Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Automatic analysis of deep-water remotely operated vehicle footage for estimation of Norway lobster abundance*

Ching Soon TAN[1], Phooi Yee LAU[†‡1], Paulo L. CORREIA[2], Aida CAMPOS[3,4]

[1]*Centre for Computing and Intelligent Systems, Universiti Tunku Abdul Rahman, Kampar 31900, Malaysia*

[2]*Instituto de Telecomunicações, Instituto Superior Técnico, Av. Rovisco Pais, 1, Lisbon 1049-001, Portugal*

[3]*Instituto Português do Mar e da Atmosfera (IPMA), Divisão de Modelação e Gestão de Recursos da Pesca, Lisbon 1749-077, Portugal*

[4]*Centro de Ciências do Mar (CCMAR) - Campus de Gambelas, Faro 8005-139, Portugal*

[†]E-mail: laupy@utar.edu.my

Received Nov. 9, 2017; Revision accepted Aug. 22, 2018; Crosschecked Aug. 23, 2018

Abstract: Underwater imaging is being used increasingly by marine biologists as a means to assess the abundance of marine resources and their biodiversity. Previously, we developed the first automatic approach for estimating the abundance of Norway lobsters and counting their burrows in video sequences captured using a monochrome camera mounted on trawling gear. In this paper, an alternative framework is proposed and tested using deep-water video sequences acquired via a remotely operated vehicle. The proposed framework consists of four modules: (1) pre-processing, (2) object detection and classification, (3) object-tracking, and (4) quantification. Encouraging results were obtained from available test videos for the automatic video-based abundance estimation in comparison with manual counts by human experts (ground truth). For the available test set, the proposed system achieved 100% precision and recall for lobster counting, and around 83% precision and recall for burrow detection.

Key words: Object detection; Object tracking; Feature extraction; Remotely operated vehicle (ROV)
https://doi.org/10.1631/FITEE.1700720 **CLC number:** TP391

1 Introduction

As the costs associated with the use of underwater equipment drop, underwater video has been used increasingly in the context of management of commercial fish stocks. An example is the Norway lobster (*Nephrops norvegicus*), a burrowing crustacean species living in muddy sediments at depths ranging from 15 m to more than 800 m in the whole Northeastern Atlantic Ocean and Mediterranean Sea (Howard, 1989). Since Norway lobsters constitute a valuable commercial catch in the European fish market (Howard, 1989), regular, long-term, and large-scale monitoring surveys have been established for this species to assess and manage its populations. Lobsters are highly territorial species, usually spending most of their lifespan within or in the vicinity of their burrows. Emergence from burrows reflects a daily routine, affected by seasonal variations, for instance, related to reproduction cycles, directly affecting the density and sex-ratio estimates observed in monitoring surveys. Owing to the recent availability of image surveys, burrow density has become an important clue and is being used regularly as an abundance index in lobster stock assessment (Sardà and Aguzzi, 2012). Currently, underwater imaging is well-suited for the estimation of lobster stock

‡ Corresponding author

* Project supported by the UTAR Research Fund from the Universiti Tunku Abdul Rahman, Malaysia (No. IPSR/RMC/UTARRF/2013-C2/L03)

ORCID: Phooi Yee LAU, http://orcid.org/0000-0002-6329-4558

abundance through integration with burrow densities, due to its ability to perceive the burrows' visual details from the video footage (Morello et al., 2007). This paper takes this information into consideration for the development of an automatic video-based application. In earlier experimental work carried out in *Nephrops* fishing grounds off the Portuguese southern coast, the experimental setup consisted of a Kongsberg Maritime OE1324 monochrome low-light SIT camera mounted on the upper center of a trawl gear headline, angling down in the tow direction. Tows were carried out at about 3.0 knots, producing images captured from an observation angle of approximately 45° (Fonseca et al., 2008). Datasets referred to in Fonseca et al. (2008) are usually obtained by using video cameras mounted on either towed gears (Correia et al., 2007; Lau et al., 2008, 2012), or, most often, sledges towed behind a research vessel (Sooknanan et al., 2013). In practice, these setups are limited to towing in straight lines, with the video camera capturing video from a fairly fixed view point. In this paper, the dataset to be analyzed was acquired off the Portuguese south Atlantic coast using a remotely operated vehicle (ROV), belonging to the non-governmental organization OCEANA. The ROV was controlled by a technician on board of a research vessel, and could move freely in any desired direction while recording. In this new setup, the area observed by the ROV-mounted camera is unstructured and highly dynamic, unlike the experimental setups previously discussed (Fonseca et al., 2008). The ROV can capture images from various viewpoints, as the camera can be oriented in different angles and directions, conducting video analysis for lobster and burrow detections and tracking more challengingly. Thus, estimation of the monitored area becomes more difficult.

Traditional underwater video processing is carried out manually, with trained technicians, usually marine biologists, identifying the observed marine species or biogenic structures and counting them. During calibration tests, biologists evaluate video clips with good, medium, and poor visibility to check how illumination conditions affect the counting by different observers. The manual approach suffers from several shortcomings, mostly related to human operator experience and capacity for concentration, often resulting in counting bias, especially for low-visibility videos and/or densely populated lobster fishing grounds. Additionally, when a large set of relevant video clips needs to be processed, the manual approach becomes bottlenecked. To minimize human-intensive workload, a number of automated digital image processing approaches have been presented to assist in the study on Norway lobster abundance (Correia et al., 2007; Lau et al., 2008, 2012). Automatic analysis requires understanding which key characteristics of the lobsters and burrows should be looked for.

Correia et al. (2007) worked with monochrome images and noticed that lobsters correspond to brighter image areas, while burrows appear as crescent-shaped areas with shadows in the middle. They proposed a Norway lobster detection and counting solution based on the analysis of three visual feature maps: intensity, edge, and motion. A year later, a software prototype, named 'IT-IPIMAR *N. norvegicus* I^2N^2', was developed by the same team to provide a more comprehensive analysis of lobster and burrow density estimation (Lau et al., 2008). Lau et al. (2012) proposed an improved segmentation of non-uniformly illuminated regions using a local thresholding technique. Sooknanan et al. (2013) adapted a mosaic indexing representation to solve the burrow detection problem, the mosaic being the composite image generated across the frames in a video. Their mosaic-based indexing approach indirectly solves the geometric distortion and limited field view problems. However, their work requires a consistent acquisition method; i.e., the images should be registered correctly without any distortion or information loss before starting the detection. In our recent work (Tan et al., 2014, 2015), the initially released subset of the dataset acquired from ROV was considered. In Tan et al. (2014), a preliminary solution was proposed by analyzing each image individually. Later, a visual tracking scheme was added by Tan et al. (2015) to track the detected lobsters and burrows in a dynamic motion environment, and finally to estimate the numbers of lobsters and burrows in the whole footage, preventing over-counting for objects detected in consecutive frames. The integration of Tan et al. (2014, 2015) allows an automatic processing of ROV video sequences. However, when further video sequences from the ROV dataset become available, it becomes apparent that the detection algorithm presented by Tan et al. (2014) cannot effectively discriminate the targets in the

presence of other, previously unobserved, seabed structures. The features considered for detection and segmentation are not robust enough, with some of the regions of interest being missed due to poor region segmentation.

