Citation for published version

Bafti, Saber Mirzaee and Ang, Chee Siang and Hossain, Md. Moinul and Marcelli, Gianluca and Alemany-Fornes, Marc and Tsaousis, Anastasios D. (2021) A crowdsourcing semi-automatic image segmentation platform for cell biology. A crowdsourcing semi-automatic image segmentation platform for cell biology . ISSN 0010-4825. (In press)

DOI

https://doi.org/10.1016/j.compbiomed.2020.104204

Link to record in KAR

https://kar.kent.ac.uk/85323/

Document Version

Author's Accepted Manuscript

ISSN 0010-4825

Computers in Biology and Medicine

An International Journal

A crowdsourcing semi-automatic image segmentation platform for cell biology

Saber Mirzaee Bafti, Chee Siang Ang, Md. Moinul Hossain, Gianluca Marcelli, Marc Alemany-Fornes, Anastasios D. Tsaousis

PII: S0010-4825(20)30535-7

DOI: https://doi.org/10.1016/j.compbiomed.2020.104204

Reference: CBM 104204

To appear in: *Computers in Biology and Medicine*

Received Date: 29 September 2020

Revised Date: 24 December 2020

Accepted Date: 26 December 2020

CRediT authorship contribution statement

Saber Mirzaee Bafti: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original Draft, Visualization, Data Curation. **Chee Siang Ang**: Writing - Review & Editing, Supervision, Resources. **Md. Moinul Hossain**: Supervision, Writing - Review & Editing. **Gianluca Marcelli***:* Supervision, Writing - Review & Editing, Validation. **Marc Alemany-Fornes**: Investigation. **Anastasios D. Tsaousis:** Supervision, Data Curation, Writing - Review & Editing.

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

A crowdsourcing semi-automatic image segmentation platform for cell biology

Saber Mirzaee Bafti[a,*], Chee Siang Ang[a], Md. Moinul Hossain[a], Gianluca Marcelli[a], Marc Alemany-Fornes [b], Anastasios D. Tsaousis [b]

[a]*School of Engineering and Digital Arts, University of Kent, Canterbury, CT2 7NZ, UK*
[b] *Laboratory of Molecular & Evolutionary Parasitology, RAPID group, School of Biosciences, University of Kent, Canterbury, CT2 7NJ, UK*

Abstract

State-of-the-art computer-vision algorithms rely on big and accurately annotated data, which are expensive, laborious and time-consuming to generate. This task is even more challenging when it comes to microbiological images, because they require specialized expertise for accurate annotation. Previous studies show that crowdsourcing and assistive-annotation tools are two potential solutions to address this challenge. In this work, we have developed a web-based platform to enable crowdsourcing annotation of image data; the platform is powered by a semi-automated assistive tool to support non-expert annotators to improve the annotation efficiency. The behavior of annotators with and without the assistive tool is analyzed, using biological images of different complexity. More specifically, non-experts have been asked to use the platform to annotate microbiological images of gut parasites, which are compared with annotations by experts. A quantitative evaluation is carried out on the results, confirming that the assistive tools can noticeably decrease the non-expert annotation's cost (time, click, interaction, etc.) while preserving or even improving the annotation's quality. The annotation quality of non-experts has been investigated using IOU (intersection of union), precision and recall; based on this analysis we propose some ideas on how to better design similar crowdsourcing and assistive platforms.
Our platform is available at https://object-detection-a5d76.web.app/home.

Keywords: Semi-auto segmentation; Object detection; Computational biology; Crowdsourcing; Image annotation; Instance segmentation

[*] Corresponding author. Tel.:+44-772-170-4334.
E-mail address: sm2121@kent.ac.uk

1. Introduction

Accurate computerized object detection and segmentation are becoming important in healthcare. For example they have been successfully used for detection of anatomical and cellular structures, as well as diagnosis and prognosis of diseases [1]. In some studies, object detection and segmentation have been utilized for oral disease screening (such as thrush, leukoplakia, lichenplanus, etc. [2,3]), for disease diagnosis on X-ray images and for cell detection on microscopic images [4–6]. Despite the wide use of object detection tools for the identification of diseases, their application in cell biology (e.g. identification of microbes) is still quite rare. In addition, most of the current state-of-the-art object detection algorithms are based on deep neural networks [7–11], the performance of which is highly correlated with the volume of data and the quality of annotations, which can be laborious, time-consuming and expensive to generate. For everyday objects, numerous annotated datasets such as Cityscapes [12] or COCO [13] are now publicly available. However, for specialized domains such as microbiological images, the availability of adequate and accurately annotated data is very limited. Furthermore, the requirement of specialized knowledge for microbiological images is a challenge that makes their annotation process more difficult than the annotation of everyday objects. Some general approaches can be used to overcome the challenges of the annotation process: i) *crowdsourcing* the annotation process and ii) providing *assistive tools* to the annotators [14].

Crowdsourcing is used to reduce costs by outsourcing a task to a group of experts or to a group of non-experts, who can be given online training [15]. Crowdsourcing has drawn the attention of computer vision researchers, in fact, studies [16,17] in this field have explored the effectiveness of outsourcing of image classification and instance segmentation on public datasets such as Pascal VOC, LabelMe and KITTI [18–20]. Recent studies on crowdsourcing have shown promising results on biomedical images. For example, [21] applied crowdsourcing techniques for the detection of dividing cells in breast cancer histology images, while [22] used a crowdsourcing framework for lung nodule detection and annotation to aid radiologists in lung cancer diagnosis.

To achieve an easy and faster annotation process, it is essential to design an efficient annotation user interface and assistive tools, which can also maintain the motivation of the annotators and the quality of their annotations. Polygon operator is widely used for instance segmentation [18] while the use of assistive tools in conjunction with polygon operator to support annotators, e.g. to correct drawn polygons or to propose new polygons [23–25], is still an area of development.

Given that crowdsourcing frameworks and assistive tools have been used mainly in isolation, in this study we propose a novel web-based image annotation platform combining crowdsourcing and assistive segmentation tools to support non-experts in annotating microbiological images of gut parasites. We show that our assistive tools enable non-expert annotators to perform their task accurately and more quickly. We also investigate the behavior of non-expert annotators under different levels of image complexity (high and low object density) of microscopic images. Finally, we use our analysis to propose design directions for the development of state-of-the-art annotation platforms.

2. Related works

Given the importance of high-quality image annotations to train machine learning algorithms, research has looked into the design of annotation platforms to reduce the annotation cost (i.e. time, clicks, etc.) and improve its quality, e.g. by designing intelligent user interfaces which can assist human annotators to perform the task. In the following subsections, we present the key studies relating to i) annotation tools in crowdsourcing of medical or biological images, ii) assistive user interfaces and iii) annotators' behavior analysis.

2.1. Crowdsourcing medical image annotations

Following the success in images of everyday objects [16,17], crowdsourcing has been increasingly adopted for medical image annotation by both experts and non-experts. However, the lack of crowds' expertise for such specialized images is still the biggest challenge [26]. [21] investigated the performance of a novel aggregation technique (AggNet) for classification of mitosis in breast histology images based on non-expert crowds' votes. The AggNet network is trained with gold standard images (images annotated by pathologists) for classification (mitosis or not mitosis), along with an aggregation layer that has been trained to generate a ground truth from the non-expert votes. They showed how an aggregation through a CNN network can help to overcome the challenge of noisy data collected from non-experts. [27] has also used crowds' votes (i.e. from knowledge workers) for classification of abnormal fundus images of the rear of eyes. Furthermore, [28] reported the performance of a group of non-experts in annotating Malaria infected RBCs' (Red Blood Cell) images throughout a crowdsourcing game. The authors show that the public contribution in detecting the positive samples of infected RBCs through a game can lead to up to 99% accuracy compared to the experts' detection. Along with outsourcing annotations for classification problems, studies have also explored the performance of the crowd in images segmentation. For instance, [29] introduced a web-based platform for hip segmentation in MR (Magnetic Resonance) images by non-expert annotators. Similarly, Heim and O'Neil explored the performance of non-expert annotators in CT (Computer Tomography) images segmentations, aggregated with majority voting technique [30, 31]. Collectively, these studies have demonstrated promising results of outsourcing medical-images annotation tasks to the public.

2.2. Assistive user interfaces

Introducing user-friendly interfaces and assistive tools in annotation platforms is an important research direction to make the annotation process simple and engaging, hence resulting in a higher completion rate and fewer errors. For instance, [18] presented a well-known platform for image segmentation (using a polygon operator for drawing the object's outline) called LabelMe, in which polygon operators were used. Polygon operators are the most common technique for instance segmentation [25,32,33] and they are well established, therefore, most of the efforts of recent studies have been put on developing assistive approaches. Regarding assistive tools, [28] introduced an automated classification approach that generates a preliminary classification on unlabeled images to be confirmed by a non-expert crowd through a computer game. Similarly, VATIC (Video Annotation Tool from Irvine, California) and iVAT (interactive Video Annotation) are two annotation platforms with rectangular and polygon operators for bounding box and instance segmentation, respectively, where for each frame of the input video, a supervised object detection algorithm generates the preliminary annotations that need to be confirmed/modified by annotators [23,25]. In a different approach, [34] have developed a recurrent neural network that iteratively proposes segmented objects to human annotators and refines the annotations with regard to their previous modifications. [35] presented a semi-automated platform that works based on edge detection, where high quality detected instances are proposed to annotators. It is worth mentioning that other studies have looked into novel tools based on different user interactions mechanisms, e.g. the use of eye-tracking for pixel-wise probability estimation of presence of an object [36].

