2024 IEEE International Conference on Robotics and Automation (ICRA)
May 13-17, 2024. Yokohama, Japan

2024 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-8457-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRA57147.2024.10610809

Autonomous robotic re-alignment for face-to-face underwater human-robot interaction*

Demetrious T. Kutzke[1], Ashwin Wariar[2], and Junaed Sattar[3]

***Abstract*— The use of autonomous underwater vehicles (AUVs) to accomplish traditionally challenging and dangerous tasks has proliferated thanks to advances in sensing, navigation, manipulation, and on-board computing technologies. Utilizing AUVs in underwater human-robot interaction (UHRI) has witnessed comparatively smaller levels of growth due to limitations in bi-directional communication and significant technical hurdles to bridge the gap between analogies with terrestrial interaction strategies and those that are possible in the underwater domain. A necessary component to support UHRI is establishing a system for safe robotic-diver approach to establish face-to-face communication that considers non-standard human body pose. In this work, we introduce a stereo vision system for enhancing UHRI that utilizes three-dimensional reconstruction from stereo image pairs and machine learning for localizing human joint estimates. We then establish a convention for a coordinate system that encodes the direction the human is facing with respect to the camera coordinate frame. This allows automatic setpoint computation that preserves human body scale and can be used as input to an image-based visual servo control scheme. We show that our setpoint computations tend to agree both quantitatively and qualitatively with experimental setpoint baselines. The methodology introduced shows promise for enhancing UHRI by improving robotic perception of human orientation underwater.**

I. Introduction

Problems associated with humans and robots interacting, also referred to as human-robot interaction (HRI), is well-studied in controlled terrestrial environments [1]. Innovations in HRI have bolstered adoption of these technologies into many areas of life, such as manufacturing [2], medicine [3], long-term care of the elderly [4], military applications [5], and the underwater domain [6]. This is due in part to the benefit of allowing robots to take on the dirty, dull, and dangerous tasks [7] that would otherwise place humans in direct harm or assist in situations in which it is not possible for the human to provide the level of persistent attention required, as in the case of long-term care facilities. The thought goes that off-loading these tasks to robots will allow humans to interact with them from relative safety or convenience, while also performing oversight [8].

Underwater human-robot interaction (UHRI) is much more challenging, because the underwater domain presents a formidable environment for robotic sensing. It lacks many of the benefits of the terrestrial domain such as high bandwidth radio communication, *e.g.*, wireless internet and Bluetooth, consistent lighting conditions, and localization via global-positioning satellites. Underwater robots must utilize alternative methods to perform the same tasks as their terrestrial counterparts. Often, expensive sonar-based techniques are used to perform navigation and localization [9]. Utilizing visual sensors is also challenging, because differences in salinity and particulates in the water can occlude and distort imagery. Creative techniques must be employed to enhance vision underwater when the conditions are especially degraded [10], [11], [12]. However, even with the challenges of visual imagery, there are instances where utilizing camera data is preferable to expensive and invasive acoustic systems. For example, when acoustically susceptible marine life such as dolphins or whales are present [13], in these instances, utilizing visual sensing, which is less invasive, is both ethical and beneficial to the preservation of the marine life.

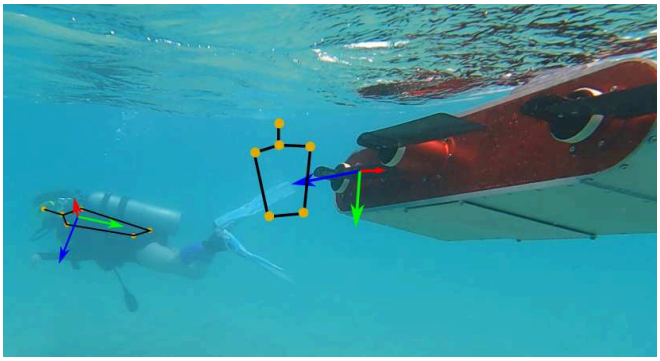


Fig. 1: Underwater human robot interaction is enhanced by the robot's ability to re-orient itself with respect to the diver, rather than requiring the diver to re-orient with respect to the robot.

Divers operate in a similarly sensory-deprived state. Scuba masks occlude peripheral vision or at the very least can reduce the diver's ability to see or perceive dynamic robotic gestures underwater; acoustic signals are degraded by inhalation and exhalation through the breathing regulator, which significantly reduces the diver's ability to hear; and environmental conditions such as strong currents, silting from sediment in the water column, and frigid water temperatures all contribute to generally high cognitive loads. Sensory-deprived states for both robots and humans means that in complex UHRI scenarios, where communication is critical from both robot-to-human (R2H) and human-to-robot (H2R), there is a high-probability of information loss. We argue that

*This work was supported in part by the Science, Mathematics, and Research for Transformation (SMART) Scholarship provided through the US Department of Defense, the University of Minnesota UROP award, and the National Science Foundation Award IIS-2220956.

