Grant Agreement No. 611373

FP7-ICT-2013-10

D2.6 Algorithms for recognition of diver symbols and special gestures for execution of compliant tasks

Due date of deliverable: 30/04/2016
Actual submission date: 03/05/2016

Start date of project: 01 January 2014 Duration: 36 months

Organization name of lead contractor for this deliverable: CNR

Revision: version 1

Dissemination level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Contents

1 Outline of the deliverable

This deliverable deals with the description of the algorithms for reliable recognition of diver hand signals by using remote sensing for execution of compliant tasks. This was primarily done using a stereo camera, with procedures described in chapters 2 (gesture detection) and 3 (gesture classification).
The robustness of the algorithms is validated through the analysis of experimental results carried out during the field trials in Biograd Na Moru (Croatia), October 2015, and Genova (Italy), November 2015, and is described in chapter 4.
Besides using the stereo camera, an approach using a multi-beam sonar was also tested, as described in chapter

2 Gesture detection

In this section, we describe the methods implemented and tested for the recognition of diver gestures underwater. It is important to mention that most of these methods have already been described in Deliverable "D3.2 – Symbolic language interpreter" as both tasks are highly correlated. For this reason, this document will describe briefly the algorithms that were already explained and focus more on the topics that have not been previously addressed.

The overall task of recognizing the divers' gestures is divided in two modules: hand detection and gesture classification. The former module is done through two algorithms: the first detects the hand thresholding the disparity map and the second one trains a Haar Cascade classifier; thus, one uses 3D information and the other 2D data. The output of both algorithms is checked for consistency. Once the hand is detected, 2D features are extracted from the image patch corresponding to the hand and are input to a Random Forest classifier to identify the gesture label.

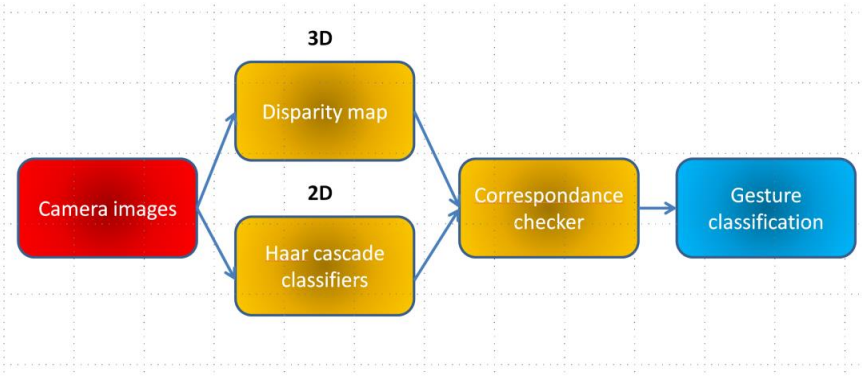

Figure 3.a. General block diagram for gesture classification. The yellow blocks are part of the hand detection task and the blue one to the classification task.

2.1 Hand detection

Current research on hand detection and recognition primarily uses dense point clouds to segment the hand and extract useful features. However, by using stereo cameras for underwater scenarios that offer rather textureless images, only sparse point clouds can be obtained and state of the art methods are not always reliable. An example of this case is shown in Figure 3.1.1, in the left image a good segmentation of the hand features can be achieved (wrist, palm and fingers), but in the right one the boundaries between palm and fingers are not adequate due to the coarse point cloud. For this reason, 3D information is only used for detection and tracking, not for classification.

