Supplementary Materials: 3D-POP - An automated annotation approach to facilitate markerless 2D-3D tracking of freely moving birds

March 21, 2023

Abstract

The following text is supplementary text for the paper "3D-POP - An automated annotation approach to facilitate markerless 2D-3D tracking of freely moving birds with marker-based motion capture". The text includes details of methods and results that are not part of the main text which are described in more detail here. We also outline detailed method that can be helpful to replicate the setup and annotation process, especially for biologists.

1 Affiliations

Here, we provide complete affiliations for the authors from the main text, with identical numbering.

1. Department of Collective Behaviour and Department of Ecology of Animal Societies, Max Planck Institute of Animal Behavior, 78464 Konstanz, Germany.

2. Department of Biology, University of Konstanz, 78464 Konstanz, Germany.

3. Centre for the Advanced Study of Collective Behaviour, University of Konstanz, 78464 Konstanz, Germany.

4. Computer Aided Medial Procedures, Informatik Department, Technische Universität München, Boltzmannstraße 3, 85748, Garching bei München, Germany.

5. Department of Biological Physics, Eötvös Loránd University, Pázmány Péter sétány 1A, Budapest 1117, Hungary.

6. MTA-ELTE 'Lendület' Collective Behaviour Research Group, Hungarian Academy of Sciences, Budapest 1117, Hungary.

2 Supplementary Methods

2.1 Camera Calibration:

The Infrared cameras of the vicon motion capture system (Vicon Vero,Vantage) are calibrated using built in software (Vicon Nexus) with a calibration wand. All cameras are also time synchronized when recording. The calibration of the Vicon system forms the basis of the whole dataset, in which we use the vicon coordinate system for all 3D coordinates.

For the 4 high definition RGB action cameras, we performed intrinsic calibration, extrinsic calibration and time synchronization independently.

2.1.1 Intrinsic and Extrinsic Calibration

For intrinsic calibration, we used an A0 (84.1 x 118.9 cm) charuco checkerboard before each day of recordings and undistorted all videos from each camera view using the obtained distortion and camera matrix from opencv.

For extrinsic calibration, we adopted a subject based approach, where we manually annotated the 2D position of a motion capture markers visible in the image (*e.g.*, backpack marker) on a moving pigeon subject over up to 30 frames. For each frame, we compute camera pose using 2D marker positions on the backpack and the 3D coordinates of the backpack in the vicon coordinate system. The combination of both provide us the extrinsic parameters for each camera. We ensure that sampled 3D positions are well distributed in the tracking volume to avoid bias in extrinsic parameters. This approach is useful as it allows us to move the tripod positions between sessions and perform fast extrinsic calibration without using the checkerboard.

In the future, we plan to make the method for marker selection automatic, which would improve accuracy as the system will recompute extrinsics in real-time and change in camera position would not require calibration.

2.2 Temporal synchronization

To synchronize the RGB action cameras, we attached a camera control box (Sony CBB-WD1) to each action camera. The control boxes has built-in functionality to synchronize video streams from multiple cameras over ethernet and a network switch.

Since our data are collected by two independent systems (Motion tracking system and RGB cameras), we designed an arduino based synchronization device with 3 RGB LED lights and 2 infra-red lights that blinks for 1 second at 5 second intervals. For the RGB videos, we computed the change in maximum pixel values of the cropped box area through time, and detected light flashes based on a change of more than 30 units. For the vicon system, we attached 4 additional markers onto the arduino box, which allow a new object to be defined within the mo-cap software together with the 2 infra-red lights. Flashes can then be detected based on the number of markers present in the object (6 markers for flash on, 4 markers for flash off). The systems are then temporally synchronized by matching the detected light flashes from both systems. In cases where a light flash were not detected, we manually filled in the flash by assuming the flash is exactly 6 seconds after the previous one.

2.3 Markerless Data

POP-3D dataset contains ground truth annotation for pigeons with markers attached to their body. However, we expect this dataset to play a role in development of markeless algorithms for tracking pigeons and other birds. In experiment 2, we show that models trained on 3D-POP dataset do work well with pigeons recorded in the same arena without any markers. To validate results in future methods, 3D-POP dataset also includes trials of freely moving birds without any marker attachment (n = 1,2,5,11). The videos will serve useful for making qualitative claim about performance of algorithms developed with mocap annotated posture or identity data. It is worth noting that we only demonstrate that position of markers do not play role in prediction of keypoints but do not show same validation for problem of identity recognition. Markers may play a role in identity recognition, we would like to test it in future work. Please see Table S1 for more details of sessions recorded with pigeons without markers.

