Subject-independent diver gesture classification using upper limb movement

Bilal Ghader, Claire Dune, Eric Watelain, Vincent Hugel

Subject-independent diver gesture classification using upper limb movement

Bilal Ghader[1,2], Claire Dune[1], Eric Watelain[2], and Vincent Hugel[1]

***Abstract*—This study focuses on categorizing diver gestures by analyzing angle features extracted from the movements of their upper limbs without exploiting information encoded by the hands, as is generally the case in the literature. Our approach is intended to be as generic as possible, in order to enable gesture recognition, whatever the diver's equipment, and to use the usual signs used by divers. New shallow RNN pipelines based on LSTM and GRU are proposed and evaluated with regard to a DTW-KNN deterministic baseline. For underwater gestures, a preliminary energy-based SVM separation stage is introduced to distinguish between one-arm and two-arm gestures. All classification strategies are validated using a leave-one-out protocol on a motion capture dataset comprising 14 divers performing 11 distinct gestures. The database was collected in-house with a total of 1078 individual gesture recordings. The SVM separation stage clearly improves the results, from 15% for DTW-KNN to 5% for RNNs. The best RNN leave-one-out classification accuracy is obtained for the proposed two-layer LSTM network combined with a 1D-convolution layer, and a fully connected layer, yielding a 89.5% good classification rate, compared with a 92% rate using the DTW-KNN baseline. The data and code are made publicly available at https://github.com/LaboratoireCosmerTOULON/DTW_KNN_RNN**

***Index Terms*—Gesture, Posture and Facial Expressions, Human-Robot Collaboration, Marine Robotics, Datasets for Human Motion.**

I. Introduction

DEPENDING on the depth, duration, and complexity of the operations to be carried out, underwater missions are performed either by divers or underwater robots. So far, joint operations remain very limited. However, for complex applications where on-site human judgment is required, human-robot interaction is essential.

In the context of underwater collaboration, divers employ a set of standardized gestures to facilitate communication, which they have acquired through training (Fig. 1). The CMAS (Confédération Mondiale des Activités Subaquatique - https://www.cmas.org) standardized the diving gesture dictionary at the Barcelona Congress in 1960 and the Singapore Congress in 1999. The aforementioned gestures are primarily concerned with safety and, as a result, must be straightforward, unambiguous and limited in number to ensure optimal memorization. These gestures may involve the use of one or both arms, with the height of the hands varying from the hips to the head. Some gestures may oscillate, while others may include a prolonged static pause phase. One gesture only communicates one piece of information. Videos demonstrating the gestures are available alongside the data[a]. Gestures performed with one hand can be performed with either hand. To interact with a human diver in a natural and intuitive way, a companion robot must be able to recognize standard gestures.

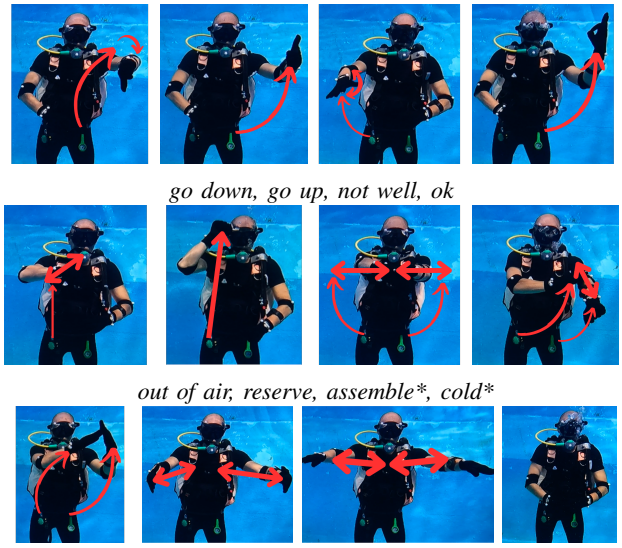


Figure 1: Subset of standard European diver gestures with 6 one-arm and 5 two-arm gestures (*), and rest position.

Most existing methods for the classification of underwater gestures heavily rely on hand pose classification. To cope with finger detection and tracking in underwater conditions, most of them use modified gloves, e.g. by adding color-tape [1] or strain sensors [2]. Others attempt to segment the hand directly in video images [3], [4]. All of them focus on the information communicated by the hand without exploiting the time dimension of the gesture.

This paper employs an alternative methodology, requiring no additional equipment, which concentrates on upper limb movements as opposed to a fixed final hand pose. The tracking

Manuscript received: April 19, 2024; Revised: July 25, 2024; Accepted: August 25, 2024.

This paper was recommended for publication by Editor Gentiane Venture upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the French PACA Région and NotiloPlus/Delair Marine company.

[1]Bilal Ghader, Claire Dune and Vincent Hugel are with COSMER Laboratory, UR N°201522018X, Université de Toulon, France, bghader@gmail.com, claire.dune@univ-tln.fr, vincent.hugel@univ-tln.fr.

[2]Eric Watelain is with the J-AP2S Laboratory UR N°201723207F, Université de Toulon, France, eric.watelain@univ-tln.fr.

Digital Object Identifier (DOI): see top of this page.

[a]https://github.com/LaboratoireCosmerTOULON/DTW_KNN_RNN

of upper limb motion is more reliable than the detection of fingers and hand poses when there is underwater turbidity and a distance between the diver and the drone. While the final position of the arm may be identical for several gestures, the trajectory is assumed to retain sufficient information to enable the classification of gestures.

