IEEE Robotics and Automation Letters (RAL) paper presented at the
2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Matching Color Aerial Images and Underwater Sonar Images using Deep Learning for Underwater Localization

Matheus M. Dos Santos[1], Giovanni G. De Giacomo[1], Paulo L. J. Drews-Jr[1], Silvia S. C. Botelho[1]

***Abstract*— Underwater localization is a challenging task due to the lack of a Global Positioning System (GPS). However, the capability to match georeferenced aerial images and acoustic data can help with this task. Autonomous hybrid aerial and underwater vehicles also demand a new localization method capable of combining the perception from both environments. This study proposes a cross-domain and cross-view image matching, using a color aerial image and an underwater acoustic image to identify if these images are captured in the same place. The method is designed to match images acquired in partially structured environments with shared features, such as harbors and marinas. Our pipeline combines traditional image processing methods and deep neural network techniques. Real-world datasets from multiple regions are used to validate our work, obtaining a matching precision of up to 80%.**

I. INTRODUCTION

Unmanned vehicle systems have become popular due to their ability to adapt to complex environments and to perform autonomous tasks. Underwater environments are challenging, dangerous, and remain largely unexplored. Autonomous Underwater Vehicles (AUVs) and Remote Operated Vehicles (ROVs) are the safest way to perform underwater tasks [1].

Autonomous underwater navigation remains an open and challenging problem due to the nature of the environment. Water attenuates the electromagnetic waves limiting wireless communication and light-based perception: cameras are limited by water visibility and light conditions. Underwater sonar does not suffer these limitations but the acoustic images are noisy and less informative [2], [3].

There are tasks beyond the capacity of vehicles that only navigate in a single environment [4], [5]; thus, researchers have developed hybrid vehicles which can navigate in both air and underwater [6], [7]. These vehicles can perform inspection tasks on partially or completely submerged structures, such as ship hulls or risers [8]. Also, hybrid vehicles can acquire both high-resolution aerial images and underwater acoustic images when diving.

This paper explores the scenario where an autonomous system performs tasks in partially submerged structures and the underwater places can be identified by combining different sensors and views from a heterogeneous environment.

Many studies have proposed a cross-view matching of aerial and ground images to improve terrestrial localization [9], [10], [11], [12]. Here, we propose a novel cross-view image matching approach using underwater and aerial data. Different from terrestrial multi-view matching methods, we used sonar and camera. Acoustic images present specific characteristics such as the nonexistence of textures and colors and lower signal to noise ratio [2], thus preventing us from applying these cross-view matching of aerial and ground images to our problem.

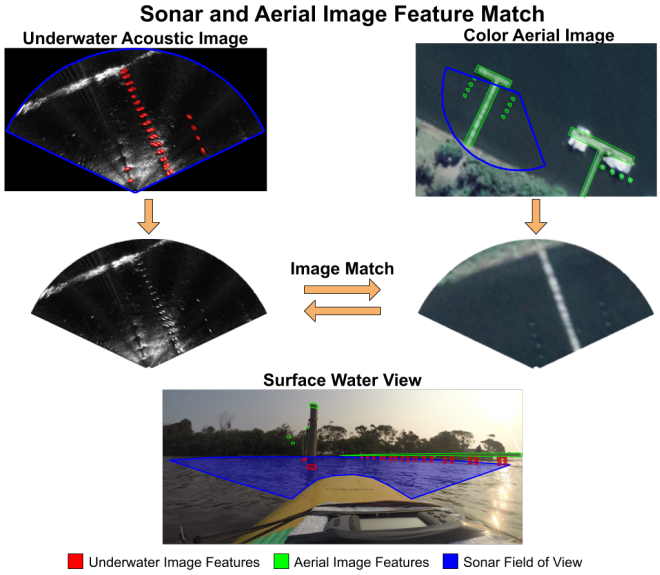

Fig. 1. Matching of color aerial and underwater acoustic images based on their features. Both images are acquired in the same place. The aerial image is cropped in the shape of the acoustic image. The greenish aerial features are detected from top view and the reddish underwater features are identified from frontal view of the partially submerged structures, as illustrated in the surface water view.

Fig. 1 presents an overview of our problem: to match a color aerial image and an underwater acoustic image, and estimate if the two images are being acquired in the same place. Both views show partially submerged structures; we are interested in identifying stable stationary structures, such as piers and piles.

