattering coefficient, which we denote as β^B:

$$B = B^\infty(1 - \frac{e^{-\beta^B(|PS_i|+|OP|)}}{(|PS_i|+|OP|)^2}) \tag{4}$$

Note, in many image formation model applications, β^D and β^B are assumed to be the same.

The veiling light B^∞ is composed of the water medium's diffuse attenuation, backscattering, and beam absorption coefficients. Roughly, the veiling light can be visualized as the color in the viewing scene that does not consist of any physical objects or the ground, or in other words, it is the color in the far 'infinite' distance if the image is taken horizontally. The veiling light can be calculated as

$$B^\infty = \gamma\, L^\infty(z) \tag{5}$$

where $L^\infty(z)$ is a function of the total light arriving at the image sensor that was reflected midwater along the line of sight. Here, the range z describes the distance where the illumination from the artificial light source overlaps with the line of sight, or, in a simplified version, the distance between the camera O and the scene point P. In many cases, $L^\infty(z)$ can be approximated as a singular value L^∞ for a given scene. γ is described, typically in function form, as the proportion of light scattered or the scattering scale in the medium.

C. Backscatter Estimation

Traditionally, backscatter is estimated to lessen the complexity of the underwater image formation model. From analysis [20], backscatter increases exponentially over distance until it reaches to a point of near saturation.

At saturation, the backscatter value stays constant and does not change. The regions in the image where there are no visible objects, or the viewing distance is infinite (background), can be assumed to be composed of more backscatter than direct signal $B \gg D$, or simply $D = 0$. By taking the average color pixel value in the image background, we can estimate the backscatter component. More intuitively, this estimated background color is assumed to be the veiling light B^∞. Accordingly, pixels that contribute to scene points will have a less saturated backscatter component value. By subtracting the estimated backscatter value from the entire image, all that is left is the direct signal component.

On the other hand, if the images are taken at very close ranges, the backscatter component is assumed to be negligible, $D \gg B$. Thus, the backscatter component is estimated to be $B = 0$.

D. Attenuation Estimation

Attenuation coefficient values are constant throughout the general area and depth. It can be estimated if one follows any underwater image model-based algorithm with ground truth targets. For example, a color chart of known size and colors is a great tool to help estimate attenuation values in the RGB spectrum prior to a robot's run or by placing them in the viewing scene as in-situ information [3], [17].

For grayscale applications, attenuation can be estimated with a white (Lambertian type surface) board [25]. The board is of known size, albedo, surface normals, and distance away from the camera. With one image taken at the interested water depth, one can calculate the attenuation for direct signal.

III. Light Models

In daylight, ambient light L_E is a significant source of illumination for the first 20-30 m deep in the water column. With increasing depth, ambient light's intensity diminishes due to attenuation and the inverse-square law. As ambient light's original intensity L_{E_0} at the surface and its color properties are unknown for the most part without specialized equipment or calibration, it is typically treated as an unknown. If ambient light is the only light source in the scene, then it can be approximated as $L_{E_0} = 1$ at the surface. At deeper depths d, it can be estimated as

$$L_E(d) = L_{E_0} \frac{K_d}{d^2} \tag{6}$$

where K_d is the diffuse attenuation coefficient. If ambient light is not the only light source in the scene – as in, the AUV has its artificial lights on – then one can capture an image of the scene with only ambient light present and use that to subtract the following images that include artificial light sources.

Each artificial light source i will be represented as a spotlight and its original intensity L_{i_0} is known at all time. Like the work by Bryson *et al.* [3], we assume that each light source's intensity is the brightest along its center directional line and it decreases with an increased angle ϕ from this center line. This can modeled as if there is a Gaussian diffuse filter in front of the light:

$$L_{i_\phi} = L_{i_0} e^{-\frac{1}{2}\frac{\phi^2}{\sigma_i^2}} \tag{7}$$

$$\sigma_i = \sqrt{\frac{\phi_{50\%}^2}{-2\log 0.5}} \tag{8}$$

where $\phi_{50\%}$ is the angle at which the power of the light drops off to 50% – this is typically provided in technical sheets of light sources or can be estimated using a white chart calibration setup.

