Multiple Animals Tracking in Video Using Part Affinity Fields

Ivan F. Rodriguez*, Rémi Mégret†, Roian Egnor ‡, Kristin Branson‡
Jose L. Agosto§, Tugrul Giray§ and Edgar Acuña¶
*Department of Mathematics, University of Puerto Rico, Río Piedras campus
†Department of Computer Science, University of Puerto Rico, Río Piedras campus
‡HHMI Janelia Research Campus
§Department of Biology, University of Puerto Rico, Río Piedras campus
¶Department of Mathematical Sciences, University of Puerto Rico, Mayagüez campus

Abstract—**In this work, we address the problem of pose detection and tracking of multiple individuals for the study of behaviour in insects and animals. Using a Deep Neural Network architecture, precise detection and association of the body parts can be performed. The models are learned based on user-annotated training videos, which gives flexibility to the approach. This is illustrated on two different animals: honeybees and mice, where very good performance in part recognition and association are observed despite the presence of multiple interacting individuals.**

I. Introduction

Automatic pose estimation of insects and animals in video is of great interest for behavioural science [1]. High precision in detection and tracking of parts of animals is crucial for quantitative measurement of social interactions of multiple individuals. The ability to measure detailed interactions and the performance of specialized tasks provides a confident baseline that contributes to the understanding of behaviour when more than one individual is present [2].

Recent developments in machine vision and machine learning have successfully approached human real-time pose estimation [3] by providing algorithms that perform precise limb detection and correct association between them, even in complex scenes containing multiple interacting persons. Given the similarity of the tasks, this makes it suitable for application in the study of behaviour of animals, especially when complex settings, such as open field conditions or close interaction between multiple animals are considered.

In this work, we present an adaptation of the Part Affinity Fields approach [3] for detection and tracking of insect and mammal body parts. Results on honeybees and mice show that this tracking-by-detection approach produces high-quality results in presence of multiple individuals and is a promising approach to obtain precise estimates of pose for behavioral studies.

II. Related work

Traditional techniques for behavioral study have been focused on using the pose extracted from generic image processing approaches, such as ellipse-based detectors. In these approaches, the body, detected by background subtraction, is fitted with an ellipse that is then tracked over time [4].

Cascaded Pose Regression [5] was applied to track mice and fish. This method relies on an initial estimate that is refined progressively using a sequence of regressors.

More recently, deep neural network architectures have shown to provide good performance for the tracking of constrained honeybee body parts, learning the mapping from the global structure and local appearance. [6]. In addition to detection and tracking, identification of large amount of individuals using convolutional neuronal networks was proposed in [7].

The Part Affinity Fields approach [3] introduced a neural network architecture to learn both how to detect the body parts and how to associate them into a complete body skeleton. A convolutional network simultaneously predicts a set of 2D confidence maps S of body parts present and a set of 2D vector fields $\mathbf{L}$ of part affinity fields (PAFs), which encode the association between the parts. A multi-stage architecture is used to refine both fields and enforce consistency between them. Greedy inference is used to select the most likely predictions for the parts and use them as candidates for the PAFs to associate them. This approach is based on a tracking by detection approach, where no assumption is made on the number or the location of the individuals during the detection phase.

III. Part Affinity Fields Adaptation

The work of [3] focused on human pose estimation. We now discuss how this approach can be applied to animal pose.

A. Detection and Association fields

We will use the same notation as [3] and denote by $S = (S_1, S_2, ..., S_j)$ the set of J confidence maps, one per body part. The PAFs $\mathbf{L} = (\mathbf{L}_1, \mathbf{L}_2, ..., \mathbf{L}_C)$ encode C vector fields, one per connection.

To accommodate animals with different numbers of body parts, our implementation includes a flexible configuration that allows the user to define custom skeletons and custom number of parts. The ability to adapt the architecture according to the number of parts improves the training time when only a few parts are needed.

For honeybees, we considered five parts including abdomen, thorax, head and the two antennae. For mice, we currently consider two parts tail and head as the flexibility of the bodies and hair and the lack of precise visual landmarks makes it harder to define other reference points. Adding additional body parts in this context is the subject of ongoing work.

B. Inference Stage

Given that honeybees may present poses on multiple directions, including upside down, and that it is common for two or more individuals to be aligned, we used the distribution of the distance between points to constrain the connections based on the scale of the honeybees' bodies. Thus, PAFs of different bodies that were aligned can be recognized as separate bodies. For example, Figure 1 shows the type of issues presented before this extra step in the inference stage was taken.

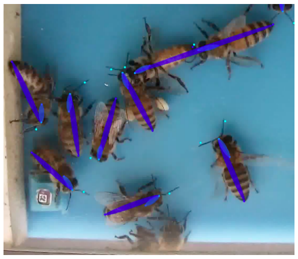

Fig. 1. Incorrect association obtained when ignoring factor $\pi_{j_1 j_2}$ when two bodies that are aligned.

