ButterFlySet: A 2D video dataset for pose estimation of the flying butterflies in the wild

Kai Amino[1][0000-0002-4359-991X] and Keisuke Fujii[2][0000-0001-5487-4297]

[1] Graduate School of Informatics, Nagoya University, Japan
kaiamino417@gmail.com
[2] Graduate School of Informatics, Nagoya University, Japan
fujii@i.nagoya-u.ac.jp

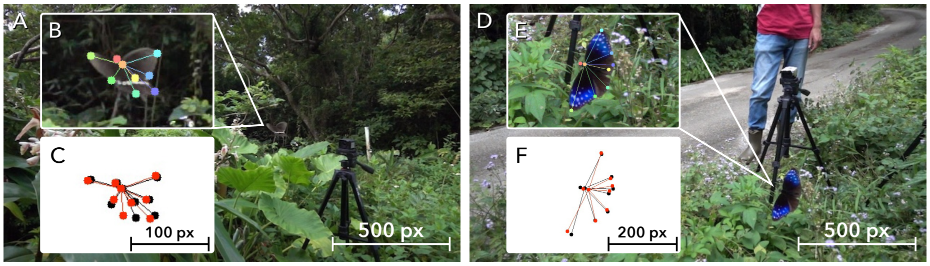

Fig. 1. The examples of the frames in ButterFlySet containing a non-toxic butterfly *Papilio polytes* (A) and a toxic butterfly *Euploea mulciber* (D). Manually annotated nine keypoints (B, E). Results of the deep-learning-based pose estimation (C, F). The black and red points represent ground truth and estimation, respectively.

Abstract. Automated pose estimation remains underdeveloped for insects compared to humans and other animals, largely due to a lack of pose datasets. This challenge should be acute for butterflies, whose unique body structures and flight patterns differ significantly from other insects. Current studies on butterflies' flight behavior rely on manual scoring, which is labor-intensive. To address this, we propose ButterFlySet, the first video dataset of flying butterflies and the largest dataset for insect pose estimation recorded in the wild, containing 7440 frames with nine key points annotation. Using ButterFlySet, we evaluated two animal pose estimation models (DeepLabCut and SLEAP) and measured wingbeat frequency. We found that toxic butterflies fly slower than non-toxic species, offering an insight into predator-prey visual communication. This dataset enables automated analysis of flight patterns of butterflies, reducing labor in data collection and filling a critical gap in datasets collected in the wild and laboratories. Our dataset is available at: https://github.com/KaiAmino/ButterFlySet.

Keywords: Animal pose estimation · Butterfly · Dataset.

1 Introduction

Automated animal pose estimation has provided useful insights into studies related to animal behavior, such as behavioral neurogenetics, animal conservation, and evolutionary ecology [1]. Compared to the ongoing innovations in human pose estimation, animal pose estimation remains underdeveloped, primarily due to the many obstacles it faces, of which the scarcity of pose datasets is viewed as the primary factor [2]. Unlike human pose estimation, the diversity in the shape of animal body parts makes it difficult to use the pose dataset of other groups in the analysis of focal subjects (for example, using a model trained on a dataset of horses in pose estimation of dogs). Moreover, this challenge is likely to be even more pronounced in insects, which exhibit a wide diversity in body shapes, including variations in body structure, leg length, and wing morphology. Because insects are easily utilized in genetic research due to their short life cycles, and understanding their ecology is critical for addressing agricultural pests and exploring potential food resources, automated behavioral analysis in insects is highly valuable [3] [4].

In contrast to the richness of the pose datasets in mammals (e.g., primates [5], dogs [6] [7], and other quadrupeds [8] [9]) the number of insect datasets is quite limited [2]. Two pose estimation models, DeepLabCut [10] and LEAP [11] introduced pose datasets of single fruit flies, whereas SLEAP [12], a multi-animal pose estimation model, was evaluated using two datasets 'flies13' and 'bees' which contains two individuals of fruit fly and bumble bee. However, to the best of our knowledge, no similar datasets exist aside from these four.