The authors are with the Department of Computer Science & Engineering and the Minnesota Robotics Institute, University of Minnesota–Twin Cities, Minneapolis, MN 55455, USA {[1]kutzk015, [2]waria012, [3]junaed}@umn.edu

because of these conditions, the robot must have the ability to autonomously establish face-to-face (F2F) communication. *F2F* communication reduces the probability of information loss by ensuring that the robot and the human are within a safe distance and in full view of each other. The diver can see the robot's movements and vice-versa. The diver also has the best chance of hearing any acoustically communicated information. To achieve this F2F configuration, we propose a stereo vision algorithm to autonomously compute a desired feature setpoint, which can be used for visual servo control schemes. This eliminates the need for human-engineered features and equips the robot with the ability to infer a desired F2F setpoint from nonstandard body poses that preserves scale. Scale is important to ensure safe and consistent approach distances for divers of different shapes. To our knowledge, the problem of autonomously establishing F2F communication underwater has not been considered for general body poses or instances where the robot is not already within a safe communication distance and can perceive the human diver's face.

Many techniques have been devised to establish both R2H [14] and H2R communication (*e.g.*, [15]). These systems are supported by complementary techniques to enable visual robot control to place the robot within a safe distance of the human [16]. This allows higher fidelity understanding from both the robot's and the human's perspective, since it is thought that the information exchange is best when interpreted from the alignment between the human's eyes and the robot's camera [16]. To ensure robust communication, authors in [17] used a transformer-driven network for detecting diver gaze based on facial mask keypoints. However, their work does not handle general poses Fig. 2, or those in which the facial keypoints are not visible.

We argue that a complementary problem to the works of [16] and [17] is utilizing a two-dimensional pose estimator, along with a stereo visual approach to establish three-dimensional positions. By doing this, we can accommodate non-standard human poses, such as those shown in Fig. 2. This will ultimately enable more complex robotic control for re-orientation; *e.g.*, when the human is conducting complex tasks and is unable to re-orient themselves with respect to the robot, the robot can come to the human. It is from this perspective that we define our primary contributions to support UHRI, which can be summarized as follows:

- The aggregation of a diverse torso keypoint dataset and results from training an off-the-shelf pose estimation algorithm for two-dimensional human pose estimates that accommodates non-standard body poses.
- The computation of an alignment vector and establishment of a convention for assigning a coordinate frame to a human's facing direction.
- Scale preserving setpoint computations which preserve different body shapes at different distances between the robot and the human.

Fig. 2: Example non-standard diver poses that are typical during scuba diving operations. Diver robot interaction scenarios must accommodate these poses to be useful for underwater missions.

II. Related Work

The work introduced in this paper exists within the boundary between UHRI and computer vision for pose estimation. Here we discuss some of the works that influence our methodology.

Underwater human robot interaction. Various methods have been proposed for robotic detection and individual identification of human divers (*e.g.*, [18], [19], [20]) with features extracted from visual, or spatio-temporal signals. Others have utilized sonar detection mechanisms to both directly detect in frequency space the presence of a diver [21] and reconstructed acoustic images [22], [23]. For explicit communication between an AUV and divers, both robot-to-human and human-to-robot methods have been proposed; *e.g.*, robots have used light [14], motion [6], [24], and other cues to communicate intent and information to divers, and fiducials [25], [26], [27], hand gestures [28], [29], [30], [31], and complex user interface devices [32], [33] have all been used by divers to control robots. However, it is conceivably challenging and constraining for divers to use tags or UI devices while underwater for certain tasks.

Human Pose Estimation. Pose estimation is the task of determining a set of keypoints that define human joint positions in an image. Various techniques exist, but most rely on convolutional neural networks (CNNs) [34] to perform feature extraction and output heatmaps over candidate locations [35], [36], [37].

The networks are trained to regress from heatmaps to perform keypoint localization by selecting the location with the highest probability as the most likely joint location.

Localizing joint locations accurately is a significant challenge underwater, which is exacerbated by inconsistent lighting conditions and the lack of saliency, or pronounced features, within the typical diver silhouette. Chavez *et al*. utilize a recurrent neural network (RNN) with long short-term (LSTM) cells to learn the sequential joint orientations affixed to the human diver, exploiting stereo vision and 17 inertial measurement units (IMUs) that communicate the diver's movements acoustically. We recognize that placing additional burden on the diver's already intense cognitive load is problematic. Instead, we endeavor to construct F2F re-orientation in such a way that the robot re-orients itself with respect to the human based off of image observations alone.

III. The Face-to-Face Reorientation Approach

The F2F scale-preserving setpoint computation comprises two components. First, a pose estimation component local-

izes torso keypoints, and second, an alignment vector computation establishes a convention for affixing a right-handed coordinate system to the human, from which we compute the transformation that anti-aligns the body frame coordinate system with the camera frame. Perspective projection allows us to recover the ideal setpoint, which is the configuration in which the human is facing the camera.

