2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)

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

Hongtao Wu, Jiteng Mu, Ting Da, Mengdi Xu, Russell H. Taylor, *Life Fellow, IEEE*, Iulian Iordachita, *Senior Member, IEEE*, and Gregory S. Chirikjian, *Fellow, IEEE*

***Abstract*—Multi-mosquito object detection and 2D pose estimation are essential steps towards fully automated extracting PfSPZ-infected mosquito salivary glands for manufacture of PfSPZ Vaccine, which has been shown to protect against malaria in multiple clinical trials in the US, Europe, and Africa. This paper presents a deep learning approach to perform cluster condition classification and bounding box detection of multiple mosquitoes in an image. It also estimates the 2D pose of each non-clustered mosquito by body part detection. This approach is based on two popular convolutional neural network (CNN) architectures, Mask R-CNN and DeeperCut. In addition, we propose a cascaded image processing approach to achieve the multi-mosquito detection, cluster condition classification, and body parts detection in a multi-step manner. We compare the two approaches in terms of their functionality, robustness, accuracy, and speed. We hope our effective approaches would push forward the automation of PfSPZ Vaccine production to facilitate the prevention and elimination of this disease worldwide.**

I. Introduction

The World Health Organization (WHO) estimates that cases of malaria worldwide have increased from 214 million in 2015 to 219 million in 2017, resulting in an estimated 435,000 deaths in 2017 alone [1]. Malaria hinders the economic growth within endemic regions [2]. It is estimated that malaria causes 1.2 billion US dollars loss every year in Africa due to healthcare expense, loss of labor, and negative impacts on tourism [3]. Despite a considerable investment of 3.1 billion US dollars in 2017, the number of malaria cases has not reduced, but rather increased from 2016 to 2017. Although control measurements have been taken and substantially decreased malaria morbidity and mortality [1], there is still an urgent demand for malaria vaccines as highly effective preventative measures to facilitate the elimination of this disease within high transmission regions and worldwide.

The Sanaria *Plasmodium falciparum* (Pf) sporozoite(SPZ)-based vaccine (Sanaria® PfSPZ Vaccine, hereafter referred to as PfSPZ Vaccine) has proven to provide significantly durable protection against infection with Pf, which is responsible for more than 98% of the deaths caused by malaria annually [4]–[6]. PfSPZ Vaccine is manufactured from PfSPZ extracted from the salivary gland of the infected mosquito.

H. Wu, J. Mu, T. Da, M. Xu, R. H. Taylor, I. Iordachita, and G. S. Chirikjian are with the Laboratory for Computational Sensing and Robotics (LCSR) at the Johns Hopkins University, Baltimore MD, USA. T. Da is also with the Xian Microelectronics Technology Institute, Xian, China. G. S. Chirikjian is also with the Dept. of Mechanical Engineering at the National University of Singapore, Singapore. G. S. Chirikjian is the corresponding author. {mpegre}@nus.edu.sg

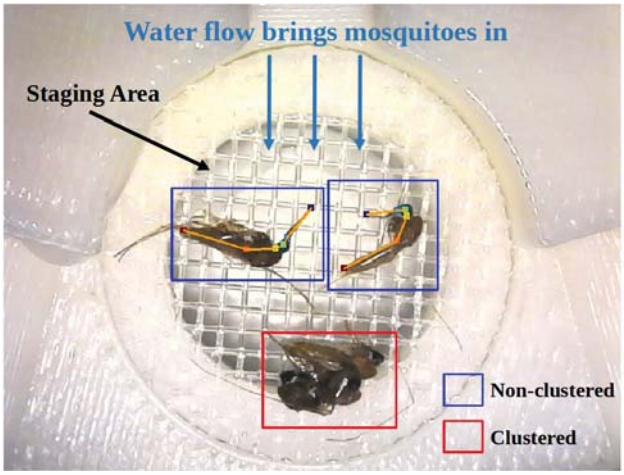

Fig. 1. Multi-mosquito object detection and body part detection for automation of PfSPZ Vaccine production. Water flow brings mosquitoes into the staging area. The blue bounding box refers to the non-clustered mosquito, while the red bounding box refers to clustered mosquito. For non-clustered mosquitoes, six body parts (proboscis tip, proboscis end, head, neck, thorax end, abdomen end) are detected for pose estimation. They are shown in colored dots. The dots are connected with orange lines to show the pose estimation. With the information of cluster condition and body part position, pick and place robots would be able to pick up the non-clustered mosquito at its proboscis. The clustered mosquito, which is non-pickable, would be flushed back to the water.

In the current manufacturing process, the salivary gland is extracted from the mosquito manually using hand tools under a microscope [7]. This process is tedious and labor-intensive, and our research team at Johns Hopkins has been working with Sanaria to improve the efficiency of this process. As a preliminary step, Schurm et al. developed a semi-automated Mosquito Micro-dissection System (sAMMS) [8], in which mosquitoes are grasped by the proboscis and placed manually into cartridges for further processing. Although the sAMMS improves the throughput of manual processing and greatly reduces the training time for technicians, we are currently developing a more automated system for very large scale production.

