Towards automated scyphistoma census in underwater imagery: a useful research and monitoring tool

Martin Vodopivec[a,1,*], Rok Mandeljc[b], Tihomir Makovec[a], Alenka Malej[a,c], Matej Kristan[b]

[a] *National Institute of Biology, Marine Biology Station, Fornače 41, 6330 Piran, Slovenia*
[b] *University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, 1000 Ljubljana, Slovenia*
[c] *M-aleja, Čevljarska ulica 35, 6000 Koper, Slovenia*

Abstract

Manual annotation and counting of entities in underwater photographs is common in many branches of marine biology. With a marked increase of jellyfish populations worldwide, understanding the dynamics of the polyp (scyphistoma) stage of their life-cycle is becoming increasingly important. *In-situ* studies of polyp population dynamics are scarce due to small size of the polyps and tedious manual work required to annotate and count large numbers of items in underwater photographs. We devised an experiment which shows a large variance between human annotators, as well as in annotations made by the same annotator. We have tackled this problem, which is present in many areas of marine biology, by developing a method for automated detection and counting. Our polyp counter (PoCo) uses a two-stage approach with a fast detector (Aggregated Channel Features) and a precise classifier consisting of a pre-trained Convolutional Neural Network and a Support Vector Machine. PoCo was tested on a year-long image dataset and performed with accuracy comparable to human annotators but with 70-fold reduction in time. The algorithm can be used in many marine biology applications, vastly reducing the amount of manual labor and enabling processing of much larger datasets. The source code is freely available on GitHub.

Keywords: Jellyfish, Scyphozoa, Automated counting, Convolutional neural networks

1. Introduction

Jellyfish represent an important component of marine biota characterized by a large numerical variability (Boero et al., 2008) and sometimes appear in huge masses (blooms). These mass events attract public attention since they prevailingly occur in coastal and shelf seas and frequently interfere with human enterprises (Lucas et al., 2014). An increase in frequency and intensity of jellyfish blooms has been reported in areas worldwide (Brotz et al., 2012; Purcell, 2012; Kogovsek et al., 2018). In these locations jellyfish are affecting tourism, clogging intakes of power-plants and desalination facilities and are also causing harm to fisheries and aquacultures (Richardson et al., 2009; Purcell, 2012; Gershwin, 2013). On the other hand, the high demand for jellyfish in some Asian food markets has opened new possibilities to fishermen in many parts of the world (Brotz and Pauly, 2016). Recent research shows that the role of jellyfish in the ecosystem is far from negligible. They are voracious predators of plankton including fish eggs and larvae, they deplete food for other plankton feeders (Graham et al., 2014) and might be crucial in transporting or-

*Corresponding author
Email address: vodopivec.martin@gmail.com (Martin Vodopivec)
[1] Present address: Slovenian Environment Agency, Vojkova 1b, 1000 Ljubljana, Slovenia

ganic carbon into deeper parts of the water column (Lebrato et al., 2013). When decomposing in large numbers, jellyfish cause significant bacterial community shifts (Tinta et al., 2010, 2012) and influence the nitrogen cycle (Tinta et al., 2016).

Seventy percent the reported mass occurrences of gelatinous taxa are attributable to scyphozoans (Lucas and Dawson, 2014). Among the key traits that facilitate production of large number of individuals is a bipartite life history – the life cycle of most Scyphozoa alternates between a free-swimming medusa and an attached polyp. Sexually-reproducing pelagic medusa generates planulae which, after settling, develop into polyps, which in turn asexually produce free-swimming ephyrae (juvenile medusae). While most medusae reach sizes in the order of tens of centimeters, the polyps are much smaller (few millimeters) and are much harder to find and monitor. Consequently most scientific research focused on the medusa phase. Understanding the polyp population dynamics is vital for explaining the mechanisms and dynamics of jellyfish blooms since polyps have the potential to produce large numbers of recruits to the medusa population (Lucas and Dawson, 2014). Despite the importance of the perennial polyp phase, the in-situ population studies are scarce (Willcox et al., 2008; Purcell et al., 2009; Makabe et al., 2014; Hočevar et al., 2018; Fox et al., 2016). Our knowledge is largely derived from laboratory experiments (Lucas et al., 2012) and the reason likely lies in the substantial and tedious work required for obtaining sufficient amounts of representative data (Olariaga et al., 2014).

Recent research shows that coastal (Malej et al., 2012; Makabe et al., 2014; Hočevar et al., 2018) and offshore (Janßen et al., 2013; Vodopivec et al., 2017) man-made structures could have a significant impact on jellyfish populations in many areas. The polyps preferably attach to downward facing hard substrate (Schiariti et al., 2014) and such structures are scarce, which results in polyp populations clustering on relatively small regions. Understading the causes of jellyfish blooms and properly evaluate the influence of marine man-made constructions on jellyfish population dynamics, requires numerous field studies, based on quantifying polyp abundance over a period of several months/seasons. The change at each temporal unit is estimated by counting the number of polyps in a wild population, which can be done using underwater photographs. Each image may contain over a thousand polyps (Hočevar et al., 2018), which demands a significant focus, accuracy and time from the expert. Therefore, a method that could speed up the annotation process and reduce the amount of manual work would significantly improve the feasibility of large scale *in-situ* experiments.

With a rapid development of digital imaging techniques manual counting issues are becoming common in various branches of biology. The capacity to analyze images has not advanced at the same pace as the ability to collect them (Durden et al., 2016a; Beijbom et al., 2015). Monotonous operations like manual tagging and counting lead to reduction of the annotator's concentration. Furthermore, the manual work is subjected to the annotator's bias and their prior experience with the annotation domain (Schoening et al., 2017). The obtained counts of target entities allow researchers to estimate their population and infer individual characteristics, yet the amount of manual work often precludes large-scale studies. Therefore, automatic counting solutions based on image-processing techniques are often sought, both in order to alleviate the manual work-load and to reduce inter- and intra-operator variance.

In most cases, automatic counting is used on in-vitro samples since it allows a high degree of control over the image background (Friedland et al., 2005). Typically the samples are further stained with contrasting agent to increase the foreground-background separation (Di Mauro et al., 2011). This significantly simplifies the problem and allows application of basic image processing techniques. Several approaches are in fact implemented as macros or plugins within available image-processing suites, such as ImageJ software (Rasband, 2012). On the other hand, jellyfish polyps are typically observed in highly diverse environments that surpass the capabilities of basic image processing (Durden et al., 2016b). Over the last five years, significant advances have been made in the field of computer vision, particularly in object detection and recognition, as demonstrated on established large-scale benchmarks (Everingham

et al., 2010; Russakovsky et al., 2015). The state-of-the-art detection approaches are based either on fast hand-crafted features (Dollár et al., 2009), or, with increasing prevalence, on deep learning (LeCun et al., 2015; He et al., 2017). For example, Girshick et al. (2014); Girshick (2015) decompose images into regions of interest using an object-agnostic region proposal algorithm, and classify them using features extracted from a pre-trained convolutional neural network (CNN) (Krizhevsky et al., 2012). Ren et al. (2015); He et al. (2017) extended this approach with end-to-end learning, in which the region proposal algorithm is also trained inside the CNN framework. We leverage these modern computer vision approaches to enable semi-automatic polyp population census. In particular, we train a fast detector from Dollár et al. (2010) to generate potential regions of interest, and classify them using the approach aking to R-CNN Girshick et al. (2014). As such, our approach addresses the polyp appearance variations and is robust to the challenging background environment.

We make the following two contributions. Our primary contribution is a two-stage detection algorithm for automatic polyp count estimation (PoCo). The PoCo approach combines state-of-the art computer vision and machine learning methods, tailored to the specific domain of polyp counting. PoCo is evaluated on a set of human-annotated images (Section 3.2). To provide accurate ground truth for our algorithm evaluation, these images have been annotated by several annotators. Our second contribution is a quantitative evaluation of human annotator variation in counting from typical polyp images (Section 3.1). PoCo was applied to a one-year population dynamics analysis of moon jellyfish polyps from Hočevar et al. (2018). Based on our in-depth analysis of the annotation problem from perspective of count variance and PoCo failure cases, image acquisition guidelines are outlined to facilitate future use in population studies (Section 4.1). To the best of our knowledge, this is the first work that holistically addresses the problem of automated polyp count estimation and quantitatively exposes the problem of human annotation errors in this task.

2. Methods

The purpose of this research is development and evaluation of a computer-vision based automated polyp counting approach. For evaluation purposes we acquired a large annotated dataset of 3894 polyps from 7 sample images using manual annotation. This dataset has two purposes: evaluation of the algorithm and evaluation of human annotation quality.

2.1. Image acquisition

Underwater polyp images were obtained by scuba divers during a 3-year survey in the Port of Koper (northern Adriatic) where polyps were found attached to under-surfaces of oysters growing on port pillars (Malej et al., 2012). The entire under-surfaces of five selected shells were photographed once per month at depth ranging from 2 m to 6 m and were taken with two cameras of different quality (Nikon D2X with Nikkor AF-D micro 60 mm f/2.8 lens and Pentax Optio WG-1 compact camera). The images varied in resolution from 1180x863 to 4288x2848 pixels. The distance varied as well, since no distance rig was used. Nikon D2X shots were made using two external lights (Subtronic Pro270 flash and Nikon SB-800 AF Speedlight flash in Sealux CX-800 housing, both mounted to Sealux CT25 flash arms), while the Pentax Optio WG-1 shots were taken with a built-in flash only. An approximate metric calibration was performed for each shell at the start of the monitoring campaign. A photograph of each shell with a ruler placed next to it was taken to estimate the shell surface area, which was subsequently used to obtain the polyp densities from the obtained polyp counts. Further details about *in situ* study and the results of manual counts are given in Hočevar et al. (2018).