Cross-view and cross-domain image matching is not a trivial problem [13]. Although we mitigated the problem by projecting both images in the horizontal plane of the surface water, the underwater image features do not directly match their aerial counterparts. The aerial images capture the top view while underwater acoustic images capture the front view of the scene. Another challenge is the detection of water bodies, objects, and structures in aerial images because color and textures change according to the weather and the

This work was supported by the Coordenacao de Aperfeiçoamento de Pessoal de Nivel Superior - Brasil (CAPES) - Finance Code 001 and INCT-Mar COI funded by CNPq Grant Number 610012/2011-8.

[1]All Authors are from Intelligent Robotics and Automation Group - NAUTEC, Center for Computational Science - C3, Universidade Federal do Rio Grande - FURG, Rio Grande, Brazil. {matheusmachado, paulodrews, silviacb}@furg.br

season of the year.

Our main contribution is a new cross-view and cross-domain image matching method where color aerial and underwater acoustic images are matched. The matching can improve underwater navigation in partially structured environments with shared features, such as harbors and marinas.

Our pipeline explores two Deep Neural Networks (DNNs) to deal with the problem of detection and matching. First, the color aerial images are semantically segmented using a neural network. Then, the segmented image is converted into a pixel-wise binary image of the stationary structures. Finally, the binary image is cropped and a matching is found with the underwater acoustic image using a second DNN.

This paper is organized as follows: Section II presents related works, Section III explains our proposed pipeline and each step, Section IV shows the experimental results with multiple real datasets, and finally, Section V summarizes our contributions and outlines our future works.

II. RELATED WORKS

According to the best of our knowledge, cross-view and cross-domain aerial and acoustic underwater image matching is a new application domain without any related works. However, some approaches have dealt with matching between aerial and terrestrial images. According to Gao *et al.* [13], these methods can be classified as structure-based and image-based methods.

Structure-based methods search for objects in the environment that can be observed by both ground and aerial views [9], [14], [15], [16], [10].

Image-based methods explore similar features in both aerial and ground views to match and locate urban images. The features can be: *self-similar*, (*i.e* the same features are detected in both views) [11], *semantic* features [17] or *learned* by DNNs [18], [19], [12].

Self-similar feature methods [11] take advantage of urban buildings, such as skyscraper facades, that can be observed in both views. Aerial images are matched with their street-view counterparts searching for shared features such as texture, color, and shape. Semantic-based methods consider information about the observed scene to perform the matching. Castaldo *et al.* [17] geolocate urban images exploring Geographic Information System (GIS) maps. The ground images are segmented and transformed into a top-down view. They proposed the semantic segment layout (SSL) descriptors that match information between the rectified images and the GIS map.

Deep learning methods have also been proposed [19], [12], [18]. Workman *et al.* [18] suggest a deep learning approach to geolocate ground images by using aerial images. They propose a cross-view training strategy that uses Convolutional Neural Networks (CNNs) to extract features from aerial images. Their approach obtained state-of-the-art results for the geolocalization of ground images on two benchmark datasets.

Tian *et al.* [19] use aerial images as a reference to geolocate urban ground images. They explore CNNs to detect and classify objects. Buildings are detected in both aerial and ground views using a DNN. The Siamese network [20] is adopted to learn features of the buildings from both views. The ground view is matched with the aerial view by comparing their buildings using a k-nearest approach.

Leung *et al.* [9] propose a ground image localization approach using aerial images as references. The method is based on orthogonal structures in urban areas. A feature map is generated from aerial orthoimages. The progressive probabilistic Hough transform is adopted to identify building boundaries. Detected lines and vanishing points are analyzed to determine wall orientations. The walls are used as observation for a particle filter [21] framework that locates the ground images.

Noda *et al.* [16] geolocalize terrestrial vehicle cameras using aerial images. A map is built by extracting roads from aerial images. The vehicle images are transformed into a top-down view. The image feature descriptor is adopted during the matching between terrestrial images and the aerial feature map. Geolocating ground images on aerial images has similar aspects to our problem. However, acoustic images present specific characteristics such as nonexistence of textures and colors, and lower signal to noise ratio [2]. Moreover, they consist of a set of object distances and shapes projected onto a plane and are distinct from ground images. Issues related to perspective distortion reported in [17], [16] are not present in acoustic images.