While a future goal is to use a three-dimensional pose estimation algorithm on monocular camera data, much of the work in three-dimensional pose estimation first uses multi-camera setups to triangulate pose keypoints to provide a z-component to ground truth labeled data. Pose estimation algorithms can then be trained to directly predict a three-dimensional vector from a single monocular image, see [38], [39], for example. In the underwater domain, instrumenting an experimental setup with calibrated multi-view cameras is a challenge and impractical for most setups. To that end, we utilize calibrated stereo cameras affixed to a robot to collect and aggregate images of human divers. We trained a deep neural network on labeled image data to localize two-dimensional human joints. During runtime, we utilize stereo reconstruction based on pose alone. Dense stereo matching is found to be ineffective for the underwater environment. The typical diver silhouette almost entirely appears the same to traditional block matching algorithms or even more sophisticated post-processing techniques that fill in gaps in the disparity map. From the scale-preserving setpoint computation, we preserve different body shapes to ensure that the robot can automatically predict the optimal alignment or setpoint for the visual control scheme, without need for human intervention or calibration before the beginning of the mission. Instead, the robot can detect the setpoint using our method and perform classic visual control schemes. A detailed account of visual servo control schemes is beyond the scope of the present work, but an extensive treatment can be found in [40], [41], [42].

Stereo Diver Pose Dataset. We aggregated and labeled $6,711$ stereo image pairs from a closed-water environment in which a stereo vision camera was used to collect images of divers in diverse poses. The images collected are all of size 640×480 pixels, and feature a *single* diver in a full wet suit and dive gear. The hands and face are exposed, except for parts of the face, which are partially obscured by the breathing regulator and mask. The poses were labeled according to the convention shown in Fig. 3. We argue that the points that encompass the torso, the midpoint of the eyes, and the base of the neck, are the primary keypoints for robotic re-orientation. The torso keypoints do not change relative to each other throughout normal diving operations, whereas the limbs are often moving to stabilize the diver's position in the water. The midpoint of the eyes was also utilized to accurately reflect the midpoint of the body itself. Without this additional keypoint, the robot would come face-to-face lower than expected with the diver, inhibiting the most efficient communication. Example poses are shown in Fig. 4.

Pose Estimation. With the aggregated dataset, we train

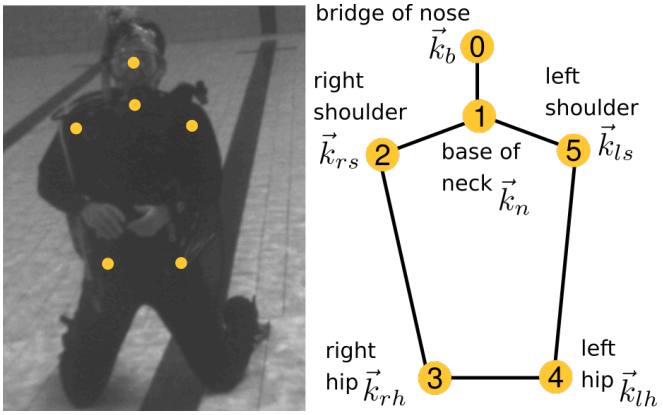


Fig. 3: Pose keypoint convention used by F2F along with a sample labeled image. The pose estimator may provide more anatomical keypoints, but F2F only requires these six.

Fig. 4: Sample raw images from the F2F pose dataset. The dataset contains diverse poses to represent the broadest possible set of orientations a diver can assume while conducting underwater operations.

a deep neural network based on the DeepLabCut (DLC) framework [43], [44] for torso keypoint estimation. Specifically, we train a ResNet-50-based neural network using 95% of the dataset, or approximately $6,375$ images, for $500,000$ training iterations. We find the test error to be 12.75 pixels, and the train error at 11.14 pixels. We then use a threshold p-cutoff of 0.05 to condition the (x,y) coordinates for future analysis. While the error might seem high, our goal is to quantify the extent to which we consider alignment as face-to-face interaction. To that end, we do not need nearly-optimal keypoint configurations, but localizations that are good enough to recover sparse 3D reconstructions are sufficient; we find that approximately 10 pixels of error provides enough accuracy for our application.

Examples from the pose estimator evaluated on the test dataset are shown in Fig. 5 below. Generally, DLC performs adequately localizing the joints. However, there are times where even variations in labeling likely contribute to issues in localization. For example, if the joint is occluded, it is possible the DLC network will be unable to predict location.

Alignment Coordinate System Convention. A key contribution of this work is the computation of the alignment vector. The *alignment vector* extends perpendicularly to the plane defined by the torso of the diver. This serves as the anchor vector, which allows us to define a coordinate system that is attached to the diver. To construct this vector, we use

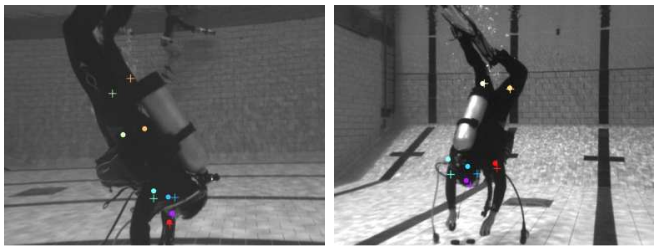


Fig. 5: DeepLabCut evaluation on the F2F test dataset. Cross markers indicate ground truth labels, and dots indicate DeepLabCut estimations with confidences $p > p_{\text{cutoff}} = 0.05$.

a set of joint location keypoints from our pose estimation network. Let $\vec{k}_{\text{b}}$ define the bridge of the nose, $\vec{k}_{\text{n}}$ define the base of the neck, $\vec{k}_{\text{rs}}$ define the right shoulder, $\vec{k}_{\text{rh}}$ define the right hip, $\vec{k}_{\text{lh}}$ define the left hip, and $\vec{k}_{\text{ls}}$ define the left shoulder.