In addition to the scarcity of datasets, it seems also undesirable that all of the datasets above were collected in laboratories. Videos captured in the wild, with their diverse backgrounds and lighting conditions, are useful in validating pose estimation models that need to perform accurately under varying conditions. Furthermore, the wild data is desirable because animal behavior recorded in the wild and laboratories can be entirely different and the former appears to reflect their behavior in natural conditions better [13].

Butterflies, known for their beautiful and diverse wing patterns, have attracted researchers worldwide who have accumulated extensive knowledge from various perspectives (e.g., morphology, host plant selection, and mating behaviors), among which flight behavior is studied by both behavioral sciences and biomechanics [14] [15]. Although previous studies describing butterfly flight pattern used manual scoring [14], automated keypoint estimation may significantly reduce the effort required for data collection. However, the challenge in pose estimation is likely to be even more pronounced in butterflies, whose body structure (i.e., large wings compared to their body) and behavior (i.e., flight pattern) is entirely different not only compared to other animals but even among other insects. Nevertheless, there has been no dataset for butterfly pose estimation.

Therefore, we propose ButterFlySet (Fig. 1), the first dataset on the videos of butterflies' flight behavior. We directly recorded the flying butterflies in the wild, and manually annotated 9 key points from a total of 7440 frames (Fig. 2), which is the largest number of frames among the insect pose datasets mentioned above.

Based on our dataset, we evaluated the accuracy of two widely used animal pose estimation models, DeepLabCut [10] and SLEAP [12]. We also measured wingbeat frequency (WBF), which is a well-used trait to quantify flight behavior. In butterflies, it is known that non-toxic species fly with high WBF, whereas toxic species fly with low WBF [16], which may contribute to the learning of the color patterns of toxic species by avian predators. Because there have been no studies quantitatively comparing flight patterns in the wild, we compared them using ButterFlySet (Fig. 2).

