



# A Survey of Underwater Human-Robot Interaction (U-HRI)

Andreas Birk<sup>1</sup>

Accepted: 8 August 2022 / Published online: 23 September 2022  
© The Author(s) 2022

## Abstract

**Purpose of Review** This review provides an overview of the current state of the art in Underwater Human-Robot Interaction (U-HRI), which is an area that is quite different from standard Human-Robot Interaction (HRI). This is due to several reasons. First of all, there are the particular properties of water as a medium, e.g., the strong attenuation of radio-frequency (RF) signals or the physics of underwater image formation. Second, divers are bound to special equipment, e.g., the breathing apparatus, which makes, for example, speech recognition challenging, if not impossible. Third, typical collaborative marine missions primarily requires a high amount of communication from the diver to the robot, which accordingly receives a lot of attention in U-HRI research.

**Recent Findings** The use of gestures for diver-to-robot communication has turned out to be a quite promising approach for U-HRI as gestures are already a standard form of communication among divers. For the gesture front-ends, i.e., the part dealing with the machine perception of individual signs, Deep Learning (DL) has become to be a very prominent tool.

**Summary** Human divers and marine robots have many complementary skills. There is hence a large potential for U-HRI. But while there is some clear progress in the field, the full potential of U-HRI is far from being exploited, yet.

**Keywords** Human-Robot Interaction (HRI) · Human-Machine Interaction (HMI) · Marine robotics · Underwater sensors · Underwater vision · Sonar · Diver · Remotely Operated Vehicle (ROV) · Autonomous Underwater Vehicle (AUV) · Deep Learning (DL)

## Introduction

Human-robot interaction (HRI) is a well-established research field covered by books, e.g., [1, 2], conferences, e.g., the Annual ACM/IEEE International Conference on Human-Robot Interaction [3], and survey articles, e.g., [4, 5]. But the situation is quite different for HRI in the context of underwater robotics.

On the one hand, collaboration of humans and robots is very desirable in this domain. There are many tasks that — despite significant advances in marine robotics — still can only be done by divers. But Remotely Operated Vehicles (ROV) and Autonomous Underwater Vehicles (AUV)

can substantially assist and mitigate risks for divers [6, 7] though they also may induce risks themselves [8, 9]. Especially in marine scenarios that involve complex manipulation or require very good situational awareness, the collaboration of humans and robots has a very high potential as they can complement each other. Examples include marine science, archeology, oil and gas production (OGP), handling of unexploded ordnance (UXO), e.g., from WWII ammunition dumped in the seas, or inspection and maintenance of marine infrastructure like pipelines, harbors, or renewable energy installations.

On the other hand, the implementation of Underwater Human-Robot Interaction (U-HRI) is more challenging than on land due to the differences in machine perception that render a straightforward application of well-established methods and technologies difficult, if not impossible (Fig. 1). In general, the human factors in the operation of underwater robots are quite different from the handling of their ground or aerial counterparts [10–12], e.g., with respect to situation awareness, trust, or human-to-machine communication. In [12], a human factors model for U-HRI is presented that can

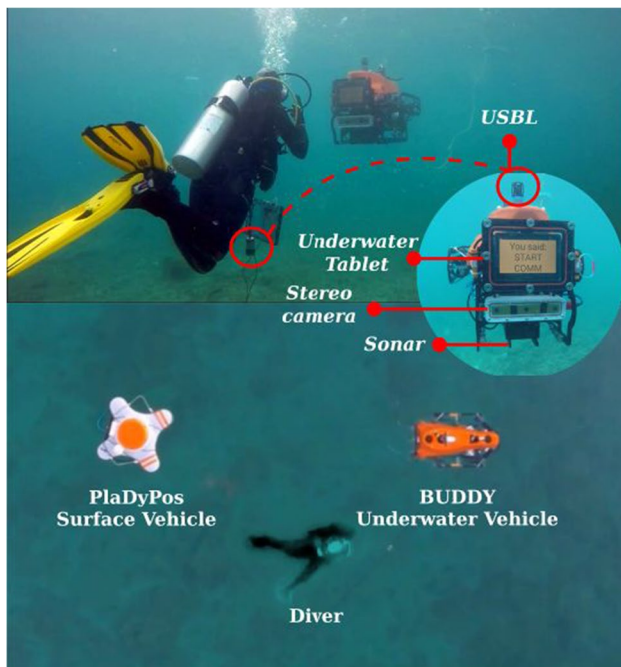
---

This article is part of the Topical Collection on *Underwater Robotics*

---

✉ Andreas Birk  
a.birk@jacobs-university.de

<sup>1</sup> Computer Science and Electrical Engineering, Jacobs University Bremen, Campus Ring 1, Bremen 28759, Germany



**Fig. 1** The CADDY system for assistance in diver missions as an example for U-HRI [7, 13–15]. The Buddy-AUV [16] is equipped with a Blueprint Subsea X150 USBL for localization, an Underwater Tablet, a BumbleBeeXB3 Stereo Camera, and an ARIS 3000 Imaging Sonar for diver tracking, monitoring and communication (right). Commands or even full missions can be signaled by a diver with gestures (top). An Autonomous Surface Vehicle (ASV) named PladyPos provides among others global positioning (bottom)

be a very useful template for taking these aspects in a design into account. The model is based on the core components human, robot, task and environment, for which selected aspects of human factors engineering are discussed.

Regarding the aspects of machine perception, image formation is, for example, in water significantly different than in air [17, 18]. There are several strong effects that challenge standard computer vision methods in underwater scenarios, e.g., (a) refraction effects, which can be quite complex in the water column itself due to temperature or salinity gradients as well as at the housing to water interface which renders standard camera models obsolete for flat-panel interfaces [19], (b) the wave-length dependent attenuation of colors [18], and (c) the almost omni-presence of turbidity due to scattering [20–22]. Nevertheless, optical cameras and the related imaging systems play a very important role for marine systems [23–27].

Sonar is also an important sensor for underwater perception as it works unlike optical cameras in low visibility conditions and over extended ranges. But it has severe limitations in its spatial-temporal resolution and there is a high presence of noise including structural noise effects. There are several main reasons for this. First, propagation of sound

in water is slow and it is influenced by many factors in the water column [28]. Second, sonar samples the environment only with very coarse approximations of rays through beamforming, which leads to side-lobes [29–35], i.e., an inhomogeneous illumination of the scene. Third, interfering echoes due to multi-paths generate structural noise, typically in form of clutter [36–38]. Last but not least, active sensing with sound has the potential to affect marine life or even divers. But the design parameters of devices used in the context of U-HRI, especially with respect to their power levels, are considered to be in a very safe and harmless range [39–41]. Nonetheless, passive sensing using, e.g., vision, is of course definitely on the safe side from a human factor as well as environmental protection perspective.