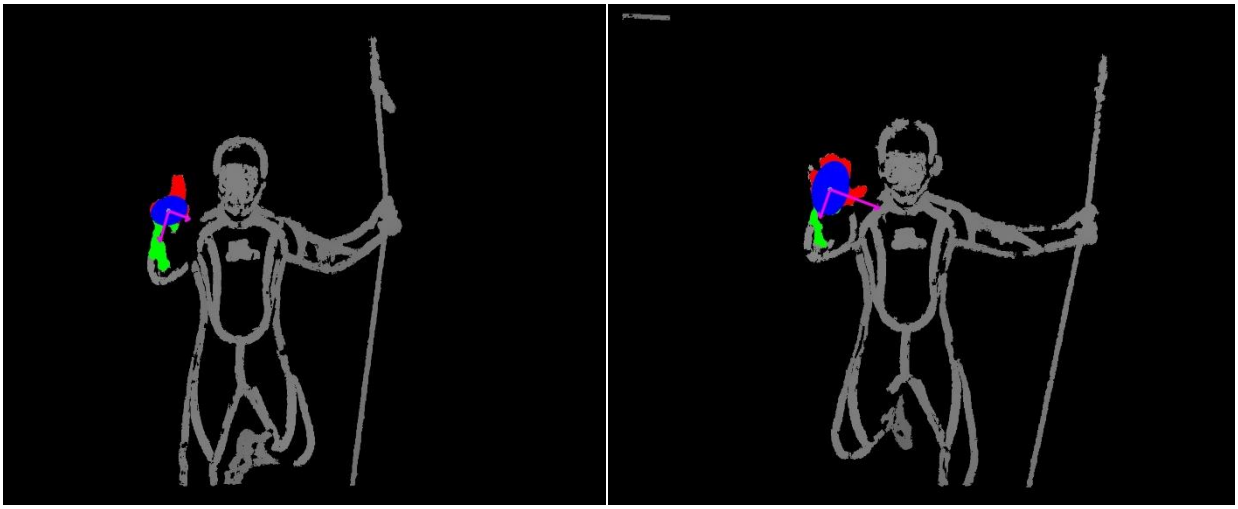

Figure 3.1.1. Examples of hand segmentation using state of the art techniques. Left image shows a good segmentation example between fingers and pal. Right image shows a bad segmentation caused by the imprecision of the obtained point clouds.

Whenever we know we have successfully located the hand, a Kalman Filter is implemented to track it because the hand will not always be the closest "object" to the camera. In the case when there is no certainty that the hand can be accurately tracked, the fact that there are scarce areas in the image with texture in the image is exploited. In the majority of the images, the hands and face exhibit more texture than the rest of the image; thus, the disparity map will have a greater concentration of points in these areas.

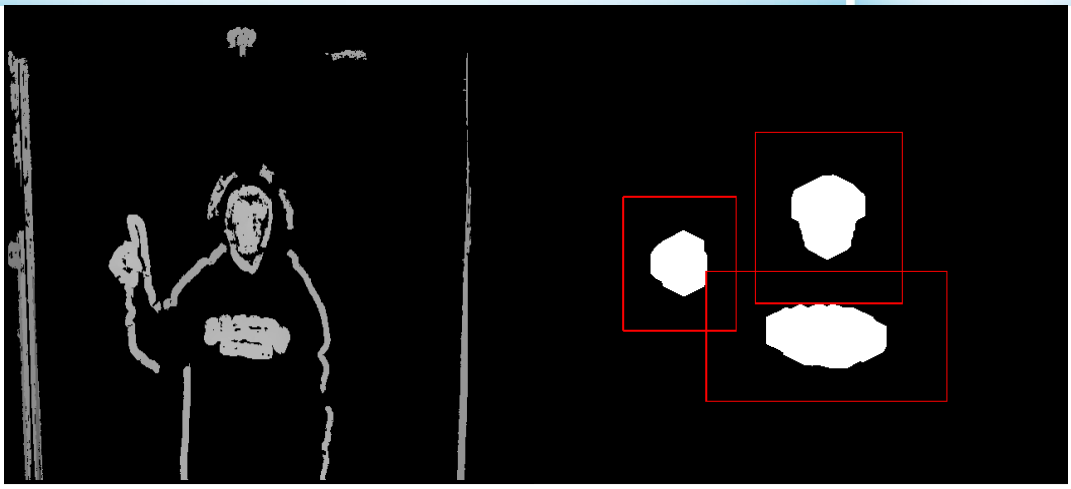

Figure 3.1.2. (Left) Disparity map obtained from stereo imagery. (Right) High density regions detected from the disparity map, these indicate the areas with more texture in the image, in most cases one of them covers the hand.