The acquired underwater images significantly vary in the jellyfish polyps appearance (Figure 1). This accentuates subjectivity and is expected to lead to a wide inter-annotator variation. These factors negatively affect the consistency of annotations, and consequently increase the uncertainty of the obtained results (Durden et al., 2016b).

Figure 1: Zoom-in on moon jellyfish polyps in different time of the year and from different angles. The images in the second row show strobilating (producing ephyrae i.e. baby medusae) polyps. Original photographs are available at (Mandeljc, 2017a).

2.2. PoCo: An automated polyp counter

In our automated polyp counter, we adopt a two-stage detection paradigm found in modern computer-vision based object detection approaches (Girshick et al., 2014; Girshick, 2015; Ren et al., 2015; He et al., 2017). In the first stage, the image is processed using an algorithm that produces a large number of regions that potentially contain objects of interest. In the second stage, each proposed region is encoded as a feature vector by a pre-trained convolutional neural network (CNN), and classified as an object of interest or discarded as background.

An overview of the PoCo algorithm is illustrated in Figure 2. As we are interested in a single class of objects (i.e., polyps), we avoid a generic region proposal algorithm in the first stage and use a fast Aggregate Channel Feature (ACF) detector (Dollár et al., 2014) instead. This approach densely scans the input image with a sliding window at multiple scales. For each window location and scale, it encodes the corresponding region using features that describe its color, gradient magnitude and gradient orientations (Dollár et al., 2009, 2010), and classifies it as a potential polyp or background using a fast AdaBoost cascade (Appel et al., 2013).

The goal of the first stage is to quickly identify regions that may contain a polyp. Therefore, the ACF detector is trained to return as many relevant regions as possible, even if many of them turn out to be false positives. Such operating regime with high recall at the cost of potentially lowered precision (see Section 2.3) is desirable for a region proposal generator, as it ensures that all relevant regions are passed to the second stage.

In the second stage, each region proposal is encoded using features, extracted from a pre-trained CNN. In particular, we use the AlexNet network (Krizhevsky et al., 2012) provided by Girshick et al. (2014), which was pre-trained on the ImageNet dataset (Deng et al., 2009) and fine-tuned on the Pascal VOC dataset (Everingham et al., 2010) for object detection. The input patch is forward-propagated through five convolutional and two fully-connected layers. A 4096-dimensional feature vector is obtained by collecting the response from the seventh, fully-connected (FC7), layer of the network. The obtained feature representation is classified as a polyp or background using a Support Vector Machine (SVM) classifier (Cortes and Vapnik, 1995). The paradigm of using generic features, obtained from a pre-trained CNN, and training a relatively simple task-specific classifier (Razavian et al., 2014) both simplifies and speeds up the training of our detector, and alleviates the over-fitting due to limited number of available training samples.

The obtained detections (bounding boxes) are further post-processed with an additional non-maxima suppression step, which removes highly-overlapping detections with low detection scores. As the last step, we compute the centroids of the filtered bounding boxes, and return them as the output of the algorithm.

Our prototype of the PoCo algorithm is implemented in a mixture of Matlab and C++ code, using open-source libraries. The ACF detector and non-maxima suppression step are implemented using a Computer Vision Matlab Toolbox from Piotr Dollar (Dollár, 2016). For the CNN implementation, we use the Caffe framework (Jia et al., 2014) with NVIDIA CUDA support, while the SVM implementation is provided by the LIBLINEAR library (Fan et al., 2008). The prototype implementation of the PoCo algorithm is publicly available at a GIT respository (Mandeljc, 2017b) and the project page [2].

2.3. Performance evaluation measures

The human and computer annotation performance were evaluated with respect to the ground truth annotations (Section 2.4). The consistency of manual annotation was primarily measured by the maximum *relative error* (Re) per image. The relative error is defined as the absolute difference between the number of manually annotated polyps (N) and the number of polyps in the ground truth set (GT), normalized with the number of polyps in the ground truth set.

$$Re = |GT - N| \,/\, GT \tag{1}$$

Additional performance measures are used to provide an in-depth evaluation of the automated approach.

[2] http://www.vicos.si/Projects/PoCo

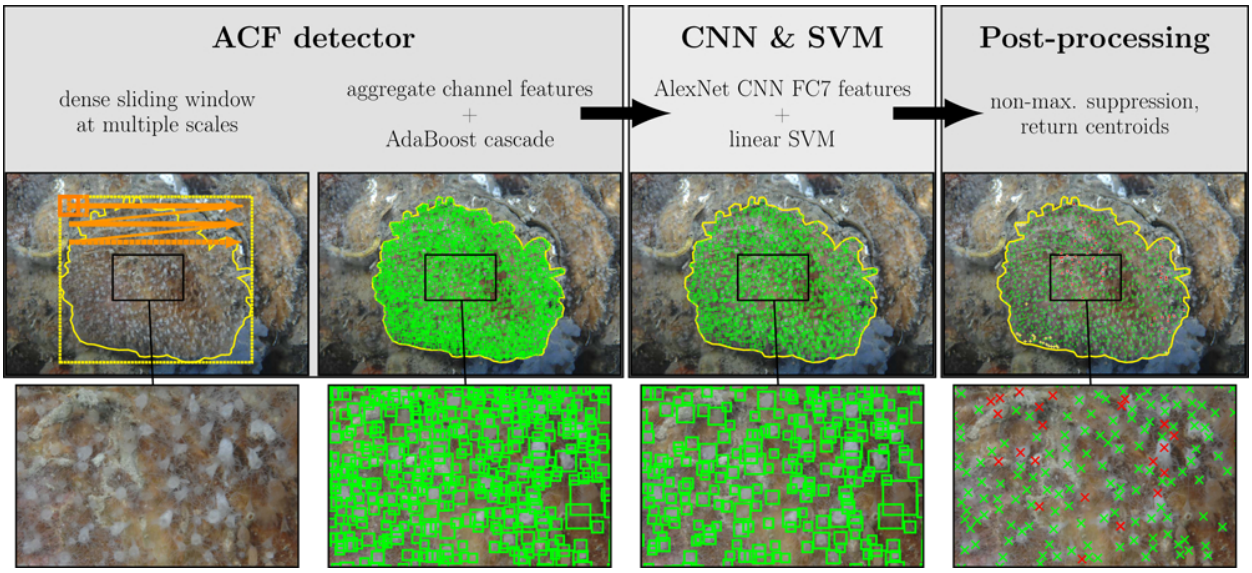


Figure 2: A schematic representation of the proposed PoCo algorithm, with main components and examples of their intermediate results (shown on image of shell #1 in March). In the first stage, the fast ACF detector performs dense sliding window scan at multiple scales, returning a large amount of image regions that potentially contain a polyp. These region proposals are verified in the second stage using CNN features and a linear SVM. The resulting detections are further processed with a non-maxima suppression to remove highly-overlapping detections. The output of the algorithm is a list of polyp detections, i.e., the center coordinates of detected bounding boxes. In the last image, false positives and false negatives are marked with red and yellow colour, respectively.

The count *ratio* is computed as the number of detected polyps, divided by the number of polyps in the ground truth set.

$$\text{Ratio} = N/GT \tag{2}$$

Two established measures from the detection literature (Fawcett, 2006) are used: *precision* (Pr) and *recall* (Rec):

$$Pr = \frac{TP}{TP + FP} \; ; \; Rec = \frac{TP}{TP + FN}, \tag{3}$$

where TP, FP, and FN stand for true positives, false positives, and false negatives, respectively. The true positives are detections that are included in the ground truth, the false positives are detections that are falsely detected (false hits), and false negatives are the ground truth data that were not detected (misses). *Precision* thus reflects the fraction of detections that have corresponding ground truth. *Recall*, on the other hand, reflects the fraction of the ground truth polyps that were captured by the detections.

In an automated detection system, it is usually possible to adjust the parameters (e.g., detection threshold) to increase the overall number of detections, thus improving recall (more true positives) at the cost of lowered precision (more false positives), or vice versa. A standard measure that combines precision and recall to summarize the overall system performance is the F_1-measure (Fawcett, 2006; Powers, 2011), which we will denote as F:

$$F = 2\frac{Pr \cdot Rec}{Pr + Rec}. \tag{4}$$

This measure places an equal importance on both precision and recall. For example, a system with perfect recall of 100 % and poor precision 50 % (or vice versa) will score $F = 67\,\%$, while a system with both moderate precision of 70 % and moderate recall of 70 % will rank higher, with $F = 70\,\%$.