Therefore, performing image matching between aerial and underwater acoustic images requires a specific solution. Our method follows the concept of learning-based methods to extract the best features of each image domain and to compare them. We are motivated by the characteristics of the phenomena acquired by color aerial and acoustic images and the recent success of DNN to identify patterns and deal with complex perception tasks.

III. METHODOLOGY

In this paper, we propose a method to estimate a cross-view and cross-domain image matching in underwater scenarios with partially submerged structures. A vehicle equipped with a Forward-Looking Sonar (FLS) obtains acoustic images in a place where there is an aerial image available. The aerial image can be acquired by satellites or drones and be preloaded into the vehicle, or previously captured in the case of a hybrid vehicle. A typical scenario and the FLS field of view are represented in Fig. 3.

The methodology is divided into three steps, represented in Fig. 2. The first step, aerial image processing, is performed before the vehicle dives. This step consists of semantic segmentation and binarization of the aerial images. A CNN segments the images into stationary objects, moving objects, and water bodies, represented in green, red, and blue colors, respectively, in Fig. 2-Step 1. The segmentation removes movable objects, such as boats. The stationary structures remain in the binary image since they are temporally stable. The second step consists of applying a threshold on acoustic images to reduce noise and remove low intensity acoustic

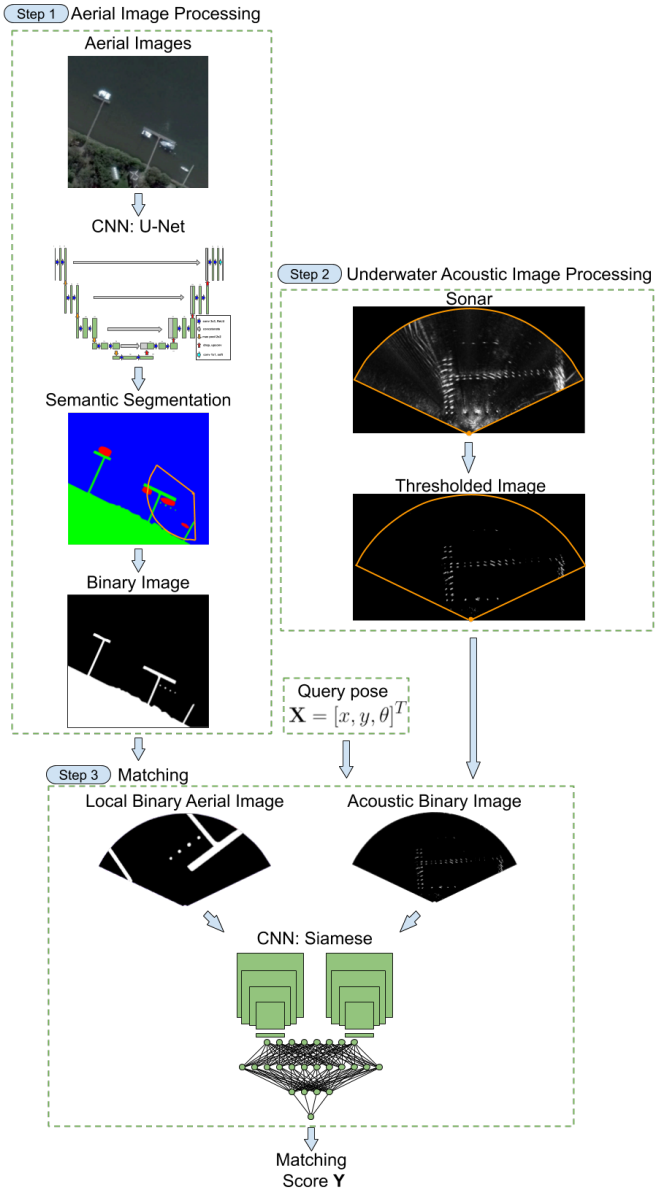

Fig. 2. A visual representation of the three steps of our methodology. An aerial image is semantically segmented into stationary structures (green), water bodies (blue), and movable objects (red), in Step 1. The image is also binarized using the stationary structures. The acoustic image is thresholded preserving relevant objects, in Step 2. Finally, the binary aerial image is cropped based on a query pose **X** and the matching is performed in a Siamese Network. The matching score **Y** identifies similar places.