To automate the micro-dissection process of the mosquito with our next-generation MMS, it is necessary to firstly detect the location of the mosquito and recognize if it is clustered with others. During storage, the mosquitoes are preserved in water. Before the pick and place step, mosquitoes are extracted from the water to the staging area. Unfortunately, sometimes two or more mosquitoes are stuck together when

978-1-7281-0356-3/19/$31.00 ©2019 IEEE

they arrive at the staging area. We call these clustered mosquitoes (see Figure 1). These clustered mosquitoes cannot be picked without damaging the body parts. The second key component is to estimate the body pose of the non-clustered mosquito by body part detection for the robot to pick and place. Since the whole body of the mosquito would be lying flat on the stage when it arrives due to the down flowing fluid (the fluid would flow down through the mesh in Figure 1), detecting the 2D pose instead of the 3D pose would be enough for the robot to localize the mosquito for pick and place.

To address the above problem, we demonstrate a deep learning approach which is able to effectively perform bounding box detection and cluster condition classification of multiple mosquitoes as well as estimating the 2D pose of the non-clustered mosquito (Figure 1). Our approach capitalizes on the state-of-art deep learning methods for object detection and pose estimation. Specifically, we firstly adopt the popular object detection network architecture called Mask R-CNN to perform bounding box detection and cluster condition classification of the mosquito [9]. DeeperCut, a network architecture developed for human pose estimation, is then used to estimate the pose of the non-clustered mosquito by body part detection in each bounding box [10], [11]. We are able to achieve excellent bounding box detection, cluster condition classification, and keypoint detection of the mosquito in laboratory setting with only 1,168 training images.

Though deep learning methods have great performance in many computer vision tasks, they are criticized for their black-box nature. On the contrary, image processing techniques are fully explainable and supported by solid mathematical proofs. Noticing that in the manufacturing setting, the background and the luminance condition is relatively fixed, we propose a cascaded image processing approach to localize the mosquito, classify its cluster condition and estimate its 2D pose by localizing the body part in a multi-step manner. We compare and analyze our two approaches in terms of functionality, accuracy, robustness to environments, and processing speeds.

The rest of the paper is organized as follows. Section II overviews the related works on object detection and body part detection in the context of insects and generally. Section III describes the two methods that we use to detect the mosquito and estimate its pose. Section IV reports our experimental results. In Section V, we conclude our work and discuss the significance of our contribution and future development for improvement.

II. Related Work

Plenty of previous works have been proposed to apply computer vision to detect insects. Fuchida et al. [12] develop a support vector machine (SVM) to distinguish mosquitoes from other species using morphological features. Huang et al. [13] use edge computing and CNN to identify two types of mosquito. Ding and Taylor [14] propose an automatic detection approach based on deep learning for identifying and counting pests in images taken inside field traps. Liu et al. [15] present a pipeline for the visual localization and classification of agricultural pest insects using the saliency map and CNN, respectively. Wen et al. [16] segment moth from field images and use deep neural network to identify the moth species. DeepLabCut, developed by Mathis et al., can perform markerless pose estimation of user-defined body parts by using the Deepercut as feature detector [17]. It has achieved excellent body part detection results on mouse and Drosophila image data by employing the transfer learning technique. Our work on mosquito detection differs from the above in that we combine localization of the mosquito, classification of cluster condition and multi-mosquito body part detection together to provide vision algorithms which facilitate the automation of the MMS-based malaria vaccine production. Besides, we present an image processing method for comparison.

In computer vision, various methods have been developed for object detection and body part keypoint detection. Before the advent of the deep learning era, one popular way of detection is the sliding window approach in which a window slides across the whole image to identify the target region. However, the sliding window approach is not invariant to diverse sizes and orientations. Another line of work performs object detection by segmentation. However, neither of the two approaches can perform classification for images. For keypoint detection, many image processing methods follow a feature extraction and feature matching pipeline. Engineering feature descriptors, such as SIFT [18] and SURF [19], have been successfully applied to various tasks. To locate the proboscis, one natural way is by finding its ending points. However, for our application, as legs and proboscises share similar line shape features, directly extracting keypoints by feature extraction can be problematic. Therefore, in our image processing approach, instead of locating the keypoint of the proboscis, we take it as a shape matching problem and adopt a multi-step approach to locate head first and then detect the proboscis orientation in a small neighborhood of head. With the boom of deep learning, R-CNN-based architectures have been successful in object detection tasks by attending to a manageable number of regions of interest (RoI) and employs CNN independently on each of them [9], [20]–[22]. Also, many network architectures have been reported for human pose estimation [10], [11], [23], [24]. However, all these works are trained and tested with massive datasets, e.g. MS COCO dataset and MPII Human Pose, and none of them have focused on insect detection. In our deep learning method, we only trained and tested with 1,460 images of mosquitoes by applying the transfer learning technique [25], [26].