Recently, progress in deep learning has led to the development of systems for 2D [5]–[7] or 3D [8] human pose estimation in video sequences. 2D methods have been exploited in water and are already giving encouraging results. In [9], the key points of the pose are extracted using DeepCutLab [10] and used to realign the robot with the diver. Similarly, in [11], [12], OpenPose [5] is used to extract human skeleton from video data. The skeleton can serve as a reference for the relative localization of several robots in a swarm in [11] and as the input of gesture classification in [12]. While results seem promising, the skeleton detection remains noisy, and the performance could be improved using more recent skeleton detectors [8].

In order to abstract from skeletal detection noise and focus on the validation of gesture classification based on arms alone, anatomical points were tracked using an underwater motion capture system inside a pool. Therefore, a new dataset was acquired for 14 divers performing a set of 11 gestures, for a total of 1078 individual gestures.Like most current research in the field of underwater human-robot interaction, we assume that the diver is in a vertical position [13], [14]. Based on this new dataset, this paper presents a novel two-step process for classifying gestures based on a well-chosen set of angles formed by the upper limbs. These features are chosen to be robust to a wider variety of diver positions, including the lying position.

The paper is organized as follows. Section II reviews the literature in the domain of underwater gesture recognition. Section III details the proposed two-stage gesture classification process. Section IV presents the experimental protocol, the database, and the classification results. Section V discusses the results, and section VI concludes the paper.

II. State of the art

Underwater Human Robots Interaction [15] is a subdomain of human robots interaction with its own specific challenges compared to airborne communication. A detailed review of airborne applications can be found in [16].

Water absorbs magnetic waves within the first few centimeters, making it impossible to use Wi-Fi or Bluetooth. Waterproof tablets can be used as input tools for divers, but they are either connected to the robot by a cable [17], or placed directly on the robot [18]. Specific diver's gloves with strain sensors can be designed [2], [19] to detect hand orientation and finger movement.

Most underwater robots use cameras and acoustic sensors to interact with divers. While acoustic sensors work well in marine environments, they currently lack the resolution and refresh rate needed to accurately capture a diver's movements. Cameras, on the other hand, are affected by water, which absorbs and reflects light differently depending on wavelength, and water turbidity increases image noise due to the particles present.

Underwater gesture recognition methods generally assume that the diver is standing in front of the camera [13], [14]. Indeed, the collaborative nature of the task implies that the diver and the robot intentionally position themselves to interact [9], [20]. The most common method involves gesture-based communication relying on the detection of hands in images. To ease finger pose tracking and classification in video sequences, the CADDY project equipped divers with color-taped gloves and defined a communication language [21]. This project produced a large dataset (more than 18,000 samples of 16 different gesture classes, with a sample mean of 1156 instances/class) [13], containing mainly stereo images of divers performing hand gestures, in several field trials in closed and open waters. Gesture classification was performed by a deterministic variant of Random Forest, Multi-Descriptor Nearest Class Mean Forests (NCMFs) chosen for its robustness, given the small sample datasets from underwater environments [1]. Later, [22]–[24] employed deep-learning classifiers on the CADDY dataset. [22] used transfer learning to assess AlexNet, VggNet, and GoogLeNet on the dataset, achieving a total accuracy of 95%. The authors of [24] are the only ones who took advantage of the stereo-camera in CADDY leveraging a bi-channel CNN. The classification phase consists of a decision tree employing multiple networks, which achieved 96% accuracy. The work of [23] brought more insights on the CADDY dataset, by training the data using various train splits of different sizes (from around 3,500 to 18,400 samples), and by testing sensitivity against artificial perturbations and data conditions. They experimented with multiple classification and feature extraction networks. All the results were compared to the original solution proposed by [1]. However, the performance heavily depends on the training data size. For the best performing network, results range from 50% for a 3,500 sample data set to 98% for the full 18,000 dataset. The network significantly surpasses the MD-NCMF (MultiDescriptor Nearest Class Mean Forests) performance only for the largest training set of 18,000 samples.

Without gloves, the assumption of bare hands can be used to detect skin color, as in [4], where a Convolutional Neural Network (CNN) classifier was used to infer the corresponding gesture, and tested against state-of-the-art networks (RCNN and SSD). The dataset contained 10 different gesture classes, used to define 30 different instructions. Their train dataset contained more than $5K$ frames per class, with $1K$ test frames. The images were acquired in a swimming pool with divers facing the camera. The accuracy of the proposed method reached $24/30$ instructions (80%), while RCNN recognized $29/30$ (97%).

The ScubaNetV2 [25] dataset contains 32 labeled gestures, and over 290,000 labeled frames, including frames with idle positions. The images were captured at sea by divers who stood in front of the camera and grasped an anchor with one arm. Using the hand detector described in [25], and a Yolov8m transfer learning method for training, SCUBANetV2 appeared to be quite successful in identifying gestures with recognition rates in the 80%-95% range, depending on the gesture.

This study differs from previous investigations of underwater gesture recognition in that it exclusively considers the temporal information embedded in arm movements, avoiding the use of hand shape and position data. This approach makes it possible to classify standard diver gestures when the static position of the hand is non-deterministic (for example, gestures such as "stabilize," "out of air," and "not well" share similar hand images). The proposed classifier uses upper-body skeleton tracking as input, which is expected to be more robust under challenging underwater visual conditions than dense image-based hand shape detection. Additionally, this method enables gesture recognition at greater distances.

III. METHOD

The overall pipeline of the proposed approach is described in Fig. 2. The first stage distinguishes between one-arm and two-arm gestures by analyzing the energy of the signals from both arms. The second stage implements three classification algorithms, namely Long Short Term Memory (LSTM), Gated Recurrent Units (GRU), and K-Nearest Neighbor on Dynamic Time Warping distance (DTW-KNN), which will serve as a baseline.