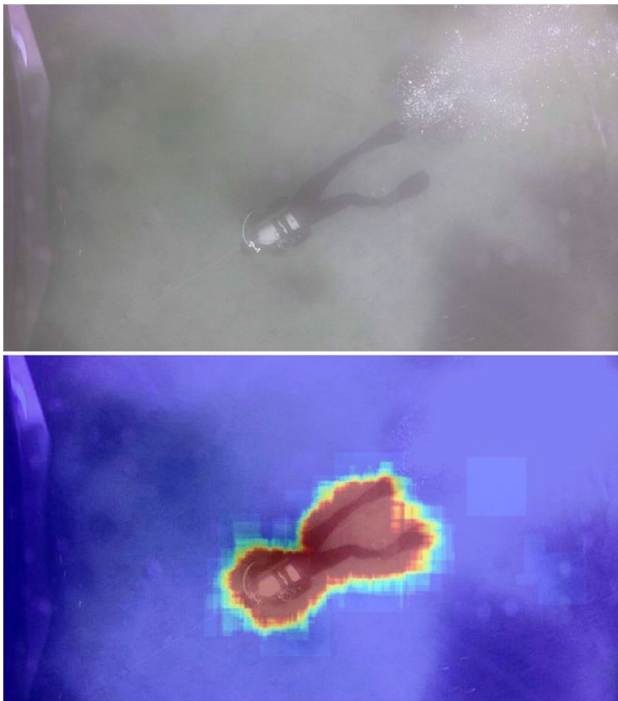
As discussed in the following sections, U-HRI uses a combination of technologies and methods that are relatively strongly adapted to the specific constraints of the marine environment, respectively the specific requirements of typical underwater missions. In “[Diver Detection and Tracking](#)”, a short overview of work on the detection and tracking of divers as a basic element in U-HRI is presented. Communication plays an essential role to enable U-HRI. Methods and technologies for communication from the robot to a diver are discussed in “[Communication from the Robot to the Diver](#)”. The more challenging other way round, i.e., research on the communication of a diver to a robot, is treated in “[Communication from the Diver to the Robot](#)”. Some attention is given to work on using gestures, which includes gesture recognition through machine perception (“[Front-ends to Detect and Recognize Diver Gestures](#)”) as well as their interpretation in form of suited languages (“[Back-ends for Interpreting Underwater Languages](#)”). As Deep Learning (DL) [42, 43] has received quite some attention in recent years for diver detection/tracking as well as underwater gesture recognition in particular, related underwater datasets that are openly available are shortly presented in “[Datasets for Training and for Evaluation](#)”. “[Conclusion](#)” concludes the article.

## Diver Detection and Tracking

The first step towards U-HRI is the detection and tracking of one or multiple divers (Fig. 2). For this purpose, the full range of sensors for underwater object recognition and tracking is used, i.e., especially optical and acoustic devices.

Vision is typically used for the near field and it includes both monocular cameras as well as stereo systems to employ range information. Acoustic methods are typically used to cover the medium to far range. An overview of the different sensor technologies is provided in Table 1.

Also, the whole range of methodologies in machine perception can be found as shown in Table 2. Please note that the USBL/pinger based approaches are not included in this



**Fig. 2** The detection and tracking of divers is an important first step as basis for U-HRI. The example shown here is based on monocular vision and a classic machine learning approach [48]

table. They provide an inherent (detection and) tracking of the divers as they are engineered as localization devices. One can observe that there is a clear shift over time from classical methods of Computer Vision (CV) over classical Machine Learning (ML) towards Deep Learning (DL), which has become the dominant paradigm in recent years.

Some work in this context is also more or less agnostic to the actual machine perception for the detection, e.g., [70] considers safe motions of an AUV among divers. Aspects of AUV-control to safely work with divers and to aid them are also studied in [60]. Furthermore, different protocols for interaction and control can also be studied and trained either fully [61, 71] or partially in simulation [72]. Aspects of the motion control also include the anticipation of the future motion trajectory of divers [57] or the combination of diver following with terrain-relative navigation [61].

**Table 1** For diver detection and tracking, different sensor modalities can be used

vision		acoustic	
monocular	stereo	USBL/pinger	sonar
[44–57]	[58–61]	[58, 60, 62–65]	[58, 60, 66–68]

Multimodal approaches [58] [69•], [60] appear in several columns according to the different sensor modalities that they use

**Table 2** The full range of different perception methods is used for detection and tracking of divers. There is a clear shift towards Deep Learning in recent years as can be seen from the range of the years of appearance of the different publications provided below

Classic Computer Vision (CV)	Classic Machine Learning (ML)	Deep Learning (DL)
[44–46]	[47–49, 50, 66]	[51–59, 61, 67, 68, 69•]
2007–2011	2013–2019	2017–2021

USBL/pinger based approaches are not included in this table

A general overview of person-following in the context of Human-Robot-Interaction is also provided in [73], which covers work using underwater as well as aerial and ground vehicles.

### Communication from the Robot to the Diver

While the communication from the diver to the robot has received quite some attention as discussed later on in more detail in “[Communication from the Diver to the Robot](#)”, the other way round, i.e., the communication from the robot to the diver, is a bit less prominent research topic. There are multiple reasons for this.

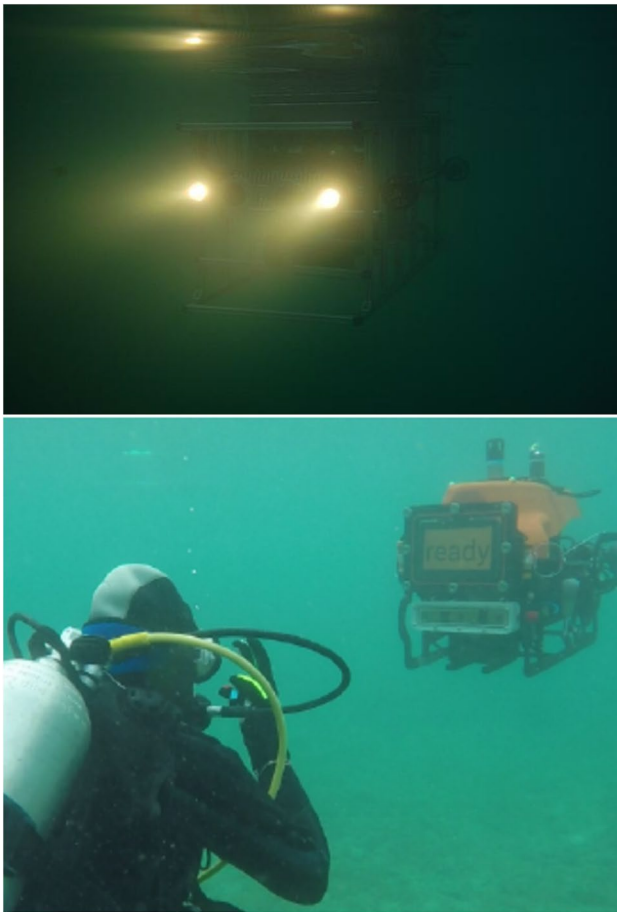
First and foremost, the diver is often mainly interested in commanding the robot to do certain tasks in typical underwater missions. The robot-to-diver communication is hence primarily seen as a confirmation channel that a command was received, respectively that the task execution has started.

Second, it is often even immediately visible to the diver if the robot starts the intended task execution or not without explicit communication feedback. Hence, the amount and expressiveness of the robot-to-diver communication are often considered to be not so essential or even completely negligible.

Third, there are relatively simple technical means of providing feedback from the robot, e.g., just blinking lights, as discussed below (“[Signaling with Lights](#)”). Nevertheless, also some more sophisticated ideas on robot-to-diver communication also exist, e.g., the use of displays, respectively tablets (“[Displays and Tablets](#)”) or of robot motion in form of kinemes (“[Robot Motion and Kinemes](#)”).

### Signaling with Lights

A very simple form of robot-to-diver feedback can be provided by using lights. Lights are a typical default hardware on ROV and AUV anyway as they are helpful for underwater vision. Also, they are by default controllable by the vehicle’s software. The disadvantages of using lights are that the illumination of the scene changes when they are switched on



**Fig. 3** The communication from the robot to the diver typically just serves as feedback that a command was received and that the task will be executed. This can be based on very simple means like just blinking lights (top) [46, 74] or more sophisticated devices like an underwater tablet (bottom) [7]

and off for communication, i.e., computer vision on the robot is difficult during these periods. Furthermore, there is a very low expressive power in using blinking lights — unless the diver is trained in understanding, e.g., Morse code.