2.3. Human annotator behavior analysis

The behavioral patterns of human annotators have been explored in different studies [37–39], although there are only a few studies that correlated the user's behavior pattern with the quality of their annotations. These are often done by capturing and analyzing user's video recordings, clickstreams, and mouse/tap dynamics, e.g. velocity and acceleration of mouse motion, time spent on clicks, etc. [30]. [40] is one of the few studies that correlated the mouse dynamics and clicks stream data with annotation quality in crowdsourced image segmentation. In that study, a regression model was trained to estimate the quality of annotations with respect to the features extracted from the clicks stream, i.e. velocity, acceleration, zoom, time, single and double clicks, contour correction, and mouse travelling distance. Similarly, [41] investigated the correlation between human annotators' effort and their performance in the annotation task, as measured by IOU (Intersection of Union), where the annotator's effort is quantified by three metrics: segmentation time, number of points and average time per point. [31] crowdsourced the task of CT lung scans annotation and investigated the correlation between users' behavior (time spent) and the quality of the annotation, and found that there is not a strong correlation between annotations' quality and annotation time or quantities such as number of regions and number of polygon vertices.

All the aforementioned studies have been conducted to facilitate the annotation process while monitoring the crowdsourced annotation quality. In our study, we aim to address three main gaps of the existing literature: i) exploring the performance in instance segmentation of non-expert annotators in the domain of cell biology; ii) studying the performance of the same non-expert annotators when they are aided by assistive tools and iii) studying annotator's behavior to glean insights to inform the design of future platforms.

3. Methodology

The aim of this study is to develop and evaluate a cost-efficient, user-friendly, and publicly available platform for instance segmentation, as well as to explore annotators' behavioral patterns. Our platform enables us to outsource the task among a group of non-expert annotators with no knowledge in the relevant cell biology domain. A polygon operator is implemented to allow annotators to draw the boundary of the objects of interest. To support the annotators in the drawing and labeling process, we have implemented a non-iterative mask proposal network that performs a preliminary detection on the input images. Preliminary detections are followed by user verification/modification steps on the computer predictions. The mask proposal network is trained with images that have been accurately annotated by an expert. The following subsections explain how the architecture of the entire platform and the different interconnected layers have been developed, how the mask proposal network is trained, and how the images have been collected, sorted and used in the study. Finally, the annotation subsection explains the procedure of image annotation by non-experts.

3.1. Platform architecture

Our platform relies on different technologies and contains three main blocks: i) the user interface, written in Typescript/HTML and deployed as a web-app, ii) the user assistive model, written in python and deployed on a python server, which is connected to the front-end through a Django gateway (shown as blue block in Fig. 1), iii) the database, which is used to store images, annotations and users' information.

In the design process of the user interface, effort has been put to make it as user-friendly as possible to ease the work of the annotators (as illustrated in section 3.3).

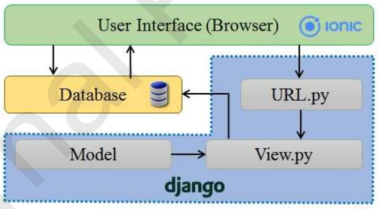


Fig. 1. Overview of the interconnection of the platform's layers

The developed platform is powered by an assistive tool to support annotators during the annotation process. The core of the assistive tool is based on the MRCNN (Mask Regional-Convolutional Neural Network, a state-of-the-art object detection, proposed by [7]) algorithm that needs to be trained (see section 3.2 for detailed information). The images and annotations are stored in a database which is directly called by the front-end (web-browser). Fig. 1 shows the workflow of the platform and the interconnections between different layers. The block, *Model*, reported in Fig. 1, represents the mask proposal network that is responsible for generating proposed polygons. The block is triggered by an Http request from the front-end layer (web-browser). The block, *View.py*, represents the auxiliary functions for refining/converting proposal masks and outputting them as polygons; the *View.py* block also stores results in the database.

3.2. Mask proposal network

In this work, we have implemented a *one-shot* mask proposal network based on the Weakly Supervised Object Localization (WSOL) technique [42], which is trained before use. Our approach is different from studies such as

[34,43], which utilized a recurrent neural network algorithm for auto-annotation that iteratively update and propose new masks. The WSOL technique has been applied (e.g. in [44]) for object detection with weakly annotated data or a subset of the entire data in many cases. In our study, instead, we have utilized a WSOL network only as a mask proposal network. The backbone of the proposed platform, which is a cutting-edge object detection algorithm (i.e. MRCNN), is trained with 20% of the total images (annotated by an expert). To facilitate the annotation of the remaining images, the weakly trained model generates proposal masks to help the non-experts. Proposed masks, which are initially generated in binary format, are converted into a tuple of polygon points using the RDP (Ramer-Douglas-Peucker) algorithm [45]. The proposed masks are provided to non-expert annotators who have the option to accept, reject or modify them. Fig. 2, shows an overview of the workflow of the assistive mask proposal network.

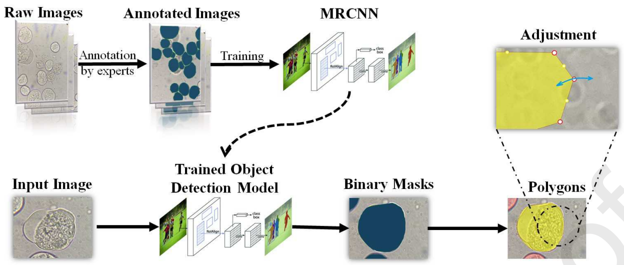


Fig. 2. The workflow of the assistive mask proposal network. The supervised object detection algorithm (MRCNN), trained with expert annotated data (gold standard), performs a preliminary detection on newly coming data and proposes masks which are accepted/modified by the annotator.

3.3. Collection, sorting and use of images

The dataset used in this study consists of bright-field microscopic images from three groups of microbial parasites, which requires domain-specific knowledge for annotation. In total, 150 microscopic images from three different groups of microbial parasites, *Entamoeba*, *Giardia* and *Prototheca*, were collected (50 images in each group). These three parasites were chosen specifically due to their distinct visual characteristics: shape, color, size, and texture (see appendix A for more information). In addition, these parasites are maintained axenically in culture (no other organism is present), avoiding any interference with the imaging process. All images were captured by an iPhone 8 smartphone, attached on top of a VWR IT 404 Inverted microscope's ocular lens (magnification of 40X) with a resolution of 4032 (H) × 3024 (V) pixels. All collected images have been directly uploaded and annotated by a postgraduate student biologist (expert), and verified by a senior academic biologist. The annotated images are then used as ground truth (GT) for training the model and testing the annotators' performances. Fig. 3, shows examples of annotated images from each group of parasites.

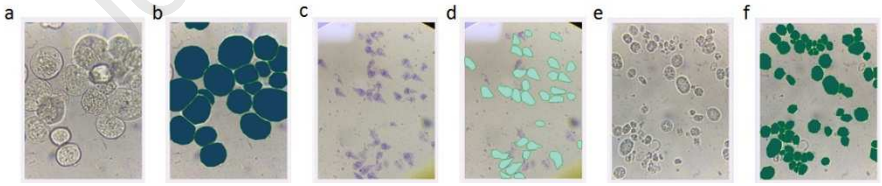


Fig. 3. Sample images of the training dataset (annotated by biologist); (a) raw *Entamoeba* image, (b) annotated *Entamoeba* image, (c) raw *Giardia* image, (d) annotated *Giardia* image, (e) raw *Prototheca* image, (f) annotated *Prototheca* image

In object detection, it is generally accepted that images which contain dense objects (*"Crowded"* images) are cognitively more demanding for human annotators than *"Non-crowded"* images. There is no a commonly accepted definition of *"Crowded"* and *"Non-crowded"* images, although in some studies (e.g. [13]) images with more than 10 objects are considered as crowded, while in some other sources (e.g. [46]) images with more than one object are considered crowded. In our study, we sorted the images in ascending order according to the number of objects in them. The first half of the images were considered *non-crowded* while the second half was considered *crowded* (see Appendix A with histograms of the number of objects in the images). Note that the platform is a crowdsourcing platform, and in some literature the annotators might be called *"Crowd"*. So, to avoid any confusion, we call the *crowded* and *non-crowded* images as *HD* (high density) and *LD* (low density) images, respectively. Fig. 4 shows examples of *HD* and *LD* images.

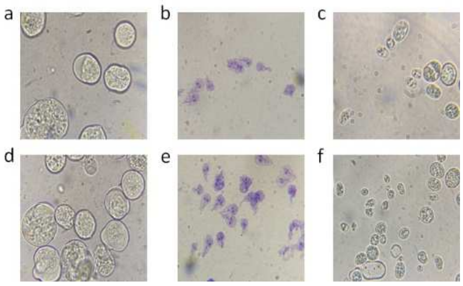


Fig. 4. Raw images for each group of parasites; (a) *LD Entamoeba,* (b) *LD Giardia,* (c) *LD Prototheca*, (d) *HD Entamoeba*, (e) *HD Giardia,* (f) *HD Prototheca.*

To train the mask proposal network, 20% of the total images (i.e. 10 images from each group of parasites) has been used, and the rest has been used by non-expert annotators to test the platform. Specifically, 20 *HD* images and 20 *LD* images for each parasite were used by the annotators to test the platform. Fig. 5 shows how the images were used in the workflow for training and testing the platform.