This paper extends the work started in Tan et al. (2014, 2015) by: (1) proposing improved lobster and burrow detection algorithms, which include additional features and consider the use of a support vector machine (SVM) classifier, (2) providing a complete framework for this study, adding tracking along with time, and (3) performing a more exhaustive evaluation, considering the additional video sequences that were made available.

2 Experimental setups

In this study, a set of video sequences of ocean seabed, captured by a camera mounted on a ROV, are used. The 'Sea-eye' Falcon ROV, belonging to the non-governmental organization OCEANA, was used on board of a research vessel off the Portuguese south Atlantic coast, during a survey dedicated to estimating the effects of continued trawling on Norway lobster deep-water fishing grounds, within the scope of project IMPACT. The ROV was submerged to a depth of about 500 m, under artificial illumination. An RGB video camera, operating at a frame rate of 12 frames/s with a spatial resolution of 640×360 pixels and storing at 24 bits/pixel, was used. An image of the ROV and the sample images acquired from Norway lobster fishing grounds are included in Figs. 1 and 2, respectively. During video footage recording, the ROV moved smoothly over the seabed, acquiring video sequences of sufficient spatial resolution for the desired analysis, without introducing significant motion blur problems. However, some seafloor video sequences presented several challenges, including: (1) 'marine snow', due to the mud clouds moving in front of the camera, thus severely limiting visibility; (2) artifacts appearing at the image borders due to the presence of the camera container used to withstand the high pressure.

3 Proposed system

In this study, we propose an improved automatic video analysis framework, whose architecture is presented in Fig. 3, for the detection and counting of

Fig. 1 Remotely operated vehicle used for image acquisition: Seaeye Falcon

(a) (b)

Fig. 2 Sample images: (a) sample image distorted by the presence of 'marine snow'; (b) sample image containing a lobster (marked with an ellipse) and three large burrow entrances (marked with rectangular boxes)

visible lobsters and burrows. It includes four main modules: (1) pre-processing, (2) object detection and classification, (3) object tracking, and (4) quantification. The goals of the pre-processing module include elimination of irrelevant image details and compensation for the effect of non-uniform illumination. The object detection module searches for the regions of interest in the image, segments them, extracts visual features, and then classifies the regions as lobsters, burrows, or others. The next module tracks objects along consecutive video frames, and is able to recognize when ROV returns to an area visited previously. Finally, the quantification module delivers a video-based estimate of the abundance of lobsters and burrows.

3.1 Pre-processing

In an oceanic environment, the deeper the water, the more light is absorbed (Johnsen and Sosik, 2004). The present work considers high depths, where natural light is too weak, requiring ROV to carry artificial illumination. From the acquired video sequences, it is possible to observe that the color information present in the images is attenuated greatly, making the image intensity component much more representative of the observed content; therefore, video frames are converted to grayscale for

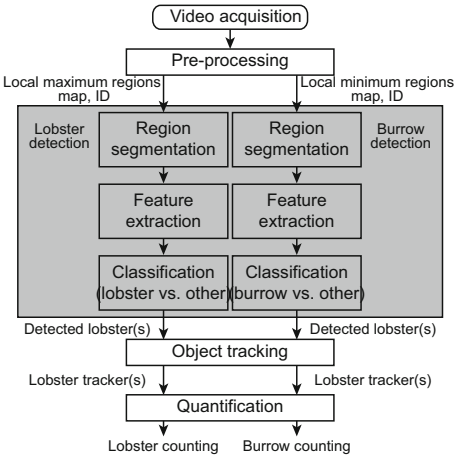


Fig. 3 Architecture of the proposed system

further pre-processing. To discard irrelevant details appearing in images due to instrumentation setup, notably the black borders near image boundaries caused by the camera container to withstand high pressure (Fig. 2), a cropping operation is conducted. The illumination sources installed in ROV propagate a light spot in the direction of the acquired images; however, they have a limited range. As illustrated in Fig. 2, the illumination is brighter in the area near the light source and gradually decreases with distance. This causes a non-uniform illumination of the imaged area. The goal of this module is to compensate for the non-uniform background illumination and to create two foreground maps, using the difference in Gaussian distribution, according to

$$\mathbf{IB} = \boldsymbol{G}_1 \cdot \boldsymbol{I} - \boldsymbol{G}_2 \cdot \boldsymbol{I}, \tag{1}$$

$$\mathbf{ID} = \boldsymbol{G}_2 \cdot \boldsymbol{I} - \boldsymbol{G}_1 \cdot \boldsymbol{I}, \tag{2}$$

where $\boldsymbol{I}$ is the input grayscale image, $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ are two-dimensional Gaussian kernels of size 55×55 and 5×5, respectively, and **IB** and **ID** are the output images containing the local maximum and minimum regions (Fig. 4), respectively. The idea behind the selection of a large value for kernel size for $\boldsymbol{G}_1(x, y)$ is to substantially extrapolate the local intensity difference.

3.2 Lobster and burrow detection

The object detection module is composed of three main steps: (1) region segmentation, (2) feature extraction, and (3) classification. Each of these steps is detailed in the following.

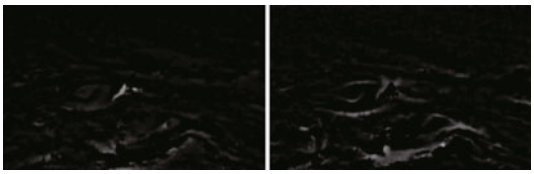


Fig. 4 Output images: (a) the lobster contains brighter pixels, in contrast to its surrounding regions; (b) the burrow contains darker pixels, in contrast to its surrounding regions

3.2.1 Region segmentation

This section discusses the identification of regions corresponding to lobster or burrow candidates. From an exhaustive analysis of imaged lobster examples, it was observed that lobster regions tend to present a high intensity contrast to their surrounding area, while burrow regions usually present a low intensity contrast. Therefore, image **IB** was used as an input for candidate lobster region segmentation, while for burrows, image **ID** was used. Candidate lobster (**CL**) region segmentation can be done by a global thresholding technique, which was applied to image **IB** according to

$$\mathbf{CL}(x, y) = \begin{cases} 255, & \mathbf{IB}(x, y) \geq \frac{m + I_{\max}}{2}, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where m is image mean intensity and $I_{\max}$ is the maximum intensity of image **IB**.

