Divers gesture recognition from upper limb tracking

Bilal Ghader, Claire Dune, Eric Watelain, Vincent Hugel

Divers gesture recognition from upper limb tracking

B. Ghader[1] C. Dune [1] E. Watelain [2] V. Hugel [1]

[1] Laboratoire COSMER, Université de Toulon
[2] Laboratoire IAPS, Université de Toulon

bilal-ghader@etud.univ-tln.fr

Résumé

La coopération sous-marine homme-robot nécessite la compréhension par le système robotique de la gestuelle des plongeurs. L'environnement sous-marin représente un défi, car la turbidité et l'absorption des couleurs dégradent rapidement la qualité de la perception du robot. Ce travail présente une classification des gestes des plongeurs basée sur la détection du squelette du haut du corps, qui peut compléter la classification de la forme de la main. Cet article décrit un pipeline pour reconnaître les gestes à partir des variations angulaires du haut du corps. La solution proposée exploite la dimension temporelle des données d'angles articulaires en utilisant un réseau neuronal de type LSTM. L'approche proposée est validée sur un ensemble de données de capture de mouvement relatives à 15 sujets à qui on a demandé d'exécuter 8 gestes de plongeur différents dans l'air. Les résultats obtenus montrent la capacité de l'algorithme à discriminer la plupart des différents gestes entraînés avec un taux de réussite jusqu'à 88%.

Mots Clef

Reconnaissance des gestes, signes de plongées sous-marine, interaction homme-robot, classification LSTM, caractéristiques géométriques

Abstract

The need for a robot to understand basic commands is crucial for human-robot underwater missions. Yet, the underwater environment is a challenge because turbidity and color absorption quickly degrade the quality of the robot perception. This work presents a classification of diver gestures based on upper body skeleton detection, which can be used in addition to hand shape classification. This paper depicts a pipeline to recognize gestures from angular variations of the upper body. The proposed solution exploits the temporal dimension of the joint angles data by using a LSTM-based neural network. The proposed approach has been validated on motion capture datasets relative to 15 subjects who were asked to execute 8 different diver gestures in the air. The results obtained prove the ability to discriminate most of the different gestures trained with a success rate of up to 88%.

Keywords

Gesture recognition, underwater diving signals, human robot interaction, LSTM classifier, geometric features

1 Introduction

Depending on the depth, duration and complexity of the operations to be carried out, underwater missions are either performed by divers or underwater robots. Usually, robots are Remotely Operated Vehicles, controlled by a pilot from a surface vessel. Recent advances have given Autonomous Underwater Vehicles (AUV) the computation power that is crucial for autonomy in tasks of navigation and localization. However, human-robot interaction appears to be vital for more complex applications where the judgement of human operators is required on site.

In underwater environments, the main bottleneck remains the ability of the diver and the AUV to exchange information in an understandable way (the AUV's understanding of the diver's commands or information, and the notification to the diver of its correct understanding). Most of the current solutions require additional devices/equipments, e.g. tablets for input or confirmation [1]. Unfortunately, these solutions are not very intuitive and require the diver to accept adaptations to communicate with the robot, which can be tedious for him.

Regarding diver's gesture recognition, a majority of existing studies rely on hand detection in monocular images. Hand detectors usually involve either skin color-segmentation [2] or predefined colors for the finger to be easily detectable, as in [3] where the authors put yellow markers on fingers. Other approaches aim at simplifying the classification by developing a new sign language, which leads to even more complex commands [2, 4].

This study assumes that the skeleton detection can be achieved by techniques like OpenPose [5] as in [6]. The setup uses a motion capture system to simulate a perfect skeleton detection.

The main contribution of this paper is the classification of diver gestures based on upper limbs' movements, hand excluded, by exploiting time frames of the orientation of the

forearm and the arm with respect to the pelvis, and to identify the different types of gestures. This method does not require the use of additional equipment. In addition to the intuitive nature of the proposed classification method, another advantage is that it does not have to take into account any physical characteristics of the subject (height, skin, color etc.).
Section 2 focuses on the state of the art in gesture classification, starting with airborne applications, which represent the mainstream application domain, then focusing on existing approaches in underwater gesture recognition. Section 3 describes the methodology regarding data processing, and the architecture of the neural network used. Sections 4 and 5 present the experimental setup and the results, respectively. Conclusion and future work constitute the last section of this paper.