2.4. A dataset for count variability analysis - Dataset A

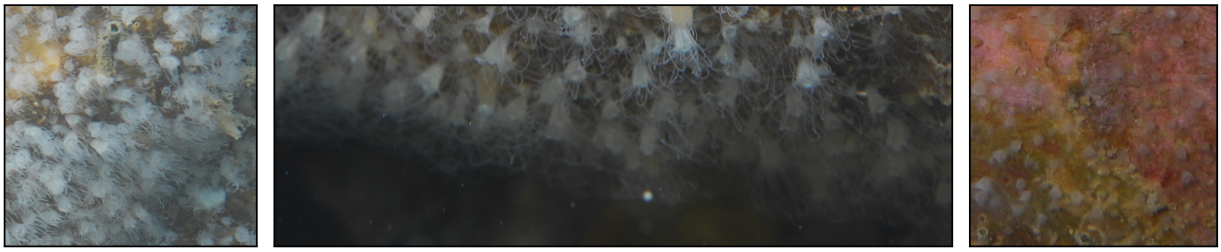
Seven diverse images were selected from a three year survey to reflect realistic variations in polyp appearance. We paid attention to select sharp images with clearly distinguishable polyps. All of them were taken using the Nikon D2X and additional lighting. To see the images from dataset, we refer the reader to the supplementary material (Figure S1) or our publicly available datasets (Mandeljc, 2017a).

In each image, a region of interest (RoI) was selected for counting. The selected region covers a large part of the shell with well-discernible polyps, while avoiding ambiguous out-of-focus regions and regions with significantly overlapping polyps (Figure 3). Another constraint was that the region had to be approximately flat, and that the line of view was perpendicular to the surface in order to mitigate the perspective projection distortion. Such setup reduces the errors in estimating the polyp density in terms of number of polyps per square centimeter, but of course also demands certain amount of work and time from the user. This task requires 10 to 30 seconds, depending on the complexity of the image. The regions of interest in our dataset contain roughly 350–770 polyps per image, which adds up to 3894 polyps. An example image with corresponding RoI polygon is shown in Figure 4.

Care was taken in obtaining the annotation ground truth, i.e., the position of each polyp within the RoI. We prepared a graphical user interface that allows marking and un-marking the polyp position in image coordinates. To resolve potential visual ambiguities, the interface allows zooming in and out of the image.

Several annotators with various levels of expertise annotated all images using our annotation tool. The annotators were: an expert diver that acquired the images, an experienced annotator, and a volunteer with limited experience. To increase the consistency across the annotators, each annotator was shown several examples of polyps. For intra-annotator consistency check, one of the annotators was required to annotate a selected image (image #5) several times.

Annotations from each image were consolidated. For a given image, the union of all annotations was created. Annotations with centers closer than half of polyp size were considered to belong to the same polyp and were averaged into a single center, while other were left unchanged. This ensured that polyps missed by some annotators, but captured by others, were identified and introduced into the ground



(a) (b) (c)

Figure 3: Examples of problematic areas that should be avoided when selecting the region of interest for automated detection: (a) surfaces not perpendicular to the line of sight, which results in polyp overlap and defocused areas, as well as incorrectly estimated surface area; (b) edges of the shell; (c) poorly lit areas with low contrast.

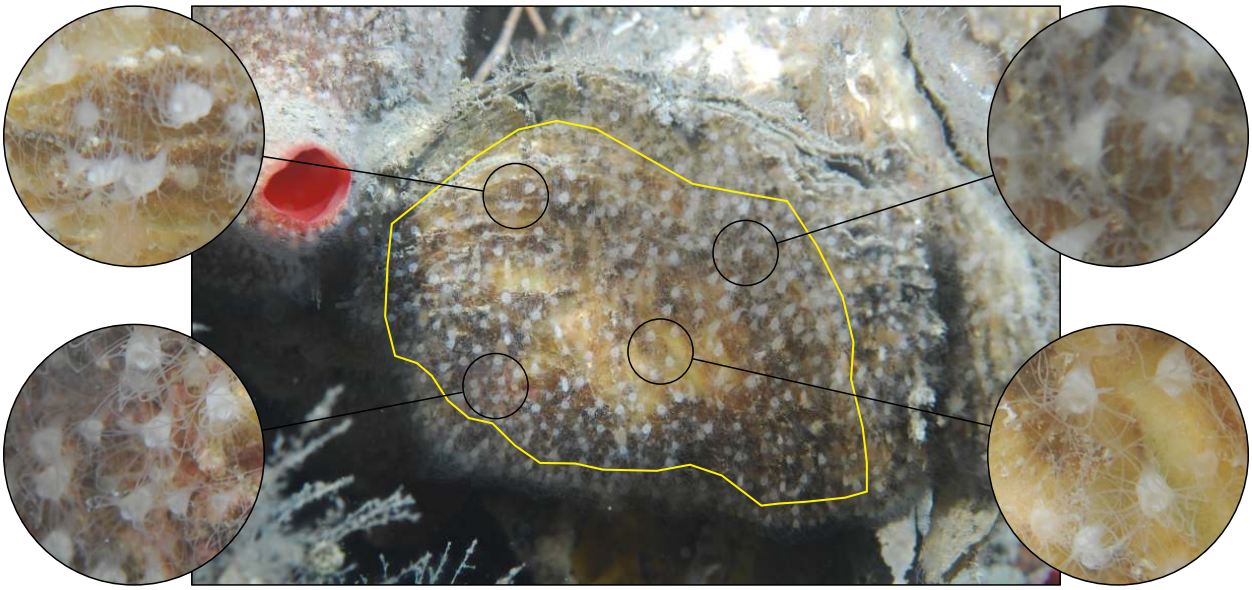


Figure 4: Image #1 from the polyp dataset. During the leave-one-out evaluation, our PoCo algorithm performed best on this image. Therefore, this image may serve as a good example for image acquisition guidelines. The polygon denoting the chosen area of interest is drawn in yellow. The edges of the shell, where polyps are out of focus and overlapping each other, are excluded. Note the variation in both polyp and background appearance in the zoomed-in areas of the image.

truth annotation. The merged annotations for all images were carefully reviewed by the authors in several passes and false annotations were removed from the union. This reviewed and consolidated ground truth annotation was used as a "gold standard" in our subsequent evaluation. Despite our significant care invested in constructing the ground truth, it turned out that a certain level of ambiguity remains and a few features in each image cannot be reliably identified.

2.5. Population dynamics analysis dataset - Dataset B

Reproducing of the population dynamics by using PoCo was analyzed on a larger collection of 48 images, containing roughly 40.000 polyps. The dataset contained images of four shells, taken over a period of one year (from February 2012 to January 2013) at monthly intervals. Most of the images were taken using the Nikon D2X, but a small number of them were taken using Pentax Optio WG-1 (in September and October). These images were annotated only once by Sara Hočevar (Hočevar et al., 2018), thus the ground truth of these counts is unknown due to annotation subjectivity. This dataset is not appropriate for a rigorous analysis of PoCo polyp detection quality, but it can be used to analyze polyp density dynamics estimation properties in comparison to the results obtained in Hočevar et al. (2018).

This time PoCo was trained on the ground truth from the Dataset A (7 images, containing 3894 polyps) and additional 11 imaged were used to accommodate for a larger variance in polyp appearance through the period of one year. The 11 images were taken during the period from February 2012 to January 2013, but were not used in the census analysis. To reduce the amount of annotation in the training phase, a region with approximately 100 polyps was annotated and added to the training set, thus enlarging it to a total of 4960 polyps.

A preliminary analysis showed that PoCo trained on the Dataset A performed well on majority of the 48 images in the Dataset B except on a small subset. Visual inspection revealed the images with poor performance considerably differed from the ones in the training set. In particular, the polyps appearance, the lighting, and the shell fouling varied over the year, and were not sufficiently represented in Dataset A. All these images corresponded to the fifth shell in Hočevar et al. (2018), located at the shady side of the pillar and close to the bottom resulting in poor lighting and decreased visibility which could be a consequence of resuspension of the sediment. The fifth shell was thus not considered in our experiment with the Dataset B. All images of the Dataset B including the fifth shell can be found in the on-line dataset repository (Mandeljc, 2017a).

3. Results

3.1. Performance analysis of human annotators

This experiment considered the Dataset A from Section 2.4. On average, each annotator required 17 min to annotate a single image. A couple of additional hours were spent reviewing the ground truth annotations. Even though significant effort was invested in the ground truth annotation, there were several unclear cases in which the presence of a polyp could not be decided. Even after multiple examinations we could not annotate all polyps with a 100 % accuracy.

Table 1 shows the counts obtained by three annotators on the seven test images from Dataset A. The counts significantly vary between the annotators, which is consistent across all images. The average maximum relative error per image is 18 % ($\sigma = 4$ %) and can be expected to be as large as nearly 25 % in some cases.

Intra-annotator consistency (Table 2) was evaluated by re-annotating image #5 four times by a single annotator. Since the same image was annotated, we can expect the annotator to become biased from *learning* the annotation. The learning might introduce correlation across the different annotation events. To reduce this correlation, consecutive annotations were performed at intervals at least 24 hours apart. Table 2 shows the results, in which the annotator is denoted as *Volunteer 1*. By comparing the annotations on Day 2 and Day 4, we see that the difference between two counts can be as large as 20 %. The same image was re-annotated by another volunteer (*Volunteer 2* in Table 2). The worst-case difference between the two volunteers amounts to over

Table 1: Manual counts in our dataset of seven test images.