IV. Scene Reflection

The underwater scene is assumed to be composed of Lambertian type surfaces [3], [24], [25]. Thus, the amount of light reflected from a surface point x is only dependent on the angle or light direction relative to the surface normal – it is not dependent on the viewing direction. In other words, given a surface normal and light direction, the same amount of reflected light will be observed in any viewing direction as:

$$L_{R_x} = \sum_{i=1}^{N_S} (L_{i\phi} \cos(\theta_i)) \tag{9}$$

where $\cos(\theta_i)$ is the angle between the surface normal and incident light direction $\hat{\mathbf{l}}_{\mathbf{PS_i}}$. Here, N_S is the number of light sources that reach the surface point x. Note, reflected light is also composed of the albedo a, but it is not shown in the above equation, as it is included when calculating J in Equation (3). Likewise, the attenuation that occurs during the light travel through the water medium is included in the calculations for D in Equation (2) (and B in Equation (4)). The inverse-square law is also applied to both equations, as light loses intensity over distance.

V. Underwater Photometric Stereo

The complete underwater image formation model, for *one* light source, also termed as the reflectance model, is as follows

$$I = k \frac{C(\lambda)\, \mathbf{n}\, L_{i\phi} \cos(\theta_i)\, e^{-\beta^D z_i} + B^\infty (1 - e^{-\beta^B z_i})}{z_i^2} \tag{10}$$

where $z_i = |PS_i| + |OP|$, k is the scalar directing image exposure, and $C(\lambda)$ comprises of additional camera properties, such as vignetting – it can also be included to account for missing attenuation factors, including forward scattering, where light rays reflected from scene surface points interact with particles and are slightly veered off in their original direction, causing distinct blurriness in the final captured image. More lights can be incorporated, which would require that the reflected light is the sum of all lights that reach the surface point and to include all distance values corresponding to each light source.

A. Near and Distant Lighting Models

There are two types of lighting model assumptions: distant-lighting and near-lighting, as shown in Fig. 3. As explained below, they are dependent on the distance between the camera and the scene as well as the size of the scene (target object).

The distant-lighting model is the simpler of the two cases. It is applied when the distance between the camera and the target object is much larger than the relative size of the target object. In this case, the light direction from a source to each surface point in the scene can be assumed to be the same. Hence, the distances between a light source to each surface point are assumed to be all the same and the distances between the camera to each surface point are assumed to be all the same. Ambient light (sunlight) is always assumed to be based on distant-lighting.

On the other hand, the near-lighting model considers when such distances are smaller or when the size of the target object is much larger. This is important when the illumination from the on-board light sources are non-uniform. In this case, no parameter values should be assumed to be the same with respect to each surface point.

Typically, the near-lighting model produces the highest accuracy of results, however, as one can imagine, it is also the most complex of the two. A clever approach [24] is to setup an iterative convergence algorithm for solving the estimated PS parameters. For the initial round, the distant-lighting assumption is used, which will most likely provide substantial reconstruction error. Consecutive iterations in the optimization scheme are then based on an intuitive observation that the normals calculated for a surface point will match more

