2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Diver Interest via Pointing: Human-Directed Object Inspection for AUVs*

Chelsey Edge[1] and Junaed Sattar[2]

***Abstract*— In this paper, we present the Diver Interest via Pointing (DIP) algorithm, a highly modular method for conveying a diver's area of interest to an autonomous underwater vehicle (AUV) using pointing gestures for underwater human-robot collaborative tasks. DIP uses a single monocular camera and exploits human body pose, even with complete dive gear, to extract underwater human pointing gesture poses and their directions. By extracting 2D scene geometry based on the human body pose and density of salient feature points along the direction of pointing, using a low-level feature detector, the DIP algorithm is able to locate objects of interest as indicated by the diver. DIP makes it possible for scuba divers and swimmers to use directional cues, through pointing, to an AUV for inspection, surveillance, manipulation, and navigation. We examine the elements that make up our method, provide quantitative and qualitative evaluation, and demonstrate AUV actuation based on diver pointing gestures in closed-water human-robot collaborative experiments. Our evaluations demonstrate the high efficacy of the DIP algorithm in correctly identifying the direction of a pointing gesture and locating an object within that region of interest. We also show that the findings of the algorithm qualitatively conform with human assessment of pointing gestures, directions, and targets.**

I. INTRODUCTION

The use of autonomous underwater vehicles (AUVs) to perform tasks underwater has become increasingly relevant in recent years. From inspection and maintenance of underwater infrastructure [1], biological monitoring [2], and archaeology [3], the utility of these AUVs heavily relies on working and cooperating with human divers. For each of the above-mentioned scenarios, it is highly likely that a robot will be directed to specifically navigate along a particular direction or inspect an object; this is referred to as the site acquisition and scene re-inspection (SASR) task. Such tasks could include inspecting a specified region of a pipeline, taking pictures of particular coral for a conservation biologist, or recovering an artifact.

While Remotely Operated Vehicles (ROVs) are able to receive instruction through a tether, AUVs require different communication methods [4]. These methods include using robotic vision or audio to receive directives from divers through the use of tools such as fiducial markers [5], or special dive gloves [6], [7] as well as with gesture-based languages of hand signals with ([8], [9]) and without ([10], [11]) gloves. As many divers working alongside AUVs have their own areas of expertise and may not be robotics experts, a communication vector that is easily understood and capable of being performed naturally is essential. For this reason, we introduce a system for divers that uses pose-based pointing gestures which do not require additional equipment to instigate a response from an AUV towards an object or location of interest (Fig. 1).

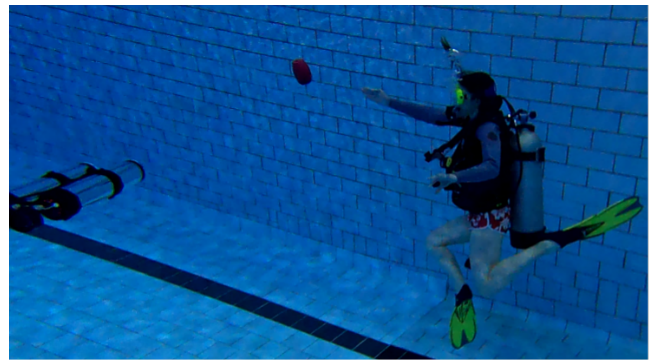


Fig. 1: LoCO AUV running DIP to navigate towards an object in response to diver pointing gesture.

Pointing is used as a natural form of human communication to share location and interest. Expanding this communication vector to robotics is not itself an unexplored area of interest (see Sec. II). In the subtopic of marine robotics, however, the use and even exploration of this topic is almost non-existent (other than very few exceptions, *e.g.*, [12] [9]). Different methods of indicating interest are often used by scuba divers, *e.g.*, shining lights, carrying physical objects such as sticks or poles, and dropping strobes or beacons. However, these methods come with significant limitations. The underwater domain is known for its high degree of sensory signal attenuation and dispersion, and thus methods such as shining lights on objects of interest may not be precise enough for many needs. Without necessitating extra tools, pointing gesture communication relies on some form of visual perception. Underwater vision is challenging due to natural environmental factors (*e.g.*, scattering, absorption, and refraction of light among others). Image enhancement techniques (*e.g.*, [13], [14]) have been shown to mitigate some of these factors and improve the ability to use RGB cameras. These factors still significantly impact the *reliability* of commonly used terrestrial sensors, such as light-wave based Time-of-Flight RGB-D cameras, used in many pointing gesture based systems ([15], [16], [17], [18], [19]). As vision itself is challenging even in the best of water conditions, adaptation of ideas and implementation

The authors are with the Department of Computer Science and Engineering (CSE), Minnesota Robotics Institute (MnRI), University of Minnesota Twin Cities, Minneapolis, MN, USA. {[1]edge0037,[2]junaed}@umn.edu

*This work was supported by the US National Science Foundation Award IIS-#1637875, and the MnRI Seed Grant.

of terrestrially designed algorithms can be difficult.