The typical use of lights in U-HRI is hence limited to the confirmation that a command was received. A publication that explicitly mentions the use of lights for robot-to-diver communication is [75]. But this idea is so straightforward and basic that it seems to be not explicitly mentioned in the literature; though pictures and videos suggest that it is widely used as feedback loop in research on diver-to-robot communication. This includes, for example, also early own work on U-HRI [46, 74] (Fig. 3, top).

### Displays and Tablets

A more sophisticated form of robot-to-diver communication can be based on the use of displays, respectively tablets

(Fig. 3, bottom). Early work in that direction is presented in [76]. There, a tethered underwater tablet carried by the diver is used. Concretely, an optical fiber is employed as tether, which has the advantage of a high bandwidth and a low latency. Detailed graphical information can hence be provided for the diver. But the use of a tether from a diver-tablet to a robot leads to practical limitations, e.g., with respect to the maximum distance of the diver to the robot, or the risk of entanglement of the system or even of the diver in the cable.

One alternative is therefore to use wireless communication. Though not specifically developed for U-HRI, the underwater tablets with multimedia communication [77] developed in the project “Autonomous underwater Robotic and sensing systems for Cultural Heritage discovery Conservation and in situ valorization (ARCHE-OSU<sub>b</sub>)” could be used for that purpose. As radio-frequency (RF) signals are heavily attenuated underwater, acoustic communication needs to be used. This is linked to limitations in, e.g., bandwidth and latency [78–80]. Furthermore, it can be tedious for a diver to constantly carry a tablet with her/him.

One option is hence to place the display or tablet directly on the robot (Fig. 3, bottom). The underwater tablet developed in the project “Cognitive autonomous diving buddy (CADDY)” was, for example, used for this purpose [7, 13–15]. The CADDY-tablet can also be carried by the diver and it is in addition capable of getting inputs via a waterproof pen, i.e., it can also be used for diver-to-robot communication (c**Communication from the Diver to the Robot**”).




Displays or tablets on the robot have the disadvantage that they require dedicated hardware and take space on the vehicle. But they do not interfere with other components and they provide a high expressive power, especially if their full potential is used in terms of displaying text as well as using colors and temporal elements, e.g., blinking.

### Robot Motion and Kinemes

An interesting option is also to use the motion of the robot itself for the communication. As mentioned before, the diver can relatively easily perceive in many application scenarios if the robot starts a task execution; typically, this is just a transition from station-keeping while receiving the command to any kind of motion that relates to (the start of) the execution of the task. But this is of course a quite informal and potentially error-prone method for a kind of robot-to-diver “communication”.

The use of robot motions as explicit, dedicated signals to the diver is in contrast proposed in [82, 83•]. While [82] concentrates on specifically an underwater robot, an application of the concepts to aerial and terrestrial robots is in addition included in [83•].

**Fig. 4** An Aqua robot [81] using different Kinemes, i.e., dedicated signaling motions, which can be used for robot-to-diver communication [82, 83•] (figure courtesy of Michael Fulton)

Kineme	Meaning	Visuals
$K_{Affirmative}$	Yes, Okay.	
$K_{Negative}$	No.	
$K_{Danger}$	Danger in the area.	
$K_{Attention}$	Pay attention to the robot.	
$K_{Malfunction}$	Internal malfunction.	
$K_{WaitCMD}$	Waiting for instructions.	
$K_{Indicate\_Direction}$	Go to direction.	
$K_{WhichWay}$	Which way should we go?	
$K_{Stay}$	Stay where you are.	
$K_{ComeHere}$	Come to the AUV.	
$K_{FollowMe}$	Diver should follow AUV.	
$K_{FollowYou}$	AUV will follow diver.	
$K_{BatteryLevel}$	Battery level is	

The robot motions are denoted in [82, 83•] as kinemes, which is a term from kinesics, i.e., the field of study of (human) body motions in communication [84, 85]. An overview of several kinemes executed by an Aqua robot [81] in Fig. 4. The advantage of kinemes is that they do not require additional hardware. And there is the option to try to exploit motions that may be naturally understood by humans like a kind of nodding behavior for a “Yes” or a shaking for a “No”.

But the according motions can also be challenging depending on the platform used. For example, roll and pitch motions are typically intentionally suppressed by the mechanical design of underwater robots to increase their stability within a horizontal 2D motion plane to only use one additional degree-of-freedom for diving, i.e., to only actively change the elevation above the sea-floor. Also, these motions can interfere with the perception components on the robot, e.g., by inducing motion blur or by simply not orienting the robot, and hence the sensors, towards a target while a nodding or shaking behavior is executed. Last but not least, there are some limitations in the overall expressive power of robot motions — at least in comparison to tablets/displays.

Nonetheless, the concept is very interesting and it has quite some room for further work in the context of cooperative human-robot mission execution, e.g., when the diver is further away from the robot or when reading information on a table generates too much overhead.

### Communication from the Diver to the Robot

The communication from the diver to the robot is as discussed before an essential part of U-HRI as the human is often primarily interested in commanding the execution of tasks or even of complete missions to the robot.

Some standard forms of in-air HRI have unfortunately severe limitations underwater. For example, the use of speech recognition is tremendously difficult for U-HRI. The diver has a breathing apparatus that makes it challenging to clearly voice anything. It is even a challenge for communication between humans using high-end commercial speech communication links for divers [86, 87].

One option is dedicated input devices like underwater tablets, which are discussed in “[Underwater tablets for User Input](#)”. A more natural way is to use visual

interactions. In early work on U-HRI [88], artificial fiducial markers are used to provide signals to the robot. The cards with the artificial markers have the significant advantage that they ease the challenges of underwater vision. But there are also significant disadvantages like the number of cards that the diver must carry and the effort to handle them.

The standard form of communication among divers is based on gestures. It is hence a natural choice to also use them in the context of U-HRI. As pointed out in [89•], gesture recognition can be divided into two parts that both have received some dedicated attention in research, namely a front-end (“[Front-ends to Detect and Recognizes Diver Gestures](#)”) where the machine perception of individual gestures takes place and a back-end (“[Front-ends to Detect and Recognizes Diver Gestures](#)”) that deals with higher-level language aspects, i.e., which serves for the interpretation of symbol sequences, error detection and correction, as well as feedback to the diver.

### Underwater Tablets for User Input

As already discussed in “[Display and Tablets](#)”, tablets are interesting options to provide information from the robot to the diver, i.e., to serve as displays. The other way round is more challenging, i.e., to provide input to an AUV or ROV with a tablet. Commercial off-the-shelf (COTS) tablets have to be put into water-proof housings, which renders touch-based input with the standard capacitive sensing extremely difficult, if not unusable. Water-proof buttons and dials exist that may be used as an alternative [90], but they tend to significantly add to the system’s complexity and can easily become point of failures in terms of functionality as well as water-tightness.

As discussed in “[Display and Tablets](#)”, a tethered tablet is used as display in [76]. The option of diver-to-robot communication is only mentioned in that context by possibly using the Inertial Measurement Unit (IMU) for joystick or Wii-like inputs. This may be an idea to test, but this requires the tablet to be carried by the diver. This can be tedious. Furthermore, it involves the challenges of underwater wireless networking [77] or the limitations of a tether [76].

