Image recognition based on deep learning in *Haemonchus contortus* motility assays

Martin Žofka, Linh Thuy Nguyen, Eva Mašátová, Petra Matoušková*
Department of Biochemical Sciences, Faculty of Pharmacy, Charles University, Heyrovského 1203, 500 05 Hradec Králové, Czech Republic.

Correspondence:
*Corresponding Author
matousp7@faf.cuni.cz

Abstract

Poor efficacy of some anthelmintics and rising concerns about the widespread drug resistance have highlighted the need for new drug discovery. The parasitic nematode *Haemonchus contortus* is an important model organism widely used for studies of drug resistance and drug screening with the current gold standard being the motility assay. We applied a deep learning approach Mask R-CNN for analysing motility videos and compared it to other commonly used algorithms with different levels of complexity, namely Wiggle Index and Wide Field-of-View Nematode Tracking Platform. Mask R-CNN consistently outperformed the other algorithms in terms of the forecast precision across the videos containing varying rates of motile worms with a mean absolute error of 5.6%. Using Mask R-CNN for motility assays confirmed the common problem of algorithms that use Non-Maximum Suppression in detecting overlapping objects, which negatively impacted the overall precision. The use of intersect over union (IoU) as a measure of the classification of motile / non-motile instances had an overall accuracy of 89%. In comparison to the existing methods evaluated here, Mask R-CNN performed better and we
can anticipate that this method will broaden the number of possible approaches to video analysis of worm motility. IoU has shown promise as a good metric for evaluating motility of individual worms.

Author summary

Deep Learning has made significant improvements over the past years reaching a level where models are able to successfully detect heterogeneous object instances in an image. The ability to precisely detect larvae in an image then allows the researchers to perform multiple types of analyses that work with the viability of the parasites and are not only limited to motility assays, but also include larval development tests and egg hatch tests which are common in vitro assays for diagnosis of drug resistance. Although the other types of tests were not the main subject of this study, they show the versatile applications that are at hand with the use of Deep Learning models. The most important aspect of these applications is the possibility to automate these in vitro assays and scale them, therefore allowing researchers to avoid tedious repetitive tasks and focus on activities where they provide a higher added value.

Introduction

Parasitic nematode infections present a substantial problem to human and veterinary medicine. Anthelmintics are often used as a weapon of choice to control or eliminate parasitic worms. With increasing reports of drug resistance, there is an ongoing effort put into the development of new anthelmintics [1, 2]. In view of this, drug screening and drug testing methods are clearly necessary. A number of important human and animal nematodes were already studied in large chemical library-scale screenings, although, nearly all of the studies were aimed at well-established non-parasitic Caenorhabditis elegans because of the scientific interest and medical applications as potential human disease models [3-7].
Nowadays, in vitro assays can measure the effects of compounds on development, growth, behaviour, and motility. Several approaches have been pursued in attempts to deliver robust, automated assays on nematode phenotype [8-11]. The current gold standard for measuring drug effectiveness is in vitro assessment of worm motility, as measured via microscopy. The automated phenotyping from short video recordings offers a useful assay for high-throughput whole-organism phenotypic screening and avoids manual scoring. In assays with parasitic worms under experimental conditions, several challenges need to be overcome to enhance performance of motility measurements. These challenges are associated with the different worm lengths, their tendency to clump, and also the range of diverse complex patterns of movement different from that seen in C. elegans.

Focusing on veterinary importance, the parasitic nematode Haemonchus contortus, known as barber’s pole worm, is a model organism often used for drug screening and studying drug resistance; due to its socio-economic importance, known genome and transcriptome [12, 13]. H. contortus infects predominantly small ruminants, it resides in abomasum where it feeds on blood. It has a direct life cycle, the third-larval stage (L3s) can be stored, and researchers commonly use L3s or in vitro fourth larval stage for motility studies. Many automated or semi-automated approaches to measure motility have already been described such as via electrical impedance [14] or infrared light beam-interference [5]. Nevertheless, for our study, we have compared various (representative) algorithms on how to analyse the video recordings in H. contortus with a different level of complexity in terms of model preparation as well as computational cost. Therefore, we were interested to assess if the additional complexity of the methods leads to a sufficient increase in precision, that would justify the time/performance trade-off of these methods.