2 Related work

2.1 Airborne application of gesture recognition

Gesture recognition is a very popular field of research, as shown by the multiple surveys [7, 8, 9, 10] carried out in this field. The work developed in [9] presents the different stages of visual gesture recognition and their limitations. While this work was limited to visual gestures, the same logic applies to the general gesture recognition problem. The three different stages are (1) detection-data acquisition-pre-processing, (2) gesture representation and feature extraction, and (3) gesture classification. The data acquisition and pre-processing phase results from multiple approaches of how the gesture is acquired. Actually, [7] distinguishes two methodologies, based on the data acquisition techniques, namely glove-sensors-based (wearables instrumented with different sensors such as strain sensors or IMUs in order to reconstruct the movement of the hand), and vision-based. The work of [9] divides the vision-based approach into multiple techniques based on the nature of the acquisition instrument (monocular, stereo, RGBD ...). The pre-processing consists of data treatments, preparing it for feature extraction, such as image segmentation in case of visual data. For instance, skin-color segmentation, or segmentation based on movements, are techniques employed for hand gestures [7, 10].
The second phase, namely gesture representation, depends on the model used to represent the gestures. Different approaches can be noted, based on multiple selection criteria, whether the model is 3D or 2D [7, 8, 9], and is static or dynamic [9, 10]. This phase goes hand in hand with the feature extraction phase of the algorithms, since they are interdependent. Some examples of the model choices are: Deformable Gabrait model, or silhouette Geometry model in the case of 2D gestures, and skeleton model, or motion models in the case of 3D gestures [8, 9].
Finally, the recognition/classification phase implies different classifiers choices, starting with traditional approaches such as curve fitting, Finite State Machine (FSM), Hidden Markov Model (HMM) [7, 9, 10], then unsupervised (K-means) [7] supervised classifiers such Support Vector Machine (SVM), random forest, or K-Nearest Neighbors (K-NN) approaches [7, 10], and neural networks [7, 8, 9, 10]. The work of [8] focuses on deep-learning in sequence of images. It explores different techniques to exploit the temporal dimension of the data. In fact, several approaches are possible to achieve this goal, such as using 3D convolutional neural networks, fusion strategies (with streams focused on the temporal dimension for instance), or using a temporal deep learning model such as Recurrent Neural Networks (RNN), or Long Short Term Memory networks (LSTM).
In particular, gait analysis in biomechanics can be related to gesture classification to some extent. For example, for normal/pathological gaits' detection [11], angles of lower limbs were estimated from kinect data and classified using a LSTM neural network. Our method is based on a similar classification technique.

2.2 Underwater human robot interactions

In underwater environments, gesture recognition can also be achieved using a multiple acquisition device. For instance, [12] proposes a smart glove in order to detect the different gestures and classify them. The glove includes multiple strain sensors that detect finger movement, and IMU units that determine the orientation of the hand.
In the work presented in [13], ScubaNet, a dataset of divers making different gestures in front of an underwater vehicle was constructed. Although the work presented in [13] is limited to the detection of the diver and his two hands, the authors present in [14] a transfer learning method that was tested on a multiple convolutional neural network architecture in order to classify the different gestures. In this work, the classification is carried out in two steps, the first step consists of isolating hand images, while the second step proceeds to their classification.
Similarly, [4] aimed to create a data set of stereo images of divers executing gestures. The dataset was based on a specifically developed communication language presented in [1]. It also included a part with stereo-footage-synchronized IMUs measurements located throughout the diver's suit (DiverNet) which serves as a ground-truth for human pose and tracking methods. The work of [3] also aims to identify the hands of the diver, to feed a classifier pipeline to obtain the gesture command.
In another work presented in [2], the authors start by extracting the hand patches of the diver, and resize the hand images into 32×32 images, to feed a Convolutional Neural Network (CNN) classifier to infer the corresponding gesture. The gesture sets are then used into an FSM to decode a set of instructions.
These works base their gesture recognition on a first stage of hand detection and tracking. In turbid water with the diver equipped with black gloves on a black suit, hand de-

(a) Tracked points (b) Cameras setup

Figure 1 – Motion capture system.

tection can be difficult or even fail. However, there are many airborne silhouette-tracking approaches that represent the human as a skeleton that could be adapted to underwater, such as the work presented in [15] or [16]. Unlike sign language, most diving gestures involve large, distinct arm movements. The recognition of gestures from the movement of arms and forearms can be more robust to the conditions imposed by the aquatic environment. This paper aims to identify which gestures can be recognized without using hand tracking.

