Towards Visual Insect Camera Traps

Anonymous Manuscript
Secret Department
Unknown University
Nowhere Land
Email: anonymous@no.where

***Abstract*—Camera traps have become a standard tool to survey wildlife distribution, abundance and behaviour. Unfortunately, the detection mechanisms are neither sensitive nor selective enough to trigger in case of insect visitations so that current systems can only be used for bigger vertebrates. In this progress report, we present our effort towards a visual insect camera trap. In particular, we discuss why current systems fail, summarise the involved challenges, trained several models on a novel realistic wildlife insect dataset and present the results of our current prototype. Our dedicated deep learning based small object detectors achieve an average precision of 78% while being trained on colour and motion features to identify insects within the field of view of the camera. Finally, we discuss which technical requirements and steps will be necessary to provide a versatile tool for future behavioural, ecological and agricultural studies.**

I. Introduction

Camera traps are a powerful and widely used method for the visual assessment of wildlife animal populations [1]. These imaging systems have been used for a variety of studies, for example to assess the biodiversity [2] or to measure animal travel speed and day range [3]. In most cases, passive infrared based motion sensing is used to trigger the collection of an image sequence [4]. Unfortunately, this triggering mechanism is not sensitive enough to detect small objects that do not generate own body heat, such as insects, so that biodiversity monitoring of insects still heavily relies on invasive techniques such as malaise traps [5].

The absence of any non-invasive visual insect camera trap (VICT) is particularly problematic since we live amid a global wave of anthropogenically driven loss of insect biodiversity [6]. This loss is expected to provoke cascading effects on food webs and to jeopardize ecosystem services since 90% of flowering plant species benefit from insect pollinators [7]. In fact, insects promote 75% of major global agricultural crops, thereby contributing ~ 150 billion EUR to the global economy [8], [9]. Moreover, insects are an important food source for other organisms so that its loss threatens the overall functioning of our ecosystems all over the world. Last but not least VICTs could also be used to monitor crops allowing more targeted use of pesticides [10].

In particular, five algorithmic (A1 – A5) and three hardware challenges (H1 – H3) have to be considered to develop a robust triggering mechanism for a versatile insect camera trap:

- A1 The trigger mechanism has to be sensitive enough to detect small objects in cluttered natural environments.
- A2 The mechanism must be robust to dynamic and omnipresent background motion such as wind-induced plant movements.
- A3 The entering speed of insects into the field of view can vary severely, ranging from slow crawling to fast jumping and flying (e.g. inducing motion blur).
- A4 Many insects show a camouflaged appearance inducing low contrast differences between the target and the background.
- A5 The decision to trigger must be made within milliseconds to not miss short events.
- H1 The visual properties of such a system must be a trade-off between a reasonably big field of view (fov), image magnification and depth of field.
- H2 The energy consumption of this triggering mechanism has to be low in order to allow surveys in remote areas.
- H3 Real-time image transmission is required if these systems are to be used for pest control.

Here, we are mainly focusing on the algorithmic challenges by training customised deep learning models to detect tiny, fast moving and low contrast objects in front of dynamic backgrounds. The hardware difficulties will however be discussed in the context of our proposed methodology.

A. Related Work

The use of deep learning models for image-based insect classification and detection is not new. In particular, for species classification, a variety of different studies can be found. For example Squeeze-and-Excitation Networks in combination with attention modules have been used to recognise different insect species [11]. The images used for this study, however, were of comparatively high quality and thus do not address the challenges A1 to A4. This is true for almost all visual insect classification strategies [12]. Others have tried to circumvent this problem by analysing audio data for insect recognition, which however can only be applied for particular species [13]. In a different approach, challenges A1 to A4, as well as H1 have been addressed by constraining the imaging conditions using a white box [14].