After applying some image filters to remove noise, it is easy to cluster high density regions. Then, these patches are extracted from the image and pass to the classifier, which can filter out the areas do not have the hand on them (see Fig 3.1.2).

The hand detection methods presented are reliable and fast to compute given that there is not much noise in the image or sudden fast movements. However, underwater we cannot avoid sudden noise peaks due to a source of light, bubbles, etc.; and under bad weather conditions, the camera or the diver can continuously move around. This is more likely to cause inaccurate disparity maps due to errors in the feature matching process.

To make the system more robust a Haar cascade classifier was trained using monocular images from the hands. This classifier can scan the whole image for matches in real time and in different scales, which is useful if the diver is moving back and forth/up and down. The disadvantage of this algorithm is that it outputs a great number of false positives, this is due to the fact that the method normally needs thousands of images to be very precise, and it is hard to build a dataset of this size for underwater applications. Thus, Haar Cascade will almost guarantee to locate the hands within the image at the expense of also detecting false positive regions; but these can be discarded in the next classification step.

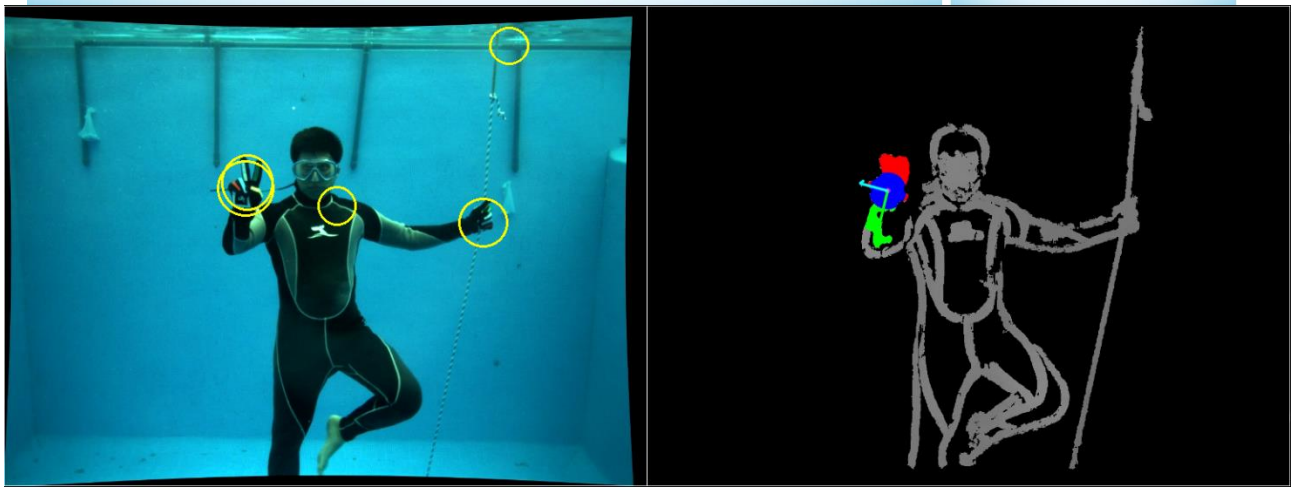

Figure 3.1.3. (Left) Monocular image with the hand candidate patches output by the Haar classifier. (Right) Disparity image with the detected hand segmented

3 Gesture classification

The output of Section 3 are a set of image patches from the original stereo imagery, some of them containing the hand and others don't. These image candidates are encoded into feature vectors and used to trained a Multi-Descriptor Nearest Class Mean Random Forest (MD-NCMF), which can aggregate different descriptors without compressing; thus, losing no information. This is beneficial because different descriptors are invariant to different type of image distortions, if they are aggregated, the classifier will be more robust against changes in objects appearance. The MD-NCMF achieves this by representing query images and classes with the mean (average) of their representative feature vectors; rather than using a Bag of Words approach or using all of the feature vectors found in an image. This makes comparisons and online changes in the classifier easier. Figures 4.1 and 4.2 show examples of a regular Random Forest and the proposed MD-NCMF to represent the main idea of the algorithm, the details of this classifier are explained in the paper published in OCEANS Genova 2015: Visual diver detection using MD-NCMF in the context of underwater Human Robot Interaction.

