Machine Learning for the Evaluation of the *Nephrops norvegicus* Population

Marco Reggiannini[1 [0000−0002−4872−9541]], Michela Martinelli[2 [0000−0003−1231−2346]], Oscar Papini[1 [0000−0003−2069−5068]], Lorenzo Zacchetti[2,3 [0000−0003−3179−7473]], Filippo Domenichetti[2 [0000−0002−8052−7544]], and Gabriele Pieri[1 [0000−0001−5068−2861]]

[1] Institute of Information Science and Technologies (ISTI), National Research Council of Italy, Via G. Moruzzi 1, 56124 Pisa, Italy
[2] Institute for Marine Biological Resources and Biotechnology (IRBIM), National Research Council of Italy, Largo Fiera della Pesca 2, 60125 Ancona, Italy
[3] Department of Biological, Geological, and Environmental Sciences (BiGeA), Alma Mater Studiorum – University of Bologna, Piazza di Porta S. Donato 1, 40126 Bologna, Italy

Abstract This paper introduces computer vision methods for detecting, recognising, and estimating *Nephrops norvegicus* (Norway lobster) burrow density via Underwater Television surveys. The current manual approach involves human operators visually assessing videos, which is prone to errors and subjectivity. Automated machine learning systems show promise in identifying and counting burrows, potentially standardising recognition and reducing operator errors. However, challenges exist in implementing computer vision techniques. An automated system aims to process video streams, detect seabed openings, extract visual features, and classify *N. norvegicus* burrows, significantly advancing the automation of underwater video reading. The primary processing presented in the paper lies in a boosting algorithm capable of extending the original annotated ground truth and assessing the improved performance of the extended data set with respect to the original one.

Keywords: Computer vision · Video annotation · Deep learning · Fisheries · *Nephrops norvegicus*

1 Introduction

Nephrops norvegicus (Norway lobster), a benthic burrowing decapod crustacean, is found on the continental shelf of the northeastern Atlantic Ocean and the Mediterranean Sea, typically at depths ranging from 50 to 800 metres [1]. This species is pivotal in European fisheries, contributing significantly with landings nearly reaching 60 000 tonnes; in the Mediterranean Sea, during the last twenty years an average of 3700 tonnes of Norway lobster have been landed every year, but since 2010 there has been a notable decrease in the landings [2]. Conventional techniques for evaluating *N. norvegicus* stocks involve animal sampling through

fishery-dependent methods and/or underwater video cameras. To assess burrow densities, Underwater Television (UWTV) surveys are carried out [3, 4]. UWTV surveys began in 1992 in Scotland and today Iceland, Denmark, Sweden, United Kingdom, Ireland, France, Spain, Italy and Croatia are involved in this type of activity. In the Adriatic Sea, in particular in the area of the Pomo/Jabuka Pits, UWTV surveys were conducted since 2009. Burrow densities, measured in burrows per square metre, provide valuable insights into stock abundance [5].

The UWTV approach is based on the inspection of a known seafloor surface to count the number of *N. norvegicus* burrows distinguishing them from other animal burrows or artefacts. Currently, this procedure relies on the direct visual assessment of hours of videos manually carried out by human operators, that is a demanding task, significantly prone to errors and possibly influenced by human subjectivity. Several workshops organised by the International Council for the Exploration of the Sea (ICES) were held with the aim of standardising video reading methodologies and quantifying the related uncertainties [6, 7, 8, 9]. Among the procedures developed and generally adopted by UWTV operators is that trained scientists who analyse the videos must first undergo a training procedure involving the use of reference footage from specific areas; scripts were developed to analytically test the concordance of the results produced by various readers and normally the final result is considered the average of their counting.

The development of automated systems based on machine learning methods for these purposes already showed promising results [10, 11, 12]. However, numerous issues must be taken into account when using computer vision techniques. For instance, concerning optical sensing, images captured by cameras in an underwater environment are typically affected by the presence of suspended matter in the water column. This increases the water turbidity and lowers the acquired imagery's contrast. Additionally, light reflected by the environment is effectively filtered by the water medium by an amount proportional to the length of its propagation trajectory towards the camera sensor. This process significantly alters the colour content in the image, partially preserving blue and green components. Furthermore, the typical depths (normally below 50 m) at which UWTV surveys take place constrain the observation of the scene to be carried out with artificial lights, introducing luminosity imbalances and contrast unevenness in the images.

Operationally, detecting burrows concerns identifying openings in the seabed (examples are visible in Figure 1) that constitute the crustaceans' burrows. A first requirement concerns the robustness of the detector with respect to environmental artefacts that may resemble the animal nests and their features, such as sand aggregations, rocks or shadows. Secondly, several crustacean and fish species are present in the surveyed areas, each potentially generating one or more types of burrows. Therefore, the system must be endowed with the capability to automatically recognise the specific type of cavities that refer to *N. norvegicus*. A final challenge lies in identifying burrows, which are systems with more than one opening connected via underground tunnels. These should be considered only once in the counting process.

