

MEASURING ARTIFICIAL REEFS USING A MULTI-CAMERA SYSTEM FOR UNMANNED UNDERWATER VEHICLES

R. Rofallski^{*1}, C. Tholen², P. Helmholz³, I. Parnum⁴, T. Luhmann¹

¹ Institute for Applied Photogrammetry and Geoinformatics, Jade University of Applied Sciences, Oldenburg, Germany, (Robin.Rofallski; Thomas.Luhmann)@jade-hs.de

² Dept. of Engineering Sciences, Jade University of Applied Sciences, Wilhelmshaven, Germany, Christoph.Tholen@jade-hs.de

³ Spatial Sciences, Curtin University, Bentley, Western Australia, Petra.Helmholz@curtin.edu.au

⁴ Centre for Marine Science and Technology, Curtin University, Bentley, Western Australia, I.Parnum@curtin.edu.au

Commission II, WG II/9

KEY WORDS: Underwater Photogrammetry, ROV, Multi-Camera System, Automated Image Masking, Transfer Learning, Convolutional Neural Network

ABSTRACT:

Artificial reefs provide an efficient way to improve marine life abundance in the oceans, including growth on the structure itself. Photogrammetric methods provide suitable tools to measure marine growth. This paper focusses on cubic reefs placed in Western Australia. The capturing platform featured a photogrammetric multi-sensor system for unmanned underwater vehicles attached to a low-cost vehicle BlueROV2. The multi-sensor system and its photogrammetric data captured was calibrated, adjusted and analyzed employing a structure-from-motion processing pipeline. Novel automated image masking techniques were developed and applied to the data to significantly reduce noise in the derived dense point clouds. Results show improvements of signal to noise ratio of more than 50 %, while maintaining a complete representation of the observed artificial reef.

1. INTRODUCTION

Artificial reefs are purpose-built submerged structures that enable marine life, such as corals, oysters or algae to settle on, and therefore attract fish to feed and shelter from predators. Usually built from concrete, steel or limestone, they are used around the world to create marine habitats and counter the global problem of decreasing marine life abundance (Carr and Hixon, 1997). Among others, several artificial reefs sized $3 \times 3 \times 3 \text{ m}^3$ (Figure 1) have been deployed off Western Australia's coast in order to avoid recreational overfishing and create new fishing spots around Western Australia (Florisson et al., 2018).



Figure 1. Artificial reef with BlueROV2 and scale bar mounted on a separate ROV

As part of the management and understanding the impact of artificial reefs, it is important to quantify marine biomass growing on these reefs. It is desirable to measure this growth using a non-destructive technique and, if possible, without

human proximity for reasons of safety and avoiding stress or damage to the marine flora and fauna. Thus, camera systems deployed on remotely operated vehicles (ROVs) are, provided good visibility, well suited to perform this task. The long term goal of this study is to develop and test a method for estimating the volume of marine biomass growing on underwater artificial reefs. This paper's contribution aims to develop a workflow to process photogrammetric data from two or three synchronized cameras in an underwater environment with spatial structures.

After introducing related work in the next section, section 3 introduces the study site, the multi-camera system as well as the calibration procedure used to calibrate the system. Then, section 4 introduces two novel image masking approaches – one based on image processing and one based on machine learning. After evaluating the results, the paper closes with a conclusion.

2. RELATED WORK

Image acquisition in or through water suffers from many quality degrading and geometry altering influences, compared to air. Firstly, the light from the object travels through multiple media (air, glass, water) and thus alters the ray path, rendering the pinhole camera model invalid. Strict modelling of the ray path was developed e.g. by Kotowski (1988), Maas (1995) and Jordt-Sedlazeck and Koch (2012), taking interfaces and refractive indices into account for photogrammetric analyses. However, several authors showed that for cameras facing almost perpendicular and very close to the interface, these effects can be compensated by radial and tangential distortion parameters (Kotowski, 1988; Shortis, 2015; Kahmen et al., 2019). Thus, standard photogrammetric and structure-from-motion processing may be used.

* Corresponding author

Secondly, optical degradation from wavelength dependent light absorption, chromatic aberration or dispersion reduce image quality. This results in images with low contrast, color cast, blur and haze (Wang et al., 2019). To overcome some of these effects, several image enhancement and image restoration algorithms, taking the specific characteristics of water into account, have been developed over the years. These account for the actual image formation model of underwater images or employ suitable image processing methods to enhance contrast, decrease color cast, etc. (e.g. LAB: Bianco et al., 2015; Sea-Thru: Akkaynak and Treibitz, 2019). Mangeruga et al. (2018) compared five state-of-the-art image enhancement algorithms for underwater photogrammetry and provided a metric to benchmarking these. It was concluded that for 3D reconstruction purposes, images enhanced with the LAB algorithm or the original images perform best. An up-to-date list of state-of-the-art image enhancement algorithms was recently compiled, reviewed and their implementations made openly accessible by Wang et al. (2019). It was concluded that none of the investigated algorithms is generic enough to create improvements under all visibility conditions that may occur in underwater imagery. Thus, algorithms have to be specifically evaluated for any given application.

Mapping underwater structures using photogrammetric techniques as the only acquisition method, or as part of a multi-sensor system, has been widely performed in tasks such as reef monitoring (Fabri et al., 2019), inspection of ship hulls (Kim and Eustice, 2013) or cave surveying (Nocerino et al., 2018). All these applications have the need to observe and robustly map submerged structures. During these processes, images are often taken at a predetermined frame rate and contain background areas without structure. Thus, an automated process, determining imaging areas or entire images that are unappealing for further analysis is desirable, especially in underwater environments with low contrast and visibility.

Segmenting regions of interest (ROI) and data labelling are common techniques in semantic analysis. The advent of machine learning in image processing increased interest to a wide extend. Convolutional neural networks (CNNs) are used widely for image segmentation due to their versatility and capability to deal with complexity. Underwater applications include recognizing underwater fauna (e.g. using DeepLab (Liu and Fang, 2020)) or obstacle avoidance (e.g. by combination with stereo matching (Arain et al., 2019)). Rizzini et al. (2015) developed an algorithm to extract man-made objects (pipes) from underwater imagery and performed image orientation with respect to the objects using a multi-feature object detection algorithm, based on silhouettes. Verhoeven (2018) compared several edge-based algorithms for segmenting sharp from blurry areas and mask these for further analysis. His findings were that the quality is hardly generic enough to transfer to randomly chosen real-world images, even though many authors claim state-of-the-art performance.