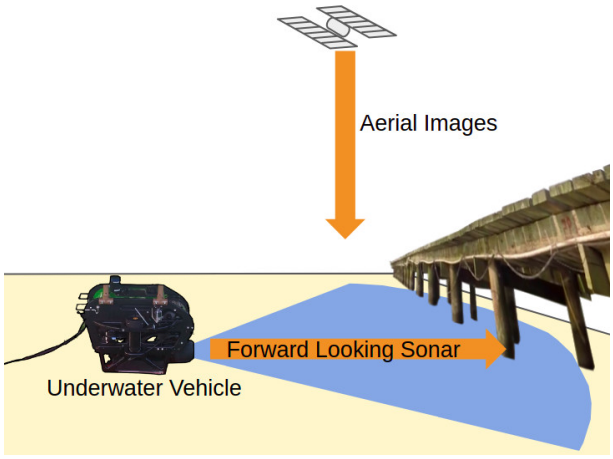

Fig. 3. A typical scenario for our cross-view and cross-domain image matching method. An underwater vehicle equipped with a FLS acquired underwater acoustic images. The FLS field of view is represented by the blue circle sector. The acoustic images overlap the color aerial images.

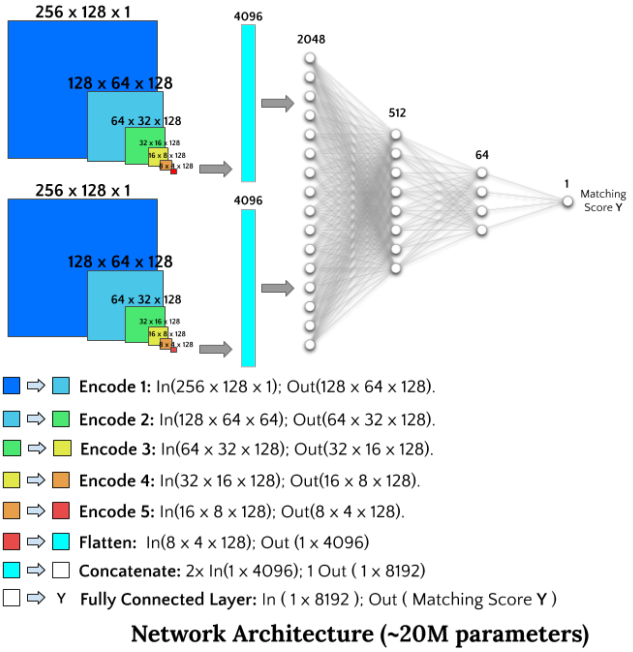

Fig. 4. Siamese-inspired model for matching binary aerial and acoustic images.

returns. Only high acoustic returns representing the object shapes remain in the acoustic binary image. The third step is the matching process. Initially, the binary aerial image is cropped based on a query pose. The comparison uses a CNN capable of obtaining the matching score **Y** to identify similar places.

A. Step 1: Aerial Image Processing

Navigation in dynamic scenarios requires the identification of stationary landmarks in the environment. In this work, we are interested in identifying partially submerged structures, such as piers, piles, and shorelines. Our color aerial image is obtained from a satellite before robot navigation. Movable objects can be summarized as boats on harbors or marinas, as shown in Fig. 1. The semantic segmentation of aerial images extracts those that are important for the matching process. We adopted a CNN to perform semantic segmentation of the aerial images. The network is based on the U-Net architecture [22] with the addition of image padding to keep the original image size. The U-Net [22] is known for its skip connections that transmit residual information. Its architecture has been used for many tasks, such as image classification, reconstruction, and translation [23]. A $256 \times 256 \times 3$ color aerial image is the input to the model, while the output is a $256 \times 256 \times 3$ segmented image

TABLE I
SEMANTIC SEGMENTATION - TRAINING DATASET.

Place	Acquisition Date	Images (256×256)
Yacht Club of Rio Grande Rio Grande - Brazil	April 2013	1000
	March 2014	1000
	November 2015	1000
	August 2016	1000
	October 2016	1000
	August 2017	1000
	Total	6000
Kansai Yacht Club Tokyo - Japan	June 2007	1000
	May 2018	1000
	Total	2000
Southport Yacht Club Gold Coast - Australia	June 2008	1000
	May 2016	1000
	Total	2000
	Overall Total	**10000**

in one-hot encoding. The architecture consists of five encode layers and four decode layers with the original padding being preserved to avoid cropping. A softmax layer is used in our final activation function. We use softmax cross entropy to perform multi-categorical segmentation. The training was performed in a NVIDIA GTX Titan X using the Adam [24] optimizer and a batch size of 16, with the architecture being implemented in the TensorFlow framework.