3 Methodology

The method introduced in this paper exploits the upper limbs angle trajectories instead of relying on a static hand image to classify gestures. This section introduces the methodology used to (1) acquire data, (2) extract features, and (3) classify the divers' gestures.

3.1 Data acquisition

In order to obtain the dataset, a *Qualisys* motion capture system was put in place. The setup (Fig. 1b) contains 8 cameras that were placed in an arc configuration to capture the central portion of the room, at different heights to cover all the upper part of the body.

Several keypoints were tracked using reflective markers, namely both shoulder joints, both elbows joints, both wrists, and two lateral pelvic points (Fig. 1a, 3). Three additional markers were placed on the face side and chin. They are not used in the gesture description.

Each subject is requested to perform different gestures, and the recording was then segmented using the *Qualisys Track Manager Software*. It allows extrapolating the data in case one of the tracked points was lost for a short period of time. Since our goal is to exploit the temporal variation of the different angles, each gesture frame was defined by a start position and an end position. The beginning and end positions were manually identified using the *Qualisys Track Manager Software*. The frame starts from a predetermined reference position of the user's arm (Fig. 1a). The frame ends when the arm comes back to that predefined initial position. In order to have a uniform input length for the classification method, all the data samples were resampled to a constant sample length of $n = 400$.

3.2 Features selection and extraction

In order to fully describe the motion of the shoulders, arms and forearms, three different orientations were tracked for both left and right sides (Fig. 3). The first two rotations are those formed by the (pelvis-10, shoulder-4, elbow-5) tuple, and the (shoulder-4, elbow-5, wrist-6) tuple representing two real joints, which are presented in Fig. 3 by the blue and yellow angles respectively. In addition, a third virtual joint defined by the tuple (pelvis-10, shoulder-4, wrist-6) was tracked since it encodes relevant information about the rotation of the forearm around the arm. The corresponding rotation angle is displayed in green on Fig. 3. The elbow joint has one degree of freedom only, while both shoulder angles have three degrees of freedom each. The same process is carried out symmetrically on the left side of the body.

To begin with, all tracked points have to be expressed in a frame attached to the subject body to be independent of the subject position or orientation in the motion capture coordinate frame. A local frame attached to the right side of the pelvis is defined (point 10, Fig. 1a). Two vectors $\mathbf{u}$ and $\mathbf{v}$ are defined as the normalized unit vectors going from the right side of the pelvis to the left side (tuple $10; 11$ on Fig. 1a) and from the right side of the pelvis to the right shoulder (tuple $10; 4$ on Fig. 1a) respectively.

The frame $(\mathbf{u}, \mathbf{v}, \mathbf{w})$ is defined, where $\mathbf{w}$ is the cross-product of $\mathbf{u}$ and $\mathbf{v}$. In order to obtain an orthonormal frame, the vector $\mathbf{v}$ is replaced by $\mathbf{v}' = \mathbf{w} \times \mathbf{u}$, defined as the cross-product of $\mathbf{w}$ and $\mathbf{u}$. We obtain the orthonormal frame $(\mathbf{u}, \mathbf{v}', \mathbf{w})$, where $\mathbf{u}, \mathbf{v}'$ and $\mathbf{w}$ are defined in the motion capture reference frame ${}^{w}\mathcal{F}$, the rotation matrix between the motion capture frame and the body frame ${}^{b}\mathcal{F}$ is the following:

$$ {}^{b}\mathbf{R}_{w} = [\mathbf{u}|\mathbf{v}'|\mathbf{w}] \tag{1} $$

In order to describe the relative orientation of two successive body segments, the axis-angle (*aka* rotation-vector) representation was chosen. This representation allows the parametrization of a rotation in a three-dimensional space, this is done using two entities, a unit vector $\mathbf{e}$ representing the direction of the axis of rotation, and an angle value θ. This representation can either be expressed as an ordered pair $([e_x, e_y, e_z]^T, \theta)$ or as one entity $\mathbf{e}\theta$ [17].