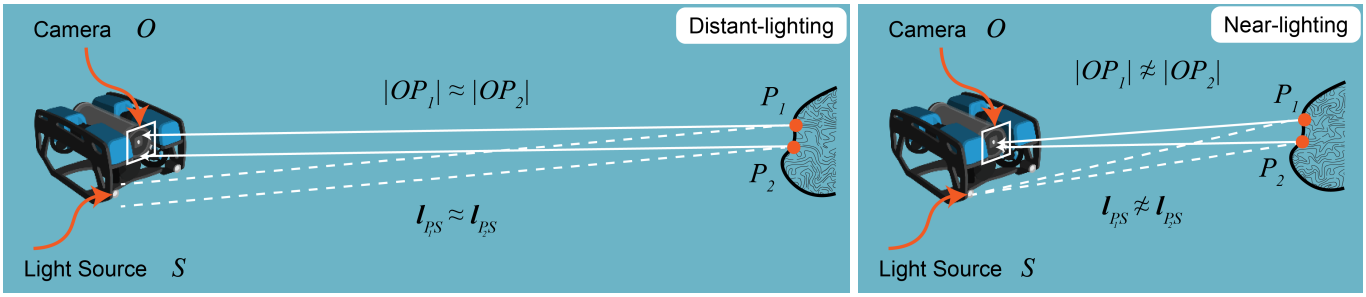

Fig. 3: (**left**) Distant-lighting assumption is applied when the distance between the camera and target object is much greater than the relative size of the target object. In this case, the parameters corresponding to distance and lighting direction for each surface point will be assumed to be the same. (**right**) On the other hand, near-lighting assumption is applied when the distance between the camera and target object is relatively small or the illumination from the light sources is non-uniform on the scene. In this case, all parameters corresponding to each surface point cannot be assumed to be the same.

closely if the lighting model used is correct. If the normals for many surface points are as close as possible but still off, given a set of predicted values for the other unknown parameters, then converting the model to a near-lighting model will lessen the restriction of the distances between the light sources and the surface points, allowing for more adjustment in calculating the surface normals.

B. Image Depth Estimation

Generally, the distance of the robot to the scene is unknown. While we can assume that the initial optimization scheme assumes a distant-lighting model to account for less unknowns, we still cannot accurately estimate what these distance values should be set as by only using monocular camera information. As shown in our previous work, we can use a forward-looking single-beam echosounder to provide a more accurate initial guess of the scene depth [5]. A single-beam echosounder emits an acoustic wave in a cone shape and will return the distance value of the most prevalent reflected signal back. Thus, it is unclear where in the cone this distance value is associated with, especially with larger distance values. One can simply estimate that for the entire given viewing scene, the distance between the camera/lights to the surface points are proportional to the echosounder's measurement – useful for first iteration assumption of distant-lighting model.

C. Photometric Stereo Unknowns

The main unknowns that we are interested in estimating are the normal vectors $\mathbf{n}$ for each visible surface point denoted by its own image pixel and the depth Z of the scene point relative to some plane parallel to the image plane.

Other unknowns include:

- distance between camera and each surface point $|OP|$,

$$|OP| = \sqrt{(\frac{uZ}{f})^2 + (\frac{vZ}{f})^2 + Z^2} \tag{11}$$

  where f is the camera's focal length, and (u, v) is an image pixel coordinate.

- distance between each light source and each surface point $|PS_i|$, if not obscured,

$$|PS_i| = \sqrt{(X_i - \frac{uZ}{f})^2 + (Y_i - \frac{vZ}{f})^2 + (Z_i - Z)^2} \tag{12}$$

- light direction relative to a light source and a surface point $\hat{\mathbf{l}}_{\mathbf{PS_i}}$, or the angle between a light source and a surface point $\cos(\theta_i)$

$$\hat{\mathbf{l}}_{\mathbf{PS_i}} = \frac{(X_i - \frac{uZ}{f}, Y_i - \frac{vZ}{f}, Z_i - Z)}{\sqrt{(\frac{uZ}{f}^2 + (\frac{vZ}{f}^2 + Z^2)}} \tag{13}$$

- and attenuation coefficients for direct signal β^D and backscatter β^B.

D. Photometric Consistency

The PS problem consists of images taken under multiple different light positions or intensities. Thus, the final estimated values for the unknown parameters must lead to a photometric (brightness) model that is as close to the observed brightness values in the set of images. The objective function, also called *photometric consistency*, minimizes the difference between the predicted and the observed brightness values as follows:

$$o(\beta_B, \beta_D, C(\lambda), \mathbf{n}, Z) = \sum_{n=1}^{N_I} (I_n - I'_n(\beta_B, \beta_D, C(\lambda), \mathbf{n}, Z)) \tag{14}$$

where N_I is the number of images taken in a specific camera pose with different lighting positions and intensities.