The first one mentioned is an algorithm which calculates a motility index value based on measuring the standard deviation of the pixel’s light intensity averaged for a number of
frames. This Wiggle Index (WI) assay led to many “hit compounds” in screening of a large variety of compound libraries [15, 16]. The second approach called Wide Field-of-View Nematode Tracking Platform (WF-NTP) is built on the use of the Gaussian adaptive threshold of the image frames to identify worms and was applied in model free-living nematode \textit{C. elegans}. It was initially developed for drug screening purposes, nevertheless, it has also found applications in other studies such as detailed genetic or behavioural studies [17]. In our study, we customized the associated WF-NTP software written in Python and compared it to the last approach which involves the application of deep learning.

The advent of machine learning has revolutionized the surveying and classifying of biological data including image recognition. Large-scale computerization reduced the time that needs to be spent in the laboratory, however the rate-limiting step is the analysis and interpretation of data. In the field of computer vision, deep learning algorithms with a convolutional neural network (CNN) architecture have made rapid advancements on a variety of image classification tasks in cell cultures [18], yet none of them have been applied to parasitic nematodes. Object-detection-based deep learning methods have also been adapted for instance segmentation. They work on the principle of predicting bounding boxes for all objects in an image and use non-maximum suppression to remove redundant bounding-box predictions. Here we used a state-of-the-art region-based CNN called Mask R-CNN [19] which predicts an object mask for each bounding box, and compared it to the WI and WF-NTP methods. The aim was to evaluate the application of the Mask R-CNN method to the field of drug discovery of anthelmintics via measuring motility of \textit{H. contortus}, using a metric called intersection over union (IoU).
**Results**

**Motility forecast errors**

In the manual video processing, we observed the motility and the live rate. To measure the performance of the algorithms, we used the motility rate rather than the live rate even though the stillness does not implicate a dead worm. The reason for this choice was that the algorithms detect movement and are not able to differentiate if the movement was caused by the object itself or by an interaction with another object. The general manual motility rate was always higher than the manual live rate (Supplementary Table S1). We measured the performance of the individual algorithms by evaluating the errors, defined as the differences between the manual processing of the file and the respective algorithm. Out of the three methods, Mask R-CNN had the lowest mean absolute error (MAE) with a value of 5.6%, followed by WF-NTP (8.76%) and the WI (14.2%). For WI we can apply a correction to limit the maximum motility value to 1, then the capped MAE is reduced to 11.8%. In terms of the bias of the predictions, the WI had the lowest mean error (ME) of -0.71% therefore slightly overestimating the motility rate on average, while both WF-NTP and Mask R-CNN underestimated the motility rate with values of 6.52% and 1.95%, respectively (Table 1).

WI had an increasing volatility with an increase in the ratio of alive worms in the group. WF-NTP and Mask R-CNN had a tendency to systematically underestimate for the motility group 100 as can be seen in Fig 1. The mean and standard deviation of motility rates per motility group and algorithm is shown also in Supplementary Table S2.

**Table 1: Motility error metrics for algorithms**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>MAE, %</th>
<th>ME, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiggle Index</td>
<td>14.2</td>
<td>-0.71</td>
</tr>
<tr>
<td>WF-NTP</td>
<td>8.76</td>
<td>6.52</td>
</tr>
</tbody>
</table>
The differences between the manual processing and the respective algorithm were expressed as the mean absolute error (MAE) and mean error (ME).

**Fig 1**: Motility rates comparisons for individual motility groups per algorithm. The number of the motility group indicates the percentage of live larvae. The motility rate is visualized as box plots.