Figure 1. Example of a UWTV survey frame recorded in the central Adriatic Sea. Burrows related to various species are visible.

In a nutshell, the main objective addressed by a *N. norvegicus* automatic counting system consists of processing a video stream to detect the openings locations, extract discriminating visual features, such as geometry, texture or radiometric properties of the image, use these features to identify the burrows belonging to the *N. norvegicus* class among the several observed ones and count the resulting detections. The availability of such a tool would represent a significant advancement towards the automation of the UWTV reading.

This work mainly deals with the problem of building sufficiently extensive datasets, suitable for being used as ground truth datasets for machine learning algorithms, employing only a minimal level of manual work. In particular, a technique is described that is able to automatically augment a handcrafted, small dataset of annotated *N. norvegicus* burrows imagery.

The paper is structured as follows: Section 2 summarises different approaches that can be used in the context of *N. norvegicus* burrows detection in UWTV footage; Section 3 describes the dataset boosting technique; Section 4 shows the advantages of this extension method by comparing the predicting capabilities of two instances of a deep neural network, one trained using a manually created dataset and the other trained using the extended version of that dataset; finally Section 5 concludes the paper and presents possible future developments on this work.

2 Classification Approaches

For the purposes of this research activity, automatic systems aim at assessing whether an input image contains targets of interest, namely burrows related to crustacean fauna, with one or more openings located over the seabed surface, and to specify further if the candidate target belongs to the *N. norvegicus* class. A usual approach to this problem envisages the extraction of image features that represent specific properties of the image content. These quantities are later fed to an algorithm in charge of assigning a class label based on the values of the features. Classically, the typology of the extracted features is defined *a priori*, regarding, for example, geometrical, textural or colour attributes of the input data. A complete control of the feature typology ensures awareness about the informative properties that are later exploited as discriminating factors in the classification process. On the other hand, this approach requires the capability of selecting the right features, i.e. those that carry the largest discriminating power for the considered task.

Currently, powerful detection and classification methods rely more and more on deep neural networks architectures—algorithms based on self-driven features learning. Indeed, a network learns the features for a specific classification task, mainly through the injection of annotated data and by updating its internal parameters in order to align the algorithm output with the annotated data. This process is completely automatic except for the generation of the starting annotated dataset. Related literature unanimously claims that the modern deep learning approach ensures outstanding performances [13]. In the latter case, it is worth pointing out that complete awareness and control over the typology of features employed by the classifier is not possible. Moreover, a huge amount of annotated data is required in the first place to train the network model for the specific task of interest.

No matter which of the presented techniques is adopted, the availability of a properly numerous and annotated dataset is a starting requirement of paramount relevance in every classification task.

3 Imagery Annotation Booster

The generation of the ground truth dataset is an important objective that must be pursued thanks to the collaboration with experts in the specific field. In this case, the annotation consists in the labelling of frames extracted from the video streams collected through UWTV campaigns. Given a UWTV frame, the labelling operation basically consists in manually delimiting one or more areas of interest (e.g. by means of bounding boxes) where the annotator recognises an object belonging to one of the categories of interest (see Figure 2). Assuming 25 images per second of footage as a typical framerate, the number of frames to be annotated sums up to 1500 per minute of the UWTV campaign. This implies a considerable amount of human effort to carry out a boring manual task, prone to errors due to possible attention drops after long annotation periods.

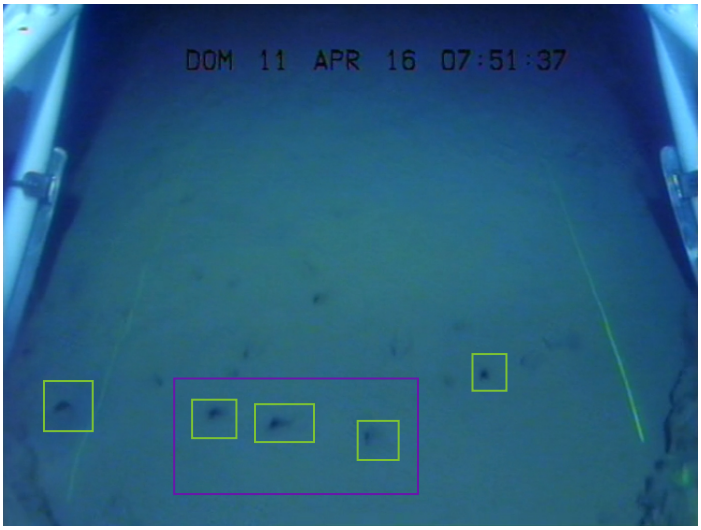


Figure 2. Example of ground truth annotation on the frame in Figure 1. Individual *N. norvegicus* entrances are tagged by green boxes while burrows are annotated through purple boxes.