For candidate burrow (**CB**) region segmentation, a local thresholding operation is preferred since burrow regions exhibit a range of different graylevels. For this purpose, integral images were considered, and the technique proposed by Sauvola and Pietikäinen (2000) was applied to image **ID**. Sauvola and Pietikäinen (2000)'s binarization method selects a local threshold T_2 for each pixel, taking a local window of size $W \times W$ as a context:

$$T_2(x, y) = m(x, y) \left[1 + k \left(\frac{\delta(x, y)}{R} - 1 \right) \right], \tag{4}$$

$$\mathbf{CB}(x, y) = \begin{cases} 255, & \mathbf{ID}(x, y) \geq \max(T_2, T_3), \\ 0, & \text{otherwise}. \end{cases} \tag{5}$$

In the literature, the value of 128 is suggested for R with 8-bit grayscale images, and for k, a value in range [0.10, 0.50] is suggested for document segmentation (Sauvola and Pietikäinen, 2000; Badekas and Papamarkos, 2005; Shafait et al., 2008). Small values of W (e.g., 3, 5) cannot preserve too many

secondary details, while larger values (e.g., 11, 13) are prone to over-fitting. In this study, after some initial experiments, it was concluded that Sauvola and Pietikäinen (2000)'s thresholding technique can be applied with $k = 0.35$, $R = 128$, and $W = 9$, to achieve the desired segmentation results. To reduce complexity in the local threshold computation T_2, an integral image is used to calculate the sum of all pixel intensities fast within a selected window, and to derive the corresponding mean $m(x, y)$ and standard deviation $\delta(x, y)$. Additionally, to avoid the effect of noise in image **ID**, a minimum value, T_3, is imposed on the local threshold with its value experimentally set to five.

For the image, a morphologic close operation is then applied as a post-processing step to both **CL** and **CB**, to merge any small region fragment and to fill in holes in the candidate region. Last, a connected component analysis is performed to merge adjacent pixels and to analyze the resulting contours as appropriate (Suzuki and Be, 1985). As a result, the sets of lobster and burrow candidate regions are identified (Fig. 5).

Fig. 5 Results of image region segmentation: (a) IB; (b) ID

3.2.2 Feature extraction

For each lobster or burrow candidate region, a set of features is extracted to describe its visual structure: F1 is the curvature, F2 is the local intensity contrast, F3 is the aspect ratio, F4 is the diameter, F5 is the region area, and F6 is the orientation. Features F2 and F6 are considered for the first time in this context. Each feature is detailed in the following.

F1 (curvature): When ROV moves, it provides dynamic observation of the scene. For instance, the visual appearance of a burrow entrance changes considerably according to the viewing angle and distance from the camera. From afar, the burrow entrance can often be characterized by its downward oriented crescent-like shape, permitting a partial view of the burrow tunnel. Curvature feature F1 proposed herein describes whether the burrow opening is downward oriented. This feature replaces the one used for the same purpose as in our previous work (Tan et al., 2014), as it is proved to have a superior performance. Computation of F1 (Fig. 6) relies on knowing, for each candidate region, its contour as well as the smallest rectangular bounding box. In case that the rectangular bounding box presents a slant angle α, being greater than 0, then the candidate region is rotated to make it horizontal. Then, the following second-order polynomial is fit to the region's contour pixels $C(x_i, y_i)$ $(i \in \{1, 2, ..., N\})$:

$$y = ax^2 + bx + c. \tag{6}$$

Curve fitting can be solved by the least-squares method, minimizing ε:

$$\varepsilon = \sum_{i=1}^{N} (y_i - ax_i^2 - bx_i - c)^2. \tag{7}$$

If quadratic coefficient a is positive, it means that the parabola opens upwards (Denise, 2007); otherwise, it opens downwards. Since burrows typically present a down-oriented curvature, the curvature feature can be obtained from coefficient a:

$$\mathrm{F1} = a. \tag{8}$$

F2 (local intensity contrast): Both lobsters and burrows typically present with a significant intensity contrast compared with their surrounding area. Feature F2 tries to capture this visual characteristic, corresponding to a significant local intensity contrast between the candidate region and the neighboring pixels at a certain distance. To compute feature F2, a median filter $\boldsymbol{H}$ with kernel size WH is applied according to

$$\mathrm{IM} = |\boldsymbol{I} - (\boldsymbol{H} \cdot \boldsymbol{I})|, \tag{9}$$

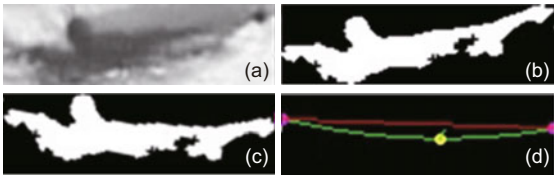


Fig. 6 Feature F1 extraction: (a) grayscale image of a candidate region; (b) segmentation of the candidate region; (c) rotation of the candidate region to make it parallel to the image horizontal axis; (d) results of curve fitting using the least-squares method

producing IM as the output. The value of WH will be determined in Section 3.2.3 and the local intensity contrast (F2) can be measured by

$$\mathrm{F2} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{IM}(x_i, y_i),\ \mathrm{IM}(x_i, y_i) \in C(x_i, y_i), \tag{10}$$

where $\mathrm{IM}(x, y)$ is the absolute value of the difference between the candidate region image and the median filter output, and $C(x, y)$ is the set of contour pixels of the candidate region under analysis.

F3 (aspect ratio): This feature is used to describe the relationship between the width and height of candidate lobster and burrow regions. For instance, it is known that burrows usually have an elongated appearance, with a greater length horizontally than vertically. For lobsters, this relationship typically yields even higher values of the aspect ratio. Therefore, feature extraction and subsequent classification are carried out separately for lobsters and burrows. This feature is used to select preliminarily between the target (lobster or burrow) and other structures observable in the input image, and not to discriminate between lobsters and burrows. It has been observed that typical aspect ratio values are above 1.5 for lobsters and above 1.0 for burrows. For burrows, the region is first rotated to make it parallel to the image horizontal axis; thus, F3 is computed as the ratio between the longest lengths of the horizontal (l_{H}) and vertical (l_{V}) axes of the object:

$$\mathrm{F3} = \frac{l_{\mathrm{H}}}{I_{\mathrm{V}}}. \tag{11}$$

Since lobsters are often found in different locations and with different orientations, due to their movements, the F3 feature for lobsters is computed according to

$$\mathrm{F3} = \frac{I_{\mathrm{major}}}{I_{\mathrm{minor}}}, \tag{12}$$

where $l_{\mathrm{major}} = \max(l_{\mathrm{H}}, l_{\mathrm{V}})$ and $l_{\mathrm{minor}} = \min(l_{\mathrm{H}}, l_{\mathrm{V}})$. The extraction of F3 is shown in Fig. 7.

F4 (diameter): This feature is important for burrow morphology study. Since burrows are recognized mostly by their dark entrances and large diameters (Fonseca et al., 2008), F4 is defined as the longest diagonal of the candidate region. Additionally, this feature will help in distinguishing burrows from other muddy sediment structures frequently found on the seabed, including marks from fishing

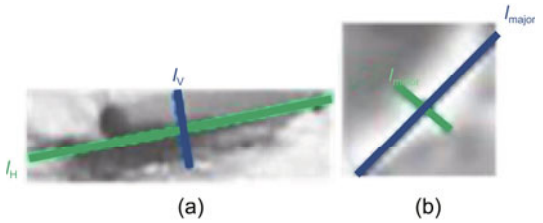


Fig. 7 Feature F3 extraction: computation of feature F3 for a burrow (a) and a lobster (b) candidate region (References to color refer to the online version of this figure)

trawls which present a long straight and narrow appearance. F4 is defined as

$$\mathrm{F4} = \frac{I_{\mathrm{H}}}{W_I}, \tag{13}$$

where W_I is the image width. To avoid false alarms, F4 should take values in range [0.05, 0.50].

