Critical Paper Review for Computer Integrated Surgery II

Henry Phalen

Group 1

April 30, 2019

Multi-Mosquito Object Detection and 2D Pose Estimation for Automation of PfSPZ Malaria Vaccine Production [1]

Hongtao Wu, Jiteng Mu, Ting Da, Mengdi Xu, Russell H. Taylor, Iulian Iordachita, and Gregory S. Chirikjian

Note: The paper reviewed in this report is currently under review for CASE 2019. Due to this submission and pending filing of IP, the contents of the paper and this review are confidential until its disclosure, either at the conference on August 22, 2019 or some later date.

Overall Paper Summary:

This paper describes the development, testing, and comparison of two computer vision algorithms for the purpose of detecting the location and orientation of mosquitoes. This task is intended to be used in concert with an automated mosquito dissection system to aid in the production of malaria vaccines. The authors tested a more conventional image processing pipeline that identifies mosquito location by locating the head as an approximate circle using a Hough transform. They compared this with a deep-learning approach which used about 1200 training images. The authors purport their deep-learning solution yields better accuracy, albeit at what they deem an acceptable 8-fold reduction in processing speed.

Paper Selection / Relevance to Project:

My group is working with the authors of this paper as part of a larger effort at Johns Hopkins' Laboratory for Computational Sensing and Robotics (LCSR) in collaboration with Sanaria Inc. to develop a semi- or fully-automated mosquito dissection system. Specifically, our group is developing the mechatronics systems for manipulating the mosquito: a robotic pick-and-place system that grabs mosquitoes from a staging system and deposits them onto an assembly line where they can be processed for gland extraction. The literature in this specific area is extremely sparse. Very few outside our group have published regarding the manipulation of insects, let alone mosquitoes, robotically. One potential paper for review was on the robotic handling of silkworms [2], but the soft robotics solution here was trying to solve a completely different, non-destructive, live-organism, macro-scale handling task different from our dissection goal. A group at Harvard previously began working with Sanaria on trying to solve this problem, but they never published solutions beyond a short promotional video [3]. While Sanaria has published many papers on the *Plasmodium falciparum* sporozoite-based vaccine (PfSPZ), (e.g. [4-6]), these also have little relevance to my technical involvement in the project. I chose this specific paper out of those written by members of our research team to review as I worked directly with the main author to integrate the deep-learning method with our robotic system.

Content Summary and Review:

I. Motivations & Related Works

The paper begins by describing the motivation for the larger project goals, citing statistics to make the case that malaria is a global health crisis and a tremendous humanitarian and financial burden: over 200 million cases annually with some $3.1 billion in cost. This description of broad motivation is an important start to any research paper, and certainly one related to the medical field. The authors spend perhaps a bit too long arguing their way into the need for a malaria vaccine, citing other control measures and stagnating reductions in global cases numbers. I think the case makes itself through the statistics without the extra effort.

The authors transition into making a case for their technology. While members of the LCSR have worked on a mechanical solution that improved production efficiency, in order to automate any system, computer vision is needed, they argue. A robotic system needs to be able to recognize mosquitoes in order to manipulate them. They present this as a two-fold problem. First, a test must occur for individual mosquitoes as they suppose the mosquitoes may become tangled up and unable to be grasped safely. Then, the individual parts of each pickable mosquito must be recognized. They suggest that their implementation of two specific networks, Mask R-CNN and DeeperCut well-address these problems.

This begins a long discussion on the merits of deep learning in computer vision versus more traditional methods, as well as the history of many computer vision techniques used in everything from insect detection to human pose differentiation. While the first distinction may be worth mentioning given the use of both kinds of methods, they never ultimately address the concerns they provide for deep learning systems (namely the 'black box' nature of them), so such a discussion could be removed for more specific details later. I do appreciate the obvious research the authors conducted into others' work in the field, but I think they stretch this section further than is needed, again limiting their space for more technically relevant discussions later.

II. Methods

Two techniques are described, a deep-learning approach and a so-called traditional image processing approach. For clarity, I have included a figure from the paper demonstrating the physical setup used in data collection (Figure 1). A camera captures the mosquitoes from above. They are situated on a mesh surface, in the project often referred to as a "cup". Mosquitoes in fluid will run down a staging apparatus over this mesh, leaving them in position to be viewed by the camera. On this cup, they can be aligned so that they are easy to pick up robotically as part of the automated dissection process. As a critique, I would note that the staging apparatus and the computer vision's direct role in the automated process is not well-explained in the paper. While I have been able to add commentary here from my personal knowledge based on my involvement in the project, the average reader may be confused by the objects they see in this figure.