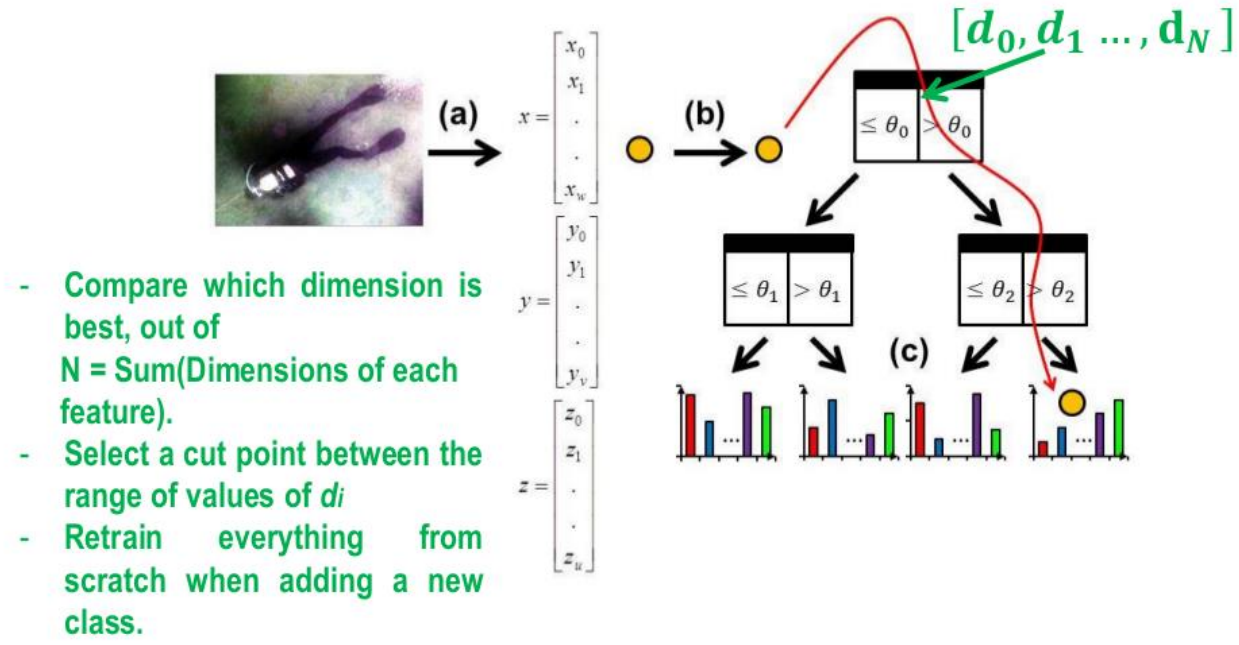

Figure 4.1. Diagram of a regular Random Forest. When different features x,y,z are computed, they are concatenated and the compressed. For classification the best dimension of the resultant vector have to be found.