Also unique to the underwater environment is the high probability that the diver, robot, and even object of interest may have a free-floating ability. For example, during terrestrial and pick-and-place applications, it can generally be assumed that anything static on a ground or table environment will remain there until moved deliberately. Even with relation to drones, the human is potentially able to remain on the ground plane and calculations for direction can take advantage of this. In open water, however, it is highly unlikely that the diver will remain on the floor of the body of water; in fact, they may even be moving along with water currents. Our approach takes this into account by leveraging only the pose of the diver and robot view space to find the area of interest.

In this paper, we introduce **DIP** (*Diver Interest via Pointing*), a method to inform an AUV of the location of an unknown object via a pointing gesture. We show that the use of a human pose estimator in conjunction with a single camera is sufficient to create an area of interest from which an AUV can detect an object of interest. In addition to locating the area of interest as indicated by a diver in still images, we provide a working example of how our method can be integrated into an object inspection system, utilizing the LoCO AUV [20].

II. Related Work

In this section, as the use of pointing control with AUVs is a novel idea, we focus our attention on the use of pointing gestures within the scope of robotics in general. DIP requires the use of human pose estimation, and also fits within the broader definition of human-robot interaction (HRI); as such, we also provide some context on the current state of gestural communication with AUVs.

A. Gestural Communication with AUVs

Using gestures to interface with AUVs, while relatively under-explored in comparison with aerial and terrestrial vehicles, is beneficial for divers as it requires no extra equipment. Currently deployed systems ([10], [11], [8], [9]) have been designed with ease of use of divers in mind, including some with directional signals (*e.g.*, [9]). Expanding this interface to convey locational cues can be seen as a next step. While Walker [12] provides an analysis of the ability to recognize pointing gestures underwater on state-of-the-art deep-learned object detectors (*e.g.*, Single Shot Multibox Detector (SSD) [21], Faster R-CNN [22]), there is no current literature on directing an AUV to a location or object of interest through pointing gestures.

We contribute the first look at using pointing to provide specific locational cues of a diver's area of interest to an AUV, and exploit these cues for the purpose of AUV-assisted task completion in that location.

B. Applications of Human Pose Estimation

Human pose estimators are often used to provide landmarks for pointing gestures ([17], [23], [19], [24]). The pose estimator locates landmarks, typically joints, on a human body. While human pose estimation in both 2D and 3D is widely studied (see [25], [26], [27]), there does not exist a vision-based human pose estimator dedicated for underwater use. Chavez *et al.* [28] tracks a diver using point clouds of diver pose; however, specific pose landmarks are not found. Terrestrial-based, out-of-the-box monocular 2D pose estimators have been applied with some success; *e.g.*, Islam *et al.* [29] use estimated human body pose (using OpenPose [30]) to find relative positions of two robots, and Fulton *et al.* [31] create an autonomous diver approach method using TRT pose [32]. The success of human pose estimators is severely impacted not only by underwater vision degradation, but also by the positions of the human body with relation to training data and additional gear worn by scuba divers (Fig. 2). We choose to incorporate the Mediapipe [33] Pose [34] framework as the foundation of DIP, because of the cited low latency and impressive results on physical activities like yoga and dance, which also deal heavily with uncommon human poses.

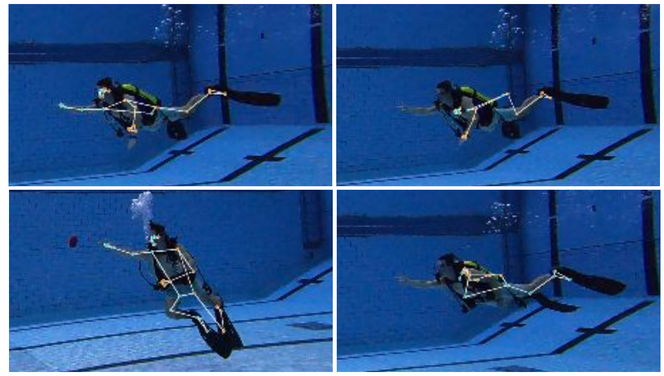


Fig. 2: Example of some failure cases in underwater human pose estimation. Even in closed-water conditions, a combination of environmental factors, non-standard body pose, and SCUBA attire can affect robust 'terrestrial' vision algorithms.

C. Pointing for Purpose

In terrestrial and aerial robotics, the use of pointing gestures to relay information has been widely studied. For example, terrestrial and aerial vehicles can be given direction to park in a certain location determined by a pointing pose estimate [17], [23], [19]. For these applications, the user is able to point at a location on the ground plane and direct the robot there. This is an impossibility in many underwater robotics applications as touching the ground plane, if it exists in view, goes against the task that is being accomplished (*e.g.*, coral reef inspection, or invasive species detection). Aerial and terrestrial robotics have also seen application of finding an object of interest through pointing gestures. Pick-and-place tasks with robotic arms have seen success through hand gesture pointing [16], [35]. Großmann *et al.* [15] use a pointing gesture to indicate an object on a shelf that should be of interest to a mobile manipulator. Medeiros *et al.* [24] is able to direct a drone towards an object based on the