**Instance detection performance**

We only compared Mask R-CNN and WF-NTP for the total worm count as the WI does not detect individual instances of worms and therefore this information was not available. Similarly as for the motility rate, the precision was measured based on the error term between the manual counts and the respective algorithm. In terms of the size of the error, the mean absolute percentage error (MAPE) was used and we observed that the error was significantly lower (P < 0.0001, Wilcoxon signed-rank test) for Mask R-CNN (7.6%) than for WF-NTP (40.23%). Both algorithms had a tendency to underestimate the number of worms on average, therefore not detecting all the worm instances in a video. This was confirmed by the mean percentage error (MPE) with values of 5.61% for Mask R-CNN and 40.23% for WF-NTP (Table 2). The reported MAPE and MPE for WF-NTP had the same value because all the detections underestimated the total number of worms. WF-NTP underestimated the worm count similarly for all motility groups, while Mask R-CNN had similar box plots in comparison to the manual counts apart from underestimating the motility group 0 as shown in Fig 2. The mean error and standard deviation of the differences between the manual count and the respective motility algorithm is shown in Supplementary Table S3.

| Mask R-CNN | 5.61 | 1.95 |
Table 2: Worm count error metrics for algorithms

<table>
<thead>
<tr>
<th>Algorithm / Metric</th>
<th>MAPE, %</th>
<th>MPE, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>WF-NTP</td>
<td>40.23</td>
<td>40.23</td>
</tr>
<tr>
<td>Mask R-CNN</td>
<td>7.6</td>
<td>5.61</td>
</tr>
</tbody>
</table>

The differences between the manual counts and the respective algorithm were expressed as the mean absolute percentage error (MAPE) and mean percentage error (MPE).

Fig 2: Worm count comparisons for individual motility groups per algorithm. The number of the motility group indicates the percentage of live larvae. The number of worms is visualized in box plots.

Mask R-CNN detection and classification precision

The precision of detecting worm instances in a video was on average 98.6%. The majority of cases where the detection was incorrect was due to debris that had a similar size and shape as a worm. For the detected worms there was an overall accuracy of 89% to correctly label them as either motile or non-motile based on their mean IoU values. In Fig 3, the distribution for different IoU values clearly showed that most of misclassified cases were located around the threshold 0.8, while the classification accuracy increased for values further away from the threshold. The precision and recall values (Table 3) confirm that the prevalent case was the misclassification of motile worms as non-motile.

The errors defined as differences between Mask R-CNN and the manually evaluated motility rates were tested for normality using the D’Agostino-Pearson test as well as Shapiro-
Wilk test. In both cases we could not reject the null hypothesis that the data were from a normal
distribution with p-values 0.85 and 0.35, respectively (Table 4).

**Fig 3:** The distribution for different Intersection over Union (IoU). Both (a) histogram of IoU
and the associated number of worms per bin classified into motile / non-motile and (b) histogram of IoU and the associated probability per class (motile / non-motile) show that the
majority of misclassified cases were located around the 0.8 IoU.

**Table 3:** Classification performance metrics

<table>
<thead>
<tr>
<th>Classification</th>
<th>Precision, %</th>
<th>Recall, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motile</td>
<td>94</td>
<td>88</td>
</tr>
<tr>
<td>Non-Motile</td>
<td>84</td>
<td>92</td>
</tr>
</tbody>
</table>

Precision and recall metrics for classification performance

**Table 4:** Normality tests for the motility rate error terms between Mask R-CNN and the manual
processing

<table>
<thead>
<tr>
<th>Normality test</th>
<th>Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>D’Agostino-Pearson</td>
<td>0.32</td>
<td>0.85</td>
</tr>
<tr>
<td>Shapiro-Wilk</td>
<td>0.97</td>
<td>0.35</td>
</tr>
</tbody>
</table>

For both normality tests we cannot reject the null hypothesis that the data were from a normal
distribution.
Discussion

The motility phenotype is often used for drug screening and drug testing of anthelmintics. In our study, we deployed and compared three algorithms to measure motility in the parasitic nematode *H. contortus*. The results showed that the performance increased with an increasing complexity of the algorithm. WI apart from requiring very little configuration and being the fastest algorithm had some limitations in comparison to the other methods; it was by design not able to detect individual instances of worms and a control group with only motile worms needed to be created in order to be able to translate the WI into a percentage for the motility rate for each video, else it was hard to interpret the numerical value of the WI. Because of these limitations and the worse precision, it would be advisable to use WI only for initial studies to quickly test experimental results as the method is the easiest to set up.