1) Training: The training dataset was created by manually annotating satellite images of a marina using shots from over several years in three different places around the world, as shown in Table I. Since the sensor resolution changes over the years, a scale image resize is applied to keep a fixed resolution of 0.2 meters per pixel. However, small scale differences are present due to inaccurate satellite data. Also, the challenge of segmentation increases due to the low-resolution of the publicly available satellite images. The dataset is composed of 10 satellite images, as shown in Table I. A data augmentation strategy was adopted[1]. Several 256×256 patches are automatically cropped from the original satellite images following a random motion model. The extraction process registers the occurrence of each class and controls the patch extraction. Thus, our final training dataset is balanced. In total, the training dataset is composed of 10,000 image patches. The network weights are initialized using Xavier initialization [25].

2) Transforming Aerial Image into Segmented Image: A splitting process is conducted before the segmentation to transform a full-size original image into small patches. A sliding window size 256×256 performs image crop row by row with a stride of 86 pixels. Thus, each pixel of the original image is extracted at least nine times in different patches. Our network evaluates all patches and the result is projected into a voting field with the same size as the original full-size image. The final class of each pixel is determined by the winner of a voting process that combines the nine outputs of our model to the same pixel. This approach improves the robustness of our classification method. Finally, a binary image is created. The white pixels are the stationary structures and the black pixels are the water bodies or movable objects, as shown in Fig. 2.

[1]A video demonstration of the data augmentation method is available at `https://bit.ly/2kyC3UZ`.

B. Step 2: Underwater Acoustic Image Processing

A threshold is applied to the sonar image $\mathbf{I_u}$, transforming a 16 bit image into a binary image $\mathbf{B_u}$. The process reduces noise, eliminates seabed information, and removes small objects [26]. After several trials, a threshold value of 355 was adopted for all images. Images with less than one percent of non-zero pixels are dropped due to lack of information to match.

$$\mathbf{B_u}(x,y) = \begin{cases} 1 & if\ \mathbf{I_u}(\mathbf{x},\mathbf{y}) > 355, \\ 0 & otherwise. \end{cases} \quad (1)$$

C. Step 3: Matching

The matching process compares one binary underwater acoustic image and a cropped binary aerial image. The cropping is based on a query pose. The result is a matching score that identifies similar places. A CNN architecture inspired by the Siamese network [20] was developed for the matching problem. The network's input is two $256 \times 128 \times 1$ images, (*i.e.* one binary underwater acoustic image and another binary aerial image). The network contains encoding layers for both inputs. However, unlike the Siamese network, there is no sharing of weights since the domains of the input images are distinct. An encode layer consists of five operations: two 3×3 convolutions with 128 filters, followed by ReLU activation functions and a 2×2 max-pooling operation. Afterward, the output of the two encoding layers is flattened, each into a 4096-dimensional vector. Then, they are concatenated into an 8192-dimensional vector and fed into a sequence of four fully connected layers of sizes: 2048, 512, 64, and 1. The output of the final fully connected layer is the matching score. The network was also implemented using TensorFlow. We also use Adam optimizer [24] and Xavier initialization [25]. Fig. 4 shows a visual representation of our architecture.

1) Training: The model was trained using real acoustic images from the dataset ARACATI 2017 [3] and the ground truth semantic segmentation of the Yacht Club of Rio Grande, August 2017.

The acoustic images were binarized as described in Section III-B. The process resulted in 1521 underwater acoustic images with accurate position and orientation acquired by a Differential Global Position System (DGPS) and the vehicle's compass. The data collection is detailed in [3].

Fan-shaped cropped images were extracted from the aerial segmentation, as shown in Fig. 2, considering the acoustic image position, orientation, and FLS field of view. A reference matching score $\mathbf{R}$ is used to train the Siamese network. The score $\mathbf{R}$ considers the position and orientation error between the acoustic image and the local binary aerial image shown in Fig. 2-Step 3.