In order to localise insects in images, object detection is required. For this reason, the MPest dataset was published in 2018 [15]. By using an adapted version of Faster RCNN, the authors achieved a mean average precision (mAP) of 89%. As the MPest dataset only comprised imagery in which the animals covered a relatively large proportion of the image,

Fig. 1. Insect camera trap dataset. The first image does not comprise an insect visitation. Insect locations in the second and third image are specified by a red box. Note the small size, cluttered background and motion blur of the animal.

this approach is still insufficient for real-life applications such as insect biodiversity monitoring.

A more realistic dataset was proposed by Grant et al. called iNaturalist [16]. This dataset includes $\sim$ 1,000 insect classes and more than 125,000 bounding boxes. Using the Inception Resnet V2, the authors achieved a top-1 accuracy of 77.1%. However, since only selected image stills (i.e. no image sequences) are available in the iNaturalist dataset, it can not be used to develop a VICT.

II. Experiments

To derive insights into the possibilities and limitations of deep learning based VICT, we implemented the following steps: We (i) built a realistic visual insect camera trap database (Section II-A); (ii) evaluated a binary classifier for insect visitations (Section II-B); (iii) implemented and evaluated a more targeted insect detection approach (Section II-C); (iv) evaluated the impact of temporal information onto the detection accuracy (Section II-D) and (v) evaluated the performance of a deployable small network which could be used for an embedded stand-alone VICT system (Section II-E).

A. VICT Dataset

Insect imagery was acquired in the field within the framework of an agricultural field experiment, where experimental field plots of $\sim 2 \times 4.5$ m^2 were recorded using a waterproof outdoor camera (Ricoh WG-50). In total we installed 16 cameras to monitor different crop combinations in monocultures and mixtures. These cameras captured continuous 30 frames per second (fps) interval recordings resulting in $1920 \times 1080 \times 3$ pixel images[1]. For classification, we extracted 35,129 frames from these videos of which 22,008 images included at least one insect. To enable insect detections, we added location information to 50% of the videos resulting in 14,847 bounding boxes in total. In contrast to the existing data, our dataset included the challenges A1 – A4 and H1: With respect to the entire image, the average bounding box had a height of 7.5%, a width of 4.4% and covered an area of 0.4% pixel. Exemplary images are given in Figure 1.

[1]An additional dataset was created using the same camera setup in a field experiment in salt marsh ecosystems (not shown here); datasets are available upon request.

Fig. 2. Exemplary image after highlighting the pixel most relevant for the overall classification result in red (hoverfly, Syrphidae). As can be seen almost no pixel within the blue bounding box (i.e. animal location) were responsible.

B. Binary Classification

To analyse the performance of a deep learning based image classifier, we trained a ResNetV1 on this dataset [17]. We changed the input dimension to $(700, 700, 3)$ by rescaling the images and the output dimension to 2 (visitation, no visitation) resulting in 232,354 parameters. The batch size was set to 25 and the network was trained for 175 epochs using Adam with a learning rate of 0.001 on 30,736 training images. The remaining images were used as a validation set on which a maximum accuracy of 75.7% was reached after 77 epochs.

Next, we used the resultant model to investigate which pixel in the input image were responsible for the classification result. In particular, we used guided backpropagation [18] and manually inspected hundreds of images of the validation set. An exemplary image is given in Figure 2: The prediction accuracy for this frame was 97.87% for the visitation class. However, as highlighted in red almost no pixel responsible for this result are actually located on the insect. This confirms that binary classification can be error prone for the underlying task and suggests that more targeted object localisation is preferable.

C. Spatial Insect Detection

In order to evaluate state of the art object detection models for our insect camera trap dataset, we trained two different algorithms on the annotated images, namely YoloV3 [19] and Faster RCNN [20]. The results are summarised in Table I: The first block indicates the results if trained on the RGB images and the second block shows the results if trained on the modified HSV image space (c.f. Section II-D).

TABLE I
QUANTITATIVE EVALUATION OF THREE DETECTION NETWORKS. AP: AVERAGE PRECISION WITH AN IOU $\geq 50\%$.