3. DATA ACQUISITION ON A MULTI-SENSOR SYSTEM

In this contribution, an artificial reef deployed off the coast near the Australian town of Dunsborough (33° 33.962' S 115° 9.980' E), about 200 km south of Perth, was observed by deploying a BlueROV2¹ equipped with a photogrammetric multi-sensor system. The ROV used in this study and the attached

multi-sensor system result from the interdisciplinary project EITAMS². Even though, several cubes and more reefs were measured during this campaign, this contribution shows data from one exemplary dive at this reef and all further processing steps were applied to this dataset.

3.1 Multi-sensor system for unmanned underwater vehicles

This study used a multi-camera system consisting of three industrial-grade cameras (two forward and one backward facing) in order to localize and map observing vertical structures as well as the seafloor (Rofallski and Luhmann, 2018). The cameras were equipped with a wide-angle lens ($f = 4.8$ mm). Further relevant camera parameters are summarized in Table 1. The backwards facing camera was attached to improve underwater positioning in sequential analyses, i.e. Simultaneous Localization and Mapping (SLAM).

Cameras	Basler Ace acA1920-48gc
Sensor size	9.2 mm × 5.8 mm
Resolution	1920 px × 1200 px
Max. frame rate	50 Hz
Pixel pitch	4.8 μm × 4.8 μm
Focal length	4.8 mm

Table 1. Camera and lens parameters

As the cameras did not offer an integrated data saving unit and due to the high frame rate needed for SLAM applications, the camera system was attached to three Ethernet cables, running into an Intel Core i7 laptop on the surface. Current developments in embedded systems show a significant increase in computing power, almost at the size of a credit card³ which will enable complex computations on the system without extra cables.

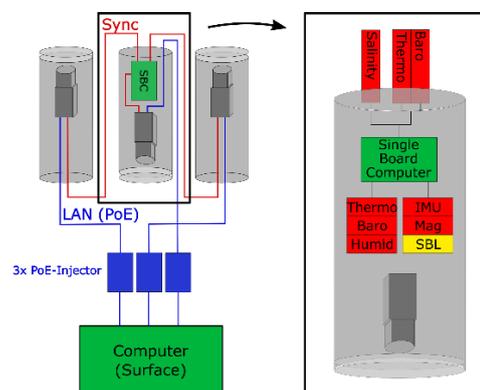


Figure 2. Schematic drawing of the used multi-sensor system

Apart from the cameras, the multi-sensor system was equipped with several further sensors to observe environmental parameters and enabling global localization (Figure 2). The sensors are summarized in Table 2. Since it is often not feasible to calibrate a camera system *in situ*, environmental sensors were placed on the multi-sensor system measuring water temperature, pressure and salinity; as these affect the refractive index of water as stated by Höhle (1971). Further sensors for measuring temperature, pressure and humidity were placed inside one of the camera housings in order to gain parameters for calculating a refractive index of the air in the tubes.

¹ www.bluerobotics.com

² www.eitams.de

³ www.aeon.com/en/p/pico-itx-boards-pico-whu4

Sensor	Data	Freq.
Camera	Images [px]	20 Hz
Short Baseline System	3D Position [m]	1 Hz
Inertial Measurement Unit	Lin. Acceleration [m/s ²] Ang. Velocity [°/s]	10 Hz
Electrical Conductivity	Salinity [ppt]	1 Hz
Barometer	Pressure [hPa]	10 Hz
Thermometer	Water Temperature [°C]	10 Hz
Internal Hygrometer	Relative Humidity [%]	1 Hz
Internal Thermometer	Temperature [°C]	1 Hz
Internal Barometer	Pressure [hPa]	1 Hz

Table 2. Sensors on the multi-sensor system and respective acquisition frequency

The system's central processing and synchronization unit was a Raspberry Pi Zero W single board computer⁴. It was equipped with a 1 GHz single core processor, 512 MB RAM and wireless connectivity. Thus, all data saved on the system was accessible without opening the tube reducing the risk of water intrusion, due to improper sealing. The computer generated a hardware trigger signal for the cameras at any frame rate up to 50 Hz. All other sensors sent their data at their own frequency to the Raspberry Pi time stamped internally and synchronized with the camera trigger signals. Apart from the camera data, all sensor data was stored on the single board computer.

The entire system was housed in three 3" cast acrylic tubes, fixed on an aluminium plate. All cameras viewed through 8.42 mm thick acrylic flat ports, which increases the effective principal distance approximately by factor 1.34 to 6.3 mm (Kahmen et al., 2019). Power supply was provided by a 14.8 V LiPo battery placed inside one of the tubes. Including power for lighting, the battery lasted for dives up to 60 minutes, exceeding an average battery charge of the BlueROV2.

3.2 Calibration

To obtain the relative orientation of the multi-camera system a customized calibration frame was built. Since the third camera's field of view did not overlap with the other two cameras', a spatial frame that is observable from the inside was constructed and attached with photogrammetric targets. During calibration the camera system was rotated around all axes and observing the predetermined markers used as ground control points. The relative orientation was then calibrated using self-calibration. Basically, it is desirable to calibrate both relative and interior orientation as closely as possible to the actual measurement, preferably *in situ*. However, due to practical reasons the relative orientation was pre-calibrated before the field data capture and the calibrated parameters assumed to remain constant.

The interior orientation parameters were calibrated on site before conducting the survey. Again, due to practical reasons the calibration took place just beneath the water surface off the boat using a flat calibration fixture with ring coded photogrammetric targets and not in 25 m depth at the reef location. The parameters of the interior orientation were determined employing distortion parameterization according to Brown (1971), i.e. principal distance, principal point, radial-symmetric and tangential distortion, and affinity and shear. Thus, no explicit modelling of

refractive effects was performed, enabling the use of standard SfM software such as Agisoft Metashape.

Camera	c	σ_c
1	-6.7108 mm	2.8 μ m
2	-6.7111 mm	2.1 μ m
3	-6.7407 mm	2.7 μ m

Table 3. Principle distance (c) and respective standard deviation of all three cameras, calibrated in water

To evaluate whether the calibrated principal distances (Table 3) could be assumed constant during the dive with respect to the environmental conditions, the additional sensor data (temperature, salinity and pressure, converted to depth) was used to calculate the refractive index, using the empirical formula by Höhle (1971).

$$n_w = 1.338 \cdot 4 \cdot 10^{-5} \cdot (486 + 0.003 \cdot d + 50 \cdot s - \lambda - t) \quad (1)$$

with ddepth [m]
ssalinity [%]
 λ wavelength of light [nm] - assumed 540 nm
ttemperature [°C]

This was performed for the calibration near the surface and for the observations at the reef. The resulting data is shown in Figure 3, where the red line shows the average refractive index of the calibration and the orange line shows the determined refractive index during data capture. On the same timeline, the depth of the ROV is depicted in blue.