Thus, the training dataset was composed of 1521 pairs of binary underwater acoustic and aerial images with a 100% match. Aerial images with a 0%, 25%, 50%, 75% match were extracted, resulting in a balanced training dataset with

TABLE II

SEMANTIC SEGMENTATION - TEST DATASET

Place	Collected Date	Size
Yacht Club of Rio Grande	November 2009	2944×1920
Rio Grande - Brazil	July 2019	3146×1871
Southport Yacht Club Gold Coast - Australia	August 2014	2560×1536
San Francisco Yacht Club Belvedere Tiburon - USA	October 2009	2432×1792

7605 image pairs. The network loss function adopted was the cross entropy. We divide 10% of the training dataset to evaluate the performance of the network, (*i.e.* the validation dataset).

The network was trained over 32 epochs defined by observing the performance of the network on the validation dataset. Note that our Siamese network does not achieve the same performance of the reference matching score **R**, since **R** depends on perceived location data and the network depends on cross-view and cross-domain image features.

IV. RESULTS

The performance of our approach was evaluated with real underwater acoustic images from the dataset ARACATI 2017 [3] and the satellite images from Google Earth displayed in Table II.

A. Semantic Segmentation

Our segmentation CNN was trained with the data shown in Table I and evaluated with the test data shown in Table II. The test dataset includes images from different places, for example San Francisco, and the same place on another date. A qualitative result is shown in Fig. 5 and quantitative results in Fig. 6.

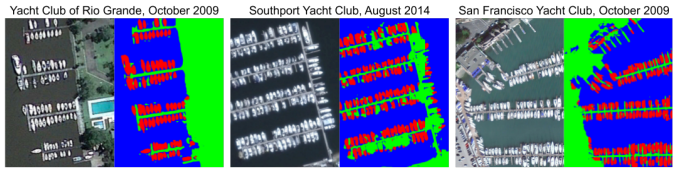

Fig. 5. Qualitative results of Step 1: Aerial image processing. The aerial images on the left column are semantically segmented by our CNN network resulting in the images on the right column. Stationary structures are represented in green, movable objects in red and water bodies in blue.

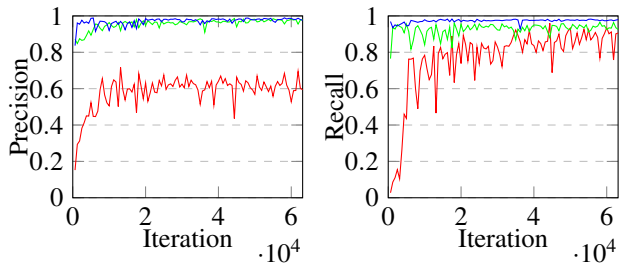

Fig. 6. Precision and Recall of the semantic segmentation on the validation dataset during the training time. Movable objects in red, stationary structures in green and water in blue.

The result shows that our model is able to segment different images. It is a non-trivial problem even for images from a trained place captured on another date. Natural effects such as weather conditions, position of the sun, and different seasons change the color and appearance of the images.

Our network presented difficulties to segment boats near to the piers on the Southport Yacht Club, August 2014, Australia. We believe it happened because of the similarity between the boats on the ground (stationary structure) and boats in the water (movable objects). This problem may be mitigated by eliminating them on the ground images from the training dataset.

B. Matching

The matching network is validated using acoustic images from the dataset ARACATI 2017 [3] and a satellite image of the Yacht Club of Rio Grande, July 2019. The final result of the matching problem is to see if the match occurs or not. We use a fixed threshold of 0.5 to classify each pair as a match or not.

After the segmentation of the satellite image, the validation dataset is created following the same methodology as the training data, described in Section III-C.1. This results in a balanced validation dataset with 3250 pairs of images.

The segmentation performance of a new satellite image and a positive and negative match case is shown in Fig. 7. The results shows how the network deal when the images are from the same place and from a different place. The results are prominent since the matching is highly dependent on the segmentation of the first network.

The complete performance of the matching network is shown in Fig. 8. The Receiver Operating Characteristics (ROC) and Precision-Recall curves are generated by varying the threshold from 0 to 1. All 3250 pairs of images are evaluated. Our approach is able to achieve a precision rate of up to 80%. The ROC curve above the main diagonal shows that our method is able to classify most of the correct matches. The matching network can be evaluated approximated 100 times per second using an NVIDIA GTX 1080.

V. CONCLUSIONS

This work proposed a new cross-view and cross-domain matching problem based on color aerial and underwater acoustic images. The methodology includes traditional image processing techniques and DNN. The method is applicable in partially structured environments such as marinas and harbors. The matching enables a place recognition using heterogeneous sensors. Qualitative and quantitative results show our method is able to perform matching using real-world datasets from multiple regions. The precision obtained in detecting a place is up to 80%.