For WF-NTP, overlapping parasites had the highest impact on the decline in the detection ability, as the algorithm filters out objects that are above or below a certain pixel size. This resulted in detecting overlapping worms as a single object, where these objects were then excluded based on the size threshold. In consequence, the total number of detected instances in an image was significantly lower than the actual number obtained by manually processing the videos. Although WF-NTP struggled with detecting all instances, the classification motile / non-motile worked well, once an instance was detected.

Mask R-CNN had the lowest MAE, nonetheless similarly to WF-NTP it struggled with overlapping objects, primarily due to the fact that the algorithm uses Non-Maximum Suppression. As a result, some instances were not detected, although in contrast to WF-NTP the worm clews were not discarded as a whole. The problem of detecting overlapping objects is currently an area of active research in object detection, particularly for biological data. For instance, Böhm et al. [20] are working on algorithms that can handle overlapping objects and have had success with applying their algorithm to *C. elegans*. Motile worms had a higher chance
of being detected, because they had higher variability across the frames due to their movement.

In terms of precision, the algorithm was slightly biased towards underestimating the motility rate, however the errors were normally distributed enabling us to increase the number of acquired videos in order to get to a desired level of confidence and therefore successfully replace manual processing.

The results confirmed that IoU can be used as a metric for determining the motility of larvae, although some modifications might be required in order to deal with slow-moving worms, such as including the total distance travelled during the video to complement the frame-by-frame information. Furthermore, the IoU threshold should be considered a parameter to be optimized as it can vary based on the age of the larvae, the temperature of the environment and other factors. The downside of using Mask R-CNN is that it requires a quite big amount of annotated data (provided in the Data availability), which is time demanding to create and that the training runs 1-2 days on a GPU instance.

In general, for quick low-precision experiments, the best option is to use the WI as it can be used out-of-the-box with slight modifications to a few parameters. If we need to detect individual instances, Mask R-CNN is the best choice, although it is more time-demanding to prepare the data and train the model.

**Materials and methods**

**Procurement of L3s of *H. contortus***

Six-month-old sheep was orally infected with 8,000 infective L3s of MHco3 strain (Inbred Susceptible Edinburgh, ISE) [21]. Four weeks post-infection, faecal samples were collected. L3s were produced from eggs by incubating faeces in a plastic box covered with foil at 27 °C for 7 days. Then, faeces were rinsed twice in tap water which was poured into 1,000mL conical measuring cups in which the larvae sank to bottom. To remove dirt or dead individuals,
the pellet of larvae was pipetted out and filtered through a 20µm sieve submerged in water (27
°C). Clean L3s at a concentration of approximately 4,000 L3 per mL were stored in culture
flasks in water at 10 °C for several months.

During the storage, some larvae naturally die. Prior to the experiment, we filtered larvae
through a 20µm sieve (for 12 h), this time collecting both live larvae from the bottom of the
sieve and dead individuals remaining on the sieve.

All experimental procedures were examined and approved by the Ethics Committee of

**Video processing and manual counting**

Videos from a microscope camera were recorded for a duration of 10 seconds with a
framerate of 20 frames per second (fps) in a quality of 2560 by 2160 pixels. The videos (.avi
with MJPEG compression, ~ 0.45GB) were acquired using NIS-Elements Imaging Software
(version 4.20) using a Nikon Eclipse Ti microscope (4x magnification) with camera (Andor
Zyla 5.5 sCMOS, 12bit: 2560 x 2160 pixels). The videos were recorded for different ratios of
live and dead larvae (100:0, 75:25, 50:50, 25:75, 0:100; denoted as motility group 100, 75, 50,
25 and 0, respectively) in 9 replicates and they contained on average 60 worms. The live larvae
were kept in 37 °C to maintain their motility.