Image	Expert diver	Expert annotator	Volunteer	Ground truth	Relative error (max.)
#1	358	378	397	455	17 %
#2	617	571	561	655	14 %
#3	455	453	462	543	17 %
#4	637	678	715	770	17 %
#5	622	676	744	723	14 %
#6	336	296	270	350	23 %
#7	384	304	323	398	24 %

40 %. Note that the percentage difference does not involve errors in the ground truth but solely between the two volunteers.

Table 2: Manual counts in the image #5. Four counts were performed by a volunteer (denoted by *Volunteer 1*) to evaluate the consistency of a single annotator. The ground truth was estimated at 723 polyps.

Image	Volunteer 1				Volunteer 2
	Day 1	Day 2	Day 3	Day 4	
#5	490	472	576	597	744

In addition to the count estimation accuracy, we have also investigated the accuracy of polyp detection across different annotators. The complete table with precision and recall values was too large to be placed in the paper, and is available in the supplementary material (Table S1). We noticed that human annotators attain high precision values, meaning that they typically do not annotate structures which are not polyps. On the other hand, the recall values are lower, meaning that the annotators miss some of the true polyps.

3.2. Performance analysis of the PoCo algorithm

3.2.1. Polyp detection accuracy analysis on Dataset A

Our polyp detection algorithm (PoCo) was evaluated on Dataset A (Section 2.4) by performing the leave-one-out evaluation, which ensures that polyps from the same images are not used in train and test set. In this setup, both stages of the PoCo algorithm (the ACF detector from the first, and SVM classifier from the second stage) were trained on six out of seven images and tested on the remaining one. The evaluation is repeated for all seven possible training/test splits. Results are shown in Table 3. PoCo delivers highly accurate estimates of the polyp counts. The relative error was 17 % ($\sigma = 9$ %), which is comparable to the average maximal human annotation error (18 %, $\sigma = 4$ %).

Note that when considering only the overall number of detected polyps, the failure in detection (i.e., false negatives) may be compensated with detection of false positives. Such failures are better reflected in evaluation with precision and recall. The average precision of PoCo is 96 % and average recall is 83 %. This means that, on average, 96 % of the PoCo detections match the ground truth, but 17 % of the polyps were not detected. This is inferior to human performance (Table S1 — average precision is almost 100 % and average recall is 87 %). On the other hand, PoCo is a deterministic algorithm and always delivers the same number of counts for the same image, meaning that the consistency is expected to outperform a human annotator (see our results in Table 2). We therefore find the precision/recall performance acceptable for most practical applications.

The consistency of PoCo is supported by the F-measure values in Table 3. The F-measure ranges from 85 % to 94 %, without significant deviations in any of the images. The average F-measure is 88 % with standard deviation of 3 %. We can conclude that in all cases, the learning sample of six images provided sufficient information for successful detection in the remaining image. The highest F-score is

Table 3: PoCo detection performance during the leave-one-out evaluation on our polyp dataset.

Image	Ground truth	Detection	Ratio	Relative error	Precision	Recall	F-score
#1	455	451	99 %	1 %	94 %	94 %	94 %
#2	655	494	75 %	25 %	100 %	75 %	86 %
#3	543	441	81 %	19 %	97 %	78 %	87 %
#4	770	571	74 %	26 %	100 %	74 %	85 %
#5	723	647	89 %	11 %	98 %	88 %	92 %
#6	350	274	78 %	22 %	99 %	78 %	87 %
#7	398	464	117 %	17 %	82 %	95 %	88 %

obtained for the image #1 (Figure 4), meaning that the visual properties were most convenient for polyp detection. Therefore, this image could serve as an example of a photograph that is suitable for computer-assisted annotation.

The average time required for automated detection with PoCo was less than 15 seconds per image: it took 11.3 s to process an input image with resolution 4288×2848 (and effective RoI size of 2722×2335) on a mid-range desktop workstation[3]. The time complexity is dominated by the ACF detector and the CNN feature extraction. In the previous example, the ACF detector took 4.5 s and produced 3000 region proposals. CNN feature extraction from these proposals took 6.4 s on a GPU (a CPU-only implementation takes longer, 162.2 s). Overall, the automatic polyp counting with PoCo is roughly 70 times faster than manual annotation, providing a vast improvement in the annotation speed and even more substantial reduction in *human annotation effort*. The quoted speed improvement accounts for the counting process only and does not include preparation and selection of RoI. The latter should be carefully chosen no matter whether doing the counting manually or by an automated algorithm.

3.2.2. Population dynamics analysis on Dataset B

This experiment considers polyp count dynamics reproduction of the results in Hočevar et al. (2018) by applying the PoCo algorithm on the Dataset B from Section 2.5. We provide a detailed comparison to the human annotations of this dataset in the supplementary material (Table S4), but focus here on a practical comparison of the end-results. The tests were performed on two types of RoI polygons. The first were encompassing the entire shell as in the original paper Hočevar et al. (2018), while the second polygons were smaller to avoid poorly focused areas with significant polyp overlaps.

3.2.2.1. Counts over the entire shell surface. Using the large RoI polygons, the best agreement was observed on shell #1 (Figure 5). The differences between manual and automated counts are within the interval of the human error, and the automated counts have captured the dynamics of polyp density variation. The automated and manual counts agreed very well on shell #2 in all months except April, September and October. A similarly good agreement was observed for shells #3 and #4 in April, May, and June. We observed a discrepancy in September and October images of shell #3, and in September, October and November images of shell #4. A closer inspection of images revealed that on shells #3 and #4 in September and October, the polyps appear to be almost transparent, which is likely due to unfavorable lighting conditions. These images were taken with the Pentax Optio WG-1 compact camera, using only the built-in flash. On the other hand, shell number #1 was located on the exposed side of the pillar, close to the surface, and was likely better lit. The inconsistent image resolution and camera quality likely influenced the results as well.

[3]Quad-core Intel i5-3570K CPU @ 3.40 GHz; 16 GB RAM; NVIDIA GeForce GTX 970 GPU; 64-bit Fedora 25 linux; Matlab R2016b

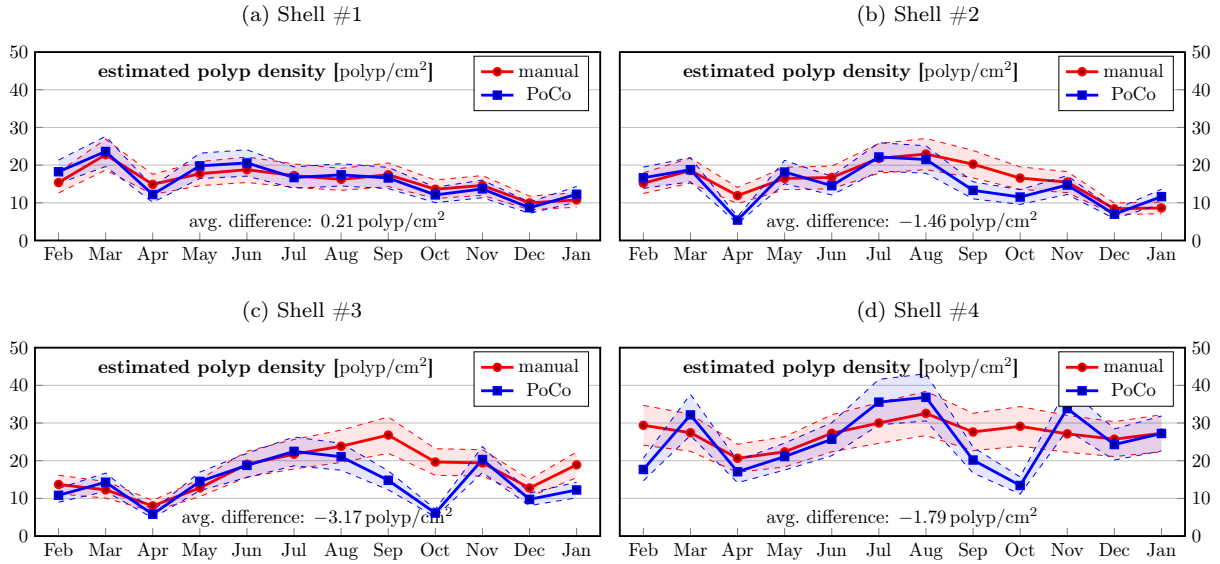


Figure 5: Estimated polyp density in year-long population dynamics analysis (Hočevar et al., 2018). Polyps were counted over the entire visible surface of each shell. The densities, estimated from manual counts, are shown in red, while densities, estimated from the automated PoCo counts, are shown in blue. The corresponding shaded areas denote the expected error. The plotted values are listed in Table S2.

3.2.2.2. Counts over a shell surface region. Performance on reduced RoI polygons is shown in Figure 6. We observe an improvement in agreement between the densities estimated from the automated and from the manual counts. The average difference between the estimated polyp density, obtained by PoCo, and the density, obtained by manual counts, went from $-1.55\,\text{polyps/cm}^2$ when counting over the entire surface of the shell to $-0.15\,\text{polyp/cm}^2$ when using the reduced RoI polygons.