VI. Research Questions and Preliminary Results

As mentioned before, our main research questions are:

- **R1:** *Can images taken with the light position in place, but whose light intensity changes, be as informative as images taken with the light position changing?*
- **R2:** *Can an object be well reconstructed using a PS model, even if the images were taken while the robot was moving?*

R1 relates to hardware choices and the physical properties of light and its travel. PS models generally perform better when the distance between the camera and the light sources increase [24], as an increase in distance between the camera and light sources will decrease the amount of light rays that may scatter back to the camera – the backscatter component becomes less prominent. However, the size of the inexpensive BlueROV2, which we use, as described in the next subsection, is not large enough to make any meaningful impact on the distance between the camera and the light sources. Thus, backscatter will be significant. In addition, installing four individually controlled lights is a waste of cable port usage (i.e., it would take up 4 of the 14 penetrators of a typical end cap on a BlueROV2). To minimize the number of lights required, we will look into a PS model that considers how different light intensities may change the scale of brightness in the scene, but, more interestingly, it may also identify surface patches whose brightness is at saturation – the reflected light (brightness) will not increase with a stronger light intensity.

R2 leads to a major difference with respect to literature; unlike previous works where the robot was settled on the bottom [24], [25], our PS model is applied to an underwater robot that is *suspended* in the water. The underwater robot cannot always stay in place while it is suspended in water; it will slightly move due to external water forces or motor usage. We will look into the following: how robot movement can be approximately estimated by short-term visual feature tracking, how much robot movement corresponds to reconstruction error, and how robot movement needs to be incorporated into the PS optimization scheme.

A. Robot Setup

We used the BlueROV2, which was installed with the Sony IMX322LQJ-C camera[1] (originally included in the BlueROV2) and the Ping echosounder[2]. The camera has a resolution of 5 MP, a horizontal field of view (FOV) of $80°$, and a vertical FOV of $64°$. The echosounder has a maximum range of $30\,\mathrm{m}$ and a cone angle of $30°$.

There are two lights[3] installed on each side of the robot, whose intensities are controlled jointly and can be adjusted by increments of 10% (max: 100%). Each light supports at max 1,500 lumens and the beam angle in water is 135°. In future work, we will compare scene reconstruction results between different lighting setups: 2 individual lights, 2 sets of 2 lights, and 4 individual lights.

Light extrinsics will need to be calibrated, either approximated by hand or by a procedure using a white board or white sphere of known dimensions and characteristics.

*B. **R1:** Changing Light Intensity*

The goal here is to substitute an image of the scene that has a new light source position with an image with the same light position, but with a change in brightness intensity. This would free up the limited number of robot cable ports and minimize the number of light sources required to calibrate.

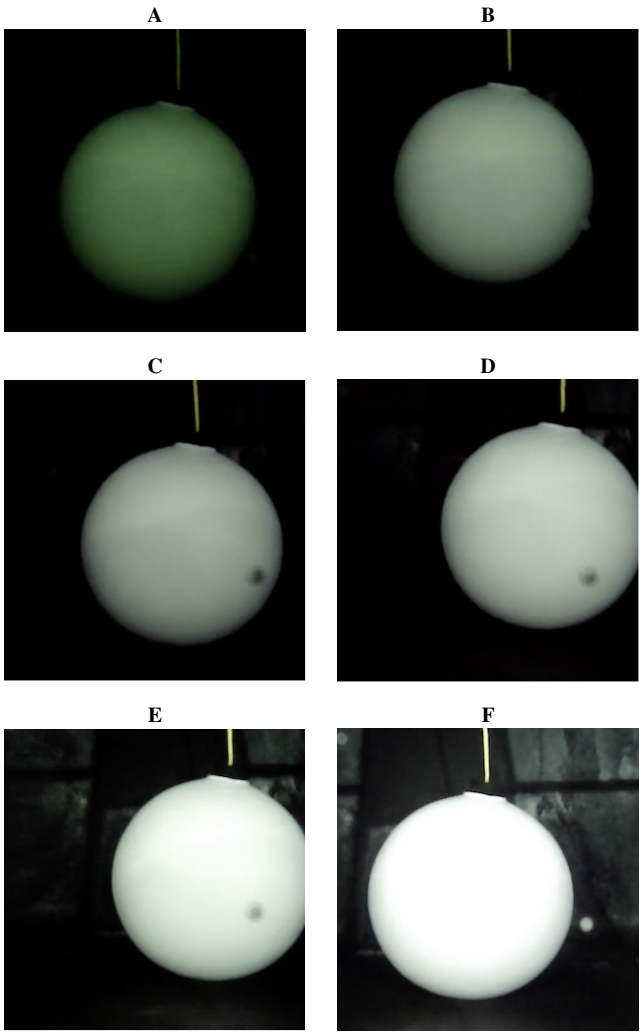

TABLE I: Series of images where the BlueROV2 changes light intensity from low to max brightness (**A**-**F**) while slightly moving. Note, the black speck in a few of the images is from a bubble on the clear dome end.