F5 (region area): This feature is used to eliminate false alarms for lobster candidate regions. The segmentation step often produces a set of false positives corresponding to noisy image regions, which are typically very small. To exclude such regions from further analysis, each candidate region's area is analyzed, with the lobster regions' area, A_{lobster}, being expected to be between 0.02 and 0.30 of the captured image area A_I:

$$\mathrm{F5} = \frac{A_{\mathrm{lobster}}}{A_I}. \tag{14}$$

F6 (orientation): A candidate region containing a single lobster typically provides a strong response to orientation detection filters. This feature is computed by applying a set of Gabor filters $G_{u,v}(x, y)$ (Struc et al., 2008) with different orientations u and scales v, as defined in the following equation, which are convolved with image **IB**:

$$G_{u,v}(x, y) = \frac{f_u^2}{\pi\gamma\eta}\exp\left[-\left(\frac{f^2}{\gamma^2}x'^2 + \frac{f^2}{\gamma^2}y'^2\right)\right] \cdot \exp\left(\mathrm{j}2\pi f_u x'\right), \tag{15}$$

where $x' = x\cos\theta + y\sin\theta$, $y' = -x\sin\theta + y\cos\theta$, and $f_u = \dfrac{f_{\max}}{2^{u/2}}$. Parameters γ and η determine the ratio between the center frequency and the size of the Gaussian envelope, set to $\gamma = \eta = \sqrt{2}$. Parameter $f_{\max}$ denotes the maximum filter frequency set to $\pi/2$. In practice, it is observed that lobster regions present strongly oriented gradients, and filtering at orientations $v = (0, \pi/4, \pi/2, 3\pi/4)$ provides a strong response as lobsters usually appear

at a slant angle, even when moving parallel to the seabed surface. Therefore, Gabor filters with three scales and four orientations, i.e., $u = 0, 1, 2$ and $v = (0, \pi/4, \pi/2, 3\pi/4)$, yielding a total of 12 feature maps, are considered. These feature maps are combined as

$$s = \sum_{3}^{u} \sum_{4}^{v} G_{u,v}. \quad (16)$$

Finally, orientation feature F6 is computed as

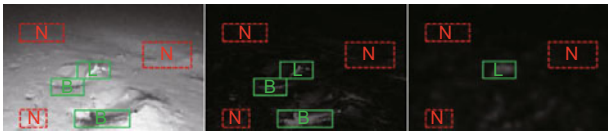
$$\mathrm{F6} = \frac{1}{N} \sum_{N}^{i=1} S(x_i, y_i), \ S(x_i, y_i) \in C(x_i, y_i). \quad (17)$$

3.2.3 Region classification

After a set of features describing each candidate region has been extracted, these features can now be used to classify the candidate regions. Lobster and burrow candidate regions are classified separately. The former are labeled as 'lobster' or 'other', while the latter as 'burrow' or 'other'. Initially, each region is evaluated based on a feature subset {F3, F5} for lobster candidate regions, and {F1, F3, F4} for burrow candidate regions. Only the regions satisfying the conditions imposed on the selected features are considered for further evaluation; otherwise, they are classified as 'other'. Then an SVM is employed to further classify the remaining regions, based on features F2 and F6 for lobster candidate regions, and on feature F2 for burrow candidate regions. The goal of SVM is to optimally separate the different classes of data using a hyperplane that maximizes the margin of separation between two different classes (Ben-Hur and Weston, 2010).

1. Classification performance

Several experiments were conducted to optimize the lobster and burrow classification performance by selecting the best parameter values using a feature set selection strategy. The dataset used for the lobster and burrow classification was obtained from a set of available video sequences acquired along the target trajectory, and labeled manually as illustrated in Fig. 8. A total of 2128 and 5949 samples, for lobsters and burrows, respectively, were identified. From these samples, subsets of 240 and 600 samples, for lobsters and burrows, respectively, were selected for training, whereby the sizes of the positive training samples are consistent with those of the negative ones (Akbani et al., 2004). The remaining samples were used as testing samples. The numbers of samples, positive and negative, for the lobster and burrow training and testing processes, are reported in Table 1.



(a) (b) (c)

Fig. 8 Training sample selection: one of the frames (a) used to select training samples, and samples selected in one image used for training of F2 (b) and F6 (c)

Table 1 Numbers of samples in the datasets for training and testing

Sample target	Number of samples			
	Training dataset		Testing dataset	
	Positive	Negative	Positive	Negative
Lobster	120	120	53	1835
Burrow	300	300	716	4633

To determine the optimal values of parameter WH for feature F2 and parameter C for the SVM classifier, a 10-fold cross-validation technique was applied. The original sample set consists of a total of 600 samples, which were partitioned randomly into 10 equally sized subsets. Of the 10 subsets, a single subset was retained for the testing model, and the remaining subsets were used as the training data. The cross-validation process consisted of repeating the procedure considering each of the 10 subsets for testing. Average results for the 10 experiments are reported in Figs. 9 and 10, leading to the selection of WH = 55 and $C = 0.1$ for both SVM classifiers, as these values provide the highest accuracy during the validation.

2. Classification strategy

To determine the best classification strategy between burrow and non-burrow regions, the performances of the feature selection and combination were examined. Results of the tests conducted using the proposed system for counting lobsters and burrows, and their comparison with the ground truth obtained by manual annotation, are reported. The evaluation metrics for these experiments were Precision, Recall, and F-score, according to

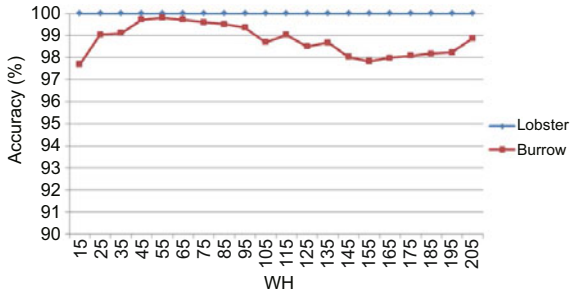


Fig. 9 Accuracy as a function of parameter WH used for feature F2

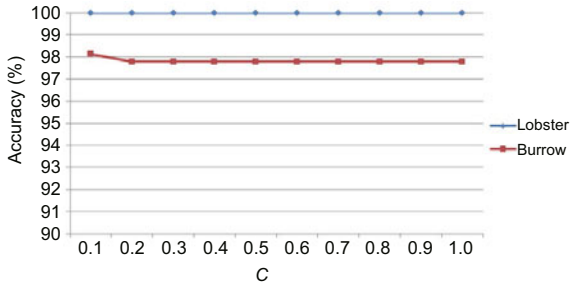


Fig. 10 Accuracy as a function of parameter *C* used by the support vector machine