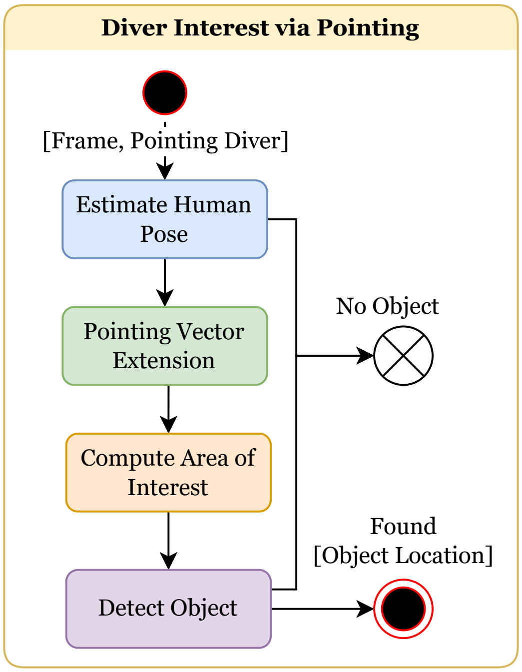


Fig. 3: Schematic diagram of the DIP algorithm. Colors denoting each step are consistent with those in Figs. 8 and 9.

intersection of a pointing direction of the user's arm and an object's bounding box. Delmerico *et al.* [36] provide a survey of many more uses for pointing gestures with an emphasis on the uses with rescue robotics. However, no work exists for the use of pose-based or locational pointing gestures for control or communicating user interest in the underwater environment.

III. Diver Interest via Pointing

The DIP approach is composed of a human pose estimator, a method for the AUV to predict a diver's area of interest, and an object detection algorithm applied to the predicted area of interest (Fig. 3). In the following section we describe the method to detect a diver's pointing pose, and using that, detail the creation of the area of interest. In addition, we enumerate the design choices that allow the implementation of the proposed algorithm on-board a physical robot.

For the purposes of this section we make the following assumptions:

- The diver is situated in an upright, pointing pose.
- The diver is pointing with their right arm.

A. Diver Pose Estimation

Vision-based AUV communication using pointing gestures from divers requires robust estimation of human pose. Human pose estimation underwater is challenging for a variety of reasons, including naturally degraded vision, non-standard body positions and the wearing of SCUBA gear by divers (see Fig. 2). To minimize the need for correctly positioned body landmarks, we use only two, the elbow and wrist. As the connection between these two points must be linear (*i.e.*, anatomically, there are no joints between them in human physiology), this creates the ability to generalize potential locations of interest in a rather straightforward manner. By simply extending the line segment connecting the elbow to the wrist, the general pointing direction can be found. The shoulder is excluded as, unless the diver's body is turned towards the camera, landmark detection can be difficult (Fig. 2). We explicitly avoid the use of the hands or fingers for finding the pointing direction, as the poor visibility underwater, particularly at the usual interaction distance between an AUV and a diver, can make it difficult for pose estimators designed for terrestrial use to identify those body parts. For the purpose of our project, we use the Mediapipe Pose [33] estimator, although any pose estimator which can provide wrist $w_{(x,y)}$ and elbow $e_{(x,y)}$ landmarks in the two-dimensional image space can be used. Once the pose with the wrist and elbow landmarks have been located, we are able to determine a direction and local area of interest.

The use of 2D pose for determining pointing direction for SASR type tasks is sufficient, as objects of interest will often be close to the diver in distance, thus we can approximate the 3D directional vector in 2D image space. As marine environments are also generally uncluttered and without many structures, the need for depth information about the scene is not necessary for many tasks. Particularly when coupled with a robust object detector, 2D pointing vectors will often suffice for such purposes. Future work looks at expanding DIP to include 3D information for use in specialized situations.

B. Determining the Area of Interest

The area of interest to the robot is defined to be a triangular region extending from the diver's wrist, shown in Fig. 4 as the white triangle. A triangular shape is chosen, as the farther away a diver is from an object, the less accurate the actual gesture may be (*i.e.*, the diver's own gesture pointing at the object of interest will be less accurate from farther away). This error can potentially be magnified in relation to above-water scenarios, as the object, diver, and robot itself may not be able to remain stationary for a variety of reasons, such as being suspended in the water, moving with a current, or constantly maneuvering to hold position. Choosing a triangular region to search makes it possible to account for such inherent inaccuracies in pointing and have a more accurate detection of the object of interest.

First, the line segment connecting the elbow and wrist landmarks (as defined by the human pose estimator) is extended $ext_{(x,y)}$ by a scaling factor sf which can be modified as needed (Eq. 1):

$$\begin{aligned} ext_{(x)} &= w_{(x)} + sf * (w_{(x)} - e_{(x)}) \\ ext_{(y)} &= w_{(y)} + sf * (w_{(y)} - e_{(y)}), \end{aligned} \tag{1}$$

A scale factor of 10 has empirically been found to be sufficient in our investigations.