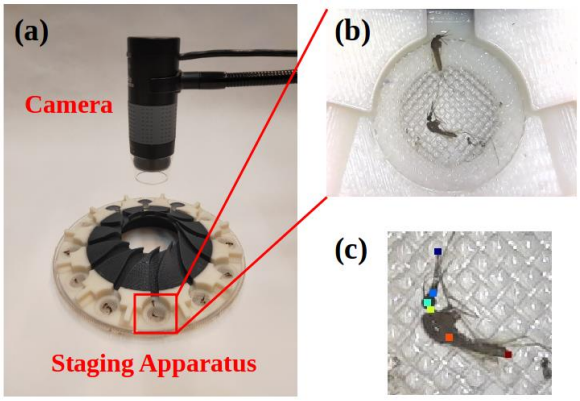

Figure 1. This figure shows the experimental setup. A camera looks down onto mosquitoes on a mesh-covered cup from above. These images are from Fig. 3 in [1].

a) Deep Learning Approach

The deep learning solution used a network model designed by the AI Research Group at Facebook called the Mask R-CNN which stands for mask region convolutional neural network. This first identifies regions of interest (ROI) in the image, then uses a convolutional neural network to classify the object in the region. The Facebook team demonstrated this by identifying people, animals, or objects in images (Figure 2). In this paper, it is used to classify objects as either separated or clustered mosquitoes, and provides the bounding boxes for free, which are used in the next step.

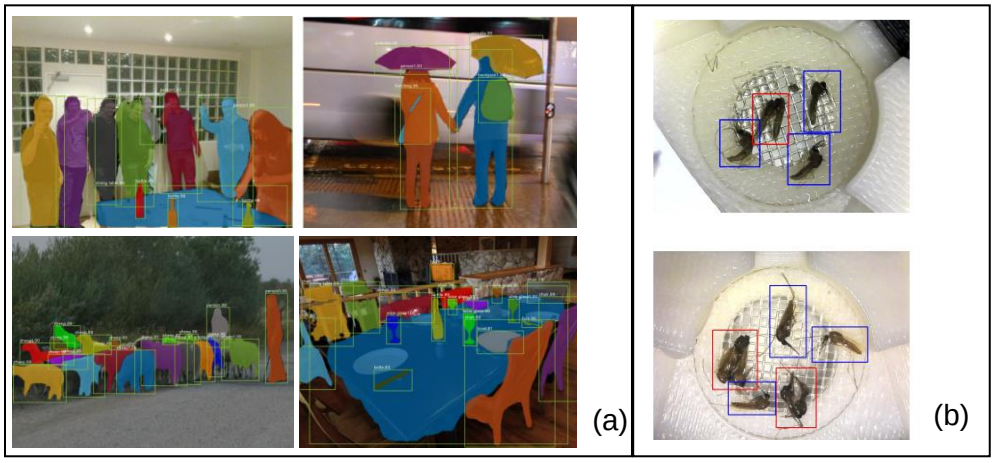

Figure 2. (a) Mask R-CNN examples provided in [7] (taken from Fig. 2). Subfigure (b) shows the implementation used in [1] (taken from Fig. 2). Blue bounding boxes surround separated mosquitoes while red surround clustered ones.

The ROIs that were determined to contain separated mosquitoes (i.e. not clustered) are fed into another network: DeeperCut. This network was first developed for human pose detection, though the authors note it has also been extended to mice and *Drosophila* (fruit flies) by previous researchers. It appears that specifically, the open source package DeepLabCut [9] was used, which employs a residual neural network where the output provides the most likely location of a given body feature. This package allowed the authors to employ a technique called transfer learning.