The primary objective of this work is to devise an alternative method that allows to obtain the largest possible ground truth dataset by exploiting a minimum effort from human experts. The procedure adopted here gets inspiration from the simple fact that a certain object captured during the survey typically appears in multiple consecutive frames. Actually, the UWTV survey is usually performed through a downward-looking camera installed aboard a sledge, which is dragged, ideally at a constant speed, by a cable-connected surface vessel. Assuming a piecewise linear trajectory, an object in the video stream usually appears in N_p consecutive frames. If any of these frames is tagged by experts due to the presence of that object, the resulting annotations can be exploited to perform a correlation-based template matching on the remaining frames (see Figure 3).

Say F_i is a UWTV frame captured at discrete time index t_i and w the template extracted from F_i, i.e. the image patch corresponding to the bounding box defined by the annotator. Assume for clarity the simple case where t_i coincides with the time of the first occurrence of w in the video. It is then expected that w will also be observed in the following $N_p - 1$ frames. The template matching step actually consists of looking for the maximum of the discrete cross-correlation function (s, t, l, m represent the image indexes)

$$(w \star F_j)(l, m) = \sum_s \sum_t w(s, t) F_j(l + s, m + t).$$

Replicating the same operation for each $j \in [i, i + N_p - 1]$ allows detecting the template w in each frame of the considered time range (see Figure 4).

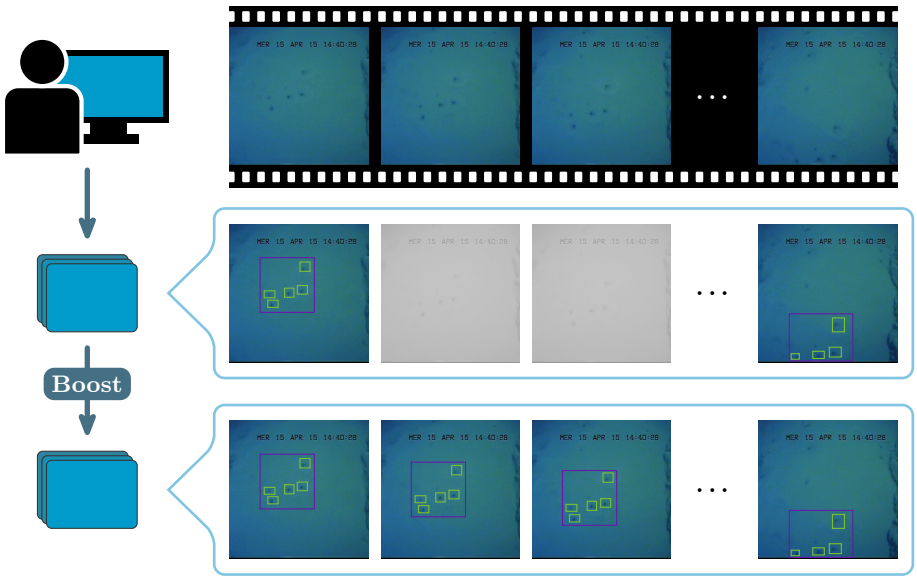


Figure 3. Concept of the annotation booster. A human annotator watches the UWTV video footage and tags only some frames; the booster algorithm then automatically adds the missing bounding boxes in the intermediate frames.

Iterating this approach for each available annotation will return a novel ground truth dataset, whose dimension is boosted, in the most favourable case, i.e. when template matching succeeds in extending the annotations to the largest possible portion of UWTV video, by an N_p multiplicative factor.

4 Experiments

As a first attempt at the automatic detection of *N. norvegicus* burrows using a deep learning approach, an annotated dataset has been prepared starting from video footage recorded during a campaign conducted jointly by the Institute for Marine Biological Resources and Biotechnology of CNR (Italy) and the Institute of Oceanography and Fisheries (Croatia) in the Pomo/Jabuka Pits area in the central Adriatic Sea [5, 10]. In particular, the source material consists of one video with a duration of about 4 minutes and 12 seconds, a resolution of 768×576 pixels and a framerate of 25 FPS. The only preprocessing performed on the video is the application of the Neural Network Edge Directed Interpolation (NNEDI) deinterlacing filter; then a set of 505 images has been obtained by extracting two frames per second and provided to the annotators.