	Architecture	AP	Precision	Recall	F1	Time
RGB	YoloV3	74.15%	96%	70%	81%	2.5 fps
	Faster RCNN	71.95%	91%	80%	85%	0.3 fps
	MobileNet	69.82%	90%	70%	79%	0.45 fps
HSV*	YoloV3	78.07%	92%	72%	81%	2.5 fps
	Faster RCNN	74.39%	86%	82%	84%	0.3 fps
	MobileNet	72.41%	92%	71%	80%	0.45 fps

YoloV3: The YoloV3 architecture was used since it was particularly improved for small object detections compared to earlier Yolo versions [19]. We used a batch size of 64 and trained it for 25,000 iterations ($\sim$ 102 epochs) by using Adam with a learning rate of 0.001. The dataset was split into 15,671 training and 3,918 validation images. The input layer size was changed to $(1216, 704, 3)$ using a subdivision of 32 and average precision (AP) was used to calculate the accuracy. After 10,340 iterations the maximal validation accuracy of 74.15% AP was reached revealing a precision of $\sim$ 96%, a recall of $\sim$ 70% and an F1 score of $\sim$ 81%. The mean processing time for inference was 2.5 fps (measured on an Intel i7-9750h CPU and a TX 1660 Ti GPU). These values reveal a low false positive detection rate but also imply many false negatives (i.e. should have detected a visitor but missed it) at relatively high frame-rates.

Faster RCNN: The Faster RCNN architecture was chosen because it has improved performances when calculating the regions of interest and it has also been the top performing model in a similar study for insect detection [15]. Images were resized to $(1280, 1280, 3)$ using the same training/validation split as described above and a batch size of 8 to fit in the GPU memory. The network was trained for 250,000 steps ($\sim$ 128 epochs) using the Adam optimiser and a learning rate of 0.00005. Best AP was reached after 32,500 steps reaching 71.95% accuracy, a precision of $\sim$ 91% and a recall of $\sim$ 80% resulting in an F1 score of $\sim$ 85%. In contrast to YoloV3, Faster RCNNs had more false positive detections but less false negatives. The mean inference time for Faster RCNN was substantially slower compared to YoloV3 by reaching only 0.3 fps.

In summary, this evaluation reveals that deep learning detection models outperform binary classification architectures while successfully avoiding the use of non-informative background pixel. However, up to now only spatial colour information were used.

D. Spatio-Temporal Insect Detection

In order to provide a more realistic VICT triggering scenario, we evaluated the performance of the above mentioned detection architectures in the presence of temporal cues. In particular, we combined spatial with temporal cues by a straight forward image processing routine. First, we took the pixel-wise absolute difference between consecutive frames $D_t = |F_t - F_{t+1}|$. The resultant difference image was converted to a mask M_t by setting all values above a threshold T to 1 and all other values to 0.3 (i.e. 30% intensity). Next, frame F_t was transformed into the HSV space, where the V (value) dimension was weighted by the mask: $V_t^* = V_t \circ M_t$. As a result, novel HSV$_t^*$ images for all frames t are generated in which pixel with high motion differences comprise higher intensity values (Figure 3). A motion enhanced image is given in Figure 4. Note that both, the insect position (red box) and edges of the plant are highlighted in this HSV* image.

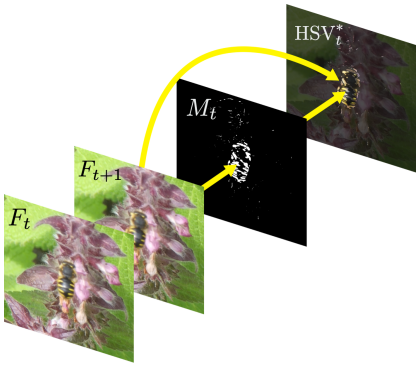

Fig. 3. Integration of motion cues into the HSV space. For details see text.