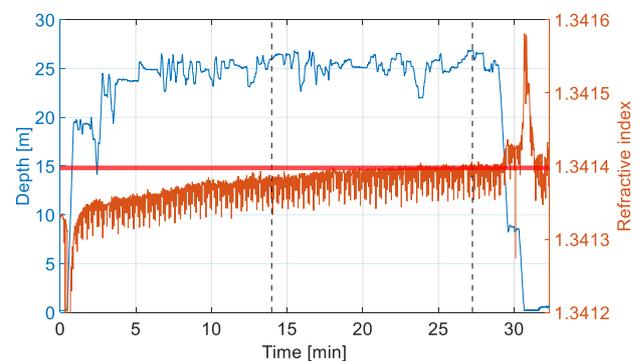


Figure 3. Refractive index (orange) with respect to the depth (blue). Red line indicates the refractive index at calibration. The dashed vertical lines indicate the time interval of analysis.

The data shows that apart from few outliers at beginning and end of the dive, the refractive index rose steadily. This may be a trend induced by a start-up curve of the sensor. Assuming error-free data, the refractive index ranged averagely between 1.34137 and 1.34140 during the image acquisition. This resulted in a maximum deviation of 0.00003 in refractive index compared to the average index during calibration. Applying Snell's law and assuming the entire ray path filled with water, this leads to an increase of 0.4 μ m in the principal distance. This is one order of magnitude lower than the standard deviation of the principal distance and thus negligible. The change in refractive index of air is also more than one order of magnitude lower than the one of water and its influence thus not discussed any further.

⁴ www.raspberrypi.org/products/raspberry-pi-zero-w/

3.3 Structure from motion

After obtaining the calibration values, the measurements were analyzed, using structure-from-motion processing. Here, the imagery acquired at a frame rate of 2 fps was evaluated. Over the acquisition time of approximately 13 minutes, 1582 images per camera were integrated into the bundle resulting in 4746 images to be aligned. To account for the predetermined relative orientation and account for scale, the three distances between the cameras (C1-C2, C1-C3, C2-C3) were introduced as scale constraints. Thus, three scale constraints per image triplet were introduced, resulting in 4746 constraints to the bundle adjustment.

A second ROV with an attached scale bar with photogrammetric targets was placed next to the artificial reef (Figure 1). The scale bar was observed and used to introduce scale into the bundle adjustment leading to a scaled sparse point cloud. The reference length of the scale bar was 825.222 mm and determined photogrammetrically prior to the field work.

For all further analysis, a single reference dataset was obtained, containing orientation data for the three cameras and a sparse point cloud. For a stereo reference, the third (backwards facing) camera was excluded and a stereo system created using the two forward facing cameras only. The orientation data of the original reference dataset and the stereo system remained largely constant and showed only minor deviations as seen in Table 4. All images were unprocessed and not masked at this point.

Cams	RMS	LME Ref. Scale	Aligned images	Number of points
2	1.18 px	1.89 mm	3094/3164	1,136,061
3	1.19 px	2.41 mm	4637/4746	1,576,634

Table 4. Statistics of the reference datasets of the two and three camera systems

The statistics are within the expected accuracy. In accordance with Shortis (2015) a relative accuracy of 0.1 % can be expected given optimum conditions in underwater photogrammetry. Maas (2015) states a loss of accuracy by factor 5 compared to equal datasets in air. This relates to an RMS reprojection error of subpixel accuracy, which is expectable from such a set of images with natural features (Luhmann et al., 2020) and is visible by the RMS values shown in Table 4.

4. IMAGE MASKING

Initial data analysis found that, although the sparse point cloud resembled the object well, the dense point cloud created by Agisoft Metashape’s dense matching algorithm had a very noisy output. It was concluded that this noise originated from large parts of the imagery being filled with unmatchable background areas. Since dense matching algorithms attempt to calculate a 3D coordinate for every pixel, the diffuse background is also taken into account leading to mismatches and consequently to a noisy point cloud. To overcome this issue, image masking was investigated, as this is a convenient way to eliminate points originating in low contrast areas.

Masking images is usually performed manually in cases with few images. However, when there is a large number of images, manual masking can be too time consuming to be feasible, and so this process is usually automated. One method to

automatically mask images for this application is by using a single-color background which can be removed automatically during analysis. Agisoft Metashape provides tools to perform both manual and automated image masking by either creating a polygon around the ROI in the software itself, or creating a binary image which fills valid areas with white (i.e. 8-bit value 255) and invalid areas with black (i.e. value 0) pixels. The latter can be generated by any image processing software; and is thus, highly customizable to the given application. Therefore, any image segmentation technique can be used to identify ROIs, and the binary images be imported into Metashape.

In a first attempt to mask ROIs, two image segmentation techniques were tried, one using an edge-based method (Chen et al., 2016) and the other a patch-based method (Golestaneh and Karam, 2017). Both methods are state-of-the-art and openly accessible algorithms for masking out-of-focus areas. An exemplary image from our dataset (left column in Figure 4) was processed using both methods. The resulting focus map (middle column in Figure 4) was then binarized using Otsu’s adaptive thresholding method (Otsu, 1979) shown in the right column of Figure 4.

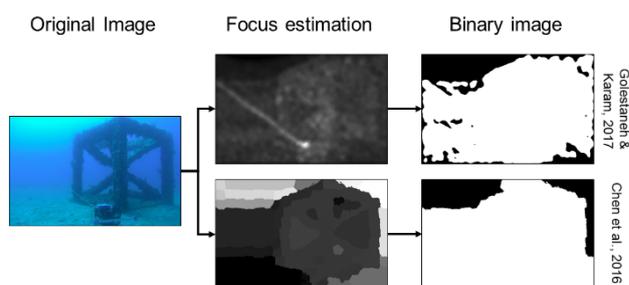


Figure 4. Exemplary image of artificial reef used with two state-of-the-art focus estimation algorithms, and the results binarized for image masking using Otsu’s thresholding method

The results show only small parts of the background being correctly masked, while still allowing a major part of the image for matching. While method 2 seemed to be able to at least segment the reef and seafloor from the rest of the imagery, it did not create unambiguous classes that enable segmentation of the background by thresholding or a bandpass selection. This does not reduce the scatter in the dense point cloud. Successful masking should follow the object’s edge tightly and allow only ROIs on the structure. Furthermore, these algorithms hardly deal with images without any structure element, identifying simply the sharpest areas in an image. Hence, the authors developed two novel methods able to distinguish the artificial reefs and seafloor from the water column more precisely and furthermore being able to identify images without matchable objects so that they can be excluded entirely from further processing. In the following, the discussed approaches are referred to, as follows:

1. Image processing (IP) approach
2. Machine learning (ML) approach

4.1 Image processing approach

The image processing (IP) approach used a combination of standard image processing procedures in a workflow shown in Figure 5. First, dark areas in the blue channel were identified, which best distinguishes the reef from the background. Secondly, noise was reduced by applying a low pass (Gaussian blur) filter to the image. The filtered image was then classified using two methods: 1) Otsu’s adaptive threshold method (Otsu, 1979) partitioned the image into dark and light areas, where dark areas

were assumed to correlate with the area of interest (in this case the artificial reef); and 2), the Canny edge detection algorithm (Canny, 1986) to identify high frequency areas, such as structures on the seafloor. The results of these two classifications were combined by a closing operation.

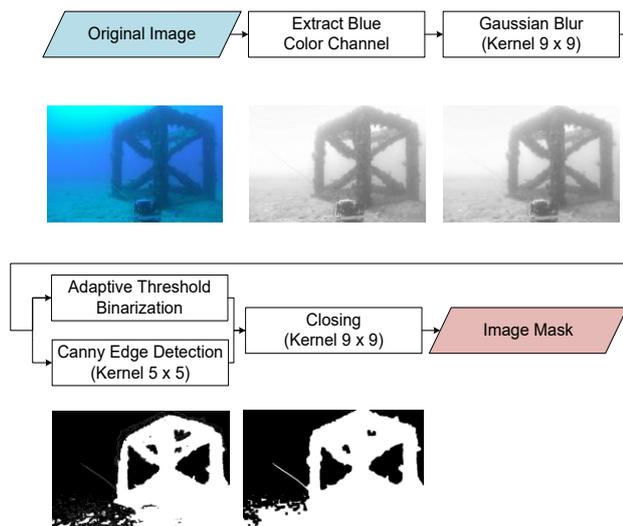


Figure 5. Workflow of the image processing approach

Like many image segmentation routines, there was a trade-off between accuracy and completeness of the mask. If the kernel for closing was too large, cut-out areas within the artificial reefs were closed as well and thus too much area masked out. On the other hand, if the kernel was too small, only a very small neighbourhood of an edge was used and thus too many features on the seafloor were omitted. However, since feature detectors usually search for high frequency areas, it was more desirable to choose the kernel size too small than too large, as distinct edges are more likely to be chosen as a key point by the feature detector in the SfM processing. The kernel sizes chosen were based on the visual performance for the application and are not further investigated, as they were outside the scope of this study. The image processing approach was applied to all captured images classifying them into structure and background.

4.2 Machine learning approach

As an alternative to the image processing (IP) approach, a machine learning approach (ML) was also implemented. The procedure for the machine learning approach consisted of two parts. Firstly, a convolutional neural network (CNN) was trained by adapting knowledge from an existing network and applying it to solve the given problem of detecting static structures. This approach is called transfer learning (Long et al., 2015). Secondly, the trained CNN was used to segment images recursively.

4.2.1 Training: The machine learning approach was based on ResNet-50, a CNN with 50 layers designed for image processing applications (He et al., 2016). Based on the pre-trained layers and weights of this network, the CNN was finely-tuned by changing the output layers to two classes: static and non-static structure. Afterwards, the weights of the other layers were fixed and the weights and biases of the newly added layers were trained.

For training, 100 images of the dataset were recursively split in to a quadtree five times to a patch size of 62×39 px. The patches were not chosen any smaller, as it became increasingly complicated to identify the image content with smaller patch sizes. For each of the split layers classified, there were about the

same number of image patches showing areas with a static structure, i.e. reef and seafloor (class 1) and areas without a static structure, i.e. background, fish and tethers (class 0). The training dataset consisted of 5816 images, of which 2977 images were labelled as class 0 and 2839 images labelled as class 1. The dataset was randomly split into 80 % of the images used for training and 20 % for validation. The training took 80 minutes on a NVIDIA GeForce GTX 1080 Ti GPU. The training validation accuracy of the CNN was 91.5 %.

4.2.2 Segmentation: The segmentation was based on a quadtree structure by split-and-merge segmentation. Thus, the imagery was recursively split in quadrants n times and each of the quadrants labelled, according to the findings of the trained network. If one quadrant did not contain structure (i.e. class 0), no further splitting was performed and the entire quadrant labelled accordingly. After all quadrants were classified they were merged back to the full image, resulting in an image mask with a resolution corresponding to the size of the lowest quadtree level. In this case, the maximum level n was 7, corresponding to a mask resolution of 15×10 px. Figure 6 shows pseudocode for the algorithm described above.

```

image = load all images
num = number of images
n = 7
for i = 1 : num do:
    label = classify(image(i))
    if label == 1 do:
        image(i) = split_predict(image(i), n)
    else
        image(i) = black
    end if
    save image(i)
end for

function image = split_predict(image, n)
    if n == 0 do:
        image = white
        return image
    end if
    quarter = split image in quarters
    for j = 1 : 4 do:
        label = classify(quarter(j))
        if label == 1 do:
            quarter(j) = split_predict(quarter(j), n-1)
        else
            quarter(j) = black
        end if
    end for
    return image
end function

```

Figure 6. Pseudocode for the image segmentation process

5. EVALUATION OF AUTOMATED IMAGE MASKING FOR DENSE MATCHING

To assess the improvement of the created dense point clouds based on the images preprocessed with the different image segmentation algorithms, the reference datasets as introduced in section 3.3 were used. The first reference dataset is a reconstruction using images from all three available cameras, the second reference dataset was created using the forward-looking cameras forming a stereo system only. Using the same orientation data ensured that the alignment parameters do not affect the dense image matching and an objective comparison between the generated point clouds can be achieved. Furthermore, the area of the investigated point cloud was equal over all datasets. Thus, the absolute points numbers refer to the same area of interest and are comparable.

Since ground truth data was not available from the reefs, an independent measure of quality was not possible. Instead, a reference dataset of the reef was created by manually improving

the visually least noisy dataset produced. Overall, 11.7 % of the total point cloud was removed (noise and obviously erroneous points). Subsequently, a mesh was created, using Cloud Compare's *Poisson Surface Reconstruction* method (Kazhdan et al., 2006) with an octree level 6. The resulting reference mesh had visibly less noise than the original point cloud. All processed point clouds were then compared against this reference mesh.