Once the segment is extended, the resulting point $ext_{(x,y)}$ is defined as the end of the pointing vector. To create an area of interest in the shape of a triangle in image space, two vertices are defined by adding and subtracting a *vertical*

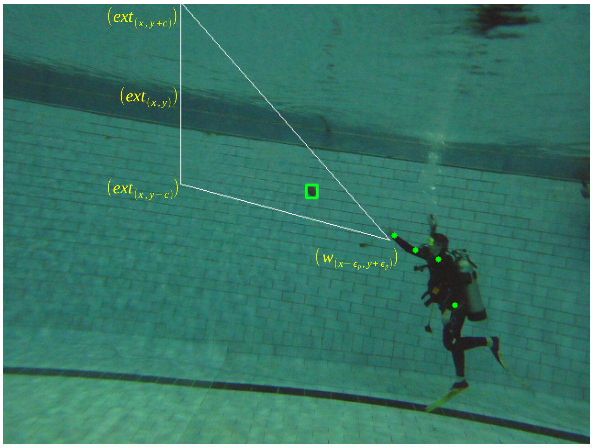


Fig. 4: Visualization of the DIP algorithm: Human pose landmarks are located, an area of interest is created based on pose pointing vector, and an object is located within the area of interest.

constant c, to the y value of the point extension (*e.g.*, $ext_{(x,y\pm c)}$) (see Fig. 4 for a visualization).

The third vertex of the triangle is defined by the wrist landmark of the human pose detector. This vertex may also be offset by a small amount ϵ_p to reduce false positives of hand detection to compensate for detector errors. The vertices of the triangle are therefore defined as: $(w_{(x-\epsilon_p,y+\epsilon_p)}), (ext_{(x,y\pm c)})$.

Once the area of interest is computed, search for the object of interest via object detection is confined to this area.

C. Detecting the Object of Interest

As the main goal is to identify the diver's intent to highlight an area of interest, the consequent object detection step is highly dependent on the actual task given to the AUV and thus should be changed accordingly, *i.e.*, a trained trash detector should be used if the AUV's task is to locate trash in an area of interest. However, even without knowing the object type ahead of time, it is possible to exploit low-level image features to identify objects pointed at by divers within the triangular area of interest. For example, point features (*e.g.*, SIFT [37], ORB [38]) and edge detectors (*e.g.*, Canny [39]) can be used to identify regions with high probability of being objects of interest. The motivation to use such low-level feature extractors stems from the fact that the underwater environment often has little background variation, and only salient features will be on or in close proximity to the object. Due to the challenges of detection in an underwater environment, we demonstrate detection through two methods: keypoint and contour detection. Via keypoint detection, we choose the *keypoint with the greatest strength* to represent the object, even if this has the potential for occasional false positive detections. The Canny edge detector is used to extract object contours. If the centroid of the contoured region is within the triangle of interest, the target is chosen as the object of interest. Fig. 4 shows an example of DIP successfully finding the object pointed to by the diver.

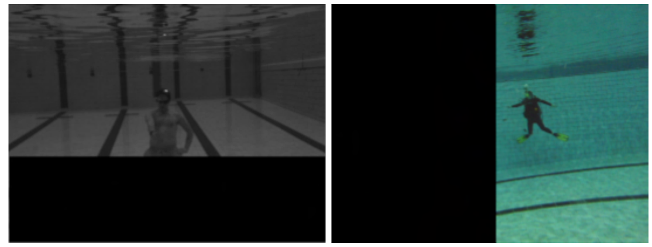


Fig. 5: Two of the eight images the human correlation study participants were asked to annotate. The locations of the target objects are blackened out to not bias the participants' assessment of the pointing direction.

IV. Evaluation

Obtaining ground truth is difficult in underwater environments as there is constant motion between diver, robot, and objects. In addition, accuracy of DIP as a whole rather than the accuracy of the pose and object point location is a necessity; therefore, for the following evaluations, visual cues rather than straightforward landmark matching will be used in assessment.

All evaluations are performed using the same parameters with input image sizes of 640×480 pixels. We use an empirically found vertical constant of 100. In other words, the height of the triangular area of interest will be 200 pixels, just under half the image height. We also offset the wrist vertex by 5 pixels (*i.e.*, $\epsilon_p = 5$, see Sec. III-B) to reduce false positives of detecting the hand. The vertices of the triangular area of interest therefore become $(w_{(x-5,y+5)}), (ext_{(x,y\pm 100)})$.

A. DIP and Human-based Pointing Correlation

We present a small human study in which participants annotate a pointing vector on 8 images. Each image consists of a submerged diver pointing towards an object, either suspended in the water, or resting on the bottom. The general location of an object of interest was removed (Fig. 5) in order to prevent preconceptions of what the diver is pointing towards. Nine participants annotated a line segment of the assumed pointing vector, beginning at the diver's wrist. Fig. 6 shows a compilation of the results: the green line segments represent the human-annotated vectors, the pink segment represents DIP's foundational pointing vector, and the white triangle represents the location of diver interest as produced by DIP. As can be seen from the quantitative data in Table I, where image numbers match the vector images, fitting a 2D vector to a point can be highly variable even by human annotators. Angles are not included for DIP where pose estimation is incorrect.

Fig. 6 also conveys some of the challenging situations that can occur within a highly-specified underwater setting. As our method is dependent on human pose, if the pose captured is incorrectly such as in Image 2, the area of interest will be so as well. Failure cases were found to occur frequently when the diver's pointing arm intersects with the torso, as seen in diver center and left-pointing evaluation cases (See Images 5, 7 in Fig. 6).