The video recordings were processed so that each worm was labelled with a numeric
identifier. A trained human accessed each worm from each video and information about motility
and dead phenotype was recorded. The discrepancies were dealt with as following (i) a dead
worm which was moving due to interference of a motile worm was counted as motile; (ii) a live
coil non-motile worm was counted as non-motile; (iii) the worms which were not labelled by
an algorithm were given an identifier manually; (iv) worms were excluded from the counts
when they appeared in less than 20 frames; (v) dead phenotype was evaluated when overt
microscopical pathologies, such as the presence of large vacuoles in the tegument and straight phenotype, were observed.

This manual counting was taken as a referential method for comparison of accuracy among the three models.

**Wiggle Index analysis**

We used an ImageJ implementation of the WI for the analysis [22]. We did not apply any scaling of the videos, the Gaussian image blur was set to 2, the number of frames to be averaged was set to 50 and the motility output was based on the standard deviation. Since the values from the WI are relative, we used motility group 100 as a control group to be able to convert these relative values to percentages. We calculated the mean value for the group 100 and divided all values by the obtained mean. However, after the conversion some values may be higher than 1, which in the context of motility does not make sense, therefore those values can be capped to 1 for the calculation of the error metrics.

**WF-NTP analysis**

WF-NTP presents an automated evaluation of motility in *C. elegans*. For its successful implementation, we adopted the method for the *H. contortus* motility assay. However, as these parasites differ in their sizes as well as their movement, some modifications to the original configuration and algorithm were necessary. These modifications belonged generally to two categories, where the first was a recalibration of the available input parameters of the WF-NTP method. This included aspects such as the image size, the size of the worms that we want to detect and other configurations (see Data availability). The other modifications were changes in the evaluation of the conditions for worm motility and a limit on the minimum number of frames that a worm needs to be detected in for it to be included in the evaluation.
The original method relied on bends per second of individual worms to classify them as either motile or non-motile, however as the movement of *H. contortus* is more complex, we replaced the bends per second with an evaluation metric that is regularly used in image recognition, namely Intersection over union (IoU) which quantifies the percent overlap between the target mask and our prediction output [23]. The use of IoU allowed us to use the capabilities of WF-NTP in terms of the detection of individual worm instances in an image and then apply the IoU metric to determine the worm motility. The calculation of IoU was done for the same worm instance in neighbouring frames to detect the change on a frame-by-frame basis and then the mean was taken and worms above a certain IoU threshold were labelled as non-motile. To ensure that the comparison of WF-NTP and Mask R-CNN was aimed at the quality of the detection of individual worm instances rather than the approach to determine the motility, we used the same IoU conditions for both methods.

Another modification of the original method was the introduction of a threshold for the minimum number of frames, that a worm needed to be detected in order to be included in the analysis. The reason for this restriction was that as the individual worms move, they have a tendency to either change the shape or overlap with other worms, therefore in terms of the detection the individual worm can be lost for a certain number of frames and reappear later, however this can lead to counting some instances multiple times. To limit these potential duplicities (Supplementary Fig S1) as much as possible and not bias the total number of worms detected in the video, we selected the highest number of worms detected in a given video across all frames and added a buffer of 10% in order to account for situations where some worms were not detected, for example due to their overlaps. This maximum number then served as a limit for selecting the number of worms to be evaluated as motile or non-motile. The instances were selected based on the number of frames in a decreasing order, therefore ensuring that we had
data from enough frames to classify the worm correctly. Again, this restriction was applied to
detections from Mask R-CNN as well to enable a comparison under the same conditions.