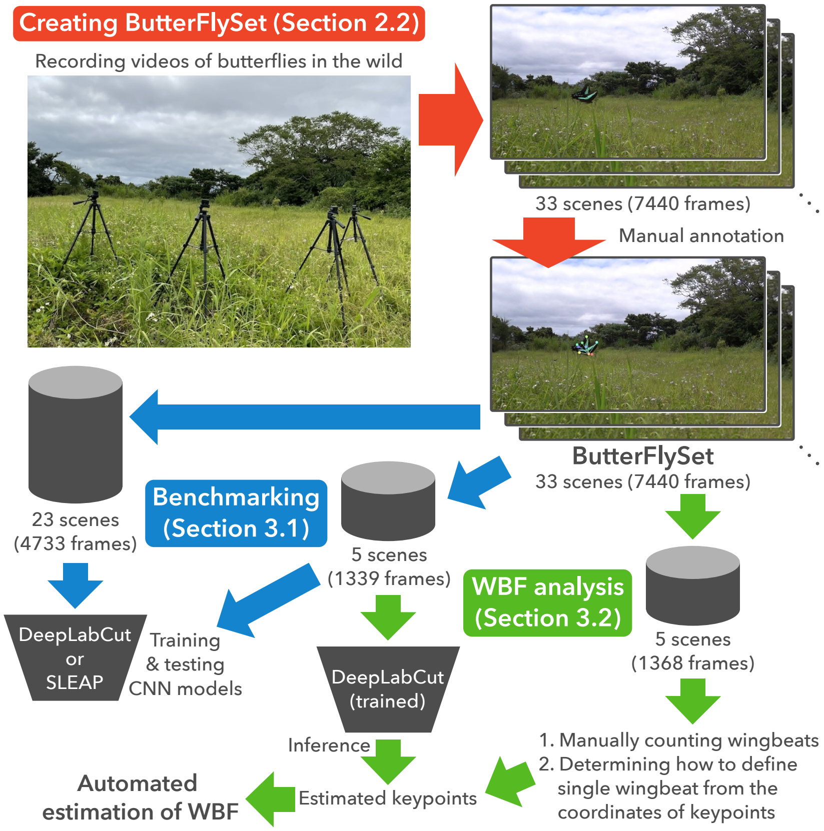

Fig. 2. Overview of our study. ButterFlySet was created using 7440 frames of videos recorded in the wild and manually annotated nine keypoints (red arrows). Using ButterFlySet, the accuracies of two CNN-based pose estimation models were evaluated (blue arrows). Using the inference by trained models and the definition of single wingbeat determined by mannual count of wingbeats in separated datasets, we automated the estimation of wingbeat frequency (WBF) from videos of flying butterflies in the wild (green arrows).

2 ButterFlySet: a pose video dataset of flying butterflies

2.1 Dataset characteristics

Although there are increasing numbers of 2D datasets on mammals [1], number of insect datasets is scarce, with only four [2]. In developing DeepLabCut, Mathis et al. [10] introduced a pose dataset of fruit flies, which consists of 589 frames with 12 keypoints. LEAP [11], the predecessor to SLEAP [12], has been evaluated using 'fly32' dataset, containing 32 keypoint annotations performed on 1500 frames collected from 59 videos of fruit flies. In addition to fly32, Pereirra et al. [12] introduced two multi-animal datasets, 'flies13' and 'bees' in developing SLEAP. The 'flies13' dataset contains 2000 frames of two fruit flies with 13 keypoints, whereas 'bees' contains 804 frames of two bumble bees with 21 keypoints. Compared to these four datasets for insect pose estimation, our ButterFlySet, which contains 7440 annotated frames, is the largest dataset in terms of frame numbers.

In terms of the computer vision analysis in butterflies, there have been several investigations on butterfly image recognition [17], which is accompanied by development of large-scale datasets. For example, 'Leeds Butterfly Dataset' [18], which contains 832 wild butterfly images, has been used by several studies benchmarking classification and segmentation accuracy [17]. Lin et al. [19] developed a dataset of 24836 butterfly specimen images for 56 species classification. Adityawan et al. [20] used a dataset of 13594 wild butterfly images from Kaggle repository for 100 species classification. However, there has been no dataset available for videos or pose estimation.

2.2 Dataset collection

Video data was collected between May 17 and May 20, 2024, in grassland and forest road near Mt. Otoha, Okinawa, Japan. We used digital still cameras (DSC-RX0M2, SONY, Tokyo, Japan) equipped on tripods, which were adjusted to a height corresponding to the flying height of butterflies (about $0.3 - 1.5$m above ground). The videos were recorded at 240 frames per second (fps) in XAVC S format with a resolution of 1920 × 1080 pixels. We recorded five species commonly observed in the Okinawa Island, including three non-toxic species; *Papilio polytes*, *Papilio memnon*, *Graphium sarpedon*, and two toxic species; *Ideopsis similis* and *Euploea mulciber*. We set multiple cameras from various angles to capture each butterfly individual freely flying in an open space from a closer distance and selected 33 scenes in which individuals are captured from close distances.

We manually annotated nine keypoints; head, thorax, abdomen, left/right forewing 1 (tip), left/right forewing 2 (anal edge), and left/right hindwing (Fig. 3). Some keypoints which were difficult to identify due to rapid movements and varying postures of the butterflies were estimated by human annotators.