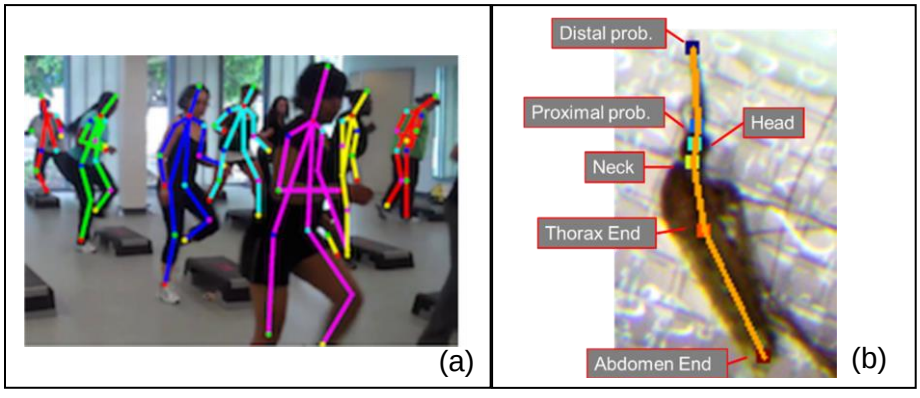

Figure 3. (a) DeeperCut example provided in [8] (taken from Fig. 6) Subfigure (b) shows the implementation used in [1], though I made this figure myself using data I collected with their algorithm. The 6 keypoints considered are labeled.

Essentially, a model that is already well-trained for some specific subject (e.g. human or fruit fly) can be adapted to work for another subject (e.g. mosquito) requiring much less training data. This seems to work because some early layers in these networks can be thought of as more “general”, while later ones are more “specific”. If you replace or

retrain some of these more specific layers, you can achieve good performance with relatively low training data. In this paper, 1460 images were used for training, validation, and testing in a 7:2:1 split, a fairly low number as the pre-trained Mask R-CNN and DeeperCut networks were used excluding their final layer or “head” to initialize the training process. The network architectures are given in Figure 4.

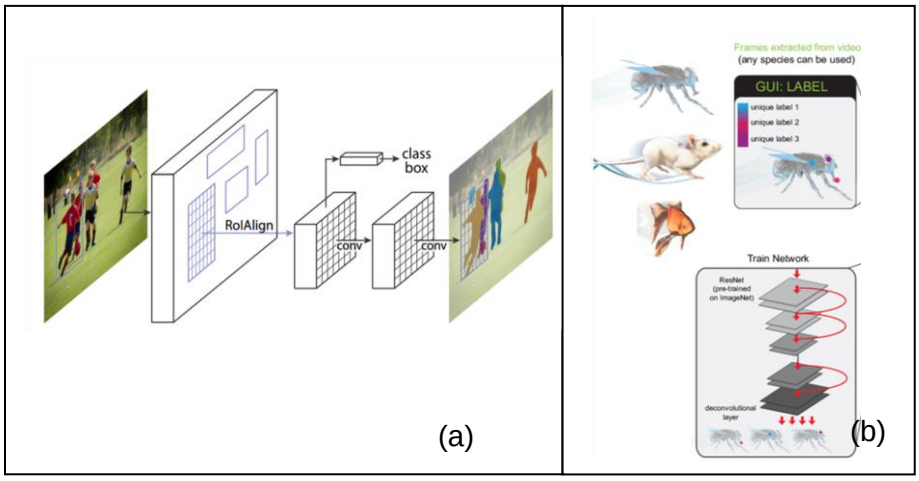

Figure 4. Network architectures (a) Mask R-CNN (from [7]) and (b) DeepCut (from [9]). The authors do not appear to use the masking results from the Mask R-CNN.

Great effort is made to establish that a diverse training set was used, which the authors felt would contribute to system robustness. They varied lighting conditions, the flow rate of fluid in which the mosquitoes were deposited onto the substrate they were imaged from (to vary mosquito presentation), the distance of the camera to the mosquitoes (to vary their size in the images) and they included both clustered and separated mosquitoes. This dataset creation is quite insightful and very important in a situation like this where you want to prevent overfitting. Having to retrain an entire system just because you moved a camera closer to the mosquitoes for example, would be burdensome. I do think the authors missed out on the opportunity to do some analysis of the effects of these variations, however. I would have like to see them evaluate accuracy within each of these subgroups and use that as an indicator for robustness and to validate that they had enough sample images for each scenario. They also provide no detail of these variables and to what degree they were modulated. Even pictures of a few scenarios are not included.

b) “Traditional” CV Approach

The traditional method is best described using Figure 5. The technique was to first apply a threshold to detect the dark objects in the image, then separate into potential mosquitoes using the watershed algorithm. These regions of interest were evaluated against assumed ranges for size to remove clustered mosquitoes. Then the head, the most basic and consistent part of the mosquito is found using a Hough circle transform. This is not straightforward due to other complex shapes being picked up falsely, so first the centroid of each region is found, an erosion process is started, and the mosquito body is split into two parts that can be classified into head versus thorax with a density

classification algorithm DBSCAN. Finally, the proboscis, which is assumed to appear as a line is searched for using the Hough line transform.