The annotation process has been carried out using the Microsoft VoTT annotation tool (https://github.com/microsoft/VoTT). Two classes have been

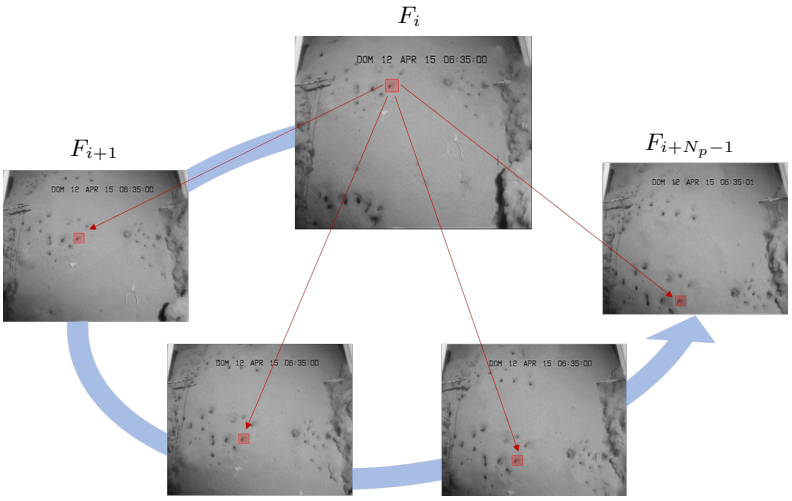


Figure 4. Template matching concept. The template w is localised in the neighbouring frames.

chosen (see Figure 2): the class "Opening", used to mark the single entrances of a *N. norvegicus* burrow (in green in the figure), and the class "Burrow" that groups all the openings belonging to a single burrow (in purple). In the end, the whole annotation process resulted in a dataset which will be called "original" from now on. This dataset has then been extended using the boosting tool described in Section 3 applied to the full 25 FPS video, obtaining a richer dataset, referred to as "extended" in the following. Table 1 displays the characteristics of the two datasets.

Table 1. Comparison of the original and extended datasets.

	Original	Extended
Total number of images	156	2406
Total number of objects	394	5528
of which "Openings"	315	4844
of which "Burrows"	79	684

Both the original and the extended datasets were used to train two separate instances of YOLOv4 [14], a deep neural network designed for the object detection task. In both cases, the dataset is randomly split into a training set (80% of images) and a validation set (20%), and the network has been trained for 10 000 iterations, applying a 5-fold cross-validation, obtaining two detection models M_{orig} and M_{ext}; these can be seen as functions that take as input an image, e.g. a frame from UWTV footage, and return a list of (predicted) bound-

ing boxes, each with a label (either "Opening" or "Burrow") and a score, i.e. a number between 0 and 1 that represents the level of confidence given to the box by the model, where a higher number represents a higher level of certainty. A box that receives a score below 0.25 is not considered as a detection.

The performance of an object detection model is usually assessed with a parameter known as *average precision* (AP), that is a number between 0 and 1 (the higher, the better) computed starting from a set of images for which there are both ground truth and predicted bounding boxes. The AP is computed separately for the boxes belonging to each class, and then the *mean average precision* (mAP) is defined as the average of the APs of all classes. The exact method of computation of the AP has changed over time (see, for example, [15]), so it is useful to summarise here the one used for this work.

Recall that, given two rectangles A and B, their *intersection over union* (IoU) is the ratio between the area of their intersection and the area of their union (see Figure 5); in the following it is denoted by $\mathrm{IoU}(A, B)$. Consider a set of images as above and let $\mathcal{G}$ and $\mathcal{P}$ be respectively the set of all ground truth and predicted bounding boxes for a given class. Assume that $\mathcal{P} = \{P_1, \dots, P_N\}$ is sorted by score decreasing, so that P_1 is the predicted box with maximal score, and choose a threshold k for the IoU (a typical value is 0.5).

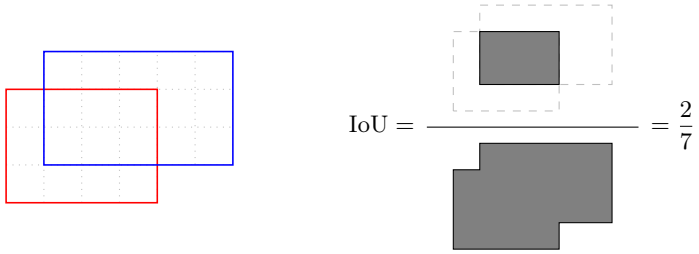


Figure 5. The IoU of the two rectangles is the ratio between the areas of their intersection and their union.