By comparing the population dynamics graphs between the large (Figure 5) and small RoI polygons (Figure 6), we observe a sensitivity of the estimated polyp density to the RoI size. The RoI polygon should be selected carefully, and should be representative for the whole shell. The smaller number of polyps contained within the reduced RoI polygons inherently increases the standard error ($SE \propto n^{-1/2}$), which should also be taken into the account when evaluating the census results. For example, the sharp peaks in polyp density for shell #1, appearing at reduced polygons (Figure 6a) are probably unrealistic and are likely the result of the too small polygon areas. The changes in density are much smoother when using larger RoI polygons (Figure 5a), and the match between manual and automated counts is already satisfactory. On the other hand, the results for shells #3 and #4 (Figure 6c and Figure 6d) do not exhibit such behavior. In fact, the automated counts in these cases have improved with area size reduction, which means that the smaller RoI polygons for shells #3 and #4 were properly chosen.

4. Discussion

Our findings support previous published work on evaluation of manual annotators in other areas of marine biology. Culverhouse et al. (2003) evaluated the inter- and intra-consistency of human annotators on the task of visual identification of six species of marine dinoflagellates. They reported that trained

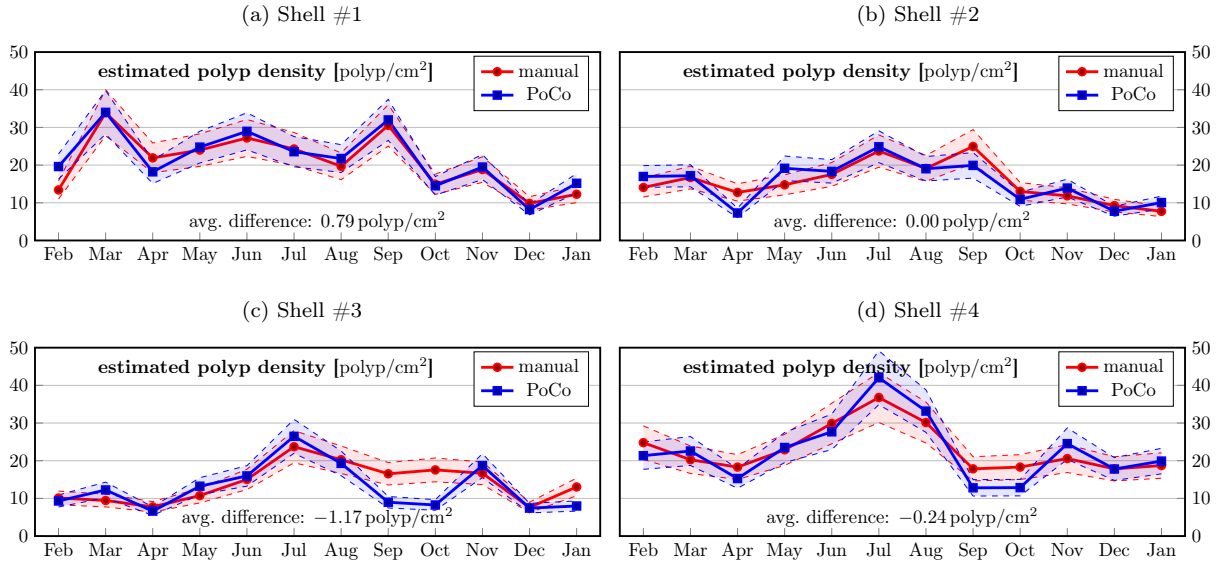


Figure 6: Estimated polyp density in year-long population dynamics analysis (Hočevar et al., 2018), obtained via manual and automatic counts inside revised (smaller) regions of interest (see text for details). The densities, estimated from manual counts, are shown in red, while densities, estimated from the automated PoCo counts, are shown in blue. The corresponding shaded areas denote the expected error. The plotted values are listed in Table S3.

personnel can be expected to achieve self-consistency of 67 % to 83 % and inter-expert consensus of 43 %, while experts that routinely engage in the task can achieve accuracies in range of 84 % to 95 %. In a more recent study, Culverhouse et al. (2014) evaluated consistency of both manual counting and visual identification of mesozooplankton. They found that self-consistency in counting ranged from 68 % to 99 %, with 20 % of analysts returning counts that varied by more than 10 %. Durden et al. (2016a) investigated the consistency in manual annotation of megafaunal morphotypes and reported a 78 % detection success. In another study, Kelly (2001) found that Trophic Diatom Index values, obtained from manual counts of benthic diatom specimens in 58 UK river samples, varied as much as 30 % between a trained "counter" and an "auditor". Schoening et al. (2012) evaluated consistency of human annotators on quantification of several megafauna taxa in deep-sea floor images. For "small white sea anemone", which is visually similar to the polyps, they reported average inter- and intra-observer agreements of 70 % and 79 %, respectively.

Manual counting is a highly time-consuming and arduous task. Consequently, an annotator usually processes images only once, and the variance or the annotation error is often neglected in the published research (Schoening et al., 2017). Our findings show that single-pass annotations in a typical annotation task result in errors that are far from negligible, and are expected to increase when considering even larger datasets. The surprisingly high differences in manual counts show that care has to be taken when comparing results obtained by different human annotators. The discrepancy comes from personal bias in annotation, as well as the variance in the annotation process. Furthermore, we have observed considerable variation in annotations made by the same annotator. These variations come from expertise in recognizing the objects of interest, as well as from reduced focus when annotating large, densely-populated areas. Our results indicate a non-negligible level of uncertainty involved in the studies based on manually-obtained

counts. Comparison across different sites likely further increases this variance. This calls for an automated method which, as imperfect or biased as it may be, can at least be expected to perform consistently. Our attempt at this is proposing the polyp counting algorithm (PoCo).

Even though PoCo annotation accuracy is slightly inferior to the manual annotation, it enables processing of much larger datasets and offers a much higher consistency. This should result in higher precision but only under the condition that the automated process is supervised by an expert who would identify cases where the algorithm went astray and would exclude such images or add additional learning samples to improve performance in these situations.

Given that the images in our datasets were captured without considering the requirements of automated counting and considering significant variations in image quality (two different cameras, different lighting, varying distance and water properties), the PoCo algorithm performed very well in polyp population dynamics experiment (Section 3.2.2). In fact, all cases with significant deviation between the automatic and the manual counts can be explained by inferior image quality, which is expected to influence not only the automated detection, but also the accuracy of manual census.

4.1. Guidelines for image acquisition

The image acquisition protocol should be adjusted to produce images with clearly visible and distinguishable polyps. The region of interest should be in-focus and well lit, ensuring a high level of contrast between polyps and the background. The water should be as clear as possible, and the divers should avoid sediment resuspension. The polyp appearance should be consistent over the images to increase the consistency with the training set. This could be achieved by consistent use of the same adequate artificial lighting, the same camera, and by taking images at a constant distance from the camera. Constructing a distance rig with a camera mount is highly advisable.

An exemplar image on which the PoCo performed very well is shown in Figure 4, while an example where the PoCo performed poorly is shown in Figure 7. It is obvious that the polyp visibility in Figure 7 much poorer than in Figure 4. In fact, the polyps appear to be nearly transparent, with poor contrast to the background. We believe this is due to poor visibility and insufficient lighting, as the visibility of polyps in images of the two other shells, taken on the same day, is significantly better.

In addition to paying attention to photograph quality, care has to be taken to capture photographs with polyps sufficiently well visible. Polyps have to occupy sufficient number of pixels to extract discriminative information to separate from the background. In our experiments, the smallest polyp size was approximately 5x5 pixels. Experiments have shown that the algorithm successfully detects polyp even in cluttered regions with many polyps touching. Nevertheless, in regions where polyps overlap to such an extent that the background is no longer visible, the detection quality is expected to drop.

We discourage counting polyps in the whole image. Care should be taken when selecting the region for polyp counting. On one hand, the region should be large enough to encompass a sufficiently large amount of polyps and thus reduce the standard error. On the other hand, the region should be limited to the area with sharp, clearly visible polyps that minimally overlap each other (Figure 3). The same guidelines should also be followed when acquiring images for manual annotation.

5. Conclusions

We addressed the problem of automated counting of a large number of small objects located on heterogeneous background. As a case study, we considered polyp counting, which is used in the jellyfish research. Understanding polyp population dynamics is vital for understanding the causes of jellyfish blooms and such counting tasks are usually performed manually. The work is tedious, time consuming and, as our experiments show, produces rather inconsistent and unreliable counts. According to our measurements, the manual annotation of an image containing between 300 and 800 polyps requires 17 minutes on average and the obtained counts can differ up to 25 % from the ground truth. Analysis of large datasets there-

