Underwater Vision-Based Gesture Recognition: Robustness Validation for Safe Human-Robot Interaction

Anonymous LXAI CVPR 2021 Workshop submission

Paper ID ****

Abstract

Underwater robot interventions require high levels of safety and reliability, specially when used for human-robot collaboration missions. For this reason, we propose a robust gesture-based communication pipeline for an AUV to assist divers in their missions, which provides feedback in every step of the process to ensure valid operations. In this work, we prove that such human-in-the-loop system is still necessary despite the recent success of deep learning vision models because of the difficulty to predict the amount of data that will be available for training and how representative of the actual environment it will be. We highlight these issues by replaying the development of the EU-FP7 CADDY project and testing the performance of several state-of-the-art methods with the historically available data at each development stage. These methods include both classical machine learning and deep learning frameworks. We can gain insights into which DL architectures are more robust to high intra-class variations or smaller datasets for underwater image applications. Further work will consider image distortions based on the light behavior underwater.

1. Introduction

In the past decade, there have been significant research efforts to boost Autonomous Underwater Vehicle (AUV) capabilities that are essential for inspection and intervention tasks in different application domains including industrial, oceanographic, archaeological and environmental scenarios. But the level of autonomy required to perform safe and reliable missions involving for example dexterous manipulation, cautious biological sampling, or exhaustive environment exploration in a fully unsupervised manner is not yet persistently available and still a matter of research. Hence, human intervention by divers or tele-operation of a system is often necessary. Nevertheless, human divers can profit from AUV assisting and monitoring, i.e., there is ample potential for underwater human-robot collaboration. To this end, we present work on a robust gesture-based communication pipeline for an AUV to assist divers in missions.

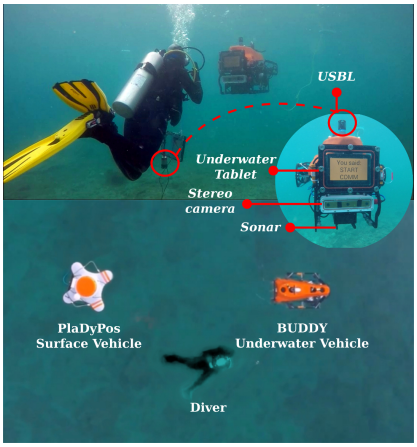

Figure 1: The CADDY system for assistance in diver missions. (Right) The Buddy-AUV equipped with a Blueprint Subsea X150 USBL, a Underwater Tablet, a BumbleBeeXB3 Stereo Camera, and an ARIS 3000 Imaging Sonar for diver tracking, monitoring and communication. (Top) Diver gesturing a command. (Bottom) Aerial view of the system with PladyPos surface vehicle for global positioning.

In order to achieve high performance levels and to meet safety-critical standards, a system comprised of two robot "companions" – an Autonomous Underwater Vehicle (AUV) and an Autonomous Surface Vehicle (ASV) – was proposed in the EU-CADDY FP7 project (Cognitive Autonomous Diving Buddy) to both monitor and support divers operations (http://www.caddy-fp7.eu/). To achieve safety-critical standards, an AUV called Buddy constantly tracks the diver through a sonar and a stereo camera. The AUV also receives updates about the diver's health, position, heart-rate, breathing and IMU sensors through acoustic modems. This information is parsed and broadcasted to the ASV called PlaDyPos, which relays the data to an offshore vessel or an onshore control center (see Fig. 1).

We focus on the gesture-based communication pipeline used by the divers to instruct the AUV to assist with a particular task or in an emergency. The gesture language syntax and grammar, named Caddian [2][3], was specifically

developed to enable a wide variety of tasks and quick message correction. Each hand signal is detected and classified based on the stereo camera input. We augment the work done by Chiarella [3] by adding an underwater tablet on the AUV that provides feedback to the divers (see Figure 1). Based on this, the divers can adapt on the fly, e.g., reconfigure mission parameters, abort current tasks or ask for help.