No. individuals	Available frames	Video length (min)
1	36,825	20
2	36,600	20
5	9,810	5.5
11	36,225	20

Table S1: Markerless Data Summary: Total number of frames and video length for markerless data.

2.4 Post-processing Mo-cap data

The data obtained from motion capture is often not directly usable for the annotation pipeline. It requires a post-processing step to ensure smooth annotation process. Marker-based motion capture has a limitations related to tracking loss and marker identification (for 6-DOF tracking). This results in error of identification of pigeons and 6-DOF pose of rigid bodies defined by the mo-cap system. The problem stems from limited availability of space on the pigeon body. The markers are forced to be close to each other (in 4-marker pattern) which leads to error in correspondence matching required for pose computation. To solve for these changes, we applied a post-processing pipeline introduced by Kano *et al.*[1] to fix mis-labelled frames by detecting large changes in the distance between defined markers, then determining the correct labels through permutation techniques. Detailed description and code used is provided in Kano *et al.*[1].

2.5 Computing Pose Variation:

Computing variation in pose is important to understand how many different postures are recorded in the dataset. Definition of coordinate system defined for each pigeon's body part (head, body) is slightly different because the definition is based on marker positions, which is different for each pigeon each day. Therefore, the orientation angles defined by the 6-DOF pose w.r.t the canonical frame (world coordinate system) are different for each pigeon even if absolute posture of pigeons is the same. This problem does not allow us to compute variation of pose for each pigeon in a standard manner. We argue that pigeon posture can be measured in a standard way if orientation of the body parts are defined using positions of 3D keypoints in a unified coordinate system *i.e.*, world coordinate system. The keypoints such as beak, eyes or shoulders are common features recorded for each bird and thus a posture representation involving these features would provide means for comparing postures of different pigeons. Using this logic we designed a new technique to measure 3D orientation (rotational angles) of pigeon body parts relative to each axis. Please refer Figure S1 for a pictorial representation.

In this text, we will describe the steps to obtain rotation angles of the head in the world coordinate system. Firstly, we select beak as primary keypoint and shift the origin of world coordinate system to this point (translation). This is done because we are interested in comparing rotation units and shifting origin reduces complexity of pose representation from 6-DOF (rotation and translation) to 3-DOF (rotation). In other words, we change the representation of a rigid body (pigeon head) from 6-DOF pose representation to a plane representation (passing through origin). The plane is defined using 3D positions of beak, left eye and right eye (Figure S1). The normal of the plane is the cross product of two vectors originating from beak to left and right eye. This normal

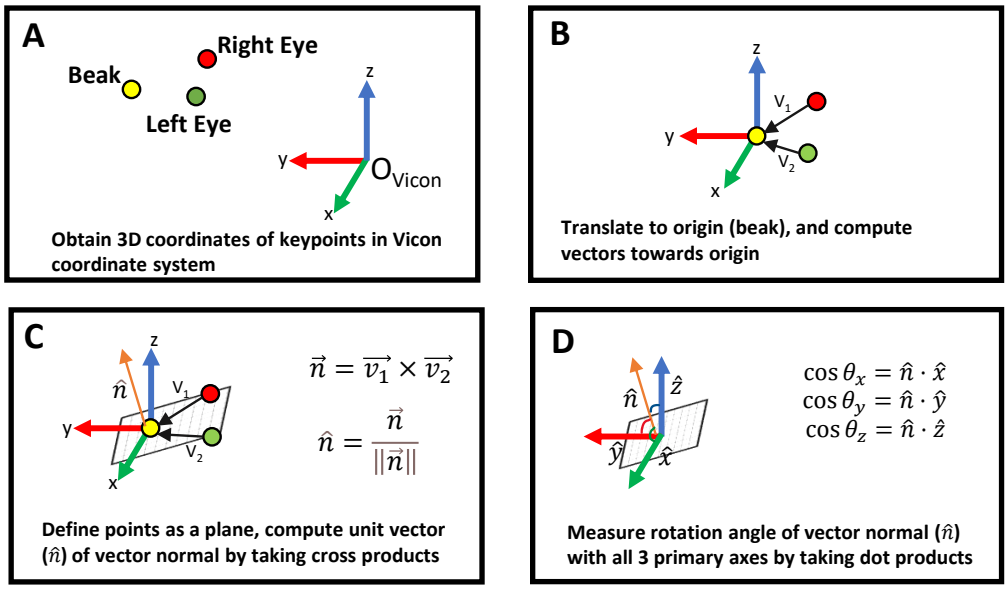

Figure S1: Schematic for computing 3D orientation angles of the pigeon's body parts (head and body) for a given frame. A) Obtain 3D coordinates of keypoints in the world coordinate system (vicon). B) Translate all keypoints to new origin (beak) and compute vectors towards the origin using keypoints. C) Define a plane and compute surface normal. D) Calculate angle between surface normals and the 3 primary axes.