1. For each $i = 1, \dots, N$, decide if P_i is a *true positive* (TP) or a *false positive* (FP) in the following way:
   (a) P_1 is a TP if and only if the set $W_1 \coloneqq \{G \in \mathcal{G} \mid \mathrm{IoU}(P_1, G) > k\}$ is not empty; if P_1 is a TP, let $G_1 \in W_1$ be the box with maximal IoU with P_1 and let $\mathcal{G}_1 \coloneqq \mathcal{G} \setminus \{G_1\}$, otherwise let $\mathcal{G}_1 = \mathcal{G}$;
   (b) consider now P_2: again, it is a TP if and only if the set $W_2 \coloneqq \{G \in \mathcal{G}_1 \mid \mathrm{IoU}(P_2, G) > k\}$ is not empty; if P_2 is a TP, let $G_2 \in W_2$ be the box with maximal IoU with P_2 and let $\mathcal{G}_2 \coloneqq \mathcal{G}_1 \setminus \{G_2\}$, otherwise let $\mathcal{G}_2 = \mathcal{G}_1$;
   (c) continue in order for all P_i.
2. For each $i = 1, \dots, N$ compute the cumulative precision

$$\tilde{p}_i \coloneqq \frac{\text{number of TP among } P_1, \dots, P_i}{i}$$

and the cumulative recall

$$\tilde{r}_i \coloneqq \frac{\text{number of TP among } P_1, \ldots, P_i}{|\mathcal{G}|}.$$

3. For each *unique* recall value consider only the maximal precision associated with it. In other words, let $0 = r_1 < r_2 < \cdots < r_n = 1$ be the set of the unique recall values among $\{\tilde{r}_1, \ldots, \tilde{r}_N\} \cup \{0, 1\}$ and for each r_i, $i = 1, \ldots, n$, define
$$\begin{cases} p_1 = 1 & \text{if } r_1 = 0 \notin \{\tilde{r}_1, \ldots, \tilde{r}_N\}, \\ p_n = 0 & \text{if } r_n = 1 \notin \{\tilde{r}_1, \ldots, \tilde{r}_N\}, \\ p_i \coloneqq \max\{\tilde{p}_j \mid \tilde{r}_j = r_i\} & \text{otherwise.} \end{cases}$$
The set $\mathcal{C} = \{(r_1, p_1), \ldots, (r_n, p_n)\}$ is the so-called *precision-recall (PR) curve*.
4. Compute the *non-increasing* PR curve, i.e. consider the set of pairs $\mathcal{C}' = \{(r_1, p'_1), \ldots, (r_n, p'_n)\}$ where $p'_i = \max\{p_i, p_{i+1}, \ldots, p_n\}$ for $i = 1, \ldots, n$.
5. The AP is the area below the non-increasing PR curve, computed with the lower rectangle approximation:
$$\text{AP} = \sum_{i=2}^{n} (r_i - r_{i-1}) p'_i.$$

The plots in Figure 6 show the trend of the mAP computed for the validation sets during the training process, meaning that every 100–120 iteration steps (after the first 1000) the network is "freezed" and its performance at that moment is assessed. The difference really stands out: the mAP of the network trained with the original dataset almost never exceeds 40%; furthermore it seems to *decrease* with the number of iterations, eventually falling to between 20% and 25%. On the other hand, the mAP of the network trained with the extended dataset shows a more regular trend, constantly increasing (albeit with some oscillations) up to about 75% with peaks around 80%.

The following Figures 7 and 8 show some examples of predictions. To allow a direct comparison, the same frames have been fed to the two models. The numbers appearing in the pictures are the scores of the boxes.

Figure 7a represents a frame extracted from the datasets (it is present in the validation sets of both of them); the boxes predicted by M_{orig} and M_{ext} are shown in Figures 7b and 7c respectively. In this example, M_{orig} fails to detect most of the openings and the system on the left of the frame (and gives a low score to the only opening it detects), and produces a system on the right which is a false positive; on the other hand, M_{ext} is able to locate all and only the objects annotated in the ground truth with high confidence.

The frames of Figure 8 come from another video with length of about 4 minutes and 41 seconds (7024 frames) and features similar to the one used for the training of the networks. The pictures in the first column (Figures 8a, 8c and 8e) show the bounding boxes predicted by M_{orig}, while the second one (Figures 8b, 8d and 8f) displays the predictions of M_{ext} on the same frames.