The original approach measured the association between two parts by computing the line integral over the corresponding PAFs or, in other words, by measuring the alignment of the body parts detected. For instance, considering two body parts $\mathbf{d}_{j_1}$ and $\mathbf{d}_{j_2}$ that they are candidates for the association, the confidence of such election is expressed as:

$$E = \int_{u=0}^{u=1} \mathbf{L}_c(\mathbf{p}(u)) \cdot \frac{\mathbf{d}_{j_2} - \mathbf{d}_{j_1}}{\left\|\mathbf{d}_{j_2} - \mathbf{d}_{j_1}\right\|_2} \pi_{\mathbf{j_1 j_2}} du \tag{1}$$

where $\mathbf{p}(u)$ interpolates the position of the two body parts d_{j_1}, d_{j_2}, $\mathbf{p}(u) = (1-u)d_{j_1} + ud_{j_2}$. The probability $\pi_{j_1 j_2}$ is defined as the empirical probability that d_{j_1} connects with d_{j_2} conditionned on $\left\|d_{j_2} - d_{j_1}\right\|_2$, and is evaluated on the training data. This factor is important in scenarios where PAFs can be aligned, since all individuals share the same PAF channels in the network.

This particular assumption works well for honeybees, as their body is usually quite rigid, so the variance between the distances of each of the body parts is small. However, for mice body part detection, we did not include this score, as the flexibility of their bodies introduces high variability in the distance between nose and tail, which may affect the correct detection when nose is close to the tail. Instead, we did use the information of a fixed number of individuals to reduce the expression of false positives or wrong connections. This number was used as follows: if too many individuals were detected after the inference, the incomplete skeletons were removed and only the one that matched with past detection were kept; if not enough individuals were detected, the detection threshold was lowered.

IV. Tracking

A. Temporal matching

For tracking, we rely on the precision of the detection on consecutive frames and the Hungarian algorithm [8]. First we create a $N \times T$ matrix M, where N is the maximum possible observed bees at a time, and $T = (t_1, ...t_m)$ are the frames in chronological order. An unique id is assigned to each visible bee at t_1, incrementally from 0 up to the number of bees found. Finally, the matrix was filled such that M_{ij} contains the track id of the ith detection in frame t_j.

The distance metric used for the Hungarian algorithm in the case of honeybees takes into account not only point to point distance to thorax, but also the distance to antennae and to the head. Using the correspondences for all parts reduced incorrect matches due to closeness between several individuals. Missing parts were assigned a fixed penalty of 100 pixels.

B. Behaviour Classification

In the case of honeybees, the quality of detection for the frame-to-frame tracking enabled us to define a preliminary rule based classification of foraging behavior using the starting and ending points of the tracks, as illustrated in Figure 2. Trajectories ending at the bottom correspond to leaving bees; those ending at the top correspond to entering bees; and bees that stay a long period of time in a fixed position, usually are cooling the colony using their wings (fanning behavior).

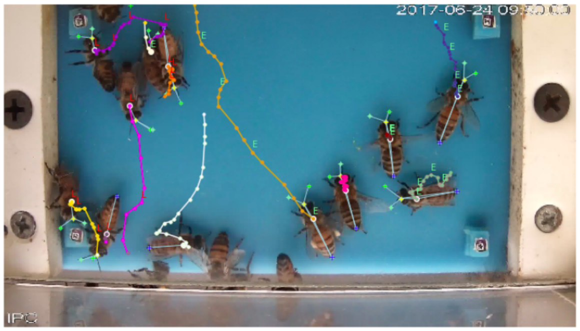

Fig. 2. Detection, tracking and behaviour classification. E indicates ”entering,” L ”leaving,” and F ”fanning.”

V. Experimental results

A. General considerations

We based our part affinity field implementation on a Keras-TensorFlow open source project[1]. Specific changes were made to the architecture definition and the inference part, as was explained in Section I. Our practical customizations include

[1] https://github.com/michalfaber/keras_Realtime_Multi-Person_Pose_Estimation

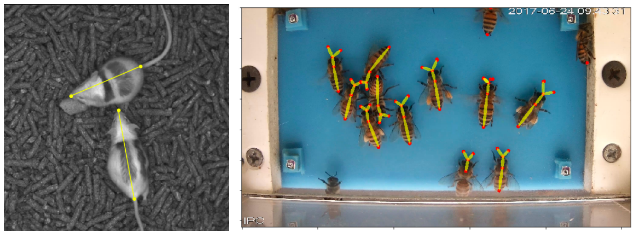

Fig. 3. Skeleton detected for Mice and Honey bee frames.

a flexible structure definition that allows us to use a smaller network when only detecting five parts or fewer are being detected. In the cited implementation, a fixed 19 channels for parts and 38 for PAFs is used, making it too slow when predicting for smaller conditions.