is defined at the origin and it's angles with respect to the primary x-y-z axis, which represent one unique head orientation in the world coordinate system. This process is repeated to compute head posture of all pigeons in all sequences. The comparison between all posture is only possible because the world coordinate system is consistent for all sequences. Finally, we show a histogram of the occurrences of the different rotation angles to indicate pose variation (Figure S2). The same process is repeated for body pose by defining a plane using shoulder and tail keypoints, with tail as origin.

2.6 Dataset comparison

There are many datasets available with animals that target one or more problems. We provide a short overview 18 different datasets and provide a table at the end of this text as auxiliary information.

3 Experimental results

3.1 Outlier Detection and Filtering Pipeline

Our outlier detection and filtering pipeline introduces gaps in the dataset. Here is a quantification of the number and length of gaps that are present in the dataset. (Table S2)

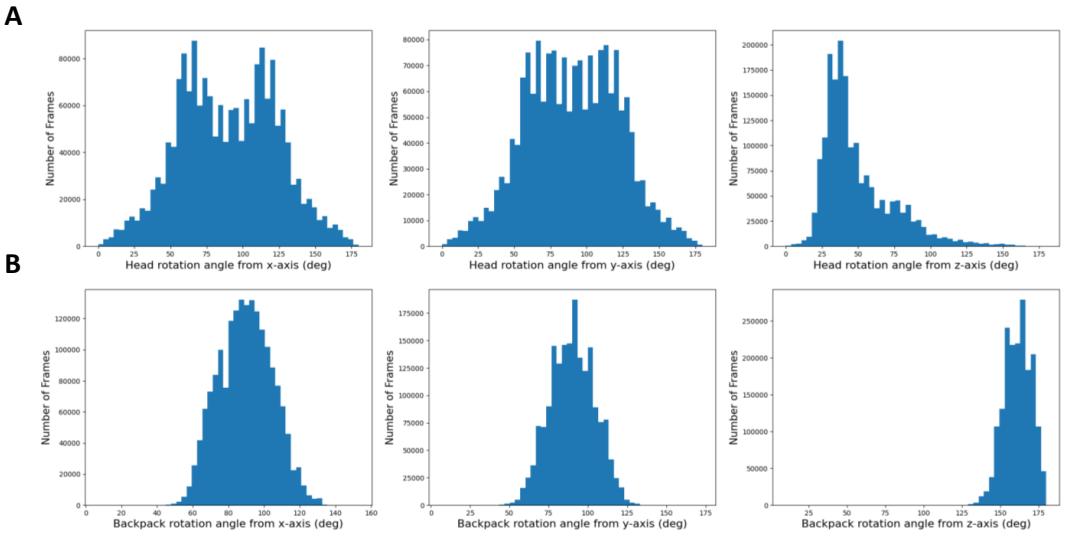

Figure S2: Frame distribution of the head and backpack rotation angles with respect to the 3 primary axes of pigeon subjects present in the dataset. A) Distribution of head rotation angles. B) distribution of backpack rotation angles.

1 frame	2-30 frames	>30 frames
3585	5062	351

Table S2: Frame length of consecutive gaps present in the dataset

3.2 Experiment 1 - Hybrid Approach:

In the paper, we performed an experiment to show that markerless tracking is possible for pigeons using the 3D-POP dataset. First, experiment was performed with pigeons with marker attached, like experimental scenario of Kano *et al.*[1]. We claim that markerless solution is directly useful to improve tracking performance for cases where motion capture is already in use. We took a sample data from our recording sequence and introduced gaps in trajectories to simulate loss of tracking, then further used markerless approach to fill the gaps to evaluate quality of markerless 3D tracking in comparison with ground truth. The results are demonstrated in Table S3, which show that using a 2D keypoint detection model and simple triangulation, the gap filling algorithm provides good accuracy. This is useful already for experiments where missing data with mo-cap setup is a consistent problem. We also show a simple comparison with interpolation approach to show that markerless solution have higher chance of filling gaps than data interpolation methods. In future work, we want to try different strategies to combine multi-view data to get higher prediction accuracy.