While the change in light intensity should not provide substantial photometric information, as it is simply changing the scale of brightness in the scene, we did observe brightness saturation. In the images, see Table I, there were areas on the white sphere that were saturated, identified as Local Diffuse Maxima (LDM) regions – image pixels whose local brightness intensity is maximum due to the surface points' normal vectors coinciding with the direction of the incident light rays. With an increase in light source brightness, the previous saturated pixels stayed constant, while the general saturated area increased. We believe that this observation of the LDM regions is key for integrating images with various light intensities into the PS model system. It can also be used as a constraint in case there are multiple optimal solutions

[1]https://www.bluerobotics.com/store/sensors-sonars-cameras/cameras/cam-usb-low-light-r1/

[2]https://bluerobotics.com/learn/ping-sonar-technical-guide/

[3]https://bluerobotics.com/store/thrusters/lights/lumen-sets-r2-rp/

to the objective function – this is similarly done by adding discrete ranges to parameters, such as the albedo [25].

In addition, we are able to collect multiple images under different light intensities – each light intensity output is measured from 0% to 100% with increments of 10%. Though, this might increase the number of images used in the photometric consistency optimization scheme from the traditional 3-4 images to about 10-20 images, depending on the number of controllable light sources and ambient light.

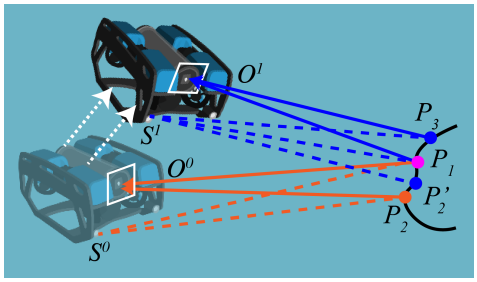

Fig. 4: Due to robot movement, feature points will need to be detected and matched across sequential images. The goal will be to use surface points that are detected as feature points in multiple images, such as P_1. It will be important to discard surface points from the framework that are incorrectly matched, such as P_2 and P_2'. Of course, some points will be non-usable, as they are only detected once, such as P_3.

C. ***R2:*** *Non-Stationary Robot*

As soon as the underwater robot moves, intentional or not, it rebukes the main constraint of the PS model that the camera must stay in place. The goal here is to either loosen the constraints in the PS optimization scheme, as the distances between robot and scene will slightly vary across images, or include camera/light movement estimations in the parameter estimation procedures.

We believe that with small robot movement and quick light position or intensity changes, camera/light movement should be minimal and easily calculated. Some possible approaches are to apply strip-down versions of Simultaneous Localization and Mapping (SLAM) systems, such as Monocular ORB-SLAM [9] and SVO [28]. For example, ORB features are known to be quick to calculate and match between sequential images, and ORB-SLAM was shown to work well in underwater applications [5], [29]. If we collect a short and informative sequence of images, all images should have an overwhelming number of overlapping feature points – similar to how a group of images correspond to a keyframe, due to their feature point similarity. The camera movement can be estimated if enough overlapping feature points among the images are matched, as illustrated in Fig. 4. Though, we will need to conduct experiments in simulation and in the pool to analyze how robot movement affects the scene reconstruction model and how much can estimated camera movement help mitigate the reconstruction error.

One issue that we are looking into is that estimating robot movement will lead to an ambiguous depth scale estimation. As mentioned before, the monocular depth estimation can be improved with an echosounder [5], but it might not be enough to produce an accurate scene reconstruction model. The reason is that the echosounder cone shape is not as tight as one would like, so distance values can be arbitrary.