Since underwater navigation is slow and the pipeline is fast, our future works will be focused on integrating the method into a particle filter framework to improve underwater localization. We plan to evaluate the matching process on the particle filter algorithm as an observation model. The

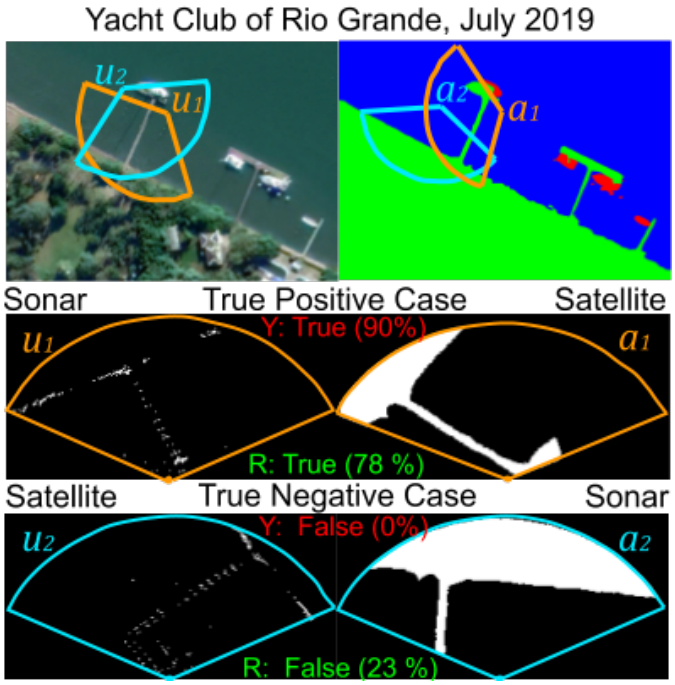

Fig. 7. Qualitative results of Step 3: Matching. Two underwater images on poses u_1 and u_2 and two aerial images on poses a_1 and a_2 are evaluated by the Siamese network resulting in the red **Y** scores. The matching reference score **R** is shown in green. The aerial image, its semantic segmentation after Step 1, and the evaluated poses are shown in the first line. A true positive case is shown in the second line, (*i.e.* both images are captured in the same place). The third line shows a true negative case. The cropped aerial images are extracted in poses a_1 and a_2 considering their geolocation. The acoustic underwater images are collected by the vehicle on poses u_1 and u_2 estimated by its compass and DGPS in a shallow water experiment.

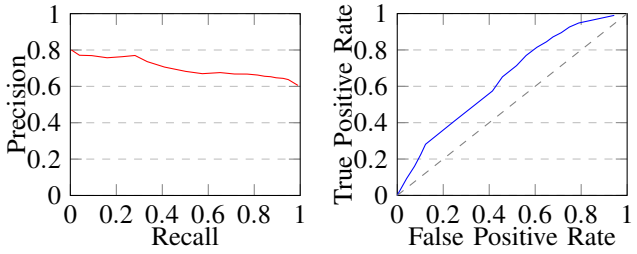

Fig. 8. Matching performance, the precision-recall and ROC curves.

matching can be employed for updating the weights of the particles in the perception step. We also believe the method can be easily extended to regions such as military areas, large forests, *e.g.* Amazon Rainforest), and iced places, *e.g.* Antarctica. Thus, an evaluation in these environments will be conducted.

References

[1] Y. R. Petillot, G. Antonelli, G. Casalino, and F. Ferreira, "Underwater robots: From remotely operated vehicles to intervention-autonomous underwater vehicles," *IEEE RA Magazine*, vol. 26, no. 2, pp. 94–101, 2019.

[2] N. Hurtos, S. Nagappa, X. Cufi, Y. Petillot, and J. Salvi, "Evaluation of registration methods on two-dimensional forward-looking sonar imagery," in *OCEANS - Bergen, 2013 MTS/IEEE*, June 2013, pp. 1–8.

[3] M. Dos Santos, G. Zaffari, P. O. Ribeiro, P. Drews-Jr, and S. Botelho, "Underwater place recognition using forward-looking sonar images: A topological approach," *JFR*, vol. 36, no. 2, pp. 355–369, 2019.