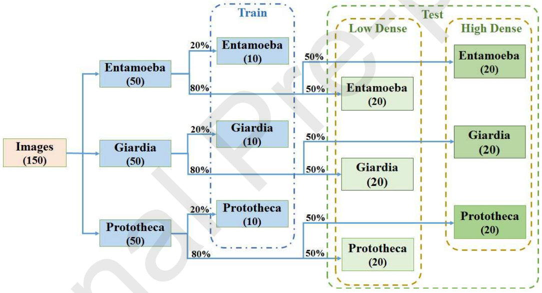


Fig. 5. Use of images in the workflow for training and testing the platform.

Fig. 6 shows the annotation interface of the platform. The annotation tools and options (previous/next image buttons, classes' buttons, etc.) are placed on the left of the interface, and the annotation environment is on the right. In Fig. 6, two parasites (blue polygons) are drawn and accepted, while one drawn polygon, in yellow, is selected for revision.

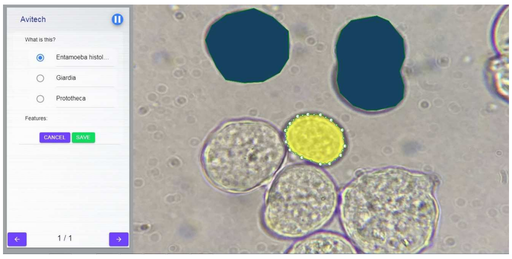


Fig. 6. A screenshot of the annotation interface (*Entamoeba* cells). (Colorful)

3.4. Train the proposed assistive Mask Proposal Network

The proposed assistive mask proposal network is trained with 10 images (i.e., 20%) for each parasite where the training *Entamoeba* images contain 149 objects and the *Giardia* and *Prototheca* images contain 135 and 665 objects, respectively. The purpose of this training is to generate proposal masks for annotators by the weakly trained model (see section 3.2). The model is trained with the following hyper parameters: learning rate = 0.0001, step per epoch = 2000, epoch =10, ROIS (region of interest) per image = 200, and image size = 1024 (h) ×1024 (v). Along with the training dataset, a sequential horizontal flipping, vertical flipping, horizontal and vertical rescaling, and ±90º rotating augmenter have been applied on all images to increase the volume of training dataset and model's generalization. The backbone of the MRCNN model is set based on Resnet101. The trained model and the core of the mask proposal network are then deployed on a python server (See sections 3.1).

3.5. Annotation procedure

Four non-expert annotators were recruited to take part in this study. The annotators were from different geographic locations and they all have been screened to make sure no one has a background in biology. The annotators agreed to take part in this study by signing the voluntary consent form. The annotation process starts with the tutorial and assessment steps, which are followed by the actual annotation task as shown in Fig. 7. In this section, the annotator's tutorial and assessment, and the annotation task are discussed.

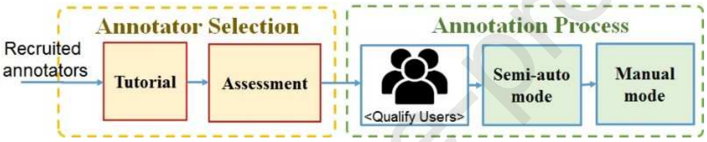


Fig. 7. Overview of user selection and annotation process

Annotator tutorial and assessment. In order to increase the annotation quality and user's understanding of the task, a short tutorial has been created to train the annotators. The tutorial contains written instructions that explain the process of annotation, followed by a short video that presents the annotation tools. In the last step of the tutorial the platform interface shows the annotators the three annotated images (one from each group of parasites), in which the objects of interest are identified with polygons. Afterwards, the annotators undergo an assessment step, in which they have to annotate a small set of images. Annotators who reached a mAP (mean average precision) higher than 80% can then proceed to the annotation task.

Annotation task. Four trained annotators start the annotation process right after they have successfully passed the assessment. We have created two different modes, "*manual*" (without assistive tool) and "*semi-auto*" (with assistive tool) in our platform and the four annotators were added to both modes. Images were imported in both modes and equally distributed among the annotators; each annotator was given 5 *HD* and 5 *LD* images per parasite (*Entamoeba*, *Giardia*, and *Prototheca*, respectively), i.e. 6×5=30 images in total. To avoid biased results due to *learning* effect and annotator's fatigue, the annotators have been asked to first complete the *semi-auto* task and the day after to complete the *manual* task. They had to use a laptop or a desktop, with a mouse for annotation and sit behind a desk. The annotators could remove and redraw the proposed masks in the *semi-auto* task if they thought it was necessary. The annotation task's results are reported and analyzed in the next section.

4. Results

In this section, the performance of non-expert annotators in both *manual* and *semi-auto* modes is analyzed. Specifically, this section presents the analyses of the annotators' performance in terms of time, clicks and annotation quality. The annotators' ability to distinguish between true and false parasites has been measured as accuracy and recall, where their effort has been quantified by three metrics i) *Tp*, true positive, ii) *Fp*, the number of falsely identified objects, and iii) *Fn*, the number of missed (un-identified) objects by annotators. The annotators' performance in terms of parasites' border delineation has been measured with IOU (intersection of union), since it is the most common segmentation evaluation metric [10,14,18,34,43,47–51]. In the following subsection time, clicks, and annotation quality are discussed in detail.

4.1. Time analysis

Time is an important factor in the annotation process which can affect the annotator's motivation and performance. In this study, we measure the time-cost as defined by the amount of time that annotators have spent on *manual* or *semi-auto* mode, respectively. Specifically, we define as *gross-time* the total time spent by the annotators to complete their task, from turning on the interface to the end of the task (i.e. including image loading time, time to choose the different tools in the interface, time to move from one image to the next, drawing parasites, etc.). The annotators were asked to measure the *gross-time* manually by themselves and report it to the researchers. Furthermore, for more accurate, standardized, and detailed information, we define as *net-time* the time spent just for annotation, which was measured automatically by the platform (i.e. time spent to draw polygons around objects plus the time to modify polygons, which are indicated as *Drawing-time* and *Modifying-time*, respectively). Finally, we define as *observation-time* the difference between *gross-time* and *net-time* that represent the time spent to observe images, choosing tools, moving images, etc. Fig. 8 shows the *gross-time* spent by four annotators on the three groups of parasites. Fig. 8 reports also the *observation-time* and the *net-time*.

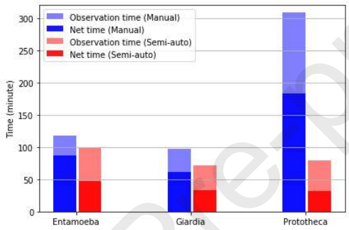


Fig. 8. *Gross-time* for each group of parasites, calculated as the sum of the *gross-times* (*net-time* + *observation-time*) of each annotator. Blue bars refer to *manual* mode, red bars refer to *semi-auto* mode. Light color (blue and red) represents the *observation-time*, while the dark color represents *net-time*. (Colorful)

As Fig. 8 shows, for the first two parasite groups (*Entamoeba* and *Giardia*) the *gross-time* in the *semi-auto* mode is 16% and 25% lower than the *manual* mode respectively; the gross-time for the *Prototheca* is 74.4% lower in the *semi-auto* mode. In comparison with the other two groups of parasites, *Prototheca* shows a much larger reduction in *gross-time*. From Fig. 8 a consistent trend emerges: the *gross-time* in *semi-auto* mode is shorter than in the *manual* mode's one. Importantly, Fig. 8 shows that in the *manual* mode, most of the time is spent on drawing and modifying polygons (i.e. *net-time*), while in the *semi-auto* mode, most of the time is spent to observe the images (i.e. observation time). This is because the annotators spent more time studying the polygons proposed by the mask proposal network to decide if they are real parasites and if they need to correct any mistakes (see appendix B for more detailed information).

Fig. 9 reports the mean *net-time* for annotation of a single object (i.e. a parasite cell) over all four annotators (for each parasite group, and for *HD* and *LD* images, respectively). In order to calculate the mean *net-time* reported in Fig. 9, we calculated firstly the mean *net-time* per image, by each annotator:

$$net\_time_{j,m} = \frac{1}{N_{j,m}} \sum_{i=1}^{N_{j,m}} Drawing\_time_{i,j,m} + Modification\_time_{i,j,m} \qquad (1)$$

Where i is the index for the object in image j, and m represents the index for the annotator. $N_{j,m}$ is the number of objects (parasites) within image j, which have been identified by annotator m. Therefore, the mean *net-time* of an object (for each parasite group, and for *HD* and *LD* images, respectively) reported in Fig. 9 is calculated according to Eq. (2):

$$mean\_net\_time = \frac{1}{N}\sum_{m=1}^{w}\sum_{j=1}^{v} net\_time_{j,m} \qquad (2)$$

Where the image-index, *j*, goes from 1 to *v*, i.e. the number of images given to each annotator (*v*=5), and the annotator-index, *m*, goes from 1 to *w*, i.e. the number of annotators (*w*=4). In Eq. (2), *N* is the total number of images annotated by four annotators in each group (in this case, N= 4×5=20). See Appendix B for more information.