For evaluation, the following statistics are used to obtain performance metrics, mostly following recommendations by Mangeruga et al. (2018) but with some different interpretations of those:

Number of 3D points (NB 3D): The number of points represents a measure of points that could be matched. The evaluated region of the point cloud was equal for all datasets and thus comparable amongst each other. However, since the basic problem was high noise and too many points being matched, a high number does not necessarily represent a high quality.

Mean cloud to mesh distance (C2M): The mean distance from the point cloud to the reference mesh represents a measure of quality. A mesh, rather than the point cloud was used, as this posed a more generalized representation of the reef surface.

Signal to noise ratio (SNR): SNR is the ratio of the points within 5 cm (i.e. $2 \times \text{GSD}$) from the reference mesh to the total amount of points computed. This is a measure of the amount of noise present in each point cloud.

Surface density (SD): The surface density was estimated aggregating neighbouring points in a radius R and extrapolating this number to 1 m^2 . The assumption is that for noisy point clouds, the surface density will average be lower than for point clouds with less noise. The radius R chosen was 5 cm, i.e. $2 \times \text{GSD}$.

Integrity (I): After eliminating all points with a C2M distance of more than 5 cm the integrity is subsampling the remaining points equally spaced over a grid of 5 cm. The resulting amount was then used to calculate the ratio to the rastered reference mesh at equal resolution. Since only areas containing point cloud data will have a corresponding subsampled point, it is assumed that this represents a suitable measure of integrity.

Ratio of masked to unmasked pixels (M): The number indicates the percentage of pixels that were excluded from the dense image matching. Thus, without masking, the ratio is 0 %.

5.1 Influence of masking

First, the influence of masking images on the quality of the resulting point clouds was investigated. In addition, the camera configuration (all three cameras vs. stereo-camera pair) was changed to investigate their influence on the dense matching. Thus, six datasets (Table 5) were evaluated against the reference dataset employing the aforementioned metrics.

	Stereo pair	3 cameras
No masking	NM2	NM3
Image Processing (IP)	IP2	IP3
Machine Learning (ML)	ML2	ML3

Table 5. Overview of the processed datasets which are compared to the reference mesh

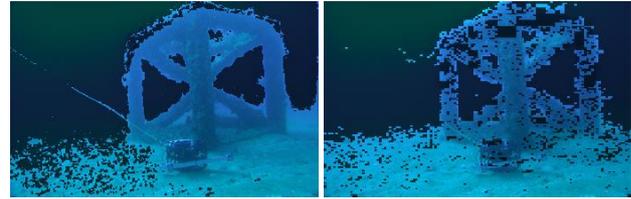


Figure 7. Mask overlay on an exemplary image of the dataset. Left: IP approach; Right: ML approach

Figure 7 shows the results of the two masking approaches (IP and ML) applied to an exemplary image of the dataset. Both methods recognized the structure of the reef and worked in the intended way. The ML approach masked more parts of the reef with a more speckle-like pattern, while the IP approach followed the edges more tightly. However, the seafloor was masked rather coarsely. Furthermore, lower cut-out areas, the cable running diagonally through the image and parts of the background on the right-hand side were not masked correctly with the IP method. In contrast, these parts were mostly covered by the ML method, though the cable in front of the seafloor was not masked, neither. Furthermore, parts like moving fish, tethers, etc. could be masked out individually by the ML method, which poses a complexity that can hardly be accounted for by standard image processing methods because of the similar structure in the frequency domain.

The processing time for single image masking differed significantly. For the IP method, a single image took about 0.1 seconds to process and to write the resulting binary JPEG file. On the other hand, the ML approach also processing and writing binary JPEG files took in average four minutes per image on the same machine. Though this can still be improved by parallelization and code optimization, the ML method takes significantly longer, while the IP method may be integrated to online systems, such as SLAM. Furthermore, data labelling for the ML approach took about two hours of manual work, in order to prepare the training data.

Data	NB 3D [10^3 pts]	C2M [m]	SD [pt/m ²]	SNR [%]	I [%]	M [%]
NM2	1197.9	0.325	2580.5	18.7	100	0
NM3	1191.4	0.323	2579.0	18.7	99	0
IP2	717.6	0.153	2853.2	32.1	95	56
IP3	717.5	0.152	2856.1	32.1	94	54
ML2	602.0	0.078	3665.7	48.5	96	58
ML3	604.6	0.078	3644.9	48.3	96	63

Table 6. Results of masking for dense image matching. NM: No masking; IP: Image processing, ML: Machine learning. Suffix number indicates number of used cameras (2 or 3).

Table 6 summarizes the results of this first investigation. It can be observed that no significant differences were found between using two and three cameras. This is likely because the reef was mainly observed by the two forward facing cameras, hence the third camera hardly influenced the processing results. However, the two masking procedures (IP and ML) reduced the mean C2M distance by up to a factor of 4. Furthermore, the number of matched 3D points (NB 3D) decreased with the algorithms down to almost 50 % compared to the unmasked dataset. This correlated with the amount of masked pixels (M), which was about the same order of magnitude. The surface density (SD) was higher for the masking methods which points towards more points being in close neighbourhood. This indicates a better and