A. Geometric features extraction

In [26], *joint-lines* projections are found to be the optimal features for action classification, while angle features also prove to be competitive candidates. In addition, angles combined with LSTM are a popular solution. In [27], angle features are extracted from RGBD, and motion capture data are used to classify normal and pathological gaits. In [28], the input data are Joint Center Positions (JCP), which are estimated using an inertial suit, and from which angle features are extracted in order to classify the different actions. Taking into account these results, each arm movement is encoded by five angles, one at the elbow, three at the shoulder, and one at the wrist, which represents the rotation of the forearm along its axis.

Figure 3 describes the chosen anatomical points: shoulders-4/9, elbows-5/10, pelvis-14/15, and wrist pairs of points (6,8)/(11,13), while 7/12 (in blue) are calculated as center of the two wrist points.

Given three 3D points, $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$, the vectors $\mathbf{x} = \mathbf{a} - \mathbf{b}$ and $\mathbf{y} = \mathbf{c} - \mathbf{b}$ are used to define the angle relative to point $\mathbf{b}$ as

$$\theta = atan2(\|\mathbf{x} \times \mathbf{y}\|, \mathbf{x} \cdot \mathbf{y}) \tag{1}$$

On the right-hand side of the skeleton, the point tuple (7,5,4) is used to calculate the elbow angle. The point tuples (14,4,5), (9,4,5) and (9,4,7) are used to get the three shoulder angles. The wrist angle is calculated using segments (6,8), (7,5), and (4,5) with $\mathbf{x} = (6, 8)$ and $\mathbf{y} = (7, 5) \times (4, 5)$ in Eq. 1. The left angles are defined symmetrically.

These five angles provide enough information to represent the position of the diver's arm, assuming that the biomechanics of the human body constrain the arm position in front of the body.

B. Energy features extraction

When the gesture is performed in the air, the passive arm remains inert. However, underwater, the passive arm can be active to help balance keeping. If gesture classification is carried out with both arms, this balance keeping movement will encode information that will affect the classification process. Therefore, we proposed to add a gesture type separation to distinguish one-arm gestures from two-arm gestures. In the air, the signals corresponding to the passive arm have low variance, which can be easily detected and filtered out using a variance threshold [29].

To identify the arm encoding gesture information, two kinds of energy are calculated for each feature signal s, namely *amplitude energy* (E_s) and *kinetic energy* (VE_s).

$$E_s = \sum_{k=0}^{k=N} (s[k] - s[0])^2, \quad VE_s = \sum_{k=0}^{k=N} (v[k] - v[0])^2 \tag{2}$$

where N is the number of samples and $k \in \{1..N\}$, v is the time derivative estimation of the signal s obtained using an alpha-beta filter. The initial value $s[0]$ and $v[0]$ of the signal are subtracted to remove the initial offsets among subjects.

The energies of each characteristic feature are summed to obtain the amplitude energy for left and right arms, namely E_l and E_r. In the same way, kinetic energies are defined for the left and right arms, namely VE_l and VE_r.

C. Gesture type separation

Then, for each gesture, two energy-feature ratios r and rv are calculated:

$$\begin{aligned} r &= \frac{max(E_l, E_r) - min(E_l, E_r)}{max(E_l, E_r)} \\ rv &= \frac{max(VE_l, VE_r) - min(VE_l, VE_r)}{max(VE_l, VE_r)} \end{aligned} \tag{3}$$

One-arm gestures are expected to have ratios close to 1, because the inactive arm has much lower amplitude energy and lower kinetic energy. Two-arm gestures are expected to have ratios close to 0 as both arms are active and spend similar energies.

An SVM classification is applied to the (r, rv) ratios to separate one-arm from two-arm gestures. After separation, the non-active arm is withdrawn from the one-arm gestures.

D. DTW-KNN classification

The DTW algorithm calculates a similarity distance between two temporal sequences, which may have local variations [29]. A K-Nearest-Neighbors (KNN) technique is then used to classify the signal. Let S_1 and S_2 be respectively a $(N \times L)$ and $(M \times L)$ signal, with N and M being the signal lengths, and L the signal dimension. Let $S_1[i]$ be used as a shorthand notation for the i^{th} time step of the signal S_1, $S_1[i] = [S_1[i, 1], S_1[i, 2], ..., S_1[i, L]]^T$.

Here, the signal is a set of 5 angular amplitudes for one-arm gestures, and 10 angular amplitudes for two-arm gestures.

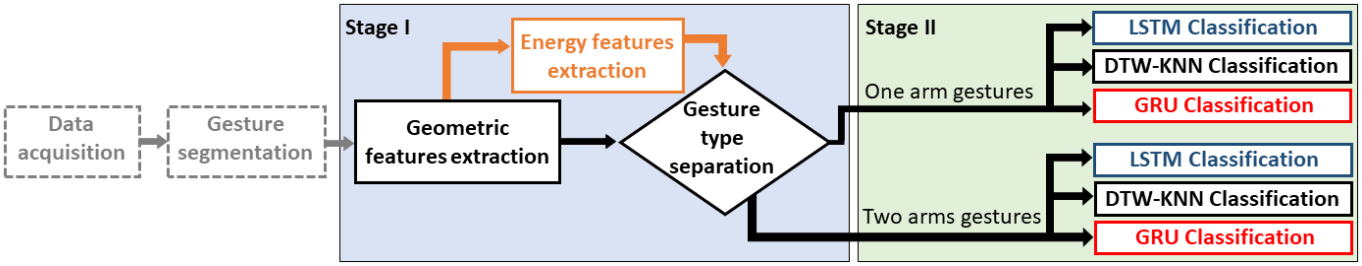


Figure 2: Classification pipeline: Stage I is an energy-based one-arm, two-arm gesture separation, and Stage II is a classification with DTW-KNN or RNN networks. Solutions for data acquisition and gesture segmentation can be found in [5]–[12].