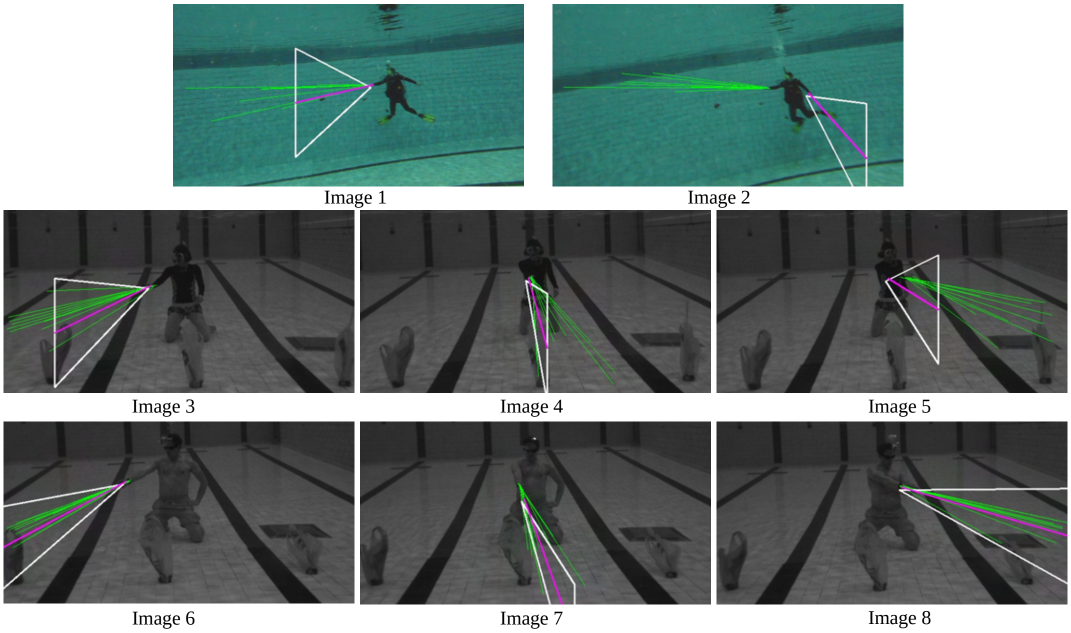


Fig. 6: Results of human-annotated vectors (green) shown with the pose estimator-based resultant vector (pink), and DIP area of interest (white triangle). DIP's output predominantly aligns with human assessment of pointing directions.

	Mean (Human)	DIP	Difference	Variance (Human)
Image 1	3.059 rad	2.912 rad	0.147 rad	0.008
Image 2	3.198 rad	—	—	0.003
Image 3	2.849 rad	2.677 rad	0.121 rad	0.017
Image 4	1.097 rad	1.309 rad	0.211 rad	0.069
Image 5	0.357 rad	—	—	0.024
Image 6	2.717 rad	2.643 rad	0.074 rad	0.004
Image 7	1.249 rad	—	—	0.028
Image 8	0.286 rad	0.288 rad	0.002 rad	0.008

TABLE I: Results of human study (Fig. 6): Mean human-annotated angle measure, DIP angle measure (if correct pose), Difference in Mean and DIP angles, and variance in human-annotated responses are recorded. All angles are in radians.

B. DIP-based Pose and Object Detection

We evaluate DIP for the ability to use human pose estimation and low-level feature detectors to first compute a diver's specified region of interest and then locate an unknown object within that region. Due to the challenging environment, we consider the ability to identify an accurate region of interest more beneficial than exact landmark matching. In regard to unknown object detection, we provide results for both the SIFT [37] algorithm and Canny [39] edge detection as defined in Sec. III-C.

A selection of 650 frames were taken from a ROS [40] bag file recorded with the LoCO AUV [20] in a closed-water environment (Fig. 7). All images in sequence are included unless the entirety of the diver's body is not in the scene. Fig. 8 shows results of a visual examination of this dataset. While not all images may be considered "good quality" by human standards, we include these images to demonstrate the effectiveness of DIP from the AUV viewpoint. The evaluation dataset can be considered to represent fairly optimal conditions in terms of water quality and distance from the diver.

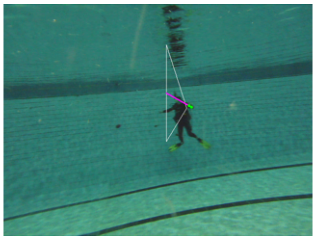
Taking into account only those images that produce a "correct" pose (349/650 images), we see that the object of interest is included in the region of interest 345 out of 349 times or 98.85%. On the other hand, the pointing vector itself lands on the object only 74 times, or 21.2%. As supported by the results of Sec. IV-A, due to variations in pointing, locating an area of interest produces better detection results than choosing a single vector. As to the evaluation of object detection, the inclusion of a good specified object detector is essential for each task. However, due to the nature of the underwater environment, objects of interest are able to be located far better than random chance with low-level features, with the SIFT algorithm failing to locate a potential object of interest in only 13 out of 345 images.

C. Runtime

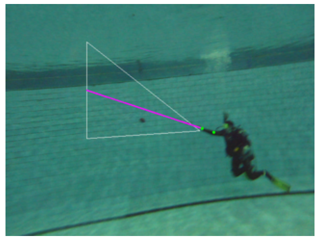
With Mediapipe Pose as a backbone, the DIP algorithm (pose estimation and SASR creation) runs at 4.714 fps on an Intel^{TM} i7-6700 CPU, which is acceptable for AUV operations.