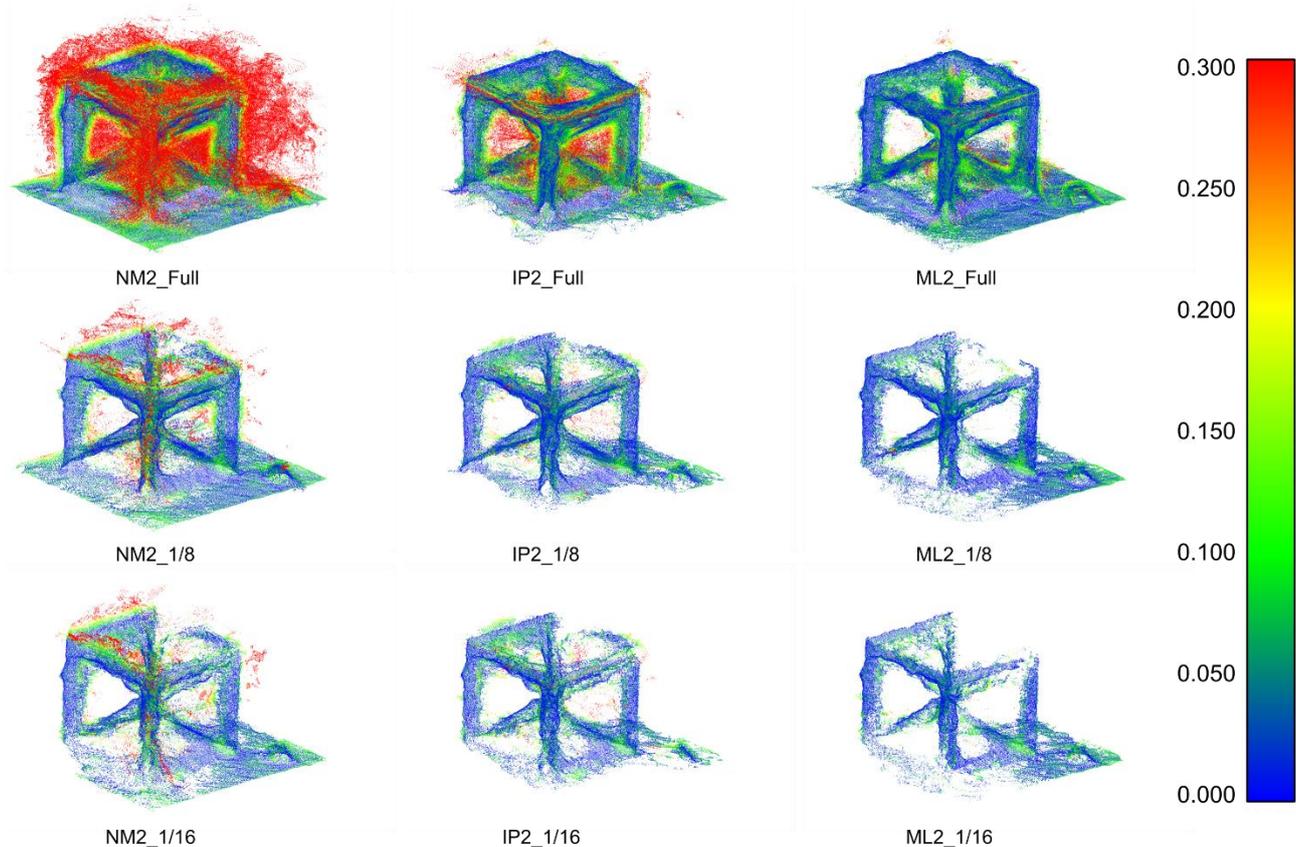


Figure 8. Cloud to mesh distances of some representative datasets. Datasets are ordered columnwise: Left: No masking (NM), Middle: Image processing method (IP); Right: Machine learning approach (ML). The rows indicate different numbers of images used for matching. All distances > 0.3 m are also labelled red.

less noisy representation of the structure (i.e. higher SNR). Interestingly, the integrity (I) remains almost constantly high at around 100 % for all datasets.

The upper row of Figure 8 shows the corresponding point clouds for the datasets from two cameras with C2M distances. The datasets with three cameras show comparable results. It is obvious that the unmasked point cloud suffered from high noise, though covering the entire area, including seafloor in about the same density. The IP method showed considerably less noise, whereas parts of the seafloor on the left of the reef were left out. However, cut-out areas inside the reef remained noisy. The ML approach visually shows the best results, which is backed by the values presented in Table 6. Here, cut-out areas were mostly free of noise and almost the entire seafloor was mapped, as well.

5.2 Influence of image enhancement

The influence of image enhancement on the matching results was investigated. Mangeruga et al. (2018) compared various image enhancement algorithms and concluded that the LAB image enhancement algorithm and unprocessed images performed best in underwater applications, accordingly to their benchmark. This is concurrent with our findings from other algorithms that were evaluated in their contribution. Other openly accessible algorithms (ACE, CLAHE, NLD, SP) introduced higher noise in the images, color artefacts or no visible contrast increase. Thus, this study limited its comparison to the LAB method.

Next, as the results from two and three cameras showed almost identical results, only the datasets utilizing two cameras are discussed further. Findings are transferrable to the dataset with

three cameras, unless stated otherwise. Figure 9 shows an exemplary unprocessed image next to a LAB-processed image. The blue color cast disappeared and visually a higher contrast is present in the image.



Figure 9. Original image (left) and LAB enhanced image (right)

Maintaining equal orientation data, the three datasets (NM, IP and ML) were processed with LAB enhanced images. Comparing the processed dense point clouds with and without enhancement show very similar results, as visible in Table 7. For comparison, the respective results without image enhancement from Table 6 are shown again, as well.

No metric varied significantly from the unprocessed imagery. For the unmasked dataset, slightly worse results were achieved with higher C2M distances (0.325 m vs. 0.338 m) and lower SNR (18.7 % vs. 18.5 %). The two masking approaches showed slightly better results with lower C2M distances of few millimeters and a marginally higher SNR (32.1 % vs. 32.9 % and 48.5 % vs. 48.6 %). This correlated with slightly fewer matched 3D points, whereas the unmasked dataset produced slightly more 3D points. However, these results show no measurable improvement over the unprocessed data, which is why unprocessed images are used for further investigations.

Data	NB 3D [10 ³ pts]	C2M [m]	SD [pt/m ²]	SNR [%]	I [%]	M [%]
NM2	1197.9	0.325	2580.5	18.7	100	0
NM2 LAB	1213.7	0.338	2579.5	18.5	100	0
IP2	717.6	0.153	2853.2	32.1	95	56
IP2 LAB	697.8	0.147	2859.1	32.9	95	56
ML2	602.0	0.078	3665.7	48.5	96	58
ML2 LAB	599.6	0.077	3672.6	48.6	96	58

Table 7. Results of image processing with LAB image enhancement compared to unprocessed images

5.3 Influence of reduced image numbers

The amount of images was reduced to investigate performance with lower object coverage. Especially considering the long processing time for masking images using the ML approach, it is of interest to investigate whether the entire object can still be reconstructed when fewer images are used. Three datasets were created by reducing the dataset with two cameras by factors 2 (to 1582 images), 8 (to 395 images) and 16 (to 197 images) while relying on the same orientation parameters and area covered by the reconstruction. Table 8 depicts the results of these datasets separated by bars for each approach. The point clouds with C2M distances corresponding to the 1/2, 1/8 and 1/16 dataset are shown in the middle and bottom row of Figure 8.

For all methods, the number of calculated points decreased with decreasing number of images to about one quarter compared to the full dataset. The ML approach constantly had the lowest number of matched points, though the quality of these point clouds were the highest compared to the other approaches. In contrast, the SNR level in the NM dataset (no masking) with 1/16 of the images was even with the ML method with all images but visible and significant outliers were still present in the unmasked dataset. However, the masking approaches (IP and ML) nearly eliminated the noise, also in the full dataset with all images.