Then, the distance is kept in $[0, \pi]$, and the local distance difference between the signals is computed as follows:

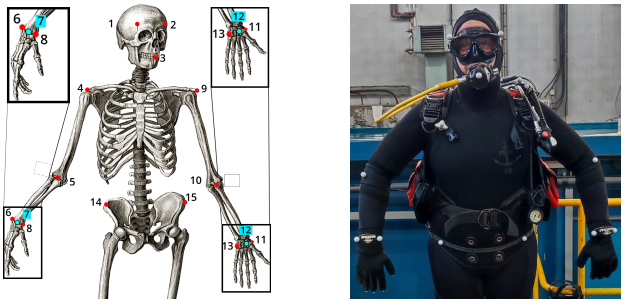
$$d(i,j) = \sum_{l \in 1..L} [(S_1[i,l] - S_2[j,l] + 3\pi)\%2\pi] + \pi \quad (4)$$

with S_1 and S_2 in radians, and the subscript l referring to the dimension of the gesture.

Two KNN classifications are then performed with different data splits, an *intra-subject*, and an *inter-subject* classification. In the *intra-subject* classification, the distance of each signal performed by a subject A is calculated with respect to all the signals performed by the same subject A. This classification is useful to quantify underwater gesture repeatability, which may be affected by sensory changes induced by submerging. It can also reflect dataset specificity. The number of signals per diver varies a lot, with a mean of 68 signals/diver and a median of 57 signals/diver. The number of nearest neighbors K is set to 3 due to the limited number of signals per diver. A leave-one-out test protocol is used for the DTW-KNN classification to allow for the deterministic measurement of *inter-subject* variability and its impact on gesture classification. The value of K is increased to 10 as the number of data increases.

E. Recurrent Neural Network (RNN) classification

RNN classification is preceded by a resampling step to have standardized input size of 400 samples, and a normalization step with a 0 average and a standard deviation of 1. The implemented RNN networks share a common architecture composed of two-layer bidirectional RNN with dropout layers in-between, and a dense final layer (in white on Fig. 4). Each



(a) (b)

Figure 3: a) Anatomical model: tracked markers in red and computed markers in blue. b) Diver's equipment.

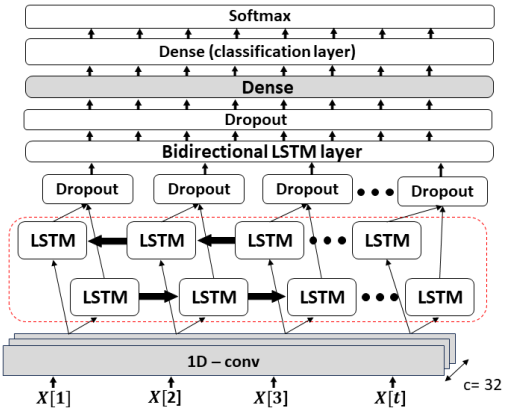


Figure 4: Architecture of the LSTM-CL network. Gray layers refer to additional layers w.r.t. the initial LSTM network.

layer includes 120 cells per direction, which makes 240 in total. This architecture was configured with LSTM or GRU cells.

LSTM and GRU configurations were augmented by two non-recurrent networks: a 1D convolution layer with a kernel size of 50 and 32 filters, and an additional dense layer inserted after the RNN layers but before the final classification layer (in gray on Fig. 4). The 1D convolutional layer was introduced to enhance the representation of local variations, while the extra dense layer was introduced to facilitate the exploration of more complex feature combinations before the classification layer. These modified networks are called LSTM-CL and GRU-CL.

F. Statistical Analysis

The results obtained are presented as average and standard deviation of classification accuracy in the next section. After verification of the normality conditions, the observation of a statistical significance ($p < 0.05$) of the differences between model type (DTW-KNN, LSTM, LSTM-CL, GRU, and GRU-CL) and arm separation (SVM) was conducted by a two-factor ANOVA test (Tab. I) as well as Tukey's post hoc tests (Tables II and III). A paired t-test Wilcoxon rank test was also carried out to compare each model with and without SVM. The Jamovi 2.5.6. Software was used for this analysis.

factor	Sum of Sq.	df	Mean Sq.	F	p
Arm sep.	2947	1	2947	27.36	$<$ **0.001**
Model	5793	4	1448	13.45	$<$ **0.001**
Arm sep.$\times$model	524	4	131	1.22	0.307
Residuals	14001	130	108		

Table I: Two-factor ANOVA analysis of the results.

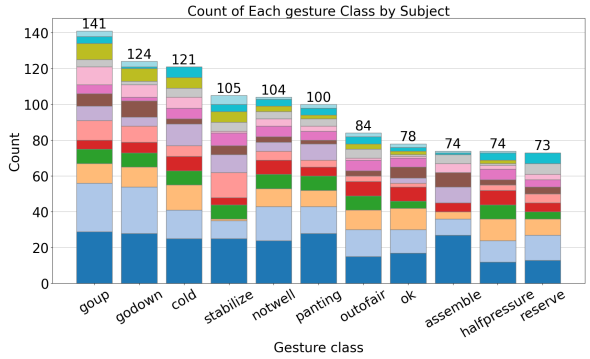


Figure 5: Distribution of the database (1078) per gesture (11) and color-coded subjects (14).