YoloV3: When training the YoloV3 model with the spatio-temporal enhanced HSV* images the AP increases to 78.07% (+3.92%; reached after 10,340 steps). Interestingly the precision drops to $\sim$ 92% while the recall increases to $\sim$ 72% resulting in the same overall F1 score of $\sim$ 81%. These measures indicate that the false negatives dropped but with an increase of false positives. Since erroneously captured empty images are less problematic than missed visitations, these results show a clear improvement.

Faster RCNN: In a similar fashion the performance of Faster RCNNs improve by +2.44% to 74.39% after 10,000 steps and the precision decreases to 86% whereas the recall increases to 82% (F1 score 84%). Again, these results are favourable if low false negative rates are required.

E. Towards an Embedded VICT via the SSD MobileNet V2

A stand-alone VICT requires that potential visitation detections are performed on the build-in hardware of the trap. Considering challenge A5 fast inferences are vital for real-time applications and also the energy efficiency is crucial (challenge H2). Therefore, we evaluated an edge TPU ready neural network which could be used to build an embedded VICT by training the SSD MobileNet V2 [21] on the dataset described above. After training, the model is quantized to enable accelerated inferences.

In this experiment an image resolution of $(1280, 720, 3)$ was used. Again we used the Adam optimiser with a batch size of 4 and a learning rate of 0.00005 and trained the network for 200,000 steps ($\sim$ 51 epochs). For the RGB images a maximum AP of 69.82% was reached after 195,000 steps. The overall mean computational time for inference was 0.45 fps. By using the HSV* dataset the performance of the SSD MobileNet

Fig. 4. Motion enhanced HSF* frame from Figure 1 (red box: insect location).

V2 increased by 2.59% to 72.41% after training for 72,500 steps. As can be seen in Table I both, the precision and recall improves when including temporal information, which is also reflected in a slight improvement of the F1 score.

III. Discussion & Conclusion

In summary, temporal motion cues can help to avoid false negative detections (i.e. missed visitations) which is of utmost importance for a VICT triggering mechanisms. All tested architectures achieved higher recall values once motion was included but at the cost of slightly reduced precision measures.

In this study, we evaluated the possibilities and limitations of deep learning based visual insect camera traps. In particular, we summarised the challenges involved in such a system and generated a novel dataset to investigate the underlying difficulties. We tested a binary classification mechanisms (visitation vs. no visitation) on this dataset revealing that too many irrelevant background pixel were involved in the classification decision. Moreover, we evaluated three detection architectures, namely YoloV3, Faster RCNN and SSD MobileNet V2. We gathered evidence that more targeted detection mechanisms have to be used to identify the location of the insects to result in a comparatively better trigger mechanism. By involving temporal cues we could further demonstrate, that this additional information can help to reduce the false negative rate in order to avoid missed visitations.

Additional research is necessary to improve the accuracy of a visual insect camera trap. Considering the highly diverse and complex imaging conditions, deep learning algorithms appear to be most promising to enable robust triggering mechanisms. Furthermore, advances in IoT hardware are also crucial to accelerate the image analysis via edgeTPUs while reducing the energy consumption necessary for an embedded image analysis system.

Acknowledgement

Acknowledgement excluded for review.

References

[1] A. C. Burton, E. Neilson, D. Moreira, A. Ladle, R. Steenweg, J. T. Fisher, E. Bayne, and S. Boutin, "REVIEW: Wildlife camera trapping: a review and recommendations for linking surveys to ecological processes," *Journal of Applied Ecology*, vol. 52, no. 3, pp. 675 – 685, 2015.

[2] B. Zaragozí, A. Belda, P. Giménez, J. T. Navarro, and A. Bonet, "Advances in camera trap data management tools: Towards collaborative development and integration with GIS," *Ecological Informatics*, vol. 30, no. C, pp. 6 – 11, 2015.

[3] J. M. Rowcliffe, P. A. Jansen, R. Kays, B. Kranstauber, and C. Carbone, "Wildlife speed cameras: measuring animal travel speed and day range using camera traps," *Remote Sensing in Ecology and Conservation*, vol. 2, no. 2, pp. 84–94, 2016.