After performing stereo image rectification and triangulating based on joint localizations from both the left and right image pairs, the vectors $\vec{k}_i \in \mathcal{K}$, where i corresponds to a specific joint, are defined in three-dimensional camera coordinates as $({}^c x_i, {}^c y_i, {}^c z_i)$. Let $\vec{s}_i$ denote the corresponding projection of the point $\vec{k}_i$ onto the image plane of the robot's camera. The projection $\vec{s}_i = (u_i, v_i)$ is a point defined in the image plane, where $u_i \in [0, N]$ and $v_i \in [0, M]$, and N, M are the image width and height, respectively. To compute the alignment vector, we define the following steps.

1) We compute the center of the predicted keypoints as $\vec{k}_{\text{o}} = \langle \vec{k} \rangle_{\vec{k} \in \mathcal{K}}$, where $\langle \cdot \rangle$ defines the vector average computation. The resultant vector $\vec{k}_{\text{o}}$ is located approximately center of mass and is skewed toward the upper part of the torso.
2) We define several difference vector quantities that exist on the torso plane as
$$\vec{k}_{\text{lsh}} = \vec{k}_{\text{ls}} - \vec{k}_{\text{lh}} \qquad \vec{k}_{\text{nlh}} = \vec{k}_{\text{n}} - \vec{k}_{\text{lh}} \tag{1}$$
$$\vec{k}_{\text{nrh}} = \vec{k}_{\text{n}} - \vec{k}_{\text{rh}} \qquad \vec{k}_{\text{rsh}} = \vec{k}_{\text{rs}} - \vec{k}_{\text{rh}}. \tag{2}$$
These quantities are needed to establish the relationships between joint locations, effectively defining the torso plane and conditioning the proceeding analysis with respect to the torso plane.
3) We compute the alignment direction (or the diver's facing direction) by taking the average direction of the cross product between the difference vectors of the torso and neck joints. This defines a direction perpendicular to the plane defined by the torso keypoints
$$\vec{k}_{\text{l}_\times} = \vec{k}_{\text{lsh}} \times \vec{k}_{\text{nlh}} \tag{3}$$
$$\vec{k}_{\text{r}_\times} = \vec{k}_{\text{nrh}} \times \vec{k}_{\text{rsh}}. \tag{4}$$
To compute the average direction and define a unit vector, we first take the average, and then we divide by the vector $L2$-norm
$${}^c\hat{z}_B \equiv \frac{\langle \vec{k}_{\text{r}_\times}, \vec{k}_{\text{l}_\times} \rangle}{\| \langle \vec{k}_{\text{r}_\times}, \vec{k}_{\text{l}_\times} \rangle \|_2}. \tag{5}$$
The alignment vector given by (5) points in a direction perpendicular to the plane defined by the torso keypoints. We now affix a right-handed coordinate system to $\vec{k}_{\text{o}}$, with ${}^c\hat{z}_B$ aligned along the direction given in (5). We choose ${}^c\hat{y}_B$ to be the vector that points along the direction of the midpoint between hip joints. This is given by computing the midpoint of the line segment connecting the hip joints
$$\vec{k}_{\text{midpt}} = \langle \vec{k}_{\text{lh}}, \vec{k}_{\text{rh}} \rangle. \tag{6}$$
4) From this we compute the unit vector that points from the center of mass vector $\vec{k}_o$ to $\vec{k}_{\text{midpt}}$. This unit vector is defined to be ${}^c\hat{y}_B$
$${}^c\hat{y}_B = \frac{\vec{k}_{\text{midpt}} - \vec{k}_o}{\| \vec{k}_{\text{midpt}} - \vec{k}_o \|_2}. \tag{7}$$
5) Finally, the ${}^c\hat{x}_B$ is computed through a cross product ${}^c\hat{x}_B = {}^c\hat{y}_B \times {}^c\hat{z}_B$. Together these constitute the body frame ${}^c\mathcal{F}_B = [{}^c\hat{x}_B, {}^c\hat{y}_B, {}^c\hat{z}_B, \vec{k}_{\text{o}}]$ of the human diver, affixed to the midpoint of the extracted pose keypoints, with the ${}^c\hat{z}_B$ aligned in the direction perpendicular to the plane defined by the torso keypoints. It is from this that we can then compute the ideal pose configuration by anti-aligning the body frame with the camera frame unit vectors.