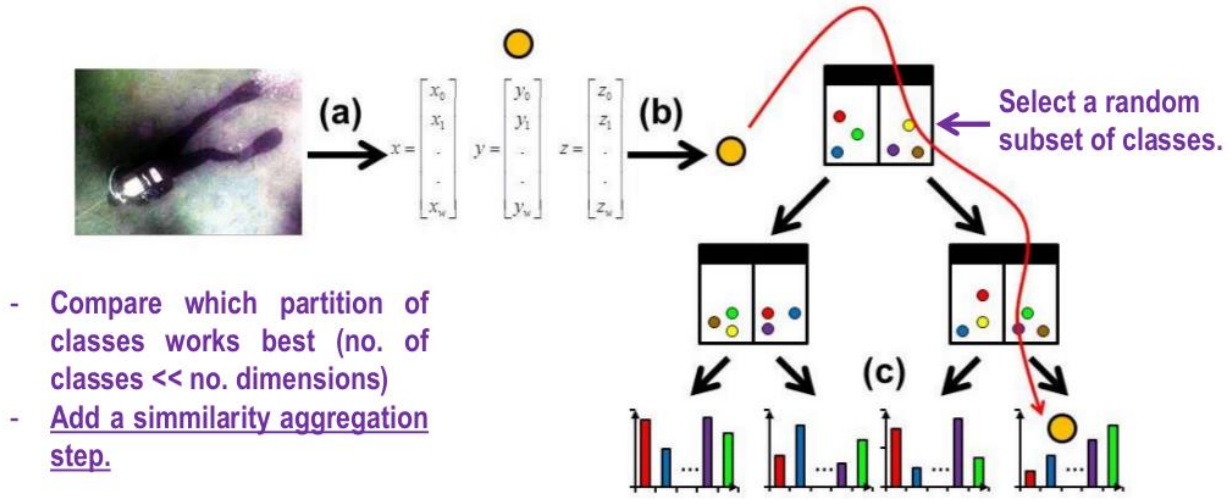

Figure 4.2. Diagram of a Multi-Descriptor Nearest Class Random Forest (MD-NCMF). Each set of features and/or classes can be represented with their mean, which make comparisons in the classifier easier and reduces dimensionality; it is also easier to retrain the classifier when adding features or new labels.

After some experimentation with a part of the built dataset, the feature descriptors Histogram of Oriented Gradients (HOG) and Laplace of Gaussian-SURF (LOG-SURF) output the best results as shown in Table 4.1. Of course, this type of experiments can be done more extensively and looking for combinations of more than 2 types of features to achieve better results.

	LOG-SURF	LOG-DAISY	MSER-SURF	MSER-DAISY	HOG
LOG-DAISY	77.4%				
MSER-SURF	72.3%	83.4%			
MSER-DAISY	78.3%	76.5%	75.1%		
HOG	**94.1%**	90.2%	86.2%	87.2%	
HAR-SIFT	76.2%	74.7%	77.1%	79.8%	83.7%

Table 4.1. Classification rate of of MD-NCMF with different combination of image features (keypoints+descriptors).

Overall, 16 different type of gestures are recognized; which are then the input of the CADDIAN syntax checker explained in deliverable "D3.2 - Symbolic language interpreter", and which is able to recognize complex messages. This classification pipeline has proved to be robust against scale (different distance of the diver to the camera) and continuous movement of the diver to harsh weather conditions. Figure 4.3 shows different detected and classified gestures.

