3D-MuPPET: 3D Multi-Pigeon Pose Estimation and Tracking

Supplemental Material

Urs Waldmann[1,2*†], Alex Hoi Hang Chan[2,3,4*†], Hemal Naik[2,4,5], Máté Nagy[2,3,4,6,7], Iain D. Couzin[2,3,4], Oliver Deussen[1,2], Bastian Goldluecke[1,2], Fumihiro Kano[2,3,4]

[1]Department of Computer and Information Science, University of Konstanz, Germany.
[2]Centre for the Advanced Study of Collective Behaviour, University of Konstanz, Germany.
[3]Department of Collective Behavior, Max Planck Institute of Animal Behavior, Konstanz, Germany.
[4]Department of Biology, University of Konstanz, Germany.
[5]Department of Ecology of Animal Societies, Max Planck Institute of Animal Behavior, Konstanz, Germany.
[6]Department of Biological Physics, Eötvös Loránd University, Budapest, Hungary.
[7]MTA-ELTE 'Lendület' Collective Behaviour Research Group, Hungarian Academy of Sciences, Budapest, Hungary.

*Corresponding author(s). E-mail(s): urs.waldmann@uni-konstanz.de; hoi-hang.chan@uni-konstanz.de;
Contributing authors: hnaik@ab.mpg.de; nagymate@hal.elte.hu; icouzin@ab.mpg.de; Oliver.Deussen@uni-konstanz.de; bastian.goldluecke@uni-konstanz.de; fumihiro.kano@uni-konstanz.de;
†These authors contributed equally to this work.

Abstract

In the supplemental material, we report detailed results on our network training and ablation studies. Also, we briefly explain the metrics used in our main paper.

1 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.

1.1 Data Augmentation for Pigeons

For data augmentation for the KeypointR-CNN (He, Gkioxari, Dollar, & Girshick, 2017) we find that changing brightness, flipping or scaling do not enhance performance, but changing sharpness with a probability of 0.2 results in the best performance in terms of RMSE (for numbers cf. Tabs. 1 and 2). This is intuitive since we train on the single pigeon data where the training data already contains a wide range of different pigeon positions and lighting conditions and thus covers most of the scaling and brightness. Also, the training data already include most body orientations (with respect to the camera), thus flipping does not improve test accuracy. Since the depth of field of the cameras is limited the pigeons are sometimes slightly out of focus and therefore blurring the input image with a small probability of 0.2 improves the accuracy of the test set.

In the case of multi-pigeon video sequences, however, we find that the best data augmentation parameters are not the same as for the single pigeon data. We keep the parameters from the single pigeon analysis but find that randomly jittering brightness by a factor chosen uniformly from $[0.4, 1.6]$ and a flipping probability of 0.5 produces the best outcome. This is intuitive because the single pigeon data does not cover the range of brightness found in the multi-pigeon data and the flipping makes the pose estimation in new situations more robust. A small scaling range of $\pm 5\%$ is sufficient since the single pigeon data covers already a large range of pigeon sizes. Also, if the scaling range is too large, we find multiple (mis-)detections if pigeons are nearby. This is also the case in situations where the pigeons occlude or are close to each other even if we do not apply scaling.

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= $[1, 1]^*$, 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

1.2 Data Augmentation for Cowbirds

In Tab. 3 you find detailed results on experiments for data augmentation in case of the single cowbird data from Badger et al. (2020). 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).

1.3 Training Hyperparameters

In Tab. 4 you find detailed results on experiments for hyperparameter tuning for the KeypointR-CNN (He et al., 2017). A step size of 50 and a multiplicative factor of learning rate decay $\gamma = 0.5$ yield the best result (cf. Tab. 4).

2 Metrics

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

2.1 Pose Estimation

The RMSE is the L2 distance between the predicted and ground truth positions of keypoints. We average over samples and keypoints like Mathis et al. (2018).

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 (Badger et al., 2020) 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, in our comparison on the cowbird data. For our comparison on the pigeon data instead, the normalized distance is again a fraction (0.05 and 0.1) of the largest dimension of the ground truth bounding box for the 2D evaluation and the maximum distance between any two ground truth keypoints for each individual in 3D.

2.2 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 (Bernardin & Stiefelhagen, 2008). 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 (Bernardin & Stiefelhagen, 2008; Dendorfer et al., 2021). 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

Table 3 *Ablation Study.* Data augmentation ablation study (cowbird data from Badger et al. (2020)) 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

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 the 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, γ = 0.5	5.5
sz = 25, γ = 0.5	4.6
sz= 50, γ = 0.5	**3.8**
sz = 75, γ = 0.5	4.3
sz = 25, γ = 0.7	4.5
sz = 50, γ = 0.7	4.3
sz = 75, γ = 0.7	4.4
sz = 25, γ = 0.95	4.6
sz = 50, γ = 0.95	4.7
sz = 75, γ = 0.95	4.4

total output objects respectively (Li, Huang, & Nevatia, 2009). 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 (Li et al., 2009). Frag is the number of fragmentations where a track is interrupted by miss detection (Bewley, Ge, Ott, Ramos, & Upcroft, 2016). The trajectory-based metric IDF1 is the ratio of correctly identified detections over the average number of ground-truth and computed detections (Ristani, Solera, Zou, Cucchiara, & Tomasi, 2016).

References

Badger, M., Wang, Y., Modh, A., Perkes, A., Kolotouros, N., Pfrommer, B.G., ... Daniilidis, K. (2020). 3d bird reconstruction: A dataset, model, and shape recovery from a single view. *Eur. conf. comput. vis.* (pp. 1–17).

Bernardin, K., & Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, *2008*, 1–10,

Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B. (2016). Simple online and realtime tracking. *Ieee int. conf. image process.* (p. 3464-3468).

Dendorfer, P., Osep, A., Milan, A., Schindler, K., Cremers, D., Reid, I., ... Leal-Taixé, L. (2021). Motchallenge: A benchmark for single-camera multiple target tracking. *Int. J. Comput. Vis.*, *129*(4), 845–881,

He, K., Gkioxari, G., Dollar, P., Girshick, R. (2017). Mask r-cnn. *Int. conf. comput. vis.*

Li, Y., Huang, C., Nevatia, R. (2009). Learning to associate: Hybridboosted multi-target tracker for crowded scene. *Ieee conf. comput. vis. pattern recog.* (p. 2953-2960).

Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., Bethge, M. (2018). Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.*, *21*, 1281–1289,

Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. *Eur. conf. comput. vis.* (pp. 17–35).