The preceding analysis computes the body frame ${}^c\mathcal{F}_B$ with respect to the camera frame. To compute the pose setpoint for a visual control scheme, we need to compute the transformation that anti-aligns the camera frame and the body frame. That is, there exists some transformation $\tilde{\text{T}}$ that yields an ideal configuration of the body frame ${}^c\mathcal{F}_B^*$
$${}^c\mathcal{F}_B^* = \tilde{\text{T}}{}^c\mathcal{F}_B. \tag{8}$$
This transformation is not known *a priori*.

For a camera that has z-axis along the optical axis, a right-facing x-axis, and a down-facing y-axis, the following constraints hold
$${}^c\hat{z}_B^* \cdot {}^c\hat{z} = -1 \qquad {}^c\hat{x}_B^* \cdot {}^c\hat{x} = -1 \qquad {}^c\hat{y}_B^* \cdot {}^c\hat{y} = 1. \tag{9}$$
From these equations, we compute the rotation matrix required to align the body frame with respect to the camera frame to be in an F2F orientation as
$${}^c\mathcal{F}_B^* = \tilde{\text{T}}{}^c\mathcal{F}_B. \tag{10}$$
The transformation that aligns the axes can be computed using the Kabsch algorithm [45] which minimizes a root-mean-square error function to find the optimal rotation matrix that aligns a set of vectors. The keen observer will note that this alignment does not quite yield the configuration desired. In fact, it aligns the coordinate systems such that the human would be facing away from the robot's camera. To anti-align the coordinate systems, we define $\tilde{\text{T}} \leftarrow \text{R}_y(\pi)\tilde{\text{T}}$. This produces the desired rotation.

Together, along with a translation constraint, which defines how close the keypoints should appear to the camera frame,

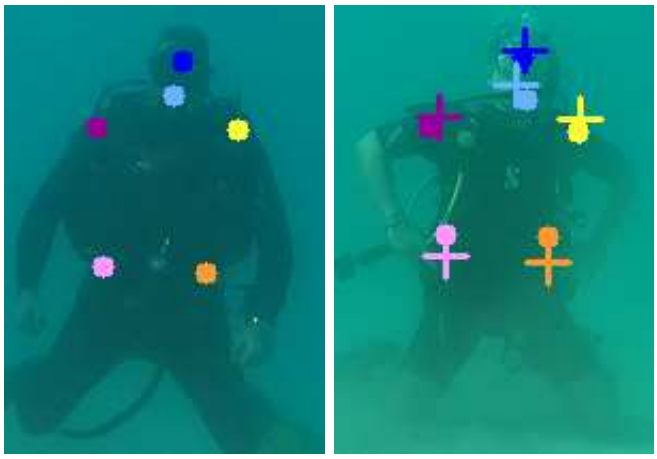


Fig. 6: Two divers of different body shapes have different setpoints at the same scale, in this case 2 m from the camera. The crosses in the right-hand figure are the baseline setpoints from the diver in the left-hand image. The points have been shifted to center on the mean of the keypoints in the right-hand image. F2F produces scale aware setpoints, such that the robot can utilize visual servo techniques to ensure safe approach from non-standard body poses.

1 m 2 m 3 m

Fig. 7: Setpoints used as baselines for comparing projections from reconstruction and alignment. Note that setpoint baselines (colored dots) have been shifted to the center of the image to accommodate for differences in image capture that occurred due to strong ocean currents that made station keeping challenging for the divers.

or the robot, these can be used to compute the components of the transformation matrix. Finally, perspective projection using the camera intrinsics yields the appearance of the points in the image plane by $\vec{s}^* = \mathrm{K}\tilde{\mathrm{T}}\vec{k}$. The vector of points $\vec{s}^*$ is the scale-preserved setpoint.

The benefits of this approach are two-fold. By computing a body-fixed frame, we can compute the setpoint for a visual-servo control scheme that is scale-aware. This means that different diver body shapes will appear as different sizes depending on the scale heuristic. For example, the divers shown in Fig. 6 are at the same distance from the camera, but the ideal setpoint is very different.

As a result, the control scheme would indicate that the robot should move closer to achieve the same level of error between the setpoint and the observed pose. Of course, this could potentially place the diver in harms way if the robot malfunctions.

IV. Experimental Results

Experiments were conducted using data collected from divers in both closed-water (*i.e.*, pool) settings and in the ocean waters off the coast of Barbados, West Indies. For this work, we focus on measuring the error between baseline setpoints and constructed setpoints using our alignment vector convention.

To collect setpoint baselines, a diver for whom we had existing pose data was asked to station keep above an experimental trackline. The trackline was measured to 1, 2, and 3 m distances. The camera operator asked the diver to spend approximately 15 seconds station keeping during acquisition of stereo image data. The camera operator then signaled the diver to move forward. This process was repeated until the diver had been recorded at all three distances. The setpoint baseline was computed by using a simple by-hand label technique that mimicked what a field operator or end-user would do during calibration of a visual servo system. This hand measurement is shown at the three distances in Fig. 7.