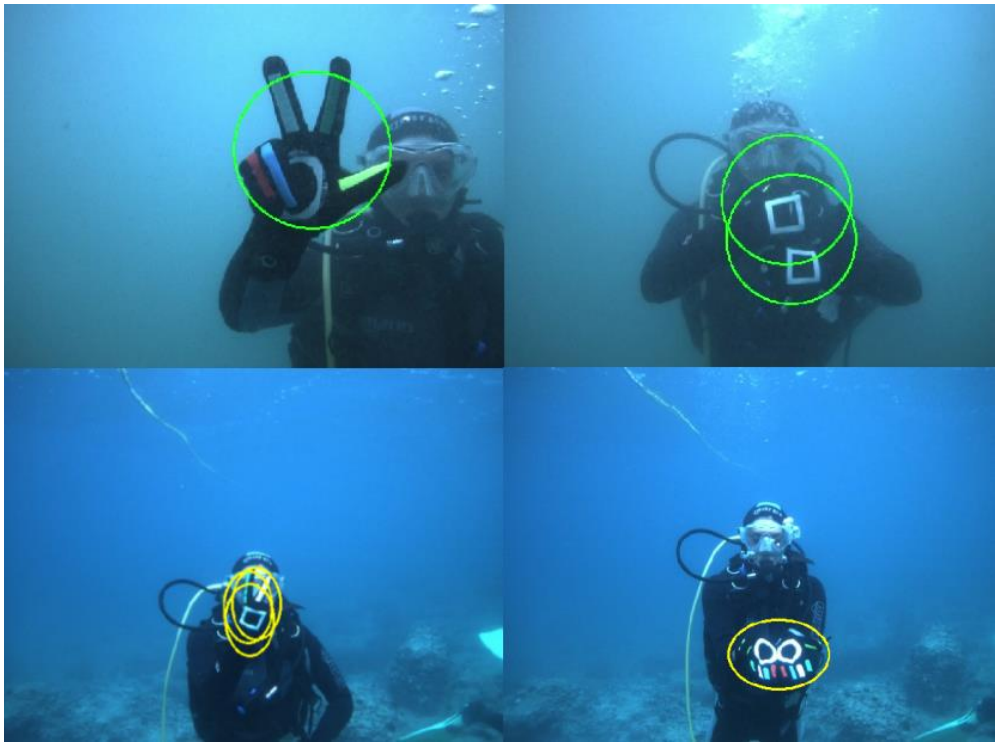

Figure 4.3. Detected gestures: take a photo, carry equipment, start communication, go to the boat. Starting from the top left in clockwise fashion

Finally, in Figure 4.4, a screenshot of the system output is displayed; the system detects each gesture performed by the diver if it is stable through several frames; this means that the same gesture is detected continuously, in this way sporadic false positives are avoided. The system also keeps track of the last recognized gesture as seen in the image. When the AUV recognized the gesture, it signals the diver through lights.

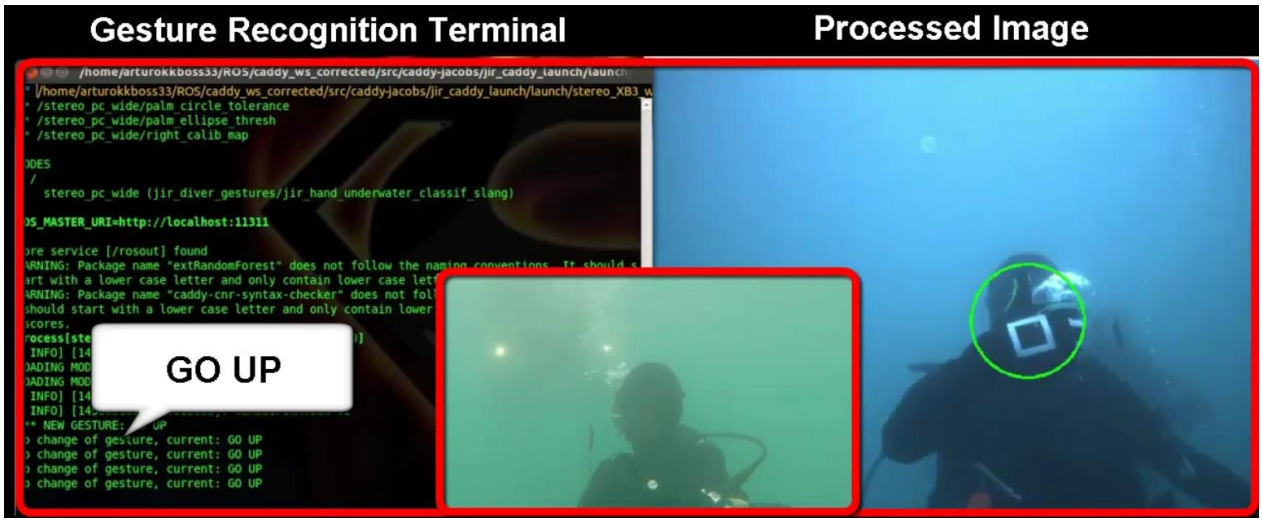

Figure 4.4. Classification system processing the raw stereo images, processing them to detect the hands and classify them. The output is shown in the gesture recognition window

4 Validation trials

Validation trials are described in Section 6 of Deliverable "D3.2 – Symbolic language interpreter".

5 Gesture recognition with multi-beam sonar

The gesture recognition task using a multi-beam sonar is divided into two steps, similarly as in camera gesture recognition. First, the hand detection is performed. After that the gesture classification is done. The gestures tested in this approach were diver showing between one and five stretched fingers, as shown in Figure 5.1.

Figure 5.1. Gestures used in multi-beam sonar gesture recognition.