$$\text{Recall} = \frac{N_{\text{true\_positive}}}{N_{\text{positive}}} \times 100\%, \tag{18}$$

$$\text{Precision} = \frac{N_{\text{true\_positive}}}{N_{\text{true\_positive}} + N_{\text{false\_positive}}} \times 100\%, \tag{19}$$

$$\text{F-score} = 2 \cdot \frac{\text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \times 100\%, \tag{20}$$

where N_{positive} denotes the actual total count of manually labeled objects in the video, and $N_{\text{true\_positive}}$ and $N_{\text{false\_positive}}$ are the total counts of true and false positives obtained by the proposed system, respectively. As shown in Table 2, from individual feature analysis, individual feature F2 obtained an F-score of 48.48% with smaller classification errors than the other individual features, while feature F4 was the worst with an F-score of 11.59%. From the analysis of different feature combinations using SVM, one of the interesting points that should be noted is that most of the combined feature sets performed better than using an individual feature, especially for feature subset {F1, F2, F3} or {F1, F2, F3, F4}, which both obtained an F-score of 48.67%. As such, the proposed burrow classification strategy (Fig. 11) is even superior to the SVM feature subset, since it clearly showed that an F-score of 60.47% is achieved with the best recall of 51.06%.

Table 2 Lobster and burrow classification performance with different feature combinations

Target	Feature	Precision (%)	Recall (%)	F-score (%)
Burrow	F1	100.00	13.97	24.51
	F2	84.50	33.99	48.48
	F3	85.06	26.07	39.90
	F4	11.59	11.42	11.50
	{F1, F2}	84.50	34.01	48.50
	{F1, F3}	85.75	26.65	40.66
	{F1, F4}	100.00	14.06	25.48
	{F2, F3}	84.63	34.14	48.66
	{F2, F4}	84.50	33.99	48.47
	{F3, F4}	84.63	29.66	43.93
	{F1, F2, F3}	84.64	34.16	48.67
	{F1, F2, F4}	84.50	34.01	48.50
	{F1, F3, F4}	85.06	30.02	44.37
	{F2, F3, F4}	84.64	34.14	48.66
	{F1, F2, F3, F4}	84.64	34.16	48.67
	Proposed strategy	74.16	51.06	60.47
Lobster	F2	75.47	26.31	39.02
	F3	32.08	1.35	2.59
	F5	41.51	10.63	16.92
	F6	100.00	20.00	33.33
	{F2, F3}	77.36	26.28	39.23
	{F2, F5}	75.47	26.32	39.02
	{F2, F6}	96.23	27.27	42.50
	{F3, F5}	32.08	1.35	2.59
	{F3, F6}	100.00	19.92	33.23
	{F5, F6}	100.00	20.00	33.33
	{F2, F3, F5}	77.36	26.28	39.23
	{F2, F3, F6}	96.23	26.98	42.15
	{F2, F5, F6}	96.23	27.28	42.50
	{F3, F5, F6}	100.00	19.92	33.29
	{F2, F3, F5, F6}	96.23	26.98	42.15
	Proposed strategy	96.23	34.23	50.50

For the classification between lobster and non-lobster regions, the individual features F2 and F6 obtained F-scores of 39.02% and 33.33%, respectively, showing a better performance compared with F3 and F5. Feature subsets {F2, F6} and {F2, F5, F6} obtained the same F-score of 42.50%, which is the best among all the feature combinations for lobster classification. As such, the proposed strategy (Fig. 11) is superior to the strategies using the feature subsets above, since it clearly achieved an F-score of 50.50%, with the best recall at 34.23%. Again, it should be noted that all the samples were acquired along the target trajectory. When either a burrow or lobster spatially leaves the scene or the distance between the targets to the camera is long, we still consider this region as the target and mark its label with the relevant class. This is the reason why in these cases the

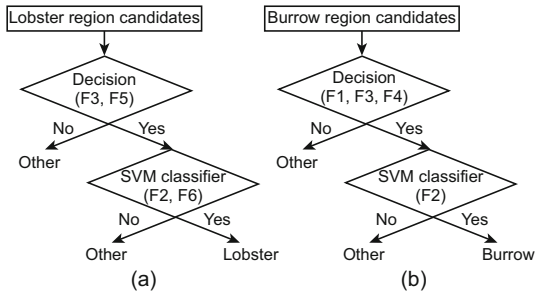


Fig. 11 Proposed classification schemes: (a) lobster; (b) burrow

classifications for lobsters and burrows did not perform well, because the analysis was performed based on a single image instead of consecutive frames. This will be solved in the next step by integration with the object tracking module, the objective of which is to assign into the same tracker those objects (lobsters or burrows) of interest that are the same as appearing in consecutive frames.

3.3 Object tracking

The aim of the object tracking module is to assign the same unique label to an object that appears in consecutive video frames. This will then be explored by the final module of the proposed system for computing a video-based estimation of lobster and burrow abundance. The visual-tracking scheme proposed by Tan et al. (2015) was applied to solve the lobster and burrow tracking problem, relying on the use of an image mosaicing technique. This technique estimates the motion between the consecutive images, relying on the identification of the same textured patch or visual clue in consecutive images, after eventually applying some geometric transformation to compensate for the differences in camera observation angles due to camera motion and eventually to the remote operator actions, such as zoom.

Since burrows are static elements, burrow tracking is solved using the above strategy, taking into account the camera motion, which is a dynamic observation. In this case, an affine transformation model is used to describe the observed motion, which is estimated using a set of local spatial flow fields over pairs of consecutive images and the iterative Lucas-Kanade optical flow method (Bouguet, 2003). Once the affine transformation parameters are available, a burrow position in the next frame can be predicted. Note that a tracked burrow location will be available when analyzing the following frames. Lobsters, on the other hand, may either move or idle with the translation under the observation by a non-stationary camera, e.g., carried out by a ROV with translation at time t. This means that a lobster's location is not only defined with respect to the motion of the camera, but also affected by its own motion activity. In this case, the lobster may move additionally from the position where it was observed originally. As such, the particle filter based method (Yang et al., 2005) is used to predict the state of a tracked lobster candidate in consecutive frames. Subsequently, data association is performed for the lobster and burrow regions separately. The tracking cycle thus involves the steps of (1) state prediction and (2) data association. Given a set of lobsters L_t and burrows B_t detected at frame t, and the estimated lobster and burrow trajectories $L\_\mathrm{Tr}_{1:t-1}$ and $B\_\mathrm{Tr}_{1:t-1}$, respectively, object correspondences between frames t and $t-1$ can be established using the information available from L_t and $L\_\mathrm{Tr}_{1:t}$, B_t and $B\_\mathrm{Tr}_{1:t}$, respectively, considering the centroids of the lobster and burrow regions. This data association can be solved by the Hungarian algorithm (Kuhn, 1955). If a region detected at frame t and a region tracked from frame $t-1$ are overlapping by more than 30% of the maximum area, they receive the same identification label.