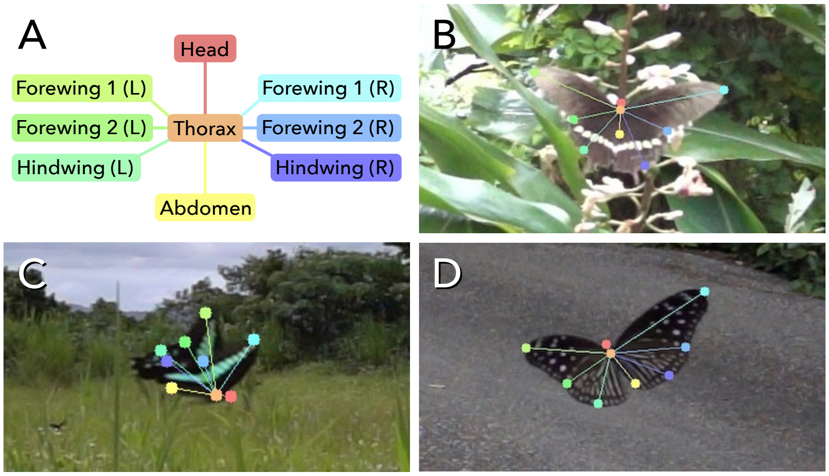

Fig. 3. The names of the nine keypoints (A) and the examples of annotated butterflies. (B) *Papilio polytes*. (C) *Graphium sarpedon*. (D) *Euploea mulciber*.

3 Experiments

3.1 Benchmarking ButterFlySet

To verify animal pose estimation performance, we evaluated two most widely used [2] models (DeepLabCut [10] and SLEAP [12]). For DeepLabcut, we used a single animal pose estimation pipeline with the backbone of ResNet50 with default hyperparameters. For SLEAP, we used U-Net backbone with 'baseline–large' hyperparameters, whose 'input scaling' was set to 0.30 because the receptive field size covers the size of butterfly individuals under this parameter. Out of 33 scenes, five pairs of scenes featuring same individual from two different angles (a total of 10 scenes) were removed, and half of these were used for testing. Remaining 23 scenes (4,733 frames) were used for training with a 70/30 train-validation split (Fig. 2). For DeepLabCut, we found that test error was 18.61 ± 82.8 pixels. A large standard deviation compared to the mean would suggest that in a small number of frames, keypoints were detected at locations significantly different from correct ones, such as in the background. For SLEAP, we found that the test error was 108.44 ± 271.3 pixels, which may be due to smaller size of our training data. It might be possible that default hyperparameter of DeepLabCut makes it more accurate than SLEAP, which means that further tuning may narrow gap between the two models. In addition, for SLEAP, we had to down-sample input frames in order to fit size of butterflies in 'receptive field size' of SLEAP, which might have lowered the model performance. To assess the detection performance, we also show Percent of Detected Joints (PDJ) [21] for nine keypoints (Table 1). Because DeepLabCut [10] exceeded SLEAP [12] in performance on ButterFlySet, we used the former model in the following WBF analysis.

Keypoints	Head	Thorax	Abdomen	FW1(L)	FW2(L)	HW(L)	FW1(R)	FW2(R)	HW(R)
DeepLabCut	99.5%	99.4%	99.5%	99.4%	99.5%	98.1%	98.7%	99.8%	99.6%
SLEAP	69.3%	69.6%	71.4%	71.6%	72.9%	73.0%	73.1%	72.4%	72.5%

Table 1. PDJ of nine keypoints. Keypoints are considered detected if the normalized distance between the predicted and ground truth keypoint is under 0.2 [21]. FW and HW stand for forewing and hindwing, respectively.

3.2 Wingbeat frequency (WBF) analysis

The overview of our analysis is shown in Figure 2. To estimate WBF automatically in the wild, at first, we manually identified 'Wings closed' frames, where the wings are most closed during a single wingbeat, from five scenes which were not used for training and testing phases. Using the ground truth of the coordinates data and reffering to the WBF mannually counted, we found that the frames when the difference in height between Forewing 1 (averaged left and right) and Thorax is greatest within ± 12 frames can detect 'Wings closed' within ± 3 frames with 0.826 F1–score. Using this threshold, we automatically estimated 'Wings closed' from five scenes used for the test phase, of which butterfly pose was estimated by DeepLabCut, and detected 44 frames with 0.893 F1–score.