III. Methods

Figure 2 and Figure 4 illustrate the pipelines of our two methods. For the non-clustered mosquito, we are especially interested in the position of its proboscis, its head and its neck. The reason is as follows: to decapitate the mosquito in the micro-dissection manufacturing process, the exact

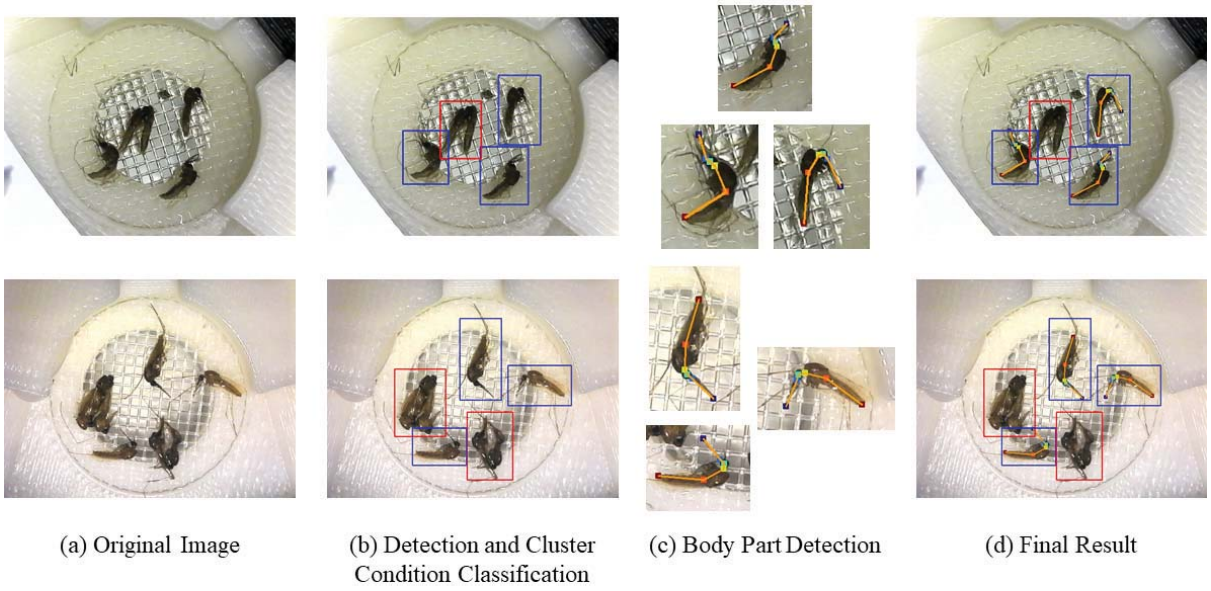

Fig. 2. Overall pipeline of the deep learning method. The original image (a) is firstly input into Mask R-CNN for cluster condition classification, shown in (b). Non-clustered mosquitoes are labeled with blue boxes while clustered mosquitoes are labeled with red boxes. Each non-clustered mosquito is then input into DeeperCut for body part detection, shown in (c). We finally assemble (b) and (c) to generate the object detection and pose estimation for all mosquitoes, shown in (d).

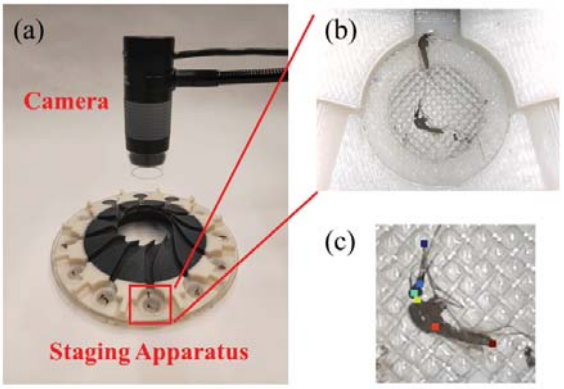

Fig. 3. Data Collection. (a) We use the staging apparatus of the MMS to collect data. (b) The image captured by the camera covers the whole area of the staging area (the circle mesh). The image resolution is 640×480. (c) The body part labels are shown in different color dot. We label the proboscis tip (dark blue) and end (blue), the head (green), the neck (yellow), the thorax end (orange), and the abdomen end (dark red) for each non-clustered mosquito.

position of the neck and head are needed; to grip and manipulate the mosquito, the position and orientation of the proboscis is desired. The proboscis is the best body part for the pick and place robot to grip because gripping any other body part would risk damaging the salivary gland.