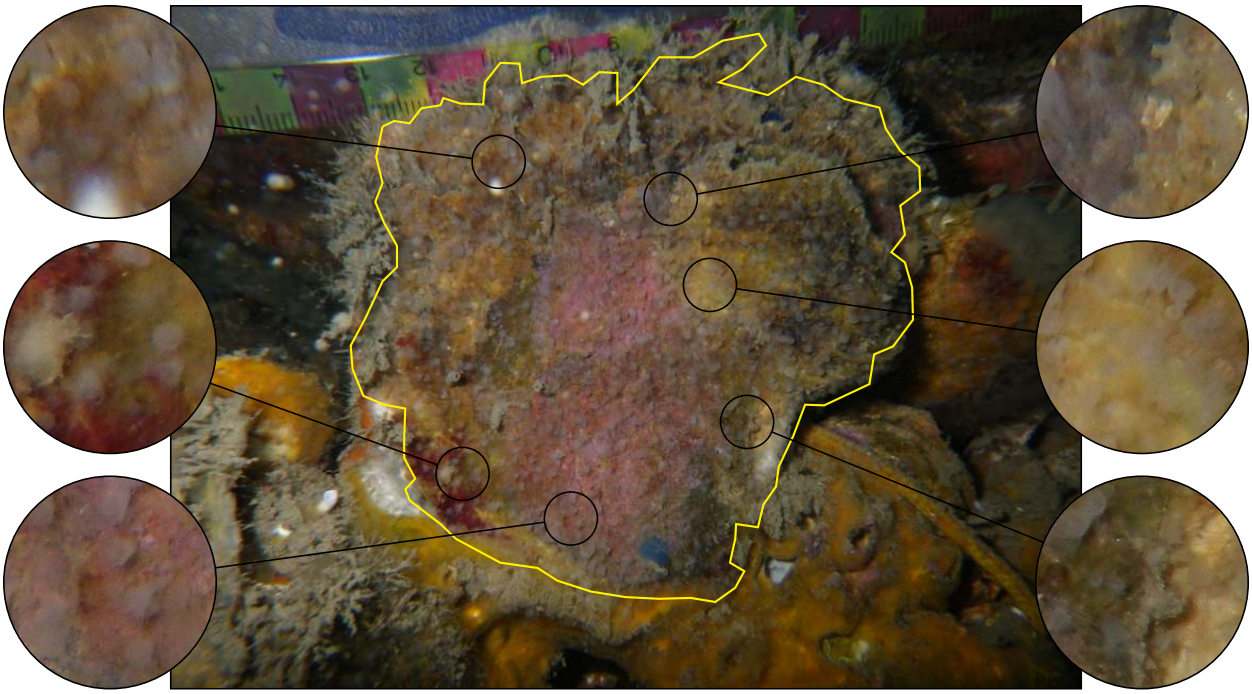


Figure 7: Example of an image where the algorithm performed poorly (shell #4 in October). As can be seen, the polyps are hardly visible, even in the zoomed-in areas of the image. The contrast is poor and the polyps appear to be transparent. As performance was better on other images shot during the same day, we attribute these issues to insufficient lighting.

fore requires significant amounts of manpower, and is often not feasible given the available resources.

We propose an automated polyp counting algorithm (PoCo) to address the annotation difficulties. PoCo applies state-of-the-art object detection and recognition approaches from computer vision and machine learning. Results on a carefully annotated dataset indicate that the PoCo performance in polyp counts is on par with the human annotation on images that meet reasonable acquisition quality standards. Although the precision and recall values are slightly lower compared to manual annotations, the reduction in annotation time is roughly 70-fold, enabling the annotator to process much larger datasets than previously possible. With the use of PoCo, the experiments could be designed with much larger datasets and with more frequent image acquisition. This should enable in situ research that was not feasible using manual annotation. Since PoCo can be trained on different kinds of samples, it is highly flexible and affords a straight-forward specialization to a wide range of counting problems reaching beyond moon jellyfish polyps.

PoCo displayed very good performance on a large dataset of 48 images containing approximately 40.000 polyps (Section 3.2.2). Note that these images were captured by two different cameras, at different distances, different water properties and different lighting conditions. The discrepancy from manual counts in majority of cases was well within margins of human error, which speaks of a considerable robustness to the input variability and of a wide applicability.

PoCo is not (yet) available as a part of a computer program or as a stand-alone application, but a prototype Matlab implementation of the PoCo algorithm is freely available (along with installation notes and sample cases) at GitHub (Mandeljc, 2017b). The code was successfully tested by a regular Linux/MATLAB user, but modification and retraining of the algorithm would require a computer-vision specialist. In our future work we plan to streamline the algorithm and library dependence to increase the practical accessibility. A potential user is strongly encouraged to contact the authors of this paper.

Performance of the automated algorithm decreases with degraded image quality, and in cases when the polyp appearances significantly differs from the training samples. Such cases could be naturally addressed by the user-in-the-loop regime. The automatic annotations would be visually verified, and potentially corrected by an experienced annotator. On one hand, this would ensure a high quality of the results. On the other hand, the additional manual annotations could be used to re-train PoCo, thus adapting it to the current setup. An experienced annotator would also easily recognize images that are inappropriate for automated detection, and exclude them from the dataset. Alternatively, the algorithm could attempt to automatically detect poor imaging conditions and notify the user. These will be the topics of our future work.

Software and data availability

Project name: PoCo
Project page: http://www.vicos.si/Projects/PoCo
PoCo v1.0 code developer: Rok Mandeljc (Visual Cognitive Systems Laboratory, University of Ljubljana, Faculty of Computer and information Science)
Code URL: https://github.com/rokm/polyp-detector
Software required: Linux operating system, Matlab 2015a or higher, Piotr's Computer Vision Toolbox, Caffe framework, LIBLINEAR library.
Dataset URL: https://vision.fe.uni-lj.si/~rokm/polyp-detector

Acknowledgments

The authors are grateful to Sara Hočevar, who kindly provided the annotated images from her survey, and to Tjaša Kogovšek, who led and supervised the fieldwork. We would also like to thank Kristijan Shirgoski for his preliminary work in polyp detection, and all the volunteers who manually annotated the images: Anja Fettich, Ana Križaj, Peter Smerkol, Mavricij Grbec, Jernej Uhan, and Benedikt Strajnar. We are thankful to the Slovenian Environment Agency (ARSO) for the use of their facilities. Funding: This work was supported by the EU

FP7 project PERSEUS and the Slovenian Research Agency (ARRS, P1-0237, P2-0214 and J2-8175).

References

Appel, R., Fuchs, T. J., Dollár, P., Perona, P., 2013. Quickly boosting decision trees-pruning underachieving features early. In: Proceedings of the 2013 International Conference on Machine Learning (ICML 2013). pp. 594–602.

Beijbom, O., Edmunds, P. J., Roelfsema, C., Smith, J., Kline, D. I., Neal, B. P., Dunlap, M. J., Moriarty, V., Fan, T.-Y., Tan, C.-J., et al., 2015. Towards automated annotation of benthic survey images: Variability of human experts and operational modes of automation. PloS one 10 (7), e0130312.

Boero, F., Bouillon, J., Gravili, C., Miglietta, M. P., Parsons, T., Piraino, S., 2008. Gelatinous plankton: irregularities rule the world (sometimes). Marine Ecology Progress Series 356, 299–310.

Brotz, L., Cheung, W. W., Kleisner, K., Pakhomov, E., Pauly, D., 2012. Increasing jellyfish populations: trends in large marine ecosystems. Hydrobiologia 690 (1), 3–20.

Brotz, L., Pauly, D., 2016. Studying jellyfish fisheries: toward accurate national catch reports and appropriate methods for stock assessments. In: Gian Luigi Mariottini (editor). Jellyfish: Ecology, Distribution Patterns and Human Interactions. Nova Publishers, Hauppauge, NY.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20 (3), 273–297.

Culverhouse, P. F., Macleod, N., Williams, R., Benfield, M. C., Lopes, R. M., Picheral, M., 2014. An empirical assessment of the consistency of taxonomic identifications. Marine Biology Research 10 (1), 73–84.

Culverhouse, P. F., Williams, R., Reguera, B., Herry, V., González-Gil, S., 2003. Do experts make mistakes? a comparison of human and machine indentification of dinoflagellates. Marine Ecology Progress Series 247, 17–25.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009). pp. 248–255.

Di Mauro, R., Cepeda, G., Capitanio, F., Viñas, M., 2011. Using zooimage automated system for the estimation of biovolume of copepods from the northern argentine sea. Journal of sea research 66 (2), 69–75.

Dollár, P., 2016. Piotr's Computer Vision Matlab Toolbox (PMT). https://github.com/pdollar/toolbox.

Dollár, P., Appel, R., Belongie, S., Perona, P., 2014. Fast feature pyramids for object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (8), 1532–1545.

Dollár, P., Belongie, S., Perona, P., 2010. The fastest pedestrian detector in the west. In: Proceedings of the 2010 British Machine Vision Conference (BMCV 2010). BMVA Press, pp. 68.1–68.11.

Dollár, P., Tu, Z., Perona, P., Belongie, S., 2009. Integral channel features. In: Proceedings of the 2009 British Machine Vision Conference (BMCV 2009). BMVA Press, pp. 91.1–91.11.

Durden, J. M., Bett, B. J., Schoening, T., Morris, K. J., Nattkemper, T. W., Ruhl, H. A., 2016a. Comparison of image annotation data generated by multiple investigators for benthic ecology. Marine Ecology Progress Series 552, 61–70.

Durden, J. M., Schoening, T., Althaus, F., Friedman, A., Garcia, R., Glover, A. G., Greinert, J., Stout, N. J., Jones, D. O., Jordt, A., et al., 2016b. Perspectives in visual imaging for marine biology and ecology: from acquisition to understanding. In: Oceanography and Marine Biology. CRC Press, pp. 9–80.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. International Journal of Computer Vision 88 (2), 303–338.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J., 2008. Liblinear: A library for large linear classification. Journal of Machine Learning Research 9 (Aug), 1871–1874.