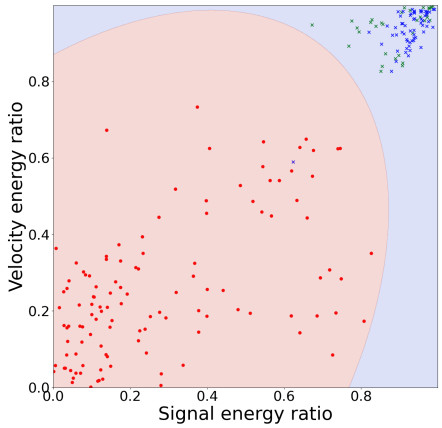


Figure 6: SVM separation results for one diver. Round, resp. crossed, data points refer to two-arm, resp. one-am gestures. Left, resp. right, arm in green, resp. blue.

IV. Experiments and Results

A. Data acquisition and pre-processing

From the 11 gestures selected to represent the variety of gestures found in standard diver communication, five of them are two-arm gestures, 6 are one-arm, 6 are oscillating, and 5 have long static phases, and they are performed at different heights (Table IV). Figure 5 shows the distribution of gestures by diver.

The data acquisition took place in a 2.5 m deep pool where gesture indications were displayed to divers through a waterproof tablet using an *Qualysis* underwater motion capture system. Shoulder, hip, elbow and wrist joints were tracked using specific reflective markers (Fig. 3b). No design modifications were requested for the gestures. Except divers 4, 5 and 7, all our divers are professional military divers who perform the gestures every day. The only imposed condition was that the diver's arms return to the rest position between two successive gestures.

B. Separation between one-arm and two-arm gestures

The SVM classification was evaluated through a leave-one-out protocol, resulting in 98% accuracy for the one-arm and two-arm gestures classification. Figure 6 showcases the classification corresponding to subject 3, the SVM being trained on all other diver gestures. Only one gesture was misclassified.

C. Classification Results

Table V displays the results of the *intra subject* classification, as well as the leave-one-out classifications for the DTW-KNN and the RNNs without SVM arm separation (LSTM, GRU, LSTM-CL and GRU-CL) and with SVM arm separation (SVM-DTW-KNN, SVM-LSTM, SVM-GRU, SVM-LSTM-CL and SVM-GRU-CL). For each method, an average classification accuracy across divers and gestures is provided. An average weighted by the number of gestures per diver is also reported. All the calculations were performed on a computer cluster with Intel Xeon Gold 6142 CPUs (2.6Ghz) under a Python 3.8.15 environment. The neural networks were implemented using the TensorFlow framework (2.4.1).

As expected, the best results are obtained with the SVM-DTW-KNN, with a top 92.14% classification rate. The next best rates are respectively 89.87% for the SVM-GRU-CL, *i.e.*, a GRU RNN with convolutional layer and SVM separation, and 89.46% for the SVM-LSTM-CL, *i.e.*, a LSTM RNN with convolutional layer and SVM separation. The results obtained with the two SVM-RNN-CL are close, and are only about 2.5% of the best rate obtained with DTW-KNN. The added-value brought by the SVM separation is between 4% and 15% depending on the model. The benefit of the additional layer depends on the presence of the SVM stage. It is more pronounced when GRU cells are used, with a rate increase of 20 to 22%, compared with a rate increase of only 7 to 12% in average for LSTM cells. These results show that the combination of the SVM separation stage and the additional convolutional layer improves the classification rate.

From a statistical point of view, significant differences are observed in the classification results brought by the SVM separation ($F = 9.56$, $p < 0.001$), and by the model type ($F = 13.44$, $p < 0.001$), without interaction between model and arm separation ($p = 0.307$). The model Post hoc analyses (Tab. III) indicate differences between GRU vs (GRU-CL and DTW-KNN and LSTM – CL) and LSTM vs (GRU-CL and LSTM-CL). The paired t-test Wilcoxon rank test shows improvements ($p < 0.05$) due to the SVM for all models except for the GRU-CL where the test is non-significant ($p = 0.099$).

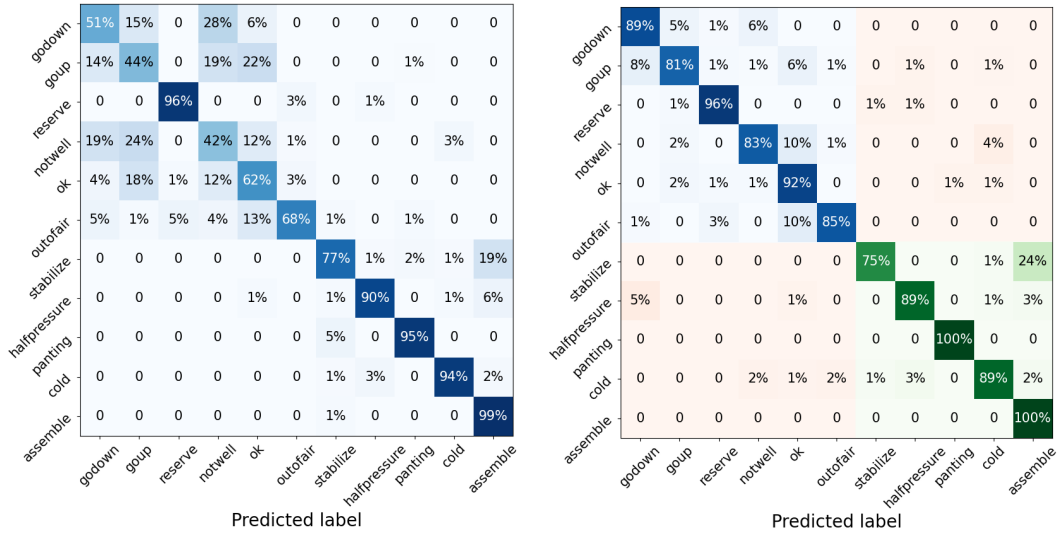
Confusion matrices relative to SVM-DTW-KNN, and SVM-LSTM-CL are presented in Fig. 7, and Fig. 8. Each of the confusion matrices is a compilation of the leave-one-out classifications. The confusion matrices with SVM are color-coded, with blue areas referring to one-arm gesture classifications, and green areas referring to two-arm gesture classifications.