3.4 Quantification strategy

A video-based estimation of lobster and burrow abundance is the ultimate goal of this study. Therefore, in the calculation of the total number of objects observable in the video footage, the information computed by the tracking module needs to be considered. Each tracked object has been assigned a unique identity and its trajectory has been computed. The quantification module should be able to analyze each tracked object and reject false positives that may have been detected. To do so, this module searches objects that can be characterized as unstable over time, meaning that they are detected for short and discontinuous time periods. To ensure the consistency of the object count, a quantification strategy is applied in this step.

3.4.1 Lobster quantification

During the observation, and given that video sequences are captured by a ROV whose trajectory

is not predetermined (as one would expect in a linear trawling survey), lobsters may leave and possibly re-enter the observed scene later. However, since a lobster's activity after leaving the scene is unknown and unpredictable (e.g., it may have entered a burrow and stood there for a long period, eventually reappearing later at a different exit of the burrowing system), the decision was to stop the tracking of a given lobster if it leaves the scene, even if the trajectory of ROV can be estimated. For counting purposes, the quantification module takes into account any lobster object that satisfies

$$N_{\text{Detection}} \geq \delta, \tag{21}$$

where $N_{\text{Detection}}$ is the number of frames in which the lobster candidate is successfully tracked, and δ is a threshold corresponding to the minimum value to consider that the object does not correspond to a false detection. Herewith, δ is set to 10 after a set of preliminary tests, which means that the object will be detected in at least 10 frames.

3.4.2 Burrow quantification

At every frame, a burrow tracker is updated based on the ROV estimated motion model. The predicted position of a tracked burrow is updated based on the estimated ROV movement, even if outside the image. Therefore, the tracking of each detected burrow is kept until the last frame of the video. We assign two variables, N_{B} and N_{O}, for each burrow trajectory. At frame t, when the tracked burrow is detected again in the visible image area, then its N_{B} is increased by one. Otherwise, if the tracked burrow centroid is supposed to be found in the image area but not detected, then N_{O} is increased by one. After processing the complete video sequence, each tracked burrow is counted if its tracker information satisfies both Eq. (21) and

$$f = \frac{N_{\text{B}}}{N_{\text{B}} + N_{\text{O}}} \geq 0.5, \tag{22}$$

where f describes how frequently the tracked object is detected correctly when revisiting a given location. If f is greater than 0.5, it means that the burrow is tracked successfully, and thus can be counted as a detected burrow.

4 Experimental results

This section describes the dataset used to test the proposed system and the results obtained. The proposed automatic analysis software was implemented in C++, using Microsoft Visual Studio, integrated with the OpenCV (http://opencv.org/) open source library. Tests were conducted in a computer with an Intel core i7-3770@3.440-GHz processor and with a 16-GB RAM, running the Windows operating system.

4.1 Available dataset

To evaluate the performance of the proposed methodology, two challenging real-time underwater video sequences, recorded on different days, were used. These video sequences were captured by a video camera installed in a ROV using the experimental setups described in Section 2. The two long video sequences, around 2-h duration each, were split into several short video clips; after removing uninterested contents, e.g., due to marine snow or the positioning of ROV, a total of 17 meaningful clips were selected (Table 3). All the video clips were tested at 12 frames/s, with a resolution of 640×360. The average playing time of these video clips was 43 s, and the total playing time was around 12 min. One of these video clips was used for training, as discussed in Section 3.2.3. In some clips, both lobsters and burrows are visible, while in others only burrows or

Table 3 Selected video clips for analysis

Video	Evaluation	Playing time (s)	Lobster (L) Burrow (B)
1	Train	75	LB
2	Test	31	B
3	Test	53	B
4	Test	16	B
5	Test	25	L
6	Test	12	B
7	Test	70	B
8	Test	17	B
9	Test	50	–
10	Test	91	L
11	Test	33	B
12	Test	23	B
13	Test	65	B
14	Test	79	B
15	Test	31	L
16	Test	32	B
17	Test	33	B

Data were taken from two different long sequences

lobsters are. Seldom, more than one lobster is found in a single image.

To validate the system's automatic operation, the available video clips were labeled manually by an expert to be considered as ground-truth data. Note that even expert marine biologists cannot convey a high confidence in burrow classification labeling with results varying from person to person. This is an ill-defined problem, due to the inconsistent geometric appearance of burrow entrances, partly because of illumination inconsistencies and distance to the camera, as well as the nature of the sediments found in the seabed. Also, Norway lobsters are not the only burrowing species. Therefore, in the context of this study, the focus was on the primary burrow system entrances, which can be recognized easily as they present a better intensity, homogeneity, contrast, and a longer diameter which was assumed to be greater than 0.05% of the image width in this test. Note that, for quantification purposes, if the same burrow was labeled manually in different images, including when ROV returned to a previously visited location, it was not labeled as a new object.

4.2 Results

To evaluate the performance of the proposed visual tracking scheme, one of the CLEAR multiple object tracking (MOT) metrics, MOT accuracy (MOTA) was used as the evaluation metric for this experiment (Bernardin and Stiefelhagen, 2008). In this case, the evaluation commenced as items were detected and tracked. The higher the MOTA score, the more precise the tracking system. Note that the objective of this experiment was to investigate the tracking performance, and thus we did not consider whether the tracked objects were labeled correctly. In addition, note that a lobster which left the scene and then re-entered was assigned to a new track.

As seen in Table 4, the proposed tracking scheme achieved 1.0000 and 0.9902 MOTA scores for lobster and burrow tracking, respectively. When tracking multiple objects at a scene, the method can be prone to occlusion problems as the tracked object overlaps with others. This possibility can lead to missed tracking or false identity switching. In our case, the testing scenario did not face this problem because there was rarely more than one lobster present in the scene (lobster tracking), while burrow structures are static (burrow tracking). However, a few false positives were detected in burrow tracking due to the inaccurate global motion of the images (Fig. 12).

Table 4 Performance evaluation of the proposed system based on the multiple object tracking accuracy (MOTA) metric

Target	$\sum_{i=1}^{N} G_i$	MOTA	ML	FM	ID
Lobster	681	1.0000	0	0	0
Burrow	7036	0.9902	0	0.0098	0

ML: ratio of the missed tracks in total frames; FM: ratio of tracks with false positive in total frames; ID: ratio of mismatched id in total frames; G_i: number of mapped objects over the entire trajectory at the i^{th} frame

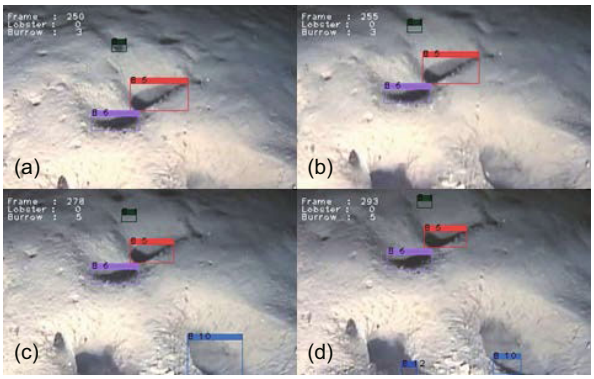


Fig. 12 False positive track for region B1 in the 250$^{\text{th}}$ (a), 255$^{\text{th}}$ (b), 278$^{\text{th}}$ (c), and 293$^{\text{rd}}$ (d) frames (References to color refer to the online version of this figure)

Bearing in mind that a proper quantification is our ultimate goal, we conducted another experiment to investigate the precision of the actual estimation of abundance of lobsters and burrows in underwater video sequences. The results of the tests using the proposed system, for counting lobsters and burrows, were compared with those obtained using the method proposed by Tan et al. (2014), and both methods were again compared with ground-truth data obtained by manual annotation (Table 5).