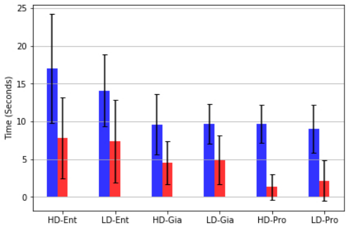


Fig. 9. Mean *net-time* for each group and for high-dense and low-dense images. Blue bars for manual mode, red bars for semi-auto mode. Error bars represent the standard deviation calculated over $net - time_{j,m}$. (Colorful)

To evaluate the significance of the mean *net-time* on groups, a statistical Wilcoxon test has been carried out on the mean net-times. According to the test, the mean-net time in *semi-auto* mode is significantly shorter than *manual* mode (P < .001). Fig. 9 and the Wilcoxon test confirm the trend from Fig 8, where the *net-time* in the *semi-auto* mode is shorter than the *net-time* in the *manual* mode. In the case of *Prototheca* (both *HD* and *LD*), the *semi-auto* mode's *net-time* is noticeably smaller than the *manual* mode's *net-time* (87.31% smaller for *HD* and 78.44% smaller for *LD*, respectively). Looking at the results for *Prototheca*, the densest group of parasites (see Fig. A.1), the comparison of mean *net-time* between *HD* and *LD* images in the *manual* and *semi-auto* modes shows that the *net-time* reduction from *manual* to *semi-auto* mode in the *HD* images is more pronounced than in the *LD* images. We believe this could be because the annotators became more fatigued and less motivated with the *HD* images. Therefore when they annotated *HD* images in the *semi-auto* mode, they tended to trust the proposed polygons by machine more often. To explore the impact of this over-trusting of the proposed mask on quality and other aspects of the annotation process, we have carried out click and quality analyses in following sections.

4.2. Clicks Analysis

Clicks are also another factor that can affect the annotation cost, annotator's motivation, and thus the annotation quality. In this study, further quantitative analysis is carried out by computing the number of clicks in the annotation task; we define as *Drawing-clicks* the number of clicks required by the annotator to draw a new polygon around an object (in both *manual* and *semi-auto* modes), and we define as *Modifying-clicks* the number of clicks required for correcting machine-proposed polygons (only in *semi-auto* mode) or user-drawn polygons (in both *manual* and *semi-auto* modes). Fig. 10 shows a consistent trend in that the total number of clicks in the *semi-auto* mode is considerably smaller than the clicks in *manual* mode; this is the case in particular for *Prototheca* images (both *HD* and *LD*). With respect to this finding, and given that the *Prototheca* is the densest group of images in comparison with the two other groups (See appendix B), we believe that annotators were less motivated when they annotated high dense images, therefore in the semi-auto they tended to do less clicks, and trust the proposed polygons by machine.

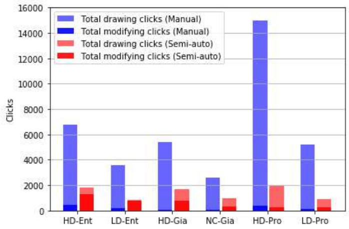


Fig. 10. Number of clicks for each group of images, calculated as the sum of the drawing and modifying clicks of each annotator. Blue bars refer to *manual* mode and red bars refers to the *semi-auto* mode. Light colors (blue and red) represent drawing-clicks while dark colors represent modifying-clicks. (Colorful)

Fig. 11 reports the mean number of clicks for each object, calculated over all the objects identified by all four annotators (for each parasite group and for *HD* images and *LD* images, respectively). In order to calculate the mean number of clicks, reported in Fig. 11, we calculated first the mean clicks per image, by each annotator:

$$num\_clicks_{j,m} = \frac{1}{L_{j,m}} \sum_{i=1}^{L_{j,m}} Drawing\_clicks_{i,j,m} + Modification\_clicks_{i,j,m} \tag{3}$$

Where *i* is the index for the object in image *j*, and *m* represents the index for the annotator. $L_{j,m}$ is the number of objects (parasites) within image *j*, which have been identified by annotator *m*. Therefore, the mean number of clicks (for each group and for high-dense and low-dense images, respectively) reported in Fig. 11 is calculated according to Eq. (4):

$$Mean\_num\_clicks = \frac{1}{L} \sum_{m=1}^{w} \sum_{j=1}^{v} num\_clicks_{\ j,m} \tag{4}$$

Where the image-index *j*, goes from 1 to *v*, i.e. the number of images given to each annotator (*v*=5), and the annotator-index, *m*, goes from 1 to *w*, i.e. the number of annotators (*w*=4). Here, *L* is the total number of images annotated by the four annotators in each group (in this case, N= 4×5=20). See also appendix B.

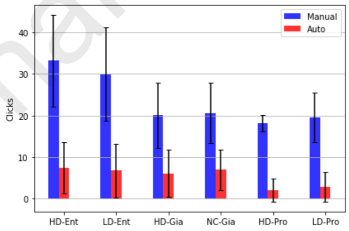


Fig. 11. Mean number of clicks per object, for each group and for *HD* and *LD* images. Blue bars for *manual* mode, red bars for *semi-auto* mode. Error-bars represent the standard deviation calculated over $num\_clicks_{j,m}$. (Colorful)

Fig. 11 shows that the number of clicks in *semi-auto* mode is smaller than in the *manuals'* one, especially for the case of *Prototheca* (88.8% smaller for *HD* and 85.4% smaller for *LD* images). This seems to reinforce what emerged from the time analysis. A statistical Wilcoxon test has also been carried out on the mean number of clicks in all groups. According to the test, the mean number of clicks in *semi-auto* mode is significantly lower than *manual* mode (P < .001).

4.3. Annotation quality analysis

As it is common in object detection [13], we computed a range of evaluation metrics to explore annotations' quality, including Precision, Recall, IOU (intersection of union, also known as Jaccard index in some literature) and Acceptance Ratio. These parameters are explained in more detail, later in this section. Here we indicate with *Tp* (true positive) the number of truly identified objects, with *Fp*, the number of falsely identified objects, and with *Fn*, the number of missed (un-identified) objects by annotators. Following the literature, we set the IOU threshold to 50% for the calculation of *Tp*, *Fp*, and *Fn*, i.e. those objects, identified with an overlap higher than 50% with GT objects, are considered positive. *Tp*, *Fp*, and *Fn* are calculated according to Equations (5). In Eqs. (5), image-index, *j*, goes from 1 to *v*, i.e. the number of images given to each annotator (*v*=5), and the annotator-index, *m*, goes from 1 to *w*, i.e. the number of annotators (*w*=4).

$$Tp = \sum_{m=1}^{w} \sum_{j=1}^{v} True\_Positive_{j,m} \tag{5}$$

$$Fp = \sum_{m=1}^{w} \sum_{j=1}^{v} False\_Positive_{j,m}$$

$$Fn = \sum_{m=1}^{w} \sum_{j=1}^{v} False\_Positive_{j,m}$$

Fig. 12 shows that the number of identified objects (both *Tp* and *Fp*) in the *semi-auto* mode is higher than the identified objects in *manual* mode for all groups of images, although, in some cases, the number of *Fp* in *semi-auto* mode is higher than the *manual* mode (see appendix C for more detailed information).

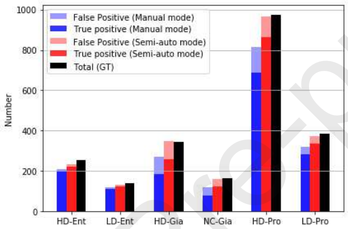


Fig. 12. True positive, *Tp* (dark color), false positive, *Fp* (light color), and total number of objects (black) in each group of images, with 50% IOU threshold. Blue-bars *manual* mode, red-bars *semi-auto* mode. (Colorful)

Precision, Recall and F1 score are calculated according to Eq. (6).

$$Precision = \frac{Tp}{Tp + Fp} \tag{6}$$

$$Recall = \frac{Tp}{Tp + Fn}$$

$$F1 = \frac{2 \times Presicion \times Recall}{Presicion + Recall}$$

Fig. 13 shows the average Precision, Recall and F1 score in both *manual* and *semi-auto* mode for each group of images. The comparison between *manual* and *semi-auto* mode in Fig. 13 shows that, unlike Precision, Recall is considerably increased in the *semi-auto* mode, which means that the *semi-auto* mode helped to reduce the number of *Fn* more than for the number of *Fp* (see appendix C for detailed information).

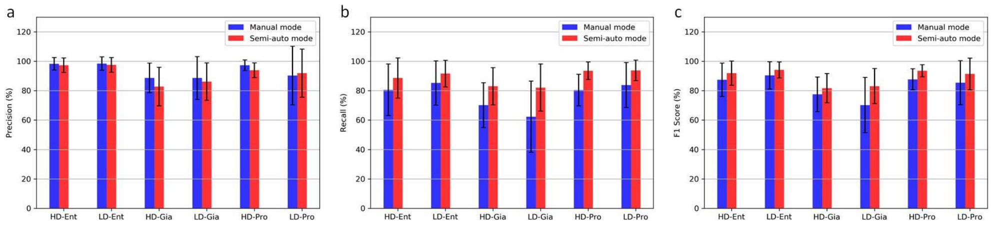
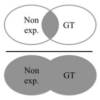


Fig. 13. (a) Average Precision for each group of images, (b) Average Recall for each group of images, (c) Average F1 score for each group of images. (Colorful)

IOU is a well-known metric that has been widely used in instance segmentation studies [10,14,18,34,43,47–51], as a measure of the annotators' accuracy in drawing objects' borders. IOU is a measure of the overlap between a drawn polygon (by non-experts in this case) and the ground truth polygon (by experts), and it is defined as in Eq. (7):

$$IOU = \frac{Area\, of\ overlap}{Area\, of\ union} = \qquad (7)$$



Note that, the mean IOU is only calculated on *Tp* (true positive) objects. We first calculate the summation of the entire objects' IOU within each image, then calculate *Mean_IOU* as shown in Eq. 8, where *m*, *j*, and *i* are the index of annotator, image, and object, respectively. Here, *L* is the total number of objects annotate by the four annotators in each group of images, and z refers to the number of objects within the image

$$Total\_IOU = \sum_{i=1}^{z} IOU_i \qquad (8)$$

$$Mean\_IOU = \frac{1}{L}\sum_{m=1}^{v}\sum_{j=1}^{w} Total\_IOU_j$$

Fig. 14 indicates that the IOUs (for *Entamoeba* and *Prototheca*, *HD* and *LD*) in *manual* and *semi-auto* mode do not show a significant difference. The IOU for *Giardia* images is 7% higher in *semi-auto* mode for *HD* images, and 10% higher in *semi-auto* mode for *LD* images (see appendix D for more information). Note that, unlike *Entamoeba* and *Prototheca*, which have a round shape (see Fig. 4), *Giardia* has a more complex shape, including sharp edges. We believe that our assistive tool is more effective (in terms of IOU) for challenging objects than for simpler objects.