We then performed experiments using the setpoints extracted from the three distances and comparing the euclidean error and standard deviation of the error between six canonical pose positions. These results are summarized in the table below. Note that 50 frames of data at each canonical pose were used during the error computation. A reasonable expectation might be on the order of the error of the pose estimation network. In this case, approximately 60 pixels of error across all projected keypoints is reasonable. Clearly we do not observe that low of error, except in exceptional cases. There are a couple of reasons for this. The pose estimation network runs inference on both the left- and right-hand image pairs, and any fluctuations caused from camera motion or inconsistent lighting conditions can cause significant errors during triangulation. To that end, notice that we have also shown not only the average error in the figures of Fig. 8 but also the minimum error observed during the frame acquisition. Some poses appear exceptionally good qualitatively and tend to agree with the baselines. There is a trend that indicates reconstruction and projection is better at distances that exceed 1 m. This is likely caused by better pose estimations. Most of the dataset on which we trained DLC was collected at distances of approximately 2–3 m from the diver. As a result, predictions from the pose network are better and more consistent at these distances, likely causing reductions in reconstruction errors throughout the frames used for analysis.

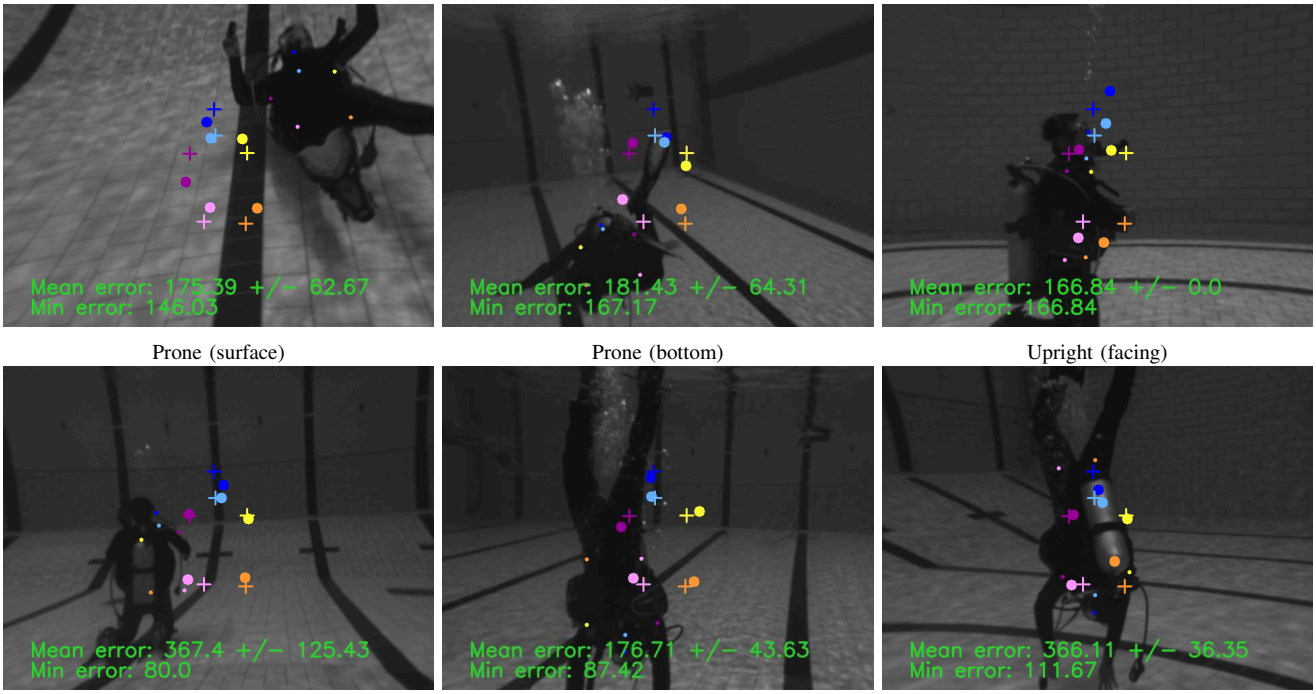
The results summarized in Table I demonstrate that we achieve reasonable projection errors for most poses. The visual results for the 1 m poses are shown in Figure 8 for the six canonical pose states.

V. Future Work

Scale-preservation and automatic setpoint computation allows a robot to autonomously detect the desired keypoint configuration, without need for prior calibration. This means in a complex mission, the robot has the ability to re-orient itself with respect to the diver to establish F2F communication. However, the work described has shortcomings. First, totally occluded joints pose a problem for reconstruction, because lateral views appear at the same distance. To subvert this,

Pose	1 m	2 m	3 m	Error across distances
Prone (surface)	175.39 ± 62.67	227.19 ± 8.3	322.61 ± 0.0	241.73 ± 23.66
Prone (bottom)	181.43 ± 64.31	192.04 ± 0.0	157.98 ± 6.51	177.15 ± 23.6
Upright (away)	367.4 ± 125.43	61.9 ± 0.98	88.81 ± 0.0	173.99 ± 52.7
Upright (facing)	166.84 ± 0.0	383.12 ± 157.83	92.67 ± 31.70	214.21 ± 63.17
Inverted (facing)	176.71 ± 43.63	101.43 ± 12.77	224.08 ± 71.63	167.41 ± 42.68
Inverted (away)	366.11 ± 36.35	172.28 ± 43.43	131.31 ± 58.75	223.23 ± 46.18
Error across poses	238.98 ± 55.40	189.66 ± 37.22	170.22 ± 33.38	

TABLE I: Summary of projection errors between setpoint baselines at 1, 2, and 3 m distances and projections from scale-preserving computations, averaged over 50 projection estimates for image resolution 640×480. Errors are reported as *mean*$\pm$*standard deviation* in pixel units. Each error is the sum over all keypoints ($\times 6$) of the Euclidean distance between the ground truth and the predicted keypoint location. A standard deviation of zero indicates observation of a single alignment over 50 projection estimates.