[4] J. L. P. Tack, B. S. West, C. P. McGowan, S. S. Ditchkoff, S. J. Reeves, A. C. Keever, and J. B. Grand, "AnimalFinder: A semi-automated system for animal detection in time-lapse camera trap images," *Ecological Informatics*, vol. 36, no. C, pp. 145 – 151, 2016.

[5] C. A. Hallmann, M. Sorg, E. Jongejans, H. Siepel, N. Hofland, H. Schwan, W. Stenmans, A. Müller, H. Sumser, T. Hörren, D. Goulson, and H. d. Kroon, "More than 75 percent decline over 27 years in total flying insect biomass in protected areas," *PloS ONE*, vol. 12, no. 10, pp. 1 – 21, 2017.

[6] G. Ceballos, P. R. Ehrlich, and R. Dirzo, "Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines," *Proceedings of the National Academy of Sciences*, vol. 114, no. 30, pp. E6089–E6096, 2017.

[7] R. Dirzo, H. S. Young, M. Galetti, G. Ceballos, N. J. B. Isaac, and B. Collen, "Defaunation in the Anthropocene." *Science*, vol. 345, no. 6195, pp. 401 – 406, 2014.

[8] N. Gallai, J.-M. Salles, J. Settele, and B. E. Vaissière, "Economic valuation of the vulnerability of world agriculture confronted with pollinator decline," *Ecological economics*, vol. 68, no. 3, pp. 810–821, 2009.

[9] E. J. Eilers, C. Kremen, S. Smith Greenleaf, A. K. Garber, and A.-M. Klein, "Contribution of pollinator-mediated crops to nutrients in the human food supply," *PLOS ONE*, vol. 6, no. 6, pp. 1–6, 06 2011.

[10] M. Cardim Ferreira Lima, M. E. Damascena de Almeida Leandro, C. Valero, L. C. Pereira Coronel, and C. O. Gonçalves Bazzo, "Automatic detection and monitoring of insect pests—a review," *Agriculture*, vol. 10, no. 5, p. 161, 2020.

[11] Y. J. Park, G. Tuxworth, and J. Zhou, "Insect Classification Using Squeeze-and-Excitation and Attention Modules - a Benchmark Study," *2019 IEEE International Conference on Image Processing (ICIP)*, vol. 00, pp. 3437–3441, 2019.

[12] M. Martineau, D. Conte, R. Raveaux, I. Arnault, D. Munier, and G. Venturini, "A survey on image-based insect classification," *Pattern Recognition*, vol. 65, pp. 273–284, 2017.

[13] D. F. Silva, V. M. A. d. Souza, G. E. A. P. A. Batista, E. Keogh, and D. P. W. Ellis, "Applying Machine Learning and Audio Analysis Techniques to Insect Recognition in Intelligent Traps," *2013 12th International Conference on Machine Learning and Applications*, vol. 1, pp. 99–104, 2013.

[14] Y. Chen, A. Why, G. Batista, A. Mafra-Neto, and E. Keogh, "Flying Insect Classification with Inexpensive Sensors," *Journal of Insect Behavior*, vol. 27, no. 5, pp. 657–677, 2014.

[15] D. Xia, P. Chen, B. Wang, J. Zhang, and C. Xie, "Insect detection and classification based on an improved convolutional neural network," *Sensors*, vol. 18, no. 12, p. 4169, 2018.

[16] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *CVPR*, 2018, pp. 8769–8778.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition." *CVPR*, pp. 770 – 778, 2016.

[18] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv*, 2014.

[19] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018.

[20] S. Ren, K. He, R. B. Girshick, and J. S. 0001, "Faster R-CNN - Towards Real-Time Object Detection with Region Proposal Networks." *TPAMI*, vol. 39, no. 6, pp. 1137 – 1149, 2017.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," in *ECCV*, 2016, pp. 21–37.