For the videos of mice and honeybee the camera was always in fixed position. The annotation of the video datasets consisted of labeling every individual animal in a selected frame according to the number of parts selected. It was assured there was a difference of at least two seconds between each of the samples training datasets to enrich the diversity of poses.

The annotation followed the Coco dataset's format [9], which consisted of labeling every fully visible bee or mouse in the frame with the desired body parts. For honeybees body parts were: Tip of the Abdomen, Head, Thorax, Left Antenna and Right Antenna and for mice: Nose and Tail. Each individual represents one separate annotation, and each body part is a tuple (x, y, v) where x, y represent the Cartesian coordinates and v the visibility (0: absent, 1: visible and present, 2: present but not visible). Once these datasets were obtained, a split with training ratio of 2/3 was used.

The dataset was augmented by a factor of 82 using linear transformations, such as translations, rotations, and scaling. All the experiments were performed using a Nvidia Titan X GPU card.

Examples of detection are shown in Figure 3.

B. Results on honeybee videos

1) Dataset: The video capture system is designed to observe the ramp through which all foraging bees must pass to exit or enter the colony. We used a 4 Mpixels GESS IP camera connected to a networked video recorder configured at 8Mbps for continuous recording. A transparent acrylic plastic cover located on top of the ramp ensures the bees remain in the focal plane of the camera. To avoid interfering with the bee's biological cycles, only natural light is used. A white plastic diffuses the natural light received, and a black mask is put around the camera to reduce the direct reflections that could be visible on the ramp cover.

The videos were acquired in June 2017 at the UPR Agricultural Experimental Station of Gurabo, Puerto Rico. Dataset consists of 100 fully annotated frames, where each frame contains from 6 to 14 individuals.

For honeybees, we considered training on five, three, and two parts. Given that in some frames incomplete bees were not labeled, we used a mask to avoid counting their detection as incorrect.

2) Effect of the number of parts: We evaluated the performance of the algorithm using mean Average Precision (mAP) as provided by pose evaluation package [2] which is based on [10]. First, multiple body pose predictions are greedily assigned to the ground truth (GT) based on the highest PCKh [11]. Since our scale is unique, we only use the distance between thorax and head for PCKh-0.5. Table I shows results for the best models for two, three and five parts and Average Precision for each of the parts.

To analyze the performance in terms of the parts considered, we trained the model up to 5000 epochs; every 20 epochs we evaluated and saved the information related to detection of the head and tail. The following figure represents the results obtained from epoch 1000 up to 5000. Showing of each case the min,25% percentile, median, 75% percentile and max.

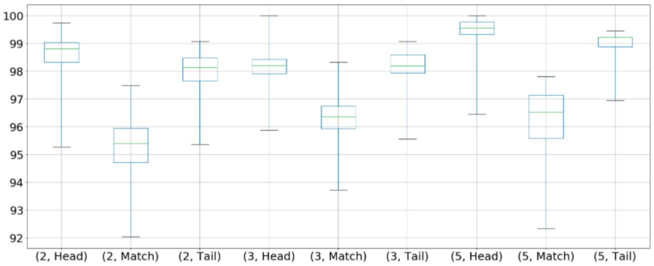

Fig. 4. Box plot results after 1000 epochs of training.

The box plot shown in Figure 4 shows that in terms of detection accuracy, higher scores are obtained on average when using five parts for training. We hypothesize that the higher number of parts may help the network interpolate poorly detected parts by using the detection of its connected parts.

TABLE I
BODY PART DETECTION PERFORMANCE (AP).

	2 parts AP	3 partsAP	5 parts AP
Head	98.7%	96.4%	98.1%
Tip abdomen	94.0%	96.2%	95.0%
Thorax	–	95.0%	98.7%
Right Antenna	–	–	94.4%
Left Antenna	–	–	90.4%
mAP	95.57	96.39	96.4

C. Results on mice videos

1) Dataset: The mouse recordings were made in February 2016 at the Howard Hughes Medical Institute's Janelia Research Campus, in accordance with approved IACUC protocols. The dataset is composed of 450 frames that always contain two individuals.

[2] https://github.com/leonid-pishchulin/poseval.git

2) Detection performance: We considered two parts: nose and tail. For training, 5000 epochs were used, reaching 93.0% mAP in the validation dataset. Since the interaction of mice may involve occlusion, complicated poses that involves curved bodies or close interactions, there is ambiguity in the matching that the PAFs cannot solve, since they are based on the assumption of straight connections. Moreover, when noses are touching, the two noses may generate a single detection from the part confidence map. Future work will evaluate the possibility of training additional intermediate body parts to alleviate these issues.

