I-MuPPET: Interactive Multi-Pigeon Pose Estimation and Tracking *

Supplemental Material

Urs Waldmann[1,3][0000−0002−1626−9253], Hemal Naik[2,3,4,5][0000−0002−7627−1726], Nagy Máté[2,3,4,6,7,8][0000−0001−8817−087X], Fumihiro Kano[3,4][0000−0003−4534−6630], Iain D. Couzin[2,3,4][0000−0001−8556−4558], Oliver Deussen[1,3][0000−0001−5803−2185], and Bastian Goldlücke[1,3][0000−0003−3427−4029]

[1] Department of Computer and Information Science, University of Konstanz, Germany
[2] Department of Biology, University of Konstanz, Germany
[3] Centre for the Advanced Study of Collective Behaviour, University of Konstanz, Germany
[4] Department of Collective Behaviour, Max Planck Institute of Animal Behavior, Konstanz, Germany
[5] Computer Aided Medical Procedures, Technische Universität München, Munich, Germany
[6] MTA-ELTE "Lendület" Collective Behaviour Research Group, Hungarian Academy of Sciences, Budapest, Hungary
[7] MTA-ELTE Biological and Statistical Physics Research Group, Hungarian Academy of Sciences, Budapest, Hungary
[8] Department of Biological Physics, Eötvös Loránd University, Budapest, Hungary
Corresponding author: urs.waldmann@uni-konstanz.de

Abstract. In the supplemental material, we first provide some more information on our annotated single pigeon data set. Next, we report detailed results on our network training and ablation studies. Also, we briefly explain the metrics used in our main paper. Finally, we report more qualitative results on the I-MuPPET tracking performance. Please check out our videos on https://urs-waldmann.github.io/i-muppet/ for further insights into the multi-pigeon pose estimation and tracking part.

* This version of the contribution has been accepted for publication, after peer review but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/[insertDOI]. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms.

Table of Contents

1 Additional Information on Data Acquisition

In Fig. 1 you can see the imaging facility where our pigeon data was recorded.

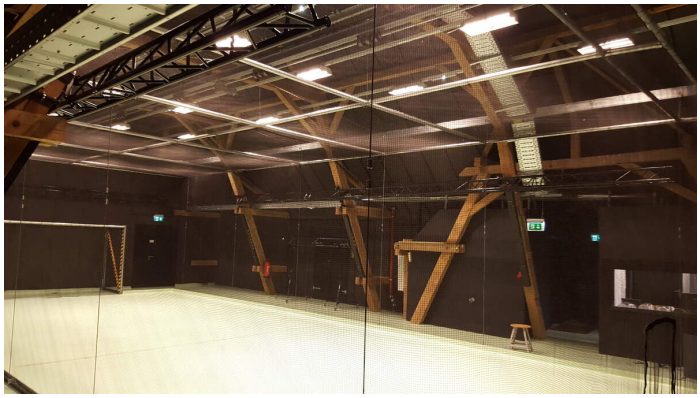

Fig. 1: *Imaging Facility for Data Acquisition.* The system consists of six Vue 2, four Vantage 5 and 26 Vero 2.2 sensors covering a volume of approximately $15 \times 7 \times 4$ meters. © CASCB Uni Konstanz.

Single Pigeon Data. The 27730 frames of our annotated single pigeon data set are from four recording sessions on two days where session one, two and four contain 183, 1614 and 2823 frames for each camera view respectively. Session three contains 9578 annotated frames for one view and 8912 for the other. This is due to the absence of the recorded pigeon in some frames in the second view.

2 Results on Network Training and Ablation Studies

In this section of our supplemental material, we give more detailed results on experiments on network training and ablation studies.

2.1 Data Augmentation for Pigeons

In Tab. 1 and Tab. 2 you find detailed results on experiments for data augmentation in the case of our annotated single pigeon data set. No change in brightness and a sharpness probability of 0.2 yield the best result (cf. Tab. 1). Scaling and flipping does not essentially improve results (cf. Tab. 2).

2.2 Data Augmentation for Cowbirds

In Tab. 3 you find detailed results on experiments for data augmentation in case of the single cowbird data from [1]. Randomly changing brightness by a factor chosen uniformly from $[0.7, 1.3]$ and a sharpness probability of 0.1 work best (cf. Tab. 3).