A. Deep Learning Approach

Dataset: To train the networks which work for the real manufacturing setting, we collected the data with the mosquito staging apparatus of the MMS (Figure 3) [27]. We collected images in which the mosquito was randomly positioned on the staging area of the apparatus manually and images in which the mosquito was transferred from the water by the apparatus automatically. To ensure robustness against uncertain water flow and luminance conditions, we varied the luminance and water flow conditions. We included images with clustered and non-clustered mosquitoes. The distance between the camera and the staging area is also varied to obtain images with different scales of the mosquito. In total, we collected 1,460 images and labeled the cluster condition, the bounding box for each mosquito in the image with the VGG Image Annotator [28]. For the non-clustered mosquito, we also labeled the body part location (Figure 3(c)).

Mosquito Detection and Cluster Condition Classification: To localize the mosquito and classify its cluster condition, we adopt the popular neural network architecture, Mask R-CNN. Mask R-CNN has achieved state-of-the-art accuracy on object detection benchmarks with a fast processing speed [9]. Here, we adopt the same two-stage procedure of the Mask R-CNN architecture. The first stage, the Feature Pyramid Network (FPN) backbone [29], is responsible for feature extraction over the entire image. For the second stage, the network head, we adopt the Mask R-CNN branches for class and bounding box prediction. Combining the backbone and the head gives excellent and fast prediction of the location and cluster condition of each mosquito in an image (Figure 2(b)).

Body Part Detection: After we localize and classify all mosquitoes in the image, we perform pose estimation for the non-clustered ones (Figure 2(c)). We adopt a network architecture designed for human pose estimation, DeeperCut [11]. Besides achieving state-of-the-art results in human pose estimation, Deepercut has proven to be also effective in animal body part detection [17]. The network backbone