[4] R. da Rosa, P. Evald, P. Drews-Jr, A. Neto, A. Horn, R. Azzolin, and S. Botelho, "A comparative study on sigma-point kalman filters for trajectory estimation of hybrid aerial-aquatic vehicles," in *IEEE IROS*, 2018, pp. 7460–7465.

[5] Y. Wu, "Coordinated path planning for an unmanned aerial-aquatic vehicle (UAAV) and an autonomous underwater vehicle (AUV) in an underwater target strike mission," *Ocean Engineering*, vol. 182, pp. 162 – 173, 2019.

[6] D. Mercado, M. Maia, and F. J. Diez, "Aerial-underwater systems, a new paradigm in unmanned vehicles," *JINT*, vol. 95, no. 1, pp. 229–238, 2019.

[7] P. L. J. Drews-Jr, A. A. Neto, and M. F. M. Campos, "Hybrid unmanned aerial underwater vehicle: Modeling and simulation," in *IEEE/RSJ IROS*, Sep. 2014, pp. 4637–4642.

[8] A. A. Neto, L. A. Mozelli, P. L. J. Drews-Jr, and M. F. M. Campos, "Attitude control for an hybrid unmanned aerial underwater vehicle: A robust switched strategy with global stability," in *IEEE ICRA*, 2015, pp. 395–400.

[9] K. Y. K. Leung, C. M. Clark, and J. P. Huissoon, "Localization in urban environments by matching ground level video images with an aerial image," in *IEEE ICRA 2008*, May 2008, pp. 551–556.

[10] A. Viswanathan, B. R. Pires, and D. Huber, "Vision based robot localization by ground to satellite matching in gps-denied situations," in *2014 IEEE/RSJ IROS*, Sep. 2014, pp. 192–198.

[11] M. Wolff, R. T. Collins, and Y. Liu, "Regularity-driven building facade matching between aerial and street views," in *IEEE CVPR*, June 2016, pp. 1591–1600.

[12] S. Hu, M. Feng, R. M. H. Nguyen, and G. Hee Lee, "CVM-Net: Cross-view matching network for image-based ground-to-aerial geo-localization," in *IEEE CVPR*, 2018, pp. 7258–7267.

[13] X. Gao, S. Shen, Z. Hu, and Z. Wang, "Ground and aerial meta-data integration for localization and reconstruction: A review," *Pattern Recognition Letters*, vol. 127, pp. 202–214, 2018.

[14] A. Li, V. I. Morariu, and L. S. Davis, "Planar structure matching under projective uncertainty for geolocation," in *ECCV*, 2014, pp. 265–280.

[15] A. L. Majdik, Y. Albers-Schoenberg, and D. Scaramuzza, "MAV urban localization from google street view data," in *IEEE/RSJ IROS*, 2013, pp. 3979–3986.

[16] M. Noda, T. Takahashi, D. Deguchi, I. Ide, H. Murase, Y. Kojima, and T. Naito, "Vehicle ego-localization by matching in-vehicle camera images to an aerial image," in *ACCV*, 2010, pp. 163–173.

[17] F. Castaldo, A. Zamir, R. Angst, F. Palmieri, and S. Savarese, "Semantic cross-view matching," in *IEEE ICCVw*, 2015, pp. 9–17.

[18] S. Workman and N. Jacobs, "On the location dependence of convolutional neural network features," in *IEEE CVPRw*, 2015, pp. 70–78.

[19] Y. Tian, C. Chen, and M. Shah, "Cross-view image matching for geo-localization in urban environments," in *IEEE CVPR*, 2017, pp. 3608–3616.

[20] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *IEEE CVPR*, vol. 1, 2005, pp. 539–546.

[21] S. Thrun, "Probabilistic robotics," *Communications of the ACM*, vol. 45, no. 3, pp. 52–57, 2002.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[23] G. Giacomo, M. M. dos Santos, P. L. J. Drews-Jr, and S. S. C. Botelho, "Sonar-to-satellite translation using deep learning," in *IEEE ICMLA*, 2018, pp. 454–459.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[25] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *AISTATS*, 2010, pp. 249–256.

[26] M. dos Santos, P. Ballester, G. Zaffari, P. Drews-Jr, and S. Botelho, "A topological descriptor of acoustic images for navigation and mapping," in *IEEE LARS/SBR*, 2015, pp. 289–294.