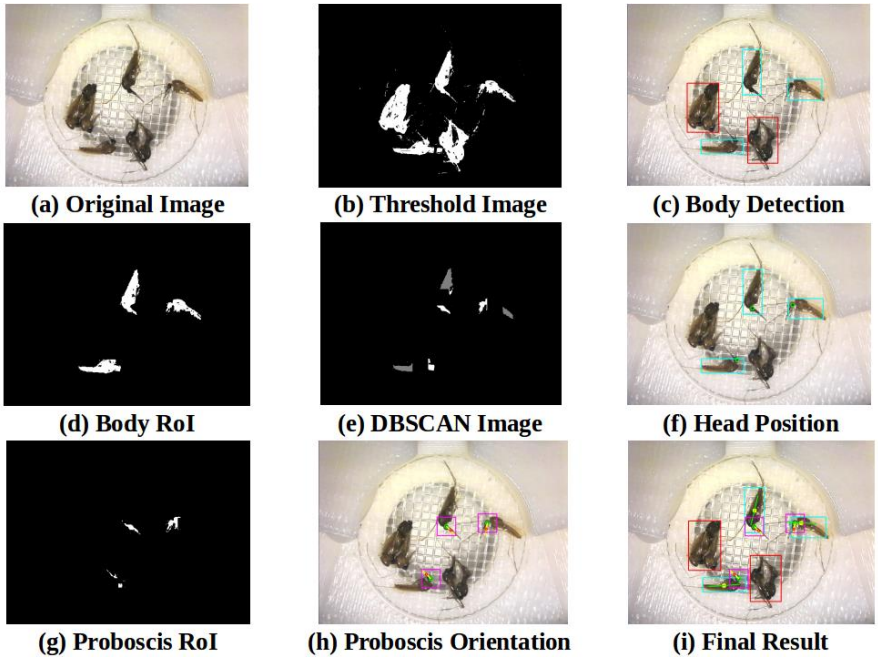

Figure 5. Taken from [1], this shows the methods used for the traditional CV approach.

III. Results & Conclusions

The authors provide some standard metrics for their results in classification (precision-recall curve, mean average precision) and for accuracy of their key point detection (root mean square error (RSME)). I was disappointed to not see a confusion matrix. I think these are the most efficient way for a reader to quickly understand a classifier's overall performance. Additionally, it is quite difficult from the paper to understand the meaning of the presented results. There is no real mention of requirements, and all accuracies are given in pixels, not distances. If you dig back through the paper and do some math with the values they present, you can come up with the estimated pixel-to-distance conversion they had of 50 microns per pixel. They should have offered these real-world values to the reader in the table copied in Figure 6.

The authors chose in several parts of the paper to make claims of system robustness yet provide no evidence of this. The argument is that if the model was trained with a variety of images, it should be able to detect well in all those conditions. However, to provide an even comparison with the traditional method, which must have thresholds tuned at different camera positions and luminosities, the evaluation of that network was only done at these constant values as well. While this is important for the comparison (e.g. the table in Figure 6), it does not demonstrate that the model is robust. For all the reader can tell, the model is applicable only at those luminosities and distances as well, which would undermine on of the authors main claims for its superiority over the traditional method.