In the axis-angle representation, the vector $\mathbf{e}$ is the unit vector pointing outwards, and θ is the angle value. Starting with three 3D points, $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$, in order to obtain the axis-angle representation of the angle formed the vectors $\mathbf{x} = \mathbf{a} - \mathbf{b}$ and $\mathbf{y} = \mathbf{c} - \mathbf{b}$, two entities are needed, the vector $\mathbf{e}$ and the rotation angle θ. $\mathbf{e}$ is defined as the cross-product of $\mathbf{x}$ with $\mathbf{y}$. To obtain $\theta \in [0; \pi]$, we compute the value of $\sin\theta$ and $\cos\theta$:

$$ \sin\theta = \frac{\|\mathbf{x} \times \mathbf{y}\|}{\|\mathbf{x}\|\|\mathbf{y}\|} \tag{2} $$

$$ \cos\theta = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} \tag{3} $$

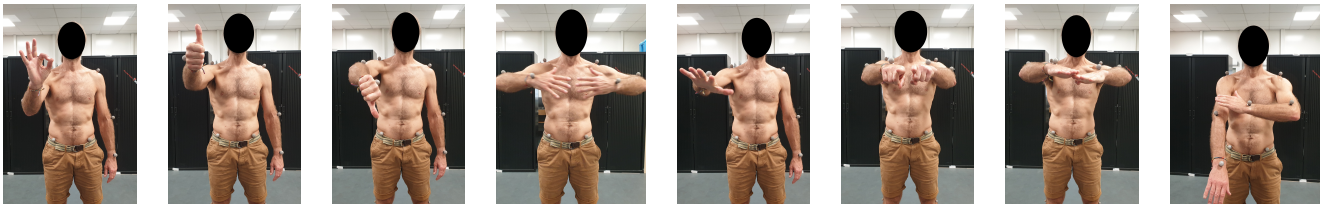


Figure 2 – The gesture shown from top left are: *ok, go up, go down, panting, not well, assemble, stabilize* and *cold*.

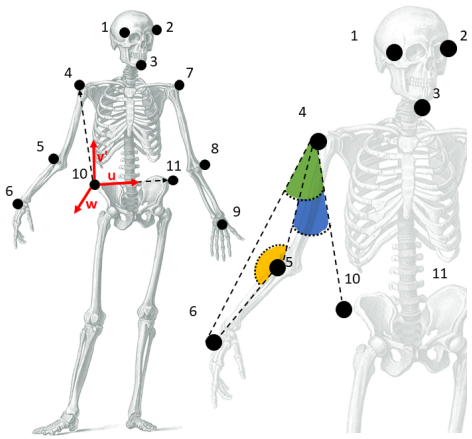


Figure 3 – Elbow angles (4,5,6) and (7,8,9), shoulder angles (10,4,5) and (11,7,8), virtual wrist-shoulders angles (10,4,6) and (11,7,9).

The angle θ is then computed using the $atan2$ function:

$$\theta = \arctan 2(\sin\theta, \cos\theta) \tag{4}$$

Therefore, the motion representation is composed of 2 axis-angle representations (angles in green and blue on Fig. 3) and one scalar angle value (the angle of the elbow in yellow in Fig. 3). Each angle-axis representation involves 4 values in the ordered pair representation. We have 9 scalar values (4 of which are independent) for each arm, with a total of 18 for both arms. Each gesture consists of $18\times n$ values, where n is the normalized number of samples.

3.3 Data augmentation

As diver gestures can be done indifferently with the right hand or the left hand depending on whether people are right or left-handed, a mirroring operation was carried out, allowing to flip the left and right sides of the body of the subject.

The aim is to obtain an additional set of points that is symmetrical to an acquired set of 3D points, where each point homogeneous representation in the world frame ${}^{w}\mathcal{F}$ is ${}^{w}\mathbf{P} = ({}^{w}X, {}^{w}Y, {}^{w}Z, 1)^{T}$. The origin of the frame ${}^{s}\mathcal{F}$ is defined as the midpoint of the two points defined by the right and left pelvic points (points 10 and 11 on Fig. 1a). The base of this frame is defined by the three vectors $\mathbf{u}, \mathbf{v}', \mathbf{w}$ previously defined.