As mentioned before, the CADDY underwater tablet [65] can be mounted on a robot [7, 13]. It has the significant advantage that the COTS tablet used as underlying basis has an inductive pen. In contrast to the standard capacitive solutions used in most COTS tablets, the inductive technology is still properly functioning when both the tablet and the pen are in their custom-made watertight housings. Though a tablet with a pen allows easy, intuitive input by humans, it also has the disadvantage that the diver has to get close to the robot to do this.

### Front-ends to Detect and Recognize Diver Gestures

Gestures are a natural basis for diver-to-robot communication. First and foremost, they are already extensively used by divers who can easily adopt them for interacting with robots, even if they did not have any according experiences before [13–15, 89•]. Second, there are inherent limitations of water as a medium, e.g., in form of its substantial damping of RF-signals, which impede the use of many communication approaches used in air.

Nonetheless, also underwater vision, respectively the use of sonar has also its particular challenges as already discussed in the introduction. Furthermore, gestures are based on motions, which induce in the water column a counter-motion of the diver. Also, there can be currents and waves inducing unintended motions of the diver and of the robot. These factors can affect the distance of the diver and especially her/his orientation towards the robot. Hence, multiple system components like station-keeping, tracking, image segmentation, or fault detection and recovery do play a role for a fully functioning, field-able system. It is therefore of interest to note whether a method only addresses a specific aspect, e.g., the recognition of gestures in pre-segmented images in a dataset under benevolent conditions, or the method, respectively a combination of methods, has been shown to also work in field trials under real mission conditions.

Early work on the use of gestures for U-HRI is presented in [74], which specifically focuses on dynamic gestures with the diver’s arms or hands, e.g., the standard gesture for “stay at the same height”. Concretely, the waving gestures are recognized by differential imaging. The improved Fourier Mellin Invariant (iFMI) is used to extract motions from subsequent frames by registering them with each other. This is followed up by classical methods of Computer Vision (CV) for segmentation, especially simple thresholding. The trajectories of arm or hand motions are then recognized with a Finite State Machine (FSM). The experiments in [74] are done in a pool.

In [92], an imaging sonar, which is also known as acoustic camera, is used as an acoustic sensor for gesture recognition in the context of the CADDY project. First, pre-processing stages with cascade classifiers and shape processing generate features. Based on this, three different classification approaches are used and evaluated, namely a convex hull method, Support Vector Machines (SVM), and the fusion of both. Experiments are conducted in a pool and during field trials with divers [13]. The selection of device parameters within a mission is a known challenge for this type of sensor, which is also reported in [92].

The main type of sensor for U-HRI in CADDY is therefore a (stereo-)camera. To cater for the special challenges

**Table 3** The full range of different methods is used for visual gesture recognition front-ends, i.e., the machine perception of divers' gestures. Also here is a clear shift towards Deep Learning (DL) in recent years

Classic Computer Vision (CV)	Classic Machine Learning (ML)	Deep Learning (DL)
[74]	[13, 15, 48, 91, 92]	[50, 89•, 93–95]
2011	2015–2017	2019–2021

of underwater vision, a modification of Nearest Class Mean Forests (NCMF) is introduced, which is dubbed Multi-Descriptor NCMF (MD-NCMF) [48, 91]. MD-NCMF partitions the sample space by comparing the distances between class means instead of comparing values at each feature dimension like in more traditional Random Forests approaches. Therefore, MD-NCMF can treat each feature-object pair as a new class, e.g., SURF-object1, SIFTobject2, SURF-object2, SIFT-background, etc. MD-NCMF can then determine which one provides the best partitioning of the sample set. MD-NCMF is used both for diver detection and tracking [48] as well as for the classification of diver gestures [91]. Within the CADDY project, the divers wear standard diver-gloves augmented with color markers and experiments include realistic field tests [14, 15].

The gesture recognition front-end in [50] is based on Deep Learning (DL) models, which in general have become quite popular in this context (Table 3). More precisely, Single Shot Detector (SSD) [97] and Faster Region-based Convolutional Neural Networks (Faster R-CNN) [98] are investigated, which achieve above 90% accuracy when being trained with a 50K dataset. It is assumed in [50] that the diver wears no gloves. Therefore, skin detection and image contour estimation can be used. While this is on the one hand more general than the use of colored gloves like in CADDY, it also must be noted that professional — as well as many sport — divers tend to always wear gloves for protection and to avoid heat loss.

In [93], several DL methods are tested on the CADDY dataset [96], i.e., an open access dataset of labeled images from various field trials of divers wearing gloves with artificial color markers. The authors use transfer learning [99], i.e., pre-trained networks, to evaluate several standard DL methods, namely AlexNet [100], VggNet [101], ResNet [102], and GoogLeNet [103]. Classification rates of 95% are achieved with VggNet.

The same dataset [96] is also used in [104]. There, Adversarial Learning [95] is used to augment the data for DL with Mask R-CNN performance of [105]. But though the use of Generative Adversarial Networks (GAN) has recently been quite successful in various cases of underwater vision, the performance gains are not that strong in this case.

Results towards a classification under a wide range of conditions including divers with and without gloves are presented in [94]. Building upon a DL-based approach dubbed SCUBANet to recognize diver body parts [59], MobileNetV2 [106] is trained to recognize 25 image classes using finger count and palm direction — though the authors also state that a significant portion of these classes are unused in most gestures [94].

DL is also investigated in [89•]. There, the focus for the front-end is on a systematic evaluation of the performance of state-of-the-art DL methodologies [97, 97, 98, 98, 102, 107, 108] in comparison to a “classical” Machine Learning method (ML) [48, 91]. Especially, the dependency of the different DL approaches and architectures on the amount and variability of training data is investigated. The training and testing are based on the CADDY dataset [96], i.e., the divers are wearing gloves with color markers. An important contribution of [89•] is to show how the real-world data can be artificially degraded in a physically realistic fashion to extend the amount of available data.

A completely different approach to the gesture front-end is presented in [67, 109, 110]. There, sensors are used to detect gestures with the diver's glove itself. To this end, a regular diver glove is augmented with strain gauges suited for underwater operation [111, 112], an Inertial Measurement Unit (IMU), and a processing unit with acoustic communication to the robot. Though exact details on the underlying classification methods are not described, some clear evidence is presented that the approach can work across different individuals [109, 110].

### Back-ends for Interpreting Underwater Languages

While front-ends discussed in the previous section deal with the machine perception of individual signs, back-ends handle the interpretation of the actual gesture-based language [89•]. Hence, back-ends can be combined with different front-ends as they are agnostic to the way the symbols are perceived and recognized.

Already in early work on U-HRI [88], a conceptually quite sophisticated back-end is proposed. Based on artificial fiducial markers as front-end, the RoboChat language is introduced, which is a bit revised in [113]. The idea is that it provides a programming language with which missions can be specified. It hence has symbols for actions as well as for related parameters. Furthermore, conditionals, i.e., the use of “if”, and functional blocks are specified. One drawback is the lack of handling possible errors in the machine perception and related feedback, i.e., it is assumed that there is an instantaneous, perfect recognition of commands in the front-end [88].