Data	NB 3D [10 ³ pts]	C2M [m]	SD [pt/m ²]	SNR [%]	I [%]	M [%]
NM2 1/2	783.8	0.245	2585.0	26.3	90	0
NM2 1/8	342.7	0.134	2627.5	45.8	64	0
NM2 1/16	247.9	0.106	2593.9	50.0	50	0
IP2 1/2	506.9	0.116	2867.4	40.9	82	56
IP2 1/8	240.2	0.070	2870.5	59.1	53	55
IP2 1/16	183.0	0.067	2820.0	59.5	40	55
ML2 1/2	430.2	0.062	3663.7	57.9	80	58
ML2 1/8	207.4	0.043	3767.8	70.5	44	58
ML2 1/16	152.8	0.044	3488.8	67.8	33	58

Table 8. Results of reduced datasets

Based on the measures presented in Table 8, the integrity (I) was reduced with every reduction of the number of images, independent of the method used. This proved to be especially true for the ML approach, where the lowest integrity (I) was observed down to one third. It becomes visible in the point cloud, where one edge of the cubic object is almost entirely missing as well as large parts of the seafloor. The IP approach showed a less crucial reduction of integrity (I) with reduced images, but on the other hand still had significantly lower SNR values compared to the ML approach. The surface density (SD) value remained almost constant between the three approaches indicating that the areas that could be mapped had a similar density, regardless of the amount of images. Even with the strongly reduced integrity (I) of the ML approach, the SD value remained high.

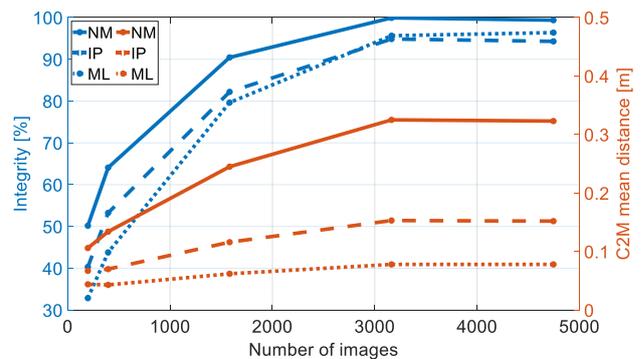


Figure 10. Different image masking methods as a function of the number of images used for dense matching. Blue lines indicate integrity; orange lines indicate mean C2M distance.

The C2M distance was reduced to less than a third for the unmasked images and more than halved for the IP dataset. In contrast, the C2M distance of the ML approach was only reduced by factor 1.75. Thus, the quality depends less on the amount of images used. Figure 10 shows integrity and C2M mean distances as a function of the number of images. It can be observed that the mean C2M distance remained constantly low for all numbers of images with the ML method, whereas the IP method and no masking showed a significant increase in C2M mean distance with increasing image numbers. Furthermore, for all datasets both measures increased with increasing number of images. The integrity values (I) basically followed the same trend as the C2M measure but with larger differences between the approaches. The ML approach had the lowest integrity (I) value except when using the full dataset (all images available). Then, the IP showed the smallest integrity (I) value. Otherwise, the integrity (I) values of the IP approach are only marginal larger than the integrity (I) value of the ML approach. The datasets using the unmasked images had always the highest integrity (I) values.

5.4 Discussion of the masking methods

The results show that unmasked imagery is not suitable for the processing of our datasets and it is expected that many other applications in underwater photogrammetry suffer from similar issues. Decreasing image numbers helped reduce the noise using unmasked images. Nevertheless, even with 1/16th of the images, the noise using unmasked images was still higher than using the full dataset of masked images.

Furthermore, it has been shown that both masking approaches (based on image processing (IP) and machine learning (ML)) provide better results than the unmasked data. Accuracy metrics such as C2M distance and SNR are improved by factors 4 and 3, respectively. This, however comes at the cost of needing more

images for a complete point cloud as integrity measures decrease using these methods. This may result from parts being masked incorrectly, as it can be observed in Figure 7. In order to improve the integrity measures, a training approach with more data would be desirable. However, most of these areas are likely to be areas of low contrast in background areas for which it is unlikely to find matches, even when no masking is used.

The improvements through masking come at the price of increased processing times. While the IP approach is capable of improving results and processing is possible in near real-time, the ML approach is computationally very expensive. Both, training and segmentation times must be taken into account when using the ML method, whereas the training is performed only once.

The ML approach is capable of distinguishing entire images without matchable areas and excluding these from the workflow. In the full dataset with three cameras, a total number of 547 images were filtered using this method. This can be used to select and reduce the number of images for SfM, to prevent processing datasets with too many images.

6. CONCLUSION

This contribution provides a workflow from calibration and acquisition to the analysis of photogrammetric data from artificial reefs. Two novel automated image masking processes are provided and their performance evaluated based on real datasets. It has been shown that image masking is a very useful tool for underwater imagery that suffers from low contrast and major parts being filled with unmatchable areas, creating noise in a point cloud.

Further investigations are necessary to estimate the transferability of the masking approaches. Especially, the image processing approach is very specifically adjusted to the used datasets, whereas the machine learning approach may potentially be easier to transfer onto other underwater datasets. However, in order to improve performance, a broader training dataset is necessary, covering a wide spectrum of visibility conditions and objects.

The ML approach may also be used for further improvements on the image selection when only a subset of a big dataset has to be analyzed. Since the approach is capable of estimating the image content and the amount covered by an image, the ML approach may be used to select and therefore decrease the number of images to be processed by the structure-from-motion approach based on a score. However, the IP approach (just as other methods based on fast image processing tools) may be integrated into real-time or online processing systems such as a SLAM application and thus improve orientation and mapping of a robot in an unknown environment.

The entire masking process was performed on unprocessed images. Since the LAB imagery seems to improve contrast, training the masks on these images may be a way to further improve the approaches. Furthermore, the influence of masking on the orientation process could not be reliably investigated in this study. The used software did not deterministically provide reproducible results with the same data. Also, performing a bundle adjustment lead to some sort of an iterative behaviour in itself, meaning the results kept improving by performing the same action several times in a row. To thoroughly investigate orientation, these effects have to be clarified beforehand and be reproducible.

Unfortunately, CAD data of the reefs was not made available by the manufacturer of the reefs. Thus, no volume of marine growth could be estimated in this study. It is hoped to be able to obtain the reef CAD data in the future to finalize the last step and be able to provide marine biologists with this data. The estimated volume of the reference mesh from this study is 8.0 m³.