no arm sep.		arm sep.	Mean Difference	SE	df	t	p_{tukey}	$p_{bonferroni}$
NOSVM	-	SVM	-9.18	1.75	130	-5.23	< 0.001	< 0.001

Table II: Post hoc comparison - arm separation factor.

model A		model B	Mean Difference	SE	df	t	p_{tukey}	$p_{bonferroni}$
DTW-KNN	-	GRU	13.63	2.77	130	4.912	$<$ **0.001**	$<$ **0.001**
	-	GRU-CL	-3.25	2.77	130	-1.173	0.767	1.000
	-	LSTM	7.62	2.77	130	2.749	0.052	0.068
	-	LSTM-CL	-1.97	2.77	130	-0.711	0.954	1.000
GRU	-	GRU-CL	-16.88	2.77	130	-6.086	$<$ **0.001**	$<$ **0.001**
	-	LSTM	-6.00	2.77	130	-2.163	0.200	0.323
	-	LSTM-CL	-15.60	2.77	130	-5.623	$<$ **0.001**	$<$ **0.001**
GRU-CL	-	LSTM	10.88	2.77	130	3.922	0.001	0.001
	-	LSTM-CL	1.28	2.77	130	0.462	0.991	1.000
LSTM	-	LSTM-CL	-9.60	2.77	130	-3.460	**0.006**	**0.007**

Table III: Post hoc comparisons - model factor.



(a) DTW-KNN: DTW-KNN classification without SVM separation.

(b) SVM-DTW-KNN: DTW-KNN classification, with SVM separation.

Figure 7: Compiled confusion matrices of the leave-one-out DTW-KNN classification baseline across subjects.

Gesture	1-arm	2-arms	Static	Oscill.	Level
Go down	x		x		S
Go up	x		x		S
Not well	x			x	S
Ok	x		x		S
Out of air	x			x	N
Reserve	x		x		H
Assemble		x		x	S
Cold		x		x	C
Half pressure		x	x		S
Panting		x		x	C
Stabilize		x		x	S

Table IV: Selected diver communication gestures to cover the variety of possible combinations between one-arm, two-arm, oscillating or static gestures and their execution height: head (H), shoulder (S), chest (C), neck (N).

V. Discussion

The results of the first column named 'DTW-KNN-intra' of Tab. V show that some divers have a low repeatability rate that obviously affects their *inter-diver* classification rates.

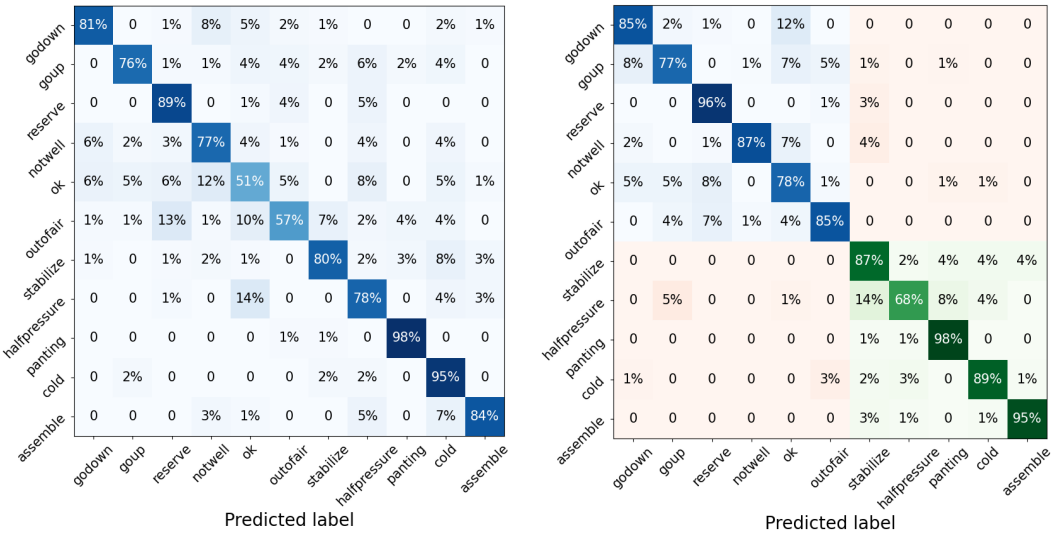
On average, the rate of the DTW-KNN leave-one-out classification, *i.e. inter-subject* (column 2), is slightly less than the rate of the intra-subject DTW-KNN classification (column 1), which is not surprising taking into account inter-diver variability. The results given in the second column can be considered as a deterministic classification baseline.

The red areas in the confusion matrices with SVM refer to the few incorrectly separated one-arm/two-arm gestures. Indeed, if a gesture is incorrectly classified in the first SVM separation phase, it will be misclassified later on. Despite this limitation, the classification with SVM separation outperforms the classification without SVM separation in all the classification methods. However, without this separation, DTW-KNN is behind RNN-CL methods.

The DTW-KNN cannot be used in real time, due to its computational demands. Each classification requires the comparison of the signal with all gestures in the training set, resulting in an exponential increase of the computation time relative to the size of the training set. Typically, the DTW-KNN classification of one gesture takes several tens of minutes. In contrast, RNN-based networks feature fast induction times, usually measured in milliseconds.