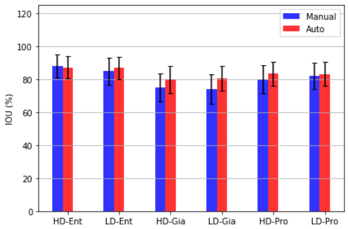


Fig. 14. Mean IOU for each group of images. (Colorful)

Fig. 15 presents a selection of samples of *Entamoeba*, *Giardia*, and *Prototheca* parasites, annotated by the expert *vs*. annotators (non-experts) in *manual* and *semi-auto* modes. As expected, the drawn masks in *manual* mode is coarser than the *semi-auto* mode, while it cost less number of points.

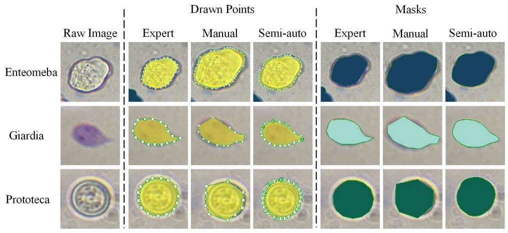


Fig. 15. Samples of raw images, of annotated images by expert and by non-expert annotators in *manual* mode and in *semi-auto* mode. “Drawn points” shows the points drawn with the polygon operator, and “Masks” shows the final generated mask.

We undertook further analysis by calculating the acceptance ratio of machine-proposed polygons by the four annotators in the *semi-auto* mode. Given a machine proposed polygons, the annotators are faced with three options: i) fully accept proposals without any modification, ii) accept with some modifications iii) reject (delete) proposals. Therefore, we define three parameters: *Fully_acceptance_ratio*, *Partially_acceptance_ratio*, and *Rejection_ratio* (calculated from all annotators) as in Eq. (9). Here the *Fully_acceptance_ratio*, represents the number of accepted proposed polygons without any modification, while the *Partially_acceptance_ratio* refers to those proposed polygons which are accepted whether with or without modification.

$$Fully\_Acceptance\_ratio = \frac{Num.of\ accepted\ polygons\,(Without\ modification)}{Num.of\ proposed\ polygons} \times 100 \qquad (9)$$

$$Partially\_Acceptance\_ratio = \frac{Num.of\ accepted\ polygons\,(With\,/\,Without\ modification)}{Num.of\ proposed\ polygons} \times 100$$

$$Rejection\_ratio = 100\% - Partially\ acceptance\ ratio$$

Table 1. Acceptance ratio of proposed polygons for each group of images. *Partially_acceptance_ratio* refers to machine-generated masks accepted by annotators, and *fully_acceptance*_ratio refers to those computer generated masks they are accepted and modified.

	Entamoeba		Giardia		Prototheca	
	HD	LD	HD	LD	HD	LD
Partially Acceptance ratio	83.84%	85%	73.42%	58.57%	95%	87.6%
Fully Acceptance ratio	41.1%	32.6%	40.3%	39%	85.8%	77%
Rejection ratio	16.16%	15%	26.58%	41.43%	5%	12.4%

Table 1 shows that in *HD* images, the annotators tend to accept proposals more often than *LD* images, which reinforces what emerged from the time and clicks analyses (for detailed information see appendix E). Based on appendix D (Tables D.2 and D.3), despite the fact that the annotators spend a significant amount of time for refining proposed masks, the final IOU of accepted/refined proposals by annotators does not show a noticeable improvement over the proposed masks.

5. Discussion

In this paper, we investigated non-expert annotators' behavior on a specialized domain (cell biology), using a bespoke segmentation annotation platform powered by a user-assistive tool. The annotators were asked to perform segmentation tasks in two modes: *manual* and *semi-auto* (assisted with a mask proposal network). Our results show that like the segmentation of everyday objects (e.g. using Cityscapes or COCO dataset), outsourcing the specialized annotation task in cell biology to non-experts can result in a decrease in the annotation cost, i.e. time spent, number of clicks, when supported by the assistive tool(see Figs. 9 and 11). Importantly, the overall IOU performance of non-expert annotations was higher with the assistive tool. Furthermore, our results show that *semi-auto* annotation resulted in consistently higher recall (which means that fewer objects/cells in the image were missed by the annotator). We have also investigated the behavioral patterns of annotators in both modes and identified some key directions for the design of future platforms.

Firstly, our analysis reveals that performing more clicks and spending more time on the segmentation of each object does not lead to significantly better annotation quality (see Tables B.2, B.6, and D.1). We believe that spending more time and more clicks on the task eventually lead to mental fatigue, which may result in poor quality annotation. This implies that the design of such platforms should focus not just on helping users to make accurate annotations, but also efficient ones with fewer clicks, hence less time. Conventional reward mechanisms of some crowdsourcing platforms calculate users' wages based on the number of clicks and time spent, which may have a perverse incentive to produce lower quality work. Hence, we suggest that wage calculations could take into account the efficiency of the annotator's work as well, in order to set the right motivation. Another way to improve user motivation may involve a system with non-monetary reward (e.g. gamification scoring system), nudging annotators toward more efficient annotations whilst maintaining the quality of the results. This reward system can be implemented in the tutorial phase, or embedded seamlessly throughout the annotation task to train annotators to do the task more efficiently.

Secondly, contrary to expectations, our results show that in the *semi-auto* mode, despite annotators spending a lot of time refining the proposed masks, the mean IOU of refined masks was not always improved. In cases where there was an improvement, it was only marginal (see appendix D, Tables D.2 and D.3). Furthermore, we observed that although the annotators tended to spend a lot of time refining a proposed mask, they did not pay sufficient attention to verify if a proposed mask contained a real parasite object, i.e. many false proposed masks were confirmed by the annotator and only a few ones were rejected (see Tables C.1 and E.1). Consequently, it resulted in a high number of *Fp* (False-positive) and low precision (see Fig. 13). The implication of this observation is noteworthy: the annotators seemed to have trusted the machine in identifying the object, but did not trust as much the segmentation that was done by the machine.

Consequently, the design of future platforms, especially for the tutorial phase, could emphasize the need to verify machine-proposed masks prior to refining them. Furthermore, the behavior we observed suggests the need to optimize the confidence threshold of the mask proposal network (set at 30% in our work). Setting a higher threshold, in fact, will force the machine to propose a mask only when it is really confident about it, to avoid the problem of over-trusting of the annotators. However, a higher threshold will mean fewer masks are proposed by the machines, potentially resulting in more time spent to segment objects from scratch. Alternatively, future platforms could present individually the generated masks to annotators, rather than in bulk within each image. We propose the exploration of these solutions as the topic for future researches. We also found that on average, the annotators spent 0.49±0.16 seconds per click when creating a new mask from scratch (for detailed information see appendix B, Table B.5), while the modification of a point took 1.5±0.9 seconds on average, in a mask either proposed by the machine or generated by themselves. This means that the modification of a few points is more efficient than creating a mask from scratch by the annotator. However, if the quality of machine-proposed mask is low, resulting in the need of modifying many points, it may be more efficient for annotators to generate a mask from scratch. From these results, we recommend that in a machine-proposed mask, if the number of points which requires modification is more than 30% of all total points, it may be more efficient to reject this proposed mask and create the mask from scratch by the annotator.

6. Conclusion

Our study sheds some light onto important behavioral features of non-expert annotators in performing segmentation tasks in the specialized domain of microbiology, when assisted by a supervised object detection algorithm. These insights can help inform the design of future systems, taking into account the performance trade-off due to human-machine interactions (e.g. human's perceived trust on machine), the complexity of images, and human factors (e.g. fatigue and motivation). However, we acknowledge that the present results are based on only four annotators (although they performed a total of 1842 and 2209 segmentations in manual and semi-auto mode, respectively, yielding a large number of activities for analysis), and are drawn from images from three parasite cells produced using a single microscope. Different cells may present different challenges for the annotation task, especially to non-experts. More specifically, different life stages of the parasites (i.e. cysts, spores, gametes), environmental stresses (that change the morphology of the parasite) and other objects could be present in the images, making the annotation task more challenging. Furthermore, it is not clear how annotators' behavior may change over a longer period of time, and if the system needs to be more adaptive to respond to this possible change. This calls for future studies to broaden the scope of the investigation, involving more participants and diverse microscopic images over a longer period of time. Crucially, a collective effort is needed to generate a public dataset for microbiology, similar to Cityscape or COCO datasets for everyday objects. Future work should also focus on how human annotators perceive machine recommendations, and how user interfaces can be designed to facilitate efficient, trusting and transparent human-machine interaction.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors

Appendix Overview

In this section, we provide detailed information about the annotators, in the annotation process in both modes of *manual* and *semi-auto*. All tables in this section present data for all the annotators

Appendix A. Data Statistics

To explore the correlation between annotations' cost and images' features such as shape, size, color, number of objects per images, and difficulty level of detecting objects in images, we computed different features of the images in each group. The number of objects in the images seems to be a factor that can influence the annotator's behavior, and consequently the cost of annotation. Fig. A.1 presents the number of parasites in each group of images.