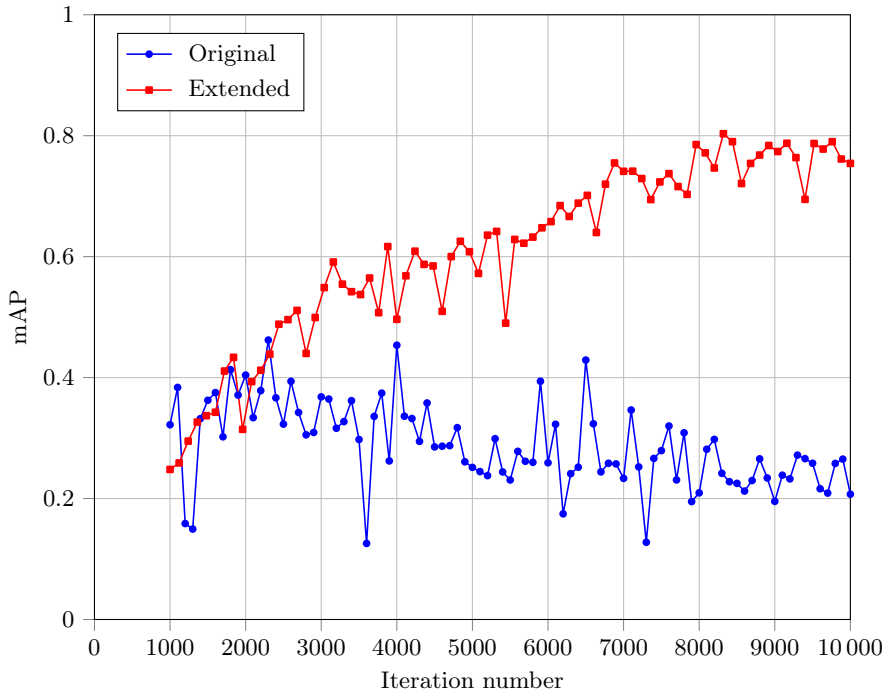


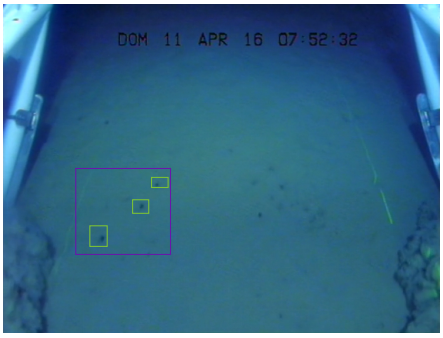
Figure 6. Plots of the mAP trend during the training with the two datasets.

A quantitative assessment of the quality of the predictions is not currently available since this new video has not been annotated yet, however a preliminary visual comparison highlights the different behaviours of M_{orig} and M_{ext}: the former tend to find more boxes with lower scores than the latter, thus suggesting that M_{ext} is more "confident" in its choices than M_{orig}. This is further confirmed by looking at the total number of bounding boxes returned by the two models (Figure 9).

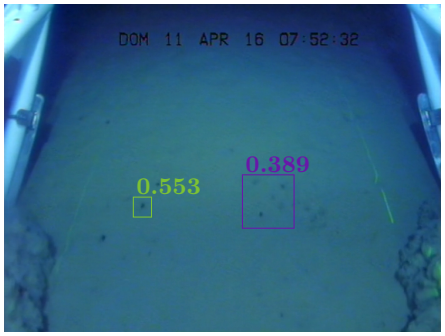
5 Conclusions

At present time, marine biologists do not have a consistent and repeatable methodology to carry out the *N. norvegicus* resource assessment. The long-term objective of this activity is the development of computer science methods conceived to detect and recognise the burrows and estimate their density. This approach might speed up technical operations and contribute to standardise the recognition of Norway lobster's burrows, also placing a limit on the human subjectivity in analysing UWTV data and reducing the human effort.

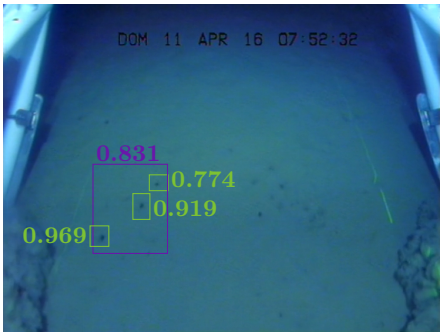
This document reports on the development of a computer vision tool to support the annotation process. Given a set of frames extracted from an UWTV video and properly annotated to generate a ground truth, the described tool



(a) Ground truth boxes.



(b) Boxes predicted by M_{orig}.



(c) Boxes predicted by M_{ext}.

Figure 7. Example of different predictions of the two models for the same annotated frame (belonging to the validation sets of both the original and the extended datasets).

allows to expand such a reference set automatically. This is obtained by considering annotated objects as templates and exploiting a correlation-based technique to match those templates in neighbouring frames. Eventually, this increases the number of ground truth annotations roughly by an order of magnitude. A noteworthy consequence was that the substantial extension in the number of annotated objects allowed to generate a sufficiently large set for the purpose of training a deep neural network model. The increase in classification performance due to this operation has been described in Section 4.