Prone (surface) | Prone (bottom) | Upright (facing)

Upright (away) | Inverted (facing) | Inverted (away)

Fig. 8: The projections we achieve through the alignment system are reasonable, given the challenges of reconstruction. The cross markers indicate the 1 m setpoint baseline computed from sea trial data. The large dots indicate projections from the alignment system. A standard deviation of zero indicates instances in which the system was unable to triangulate except for a single instance during the frame acquisition.

anthropometric data ratios (ADRs), which describe ratios between human limbs, can be used to regularize depth estimates so that totally lateral views do not appear to have joints at the same distance. Another issue with the present approach is that the algorithm does not account for instances where the diver's body is positioned behind an obstacle relative to the camera's origin. Future work will include context-aware approach strategies that utilize the entirety of the scene for understanding the diver's behavior and selecting the best place to approach from, particularly with regard to complex mission tasks. We will conduct control experiments using the F2F system in pool environments, where evaluation of the control can be better constrained. The ocean environment is susceptible to external perturbations from currents making it challenging to evaluate the efficacy of the F2F system.

VI. Conclusion

We have demonstrated that our methodology enables automatic setpoint computation from human diver pose observations, which could permit more complex UHRI scenarios where the robot must infer automatically the best orientation of target features to navigate within a safe distance of the human diver. We have shown that our method works well for 1, 2, and 3 m baselines, except in cases of poses in which a joint is totally occluded and the triangulation is unable to resolve the depth at which the occluded joint appears.

References

[1] M. Matarić, "On relevance: Balancing theory and practice in HRI," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 7, no. 1, p. 1–2, 2018.

[2] A. Cherubini, R. Passama, A. Crosnier, A. Lasnier, and P. Fraisse, "Collaborative manufacturing with physical human–robot interaction," *Robotics and Computer-Integrated Manufacturing*, vol. 40, p. 1–13, 2016.

[3] H. Su, J. Sandoval, M. Makhdoomi, G. Ferrigno, and E. De Momi, "Safety-enhanced human-robot interaction control of redundant robot for teleoperated minimally invasive surgery," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, p. 6611–6616.

[4] J. Fasola and M. J. Matarić, "A socially assistive robot exercise coach for the elderly," *Journal of Human-Robot Interaction*, vol. 2, no. 2, p. 3–32, 2013.

[5] F. Jentsch, *Human-robot interactions in future military operations*. CRC Press, 2016.

[6] M. Fulton, C. Edge, and J. Sattar, "Robot communication via motion: Closing the underwater human-robot interaction loop," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, p. 4660–4666.

[7] L. Takayama, W. Ju, and C. Nass, "Beyond dirty, dangerous and dull: what everyday people think robots should do," in *2008 3rd ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE, 2008, p. 25–32.

[8] T. B. Sheridan, "Human–robot interaction: status and challenges," *Human factors*, vol. 58, no. 4, p. 525–532, 2016.

[9] P. Corke, C. Detweiler, M. Dunbabin, M. Hamilton, D. Rus, and I. Vasilescu, "Experiments with underwater robot localization and tracking," in *Proceedings 2007 IEEE International Conference on Robotics and Automation*. IEEE, 2007, p. 4556–4561.

[10] C. Edge, M. J. Islam, C. Morse, and J. Sattar, "A generative approach for detection-driven underwater image enhancement," *arXiv preprint arXiv:2012.05990*, 2020.

[11] M. J. Islam, Y. Xia, and J. Sattar, "Fast underwater image enhancement for improved visual perception," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3227–3234, 2020.

[12] 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.

[13] M. C. Hastings, "Coming to terms with the effects of ocean noise on marine animals," *Acoustics today*, vol. 4, no. 2, p. 22–34, 2008.

[14] M. Fulton, A. Prabhu, and J. Sattar, "HREyes: Design, Development, and Evaluation of a Novel Method for AUVs to Communicate Information and Gaze Direction," *arXiv preprint arXiv:2211.02946*, 2022.

[15] S. S. Enan, M. Fulton, and J. Sattar, "Robotic Detection of a Human-Comprehensible Gestural Language for Underwater Multi-Human-Robot Collaboration," *arXiv preprint arXiv:2207.05331*, 2022.

[16] 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)*. IEEE, 2022, p. 1076–1082.

[17] S. S. Enan and J. Sattar, "Visual Detection of Diver Attentiveness for Underwater Human-Robot Interaction," *arXiv preprint arXiv:2209.14447*, 2022.