A very expressive language for U-HRI called Caddian is introduced in [114], which features a machine interpreter



**Fig. 5** Examples of images from the CADDY dataset with hand gestures of a diver [96]; more precisely, the left image from a stereo-camera is shown. The data is recorded during field trials in different environment conditions

with a phrase parser, syntax checker, and command dispatcher linked to the mission control [89, 115]. It is based on a context-free grammar (CFG), that allows the diver to specify missions as a sequence of tasks. The syntax checker is implemented as a Finite State Machine (FSM). It provides feedback to the diver and hence allows in situ corrections if there are errors in the machine perception. The back-end in combination with the MD-NCMF gesture recognition as front-end [91] are evaluated in field tests with professional divers [14, 15].

A full language for U-HRI is also presented in [50, 116]. It is syntactically a bit simpler than Caddian as it does not feature a full CFG. The instruction decoder, i.e., the FSM in its interpreter is restricted to only one possible transition from state to state, i.e., gesture to gesture, to avoid ambiguities. It is more efficient than RoboChat [88, 113] in experiments with divers [50, 116], which stems from two effects, namely (a) the use of gestures instead of artificial markers for the front-end and (b) the restriction to a simpler back-end that does not feature a full CFG and just allows to command instructions with parameters.

It is interesting to note in this context that Caddian [114, 115] is also designed to work with gesture-based front-ends [89] and that it features the fast interpretation of a sub-set of often used commands, which is denoted as CADDY-slang. The diver can hence profit from the best of the two worlds of a fast, direct signaling of (parameterized) instructions and the option of in situ programming of complex missions.

## Datasets for Training and for Evaluation

Given the attention that Deep Learning (DL) has recently received for underwater perception in general as well as diver detection/tracking (Table 2) and underwater gesture

recognition (Table 3) in particular, there is a certain importance of related underwater datasets that are openly available and that can hence be easily used for training and testing.

The CADDY dataset<sup>1</sup> [96] consists of 12K stereo-images of divers swimming around during missions plus 10K stereo-images with the explicit use of gestures for communication (Fig. 5). An interesting aspect is that in addition to annotated gestures, the ground truth poses of the divers are available. These are not derived as usual from manual annotation but from DiverNet, i.e., a combination of 17 Inertial Measurement Units (IMU) that are distributed over the diver [63]. As mentioned before, the divers wear gloves with artificial markers in the CADDY project. While gloves are a realistic assumption as they tend to be worn for protection and to avoid heat loss, the markers are a somewhat specialized solution.

The SCUBANet dataset<sup>2</sup> also features images from a stereo-camera [59]. It consists of 1K images of divers that are annotated with labels for the head, the hands and the body. The data includes divers with and without gloves. The divers always wear standard equipment, i.e., if gloves are worn, they are made of the standard homogeneously black neoprene. There are also raw monocular videos for different gestures available<sup>3</sup>.

An other more general underwater dataset is presented in [55], namely the dataset for semantic Segmentation of Underwater IMagery (SUIM)<sup>4</sup>. It consists of 1.5K monocular images with pixel level annotations. The object classes contained therein are fish, reefs, aquatic plants, wrecks/ruins, human divers, robots, and sea-floor.

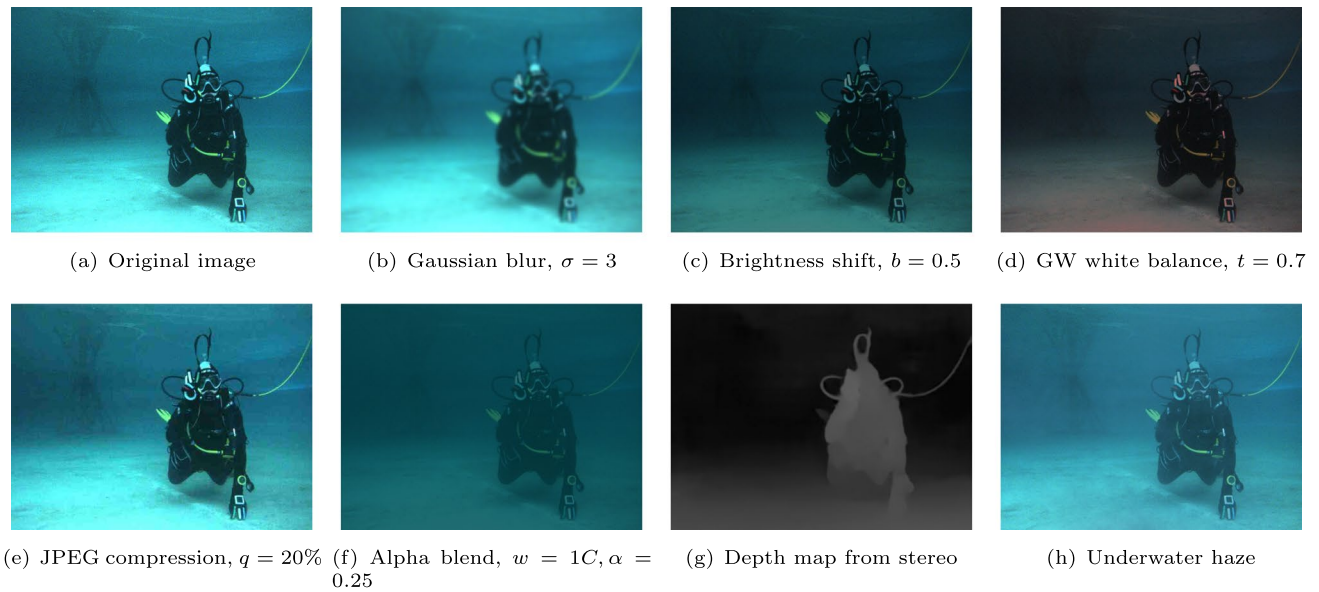
<sup>1</sup> <http://www.caddian.eu/>

<sup>2</sup> <https://vgr.lab.yorku.ca/tools/scubaznet/>

<sup>3</sup> <https://vgr.lab.yorku.ca/tools/scubanet-3-0-raw-raw-dataset-imagery/>

<sup>4</sup> <http://irvlab.cs.umn.edu/resources/suim-dataset>





**Fig. 6** Examples of artificial perturbations of underwater image to increase a data-base for training or testing [89•]. This includes a physically realistic haze-model (h) computed from the stereo-depth information (g)

The Video Diver Dataset (VDD-C)<sup>5</sup> consists of 105K annotated, monocular images of divers [117], i.e., it is a very large dataset. The divers were recorded in pools and the Caribbean off the coast of Barbados. An interest aspect is that features many images with multiple divers, which can be a challenge for diver detection and tracking, especially in the case of occlusions [53].

In [89•], it is shown how artificial image degradation can be used in the context of DL (Fig. 6). This includes physically realistic models of artifacts that commonly occur in underwater vision. These methods hence provide an alternative to the domain-agnostic use of GAN [95] and they are an option to augment visual underwater datasets in general.

## Conclusion

An overview of the state of the art in Underwater Human-Robot Interaction (U-HRI) was presented. First, the basis for U-HRI is presented in form of the detection and tracking of divers. Then, the topic of communication is discussed in some detail. It can be separated into the so-to-say channel from the robot to the diver and the one from the diver to the robot. The latter, i.e., from the diver to the robot tends to be more challenging and it has received more attention in research. Within this line of work, the use of gestures by divers plays an important role. The discussion of according research was structured into front-ends, i.e., the machine perception of gestures, and

into back-ends, i.e., the actual language interpretation. Given the increased use of Deep Learning (DL) in recent years for diver detection/tracking as well as for underwater gesture recognition, a presentation of several related underwater datasets is included in this survey.

**Funding** Open Access funding enabled and organized by Projekt DEAL. The author received funding in the EU FP7 project “Cognitive autonomous diving buddy (CADDY)”, Jan. 2014 to Dec. 2016.