Possible next steps include further analysis of the model performance under different experimental circumstances. Additional datasets will be generated using videos collected in environmental scenarios with varying water turbidity levels, seabed typology, fauna species and camera sensor set-up, and will be considered for the model training. The performances will be analysed to assess the generalisation capability of the model.

The resulting amount of annotated objects will represent a massive wealth of data that will be exploited to provide insights and possibly improve the object detection task. A statistical analysis performed on features, e.g. geometrical

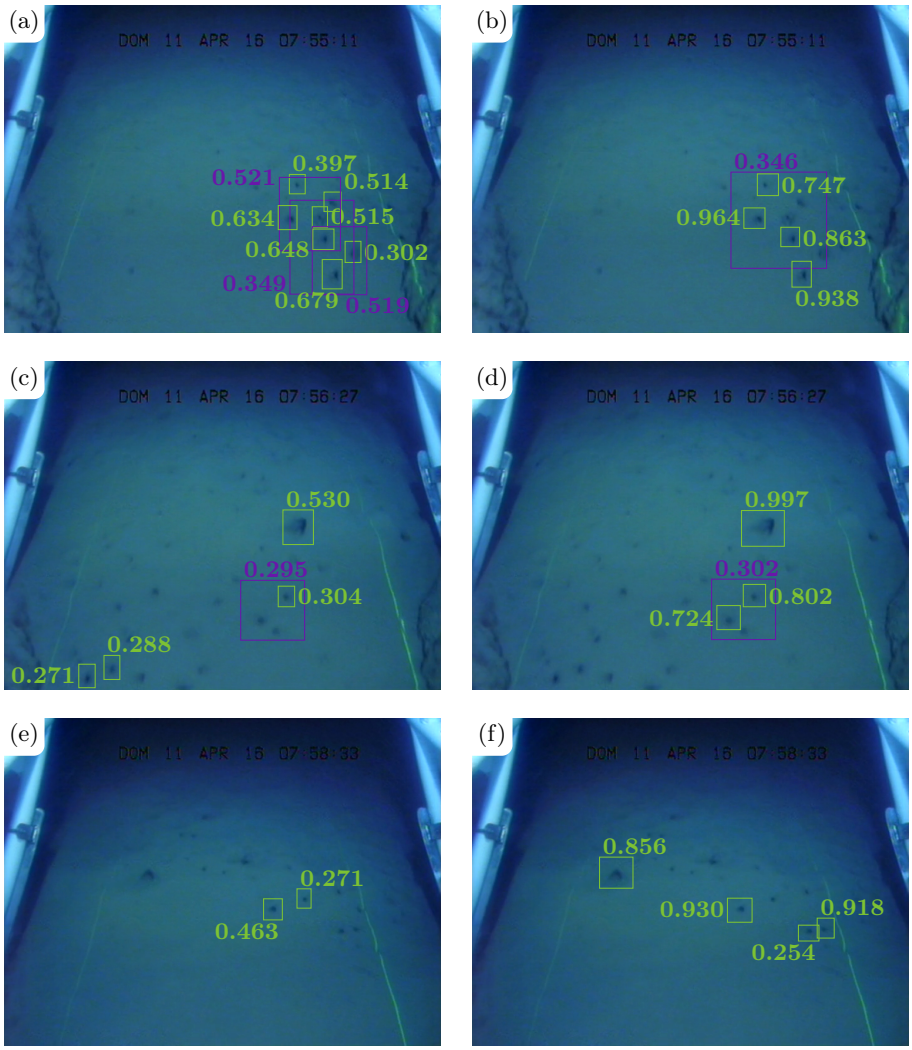


Figure 8. Some examples of detected burrows.

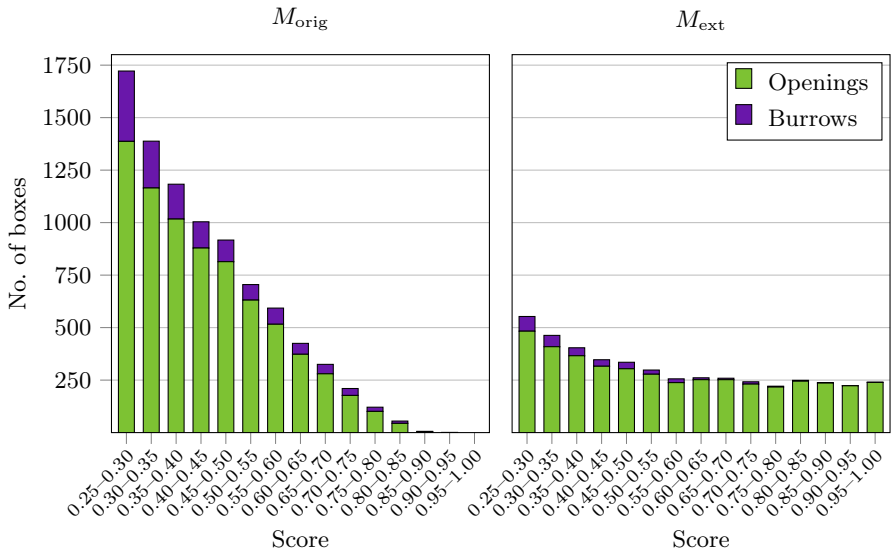


Figure 9. Number of predicted bounding boxes aggregated by score. The total number of boxes is 8657 for M_{orig} and 4592 for M_{ext}.