V. Experimental Implementation onboard the LoCO AUV

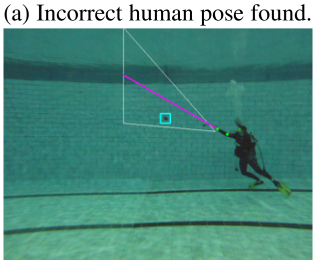
DIP works as the guiding force for an unknown object investigation task (Fig. 9). We deploy DIP, along with the rest of the system, entirely onboard the LoCO AUV, use a monocular RGB camera, and perform computations on an NVIDIA Jetson TX2 embedded system. In an enclosed pool environment, LoCO was required to detect that a diver is pointing, employ DIP to determine where an object may be located and locate an object, then finally actuate towards the object for closer inspection.



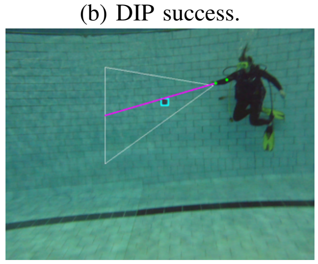
(a) Incorrect human pose found.



(b) DIP success.



(c) Unknown object located during DIP through the SIFT algorithm.



(d) Unknown object located during DIP using Canny edge detection.

Fig. 7: Samples from the pose and object detection evaluation image set. All images are sourced from the LoCO AUV and may not be considered good quality to human viewers.

Within this system, a diver is determined to be pointing through a pre-trained SSD [21] detector with a VGG-16 [41] backbone [12]. Once the pointing diver is detected, DIP proceeds to check incoming frames until the pose is detected. In the same frame the pose is detected, object detection is attempted. This cycle proceeds on successive frames until an object is located. The diver's interest and the object within the location are then considered confirmed, and the algorithm moves to initiate robot locomotion. The AUV begins moving towards the object through a Proportional-Integral-Derivative (PID) approach controller defined through a bounding box and image size ratio. We use the CAMShift tracker [42] to return a new location for the PID controller to continue AUV movement towards the object as it might be subjected to unintended motion underwater. Fig. 10 shows snapshots from LoCO's viewpoint during a successful trial. Fig. 9 diagrams the system as a whole as it runs onboard the LoCO AUV.

Out of five trial runs, the area of interest is defined (*e.g.*, the human pose is correct) four times. An object was located

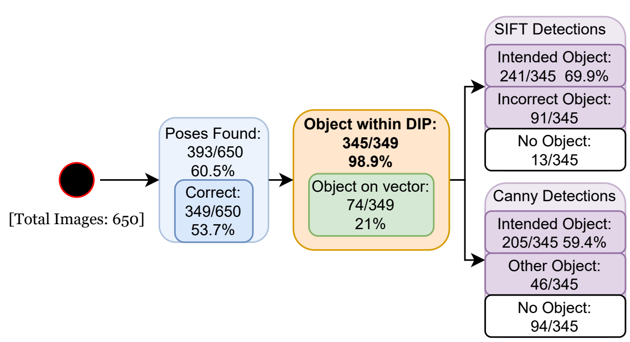


Fig. 8: Evaluation of DIP pose estimation and object detection. The ratios presented depend on success of the previous steps. Colors denote the step evaluated as in Fig. 3.

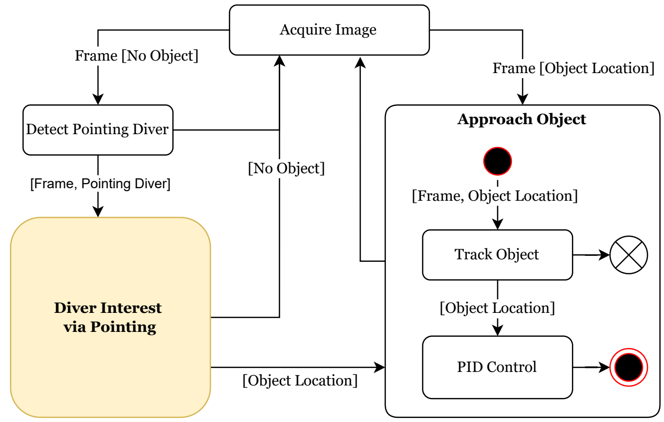
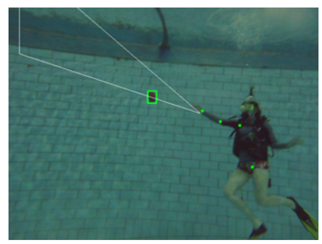
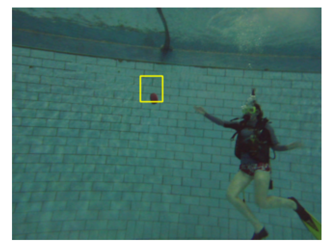


Fig. 9: Diagram showing the full system of Unknown Object Investigation guided by the DIP algorithm, as implemented on-board the LoCO AUV.



(a) An object is identified through the DIP algorithm.



(b) The object is tracked in the following frame.