	Intra-subject	Leave-one-out Inter-subject										
subject	DTW-KNN-intra	DTW-KNN	SVM-DTW-KNN	LSTM	SVM-LSTM	LSTM-CL	SVM-LSTM-CL	GRU	SVM-GRU	GRU-CL	SVM-GRU-CL	# gest.
1	73.6%	78.1%	97.9%	71.7%	100.0%	89.6%	95.9%	75.7%	95.8%	95.3%	100.0%	40
2	65.1%	83.4%	89.8%	68.3%	71.1%	84.1%	88.5%	73.5%	66.1%	84.2%	91.2%	42
3	94.4%	80.0%	91.4%	76.7%	82.1%	87.2%	91.3%	67.8%	79.6%	90.8%	90.3%	243
4	94.2%	67.8%	78.0%	58.9%	73.0%	74.9%	80.6%	37.1%	54.3%	86.1%	84.8%	62
5	42.3%	57.9%	77.5%	75.0%	82.5%	90.0%	90.0%	76.3%	68.3%	95.0%	89.2%	20
6	54.6%	68.2%	90.7%	64.2%	78.1%	70.9%	79.6%	72.0%	74.4%	75.4%	88.3%	53
7	96.4%	93.0%	96.9%	96.9%	100.0%	94.1%	98.8%	91.2%	96.9%	97.4%	96.8%	61
8	70.7%	56.4%	96.3%	46.5%	80.2%	65.2%	90.8%	48.8%	53.1%	77.6%	93.8%	43
9	97.1%	67.1%	86.1%	66.2%	80.4%	83.4%	88.4%	51.9%	80.5%	89.1%	82.1%	176
10	100.0%	76.8%	100.0%	87.8%	89.5%	95.1%	96.3%	67.3%	80.1%	96.3%	89.2%	72
11	52.3%	93.5%	98.7%	77.3%	86.3%	84.3%	92.0%	66.8%	86.7%	89.0%	91.8%	40
12	87.3%	87.0%	100.0%	76.9%	74.4%	88.1%	87.6%	74.2%	81.5%	68.5%	84.4%	52
13	93.2%	74.3%	92.8%	56.6%	69.6%	78.5%	86.1%	49.7%	65.0%	67.1%	83.3%	104
14	97.5%	93.7%	93.9%	81.0%	82.5%	84.5%	86.6%	69.0%	82.1%	88.3%	93.0%	70
average	79.9%	76.9%	**92.1%**	71.7%	82.1%	83.6%	89.5%	65.8%	76.0%	85.7%	**89.9%**	77
weight. avg	87.7%	77.0%	**91.7%**	73.4%	83.3%	84.1%	**89.5%**	63.5%	77.0%	86.1%	88.6%	
std. dev.	18.52%	11.59%	6.84%	12.12%	8.78%	7.98%	5.09%	13.19%	12.48%	9.31%	4.80%	

Table V: Results of all classifications.



(a) LSTM-CL : LSTM with convolutional layer , without SVM separation.

(b) SVM-LSTM-CL : LSTM with convolutional layer and SVM separation.

Figure 8: Compiled confusion matrices of the leave-one-out LSTM with convolutional layer classifications across subjects.

The neural networks that use the SVM separation and the additional 1D convolutional layer stand out as the most effective. The improvement mainly concerns one-arm gestures. Without SVM separation, one-arm gestures are mostly confused with each other, whereas with SVM separation, one-arm gestures show a significant improvement due to the elimination of the inactive hand, which removes unwanted information. It is worth noticing that other confused gestures remain very similar with and without SVM separation, with notably fewer classification errors for two-arm gestures.

Without the additional layers, LSTM networks outperform GRU networks. This can be explained by the LSTM structure, which offers better long-term memory than GRU. With the additional layers, the performances of GRU-CL and LSTM-CL networks are much more similar. The weakness of GRUs is offset by the additional layers. The contribution of the additional layers to the classification appears to be significant. It seems that while the LSTM is capable of efficiently extracting long-term information on the entire signal, it is not capable of doing so properly in the short term on our database. The addition of the 1D convolutional layers helps to tackle this issue.

Compared to the CADDY [23] [22] reference work on underwater gestures, there are two main differences. First, The CADDY dataset not only features gestures from real divers, but also introduces a novel gesture dictionary. This involves a learning curve for divers, potentially affecting their experience. Second, our method exclusively relies on upper limb movements, disregarding information encoded by hand shape, which could limit the amount of information available. For instance, a gesture like 'ok' (Fig. 1) would be much more recognizable if information about hand shape were available. The CADDY dataset uses colored glove markers for its construction, which is very helpful in the feature extraction phase. In addition, there is a substantial difference in the dataset size, 1080 samples in our case, compared with the 18,000 samples in the CADDY case. Notably, the performance of CADDY methods significantly decreases when the size of the training data is smaller [23], e.g. down to 64% for a data size of 6,600.

VI. Conclusion and Future Work

The DTW-KNN-based classifier is useful to understand the origin of remaining confusions between gestures, to establish a basis for comparison, and to design an improved classification pipeline based on deep learning techniques. LSTM-based or GRU-based classifiers with SVM separation can be used for real-time recognition of gestures. The average accuracy is slightly less than that obtained with the DTW-based method, but is still satisfactory, with an average rate of 89.9%.

This study presents a preliminary investigation into gesture recognition utilizing optimal motion capture data. The next step is to apply a recently developed skeleton tracker [5]–[8] on mono or stereo images to verify the proposed pipeline on real-life data [9], [12]. The long-term objective is to embed gesture recognition on an underwater autonomous robot for online underwater human-robot interaction. Future improvements include the automatic segmentation of gesture signals, and an increase of the number of gestures to be recognized. Automatic gesture segmentation can be achieved using energy features, like in the SVM separation, to detect the movement from the resting position. A gesture may be coupled with another to convey a complementary meaning, like the phrase *not well* that is often accompanied by a subsequent gesture indicating the affected body part. This observation underscores the necessity for a comprehensive phrasebook that encompasses the full range of gesture combinations.