The EU-FP7 CADDY project was developed from 2014 to 2017, under which a series of field data collections and tests were performed to gradually design, benchmark and improve diver gesture classifiers and generate a large dataset [5]. In this article, we examine how a classical Machine Learning (ML) approach to gesture recognition performs in comparison to state-of-the-art deep learning (DL) methods and how different DL architectures perform compared to each other, by replaying the CADDY development cycle. This means that we evaluate the performance of each method based on the available data at different dates and from different field experiments, which effectively impacts the data size and the environmental conditions it represents. Given the challenges of the highly variant underwater image formation and the logistical difficulties of acquiring massive amounts of labeled and representative data in this environment, the results are also of interest for underwater object detection and classification in general. The results justify the creation of a system with a human-in-the-loop module.

2. Hand gesture detection and classification

2.1. Classical ML approach

At the time of the first CADDY design (2014), deep learning methods for object perception were just starting to be competitive against methods based on engineered features like SIFT, SURF, histogram of gradients (HoG), etc. Likewise, depth calculation from stereo data was primarily done through feature matching and further optimization procedures. Accordingly, our proposed approach to hand detection is a hybrid approach combining both 3D information through disparity maps and cascade classifiers to ensure robustness against different forms of underwater image degradations. Segmentation of the disparity maps based on distance and pixel density offers quite reliable hand detection. However, it fails in the presence of bubbles from and other texture-prominent areas. Thus, 2D cascade classifiers [10] are used to filter these false positive regions.

All of the region proposals are used as input to a final classifier in form of a Multi-Descriptor Nearest Class Mean Forests (MD-NCMFs), which is first introduced for diver localization in [1]. This classifier filters out false positives generated by the previous modules and maps the true positives (hands) to a specific gesture (class) within the Caddian language. The main purpose of this variant of a Random Forest is to aggregate multiple engineered descriptors (SIFT, SURF, ORB, HoG, etc.) that encode different representations of objects, each of them ideally robust against different distortions. The methodology is shown in Fig. 2.

2.2. Deep learning approach

State-of-the-art deep learning models for visual object detection and classification often follow three meta-architectures: Single Shot Detector (SSD) [8], Faster Region-based Convolutional Neural Network (Faster R-CNN) [9], and Region-based Fully Convolutional Neural Network (R-FCN). SSD models offer fast computation speeds since they perform object detection and classification in one single pass of the network. Hence, they are often the preferred choice for embedded systems. Faster R-CNN has two stages, procedurally similar to the described classical ML approach (see Sec. 2.1): a region proposal network generates candidates for object regions and a classifier verifies each of them. Finally, the R-FCN architecture is a compromise between the previous ones as it shares learned features in the initial layers between the region proposal and the classifier network. Based on this and to test the robustness of different deep learning architectures from the literature, we consider here the following four:

Visual model	Feature extractor	Software Library	References
FCN-CNN	ResNet-50	Fast.ai/Pytorch	[6]
SSD	MobileNets	Tensorflow	[8, 7]
Faster R-CNN	ResNet-101	Tensorflow	[9, 6]
Deformable Faster R-CNN	[4]	MXNet	[9, 4]

Table 1: The evaluated Deep Learning models with pre-trained feature extractors.

3. Experiments and results

3.1. Dataset and setup

The recording of underwater gestures took place in three different locations in the open sea as well as in indoor and outdoor pools, respectively in Biograd na Moru (Croatia), the Brodarski Institute (Croatia) and in Genova (Italy). The collected data is divided into 8 scenarios representing different diver missions, locations and field experiments. Scenarios named *Biograd-A*, *Biograd-B*, and *Genova-A* represent trials organized mainly for data collection; they hence contain a high number of samples. The rest of the data was collected during test experiments of real diver missions, such as *Biograd-C* and *Brodarski-A* to *Brodarski-D*. For a detailed explanation of the number of samples in each scenario and their environmental conditions, please refer to [5].