Hand detection phase was done using a cascade of boosted classifiers based on Haar-like features. The classifier is trained with images of all five gestures.
This algorithm by itself gave good results on detection, but for classification other approaches were used. Convex hull of the hand was calculated, using the anatomy of human hand to determine the number of fingers that the diver was showing. The steps of the algorithm are shown in Figure 5.2 and 5.3. Precision rate has been quite poor on one-finger gesture (around 60%), with 90% precision on other gestures, and sensitivity was around 95% for all gestures.

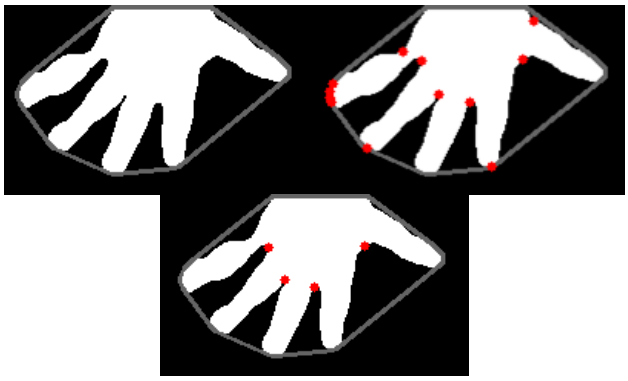

Figure 5.2 Gesture recognition from sonar by convex hull method. Convex hull (top left); all convexity defects (top right); filtered convexity defects (bottom)

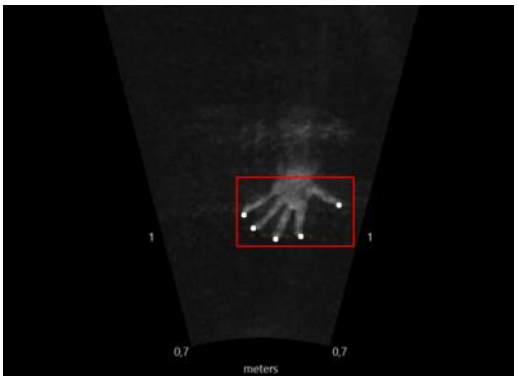

Figure 5.3 Detected fingertips within the sonar image (white dots)

The second approach in gesture classification was done using support vector machines (SVM), a well-known algorithm widely used in classification tasks. The classifier was trained using approximately 1000 images of each gesture, and the results have shown precision of around 90% with sensitivity around 96% on average.

Both approaches have shown good results, but in order to improve the precision they were used simultaneously and their output was combined. By reacting on a gesture only when the two classifiers give matching output, precision (number of false positives) is reduced, at the expense of missing out on some gestures (increasing false negatives). Practically, that means that, in case that the gesture was not recognized, diver has to keep showing it until the classifiers agree and recognize the gesture. Since one of the biggest issues in multi-beam sonar approach is placing the hand in a way that produces image of decent quality, this also means that the diver can re-position his hand until the gesture is recognized, helping the system while avoiding false recognition.

This approach has increased precision rate to over 95% for all the gestures. Sensitivity has slightly dropped, to 90-95%.

6 Conclusions

This report described the algorithms for recognition of diver symbols, which have proved to be robust, allowing the classification pipeline to recognize the 16 different gestures chosen to be executed during the trials. It is important to mention that most of the work has already been described in Deliverable “D3.2 – Symbolic language interpreter” as both tasks are highly correlated. For this reason, this document describes briefly the algorithms already treated in “D3.2 – Symbolic language interpreter” and focuses more on the topics that have not been previously addressed. To get an overall view of the work done both documents need to be read.

To summarize we can say that classification presents good accuracy rates (average of 89.8%), being the cardinality of gestures set 16: we expect more challenges as the number of gestures increases. Classification performed well also in different environmental conditions (good weather/bad weather) and with different divers (size of hands, personal preference to perform gestures, etc.).

Additionally, recognition with multi-beam sonar imagery was also explored. Given a set of good quality images (with the hand places at appropriate angle towards the sonar), it gave surprisingly good results, with around 95% precision rate.