We represent the position and orientation of the frame ${}^{s}\mathcal{F}$ in the world frame ${}^{w}\mathcal{F}$ by the homogenous matrix:

$${}^{w}\mathbf{M}_{s} = \begin{bmatrix} {}^{w}\mathbf{R}_{s} & {}^{w}\mathbf{t}_{s} \\ 0 & 1 \end{bmatrix} \tag{5}$$

With ${}^{w}\mathbf{R}_{s}$ and $\mathbf{t}_{s}$ being respectively the rotation and translation from ${}^{s}\mathcal{F}$ to ${}^{w}\mathcal{F}$. All tracked points ${}^{w}\mathbf{P}_{i}, i \in [1, 11]$ are expressed into ${}^{s}\mathcal{F}$ through:

$${}^{s}\mathbf{P}_{i} = {}^{s}\mathbf{M}_{w}\,{}^{w}\mathbf{P}_{i} = {}^{w}\mathbf{M}_{s}^{-1}\,{}^{w}\mathbf{P}_{i} \tag{6}$$

Then they are mirrored across the plane $(\mathbf{v}', \mathbf{w})$

$${}^{s}\mathbf{Pm}_{i} = (-{}^{s}X, {}^{s}Y, {}^{s}Z, 1)^{T} \tag{7}$$

Finally, they are transformed back into the world frame ${}^{w}\mathcal{F}$

$${}^{w}\mathbf{Pm}_{i} = {}^{w}\mathbf{M}_{s}\,{}^{s}\mathbf{Pm}_{i} \tag{8}$$

This set of point is then processed to compute the needed axis-angle definition for the data set inputs.

3.4 Neural network architecture

After extraction and resampling to have standardized input size, the angle trajectories are classified using a deep learning method. The neural network architecture used for gesture classification is similar to the work described in [11]. It is a 2 layers bidirectional LSTM with dropout layers in-between and a dense layer as last layer. The full architecture is depicted on Fig. 4.

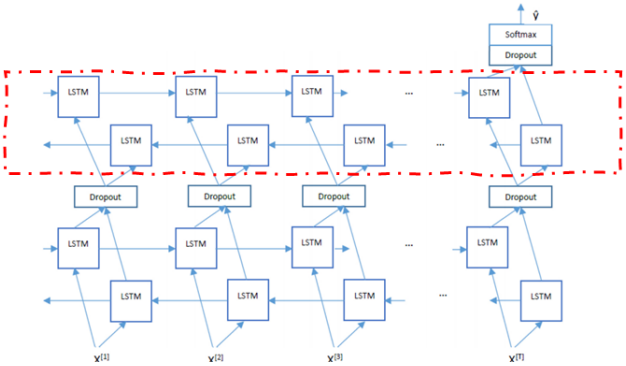


Figure 4 – Architecture of the neural network used for classification. Each layer is formed by 2 layers of 256 LSTM cells, to allow bidirectionality. The LSTM cells inside the red dotted line correspond to one bidirectional layer. Each of the two layers is fed with the output of the dropout layer, with the difference being the directionality of the layer.

4 Experimental setup and protocol

4.1 Population presentation

The dataset was recorded with 15 different subjects: 10 males and 5 females. The subjects' age range from 18 to 69 (average age = 34, standard deviation = 16). Five subjects of the population were not initiated to diving. The different gestures were presented to them in a briefing before recording the series. The rest of the population included 3 beginners (less than 20 dives) and 7 experimented divers.

4.2 Experimental protocol

The number of gestures was limited to the 8 most common diver gestures, namely *go up, go down, ok, not well, stabilize, assemble, cold,* and *panting*, which are illustrated on Fig. 2. Four of them require both arms: *stabilize, assemble, cold* and *panting*.
The subjects were asked to start and finish each gesture from a resting position with arms at the side of the body (Fig. 1a). They were requested to perform the two-armed gestures 10 times, and the one-armed gestures, 20 times, 10 times with each hand.