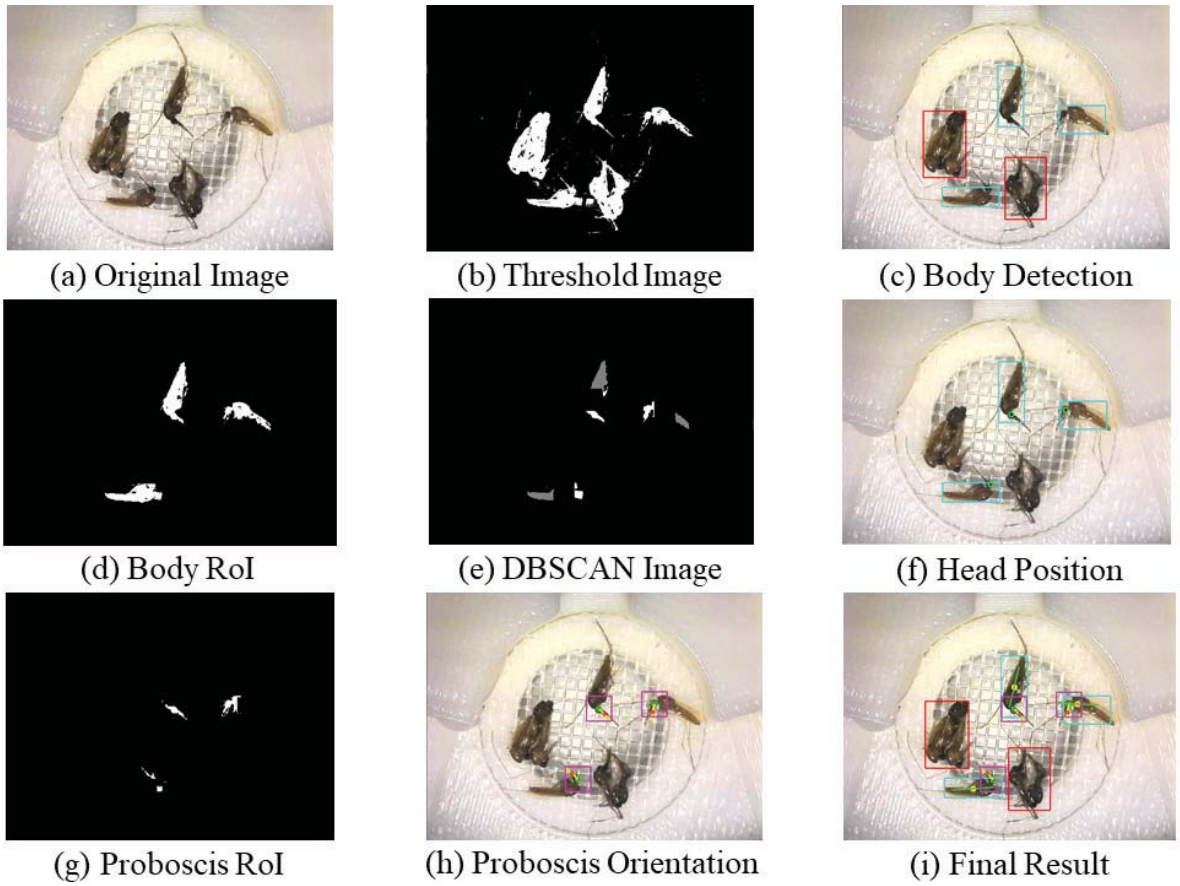

Fig. 4. Overall pipeline of the cascaded image processing method. The threshold image (b) is passed to the watershed algorithm to propose body RoIs. The algorithm removes false positives (red boxes in (c)) and propose several non-clustered candidates for further processing (cyan boxes in (c)). For each non-clustered body RoI, as shown in (d), Body Removal erodes the thorax. DBSCAN identifies the abdomen of each mosquito, as shown in (e) in grey, and excludes them for head detection. Then HCT is applied to each mosquito proposal to find heads, which are circled in green as shown in (f). HLT is employed to detect the proboscis orientation in each head-centered proboscis RoI, shown in (g). Head-centered RoIs and proboscis orientations are identified with purple boxes and yellow lines in (h). We finally assemble all the above results in (i).

is built based on the ResNet. The network head removes the final classification and adds deconvolutional layers to predict the spatial probability density of each body part. At the last stage of the network, it also performs location refinement to refine the accuracy of the body part detection. More details about the network architectures can be referred to [10], [11].

B. Cascaded Image Processing Approach

The cascaded image processing approach takes a multi-step approach to localize the mosquito, classify its cluster condition, and detect its head position and proboscis orientation sequentially.

Mosquito Detection and Cluster Condition Classification: The basic structure for detecting the mosquito is the watershed algorithm [30]. The watershed algorithm can deal with overlapping on some level without too much additional computational cost. We apply the watershed algorithm on the threshold image (Figure 4(b)) and this proposes several candidate regions. For each identified region returned by the watershed algorithm, the region area and aspect ratio are employed to remove the false positive, i.e., the clustered mosquito (red boxes in Figure 4(c)). Since the watershed algorithm processes the image based on the grayscale of each pixel, a good contrast between the mosquito and the background is required. The body orientation can be obtained by calculating the second moment within each body RoI (green lines in Figure 4(i)).

Head Detection: Noticing the mosquito's head is circular and usually darker than other body parts, we implement Hough Circle Transform (HCT) [31] to detect the head position. However, directly applying HCT to the body RoI (Figure 4(d)) is problematic because the complex curvatures would induce lots of false detections. Therefore, for each mosquito, we further implement two methods, Body Removal and DBSCAN [32], to refine the RoI.

The goal of Body Removal as its name suggests is to remove the body. The result is shown in Figure 4(e). We first apply distance transform to the body RoIs (Figure 4(d)) to find the center region of the body and then erode based on that region for a certain amount to get an erosion image. Figure 4(e) is the difference between the body RoI (Figure 4(d)) and the erosion image. It can be seen that the thorax part has been removed. Each mosquito is separated into two parts,

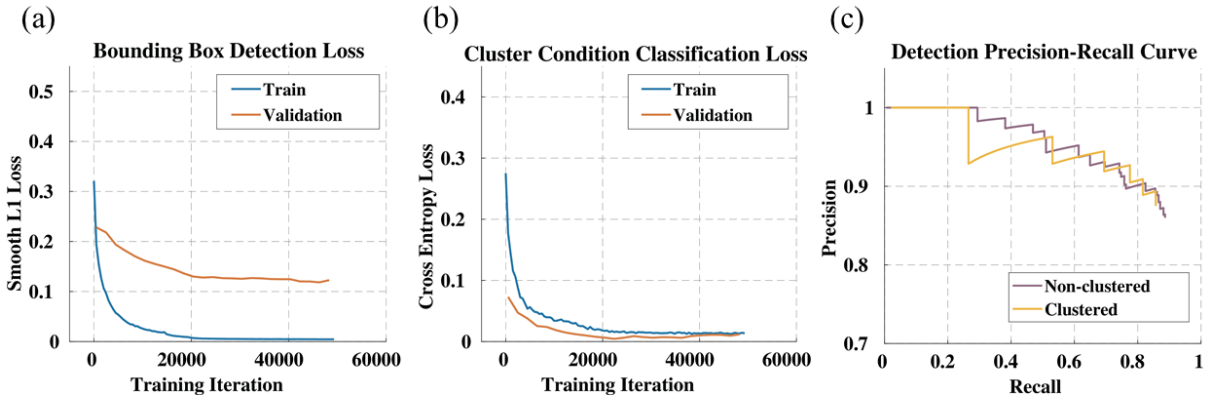

Fig. 5. Mosquito detection and classification training and testing results. The mosquito detection and classification network is trained with 49800 iterations. (a) Mosquito bounding box detection smooth L1 loss. (b) Mosquito cluster condition classification cross entropy loss. (c) Mosquito detection precision-recall curve (IoU>0.75).