Fawcett, T., Jun. 2006. An introduction to roc analysis. Pattern Recogn. Lett. 27 (8), 861–874.

Fox, C., Morten, C., Boswarva, K., 2016. Are marinas leading to jellyfish blooms? 5th International Jellyfish Bloom Symposium, Barcelona, Spain.

Friedland, K., Ama-Abasi, D., Manning, M., Clarke, L., Kligys, G., Chambers, R., 2005. Automated egg counting and sizing from scanned images: rapid sample processing and large data volumes for fecundity estimates. Journal of Sea Research 54 (4), 307–316.

Gershwin, L.-A., 2013. Stung!: On jellyfish blooms and the future of the ocean. University of Chicago Press.

Girshick, R., 2015. Fast r-cnn. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV 2015). pp. 1440–1448.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014). pp. 580–587.

Graham, W. M., Gelcich, S., Robinson, K. L., Duarte, C. M., Brotz, L., Purcell, J. E., Madin, L. P., Mianzan, H., Sutherland, K. R., Uye, S.-i., et al., 2014. Linking human well-being and jellyfish: ecosystem services, impacts, and societal responses. Frontiers in Ecology and the Environment 12 (9), 515–523.

He, K., Gkioxari, G., Dollár, P., Girshick, R., Oct 2017. Mask r-cnn. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2980–2988.

Hočevar, S., Malej, A., Boldin, B., Purcell, J. E., 2018. Seasonal fluctuations in population dynamics of *Aurelia aurita* polyps in situ with a modelling perspective. Marine Ecology Progress Series 591, 155–166.

Janßen, H., Augustin, C., Hinrichsen, H.-H., Kube, S., 2013. Impact of secondary hard substrate on the distribution and abundance of *Aurelia aurita* in the western Baltic Sea. Marine pollution bulletin 75 (1), 224–234.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093.

Kelly, M. G., 2001. Use of similarity measures for quality control of benthic diatom samples. Water Research 35 (11), 2784–2788.

Kogovsek, T., Vodopivec, M., Raicich, F., Uye, S.-i., Malej, A., 2018. Comparative analysis of the ecosystems in the northern adriatic sea and the inland sea of japan: Can anthropogenic pressures disclose jellyfish outbreaks? Science of the Total Environment 626, 982–994.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C. J. C., Bottou, L., Weinberger, K. Q. (Eds.), Advances in Neural Information Processing Systems 25. Curran Associates, Inc., pp. 1097–1105.

Lebrato, M., Molinero, J.-C., Cartes, J. E., Lloris, D., Mélin, F., Beni-Casadella, L., 2013. Sinking jelly-carbon unveils potential environmental variability along a continental margin. PloS one 8 (12), e82070.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Lucas, C. H., Dawson, M. N., 2014. What are jellyfishes and thaliaceans and why do they bloom? In: Jellyfish blooms. Springer, pp. 9–44.

Lucas, C. H., Gelcich, S., Uye, S.-I., 2014. Living with jellyfish: Management and adaptation strategies. In: Jellyfish Blooms. Springer, pp. 129–150.

Lucas, C. H., Graham, W. M., Widmer, C., 2012. Jellyfish Life Histories: Role of Polyps in Forming and Maintaining Scyphomedusa Populations. In: Michael L (ed) Advances in marine biology 63, 133–196.

Makabe, R., Furukawa, R., Takao, M., Uye, S.-i., 2014. Marine artificial structures as amplifiers of Aurelia aurita sl blooms: a case study of a newly installed floating pier. Journal of Oceanography 70 (5), 447–455.

Malej, A., Kogovšek, T., Ramšak, A., Catenacci, L., 2012. Blooms and population dynamics of moon jellyfish in the northern Adriatic. Cahiers de biologie marine 53 (3), 337–342.

Mandeljc, R., 2017a. PoCo: polyp detector/counter — datasets. https://vision.fe.uni-lj.si/~rokm/polyp-detector.

Mandeljc, R., 2017b. PoCo: polyp detector/counter — prototype & evaluation framework. https://github.com/rokm/polyp-detector.

Olariaga, A., Guallart, E. F., Fuentes, V., López-Sanz, À., Canepa, A., Movilla, J., Bosch, M., Calvo, E., Pelejero, C., 2014. Polyp flats, a new system for experimenting with jellyfish polyps, with insights into the effects of ocean acidification. Limnology and Oceanography: Methods 12 (4), 212–222.

Powers, D. M., 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning 2 (1), 37–63.

Purcell, J. E., 2012. Jellyfish and ctenophore blooms coincide with human proliferations and environmental perturbations. Annual Review of Marine Science 4, 209–235.

Purcell, J. E., Hoover, R. A., Schwarck, N. T., 2009. Interannual variation of strobilation by the scyphozoan aurelia labiata in relation to polyp density, temperature, salinity, and light conditions in situ. Marine Ecology Progress Series 375, 139–149.

Rasband, W., 2012. Imagej: Image processing and analysis in java. Astrophysics Source Code Library 1, 06013.

Razavian, A. S., Azizpour, H., Sullivan, J., Carlsson, S., 2014. Cnn features off-the-shelf: An astounding baseline for recognition. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CPRW 2014). pp. 512–519.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 28. Curran Associates, Inc., pp. 91–99.

Richardson, A. J., Bakun, A., Hays, G. C., Gibbons, M. J., 2009. The jellyfish joyride: causes, consequences and management responses to a more gelatinous future. Trends in ecology & evolution 24 (6), 312–322.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., Fei-Fei, L., 2015. Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115 (3), 211–252.

Schiariti, A., Morandini, A. C., Jarms, G., von Glehn Paes, R., Franke, S., Mianzan, H., 2014. Asexual reproduction strategies and blooming potential in Scyphozoa. Marine Ecology Progress Series 510, 241–253.

Schoening, T., Bergmann, M., Ontrup, J., Taylor, J., Dannheim, J., Gutt, J., Purser, A., Nattkemper, T. W., 2012. Semi-automated image analysis for the assessment of megafaunal densities at the arctic deep-sea observatory hausgarten. PloS one 7 (6), e38179.

Schoening, T., Durden, J., Preuss, I., Albu, A. B., Purser, A., De Smet, B., Dominguez-Carrió, C., Yesson, C., de Jonge, D., Lindsay, D., et al., 2017. Report on the marine imaging workshop 2017. Research Ideas and Outcomes 3, e13820.

Tinta, T., Kogovšek, T., Malej, A., Turk, V., 2012. Jellyfish modulate bacterial dynamic and community structure. PLoS One 7 (6), e39274.

Tinta, T., Kogovšek, T., Turk, V., Shiganova, T. A., Mikaelyan, A. S., Malej, A., 2016. Microbial transformation of jellyfish organic matter affects the nitrogen cycle in the marine water column-A Black Sea case study. Journal of Experimental Marine Biology and Ecology 475, 19–30.

Tinta, T., Malej, A., Kos, M., Turk, V., 2010. Degradation of the Adriatic medusa Aurelia sp. by ambient bacteria. Hydrobiologia 645 (1), 179–191.

Vodopivec, M., Álvaro J Peliz, Malej, A., 2017. Offshore marine constructions as propagators of moon jellyfish dispersal. Environmental Research Letters 12 (8), 084003.

Willcox, S., Moltschaniwskyj, N. A., Crawford, C. M., 2008. Population dynamics of natural colonies of Aurelia sp. scyphistomae in Tasmania, Australia. Marine Biology 154 (4), 661–670.

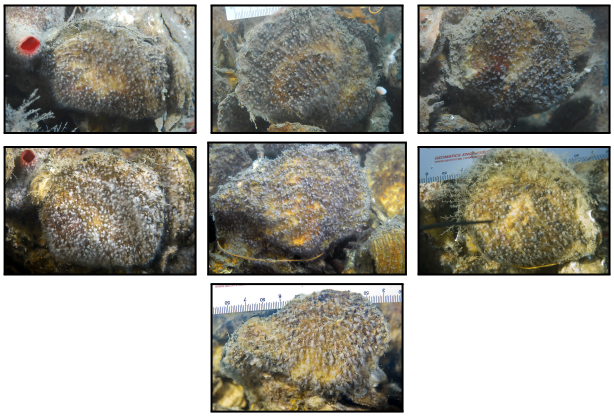


Figure S1: The 7 images from our Dataset A, used to evaluate the consistency of human annotators and used in leave-one-out experiment with PoCo algorithm.

.

Supplementary material

S1. Images from Dataset A

In Figure S1, we provide thumbnails of the seven images from Dataset A (Section 2.4), which were used to evaluate the consistency of human annotators and to perform the leave-one-out evaluation of the PoCo algorithm. The original 4288×2848 images are, along with ROI polygons and bounding box annotations, available as a part of our datasets package (Mandeljc, 2017a).

S2. Detailed evaluation of manual annotation

In Table S1, we present full evaluation of manual annotators, using the same performance measures and the same ground truth (gold standard) as in the evaluation of the PoCo algorithm (Table 3). It can be seen that the precision of manual annotators is very close to 100 %, meaning that the annotators do not mark features that are not polyps. As expected, the recall is lower (87 % on average), meaning that the annotators usually miss some polyps in the image, which results in lower count of polyps than the ground truth (ratio less than 100 %).