This study shows that while humans rely on oscillating or static movements to differentiate gestures, RNNs appear to be more responsive to gesture height and the use of one or two arms. Thus, a diver-robot interaction dictionary should focus on these features to improve RNN-based gesture recognition.

ACKNOWLEDGMENT

The authors greatly thank the PACA region, Marine Nationale, the Mesocentre of Marseille, and NotiloPlus/Delair Marine company for their help and support.

References

[1] A. Gomez, C. Mueller, T. Doernbach, D. Chiarella, and A. Birk, "Robust gesture-based communication for underwater human-robot interaction in the context of search and rescue diver missions," in *Workshop on Human Aided Rob. IEEE/RSJ Int. Conf. on Int. Robots*, 2018.

[2] C. R. Walker, Nađ, D. W. O. Antillon, I. Kvasić, S. Rosset, N. Mišković, and I. A. Anderson, "Diver-robot communication glove using sensor-based gesture recognition," *IEEE Journal of Oceanic Engineering*, vol. 48, no. 3, pp. 778–788, 2023.

[3] R. Codd-Downey and M. Jenkin, "Human robot interaction using diver hand signals," in *ACM/IEEE Int. Conf. on Human-Robot Interaction*. IEEE Press, 2019, p. 550–551.

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

[5] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 43, no. 01, pp. 172–186, Jan. 2021.

[6] 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.

[7] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[8] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, "Motionbert: A unified perspective on learning human motion representations," in *Proceedings of the IEEE/CVF Int. Conf. on Computer Vision*, 2023.

[9] D. Kutzke, A. Wariar, and J. Sattar, "Autonomous robotic re-alignment for face-to-face underwater human-robot interaction," in *IEEE Int. Conf. on Robotics and Automation*, 2024.

[10] 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, pp. 1281–1289, Sep 2018.

[11] 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.

[12] B. Ghader, C. Dune, E. Watelain, and V. Hugel, "Skeleton-based visual recognition of diver's gesture," in *OCEANS 2023 - Limerick*, 2023, pp. 1–5.

[13] A. Gomez C., 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," *J. of Marine Science and Engeeniring*, vol. 7, no. 1, p. 16, Jan. 2019.

[14] R. Codd-Downey and M. Jenkin, "Recognizing diver hand gestures for human to robot communication underwater," in *2023 32nd IEEE Int. Conf. on Robot and Human Interactive Communication (RO-MAN)*, 2023, pp. 92–98.

[15] A. Birk, "A survey of underwater human-robot interaction (u-hri)," *Current Robotics Reports*, vol. 3, no. 4, pp. 199–211, Dec 2022.

[16] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3200–3225, 2023.

[17] B. Verzijlenberg and M. Jenkin, "Swimming with robots: Human robot communication at depth," in *IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, 2010, pp. 4023–4028.

[18] N. Mišković, A. Pascoal, M. Bibuli, M. Caccia, J. A. Neasham, A. Birk, M. Egi, K. Grammer, A. Marroni, A. Vasilijević, Nađ, and Z. Vukić, "Caddy project, year 3: The final validation trials," in *OCEANS 2017 - Aberdeen*, 2017, pp. 1–5.

[19] J. Liu, L. Wang, R. Xu, X. Zhang, J. Zhao, H. Liu, F. Chen, L. Qu, and M. Tian, "Underwater gesture recognition meta-gloves for marine immersive communication," *ACS Nano*, Apr 2024.

[20] M. J. Islam, M. Fulton, and J. Sattar, "Toward a generic diver-following algorithm: Balancing robustness and efficiency in deep visual detection," *IEEE Rob. and Automation Letters*, vol. 4, no. 1, pp. 113–120, 2018.

[21] D. Chiarella, M. Bibuli, G. Bruzzone, M. Caccia, A. Ranieri, E. Zereik, L. Marconi, and P. Cutugno, "Gesture-based language for diver-robot underwater interaction," in *OCEANS*, May 2015, pp. 1–9.

[22] J. Yang, J. P. Wilson, and S. Gupta, "Diver gesture recognition using deep learning for underwater human-robot interaction," in *OCEANS 2019 MTS/IEEE SEATTLE*, 2019, pp. 1–5.

[23] 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.

[24] J. Yang, J. P. Wilson, and S. Gupta, "Dare: Diver action recognition encoder for underwater human–robot interaction," *IEEE Access*, vol. 11, pp. 76 926–76 940, 2023.

[25] R. Codd-Downey and M. Jenkin, "Finding divers with scubanet," in *IEEE Int. Conf. on Robotics and Automation*, May 2019, pp. 5746–5751.

[26] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer lstm networks," in *2017 IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2017, pp. 148–157.

[27] M. Khokhlova, C. Migniot, A. Morozov, O. Sushkova, and A. Dipanda, "Normal and pathological gait classification LSTM model," *Artificial Intelligence in Medicine*, vol. 94, pp. 54–66, Mar. 2019.

[28] A. K. Singh, M. Adjel, V. Bonnet, R. Passama, and A. Cherubini, "A framework for recognizing industrial actions via joint angles," in *2022 IEEE-RAS 21st Int. Conf. on Humanoid Robots (Humanoids)*, 2022, pp. 210–216.

[29] G. A. Ten Holt, M. J. Reinders, and E. A. Hendriks, "Multi-dimensional dynamic time warping for gesture recognition," in *conf. of the Advance School for Comp. and Imaging*, vol. 300, 2007, p. 1.