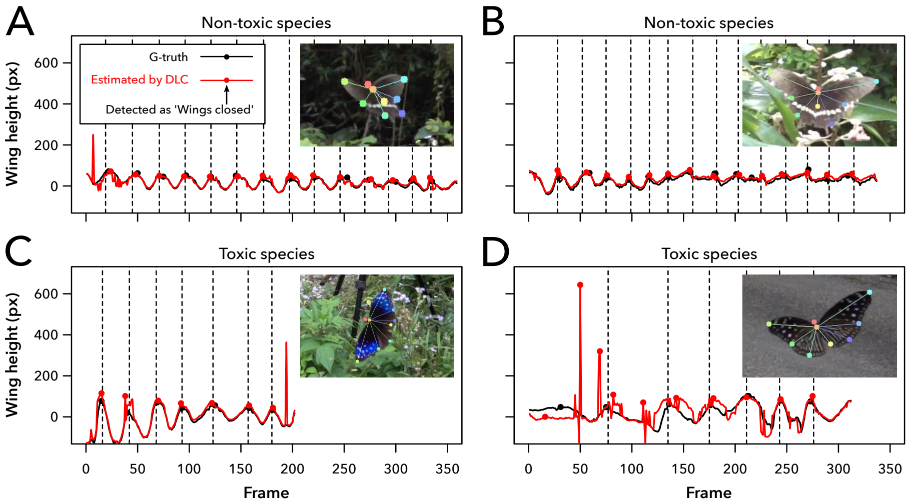

Fig. 4. Examples of the automated detection of 'Wings closed' frames in non-toxic (A, B) and toxic species (C, D). Dotted vertical lines represent the ground truth of 'Wings closed' frames identified by manual scoring.

As shown in Figure 4C and 4D, 'Wings closed' tended to be incorrectly detected when keypoints were detected with a large error, suggesting that the improvement in pose estimation accuracy may also raise the detection accuracy of ‘Wings closed.’ By defining the duration of a single wingbeat as the duration between ‘Wings closed’ frames, we calculated WBF of both toxic and non-toxic species, and found that the WBFs were 9.13 ± 3.13 Hz for toxic species and 10.46 ± 1.85 Hz for non-toxic species. Although it has relativelly small sample size (total of 43 wingbeats) and detection error of ‘Wings closed’ frames, it may suggest that toxic species fly slower than non-toxic species in the wild.

4 Conclusion

In summary, we introduced ButterFlySet, the largest video dataset for pose estimation in the wild insects. We benchmarked DeepLabCut and SLEAP on our dataset to assess the performance of existing approaches, and compared WBF between toxic and non-toxic butterflies. Our dataset can help researchers automatically examine the characteristics of butterflies’ flight behavior in the wild conditions, which conventional studies in the laboratory have not been able to access. Moreover, the relationship between the scale and diversity of our dataset and its impact on the improvement of deep learning models’ generalization performance is expected to provide valuable insights for future pose estimation of flying insects, other than butterflies, in the fields.

Acknowledgements We thank Mr. Tomohiro Suzuki for assistance in finding sites for video recording. We also thank Dr. Nozomi Nishiumi for help in designing the setup for recording. This study was supported by JSPS KAKENHI (grant no. JP24KJ0104 to KA and JP21H05300 to KF) and JST PRESTO (grant no. JPMJPR20CA) to KF.

References

1. Jiang, L., Lee, C., Teotia, D., Ostadabbas, S.: Animal pose estimation: A closer look at the state-of-the-art, existing gaps and opportunities. Computer Vision and Image Understanding. **222**, 103483 (2022)
2. Fazzari, E., Romano, D., Falchi, F., Stefanini, C.: Animal Behavior Analysis Methods Using Deep Learning: A Survey. arXiv preprint arXiv:2405.14002v1 (2024)
3. Cruz, T. L., Pérez, S. M., Chiappe, M. E.: Fast tuning of posture control by visual feedback underlies gaze stabilization in walking *Drosophila*. Current Biology. **31**(20), 4569–4607 (2021)
4. Polajnar, J., Kvinikadze, E., Harley, A. W., Malenovský, I.: Wing buzzing as a mechanism for generating vibrational signals in psyllids (Hemiptera: Psylloidea). Insect Science. (2024)
5. Labuguen, R., Matsumoto, J., Negrete, S. B., Nishimaru, H., Nishijo, H., Takada, M., Go, Y., Inoue, K., Shibata, T.: MacaquePose: A novel “in the wild” macaque monkey pose dataset for markerless motion capture. Frontiers in Behavioral Neuroscience. **14**, 581154 (2021)
6. Kearney, S., Li, W., Parsons, M., Kim, K. I., Cosker, D.: RGBD-Dog: Predicting canine pose from RGBD sensors. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 8333–8342 (2020)