S3. Detailed results of application to population dynamics analysis

In the second part of experimental section (Section 3.2.2), we compared the results of PoCo detection to human annotations indirectly, via estimated polyp density on whole shell (Figure 5) and reduced regions of interest (Figure 6). Here we present the values of polyp density that were used for plotting both figures. Table S2 lists the number of polyps per square centimeter obtained by manual and automatic counts over the whole surface of the shell, while Table S3 lists the values obtained on revised (smaller) regions of interest.

In Table S4, we directly compare PoCo detections to human annotations, using the same performance measures as in Table 3 and Table S1: the ratio between number of PoCo detections and number of manual annotations, precision, recall, and F-score. It should be noted, however, that the manual annotations were obtained in a single-pass annotation pro-

Table S1: Full evaluation of manual annotations.

Image	Gr. truth	Annotator	Count	Ratio	Precision	Recall	F-score
#1	455	Expert diver	358	78.68 %	100.00 %	78.68 %	88.07 %
		Expert	378	83.08 %	100.00 %	83.08 %	90.76 %
		Volunteer	397	87.25 %	100.00 %	87.25 %	93.19 %
#2	655	Expert diver	617	94.20 %	99.68 %	93.89 %	96.70 %
		Expert	571	87.18 %	99.82 %	87.02 %	92.99 %
		Volunteer	571	87.18 %	99.82 %	87.02 %	92.99 %
#3	543	Expert diver	455	83.79 %	100.00 %	83.79 %	91.18 %
		Expert	453	83.43 %	100.00 %	83.43 %	90.96 %
		Volunteer	462	85.08 %	99.78 %	84.90 %	91.74 %
#4	770	Expert diver	637	82.73 %	100.00 %	82.73 %	90.55 %
		Expert	678	88.05 %	100.00 %	88.05 %	93.65 %
		Volunteer	715	92.86 %	100.00 %	92.86 %	96.30 %
#5	723	Expert diver	622	86.03 %	100.00 %	86.03 %	92.49 %
		Expert	676	93.50 %	98.67 %	92.25 %	95.35 %
		Volunteer	744	102.90 %	95.30 %	98.06 %	96.66 %
#6	350	Expert diver	336	96.00 %	99.11 %	95.14 %	97.08 %
		Expert	296	84.57 %	100.00 %	84.57 %	91.64 %
		Volunteer	270	77.14 %	99.63 %	76.86 %	86.77 %
#7	398	Expert diver	384	96.48 %	97.92 %	94.47 %	96.16 %
		Expert	304	76.38 %	100.00 %	76.38 %	86.61 %
		Volunteer	323	81.16 %	100.00 %	81.16 %	89.60 %

Table S2: Polyp densities, estimated from manual counts and PoCo counts over the the entire visible surface of the shell (Dataset B). The listed values correspond to plots in Figure 5.

		Estimated polyp density [polyp/cm^2]											
Shell	**Method**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**
#1	Manual	15.4	22.8	14.9	17.7	18.8	17.2	16.2	17.4	13.6	14.6	10.0	10.8
	PoCo	18.2	23.6	12.2	19.8	20.6	16.7	17.4	16.6	12.1	13.7	8.7	12.3
#2	Manual	15.2	18.6	11.9	16.5	16.7	21.8	22.9	20.2	16.6	15.5	8.4	8.6
	PoCo	16.7	18.8	5.4	18.2	14.6	22.2	21.5	13.3	11.5	14.7	7.0	11.6
#3	Manual	13.7	12.3	8.0	12.8	19.1	21.7	23.8	26.8	19.7	19.4	12.8	18.9
	PoCo	10.8	14.3	5.8	14.5	18.8	22.5	21.1	14.8	6.1	20.3	9.7	12.2
#4	Manual	29.4	27.4	20.6	22.4	27.3	30.0	32.5	27.6	29.1	27.1	25.7	27.3
	PoCo	17.7	32.1	17.1	21.1	25.7	35.5	36.8	20.2	13.4	33.9	24.3	27.2

Table S3: Polyp densities, estimated from manual counts and PoCo counts inside revised (smaller) regions of interest. The listed values correspond to plots in Figure 6.

		Estimated polyp density [polyp/cm^2]											
Shell	**Method**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**
#1	Manual	13.4	33.9	21.9	24.0	27.2	24.3	19.7	30.5	14.9	18.8	9.8	12.3
	PoCo	19.6	34.0	18.2	24.8	29.0	23.6	21.7	32.0	14.5	19.5	8.2	15.2
#2	Manual	14.1	16.7	12.7	14.8	17.5	23.7	19.1	24.9	13.0	11.9	9.2	7.7
	PoCo	17.0	17.2	7.3	19.2	18.3	24.9	19.0	19.9	10.9	13.9	7.8	10.1
#3	Manual	10.1	9.4	7.7	10.7	15.0	23.7	20.3	16.5	17.6	16.7	7.6	13.0
	PoCo	9.3	12.2	6.6	13.2	16.0	26.5	19.3	9.0	8.2	18.7	7.4	8.0
#4	Manual	24.8	20.3	18.3	22.9	29.9	36.8	30.1	17.8	18.3	20.6	17.8	18.7
	PoCo	21.3	22.6	15.3	23.5	27.7	42.0	33.2	12.8	12.9	24.5	17.8	19.9

cess, and as such do not represent the ground truth. Nevertheless, the evaluation still offers an informative insight into the relative performance of the PoCo algorithm compared to a human annotator.

Table S4: Evaluation of PoCo detections on the year-long population observation dataset from (Hočevar et al., 2018), evaluated with respect to single-pass manual annotations.

Shell	Month	Manual annotations	PoCo detections	Detection evaluation			
				Ratio	Precision	Recall	F-score
#1	1	202	250	123.76 %	80.00 %	99.01 %	88.50 %
	2	221	323	146.15 %	68.11 %	99.55 %	80.88 %
	3	558	560	100.36 %	97.50 %	97.85 %	97.67 %
	4	361	300	83.10 %	98.67 %	81.99 %	89.56 %
	5	395	408	103.29 %	96.57 %	99.75 %	98.13 %
	6	448	477	106.47 %	93.92 %	100.00 %	96.86 %
	7	400	388	97.00 %	99.74 %	96.75 %	98.22 %
	8	324	358	110.49 %	87.99 %	97.22 %	92.38 %
	9	503	527	104.77 %	83.68 %	87.67 %	85.63 %
	10	246	239	97.15 %	96.65 %	93.90 %	95.26 %
	11	310	321	103.55 %	93.46 %	96.77 %	95.09 %
	12	162	135	83.33 %	97.78 %	81.48 %	88.89 %
#2	1	199	259	130.15 %	72.97 %	94.97 %	82.53 %
	2	362	437	120.72 %	82.15 %	99.17 %	89.86 %
	3	430	442	102.79 %	96.15 %	98.84 %	97.48 %
	4	328	188	57.32 %	98.40 %	56.40 %	71.71 %
	5	380	493	129.74 %	77.08 %	100.00 %	87.06 %
	6	451	472	104.66 %	91.10 %	95.34 %	93.17 %
	7	611	640	104.75 %	95.31 %	99.84 %	97.52 %
	8	492	490	99.59 %	100.00 %	99.59 %	99.80 %
	9	642	513	79.91 %	92.20 %	73.68 %	81.90 %
	10	335	281	83.88 %	96.09 %	80.60 %	87.66 %
	11	305	358	117.38 %	84.92 %	99.67 %	91.70 %
	12	238	200	84.03 %	90.50 %	76.05 %	82.65 %
#3	1	392	240	61.22 %	99.58 %	60.97 %	75.63 %
	2	304	280	92.11 %	89.64 %	82.57 %	85.96 %
	3	284	368	129.58 %	76.09 %	98.59 %	85.89 %
	4	233	199	85.41 %	84.42 %	72.10 %	77.78 %
	5	323	398	123.22 %	77.89 %	95.98 %	85.99 %
	6	452	480	106.19 %	93.33 %	99.12 %	96.14 %
	7	713	796	111.64 %	89.45 %	99.86 %	94.37 %
	8	609	580	95.24 %	99.48 %	94.75 %	97.06 %
	9	497	270	54.33 %	100.00 %	54.33 %	70.40 %
	10	528	248	46.97 %	99.60 %	46.78 %	63.66 %
	11	501	562	112.18 %	87.90 %	98.60 %	92.94 %
	12	229	223	97.38 %	92.83 %	90.39 %	91.59 %
#4	1	364	387	106.32 %	90.96 %	96.70 %	93.74 %
	2	482	415	86.10 %	100.00 %	86.10 %	92.53 %
	3	394	439	111.42 %	87.93 %	97.97 %	92.68 %
	4	356	297	83.43 %	99.33 %	82.87 %	90.35 %
	5	445	457	102.70 %	92.78 %	95.28 %	94.01 %
	6	581	538	92.60 %	99.81 %	92.43 %	95.98 %
	7	715	817	114.27 %	87.39 %	99.86 %	93.21 %
	8	586	645	110.07 %	90.23 %	99.32 %	94.56 %
	9	347	249	71.76 %	97.99 %	70.32 %	81.88 %
	10	356	250	70.22 %	97.60 %	68.54 %	80.53 %
	11	400	477	119.25 %	83.65 %	99.75 %	90.99 %
	12	347	346	99.71 %	86.71 %	86.46 %	86.58 %