$RMSE_{Method}$ (mm)	Beak	Nose	Left Eye	Right Eye	Left Shoulder	Right Shoulder	Top Keel	Bottom Keel	Tail
$RMSE_{Hybrid}$	8.2	6.5	7.3	6.3	13.8	9.4	13.3	8.9	9.2
$RMSE_{Linear}$	66.3	64.8	62.5	62.7	45.5	42.7	41.3	37.6	45.7

Table S3: Root mean squared Euclidean error (mm) of different approaches used to fill artificially introduced gaps in a 5 min single pigeon sequence. Hybrid approach uses a 2D DLC model from each view and triangulated and the linear approach interpolates missing data linearly with data before and after a given gap.

4 Additional Files

We have provided additional material along with this file.

- Supplementary video : The video is designed to introduce the reader with 3D-POP dataset. It shows the setup, diversity of the dataset and the annotations on video images.

  Click here for youtube link of the video.

References

[1] Fumihiro Kano, Hemal Naik, Göksel Keskin, Iain D. Couzin, and Máté Nagy. Head-tracking of freely-behaving pigeons in a motion-capture system reveals the selective use of visual field regions. *Scientific Reports*, 12(1):19113, Nov 2022.

Authors	Dataset name	No of. Individuals	No. of views	No. of species	Format	Annotation method	Annotations	Behavior	Task (see below)
Graving et al.	Zebra	1	1	1	Image (900)	Manual (semi-automated)	2D keypoints	NA	1
Wah et al.	CUB-200-2011	1	1	200	Image (11788)	Manual	Bouding box Part locations Part attributes Silhouette	NA	7
Mathis et al.	Horse-30	1	1	1	Image (8114)	Manual	2D keypoints	NA	1
Joska el al.	Acinoset	1	1	1	Image (7588)	Manual	2D keypoint 3D Triangulation	Running	1
Ng et al.	Animal Kingdom	1	1	850	Image (33k) + Video (50Hrs)	Manual	2D keypoints, Activity	Varied	1
Bala et al.	Open Monkey Studio	1	62	1	Image (195,228)	Manual	2D keypoint 3D Triangulation	NA	1,3
Dunn el al.	RAT7M	1	6	1	Video (7 million)	Automated (mocap)	2D keypoints 3D keypoint 2D-3D trajectories	13 Categories using clustering algorithm	1,2,3,8
Kearney et al.	RGB-D Dog	1	20 RGB + 6 RGBD	1	Video (73,748)	Automated (mocap)	2D keypoints 3D keypoints RGB-D, Mask	NA	1,2,8
Yao et al.	Open Monkey Challenge	1	Mixed (Single + Multiview)	26 Species (Primates)	Images + Videos	Manual (semi-automated)	2D Keypoint	NA	1,7

Authors	Dataset name	No of. Individuals	No. of views	No. of species	Format	Annotation method	Annotations	Behavior	Task (see below)
Yu et al.	AP-10K	1	1	54 (23 Families - mammals)	Image (10,015)	Manual	2D keypoints Bounding box Background	NA	1
Laurel et al.	TRI-MOUSE	3	1	1	Video (11,645)	Manual	2D keypoints	NA	1,2,4
Laurel et al.	PARENTING	2	1	1	Video (2670)	Manual	2D keypoint	NA	1,2,4
Laurel et al.	MARMOSETS	2	1	1	Video (15000)	Manual	2D keypoint Identity	NA	1,2,4
Laurel et al.	FISH	14	1	1	Video (1,100)	Manual	2D keypoint	NA	1,2,4
Labugen et al.	MacaquePose	>= 1	1	1	Image (13,083)	Manual	2D keypoints	NA	1,4,5
Badger el al.	Cowbird dataset	<= 15	8	1	Image (1000)	Manual	2D keypoints Silhouettes	Multiple interactions	1,7,8
Marshall et al.	PAIR-R24M	2	24	1	Video (24.3 million)	Motion capture (Automated)	2D-3D keypoints Behavior interactions	11 behavior, 3 interaction categories	1,2,3
Ours	**POP-3D**	**1-2-5-10**	**4**	**1**	**Video (300K frames)**	**Semi-automatic (manual + mocap)**	**2D keypoints 2D trajectory 3D keypoint 3D trajectory identity**	**NA**	**1,2,4,5,6,7**

Task	Description
1	2D Pose Estimation
2	3D Pose Estimation
3	Activity recognition
4	Identity tracking/ Reidentification
5	2D trajectory tracking
6	3D trajectory tracking
7	Object detection
8	3D reconstruction