Table 1: *Ablation Study.* Data augmentation ablation study (single pigeon data) for the parameters brightness (b) and sharpness probability (sp). Framework trained on whole session four (s4) with batch size 20, learning rate 0.005, step size 10, gamma 0.5, number of epochs 100, no flipping and no scaling. Results are given as RMSE [px] for predictions where confidence score exceeds 0.999. s1, s2 and s3 denote the different recording sessions. *: No change in brightness.

config	s1	s2	s3
b = $[1, 1]^*$, sp = 0	25.1	6.4	9.7
b = $[0.7, 1.3]$, sp = 0.1	14.3	4.4	6.9
b = $[0.4, 1.6]$, sp = 0.1	12.7	4.5	6.6
b = $[0.7, 1.3]$, sp = 0.2	13.0	4.6	6.8
b = $[0.4, 1.6]$, sp = 0.2	13.3	4.6	6.9
b = $[0.4, 1.6]$, sp = 0	13.5	4.7	7.1
b= $[\mathbf{1}, \mathbf{1}]^*$, **sp**= **0.2**	17.0	**3.9**	6.7

Table 2: *Ablation Study.* Data augmentation ablation study (single pigeon data) for the parameters flip probability (fp) and scale range (sr). Framework trained on whole session four (s4) with batch size 40, learning rate 0.005, step size 77, gamma 0.7, number of epochs 500, brightness 0.6 and sharpness probability 0.2. Results are given as RMSE [px] for predictions where confidence score exceeds 0.999. s1, s2 and s3 denote the different recording sessions. No significant improvement within sessions.

config	s1	s2	s3
fp = 0, sr = $[50, 200]$	14.3	4.7	7.5
fp = 0.5, sr = $[75, 150]$	12.3	4.6	7.0
fp = 0.5, sr = $[90, 110]$	11.9	4.6	7.0
fp = 0.5, sr = $[78, 125]$	11.8	4.7	6.8

Table 3: *Ablation Study.* Data augmentation ablation study (cowbird data from [1]) for the parameters brightness (b), sharpness probability (sp), contrast (c), saturation (s) and hue (h). Framework trained on their training split with batch size 20, learning rate 0.005, step size 9, gamma 0.5, number of epochs 45, no flipping and no scaling. Results are given as PCK and evaluated on their test split. *: No change in brightness.

config	@0.05	@0.1
b = $[1, 1]^*$, sp = 0, c = 0, s = 0, h = 0	0.37	0.55
b = $[1, 1]^*$, sp = 0.1, c = 0, s = 0, h = 0	0.35	0.54
b = $[1, 1]^*$, sp = 0.2, c = 0, s = 0, h = 0	0.38	0.55
b= [0.7, 1.3], sp= 0.1, c= 0, s= 0, h= 0	**0.39**	**0.56**
b = [0.4, 1.6], sp = 0.2, c = 0, s = 0, h = 0	0.36	0.52
b = [0.7, 1.3], sp = 0, c = 0, s = 0, h = 0	0.37	0.56
b = [0.4, 1.6], sp = 0, c = 0, s = 0, h = 0	0.37	0.55
b = [0.7, 1.3], sp = 0.1, c = 0.2, s = 0, h = 0	0.38	0.55
b = [0.7, 1.3], sp = 0.1, c = 0.4, s = 0, h = 0	0.37	0.53
b = [0.7, 1.3], sp = 0.1, c = 0.6, s = 0, h = 0	0.37	0.55
b = [0.7, 1.3], sp = 0.1, c = 0.8, s = 0, h = 0	0.38	0.55
b = [0.7, 1.3], sp = 0.1, c = 0, s = 0.2, h = 0	0.38	0.54
b = [0.7, 1.3], sp = 0.1, c = 0, s = 0.4, h = 0	0.37	0.55
b = [0.7, 1.3], sp = 0.1, c = 0, s = 0.6, h = 0	0.38	0.54
b = [0.7, 1.3], sp = 0.1, c = 0, s = 0.8, h = 0	0.37	0.56
b = [0.7, 1.3], sp = 0.1, c = 0, s = 0, h = 0.1	0.38	0.55
b = [0.7, 1.3], sp = 0.1, c = 0, s = 0, h = 0.2	0.38	0.56
b = [0.7, 1.3], sp = 0.1, c = 0, s = 0, h = 0.3	0.37	0.56
b = [0.7, 1.3], sp = 0.1, c = 0, s = 0, h = 0.4	0.37	0.53
b = [0.7, 1.3], sp = 0.1, c = 0.2, s = 0.8, h = 0.2	0.38	0.55
b = [0.7, 1.3], sp = 0.1, c = 0.8, s = 0.8, h = 0.2	0.37	0.55

2.3 Training Hyperparameters

In Tab. 4 you find detailed results on experiments for hyperparameter tuning. A step size of 50 and a multiplicative factor of learning rate decay $\gamma = 0.5$ yield the best result (cf. Tab. 4).