Fig. 10: Demonstration of the task: Unknown Object Investigation. Images are from the LoCO AUV point of view.

each time and LoCO proceeded to move autonomously towards the perceived object, making task execution successful. Three out of four times, however, markings on the pool wall were located inside the area of interest. These markings were considered to be the object and LoCO moved in that direction. Using a detector trained for specific objects of interest would help mitigate this issue.

VI. Conclusion

We present a novel communication algorithm, Diver Interest via Pointing (DIP), to signal directional intent to an AUV. By detecting the natural communication vector of pointing, we show that an AUV is able to infer a diver's area of interest. While working in the challenging underwater environment, we have shown that a triangular area of interest can be found with the use of a monocular camera and out-of-the-box human pose estimator. Knowing that the area of interest exists, an AUV is able to perform tasks specifically within that area as determined by a diver. One such task is the inspection of an unknown object. We validate DIP through an integrated system and demonstrate its feasibility onboard the LoCO AUV. Future work will continue to improve directional gesture communication through the use of 3D scene geometry, pointing gesture classification for downstream robot tasks such as manipulation and data collection, and design of robot-to-human feedback to reduce ambiguity in determining objects of interest.

References

[1] Y. R. Petillot, S. R. Reed, and J. M. Bell, "Real Time AUV Pipeline Detection and Tracking Using Side Scan Sonar and Multi-beam Echo-sounder," in *OCEANS '02 MTS/IEEE*, vol. 1, 2002, pp. 217–222 vol.1.

[2] M. Modasshir, S. Rahman, O. Youngquist, and I. Rekleitis, "Coral Identification and Counting with an Autonomous Underwater Vehicle," in *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Kuala Lumpur, Malaysia, (Finalist of T. J. Tarn Best Paper in Robotics), Dec. 2018, pp. 524– 529.

[3] B. Bingham, B. Foley, H. Singh, R. Camilli, K. Delaporta, R. Eustice, A. Mallios, D. Mindell, C. Roman, and D. Sakellariou, "Robotic Tools for Deep Water Archaeology: Surveying an Ancient Shipwreck with an Autonomous Underwater Vehicle," *Journal of Field Robotics*, vol. 27, pp. 702–717, 11 2010.

[4] A. Birk, "A Survey of Underwater Human-Robot Interaction (U-HRI)," *Current Robotics Reports*, vol. 3, no. 4, pp. 199–211, Dec 2022. [Online]. Available: https://doi.org/10.1007/s43154-022-00092-7

[5] J. Sattar, E. Bourque, P. Giguère, and G. Dudek, "Fourier tags: Smoothly degradable fiducial markers for use in human-robot interaction," in *Proceedings of the Fourth Canadian Conference on Computer and Robot Vision*, Montréal, QC, Canada, 5 2007, pp. 165–174.

[6] A. Gomez Chavez, A. Ranieri, D. Chiarella, E. Zereik, A. Babić, and A. Birk, "CADDY Underwater Stereo-Vision Dataset for Human–Robot Interaction (HRI) in the Context of Diver Activities," *Journal of Marine Science and Engineering*, vol. 7, no. 1, 2019. [Online]. Available: https://www.mdpi.com/2077-1312/7/1/16

[7] D. Nad, F. Ferreira, I. Kvasic, L. Mandic, C. Walker, D. O. Antillon, and I. Anderson, "Diver-Robot Communication using Wearable Sensing Diver Glove," in *OCEANS 2021: San Diego – Porto*, 2021, pp. 1–6.

[8] A. Gomez Chavez, A. Ranieri, D. Chiarella, and A. Birk, "Underwater Vision-Based Gesture Recognition: A Robustness Validation for Safe Human–Robot Interaction," *IEEE Robotics & Automation Magazine*, vol. 28, no. 3, pp. 67–78, 2021.

[9] D. Chiarella, M. Bibuli, G. Bruzzone, M. Caccia, A. Ranieri, E. Zereik, L. Marconi, and P. Cutugno, "A Novel Gesture-Based Language for Underwater Human–Robot Interaction," *Journal of Marine Science and Engineering*, vol. 6, no. 3, 2018. [Online]. Available: https://www.mdpi.com/2077-1312/6/3/91

[10] M. J. Islam, M. Ho, and J. Sattar, "Dynamic Reconfiguration of Mission Parameters in Underwater Human-Robot Collaboration," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 6212–6219.

[11] ——, "Understanding Human Motion and Gestures for Underwater Human-Robot Collaboration," *Journal of Field Robotics (JFR)*, pp. 1–23, 2018.

[12] A. M. Walker, "Towards Natural Underwater Human-Robot Interaction: Pointing Gesture Recognition for Autonomous Underwater Vehicles," Master's thesis, University of Minnesota, 2021.

[13] M. J. Islam, P. Luo, and J. Sattar, "Simultaneous Enhancement and Super-Resolution of Underwater Imagery for Improved Visual Perception," in *Proceedings of Robotics: Science and Systems*, Corvalis, Oregon, USA, July 2020.

[14] M. Islam, R. Wang, and J. Sattar, "SVAM: Saliency-guided Visual Attention Modeling by Autonomous Underwater Robot," in *Proceedings of Robotics: Science and Systems*, New York City, NY, USA, June 2022.