or textural, extracted from the annotated images could allow to identify the most representative visual properties of the *N. norvegicus* class, thus providing useful information that, together with the currently available knowledge, could strengthen the specific object detector.

The problem of multi-frame detection and tracking of the same object will be taken into account to implement *N. norvegicus* burrows recognition and counting.

Acknowledgments. This work is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101000825 – NAUTILOS Project (https://www.nautilos-h2020.eu).

References

1. Bell, M.C., Redant, F., Tuck, I.: "*Nephrops* Species". In: Lobsters: Biology, Management, Aquaculture and Fisheries. Ed. by B.F. Phillips. Blackwell Publishing Ltd, 2006. Chap. 13, pp. 412–461. https://doi.org/10.1002/9780470995969.ch13.
2. FAO Fisheries Division, Statistics and Information Branch, FishStatJ: Software for Fishery and Aquaculture Statistical Time Series, Accessed on 21 March 2024. https://www.fao.org/fishery/en/statistics/software/fishstatj
3. Dobby, H. *et al.*: ICES Survey Protocols – Manual for Nephrops Underwater TV Surveys, coordinated under ICES Working Group on *Nephrops* Surveys (WGNEPS). ICES Techniques in Marine Environmental Science (2021). https://doi.org/10.17895/ices.pub.8014

4. Aguzzi, J. *et al.*: Advancing fishery-independent stock assessments for the Norway lobster (*Nephrops norvegicus*) with new monitoring technologies. Frontiers in Marine Science **9** (2022). https://doi.org/10.3389/fmars.2022.969071
5. Martinelli, M. *et al.*: Towed underwater television towards the quantification of Norway lobster, squat lobsters and sea pens in the Adriatic Sea. Acta Adriatica **54**(1), 3–12 (2013)
6. Campbell, N., Dobby, H., Bailey, N.: Investigating and mitigating uncertainties in the assessment of Scottish *Nephrops norvegicus* populations using simulated underwater television data. ICES Journal of Marine Science **66**(4), 646–655 (2009). https://doi.org/10.1093/icesjms/fsp046
7. ICES, Report of the Workshop and training course on *Nephrops* burrow identification (WKNEPHBID). ICES CM 2008/LRC:03, 44 pp. 25–29 February 2008, Belfast, Northern Ireland, UK (2008). https://doi.org/10.17895/ices.pub.9828
8. ICES, Report of the Workshop on *Nephrops* Burrow Counting (WKNEPS). ICES CM 2016/SSGIEOM:34, 62 pp. 9–11 November 2016, Reykjavík, Iceland (2017). https://doi.org/10.17895/ices.pub.8651
9. ICES, Report of the Workshop on *Nephrops* Burrow Counting (WKNEPS). ICES CM 2018/EOSG:25, 44 pp. 2–5 October 2018, Aberdeen, UK (2019). https://doi.org/10.17895/ices.pub.8180
10. ICES, Working Group on Nephrops Surveys (WGNEPS; outputs from 2021 meeting). ICES Scientific Reports 4:29, 183 pp. (2022). https://doi.org/10.17895/ices.pub.19438472
11. ICES, Working Group on Nephrops Surveys (WGNEPS; outputs from 2022 meeting). ICES Scientific Reports 5:26, 125 pp. (2023). https://doi.org/10.17895/ices.pub.22211161.v1
12. Naseer, A.: Advanced Detections of Norway Lobster (Nephrops Norvegicus) Burrows using Deep Learning Techniques. International Journal of Advanced Computer Science and Applications **14**(3), 493–499 (2023). https://doi.org/10.14569/IJACSA.2023.0140357
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS'12), Volume 1, pp. 1097–1105, Lake Tahoe, Nevada (2012). https://dl.acm.org/doi/10.5555/2999134.2999257
14. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: YOLOv4: Optimal Speed and Accuracy of Object Detection, (2020). arXiv: 2004.10934 [cs.CV]
15. Terven, J., Córdova-Esparza, D.-M., Romero-González, J.-A.: A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. Machine Learning and Knowledge Extraction **5**(4), 1680–1716 (2023). https://doi.org/10.3390/make5040083. arXiv: 2304.00501 [cs.CV]