Table 4: *Ablation Study.* Hyperparameter ablation study related to training for the parameters step size (sz) and γ. Framework trained on entire session four (s4) of our single pigeon data with batch size 20, learning rate 0.005, number of epochs 250, no change in brightness, sharpness probability 0.2, no flipping and no scaling. Results evaluated on 200 randomly sampled frames from session two (s2) for predictions where confidence score exceeds 0.999.

config	RMSE [px]
sz = 10, $\gamma = 0.5$	5.5
sz = 25, $\gamma = 0.5$	4.6
sz= 50, $\gamma = 0.5$	**3.8**
sz = 75, $\gamma = 0.5$	4.3
sz = 25, $\gamma = 0.7$	4.5
sz = 50, $\gamma = 0.7$	4.3
sz = 75, $\gamma = 0.7$	4.4
sz = 25, $\gamma = 0.95$	4.6
sz = 50, $\gamma = 0.95$	4.7
sz = 75, $\gamma = 0.95$	4.4

3 Metrics

In this section of our supplemental material, we briefly explain the metrics used in our main paper.

3.1 Pose Estimation

The RMSE is the L2 distance between the predicted and ground truth positions of keypoints. We average over samples and keypoints like [6].

The PCK is the percentage of predicted keypoints that fall within a normalized distance of the ground truth. This normalized distance in 3D Bird Reconstruction [1] is a fraction (0.05 and 0.1) of the largest dimension of the ground truth bounding box containing the bird and so do we use this, too.

3.2 Tracking

The CLEAR-MOT metrics are the Multi Object Tracking Accuracy (MOTA) and the Multi Object Tracking Precision (MOTP). MOTP is the total error

in estimated position for matched object-hypothesis pairs over all frames, averaged by the total number of matches made [2]. MOTA summarizes three sources of errors with a single performance measure, i.e. the ratio of misses in the sequence, computed over the total number of objects present in all frames, the ratio of false positives and the ratio of mismatches [2,4]. The track quality measures are classified as Recall, Precision, false positives per frame (FPF) mostly tracked (MT), partially tracked (PT) mostly lost (ML), fragments (Frag) and ID switches (IDS). Recall and Precision are the frame-based correctly matched objects divided by total ground truth objects and total output objects respectively [5]. MT and ML are the percentage of ground truth trajectories which are covered by tracker output for more than 80% and less than 20% in length [5]. Frag is the number of fragmentations where a track is interrupted by miss detection [3]. The trajectory-based metric IDF1 is the ratio of correctly identified detections over the average number of ground-truth and computed detections [7].

4 I-MuPPET Tracking Performance: Additional Qualitative Results

In this section of our supplemental material, we give more qualitative results on the I-MuPPET tracking performance.

I-MuPPET is trained on our single pigeon data set. Nevertheless we are able to estiamte and track poses for multiple pigeons. Please check out our videos for further insights into the multi-pigeon pose estimation and tracking part.

We also train I-MuPPET on the odor trail tracking data set from DeepLabCut [6].

Videos can be found at https://urs-waldmann.github.io/i-muppet/.

References

1. Badger, M., Wang, Y., Modh, A., Perkes, A., Kolotouros, N., Pfrommer, B.G., Schmidt, M.F., Daniilidis, K.: 3d bird reconstruction: A dataset, model, and shape recovery from a single view. In: ECCV. pp. 1–17 (2020)
2. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing **2008**, 1–10 (2008)
3. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: ICIP. pp. 3464–3468 (2016)
4. Dendorfer, P., Osep, A., Milan, A., Schindler, K., Cremers, D., Reid, I., Roth, S., Leal-Taixé, L.: Motchallenge: A benchmark for single-camera multiple target tracking. IJCV **129**(4), 845–881 (2021)
5. Li, Y., Huang, C., Nevatia, R.: Learning to associate: Hybridboosted multi-target tracker for crowded scene. In: CVPR. pp. 2953–2960 (2009)
6. 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. Nat. Neurosci. **21**, 1281–1289 (2018)
7. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV. pp. 17–35 (2016)