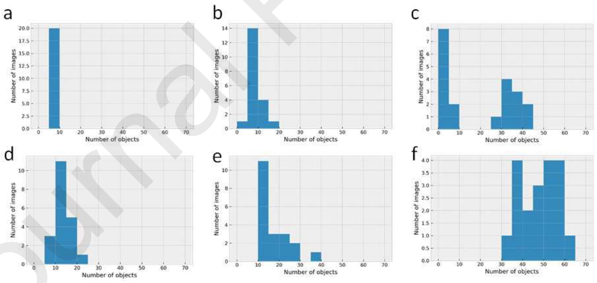


Fig. A.1. Histograms of the number of objects in images: (a) *LD Entamoeba*, (b) *LD Giardia*, (c) *LD Prototheca*, (d) *HD Entamoeba*, (e) *HD Giardia*, (f) *HD Prototheca*

The object's size is another factor that can affect the annotation's cost, including the number of clicks and time. To investigate the effect of annotating objects of different sizes on the annotator's performance, we have computed the object's size per each group of images as present in Table A.1.

Table A.1. Parasites' size - *HD-Ent*: high-dense *Entamoeba*, *LD-Ent*: low-dense *Entamoeba*, *HD-Gia*: high-dense *Giardia*, *LD-Gia*: low-dense *Giardia*, *HD-Pro*: high-dense *Prototheca*, *LD-Pro*: low-dense *Prototheca*

Image	Height (pixel)			Width (pixels)			Area (pixel)		
Group	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
HD-Ent	103	1099	560	113	1121	608	431k	1189k	355k
LD-Ent	97	1147	560	84	1160	549	36k	1169k	348k
HD-Gia	55	520	264	122	500	271	15k	206k	71k

Image	Height (pixel)			Width (pixels)			Area (pixel)		
Group	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
LD-Gia	109	524	263	126	586	263	20k	224k	69k
HD-Pro	27	460	206	89	502	214	3.4k	227k	46k
LD-Pro	56	556	217	50	524	218	8k	264k	50k

The *Entamoeba* and *Prototheca* have a round shape, while the *Giardia* has a non-round object and therefore is more challenging in terms of visibility and for drawing (see Fig. 4). *Entamoeba*, *Giardia*, and *Prototheca* are the biggest to the smallest objects in terms of pixels, based on Table A.1. On the other hand, *Prototheca* images are the most populated (dense) images, as there are 2023 objects in *Prototheca* images, 643 objects in *Giardia*, and 541 objects in *Entamoeba* images.

Appendix B. Time and clicks results

This section presents detailed results of clicks and time analysis for all participants. Table B.1 shows the *net-time* spent on each group of images by the four annotators and the expert biologist.

Table B.1. Net-time (seconds) spent on each group of images by four annotators and biologist. The first number is drawing time and second number refers to the modifying time

# user	Enteomeba				Giardia				Prototeca			
	HD		LD		HD		LD		HD		LD	
	Manual	S-auto	Manual	S-auto	Manual	S-auto	Manual	S-auto	Manual	S-auto	Manual	S-auto
# 1	440;72	26;161	285;0	0;66	490;10	73;251	133;4	10;189	878;40	37;105	694;51	63;116
# 2	525;117	52;240	235;94	44;266	356;41	136;153	201;25	88;97	1509;180	104;75	139;3	32;21
# 3	972;303	63;600	554;107	10;395	904;23	82;217	510;10	89;44	2581;178	323;273	232;0	9;79
# 4	951;88	159;624	389;14	0;167	682;15	149;277	293;18	132;39	2654;178	355;32	1420;247	223;40
Expert	4205;765	N/A	1553;210	N/A	2481;248	N/A	1565;82	N/A	8641;1112	N/A	3187;445	N/A

Tables B.2 and B.3 present the average time spent per object (drawing and modifying) in *manual* and *semi-auto* mode (calculated based on Eq. (2)).

Table B.2. Average spent time (drawing and modifying, in seconds) per object in *manual* mode. (Mean± Standard deviation)

# user	Enteomeba		Giardia		Prototeca	
	HD	LD	HD	LD	HD	LD
# 1	14.2±4.5	9.5±1.6	7.1±1.7	7.6±2.4	6.9±2.4	5.5±1.8
# 2	10.5±3.2	11.7±3.4	6.8±2	8±2.3	8.3±3.5	8.3±4.1
# 3	23.6±11.6	21.3±10.5	14.2±10.9	11.8±4	10.5±3.5	9.2±3.9
# 4	18.2±8.5	13.8±4.7	9.17±3.2	11.5±4	13.1±3.5	11.9±4
Expert	19.5±8.8	12.8±4.5	7.9±2.3	9.9±4.1	10±3.5	9.4±3.2

Table B.3. Average spent time (drawing and modifying, in seconds) per object in *semi-auto* mode. (Mean± Standard deviation)

# user	Enteomeba		Giardia		Prototeca	
	HD	LD	HD	LD	HD	LD
# 1	3.4±2.6	2.2±1.1	4.7±4.6	5.3±1.8	0.7±0.2	1.1±0.8
# 2	4.8±0.9	9.7±2.1	4.5±1.9	7.2±2.5	0.8±0.4	2.5±1.8
# 3	10.8±6.9	12.8±6.5	3.4±1.1	2.5±1.6	2±0.6	3.5±4.2
# 4	12.1±2.7	4.7±6.9	5.3±2.1	4.4±4.3	1.6±3	1.4±2.3

The average number of clicks per object in *manual* mode, for all four annotators, according to Eq. (4) are shown in Table B.4.

Table B.4. Average number of clicks (drawing and modifying) per object in *manual* mode. (Mean ± Standard deviation)

# user	Enteomeba		Giardia		Prototeca	
	HD	LD	HD	LD	HD	LD
# 1	33.4±10.54	24.7±3.9	14.8±2.8	17±3.5	16.5±4	14.2±2.8
# 2	21.8±6.3	21.1±5	15.3±3.3	15.3±3.3	17.9±5.6	19.1±7.5
# 3	42.9±14.7	45.9±16.5	33.4±9.8	30.8±7.5	19.4±5.5	24.4±6.7
# 4	25.4±8.2	22.8±7.9	15.8±3.8	17.6±4.7	17±15	15.6±3.3

In *manual* mode, when annotators are drawing parasites from scratch, the time between each click is different from person to person. Table B.5, illustrate the average time spent for each clicks for different group of images.

Table B.5. Average spent time (in seconds) per click for drawing parasites (Mean ± Standard deviation)

# user	Enteomeba		Giardia		Prototeca	
	HD	LD	HD	LD	HD	LD
# 1	0.36±0.05	0.38±0.04	0.48±0.1	0.43±0.07	0.4±0.05	0.37±0.1
# 2	0.45±0.08	0.51±0.06	0.42±0.08	0.5±0.08	0.43±0.09	0.41±0.05
# 3	0.4±0.14	0.37±0.06	0.41±0.3	0.37±0.07	0.51±0.14	0.38±0.16
# 4	0.62±0.1	0.58±0.06	0.55±0.12	0.62±0.09	0.72±0.11	0.63±0.1

The total number of clicks by annotators are presented in Table B.6. The first number shows the total number of clicks for drawing and second number shows the total number of clicks for modifying objects.

Table B.6. Total number of clicks for each group of images. (Num. of drawing clicks; num. of modifying clicks)

# user	HD Enteomeba		LD Enteomeba		HD Giardia		LD Giardia		HD Prototeca		LD Prototeca	
	Manual	S-auto	Manual	S-auto	Manual	S-auto	Manual	A-auto	Manual	A-auto	Manual	A-auto
# 1	1205;53	58;107	742;0	0;40	1041;3	191;274	306;2	26;197	2198;24	108;54	1919;35	134;102
# 2	1311;107	85;170	593;85	95;314	891;33	301;118	431;16	189;66	3628;205	290;43	326;2	61;15
# 3	2318;255	103;448	1425;78	14;251	2175;13	164;116	1357;5	182;38	5106;106	659;121	611;0	18;73
# 4	1451;30	255;571	664;4	0;137	1207;5	283;260	477;4	254;33	3674;63	661;22	2197;99	447;38

Appendix C. Precision and recall

Table C.1 shows the number of truly identified, wrongly identified, and missed objects in both *manual* and *semi-auto* is calculated (for calculation, the IOU threshold is set to 50%).