7. Biggs, B., Boyne, O., Charles, J., Fitzgibbon, A., Cipolla, R.: Who left the dogs out? 3D animal reconstruction with expectation maximization in the loop.: In European Conference on Computer Vision, Springer. 195–211 (2020)
8. Cao, J., Tang, H., Fang, H. S., Shen, X., Lu, C., Tai, Y. W.: Cross-domain adaptation for animal pose estimation. Proceedings of the IEEE International Conference on Computer Vision. 2019-Octob. 9497–9506 (2019)
9. Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., Tao, D.: AP-10K: A Benchmark for Animal Pose Estimation in the Wild. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). (2021)
10. Mathis A., Mamidanna P., Cury K. M., Abe T., Murthy V. N., Mathis M. W., Bethge M.: DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nature Neuroscience. **21**(9), 1281–1289 (2018)
11. Pereira, T. D., Aldarondo, D. E., Willmore, L., Kislin, M., Wang, S. S. H., Murthy, M., Shaevitz, J. W.: Fast animal pose estimation using deep neural networks. **16**(1), 117–125 (2019)
12. Pereira, T. D., Tabris, N., Matsliah, A., Turner, D. M., Li, J., Ravindranath, S., Papadoyannis, E. S., Normand, E., Deutsch, D. S., Wang, Z. Y., McKenzie-Smith, G. C., Mitelut, C. C., Castro, M. D., D'Uva, J., Kislin, M., Sanes, D. H., Kocher, S. D., Wang, S. S. H., Falkner, A. L., Shaevitz, J. W., Murthy, M.: SLEAP: A deep learning system for multi-animal pose tracking. Nature Methods. **19**(4), 486–495 (2022)
13. Carducci, J. P., Jakob, E. M.: Rearing environment affects behaviour of jumping spiders. Animal Behaviour. **59**(1), 39–46 (2000)
14. Page, E., Queste, L. M., Rosser, N., Salazar, P. A., Nadeau, N. J., Mallet, J., Srygley, R. B., McMillan, W. O., Dasmahapatra, K. K.: Pervasive mimicry in flight behavior among aposematic butterflies. Proceedings of the National Academy of Sciences of the United States of America. **121**(11), e2300886121 (2024)
15. Dudley, R.: Biomechanics of flight in neotropical butterflies: Morphometrics and kinematics. Journal of Experimental Biology. **150**(1), 37–53 (1990)
16. Kitamura, T., Imafuku, M.: Behavioral Batesian mimicry involving intraspecific polymorphism in the butterfly *Papilio polytes*. Zoological Science. **27**(3), 217–221 (2010)
17. Yasmin, R., Das, A., Rozario, L. J., Islam, Md E.: Butterfly detection and classification techniques: A review. Intelligent Systems with Applications. **18**, 200214 (2023)
18. Wang, J., Markert, K., Everingham, M.: Learning models for object recognition from natural language descriptions. Proceedings of the British Machine Vision Conference. (2009)
19. Lin, Z., Jia, J., Gao, W., Huang, F.: Fine-grained visual categorization of butterfly specimens at sub-species level via a convolutional neural network with skip-connections. Neurocomputing. **384**, 295–313 (2020)
20. Adityawan, H. T., Farroq, O., Santosa, S., Islam, H. Md. M., Sarker, Md K., Setiadi, De R. I. M.: Butterflies recognition using enhanced transfer learning and data augmentation. Journal of Computing Theories and Applications. **1**(2), 115–128 (2023)
21. Toshevand, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. **7**, 1653–1660 (2014)