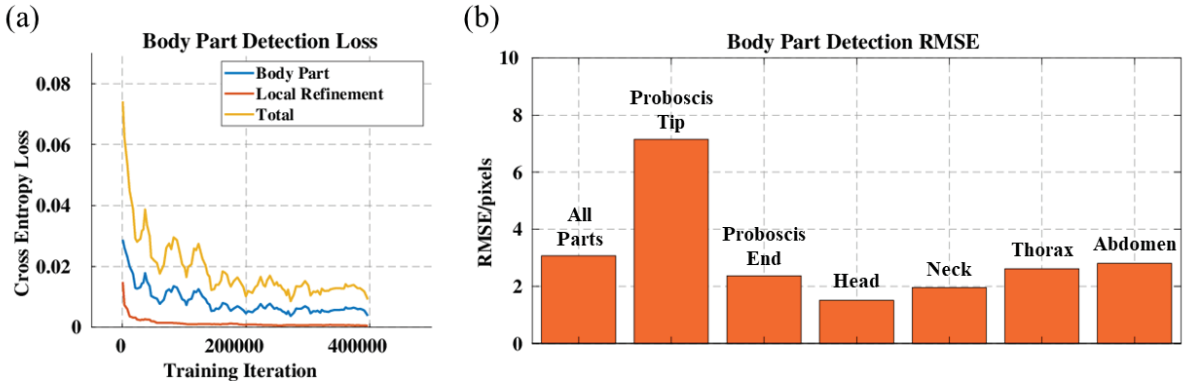

Fig. 6. Mosquito body part detection training and testing results. The body detection network is trained with 400000 iterations. (a) Mosquito body part detection cross entropy loss. (b) Mosquito body part detection root mean square error loss (RMSE) for all body parts and each body part.

the head part and the abdomen part. DBSCAN then comes into play and clusters remaining regions into the head cluster, shown in white in Figure 4(e), and the abdomen cluster, shown in grey in Figure 4(e). This allows us to eliminate all false positives in the abdomen part, where most false detection occurs. The final step is to implement HCT which will return the most likely head locations.

Proboscis Orientation Detection: After heads are detected, we define a new RoI for the detection of the proboscis orientation (purple boxes in Figure 4(h)). Noticing that the proboscis shape often appears in a straight line, the Hough Line Transform (HLT) is used for the detection [33]. For the non-clustered mosquito, the proboscis is always connected to the head. Therefore, we further calculate distances from all line candidates to the head to remove the false line proposal. Among all remaining line candidates, we select the line whose endpoint is furthest from the head and connect the endpoint back to the head to detect the proboscis orientation (Figure 4(i)).

IV. Experiments

A. Deep Learning Approach Implementation

The dataset is split by a ratio of 7:2:1 for training, validation and testing, respectively. In our pipeline, we detect the body part after the bounding box detection and cluster condition classification. Therefore, we use cropped images of all the non-clustered mosquito from the original image with body part labels to train the body part detection network. In [9]–[11], the two networks are both trained on massive datasets. However, in our case, it would be labor-intensive and unnecessary to build a dataset containing hundreds of thousands of mosquito images. We notice that the background and the luminance condition are relatively simple and invariant in the manufacturing setting. It inspires us to employ transfer learning techniques and used models pretrained on massive datasets for the feature detector of the two networks. Specifically, we initialize the feature detector of the mosquito detection network with the model pretrained on the COCO dataset and the feature detector of the body part detection network with the model pretrained on ImageNet dataset. The heads of the two networks are then trained based on the mosquito dataset. The whole networks are fine tuned in

the last stage of training. Both networks are trained on an NVIDIA GeForce RTX 2080 Ti GPU.

B. Deep Learning Approach Result

The bounding box detection and classification results are shown in Figure 5. For mosquito bounding box detection and cluster condition classification, given all predicted regions in an image, we employ non-maximum suppression which assigns the label with a higher score if the intersection-over-union (IoU) is larger than 0.75. The precision for the non-clustered mosquito class and the clustered mosquito class is 0.86 and 0.88, respectively. The average precision (AP) for the non-clustered mosquito class and the clustered mosquito class is 0.85 and 0.82, respectively. The mean average precision (mAP) is 0.84. The mAP for the case of IoU$>$0.5 is 0.96.

The body part detection results are shown in Figure 6. We measure the root mean square error (RMSE) for all body parts and each of them alone. The RMSE of all body parts is 3.1 pixels. Considering manufacturing the PfSPZ Vaccine, we are especially interested in the position of the proboscis, the head, and the neck. The detection of the head and the neck is relatively accurate, with an RMSE of 1.5 and 2.0 pixels, respectively. The RMSE of the detection of the proboscis tip is 7.1 pixels and 2.3 pixels for the proboscis end. The reason of the relatively low keypoint detection accuracy of the proboscis tip is that the proboscis' line shape is very similar to that of the leg and antennae. However, in general, considering that the mean bounding box occupied by a single non-clustered mosquito in our 640$\times$480 images is 126$\times$124 (about 6.3mm$\times$6.2mm) and the mean length of the proboscis is 36 pixels (about 1.8mm), these errors are relatively small. Also, the robot gripper in the MMS allows tolerance error of 0.5mm [34]. The keypoint detection error can be further reduced by collecting more data for training, increasing the image resolution and increasing the scale of the mosquito in the image.