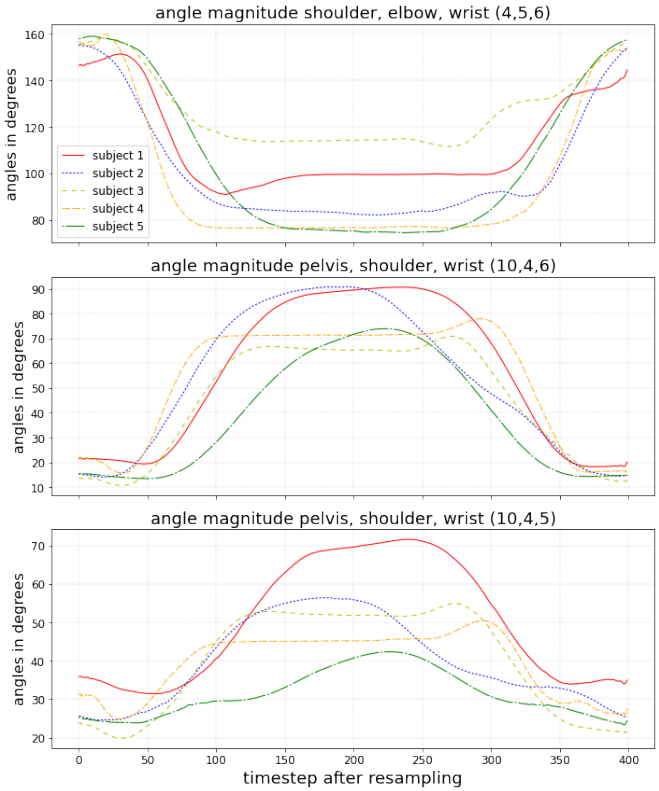
5 Results

5.1 Training

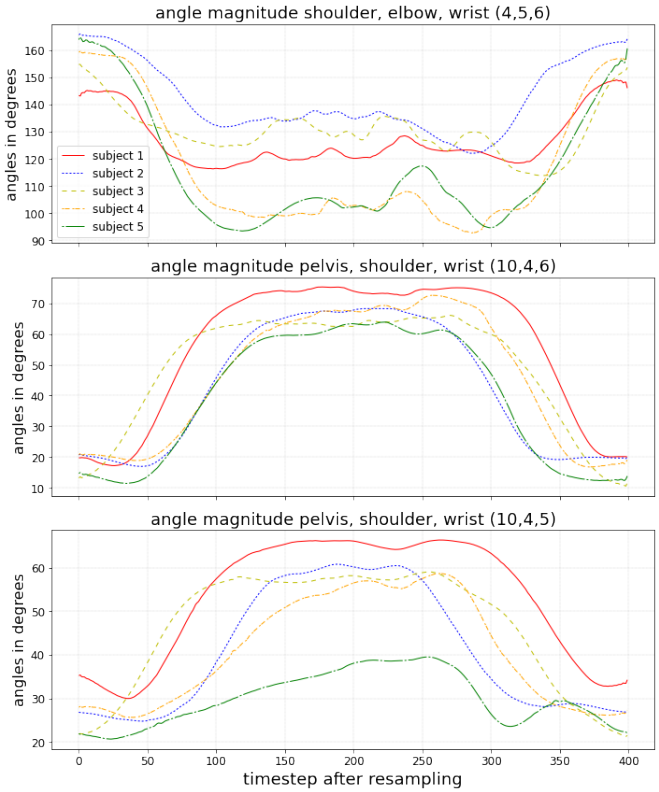
The network described in 3.4 was implemented using the TensorFlow-Keras framework. The dataset was divided into training, validation, and testing, representing 70%, 20%, and 10% of the data respectively. Data division was achieved randomly, regardless of the subject. The training process was repeated more than once to validate repeatability.
The width of the LSTM layers was chosen after iterative runs to have 256 cells, while the dropout rate was chosen to be 55%. Actually, the less LSTM cells per layer, the harder it is for the network to converge, while more LSTM cells per layers lead to much too long training times without any significant change in accuracy. Similarly, the lower the dropout rate, the more risk of over fitting on the training data. This translates into a better accuracy on the training data, but a lesser accuracy on the validation data. With a dropout rate significantly higher, the network has convergence problems.
The neural network was trained over 40 epochs. The accuracy results over the last training epochs were 92% and 90% for the training and the validation data, respectively.

5.2 Testing

The inference accuracy of the testing data is 88%. Figure 6 depicts the confusion matrix. The inference has an accuracy that is greater than 90% with two exceptions for the *not well* and *go up* gestures. Actually, these two gestures can be mistaken by the algorithm, with 17% of *not well* cases labeled as *go up* and 17% *go up* labeled as *not well*. Additional minor confusions are, for example, 9% of the *assemble* gestures incorrectly labeled as *stabilize* gestures.



(a) *Go up* movement



(b) *Not well* movement

Figure 5 – Angles magnitude.