[15] B. Großmann, M. R. Pedersen, J. Klonovs, D. Herzog, L. Nalpantidis, and V. Krüger, "Communicating Unknown Objects to Robots through Pointing Gestures," in *Advances in Autonomous Robotics Systems*, M. Mistry, A. Leonardis, M. Witkowski, and C. Melhuish, Eds. Cham: Springer International Publishing, 2014, pp. 209–220.

[16] D. Shukla, O. Erkent, and J. Piater, "Probabilistic Detection of Pointing Directions for Human-Robot Interaction," in *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Nov 2015, p. 1–8.

[17] M. Tölgyessy, M. Dekan, F. Duchoň, J. Rodina, P. Hubinský, and L. Chovanec, "Foundations of Visual Linear Human–Robot Interaction via Pointing Gesture Navigation," *International Journal of Social Robotics*, vol. 9, no. 4, p. 509–523, Sep 2017. [Online]. Available: https://doi.org/10.1007/s12369-017-0408-9

[18] D. Droeschel, J. Stückler, and S. Behnke, "Learning to Interpret Pointing Gestures with a Time-of-Flight Camera," in *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2011, pp. 481–488.

[19] S. Abidi, M. Williams, and B. Johnston, "Human Pointing as a Robot Directive," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2013, pp. 67–68.

[20] C. Edge, S. S. Enan, M. Fulton, J. Hong, J. Mo, K. Barthelemy, H. Bashaw, B. Kallevig, C. Knutson, K. Orpen, and J. Sattar, "Design and Experiments with LoCO AUV: A Low Cost Open-Source Autonomous Underwater Vehicle," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems.*, Las Vegas, Nevada, USA, October 2020, pp. 1761–1768.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 91–99.

[23] J. Richarz, A. Scheidig, C. Martin, S. Müller, and H.-M. Gross, "A Monocular Pointing Pose Estimator for Gestural Instruction of a Mobile Robot," *International Journal of Advanced Robotic Systems*, vol. 4, no. 1, p. 17, 2007. [Online]. Available: https://doi.org/10.5772/5700

[24] A. C. S. Medeiros, P. Ratsamee, J. Orlosky, Y. Uranishi, M. Higashida, and H. Takemura, "3D Pointing Gestures as Target Selection Tools: Guiding Monocular UAVs During Window Selection in an Outdoor Environment," *ROBOMECH Journal*, vol. 8, no. 1, Apr. 2021. [Online]. Available: https://doi.org/10.1186/s40648-021-00200-w

[25] Y. Chen, Y. Tian, and M. He, "Monocular Human Pose Estimation: A Survey of Deep Learning-based Methods," *Computer Vision and Image Understanding*, vol. 192, p. 102897, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314219301778

[26] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D Human Pose Estimation: New Benchmark and State of the Art Analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[27] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, "3D Human Pose Estimation: A Review of the Literature and Analysis of Covariates," *Computer Vision and Image Understanding*, vol. 152, pp. 1–20, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314216301369

[28] A. G. Chavez, C. A. Mueller, A. Birk, A. Babic, and N. Miskovic, "Stereo-Vision Based Diver Pose Estimation Using LSTM Recurrent Neural Networks for AUV Navigation Guidance," in *OCEANS 2017 - Aberdeen*, 2017, pp. 1–7.

[29] M. J. Islam, J. Mo, and J. Sattar, "Robot-to-Robot Relative Pose Estimation Using Humans as Markers," *Autonomous Robots*, vol. 45, no. 4, pp. 579–593, May 2021. [Online]. Available: https://doi.org/10.1007/s10514-021-09985-6

[30] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021.

[31] M. Fulton, J. Hong, and J. Sattar, "Using Monocular Vision and Human Body Priors for AUVs to Autonomously Approach Divers," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 1076–1082.

[32] "TRT Pose," 2020. [Online]. Available: https://github.com/NVIDIA-AI-IOT/trt_pose

[33] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A Framework for Perceiving and Processing Reality," in *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019. [Online]. Available: https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf

[34] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device Real-time Body Pose tracking," 2020. [Online]. Available: https://arxiv.org/abs/2006.10204

[35] E. Littmann, A. Drees, and H. Ritter, "Robot Guidance by Human Pointing Gestures," in *Proceedings of International Workshop on Neural Networks for Identification, Control, Robotics and Signal/Image Processing*, Aug 1996, p. 449–457.

[36] J. Delmerico, S. Mintchev, A. Giusti, B. Gromov, K. Melo, T. Horvat, C. Cadena, M. Hutter, A. Ijspeert, D. Floreano, L. M.

Gambardella, R. Siegwart, and D. Scaramuzza, "The Current State and Future Outlook of Rescue Robotics," *Journal of Field Robotics*, vol. 36, no. 7, pp. 1171–1191, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21887

[37] D. G. Lowe, "Distinctive image features from scale-invariant keypoints." *International Journal of Computer Vision(IJCV)*, vol. 60, no. 2, pp. 91–110, 2004.

[38] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2564–2571.

[39] J. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 679–698, 1986.

[40] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "ROS: an Open-Source Robot Operating System," in *ICRA workshop on open source software*, vol. 3. Kobe, Japan, 2009, p. 5.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[42] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.