## Declarations

**Conflict of Interest** The author declares no competing interests.

**Human and Animal Rights and Informed Consent** This article does not contain any studies with human or animal subjects performed by any of the authors.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

<sup>5</sup> <https://conservancy.umn.edu/handle/11299/219383>

## References

Three papers of particular interest and published in the last three years have been highlighted as:

- Of importance

- Bartneck C, Belpaeme T, Eyszel F, Kanda T, Keijsers T, Sabanovic S. *Human-Robot Interaction: An Introduction*. Cambridge University Press; 2020.
- Jost C, Le Pevedic B, Belpaeme T, Bethel C, Chrysostomou D, Crook N, Grandgeorge M, Mirnig N. *Human-Robot Interaction - Evaluation Methods and Their Standardization*, volume 12 of Springer Series on Bio- and Neurosystems (SSBN). Springer; 2020.
- Goodrich MA, Schultz AC, Bruemmer DJ. *Proceedings of the 1st ACM Conference on Human-Robot Interaction (HRI)*. ACM Press; 2006.
- Goodrich MA, Schultz AC. *Human-robot interaction: A survey*. *Foundations and Trends in Human-Computer Interaction*. 2007;1(3):203–75.
- Sheridan TB. *Human-robot interaction: status and challenges*. *Hum Factors*. 2016;58(4):525–32.
- UK Health & Safety Executive (HSE). *Offshore Safety Statistics Bulletin*. <http://www.hse.gov.uk/offshore/statistics/hsr2017.pdf>, 2017. Accessed: 2019-08-01.
- Miskovic N, Pascoal A, Bibuli M, Caccia M, Neasham JA, Birk A, Egi M, Grammer K, Marroni A, Vasilijevic A, Vukic Z. *Caddy project, year 1: Overview of technological developments and cooperative behaviours*. In *IFAC Workshop on Navigation, Guidance and Control of Underwater Vehicles (NGCUV)*; 2015.
- Loh TY, Brito MP, Bose N, Xu J, Tenekedjiev K. *A fuzzy-based risk assessment framework for autonomous underwater vehicle under-ice missions*. *Risk Anal*. 2019;39(12):2744–65.
- Miskovic N, Egi M, Nad D, Pascoal A, Sebastiao L, Bibuli M. *Human-robot interaction underwater: Communication and safety requirements*. In *IEEE Third Underwater Communications and Networking Conference (UComms)*. 2016;1–5. IEEE.
- Ho G, Pavlovic N, Arrabito R. *Human factors issues with operating unmanned underwater vehicles*. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 2011;55(1):429–33.
- Xian W, Stuck RE, Rekleitis I, Beer JM. *Towards a framework for human factors in underwater robotics*. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 2015;59(1):1115–9.
- Wu X, Stuck RE, Rekleitis I, Beer JM. *Towards a human factors model for underwater robotics*; 2015.
- Miskovic N, Bibuli M, Birk A, Caccia M, Egi M, Grammer K, Marroni A, Neasham J, Pascoal A, Vukic AVZ. *Caddy - cognitive autonomous diving buddy: Two years of underwater human-robot interaction*. *Marine Technology Society (MTS) Journal*. 2016;50(4):1–13.
- Miskovic N, Pascoal A, Bibuli M, Caccia M, Neasham JA, Birk A, Egi M, Grammer K, Marroni A, Vasilijevic A, Vukic Z. *Caddy project, year 2: The first validation trials*. In *10th IFAC Conference on Control Applications in Marine Systems (CAMS)*. International Federation of Automatic Control; 2016.
- Miskovic N, Pascoal A, Bibuli M, Caccia M, Neasham JA, Birk A, Egi M, Grammer K, Marroni A, Vasilijevic A, Nad D, Vukic Z. *Caddy project, year 3: The final validation trials*. In *OCEANS*. 2017;1–5. IEEE.
- Stilivovic N, Nad D, Miskovic N. *Auv for diver assistance and safety - design and implementation*. In *IEEE/MTS OCEANS*. 2015;1–4. IEEE.
- Marvin A, Blizard. *Ocean Optics: Introduction And Overview*, volume 0637 of *Technical Symposium Southeast*. SPIE; 1986.
- Duntley SQ. *Light in the sea*. *J Opt Soc Am*. 1963;53(2):214–33.
- Luczynski T, Pflingstorn M, Birk A. *The pinax-model for accurate and efficient refraction correction of underwater cameras in flat-pane housings*. *Ocean Eng*. 2017;133:9–22.
- Funk CJ, Bryant SB, and P J Heckman Jr. *PJ. Handbook of Underwater Imaging System Design*. Defense Technical Information Center; 1972.
- McGlamery BL. *A computer model for underwater camera systems*. In *Ocean Optics VI*, volume 0208. SPIE; 1980.
- Jaffe JS. *Computer modeling and the design of optimal underwater imaging systems*. *IEEE J Oceanic Eng*. 1990;15(2):101–11.
- Huimin L, Li Y, Zhang Y, Chen M, Serikawa S, Kim H. *Underwater optical image processing: a comprehensive review*. *Mobile Networks and Applications*. 2017;22(6):1204–11.
- Xi Q, Rauschenbach T, Daoliang L. *Review of underwater machine vision technology and its applications*. *Mar Technol Soc J*. 2017;51(1):75–97.
- Jaffe JS. *Underwater optical imaging: The past, the present, and the prospects*. *IEEE J Oceanic Eng*. 2015;40(3):683–700.
- Bonin F, Burguera A, Oliver G. *Imaging systems for advanced underwater vehicles*. *Journal of Maritime Research*. 2011;8(1):65–86.
- Jaffe JS, Moore KD, McLean J, Strand MP. *Underwater optical imaging: Status and prospects*. *Oceanography*. 2001;14(3):64–75.
- Urlick RJ. *Principles of Underwater Sound*. New York, London: McGraw-Hill; 1983.
- Chi C, Li Z, Li Q. *Fast broadband beamforming using nonuniform fast fourier transform for underwater real-time 3-d acoustical imaging*. *IEEE J Oceanic Eng*. 2016;41(2):249–61.
- Albright Blomberg AE, Austeng A, Hansen RE, Synnes SAV. *Improving sonar performance in shallow water using adaptive beamforming*. *IEEE J Oceanic Eng*. 2013;38(2):297–307.
- Pearce SK, Bird JS. *Sharpening sidescan sonar images for shallow-water target and habitat classification with a vertically stacked array*. *IEEE J Oceanic Eng*. 2013;38(3):455–69.
- Chen P, Tian X, Chen Y. *Optimization of the digital near-field beamforming for underwater 3-d sonar imaging system*. *IEEE Trans Instrum Meas*. 2010;59(2):415–24.
- McHugh R, Shaw S, Taylor N. *A general purpose digital focused sonar beamformer*. In *Proceedings of OCEANS*, volume 1, pages I/229–I/234 vol.1, 1994.
- Thorner JE. *Approaches to sonar beamforming*. In *IEEE Technical Conference on Southern Tier*. 1990;69–78.
- Albert W. Cox. *Sonar and Underwater Sound*: Univ of Toronto Press; 1974.
- Saucan A, Sintés C, Chonavel T, Le Caillec J. *Model-based adaptive 3d sonar reconstruction in reverberating environments*. *IEEE Trans Image Process*. 2015;24(10):2928–40.
- Masmoudi A, Bellili F, Affes S, Stephenne A. *A maximum likelihood time delay estimator in a multipath environment using importance sampling*. *IEEE Trans Signal Process*. 2013;61(1):182–93.
- Saucan A, Sintés C, Chonavel T, Le Caillec J. *Enhanced sonar bathymetry tracking in multi-path environment*. In *Oceans*. 2012;1–8.
- Burkhardt E, Boebel O, Bornemann H, Ruholl C. *Risk assessment of scientific sonars*. *Bioacoustics*. 2008;17(1–3):235–7.
- Xavier L, Stacy D. *Sound radiation of seafloor-mapping echosounders in the water column, in relation to the risks posed to marine mammals*. *The International Hydrographic Review*. 2011;(6).
- Ellison WT, Southall BL, Clark CW, Frankel AS. *A new context-based approach to assess marine mammal*