C. Comparison between Deep Learning and Cascaded Image Processing Approach

In terms of functionality, the deep learning approach can perform bounding box detection, cluster condition classification, and six body parts detection. The cascaded image processing approach can perform bounding box detection, cluster condition classification, head detection and proboscis orientation detection. The image processing approach's limitation in detection of body parts is due to its heavy dependence on geometry, while the mosquito may appear in various poses and shapes when they arrive at the staging area. Also, the testing data for the deep learning approach (Figure 5 and Figure 6) are variant in luminance condition and mosquito scale (the proportion of the mosquito with respect to the whole image). The results show that this method is robust to variant luminance conditions and mosquito scales. For the cascaded image processing approach, the thresholds for excluding the false positive in detection are determined by the luminance condition and the mosquito scale. Thus, it can only deal with figures with fixed luminance conditions and mosquito scales.

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

We test the performance of the deep learning approach and the cascaded image processing approach on the subset of the testing data in which the luminance condition and mosquito scale are invariant. The results are shown in Table I. The detection recall is the average value of the non-clustered and clustered class. The non-maximum suppression of the prediction results in both methods are conducted at an IoU of 0.5. We also evaluate the head position detection and the proboscis orientation detection. The predicted orientation angle $\theta_{\text{proboscis}}$ of the deep learning method is calculated with the proboscis tip position $(x_{\text{tip}}, y_{\text{tip}})$ and proboscis end position $(x_{\text{end}}, y_{\text{end}})$:

$$\theta_{\text{proboscis}} = \text{atan2}(y_{\text{tip}} - y_{\text{end}}, x_{\text{tip}} - x_{\text{end}})$$

The deep learning method outperforms the cascaded image processing method in both mosquito detection and body parts detection. However, the cascaded image processing method has a relatively fast processing speed.

V. DISCUSSION & CONCLUSION

We present a deep learning approach for multi-mosquito object detection and 2D pose estimation. We collect data with the staging apparatus used for the batch production of PfSPZ Vaccine and train two networks to perform bounding box detection, cluster condition classification, and body part detection for multiple mosquitoes. Our results show that this approach is able to distinguish the cluster condition with an mAP of 0.84 (IoU$>$0.75) and detect the body parts position with an RMSE of 3.1 pixels, running at 2.5 fps. Also, we propose a cascaded image processing approach which is able to perform multi-mosquito object detection and detect the head position and proboscis orientation. We compare the two proposed approaches in terms of detection functionality, robustness to luminance condition and mosquito scale, mosquito detection precision and recall, head position RMSE, proboscis detection error and processing speed. The results show that the deep learning approach outperforms the image processing approach in both mosquito detection and body part detection. Also, the deep learning approach is more robust and versatile, though it is running slower than the image processing approach.

As a part of future work, we are very interested in implementing the proposed vision algorithms into the MMS to facilitate the automation of PfSPZ Vaccine production. Also, as pointed out in Section IV, the proboscis of the mosquito

can be easily confused with the leg and antennae for their similar shapes. The problem can be alleviated in two ways. On the algorithm side, more data including various poses of the mosquito can be collected for training and testing. On the mechanical side, we notice in the experiment that the leg of the mosquito can be easily flushed off by the water flow without damaging other body part. The keypoint detection accuracy can be greatly increased if legs are removed before the detection.

As malaria continues to have an enormous impact on morbidity and mortality worldwide, we hope our effective approaches would push forward the automation of PfSPZ Vaccine production which further facilitate the prevention and elimination of this disease in the world.

ACKNOWLEDGMENT

We acknowledge the effort of Zeyu Lu, Guangzhi Zhu, and Guanqun Huang for their discussions on the cascaded image processing approach. This research is supported in part by NIH SBIR grant 1R44AI134500-01 and in part by Johns Hopkins University internal funds and is in collaboration with Sanaria, Inc.

REFERENCES

[1] W. H. Organization, *World malaria report 2018*. World Health Organization, 2018.

[2] J. L. Gallup and J. D. Sachs, "The economic burden of malaria," *The American journal of tropical medicine and hygiene*, vol. 64, no. 1, pp. 85–96, 2001.

[3] B. Greenwood, K. Bojang, C. Whitty, and G. Targett, "Malaria," *the Lancet*, vol. 365, pp. 1487–1498, 2005.