Table 5 showed that the proposed system correctly detected and tracked all five lobsters and 87 out of 104 burrows. In addition, 17 false positive burrows were detected. These results evidence a clear improvement over the method proposed by Tan et al. (2014), which correctly detected and tracked five lobsters and 76 burrows, additionally identifying one lobster and 34 burrows as false positives. Sample images illustrating the experimental results are included in Fig. 13, while Fig. 14 illustrates a

Table 5 Comparison between an automatic approach and a manual approach

Approach	Object	Manual count	Automatic count	True positive	False positive	False negative	Recall (%)	Precision (%)	F-score (%)
R1	Lobster	5	6	5	1	0	100.00	83.33	90.90
R1	Burrow	105	110	76	34	29	72.38	69.09	70.69
PS	Lobster	5	5	5	0	0	100.00	100.00	100.00
PS	Burrow	105	104	87	17	18	82.86	83.65	83.25

RI: methodological approach used in Tan et al. (2014) with the same tracking solution; PS: methodological approach used in this paper

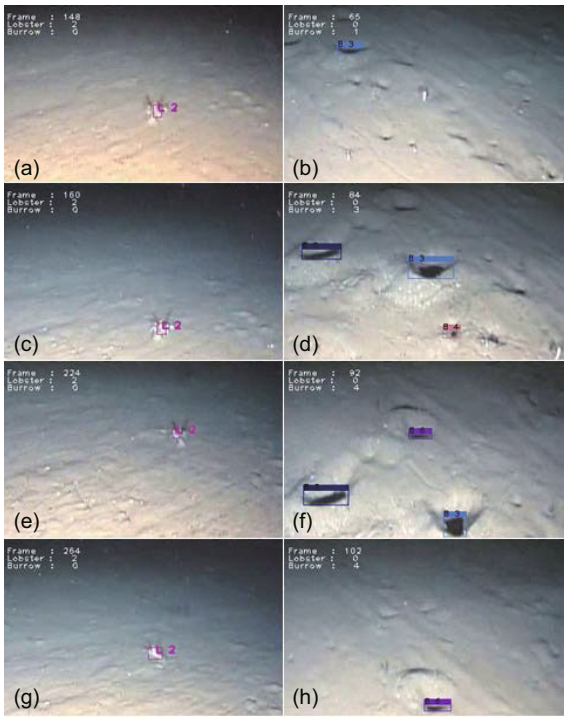


Fig. 13 Lobster tracking in the 5th video clip: 148th (a), 160th (c), 224th (e), and 264th (g) frames; burrow tracking in the 4th video clip: 65th (b), 84th (d), 92nd (f), and 102nd (h) frames

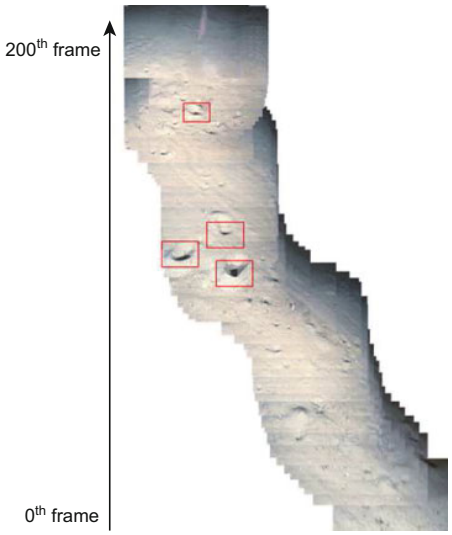


Fig. 14 Image mosaic created using one of the test video clips

mosaic image outlining the area surveyed by the movement of ROV during acquisition. As shown in Fig. 13, the lobster contour could vary according to different viewing angles and positions from the observer. However, the proposed lobster detector successfully detected and tracked the lobster across the video frames. Note that when a leaving lobster re-entered the observation scene, it was labeled with a new identity. Since the test video was obtained in the *Nephrops* fishing grounds at depths above 400 m, other benthic animals were rarely found in the scene with the exception of a few fish species appearing usually in the scene for short and temporary periods, i.e., in one or two frames. However, our proposed lobster detector was able to distinguish between them. In addition, when the distance between lobster and observer positions was large, leading to a small lobster size and a low intensity contrast compared with the surrounding region, the lobster detector did not commence. When the observation distance decreases, the visual appearance of the lobster can be noticed clearly, leading to a correct detection.

On the other hand, the geometric properties of burrows vary highly according to the distance and angle from the observer (Fig. 13). When the burrows were far away from the observer, their presence could not be noticed either because their entrance could not be seen or because most of them presented in a straight line. As the observation distance decreased, the burrow entrance became visible, with both the diameter and size becoming larger. Their shape morphology changed from a straight line to a downward crescent-like shape, and finally became a circle or ellipse. Similar to lobster detection, burrow detection was possible from only a short observation distance. The burrow detector proposed here analyzed the

burrow morphology based on the intensity contrast and several geometric features instead of shape or contour information due to the inconsistency of burrow shape appearance. In fact, the majority of burrows had downward curvature information; however, there were a few exceptional cases where the burrows presented with a sharp upper boundary (Fig. 15a). One of the geometric features used to describe the downward curvature could not be worked out since it presented with an upward curvature. This was one of the reasons that led to a missed burrow. A few challenging benthic or artifact structures found were over the seabed sediments in the test video. In particular, trawl marks were rejected correctly by the burrow detector (Fig. 16). Furthermore, we compared Norway lobster abundance estimation according to the automatic video processing techniques reported in Lau et al. (2008, 2012) and Sooknanan et al. (2013) (Table 6).