ACKNOWLEDGEMENTS

The authors would like to thank Malcolm Perry and David Belton (Curtin University) for their practical help and for sharing their valuable experience in the preparation and execution of this study. Further thanks to Kim Royce from Octopus Garden, Bunbury for providing a vessel for our experiments.

This work was funded by Volkswagen Foundation (ZN3253) and the research fund of Jade University.

REFERENCES

- Akkaynak, D., Treibitz, T., 2019. Sea-Thru. A Method for Removing Water From Underwater Images. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1682–1691.
- Arain, B., McCool, C., Rigby, P., Cagara, D., Dunbabin, M., 2019. Improving Underwater Obstacle Detection using Semantic Image Segmentation. 2019 International Conference on Robotics and Automation (ICRA). IEEE, pp. 9271–9277.
- Bianco, G., Muzzupappa, M., Bruno, F., Garcia, R., Neumann, L., 2015. A new color correction method for underwater imaging. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XL-5/W5, pp. 25–32.
- Brown, D.C., 1971. Close-Range Camera Calibration. *Photogrammetric Engineering* 37 (8), pp. 855–866.
- Canny, J., 1986. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-8 (6), pp. 679–698.
- Carr, M.H., Hixon, M.A., 1997. Artificial Reefs. The Importance of Comparisons with Natural Reefs. *Fisheries* 22 (4), pp. 28–33.
- Chen, D.-J., Chen, H.-T., Chang, L.-W., 2016. Fast defocus map estimation. *IEEE International Conference on Image Processing, IEEE, Piscataway, NJ*, pp. 3962–3966.
- Fabri, M.-C., Vinha, B., Allais, A.-G., Bouhier, M.-E., Dugornay, O., Gaillot, A., Arnaubec, A., 2019. Evaluating the ecological status of cold-water coral habitats using non-invasive methods. An example from Cassidaigne canyon, northwestern Mediterranean Sea. *Progress in Oceanography* 178, pp. 102172.
- Florisson, J.H., Tweedley, J.R., Walker, T.H.E., Chaplin, J.A., 2018. Reef vision. A citizen science program for monitoring the fish faunas of artificial reefs. *Fisheries Research* 206, pp. 296–308.
- Golestaneh, S.A., Karam, L.J., 2017. Spatially-Varying Blur Detection Based on Multiscale Fused and Sorted Transform Coefficients of Gradient Magnitudes. 30th IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Piscataway, NJ, pp. 596–605.

- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. 29th IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Piscataway, NJ, pp. 770–778.
- Höhle, J., 1971. Zur Theorie und Praxis der Unterwasser-Photogrammetrie. Deutsche Geodätische Kommission, Reihe C, Heft Nr. 163, München.
- Jordt-Sedlazeck, A., Koch, R., 2012. Refractive Calibration of Underwater Cameras. In: Fitzgibbon A., et al. (Eds.), Computer Vision – ECCV 2012. Springer, Berlin, Heidelberg, pp. 846–859.
- Kahmen, O., Rofallski, R., Conen, N., Luhmann, T., 2019. On scale definition within calibration of multi-camera systems in multimedia photogrammetry. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-2/W10, pp. 93–100.
- Kazhdan, M., Bolitho, M., Hoppe, H., 2006. Poisson Surface Reconstruction. In: Polthier, K., Sheffer, A. (Eds.), Proceedings of the fourth Eurographics symposium on Geometry processing. Eurographics Association, Aire-la-Ville, pp. 61–70.
- Kim, A., Eustice, R.M., 2013. Real-Time Visual SLAM for Autonomous Underwater Hull Inspection Using Visual Saliency. IEEE Transactions on Robotics 29 (3), pp. 719–733.
- Kotowski, R., 1988. Phototriangulation in Multi-Media Photogrammetry. International Archives of Photogrammetry and Remote Sensing (Vol. XXVII), pp. 324–334.
- Liu, F., Fang, M., 2020. Semantic Segmentation of Underwater Images Based on Improved Deeplab. Journal of Marine Science and Engineering 8 (3), pp. 188.
- Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Piscataway, NJ, pp. 3431–3440.
- Luhmann, T., Robson, S., Kyle, S., Boehm, J., 2020. Close-range photogrammetry and 3D imaging, 3rd edition. De Gruyter, Berlin, Boston.
- Maas, H.-G., 1995. New developments in Multimedia Photogrammetry. In: Grün, A., Kahmen, H. (Eds.), Optical 3D Measurement Techniques III. Wichmann, Karlsruhe, pp. 362–372.
- Maas, H.-G., 2015. On the Accuracy Potential in Underwater/Multimedia Photogrammetry. Sensors (Basel, Switzerland) 15 (8), pp. 18140–18152.
- Mangeruga, M., Bruno, F., Cozza, M., Agrafiotis, P., Skarlatos, D., 2018. Guidelines for Underwater Image Enhancement Based on Benchmarking of Different Methods. Remote Sensing 10 (10), pp. 1652–1678.
- Nocerino, E., Nawaf, M.M., Saccone, M., Ellefi, M.B., Pasquet, J., Royer, J.-P., Drap, P., 2018. Multi-camera system calibration of a low-cost remotely operated vehicle for underwater cave exploration. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-1, pp. 329–337.
- Otsu, N., 1979. A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics 9 (1), pp. 62–66.
- Rizzini, D.L., Kallasi, F., Oleari, F., Caselli, S., 2015. Investigation of Vision-Based Underwater Object Detection with Multiple Datasets. International Journal of Advanced Robotic Systems 12 (6), pp. 77.
- Rofallski, R., Luhmann, T., 2018. Fusion von Sensoren mit optischer 3D-Messtechnik zur Positionierung von Unterwasserfahrzeugen. Hydrographie 2018. Trend zu unbemannten Messsystemen., DVW-Schriftenreihe, Band 91. Wißner-Verlag, Augsburg, pp. 223–234.
- Shortis, M., 2015. Calibration Techniques for Accurate Measurements by Underwater Camera Systems. Sensors (Basel, Switzerland) 15 (12), pp. 30810–30826.
- Verhoeven, G.J., 2018. Focusing on out-of-focus. Assessing defocus estimation algorithms for the benefit of automated image masking. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-2, pp. 1149–1156.
- Wang, Y., Song, W., Fortino, G., Qi, L.-Z., Zhang, W., Liotta, A., 2019. An Experimental-Based Review of Image Enhancement and Image Restoration Methods for Underwater Imaging. IEEE Access 7, pp. 140233–140251.