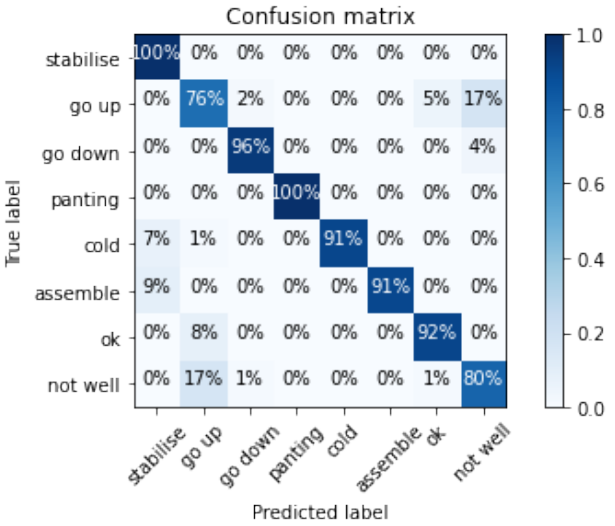
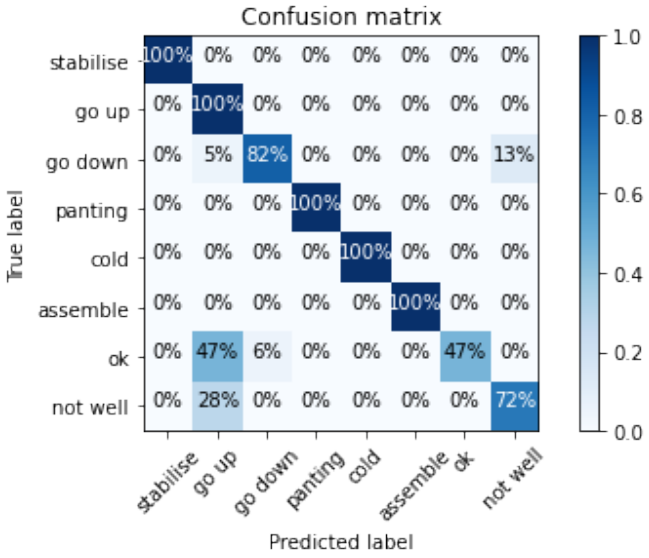


Figure 6 – Confusion matrices. Vertical axis labels are the true labels, and horizontal axis labels are the predicted labels.

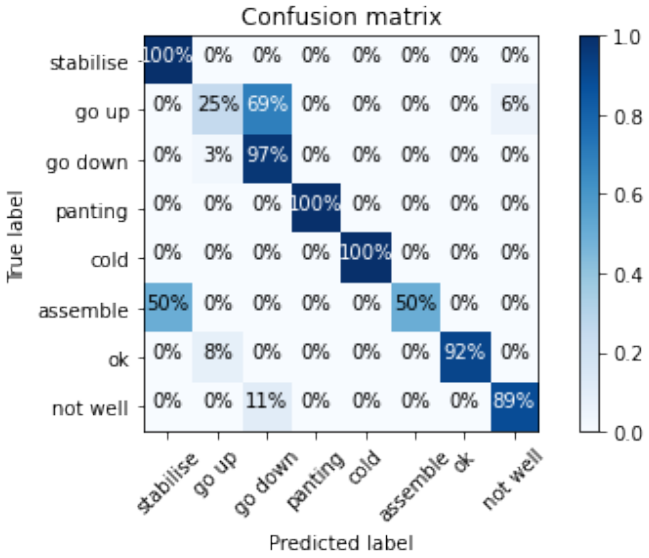
The analysis of the magnitude variations of the tracked angles (Fig. 5) shows that the variations of the different angles remain very similar among the 5 subjects selected randomly. However, there is a slight notable difference in the angle (4,5,6) of the *not well* gesture that features some oscillation. This could explain why the classifier was not capable of separating the *go up* and *not well* gestures.
In order to avoid over-fitting of specific features related to the subjects in the dataset and to confirm the inter-person transferability of our model, the data of a particular subject was removed from the initial data. The entire training and validation process was achieved again, but without the data of this particular subject. This process was conducted for two different subjects, one female (A) and one male (B), a beginner and a confirmed diver respectively. The results are presented on Fig. 7. As expected, these results are not as good as those presented on Fig. 6. We note a very high accuracy (more than 80%) for 6 out of the 8 gestures for each subject, showing a capacity of transfer by our classification. We note a confusion between *ok* and *go up* for subject A, and between *assemble* and *stabilize* for subject B. These confusions can be explained by the fact that the movement of these gestures are similar in real life, actually the same confusions can be found in Fig. 6 with smaller magnitude. One additional confusion for subject B, namely gesture *go up*, that is incorrectly classified as *go down*, can be explained as a subject-dependent variation, as it does not appear with subject A.