Note that the quantity and quality of the data from one scenario might not be representative enough to make an underwater vision system robust enough to be used "off-the-shelf". To illustrate this, the variations in image quality

Figure 2: **Hand detection**. In parallel (I.a) a Haar cascade model proposes possible image regions for the hand, while (II.b) a disparity map is computed, thresholded by distance and morphologically transformed to reduce noise, to propose more regions. (I.c) A cross-check between these methods generates the final hand image candidates. **Gesture classification**. A Multi-Descriptor NCM tree (MD-NCM) is used; each class centroid – colored dot – traverses a path through the decision tree. (II.a) The image is encoded into different types of feature vectors $\vec{x}$ $\vec{y}$ $\vec{z}$, (II.b) the sample passes down the tree following the closest centroid (aggregated similarity measure). For example, e_r^0 in the first level, and (II.c) when it reaches a leaf, the image is assigned a class distribution, when is computed when trained.

for the different scenarios are shown in Fig. 3, reprinted from [5]. Therefore, for each method described in Sec. 2, we train not just one but four classifiers on four different train/valid/test set partitions of the data as described in Table 2. Scenario F uses the complete (Full) dataset.

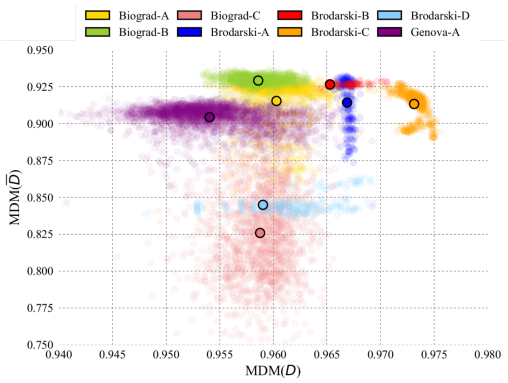

Figure 3: Image quality assessment in the dataset based on the MDM metric [5].

	Part. A	Part. B	Part. C	Part. F
Training Sets	Biograd A,B	Genova A	Brodarski A,C	All scenarios
Samples mean*	338	415	222	1156
Samples median*	151	294	206	792

Table 2: Training set partitioning based on recording scenarios. *Per class samples.

3.2. Performance evaluation

We evaluate each of the ML/DL models trained according to Table 2. The results based on accuracy are shown in Table 3 and 4. Deformable Faster R-CNN and Faster R-CNN have the lead when the complete dataset is used (Model-F), followed by FC-CNN still having an accuracy of 95%. This is an indication that with enough amount and diversity of the data, direct classifiers offer top performance, which can save efforts and time dedicated to manually segmenting object regions in the images. SSD MobileNet has a better performance than the classical machine learning approach, but it drops below 90%. However, SSD is known for its superior speed and its suitability for embedded systems. MD-NCMF, a classical ML method, comes in last with an accuracy below 80%.

As for the Models A to C, trained with data belonging to specific scenarios, Table 2 shows that the performance changes drastically compared to models trained on the full dataset and we can draw the following conclusions:

1) In general, Deformable and standard Faster R-CNN still have the lead (except for Model B), but MD-NCMF offers competitive results and it outperforms FC-CNN and SSD MobileNets. Thus, with lack of data, deep visual models suffer a greater performance drop, $\approx 40\%$, while MD-NCMF accuracy drops only by $\approx 20\%$.

2) For Model B versions, MD-NCMF does better than the other classifiers. A rigorous explanation would require a careful examination of the data including a visualization of the features learned from each of the layers of each deep convolutional model (future work). But we can reasonably postulate that data used to train Model B, i.e., from the *Genova-A* scenario, is not diverse enough for the convolutional models to learn robust features despite *Genova-A* having more samples per class than Model A and C (see Table 2), and that some of the human-engineered features used for MD-NCMF are more representative.

3) Classical models present more "predictable" performance across the test sets. DL models have strong performance drops for particular tests, e.g., see FC-CNN Model B tested against the Brodarski-B test set. A possible explanation for this is again the very different nature of the images and the features computed from the training set in comparison to those from the test set. For example, Model B (trained on the Genova-A scenario) and the Brodarski-B samples differ substantially in brightness, color tone and amount of noise (see Fig. 3).

In future work, our goal is to use image quality metrics as predictors of performance and select the best deep learning architecture accordingly, as well as generating more realistic samples based on the light formation model underwater.