**Mask R-CNN analysis**

For Mask R-CNN we used the Matterport implementation ([https://github.com/matterport/Mask_RCNN](https://github.com/matterport/Mask_RCNN)) of the algorithm including a trained model using a backbone architecture of Resnet101. The pre-trained model weights were obtained from training the model on the MS COCO dataset [24]. We selected a pre-trained model because in general deep neural networks require a lot of annotated data for training and using an existing model and re-training the top layer of the network allows us to have a smaller dataset for the training of the model, while maintaining a good performance. We manually annotated a total of 95 images and divided them into training (69) and validation (26) using the VGG (Visual Geometry Group) Image Annotator [25], these images contained three categories of objects eggs, second-stage larvae (L2) and third-stage larvae (L3), to fully use the capabilities of Mask R-CNN which is able to do instance segmentation as well as classification of those instances. The total number of annotated objects was around 10,400 divided into training (6,800 objects) and validation (3,600 objects). Then we ran the training on a GPU instance until the loss function showed no signs of improvement. To avoid overfitting, we used data augmentation and applied one or multiple of the following augmentations: flipping, contrast normalization, additive gaussian noise, multiplying the pixels by a number within a given interval to make the whole image lighter or darker. The final model had a mAP IoU of 64.1% on the validation set.

The trained model was then used on the individual frames of the video to detect the instances of parasites and we calculated the centroid of the detected object in terms of the position on the x and y axis using the scikit-image regionProps functionality [26]. These detected instances were then linked together using the trackpy library [27], which detected the
trajectories of objects based on certain conditions, in our case we restricted the distance that the centroid could move a maximum of 100 pixels in each dimension. As a result of detecting the trajectories, we were now able to identify the same parasite instance across the frames. Subsequently, we calculated the IoU and the threshold for the number of parasites to select for the evaluation as described in the WF-NTP section. We observed a decreasing MPE with an increase in the length of the video and therefore we decided on using the full length of the videos for the motility assay (Supplementary Fig S2).

Error metrics

For the evaluation of the performance of the individual algorithms we used standard metrics for the quantification of the forecast precision. These metrics work with the error term, which is defined as the difference between the actual and the forecasted value, where in our case the actual value was obtained by manually processing the videos. Two aspects were measured, the first one was the bias of the forecasts to determine if the algorithm is systematically over or under-estimating, for that purpose we used ME or MPE, which both work with the mean of the error term. The second aspect was the size of the error term irrespective if the error term was positive or negative to compare the overall performance of the algorithms, for that purpose we used MAE or MAPE, which both work with the mean absolute value of the error term.

Additional files

Supplementary Table S1: Manual counts of motility and live rates per motility group
Supplementary Table S2: Motility rates per motility group and algorithm
Supplementary Table S3: Mean error between the manual count and the respective motility algorithm
Supplementary Table S4: Mean of differences in worm counts between the manual processing and the respective algorithm.

Supplementary Fig S1: Average number of worms in the videos per motility group for WF-NTP and Mask R-CNN.

Supplementary Fig S2: The impact of the length of the video in frames on the percentage error of the detected worms for Mask R-CNN.

Abbreviations

CNN: convolutional neural network; fps: frames per second; GPU: graphics processing unit; IoU: intersection over union; MAE: mean absolute error; MAPE: mean absolute percentage error; Mask R-CNN: region based convolutional neural network; ME: mean error; MPE: mean percentage error; WF-NTP: Wide Field-of-View Nematode Tracking Platform; WI: Wiggle Index

Competing Interests

The authors declare that they have no competing interests.

Funding

This work was supported by Charles University UNCE/18/SCI/012 and SVV 260 550 as well as by the project EFSA-CDN [CZ.02.1.01/0.0/0.0/16_019/0000841], co-funded by ERDF.

Authors’ contributions

M.Ž. and L.T.N. designed this study and wrote the manuscript. M.Ž. implemented the deep learning model, wrote the code, ran the statistical analyses and generated the figures. L.T.N., E.M. acquired and annotated the images used to train the deep learning model. M.Ž. and L.T.N.
conducted the experiments and analyzed the results. P.M. managed the project. All authors contributed to literature review and critical revision of the manuscript.

Acknowledgements

We thank MVDr. Tomáš Kaňka for assisting with the sheep infection. The authors thank prof. Lenka Skálová for critical discussion and comments.

References


Figure 1
Figure 2: Box plots showing worm count across different motility groups for Manual, WF-NTP, and Mask R-CNN methods.
Figure 3