- behavioral responses to anthropogenic sounds. *Conserv Biol.* 2012;26(1):21–8.
42. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436–44.
  43. Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press;2016.
  44. Sattar J, Dudek G. Where is your dive buddy: tracking humans underwater using spatio-temporal features. In *IEEE/RSJ International Conference on Intelligent Robots and Systems.* 2007;3654–3659.
  45. Sattar J, Dudek G. Underwater human-robot interaction via biological motion identification. In *Robotics: Science and Systems (RSS).* 2009;1–8.
  46. Buelow H, Birk A. Diver detection by motion-segmentation and shape-analysis from a moving vehicle. In *IEEE Oceans;* 2011.
  47. DeMarco KJ, West ME, Howard AM. Autonomous robot-diver assistance through joint intention theory. In *Oceans.* 2014;1–5. IEEE.
  48. Chavez AG, Pfingsthorn M, Birk A, Rendulic I, Miskovic N. Visual diver detection using multi-descriptor nearest-class-mean random forests in the context of underwater human robot interaction (hri). In *IEEE Oceans;* 2015.
  49. Islam MJ, Sattar J. Mixed-domain biological motion tracking for underwater human-robot interaction. In *IEEE International Conference on Robotics and Automation (ICRA).* 2017;4457–4464.
  50. Islam MJ, Ho M, Sattar J. Understanding human motion and gestures for underwater human-robot collaboration. *Journal of Field Robotics.* 2019;36(5):851–73. <https://doi.org/10.1002/rob.21837>.
  51. Xia Y, Sattar J. Visual diver recognition for underwater human-robot collaboration. In *International Conference on Robotics and Automation (ICRA).* 2019;6839–6845. IEEE.
  52. Islam MJ, Fulton M, Sattar J. Toward a generic diver-following algorithm: Balancing robustness and efficiency in deep visual detection. *IEEE Robotics and Automation Letters (RAL).* 2019;4(1):113–20.
  53. Langis K, Sattar J. Realtime multi-diver tracking and re-identification for underwater human-robot collaboration. In *IEEE International Conference on Robotics and Automation (ICRA).* 2020;11140–11146. IEEE.
  54. Chou HM, Chou YC, Chen HH. Development of a monocular vision deep learning-based auv diver-following control system. In *IEEE/MTS Global Oceans.* 2020;1–4. IEEE.
  55. Islam MJ, Edge C, Xiao Y, Luo P, Mehtaz M, Morse C, Enan SS, Sattar J. Semantic segmentation of underwater imagery: Dataset and benchmark. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* 2020;1769–1776.
  56. Chou YC, Chen HH, Wang CC, Chou HM. An ai auv enabling vision-based diver-following and obstacle avoidance with 3d-modeling dataset. In *IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS).* 2021;1–4, 2021.
  57. Agarwal T, Fulton M, Sattar J. Predicting the future motion of divers for enhanced underwater human-robot collaboration. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* 2021;5379–5386.
  58. Arturo Gomez Chavez, Christian A. Mueller, Andreas Birk, Anja Babic, and Nikola Miskovic. Stereo-vision based diver pose estimation using lstm recurrent neural networks for auv navigation guidance. In *IEEE Oceans.* IEEE press; 2017.
  59. Codd-Downey R, Jenkin M. Finding divers with scubanet. In *International Conference on Robotics and Automation (ICRA).* 2019;5746–5751.
  60. Nad D, Mandic F, Miskovic N. Using autonomous underwater vehicles for diver tracking and navigation aiding. *Journal of Marine Science and Engineering (JMSE).* 2020;8(6).
  61. Antervedi LGP, Chen Z, Anand H, Martin R, Arrowsmith R, Das J. Terrain-relative diver following with autonomous underwater vehicle for coral reef mapping. In *IEEE 17th International Conference on Automation Science and Engineering (CASE).* 2021;2307–2312. IEEE.
  62. Glotzbach T, Bayat M, Aguiar AP, Pascoal A. An underwater acoustic localisation system for assisted human diving operations. *IFAC Proceedings Volumes.* 2012;45(27):206–11.
  63. Goodfellow GM, Neasham JA, Rendulic I, Nad D, Miskovic N. DiverNet - a network of inertial sensors for real time diver visualization. In *IEEE Sensors Applications Symposium (SAS).* 2015;1–6.
  64. Neasham JA, Goodfellow G, Sharpshouse R. Development of the “seatrac” miniature acoustic modem and usbl positioning units for subsea robotics and diver applications. In *IEEE/MTS OCEANS.* 2015;1–8. IEEE.
  65. Miskovic N, Nad D, Rendulic I. Tracking divers: An autonomous marine surface vehicle to increase diver safety. *IEEE Robot Autom Mag.* 2015;22(3):72–84.
  66. DeMarco KJ, West ME, Howard AM. Sonar-based detection and tracking of a diver for underwater human-robot interaction scenarios. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC).* 2013;2378–2383. IEEE.
  67. Dula N, Christopher W, Igor K, Antillon DO, Miskovic N, Anderson I, Loncar I. Towards advancing diver-robot interaction capabilities. *IFAC-PapersOnLine.* 2019;52(21):199–204.
  68. Kvasic I, Miskovic N, Vukic Z. Convolutional neural network architectures for sonar-based diver detection and tracking. In *OCEANS.* 2016;1–6.
  69. ● Remmas W, Chemori A, Kruusmaa M. Diver tracking in open waters: A low-cost approach based on visual and acoustic sensor fusion. *Journal of Field Robotics,* 2020. **The article presents a very good example of the aspects of diver tracking as the very first step in U-HRI; it covers the use of both underwater vision and an acoustic method.**
  70. Streenan A, Du Toit NE. Diver relative auv navigation for joint human-robot operations. In *IEEE OCEANS.* 2013;1–10. IEEE.
  71. DeMarco KJ, West ME, Howard AM. A simulator for underwater human-robot interaction scenarios. In *OCEANS.* 2031;1–10. IEEE.
  72. Nad D, Mandic F, Miskovic N. Diver tracking using path stabilization - the virtual diver experimental results. *IFAC-PapersOn-Line.* 2016;49(23):214–9.
  73. Islam MJ, Hong J, Sattar J. Person-following by autonomous robots: A categorical overview. *The International Journal of Robotics Research (IJRR).* 2019;38(14):1581–1618.
  74. Buelow H, Birk A. Gesture-recognition as basis for a human robot interface (hri) on a auv. In *IEEE Oceans;* 2011.
  75. DeMarco KJ, West ME, Howard AM. Underwater human-robot communication: A case study with human divers. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC).* 2014;3738–3743. IEEE.
  76. Verzijlenberg B, Jenkin M. Swimming with robots: Human robot communication at depth. In *IEEE/RSJ International Conference on Intelligent Robots and Systems.* 2010;4023–4028. IEEE.
  77. Bernardi M, Cardia C, Gjanci P, Monterubbiano A, Petrioli C, Picari L, Spaccini D. The diver system: Multimedia communication and localization using underwater acoustic networks. In *20th International Symposium on “A World of Wireless, Mobile and Multimedia Networks” (WoWMoM).* 2019;1–8. IEEE.
  78. Riksfjord H, Haug OT, Hovem JM. Underwater acoustic networks - survey on communication challenges with transmission simulations. In *Sensor Technologies and Applications, 2009. SENSORCOMM '09. Third International Conference on.* 2009;300–305.
  79. Cui JH, Jiejun J, Gerla M, Zhou S. The challenges of building mobile underwater wireless networks for aquatic applications. *Network, IEEE.* 2006;20(3):12–8.