[18] Y. Xia and J. Sattar, "Visual diver recognition for underwater human-robot collaboration," in *2019 international conference on robotics and automation (ICRA)*. IEEE, 2019, p. 6839–6845.

[19] M. J. Islam and J. Sattar, "Mixed-domain biological motion tracking for underwater human-robot interaction," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, p. 4457–4464.

[20] J. Sattar and G. Dudek, "Where is your dive buddy: tracking humans underwater using spatio-temporal features," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2007, p. 3654–3659.

[21] I. Kvasić, N. Mišković, and Z. Vukić, "Convolutional neural network architectures for sonar-based diver detection and tracking," in *OCEANS 2019-Marseille*. IEEE, 2019, p. 1–6.

[22] K. W. Lo and B. G. Ferguson, "Diver detection and localization using passive sonar," in *Proceedings of Acoustics*, vol. 8, 2012, p. 1–8.

[23] K. J. DeMarco, M. E. West, and A. M. Howard, "Sonar-based detection and tracking of a diver for underwater human-robot interaction scenarios," in *2013 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2013, p. 2378–2383.

[24] M. Fulton, C. Edge, and J. Sattar, "Robot Communication Via Motion: A Study on Modalities for Robot-to-Human Communication in the Field," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 11, no. 2, p. 1–40, 2022.

[25] G. Dudek, J. Sattar, and A. Xu, "A Visual Language for Robot Control and Programming: A Human-Interface Study," in *Proceedings of the International Conference on Robotics and Automation ICRA*, Rome, Italy, April 2007, pp. 2507–2513.

[26] A. Xu, G. Dudek, and J. Sattar, "A Natural Gesture Interface for Operating Robotic Systems," in *Proceedings of the IEEE International Conference on Robotics and Automation, ICRA*, Pasadena, California, May 2008, pp. 3557–3563.

[27] A. Sagitov, K. Shabalina, L. Sabirova, H. Li, and E. Magid, "ARTag, AprilTag and CALTag Fiducial Marker Systems: Comparison in a Presence of Partial Marker Occlusion and Rotation." in *ICINCO (2)*, 2017, p. 182–191.

[28] H. Buelow and A. Birk, "Gesture-recognition as basis for a human robot interface (hri) on a auv," in *OCEANS'11 MTS/IEEE KONA*, Sept 2011, pp. 1–9.

[29] 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, p. 91, 2018.

[30] M. J. Islam, M. Ho, and J. Sattar, "Understanding human motion and gestures for underwater human–robot collaboration," *Journal of Field Robotics*, vol. 36, no. 5, p. 851–873, 2019.

[31] 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 and Automation Magazine*, vol. 28, no. 3, pp. 67–78, 2021.

[32] B. Verzijlenberg and M. Jenkin, "Swimming with Robots: Human Robot Communication at Depth," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 4023–4028.

[33] M. Bernardi, C. Cardia, P. Gjanci, A. Monterubbiano, C. Petrioli, L. Picari, and D. Spaccini, "The Diver System: Multimedia Communication and Localization Using Underwater Acoustic Networks," in *2019 IEEE 20th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, 2019, pp. 1–8.

[34] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[35] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Eur. Conf. Comput. Vis.*, 2018, p. 466–481.

[36] H. Dai, H. Shi, W. Liu, L. Wang, Y. Liu, and T. Mei, "FasterPose: A Faster Simple Baseline for Human Pose Estimation," *ACM TOMM*, vol. 18, no. 4, p. 1–16, 2022.

[37] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, p. 7291–7299.

[38] Z. Yu, J. S. Yoon, I. K. Lee, P. Venkatesh, J. Park, J. Yu, and H. S. Park, "Humbi: A large multiview dataset of human body expressions," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, p. 2990–3000.

[39] J. S. Yoon, Z. Yu, J. Park, and H. S. Park, "HUMBI: A Large Multiview Dataset of Human Body Expressions and Benchmark Challenge," *arXiv preprint arXiv:2110.00119*, 2021.

[40] F. Chaumette and S. Hutchinson, "Visual servo control. I. Basic approaches," *IEEE Robotics & Automation Magazine*, vol. 13, no. 4, p. 82–90, 2006.

[41] ——, "Visual servoing and visual tracking," 2008.

[42] F. Chaumette, S. Hutchinson, and P. Corke, "Visual servoing," in *Springer Handbook of Robotics*. Springer, 2016, p. 841–866.

[43] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge, "DeepLabCut: markerless pose estimation of user-defined body parts with deep learning," *Nature neuroscience*, vol. 21, no. 9, p. 1281–1289, 2018.

[44] T. Nath, A. Mathis, A. C. Chen, A. Patel, M. Bethge, and M. W. Mathis, "Using DeepLabCut for 3D markerless pose estimation across species and behaviors," *Nature protocols*, vol. 14, no. 7, p. 2152–2176, 2019.

[45] W. Kabsch, "A Discussion of the Solution for the Best Rotation to Relate Two Sets of Vectors," *Acta Crystallographica Section A*, vol. 34, no. 5, pp. 827–828, Sep 1978.