Another issue is that there is a good chance that a small shift in robot movement and light change will cause major differences in the photometric consistency function for certain pixel/surface points. These points need to be thoroughly identified as deficient data points and be removed before they affect the overall PS model assumptions and estimations. This is a similar procedure done to remove shadow pixels.

Accordingly, with the robot movement as a constraint, an initial image will be marked as the original, while all of the other images will correspond to slight camera/light position changes. If the feature points are accurately matched across images – we should be able to calculate new distance values and light directions for these surface points. Of course, this approach might limit the number of surface points that can be used in the PS object function.

VII. Future Steps and Goals

In the future, we would like to extend our analysis to more thorough simulation experiments, pool/tank experiments with various water murkiness (i.e., milk-to-water proportions), and lake experiments. We plan to run comparison and validation experiments using other sensory suits, including a stereo camera and a mutlibeam sonar. Our final goal is to be able to reconstruct large underwater scenes, such as a shipwreck.

One of our next steps includes introducing another underwater robot into the system, such that it also has on-board lights. However, each robot may control and know about its own lights, but it will not directly know about the other robot's lights, except for estimated lighting intensity and position.

Overall, a solution to the photometric stereo model for non-stationary underwater robots is a key step to our final system for the next best view problem. Highly accurate scene reconstruction is not important for us – we are interested in a system that does not require a tight calibration or initial setup for estimating camera position relative to the scene, unlike systems such as ORB-SLAM [9], but also provides information of the scene (model) and points the robot to new directions of where to take more pictures, such that all key images collected are taken under consistent lighting – perfect for offline scene reconstruction.

VIII. Conclusion

We presented preliminary work on designing a photometric stereo framework for non-stationary underwater robots – to our upmost knowledge, a novel and unsolved problem. This paper addressed two concerns: can changing light intensity be as informative to the PS model as changing light position, and can robot movement be quickly estimated and integrated in the objective function for photometric consistency. We

showed that regions of brightness saturation (LDMs) across multiple different images with various light intensities may be advantageous in providing constraints when solving the PS objective function. In addition, quick camera movement can be calculated using ORB features or other strip-down SLAM and visual odometry approaches. Ultimately, a photometric stereo approach for non-stationary underwater robots have significant impacts – it allows for low-cost AUVs to accomplish scene reconstruction tasks with results on the same level as using high-end sonars, it provides a suitable solution to vision-based tasks under dynamic lighting conditions, and it opens up the opportunity for more exploration, monitoring, and inspection of various underwater structures.

ACKNOWLEDGMENT

We would like to thank the reviewers for their comments and suggestions, Philippos Mordohai for his invaluable advice and knowledge, and Devin Balkcom and Siddharth Agrawal for pool experimental help. This work is supported in part by the Dartmouth Burke Research Initiation Award and NSF CNS-1919647, 2024541.

REFERENCES

[1] F. S. Hover, R. M. Eustice, A. Kim, B. Englot, H. Johannsson, M. Kaess, and J. J. Leonard, "Advanced perception, navigation and planning for autonomous in-water ship hull inspection," *The International Journal of Robotics Research*, vol. 31, no. 12, pp. 1445–1464, 2012.

[2] O. Hoegh-Guldberg and J. F. Bruno, "The impact of climate change on the world's marine ecosystems," *Science*, vol. 328, no. 5985, 2010.

[3] M. Bryson, M. Johnson-Roberson, O. Pizarro, and S. B. Williams, "True color correction of autonomous underwater vehicle imagery," *J. Field Robot.*, vol. 33, no. 6, pp. 853–874, 2016.

[4] "The world's underwater cultural heritage," http://www.unesco.org/new/en/culture/themes/underwater-cultural-heritage/underwater-cultural-heritage/, Accessed 02/20/2020 2020.

[5] M. Roznere and A. Quattrini Li, "Underwater monocular image depth estimation using single-beam echosounder," in *Proc. IROS*, 2020.

[6] K. Richmond, C. Flesher, L. Lindzey, N. Tanner, and W. C. Stone, "SUNFISH®: A human-portable exploration AUV for complex 3D environments," in *MTS/IEEE OCEANS Charleston*, 2018, pp. 1–9.