80. Sozer EM, Stojanovic M, Proakis JG. Underwater acoustic networks. *IEEE J Oceanic Eng.* 2000;25(1):72–83.
81. Dudek G, Giguere P, Prahacs C, Saunderson S, Sattar J, Torres-Mendez LA, Jenkin M, German A, Hogue A, Ripsman A, Zacher J, Milios E, Liu H, Zhang P, Buehler M, Georgiades C. Aqua: An amphibious autonomous robot. *IEEE Computer.* 2007;40(1):46–53.
82. Fulton M, Edge C, Sattar J. Robot communication via motion: Closing the underwater human-robot interaction loop. In *International Conference on Robotics and Automation (ICRA)*. 2019;4660–4666. IEEE.
83. • Fulton M, Edge C, Sattar J. Robot communication via motion: A study on modalities for robot-to-human communication in the field. *ACM Transactions on Human-Robot Interaction.* 2022;11(2):Article 15. **This article presents the use of robot motion for robot-to-diver communication, which is a conceptually more advanced option than just using, e.g., blinking lights or displays for the communication from the underwater robot to a human.**
84. Birdwhistell RL. Introduction to kinesics: An annotation system for analysis of body motion and gesture. University of Michigan Library; 1952.
85. Danesi M. *Kinesics*. 2006;207–213. Elsevier, Oxford.
86. Cheng E, Huang J. Application of speech recognition and synthesis on underwater acoustic speech transmission. In *International Conference on Neural Networks and Signal Processing*, volume 2, pages 876–878 Vol.2; 2003.
87. Wisch TO, Schmidt G. Mixed analog-digital speech communication for underwater applications. In *Speech Communication; 14th ITG Conference*. 2021;1–5.
88. Dudek G, Sattar J, Xu A. A visual language for robot control and programming: A human-interface study. In *IEEE International Conference on Robotics and Automation (ICRA)*. 2007;2507–2513.
89. • Chavez AG, Ranieri A, Chiarella D, Birk A. Underwater vision-based gesture recognition - a robustness validation for safe human-robot interaction. *IEEE Robotics and Automation Magazine (RAM)*. 2021;(3):67–78. **The article presents a very good example of diver-to-robot communication; it evaluates deep learning methods for an underwater gesture recognition front-end and an efficient back-end for handling the language interpretation tested in field-trials.**
90. Speers A, Forooshani PM, Dicke M, Jenkin M. Lightweight tablet devices for command and control of ros-enabled robots. In *16th International Conference on Advanced Robotics (ICAR)*. 2013;1–6.
91. Chavez AG, Birk A. Underwater gesture recognition based on multi-descriptor random forests (md-ncmf); 2015.
92. Gustin F, Rendulic I, Miskovic N, Vukic Z. Hand gesture recognition from multibeam sonar imagery. In Vahid Hassan, editor, *10th IFAC Conference on Control Applications in Marine Systems (CAMS)*, volume 49, pages 470–475. *IFAC PapersOnLine*; 2016.
93. Yang J, Wilson JP, Gupta S. Diver gesture recognition using deep learning for underwater human-robot interaction. In *MTS/IEEE OCEANS SEATTLE*. 2019;1–5, 2019.
94. Codd-Downey R, Jenkin M. Human robot interaction using diver hand signals. In *14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2019;550–551. IEEE.
95. Wang Z, She Q, Ward TE. Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Computing Surveys.* 2021;54(2):Article 37.
96. Chavez AG, Ranieri A, Chiarella D, Zereik E, Babic A, Birk A. Caddy underwater stereo-vision dataset for human-robot interaction (hri) in the context of diver activities. *Journal of Marine Science and Engineering (JMSE), spec.iss. Underwater Imaging.* 2019;7(1).
97. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC. Ssd: Single shot multibox detector. In Leibe B, Matas J, Sebe N, Welling M, editors, *Computer Vision – ECCV 2016*. 2016;21–37, Cham. Springer International Publishing.
98. Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans Pattern Analysis Machine Intelligence.* 2017;39(6):1137–49.
99. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng.* 2010;22(10):1345–59.
100. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks; 2012.
101. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition; 2015.
102. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016;770–778.
103. Szegedy C, Wei L, Yangqing J, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015;1–9.
104. Jiang Y, Zhao M, Wang C, Wei F, Wang K, Qi H. Diver's hand gesture recognition and segmentation for human-robot interaction on auv. *Signal, Image and Video Processing (SIVIP)*. 2021;15(8):1899–1906.
105. K. He, G. Gkioxari, P. Dollár, Girshick R. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*. 2017;2980–2988.
106. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L. MobileNetV2: Inverted residuals and linear bottlenecks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018;4510–4520.
107. 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.
108. Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, Wei Y. Deformable convolutional networks. *CoRR*, abs/1703.06211. 2017.
109. Antillon DO, Walker C, Rosset S, Anderson I. The challenges of hand gesture recognition using dielectric elastomer sensors, volume 11375 of *SPIE Smart Structures + Nondestructive Evaluation*. SPIE; 2020.
110. Nad D, Ferreira F, Kvasic I, Mandic L, Walker C, Antillon DO, Anderson I. Diver-robot communication using wearable sensing diver glove. In *OCEANS*. 2021;1–6. IEEE.
111. Walker C, Anderson I. From land to water: bringing dielectric elastomer sensing to the underwater realm, volume 9798 of *SPIE Smart Structures and Materials + Nondestructive Evaluation and Health Monitoring*. SPIE, 2016.
112. Walker C, Anderson I. Monitoring diver kinematics with dielectric elastomer sensors, volume 10163 of *SPIE Smart Structures and Materials + Nondestructive Evaluation and Health Monitoring*. SPIE, 2017.
113. Xu A, Dudek G, Sattar J. A natural gesture interface for operating robotic systems. In *IEEE International Conference on Robotics and Automation (ICRA)*. 2008;3557–3563.
114. Chiarella D, Bibuli M, Bruzzone G, Caccia M, Ranieri A, Zereik E, Marconi L, Cutugno P. Gesture-based language for diver-robot underwater interaction. In *OCEANS 2015 - Genova*. 2015;1–9.
115. Chiarella D, Bibuli M, Bruzzone G, Caccia M, Ranieri A, Zereik E, Marconi L, Cutugno P. A novel gesture-based language for underwater human-robot interaction. *Journal of Marine Science and Engineering*. 2018;6(3).

116. Islam MJ, Ho M, Sattar J. Dynamic reconfiguration of mission parameters in underwater human-robot collaboration. In IEEE International Conference on Robotics and Automation (ICRA). 2018;6212–6219.
117. Langis KD, Fulton M, Sattar J. Towards robust visual diver detection onboard autonomous underwater robots: Assessing the effects of models and data. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2021;5372–5378.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.