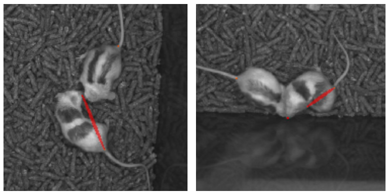

Fig. 5. Issues related with closeness and occlusion.

3) Comparison to CPR: Despite the limitations discussed previously, the proposed approach compared favorably to the Cascaded Pose Regression approach (CPR) [5] on the same challenging data. We took a short video clip with 300 frames and applied both CPR and the proposed approach. Later, an evaluator was presented with both results for each frame and asked to decide which detector did better on each of the frames or if they performed similarly. The criterion the evaluator used to perform evaluation, was to select the model that predicted the position of the body part closer to the real location in the video. In case both detected the evaluator would count them as similar performance.

These results show that out of 300 frames, Part Affinity Fields performed better in 237 frames; in 54 of them both detectors performed similarly; in 4 the CPR detector performed better; and in 6 cases, both failed in at least one nose. Figure 6 shows some examples for each case.

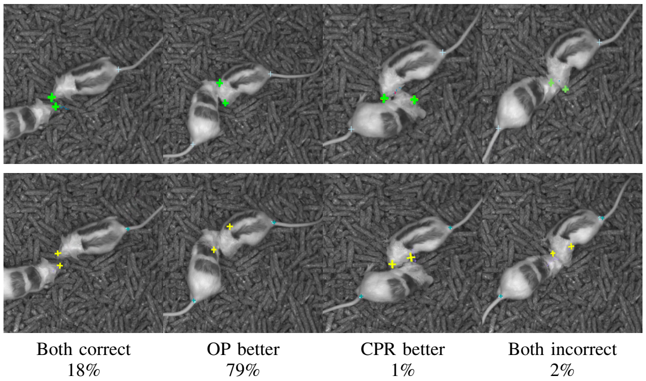

Fig. 6. Illustration of the comparison of (Upper row) Part Affinity Fields detection vs. (Bottom row) CPR detection.

VI. Conclusion

The detection of pose performed by the proposed approach offers a flexible framework that has demonstrated good performance on two different types of animal models. We have shown that the quality of the estimates for honeybees reached high accuracy and that the method outperformed a state-of-the-art tracking approach for mice.

It should be noted that the detection is performed on each frame independently. It is therefore expected that it can be improved by combining it with higher-level tracking algorithms that incorporate knowledge about the dynamics of the animals.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grants No. 1633164 and 1633184. I. F. Rodriguez acknowledges support from the Janelia Visitor Program. T. Giray acknowledges support from Puerto Rico Science and Technology Trust (PRSTRT ; 2016-00161, 2017-00164). The authors would like to thank Stephanie Feliciano and Janpierre Aleman who helped with the acquisition at the UPR Agricultural Station of Gurabo of the honeybee videos and Jeffrey Chan who helped with the annotation of the videos.

References

[1] U. Stern, R. He, and C.-H. Yang, “Analyzing animal behavior via classifying each video frame using convolutional neural networks,” *Scientific Reports*, vol. 5, pp. 14 351 EP –, 09 2015. [Online]. Available: http://dx.doi.org/10.1038/srep14351

[2] A. A. Robie, K. M. Seagraves, S. E. R. Egnor, and K. Branson, “Machine vision methods for analyzing social interactions,” *Journal of Experimental Biology*, vol. 220, no. 1, pp. 25–34, 2017. [Online]. Available: http://jeb.biologists.org/content/220/1/25

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in *CVPR*, 2017.

[4] S. R. Egnor and K. Branson, “Computational analysis of behavior,” *Annual Review of Neuroscience*, vol. 39, no. 1, pp. 217–236, 2016, pMID: 27090952. [Online]. Available: https://doi.org/10.1146/annurev-neuro-070815-013845

[5] P. Dollr, P. Welinder, and P. Perona, “Cascaded pose regression,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 1078–1085.

[6] L. Duan, M. Shen, W. Gao, S. Cui, and O. Deussen, “Bee Pose Estimation From From Single Images With Convolutional Neural Network,” in *IEEE International Conference on Image Processing (ICIP)*, Sep. 2017.

[7] B. M. H. R. H. F. d. P. Romero-Ferrero, F., “idtracker.ai: Tracking all individuals in large collectives of unmarked animals (submitted),” 2018. [Online]. Available: https://arxiv.org/abs/1803.04351.

[8] H. W. Kuhn, “The hungarian method for the assignment problem,” *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955. [Online]. Available: http://dx.doi.org/10.1002/nav.3800020109

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in *European Conference on Computer Vision, ECCV 2014*, Zurich, 2014.

[10] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele, “Arttrack: Articulated multi-person tracking in the wild,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 1293–1301. [Online]. Available: https://doi.org/10.1109/CVPR.2017.142

[11] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, “2d human pose estimation: New benchmark and state of the art analysis,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.