6 Conclusion and Future work

This paper presents a new approach regarding diver gesture classification that only relies on the 3D skeleton pose without exploiting the data encoded by the hands. This was made possible by exploiting the temporal dimension



(a) Subject A.



(b) Subject B.

Figure 7 – Leave-one-out results for two subjects.

of the upper limb angular data. The classification was achieved using a neural network based on a LSTM architecture, since it allows processing sequences of data. The results obtained show that it is possible to distinguish between 6 different gestures out of the 8 under study, while two of them, namely *go up* and *ok*, could be mistaken due to the similarity of the limb trajectories. Hand shape recognition and identification will therefore be required to make the decision between similar arm/forearm gestures. In addition, the solution proposed gives satisfactory results with respect to inter-subject transferability, with a potential for improvements.

The application of our approach to the underwater environment still needs to be tested. This requires to be able to achieve an accurate estimation of the diver skeleton on underwater images. Promising vision-based skeleton detection has been recently applied in underwater conditions [6], which will be tested to evaluate our classifier on 2D skeletons.

In addition, the accuracy of the present classifier could be augmented using a visual classifier, to be used when the confidence in the pose-exclusive classifier is not enough. The advantage of the mixed approach would be to keep the lightweight LSTM network, while taking advantage of the accuracy of the CNN that takes the image data as input.

References

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

[2] 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, pp. 851–873, 2019.

[3] C. A. Mueller, T. Fromm, A. Gomez Chavez, D. Koehntopp, and A. Birk, "Robust Continuous System Integration for Critical Deep-Sea Robot Operations Using Knowledge-Enabled Simulation in the Loop," in *International Conference on Intelligent Robots and Systems*, 2018.

[4] 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, p. 16, Jan. 2019, number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[5] Z. Cao, G. Hidalgo, T. Simon, S. 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. 01, pp. 172–186, jan 2021.

[6] 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

[7] M. Oudah, A. Al-Naji, and J. Chahl, "Hand gesture recognition based on computer vision: a review of techniques," *journal of Imaging*, vol. 6, no. 8, p. 73, 2020.

[8] M. Asadi-Aghbolaghi, A. Clapes, M. Bellantonio, H. J. Escalante, V. Ponce-López, X. Baró, I. Guyon, S. Kasaei, and S. Escalera, "A survey on deep learning based approaches for action and gesture recognition in image sequences," in *12th IEEE international conference on automatic face & gesture recognition (FG'2017)*. IEEE, 2017, pp. 476–483.

[9] B. K. Chakraborty, D. Sarma, M. K. Bhuyan, and K. F. MacDorman, "Review of constraints on vision-based gesture recognition for human–computer interaction," *IET Computer Vision*, vol. 12, no. 1, pp. 3–15, 2018.

[10] K. M. Sagayam and D. J. Hemanth, "Hand posture and gesture recognition techniques for virtual reality applications: a survey," *Virtual Reality*, vol. 21, no. 2, pp. 91–107, 2017.

[11] 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.

[12] Đ. Nađ, C. Walker, I. Kvasić, D. O. Antillon, N. Mišković, I. Anderson, and I. Lončar, "Towards advancing diver-robot interaction capabilities," *IFAC-PapersOnLine*, vol. 52, no. 21, pp. 199–204, 2019.

[13] R. Codd-Downey and M. Jenkin, "Finding divers with scubanet," in *International Conference on Robotics and Automation (ICRA'2019)*, 2019, pp. 5746–5751.

[14] ——, "Human robot interaction using diver hand signals," in *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '19. IEEE Press, 2019, p. 550–551.

[15] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "Vnect: Real-time 3d human pose estimation with a single rgb camera," in *ACM Transactions on Graphics*, vol. 36, no. 4, July 2017.

[16] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training,"

in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[17] J. Diebel, "Representing attitude: Euler angles, unit quaternions, and rotation vectors," *Matrix*, vol. 58, no. 15-16, pp. 1–35, 2006.