TABLE I
PERFORMANCE COMPARISON ON TESTING DATA WITH FIXED LUMINANCE CONDITION AND MOSQUITO SCALE

Approach	Deep Learning	Image Processing
Detection mAP (IoU>0.5)	0.97	0.80
Detection Recall	0.97	0.90
Head Position RMSE	1.61 pixels	2.70 pixels
Proboscis Orientation Error	14.3°	24.7°
Processing Speed	2.5 fps	20 fps

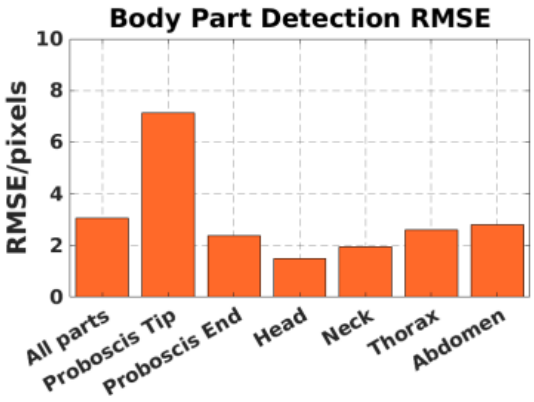

Figure 6. Main results from [1] showing that the deep learning method provided better precision, recall, and error results, while the traditional approach was about eight times as fast.

Ultimately, the conclusion, while not strongly stated, is that the deep learning method was superior to the traditional method. It certainly has better precision, recall, and error performance as presented, though the deep net solution is eight times slower than the traditional method. Additional discussion of this trade-off could have been provided. A previous discussion of the engineering requirements would have facilitated the reader in understanding this conclusion.

Summary of my Review

Overall, I think this was a good conference paper. I liked that the authors considered both types of methods and did not jump straight for a deep learning approach. The primary contribution to the field is that these methods had never been performed on mosquitoes before. I think the research described in this paper will be most useful as a tool in our project for mosquito location. It was obvious that the authors did a good literature review and made great use of the open source code in the literature through transfer learning techniques.

There were a few items in the paper that could have been improved. As a reader it is always very frustrating when you are forced to do the math yourself to convert something like a pixel accuracy to distance accuracy when the latter is more relevant. I also think some discussion of the system requirements would have been useful so that the decision regarding which properties were favorable in the evaluation could be better understood. I would recommend spending a bit less time arguing about need for malaria vaccine or describing history of computer vision early on, the authors would have had more space to provide technical details about the methods they used and decisions they made in designing their networks or picking evaluation criterion. Finally, while a head-to-head comparison of the two methods is important for comparison, the robustness claims about the deep learning solution are entirely unproved because of this. The authors might consider adding an analysis of just this network over many conditions in a future paper.

References:

[1] H. Wu, J. Mu, T. Da, M. Xu, R. H. Taylor, I. Iordachita, and G. S. Chirikjian, “Multi-mosquito object detection and 2D pose estimation for automation of PfSPZ malaria vaccine production,” 2019, Submitted to CASE 2019.
[2] Ohura, Masanobu, et al. "Development of a silkworm handling robot." 2005 ASAE Annual Meeting. American Society of Agricultural and Biological Engineers, 2005.
[3] Sanaria Inc. (2014) SporoBot - Build a Robot. Fight Malaria. Save Lives! [Online]. Available: https://www:youtube:com/watch?v=VblazNXcHFg
[4] K. E. Lyke, A. S. Ishizuka, A. A. Berry, S. Chakravarty, A. DeZure, M. E. Enama, E. R. James, P. F. Billingsley, A. Gunasekera, A. Manoj, et al., “Attenuated PfSPZ vaccine induces strain-transcending t cells and durable protection against heterologous controlled human malaria
infection,” Proceedings of the National Academy of Sciences, vol. 114, no. 10, pp. 2711–2716, 2017.
[5] B. Morduller, G. Surat, H. Lagler, S. Chakravarty, A. S. Ishizuka, A. Lalremruata, M. Gmeiner, J. J. Campo, M. Esen, A. J. Ruben, et al.,
“Sterile protection against human Malaria by Chemoattenuated PfSPZ vaccine,” Nature, vol. 542, no. 7642, p. 445, 2017.
[6] T. Bousema and C. Drakeley, “Epidemiology and infectivity of plasmodium falciparum and plasmodium vivax gametocytes in relation
to malaria control and elimination,” Clinical microbiology reviews, vol. 24, no. 2, pp. 377–410, 2011.
[7] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” in IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–
2988.
[8] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, “Deepercut: A deeper, stronger, and faster multi-person pose estimation model,” in European Conference on Computer Vision (ECCV), 2016, pp. 34–50.
[9] Online. https://alexemg.github.io/DeepLabCut/