	MD-NCMF				FC-CNN w/ ResNet-50			
	Mod-A	Mod-B	Mod-C	Mod-F	Mod-A	Mod-B	Mod-C	Mod-F
Biograd-A	0 0	0.72	0.52	0.81	0 0	0.42	0.5	0.99
Biograd-B	0 0	0.71	0.51	0.84	0 0	0.21	0.51	0.99
Biograd-C	0.74	0.75	0.68	0.85	0.53	0.45	0.51	0.97
Brodarski-A	0.76	0.76	0 0	0.78	0.52	0.24	0 0	0.95
Brodarski-B	0.81	0.79	0.71	0.73	0.63	0.12	0.68	0.86
Brodarski-C	0.7	0.65	0 0	0.77	0.57	0.48	0 0	0.98
Brodarski-D	0.69	0.61	0.55	0.71	0.71	0.48	0.53	1
Genova-A	0.52	0 0	0.48	0.69	0.34	0 0	0.24	0.89
All scenarios	0.56	0.64	0.46	0.77	0.45	0.36	0.43	0.95

Table 3: Visual models accuracy (0 1) performance on all scenarios, based on Table 2.

	SSD w/ MobileNets				Faster R-CNN w/ Resnet 101				Deformable Faster R-CNN			
	Mod-A	Mod-B	Mod-C	Mod-F	Mod-A	Mod-B	Mod-C	Mod-F	Mod-A	Mod-B	Mod-C	Mod-F
Biograd-A	0 0	0.35	0.38	0.84	0 0	0.63	0.65	0.99	0 0	0.65	0.64	0.99
Biograd-B	0 0	0.29	0.44	0.88	0 0	0.51	0.71	0.99	0 0	0.54	0.7	1
Biograd-C	0.36	0.31	0.4	0.82	0.74	0.58	0.67	0.98	0.74	0.57	0.67	0.98
Brodarski-A	0.38	0.3	0 0	0.87	0.72	0.48	0 0	0.97	0.73	0.49	0 0	0.97
Brodarski-B	0.33	0.29	0.48	0.84	0.72	0.52	0.85	0.96	0.74	0.5	0.87	0.97
Brodarski-C	0.32	0.28	0 0	0.86	0.79	0.56	0 0	0.99	0.78	0.6	0 0	0.99
Brodarski-D	0.29	0.26	0.36	0.79	0.82	0.55	0.68	0.99	0.84	0.54	0.72	0.99
Genova-A	0.25	0 0	0.23	0.75	0.69	0 0	0.44	0.94	0.66	0 0	0.41	0.96
All scenarios	0.28	0.361	0.29	0.85	0.59	0.49	0.52	0.98	0.61	0.5	0.53	0.98

Table 4: Visual models accuracy (0 1) performance on all scenarios, based on Table 2 – continuation.

References

[1] A. G. Chavez, M. Pfingsthorn, A. Birk, I. Rendulić, and N. Misković. Visual diver detection using multi-descriptor nearest-class-mean random forests in the context of underwater human robot interaction (HRI). In *OCEANS 2015 - Genova*, pages 1–7, May 2015. 2

[2] D. Chiarella, M. Bibuli, G. Bruzzone, M. Caccia, A. Ranieri, E. Zereik, L. Marconi, and P. Cutugno. Gesture-based language for diver-robot underwater interaction. In *OCEANS 2015 - Genova*, May 2015. 1

[3] Davide Chiarella et al. A novel gesture-based language for underwater human–robot interaction. *Journal of Marine Science and Engineering*, 6(3), 2018. 1, 2

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. *CoRR*, abs/1703.06211, 2017. 2

[5] Arturo Gomez Chavez, Andrea Ranieri, Davide Chiarella, Enrica Zereik, Anja Babić, and Andreas Birk. Caddy underwater stereo-vision dataset for human–robot interaction HRI in the context of diver activities. *Journal of Marine Science and Engineering*, 7(1), 2019. 2, 3

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016. 2

[7] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017. 2

[8] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016. Springer International Publishing. 2

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Analysis Machine Intelligence*, 39(6):1137–1149, June 2017. 2

[10] Paul Viola and Michael Jones. Robust real-time object detection. volume 57, 01 2001. 2