Table C.1. *Tp* (true-positive), *Fp* (false-positive) and *Fn* (false-negative) with IOU-threshold=50% for each group of images, per annotators (num. of *Tp* ; num. of *Fp* ; num. of *Fn*)

# user	HD Enteomeba		LD Enteomeba		HD Giardia		LD Giardia		HD Prototeca		LD Prototeca	
	Manual	S-auto	Manual	S-auto	Manual	S-auto	Manual	S-auto	Manual	S-auto	Manual	S-auto
# 1	33;3;23	45;2;11	24;6;11	30;3;5	55;15;49	71;41;33	10;8;33	24;13;19	103;30;89	167;34;25	115;20;57	150;13;22
# 2	60;1;8	59;1;10	27;1;9	31;1;5	33;25;44	57;6;20	9;19;22	22;5;9	167;35;59	202;18;24	16;1;2	18;0;1
# 3	50;4;4	50;10;4	30;1;1	30;1;1	37;28;33	54;33;16	36;8;7	40;12;3	209;53;82	259;28;32	13;12;4	15;11;2
# 4	56;1;21	66;1;11	28;1;7	32;2;3	60;16;32	74;12;18	22;5;26	36;7;12	208;8;56	235;23;29	136;4;42	152;13;27
Precision	95.67	94.01	92.37	94.61	68.77	73.56	65.81	76.72	84.50	89.33	88.32	90.05
Recall	78.03	85.93	79.56	91.95	53.93	74.62	46.66	73.93	70.60	89.52	72.91	86.56

Appendix D. Intersection of Union

IOUs for each group of images in both *manual* and *semi-auto* are shown in Tables D.1 and D.2.

Table D.1. Final IOU in *manual* mode for each group of images (Mean ± Standard deviation).

# user	Enteomeba		Giardia		Prototeca	
	HD	LD	HD	LD	HD	LD
# 1	85±7.9	75.5±6.2	72±11.1	68.8±12.6	77.3±8.9	79.4±7.7
# 2	85.5±8.9	85.1±10.4	69.5±10.7	64.8±10.8	77.7±8.9	80.3±9.4
# 3	87.4±12.4	90±5	71±12.3	76.3±8.8	78.2±11.9	75.9±23.3
# 4	90.1±6.5	90.8±5.5	76.2±12.8	75.6±13.4	84±6.8	85.9±5.7

Table D.2. Final IOU in *semi-auto* mode for each group of images (Mean ± Standard deviation).

# user	Enteomeba		Giardia		Prototeca	
	HD	LD	HD	LD	HD	LD
# 1	86.8±6.5	86.6±7.4	75.6±11.9	75.2±12.8	80.7±7.8	83.7±6.9
# 2	87.8±6	85.9±9.5	81.8±8.5	79±10.9	82.8±8.7	86±6.9
# 3	84.8±12.4	87.1±6.5	76.6±13.7	82.2±6.4	83.4±9.9	80.8±13.3
# 4	88.6±4.9	86.6±7.4	80.1±9.8	79.3±10.2	84.2±7.3	82±7.8

The IOUs for the masks generated in the *semi-auto* mode in comparison with the GT (ground truth) are shown in Table D.3.

Table D.3. IOU of computer generated masks (Mean ± Standard deviation).

# user	Enteomeba		Giardia		Prototeca	
	HD	LD	HD	LD	HD	LD
# 1	86±6.9	86.3±7	78±9.5	80±7.2	81.4±6.8	84.2±6.6
# 2	87±6	85.3±8.3	81.1±8.2	80±8.5	83.7±6.5	85.5±7.4
# 3	84.7±8.9	86.1±7.2	79±9.6	82.6±6.4	85.3±6.8	84.7±10
# 4	86.3±6.4	85.8±7.1	80.2±7.2	80.6±8	84.7±6.5	81.6±7.8

Appendix E. Semi-auto mode complementary results

Number of proposed objects, along with the number of added and removed parasites in *semi-auto* mode are shown in Table E.1.

Table E.1. Accepted, removed and modified mask proposals in *semi-auto* mode. (P: total number of proposed objects, A: Number of added objects by annotator, D: number of deleted objects by annotator, T: the final number of annotated objects)

# user	HD Enteomeba				LD Enteomeba				HD Giardia				LD Giardia				HD Prototeca				LD Prototeca			
	P	A	D	T	P	A	D	T	P	A	D	T	P	A	D	T	P	A	D	T	P	A	D	T
# 1	56	2	11	47	40	0	7	33	140	14	40	112	58	2	23	37	203	7	9	201	158	12	7	163
# 2	68	4	13	59	38	4	10	32	86	14	37	63	44	8	25	27	225	19	24	220	27	3	12	18
# 3	60	2	2	60	31	1	1	31	82	6	1	87	47	7	12	42	256	35	4	287	44	1	19	26
# 4	76	7	16	67	38	0	4	34	106	12	32	86	61	9	27	43	220	46	8	258	142	12	31	165

Table E.2. Number of partially and fully accepted polygons (num. of accepted proposals with modification; num. of accepted proposal without modification).

# user	Enteomeba		Giardia		Prototeca	
	HD	LD	HD	LD	HD	LD
# 1	9 ; 36	10 ; 23	39 ; 61	24 ; 11	25 ; 169	23 ; 128
# 2	34 ; 21	27 ; 1	20 ; 29	12 ; 7	15 ; 186	4 ; 11
# 3	25 ; 33	16 ; 14	40 ; 41	8 ; 37	43 ; 209	11 ; 14
# 4	43 ; 17	24 ; 10	38 ; 36	7 ; 27	0 ; 212	1 ; 133

References

[1] D. Shen, G. Wu, H. Il Suk, Deep Learning in Medical Image Analysis, Annu. Rev. Biomed. Eng. (2017). https://doi.org/10.1146/annurev-bioeng-071516-044442.

[2] R. Anantharaman, M. Velazquez, Y. Lee, Utilizing Mask R-CNN for Detection and Segmentation of Oral Diseases, in: Proc. - 2018 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2018, 2019. https://doi.org/10.1109/BIBM.2018.8621112.

[3] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2018. https://doi.org/10.1007/978-3-030-00889-5_1.

[4] C.X. Hernández, M.M. Sultan, V.S. Pande, Using deep learning for segmentation and counting within microscopy data, ArXiv. (2018). http://arxiv.org/abs/1802.10548.

[5] O. Ronneberger, P. Fischer, T. Brox, Dental X-ray Image segmentation using a U-shaped Deep convolutional network, Int. Symp. Biomed. Imaging. (2015) 1–13.

[6] Y. Xue, G. Bigras, J. Hugh, N. Ray, Training Convolutional Neural Networks and Compressed Sensing End-to-End for Microscopy Cell Detection, IEEE Trans. Med. Imaging. (2019). https://doi.org/10.1109/TMI.2019.2907093.

[7] K. He, G. Gkioxari, P. Dollar, R. Girshick, Mask R-CNN, in: Proc. IEEE Int. Conf. Comput. Vis., 2017. https://doi.org/10.1109/ICCV.2017.322.

[8] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in: Adv. Neural Inf. Process. Syst., 2015.

[9] Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You only look once: Unified, real-time object detection." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788. 2016 .

[10] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2015. https://doi.org/10.1007/978-3-319-24574-4_28.

[11] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2015. https://doi.org/10.1109/CVPR.2015.7298965.

[12] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The Cityscapes Dataset for Semantic Urban Scene Understanding, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2016. https://doi.org/10.1109/CVPR.2016.350.

[13] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: Common objects in context, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2014. https://doi.org/10.1007/978-3-319-10602-1_48.

[14] J. Jäger, G. Reus, J. Denzler, V. Wolff, K. Fricke-Neuderth, LOST: A flexible framework for semi-automatic image annotation, (2019). http://arxiv.org/abs/1910.07486.

[15] J. Howe, The Rise of Crowdsourcing, Wired Mag. (2006). https://doi.org/10.1086/599595.

[16] H. Su, J. Deng, L. Fei-Fei, Crowdsourcing annotations for visual object detection, in: AAAI Work. - Tech. Rep., 2012.

[17] A. Kovashka, O. Russakovsky, L. Fei-Fei, K. Grauman, Crowdsourcing in computer vision, Found. Trends Comput. Graph. Vis. (2016). https://doi.org/10.1561/0600000071.

[18] B.C. Russell, A. Torralba, K.P. Murphy, W.T. Freeman, LabelMe: A database and web-based tool for image annotation, Int. J. Comput. Vis. (2008). https://doi.org/10.1007/s11263-007-0090-8.

[19] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. (2010). https://doi.org/10.1007/s11263-009-0275-4.

[20] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: The KITTI dataset, Int. J. Rob. Res. (2013). https://doi.org/10.1177/0278364913491297.

[21] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, N. Navab, AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images, IEEE Trans. Med. Imaging. (2016). https://doi.org/10.1109/TMI.2016.2528120.

[22] S. Boorboor, S. Nadeem, J.H. Park, K. Baker, A. Kaufman, Crowdsourcing lung nodules detection and annotation, in: 2018. https://doi.org/10.1117/12.2292563.

[23] C. Vondrick, D. Ramanan, D. Patterson, Efficiently scaling up video annotation with crowdsourced marketplaces, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2010. https://doi.org/10.1007/978-3-642-15561-1_44.

[24] A. Bearman, O. Russakovsky, V. Ferrari, L. Fei-Fei, What's the point: Semantic segmentation with point supervision, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2016. https://doi.org/10.1007/978-3-319-46478-7_34.

[25] S. Bianco, G. Ciocca, P. Napoletano, R. Schettini, An interactive tool for manual, semi-automatic and automatic video annotation, Comput. Vis. Image Underst. (2015). https://doi.org/10.1016/j.cviu.2014.06.015.

[26] S. Ørting, A. Doyle, A. van Hilten, M. Hirth, O. Inel, C.R. Madan, P. Mavridis, H. Spiers, V. Cheplygina, A Survey of Crowdsourcing in Medical Image Analysis, ArXiv Prepr. ArXiv1902.09159 (2019). http://arxiv.org/abs/1902.09159.