[7] S. Rahman, A. Quattrini Li, and I. Rekleitis, "SVIn2: An Underwater SLAM System using Sonar, Visual, Inertial, and Depth Sensor," in *Proc. IROS*, 2019, pp. 1861–1868.

[8] H. Cho, B. Kim, and S.-C. Yu, "Auv-based underwater 3-d point cloud generation using acoustic lens-based multibeam sonar," vol. 43, no. 4, pp. 856–872, 2018.

[9] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, 2015.

[10] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *Int. J. Robot. Res.*, vol. 34, no. 3, pp. 314–334, 2015.

[11] T. Qin, P. Li, and S. Shen, "VINS-Mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, 2018.

[12] S. Hong, D. Chung, J. Kim, Y. Kim, A. Kim, and H. K. Yoon, "In-water visual ship hull inspection using a hover-capable underwater vehicle with stereo vision," *Journal of Field Robotics*, vol. 36, no. 3, pp. 531–546, 2019.

[13] R. J. Woodham, "Photometric method for determining surface orientation from multiple images," *Optical Engineering*, vol. 19, no. 1, pp. 139 – 144, 1980.

[14] J. Ackermann and M. Goesele, "A survey of photometric stereo techniques," vol. 9, no. 3–4, p. 149–254, 2015.

[15] O. Drbohlav and M. Chaniler, "Can two specular pixels calibrate photometric stereo?" in *Proc. ICCV*, vol. 2, 2005, pp. 1850–1857.

[16] M. G. Solonenko and C. D. Mobley, "Inherent optical properties of Jerlov water types," *Applied Optics*, vol. 54, pp. 5392–5401, 2015.

[17] M. Roznere and A. Quattrini Li, "Real-time model-based image color correction for underwater robots," in *Proc. IROS*, 2019.

[18] D. Akkaynak and T. Treibitz, "Sea-thru: A method for removing water from underwater images," in *Proc. CVPR*, 2019.

[19] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE Transactions on Image Processing*, vol. 29, pp. 4376–4389, 2020.

[20] D. Akkaynak and T. Treibitz, "A revised underwater image formation model," in *Proc. CVPR*, 2018, pp. 6723–6732.

[21] C. Tsiotsios, M. E. Angelopoulou, T.-K. Kim, and A. J. Davison, "Backscatter compensated photometric stereo with 3 sources," in *Proc. CVPR*, 2014.

[22] Z. Murez, T. Treibitz, R. Ramamoorthi, and D. Kriegman, "Photometric stereo in a scattering medium," in *Proc. ICCV*, 2015.

[23] A. Bodenmann, B. Thornton, and T. Ura, "Generation of high-resolution three-dimensional reconstructions of the seafloor in color using a single camera and structured light," *Journal of Field Robotics*, vol. 34, no. 5, pp. 833–851, 2017.

[24] C. Tsiotsios, T. Kim, A. Davison, and S. Narasimhan, "Model effectiveness prediction and system adaptation for photometric stereo in murky water," *Computer Vision and Image Understanding*, vol. 150, pp. 126–138, 2016.

[25] C. Tsiotsios, A. J. Davison, and T.-K. Kim, "Near-lighting photometric stereo for unknown scene distance and medium attenuation," *Image and Vision Computing*, vol. 57, pp. 44–57, 2017.

[26] T. Treibitz and Y. Y. Schechner, "Turbid scene enhancement using multi-directional illumination fusion," *IEEE Transactions on Image Processing*, vol. 21, no. 11, 2012.

[27] M. Sheinin and Y. Y. Schechner, "The next best underwater view," in *Proc. CVPR*, 2016.

[28] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO : Fast semi-direct monocular visual odometry," in *Proc. ICRA*. IEEE, 2014.

[29] A. Quattrini Li, A. Coskun, S. M. Doherty, S. Ghasemlou, A. S. Jagtap, M. Modasshir, S. Rahman, A. Singh, M. Xanthidis, J. M. O'Kane, and I. Rekleitis, "Experimental comparison of open source vision based state estimation algorithms," in *Proc. ISER*, 2016.