[4] R. A. Seder, L.-J. Chang, M. E. Enama, K. L. Zephir, U. N. Sarwar, I. J. Gordon, L. A. Holman, E. R. James, P. F. Billingsley, A. Gunasekera *et al.*, "Protection against malaria by intravenous immunization with a nonreplicating sporozoite vaccine," *Science*, vol. 341, no. 6152, pp. 1359–1365, 2013.

[5] T. C. Luke and S. L. Hoffman, "Rationale and plans for developing a non-replicating, metabolically active, radiation-attenuated plasmodium falciparum sporozoite vaccine," *Journal of experimental biology*, vol. 206, no. 21, pp. 3803–3808, 2003.

[6] J. E. Epstein, K. Tewari, K. Lyke, B. Sim, P. Billingsley, M. Laurens, A. Gunasekera, S. Chakravarty, E. James, M. Sedegah *et al.*, "Live attenuated malaria vaccine designed to protect through hepatic cd8+ t cell immunity," *Science*, p. 1211548, 2011.

[7] S. L. Hoffman, P. F. Billingsley, E. James, A. Richman, M. Loyevsky, T. Li, S. Chakravarty, A. Gunasekera, R. Chattopadhyay, M. Li *et al.*, "Development of a metabolically active, non-replicating sporozoite vaccine to prevent plasmodium falciparum malaria," *Human vaccines*, vol. 6, no. 1, pp. 97–106, 2010.

[8] M. Schrum, A. Canezin, S. Chakravarty, M. Laskowski, S. Comert, Y. Sevimli, G. S. Chirikjian, S. L. Hoffman, and R. H. Taylor, "An efficient production process for extracting salivary glands from mosquitoes," *arXiv:1903.02532*, 2019.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[10] 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.

[11] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4929–4937.

[12] M. Fuchida, T. Pathmakumar, R. E. Mohan, N. Tan, and A. Nakamura, "Vision-based perception and classification of mosquitoes using support vector machine," *Applied Sciences*, vol. 7, no. 1, p. 51, 2017.

[13] L. P. Huang, M. H. Hong, C. H. Luo, S. Mahajan, and L. J. Chen, "A vector mosquitoes classification system based on edge computing and deep learning," in *Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, 2018, pp. 24–27.

[14] W. Ding and G. Taylor, "Automatic moth detection from trap images for pest management," *Computers and Electronics in Agriculture*, vol. 123, pp. 17–28, 2016.

[15] Z. Liu, J. Gao, G. Yang, H. Zhang, and Y. He, "Localization and classification of paddy field pests using a saliency map and deep convolutional neural network," *Scientific Reports*, vol. 6, p. 20410, 2016.

[16] C. Wen, D. Wu, H. Hu, and W. Pan, "Pose estimation-dependent identification method for field moth images using deep learning architecture," *Biosystems Engineering*, vol. 136, pp. 117–128, 2015.

[17] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge, "Deeplabcut: markerless pose estimation of user-defined body parts with deep learning," *Nature Neuroscience*, vol. 21, pp. 1281–1289, 2018.

[18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[19] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European Conference on Computer Vision (ECCV)*, 2006, pp. 404–417.

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

[21] R. Girshick, "Fast r-cnn," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems (NIPS)*, 2015, pp. 91–99.

[23] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7291–7299.

[24] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 483–499.

[25] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in neural information processing systems (NIPS)*, 2014, pp. 3320–3328.

[26] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[27] M. Xu, S. Lu, Y. Xu, Kocabalkanli, J. Can, Jia, B. Chirikjian, J. Chirikjian, J. Davis, J. S. Kim, S. Chakravarty, I. Iordachita, R. H. Taylor, and G. S. Chirikjian, "Mosquito staging apparatus for producing pfspz malaria vaccines," *Submitted to the 2019 15th IEEE International Conference on Automation Science and Engineering (CASE 2019)*.

[28] A. Dutta, A. Gupta, and A. Zissermann, "Vgg image annotator (via)," *URL: http://www. robots. ox. ac. uk/˜ vgg/software/via*, 2016.

[29] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, no. 2, 2017, p. 4.

[30] J. Chanussot, P. Lambert *et al.*, "Watershed approaches for color image segmentation," in *NSIP*, vol. 99, 1999, pp. 129–133.

[31] H. Yuen, J. Princen, J. Illingworth, and J. Kittler, "Comparative study of hough transform methods for circle finding," *Image and vision computing*, vol. 8, no. 1, pp. 71–77, 1990.

[32] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[33] J. Matas, C. Galambos, and J. Kittler, "Robust detection of lines using the progressive probabilistic hough transform," *Computer Vision and Image Understanding*, vol. 78, no. 1, pp. 119–137, 2000.

[34] H. Phalen, P. Vagdargi, M. Pozin, G. S. Chirikjian, I. Iordachita, and R. H. Taylor, "Mosquito pick-and-place: Automating a key step in pfspz-based malaria vaccine production," *Accepted to the 2019 15th IEEE International Conference on Automation Science and Engineering (CASE 2019)*.