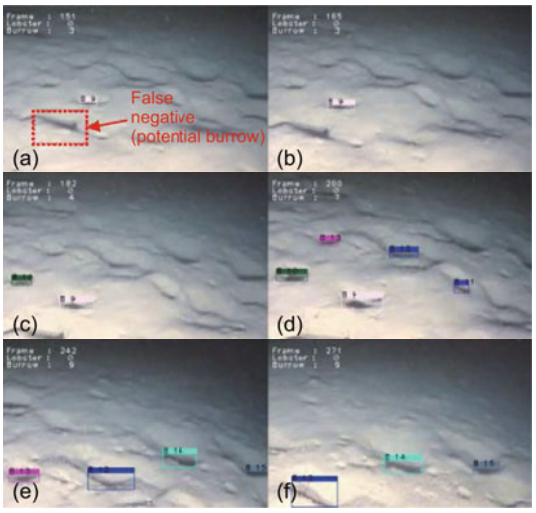


Fig. 15 Experiment results of the 13th video clip: tracking of multiple burrows in the 151st (a), 165th (b), 182nd (c), 200th (d), 242nd (e), and 271st (e) frames

Fig. 16 Trawl marks that were rejected correctly by the burrow detector in the 337th (a) and 599th (b) frames

5 Conclusions

In this paper, we proposed an automated video analysis framework, which constitutes a further step in facilitating Norway lobster stock assessment by automatically detecting, tracking, and finally quantifying both this benthic species and its biogenic feature (main burrow entrance) from the underwater video sequences obtained in deep-water crustacean grounds. The proposed method presented in this study is designed specifically to address the dynamic motion environment of video footage obtained under ROV operation. In this study, we improved the performance of lobster and burrow detection using a new algorithm. The results of testing a higher number of video sequences achieved 100% precision and recall for the lobsters and 83.7% precision and 82.9% recall for burrows. However, it must be stressed once again that fishery scientists evaluate Norway lobster abundance by counting burrow systems (assuming a single main entrance and several ventilation shafts) and not the lobsters themselves. In spite of a reasonable approximation with manual counting, the amount of missed (false negatives) and misidentified burrows (false positives) may represent a hindrance for the immediate adoption of this framework in field work. Its extensive experimentation with footage representing the different conditions found during survey cruises is mandatory, e.g., bottoms with good, acceptable, and poor visibility, and according to bottom composition and hardness which conditions the occurrence of marine snow.

Acknowledgements

The non-governmental organization OCEANA and the team of the project IMPACT 'Long-Term Effects of

Table 6 Comparison with other works estimating Norway lobster abundance

Method	Acquired method	Video vision	Target (L/B)	Analysis technique	Real time analysis	Dynamic observation	Video quality	Marine snow
Lau et al. (2008, 2012)	Towing	One-way	LB	Video	Yes	No	Yes	Yes
Sooknanan et al. (2013)	Sledge	One-way	B	Image	No	Yes	No	No
Proposed system	ROV	Free view	LB	Video	Yes	Yes	Yes	Yes

Continued Trawling on Deep-Water Muddy Ground', financed within the scope of the European Union program EUROFLEETS, are gratefully acknowledged for the authorization to use the underwater video footage analyzed herein.

References

Akbani R, Kwek S, Japkowicz N, 2004. Applying support vector machines to imbalanced datasets. Proc 15th European Conf on Machine Learning, p.39-50. https://doi.org/10.1007/978-3-540-30115-8_7

Badekas E, Papamarkos N, 2005. Automatic evaluation of document binarization results. Proc 10th Iberoamerican Congress Conf on Progress in Patt Recognition, Image Analysis and Applications, p.1005-1014. https://doi.org/10.1007/11578079_103

Ben-Hur A, Weston J, 2010. A user's guide to support vector machines. In: Carugo O, Eisenhaber F (Eds.), Data Mining Techniques for the Life Sciences. Humana Press, New York, p.223-239. https://doi.org/10.1007/978-1-60327-241-4_13

Bernardin K, Stiefelhagen R, 2008. Evaluating multiple object tracking performance: the CLEAR MOT metrics. *EURASIP J Image Video Process*, 2008:246309. https://doi.org/10.1155/2008/246309

Bouguet JY, 2000. Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the Algorithm. Intel Corporation Microprocessor Research Labs, Santa Clara, USA.

Correia PL, Lau PY, Fonseca P, et al., 2007. Underwater video analysis for Norway lobster stock quantification using multiple visual attention features. Proc 15th European Signal Processing Conf, p.1764-1768.

Denise S, 2007. Homework Helpers: Calculus (Homework Helpers). Career Press, Wayne.

Fonseca P, Correia PL, Campos A, et al., 2008. Fishery-independent estimation of benthic species density—a novel approach applied to Norway lobster *Nephrops norvegicus*. *Mar Ecol Prog Ser*, 369:267-271. https://doi.org/10.3354/meps076091

Howard FG, 1989. The Norway lobster. *Scott Fisher Inform Pamphl*, No. 7.

Johnsen S, Sosik H, 2004. Shedding light on light in the ocean. *Ocean Mag*, **43**(2):1-5.

Kuhn HW, 1955. The Hungarian method for the assignment problem. *Nav Res Log Q*, 2(1-2):83-97. https://doi.org/10.1002/nav.3800020109

Lau PY, Correia PL, Fonseca P, et al., 2008. I2N2: a software for the classification of benthic habitats characteristics. Proc 16th European Signal Processing Conf, p.1-5.

Lau PY, Correia PL, Fonseca P, et al., 2012. Estimating Norway lobster abundance from deep-water videos: an automatic approach. *IET Image Process*, 6(1):22-30. https://doi.org/10.1049/iet-ipr.2009.0426

Morello EB, Froglia C, Atkinson RJA, 2007. Underwater television as a fishery-independent method for stock assessment of Norway lobster (*Nephrops norvegicus*) in the central Adriatic Sea (Italy). *ICES J Mar Sci*, 64(6):1116-1123. https://doi.org/10.1093/icesjms/fsm082

Sardà F, Aguzzi J, 2012. A review of burrow counting as an alternative to other typical methods of assessment of Norway lobster populations. *Rev Fish Biol Fisher*, 22(2):409-422. https://doi.org/10.1007/s11160-011-9242-6

Sauvola J, Pietikäinen M, 2000. Adaptive document image binarization. *Patt Recogn*, 33(2):225-236. https://doi.org/10.1016/S0031-3203(99)00055-2

Shafait F, Keysers D, Breuel TM, 2008. Efficient implementation of local adaptive thresholding techniques using integral images. Proc SPIE, 6815:10. https://doi.org/10.1117/12.767755

Sooknanan K, Doyle J, Wilson J, et al., 2013. Mosaics for burrow detection in underwater surveillance video. OCEANS, p.1-6. https://doi.org/10.23919/OCEANS.2013.6741296

Struc V, Vesnicer B, Pavesic N, 2008. The phase-based Gabor fisher classifier and its application to face recognition under varying illumination conditions. Proc 2nd Int Conf on Signal Processing and Communication Systems, p.1-6. https://doi.org/10.1109/ICSPCS.2008.4813663

Suzuki S, Be K, 1985. Topological structural analysis of digitized binary images by border following. *Comput Vis Graph Image Process*, 30(1):32-46.

Tan CS, Lau PY, Low TJ, et al., 2014. Detection of marine species on underwater video images. Int Workshop on Advanced Image Technology, p.192-196.

Tan CS, Lau PY, Correia PL, et al., 2015. A tracking scheme for Norway lobster and burrow abundance estimation in underwater video sequences. Proc Int Workshop on Advanced Image Technology.

Yang CJ, Duraiswami R, Davis L, 2005. Fast multiple object tracking via a hierarchical particle filter. Proc IEEE Int Conf on Computer Vision, p.212-219. https://doi.org/10.1109/ICCV.2005.95