[27] D. Mitry, T. Peto, S. Hayat, J.E. Morgan, K.T. Khaw, P.J. Foster, Crowdsourcing as a Novel Technique for Retinal Fundus Photography Classification: Analysis of Images in the EPIC Norfolk Cohort on Behalf of the UKBiobank Eye and Vision Consortium, PLoS One. (2013). https://doi.org/10.1371/journal.pone.0071154.

[28] S. Mavandadi, S. Dimitrov, S. Feng, F. Yu, U. Sikora, O. Yaglidere, S. Padmanabhan, K. Nielsen, A. Ozcan, Distributed medical image analysis and diagnosis through crowd-sourced games: A malaria case study, PLoS One. (2012). https://doi.org/10.1371/journal.pone.0037245.

[29] A. Chavez-Aragon, W.S. Lee, A. Vyas, A crowdsourcing web platform -hip joint segmentation by non-expert contributors, in: MeMeA 2013 - IEEE Int. Symp. Med. Meas. Appl. Proc., 2013. https://doi.org/10.1109/MeMeA.2013.6549766.

[30] E. Heim, Inaugural-Dissertation, 2018, Large-scale medical image annotation with quality-controlled crowdsourcing, PhD thesis, Heidelberg University, 2018.

[31] A.Q. O'Neil, J.T. Murchison, E.J.R. van Beek, K.A. Goatman, Crowdsourcing Labels for Pathological Patterns in CT Lung Scans: Can Non-experts Contribute Expert-Quality Ground Truth?, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2017. https://doi.org/10.1007/978-3-319-67534-3_11.

[32] J.J. Chen, N.J. Menezes, A.D. Bradley, Opportunities for Crowdsourcing Research on Amazon Mechanical Turk, Hum. Factors. (2011). https://doi.org/10.1145/1357054.1357127.

[33] C. Halaschek-Wiener, J. Golbeck, A. Schain, M. Grove, B. Parsia, J. Hendler, Annotation and provenance tracking in Semantic Web photo libraries, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2006. https://doi.org/10.1007/11890850_10.

[34] D. Acuna, H. Ling, A. Kar, S. Fidler, Efficient Interactive Annotation of Segmentation Datasets with Polygon-RNN++, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2018. https://doi.org/10.1109/CVPR.2018.00096.

[35] X. Qin, S. He, Z. Zhang, M. Dehghan, M. Jagersand, ByLabel: A boundary based semi-automatic image annotation tool, in: Proc. - 2018 IEEE Winter Conf. Appl. Comput. Vision, WACV 2018, 2018. https://doi.org/10.1109/WACV.2018.00200.

[36] L. Lejeune, M. Christoudias, R. Sznitman, Expected Exponential Loss for Gaze-Based Video and Volume Ground Truth Annotation, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2017. https://doi.org/10.1007/978-3-319-67534-3_12.

[37] A.A.E. Ahmed, I. Traore, A new biometrie technology based on mouse dynamics, IEEE Trans. Dependable Secur. Comput. (2007). https://doi.org/10.1109/TDSC.2007.70207.

[38] C. Feher, Y. Elovici, R. Moskovitch, L. Rokach, A. Schclar, User identity verification via mouse dynamics, Inf. Sci. (Ny). (2012). https://doi.org/10.1016/j.ins.2012.02.066.

[39] G. Wang, T. Konolige, C. Wilson, X. Wang, H. Zheng, B.Y. Zhao, You are how you click: Clickstream analysis for sybil detection, in: Proc. 22nd USENIX Secur. Symp., 2013.

[40] E. Heim, A. Seitel, J. Andrulis, F. Isensee, C. Stock, T. Ross, L. Maier-Hein, Clickstream Analysis for Crowd-Based Object Segmentation with Confidence, IEEE Trans. Pattern Anal. Mach. Intell. (2018).

https://doi.org/10.1109/TPAMI.2017.2777967.

[41] D. Gurari, M. Sameki, M. Betke, Investigating the Influence of Data Familiarity to Improve the Design of a Crowdsourcing Image Annotation System, in: 4th AAAI Conf. Hum. Comput. Crowdsourc., 2016.

[42] C.R. Alex Ratner, Paroma Varma, Braden Hancock, Weak Supervision: A New Programming Paradigm for Machine Learning | SAIL Blog, Standford AI Lab Blog. (2019).

[43] L. Castrejón, K. Kundu, R. Urtasun, S. Fidler, Annotating object instances with a polygon-RNN, in: Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, 2017. https://doi.org/10.1109/CVPR.2017.477.

[44] C. Wang, K. Huang, W. Ren, J. Zhang, S. Maybank, Large-Scale Weakly Supervised Object Localization via Latent Category Learning, IEEE Trans. Image Process. (2015). https://doi.org/10.1109/TIP.2015.2396361.

[45] U. Ramer, An iterative procedure for the polygonal approximation of plane curves, Comput. Graph. Image Process. (1972). https://doi.org/10.1016/S0146-664X(72)80017-0.

[46] Create COCO Annotations From Scratch — Immersive Limit, (2020). https://www.immersivelimit.com/tutorials/create-coco-annotations-from-scratch.

[47] K.K. Maninis, S. Caelles, J. Pont-Tuset, L. Van Gool, Deep Extreme Cut: From Extreme Points to Object Segmentation, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2018. https://doi.org/10.1109/CVPR.2018.00071.

[48] D.P. Papadopoulos, J.R.R. Uijlings, F. Keller, V. Ferrari, Extreme Clicking for Efficient Object Annotation, in: Proc. IEEE Int. Conf. Comput. Vis., 2017. https://doi.org/10.1109/ICCV.2017.528.

[49] B. Adhikari, J. Peltomäki, J. Puura, H. Huttunen, Faster Bounding Box Annotation for Object Detection in Indoor Scenes, in: Proc. - Eur. Work. Vis. Inf. Process. EUVIP, 2019. https://doi.org/10.1109/EUVIP.2018.8611732.

[50] W. Lee, J. Na, G. Kim, Multi-task self-supervised object detection via recycling of bounding box annotations, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2019. https://doi.org/10.1109/CVPR.2019.00512.

[51] H. Ling, J. Gao, A. Kar, W. Chen, S. Fidler, Fast interactive object annotation with curve-GCN, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2019. https://doi.org/10.1109/CVPR.2019.00540.

Highlights

- Our assistive tool enables non-expert annotators to perform annotation of microbiological images accurately and quickly.
- Our study sheds some light on important behavioral features of non-expert annotators.
- Our findings can help inform the design of future annotation systems.
- Annotation quality is not found to be strongly correlated with either annotation time or quantities including the number of regions and polygon vertices.
- Results reveal that other than time reduction, annotation platforms should encourage annotators toward more efficient annotations to maintain their motivation.

Vitae

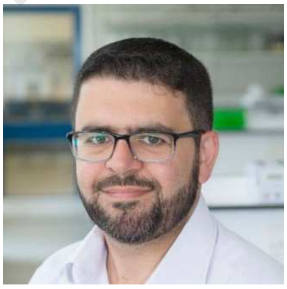
Saber Mirzaee Bafti received his B.Sc in Electronic Engineering from Chamran Technical and Vocational University, Iran, in 2010, and his M.Sc in Electronic Engineering from Sadjad University of Technology, Iran, in 2014. He is currently a Ph.D. candidate in Electronic Engineering at University of Kent, UK. His fields of research include computer vision, medical image processing, robotics, and embedded systems.

Chee Siang Ang is a Senior Lecturer in Multimedia and Digital Systems in the School of Engineering and Digital Arts, University of Kent. His main research area is in digital health, where he invest-igates, designs and develops new technologies which can provide treatment and (self-) management of health conditions, through effective prevention, early intervention, personalised treatment and continuous monitoring of the conditions.

Md Moinul Hossain received his BSc degree in Computer Science and Engineering from Bangladesh and MSc. degree in Wireless Communications and Systems Engineering from the University of Greenwich in 2005 and 2009, respectively, and his PhD degree in Electronic Engineering in the field of Instrumentation and Measurement from the University of Kent, UK, in 2014. He is currently a Lecturer of Electronic Engineering with the School of Engineering and Digital Arts, University of Kent. His current research interests include machine learning, medical image processing, sensors and condition process monitoring.

Dr Gianluca Marcelli is a lecturer in Bioengineering at the University of Kent. The main contribution of his multidisciplinary research lies in Biomechanics. He has developed computational models to understand the mechanical properties of human red blood cell and cell signalling in the ovary.

Marc Alemany Fornes pursued his BSc in Biotechnology at the University of Vic-Central University of Catalonia and during the final year joined the Molecular Photopharmacology group of the TR2lab to develop his final degree thesis, where he contributed to the creation of a system based on luminescence to analyse the response of photo-activated drugs. After that, he studied a MSc in Biomedicine at the University of Kent, where he joined the Laboratory of Molecular and Evolutionary Parasitology.



Anastasios D. Tsaousis is a Reader in Molecular Parasitology at the School of Biosciences at the University of Kent. He studied BSc in Biology at the University of Crete (Greece) followed by a PhD in Molecular Cell Evolution at the Newcastle University (UK). After that, He have undertaken several positions as a postdoctoral fellow in Canada (Dalhousie University & University of Saskatchewan) and Czech Republic (Charles University in Prague). Since 2013, He is a Principal Investigator of the Laboratory of Molecular and Evolutionary Parasitology and published more than 30 of peer-reviewed articles in the field.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: