DIPLOMAT: A tool for multi-animal tracking

Isaac Robinson[1], George Glidden[1], Neekesh Panchal[2], Nathan Insel[2], Travis Wheeler[1]

[1]R. Ken Coit College of Pharmacy, University of Arizona, Tucson, AZ
[2]Department of Psychology, Wilfrid Laurier University, Waterloo, ON

Abstract

Recent advances in computer vision have enabled the development of automated animal behavior observation tools. Despite their generally encouraging performance, multi-animal tracking tools still face challenges, particularly with "body swapping" – failure to maintain identities across time. Here we present DIPLOMAT, multi-animal tracking software tool that greatly reduces identity assignment errors by introducing a combination of (i) an automated pose estimate post-processing algorithm ("Track") and (ii) an graphical interface for efficient human supervision ("Interact"). Evaluation involving recordings of multiple moving mice shows that DIPLOMAT's automated method yields reductions in identity swaps of 80 to 95% relative to leading methods, and that these can then be almost entirely eliminated with time-efficient human editing.

1 Introduction

Recent technological developments, including advances in computer vision, have paved the way for objective measurement of high-dimensional, complex behavior. This is particularly valuable for studies of social interaction, which are by nature complex, variable, and dependent on naturalistic contexts. Relevant social information can often be signaled through subtle, moment-to-moment movements of specific muscles, such as facial expressions or posture. To understand the nature of social interaction, it is therefore necessary to use measures that take into account nuanced pose and movement changes. This can be accomplished by tracking body parts in two or more individuals over time, but only works if the tracking methods are consistent about which individual the body parts belong to.

Several tools for multi-animal tracking have recently become available, including SLEAP [1], multi-animal DeepLabCut [2], AlphaTracker [3], ID tracker [4], and TRex [5]. Two of these, DLC and SLEAP, have caught-on as particularly popular and powerful for tracking dyads and triads of rodents using a single-camera video source (Shemesh and Chen, 2023; Luxem et al., 2023; Bordes et al., 2023). Both apply convolutional neural networks to detect animal body parts, paired with differing algorithms to then assign animal identities (generally based on associations between the body parts, "skeletons"). Despite their generally encouraging performance, multi-animal tracking tools still face challenges, particularly with "body swapping" – failure to maintain identities across time when presented with similar-looking, sometimes-overlapping bodies.

To overcome these body-swapping challenges, we have developed DIPLOMAT, a Deep learning-based, Identity-Preserving, Labeled-Object Multi-Animal Tracker. DIPLOMAT introduces automated algorithms ("Track") and an efficient human interface ("Interact") to jointly eliminate identity assignment errors.

Tracking in DIPLOMAT begins by running either SLEAP or DeepLabCut to produce frame-by-frame distributions of the probabilities of possible locations for body parts. It then treats the movement of each individual as a Markov process, identifying a maximum probability (Viterbi) trace for each individual's body parts based on a custom hidden Markov model. Tracking accuracy is improved by application of skeletal constraints (which enforce body part proximity), along with a novel signal dampening strategy for ensuring mutually independent traces (see Methods). Efficient parallel computation ensures that run time for DIPLOMAT's *Track* stage is less than that of the initial pose estimation stages of DeepLabCut or SLEAP.

Following automated tracking, a researcher may use DIPLOMAT's *Interact* tool to identify potential errors, then correct those errors with a simple point-and-click error interface that enables rapid correction of multiple body parts at a time; a small number of user edits are smoothly integrated with automated algorithms for re-tracking.

Testing against stand-alone DeepLabCut or SLEAP re-

veals that DIPLOMAT's automated method alone results in statistically-confirmed reductions in identity swaps of 80 to 95%, and that these can then be eliminated with time-efficient human editing. DIPLOMAT therefore has the potential to streamline analysis of multi-animal interaction, and can be used on its own or as a complement to other methods to improve multi-animal tracking accuracy and workstream efficiency. By design, DIPLOMAT is agnostic to the source of body part maps, and can be easily extended to wrap other pose estimation tools that may appear in the future.

2 Methods

DIPLOMAT is developed as two major components. The first component, *Track*, uses either SLEAP or DeepLabCut to produce body part location heatmaps for each video frame (CNN-based probabilities that each pixel in the frame is the location of each kind of body part), then traces all animals through the full video by identifying mutually-exclusive maximum probability traces through those heatmaps. The second component, *Interact*, is a smooth and intuitive user interface for quick editing of multiple body parts across video frames, with the ability to re-integrate these edits to a quickly and smoothly re-track body parts and, when necessary, reassign identities.

The methods underlying DIPLOMAT are guided by 4 assumptions:

1. **Bodies don't teleport.** This means that a previous video frame can be used as a prior that informs probability distributions for current body part locations.

2. **Body parts stick together.** Skeletal information (distance between body parts) can further inform a posterior probability of body part location.

3. **Bodies will be easier to distinguish in some frames over others.** Some frames can serve as an anchor, with probabilities extending forward and backward from that point.

4. **Fixed body count.** In a given experiment, there is typically a pre-set number of bodies.

Although not all of these assumptions will be true for every application, they cover a wide scope of animal behavior and neuroscience protocols.

2.1 Track

The workflow for the automated *Track* component is presented in the top two boxes of Figure 1.

2.1.1 Train and apply the pose estimation model

DIPLOMAT's *Track* stage depends on a per-frame pose estimation heat map. This requires that the user select a pose estimation tool (currently choosing between SLEAP and DeepLabCut), and use the model training methods provided by that tool. This requires that the user train a tool-specific model by labeling a modest number of frames (typically 10-1000 frames). When DIPLOMAT is run, the trained model is supplied along with the tracking video, and the tool produces the requisite per-frame heat map of body part placements for each body part. Let B be the set of labeled body parts (e.g. $B = \{$ nose, left ear, right ear, tail base$\}$), and F be the number of frames in a video recording with N animals. Then for each frame $i \in \{1..F\}$, and each body pat type $b \in B$, the pose estimation tool produces a softmax-based pseudo-probability $P_{i,b}(x, y)$ for each position (x, y).

2.1.2 Identify an anchor frame.

For each frame, DIPLOMAT computes a measure of the quality of separation of the animals in the frame. First, a median distance is computed between each pair of body parts (the skeleton), across all frames. Then for each frame, an estimate is computed of the number of parts that can be reliably paired with each other according to those median distances. The frame with greatest separation reliability is chosen as an "anchor" frame.

2.1.3 Compute a maximum probability trace

Computing a most-probable trace requires definition of a probability model. In DIPLOMAT, the base form of this model motivates a recurrence in which a probability is computed that the true path for body part b for individual k runs through position (x, y) in frame i: $V_{i,b,k}(x, y)$. This probability is computed as a product of (i) the per-frame softmax-based pseudo-probability from the base model, (ii) the probabilities of all possible paths up to the preceding frame, and (iii) the probability T of transitioning (moving) from position (x', y') in one frame to position (x, y) in the next:

$$V_{i,b,k}(x, y) = \max_{x',y'} \big(V_{i-1,b,k}(x', y') \cdot T(x', y', x, y) \cdot P_{i,b}(x, y)\big) \quad (1)$$

The transition probability $T(x', y', x, y)$ is similar to a Gaussian distribution with mean at (x', y') and standard deviation based on the median body size, to discourage unreasonably large moves. Functionally, this path-based

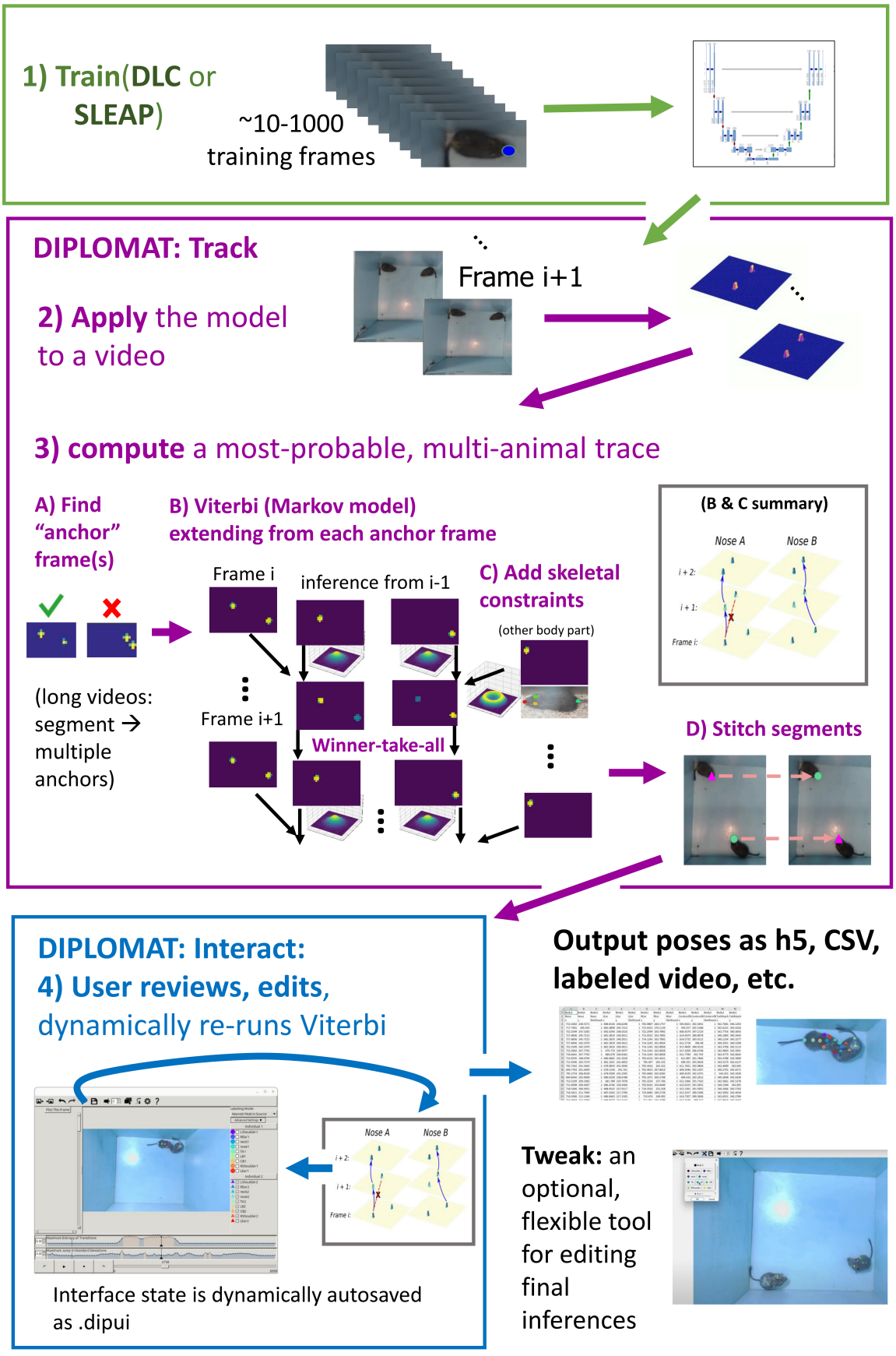

Figure 1: Flowchart illustrating the sequence of steps involved in multi-animal tracking using DIPLOMAT.

approach means that instead of using simple probability peaks, the algorithm identifies the points that optimize movement through probability fields across frames.

The recurrence allows for a natural dynamic programming tabular computation, such that the probabilities V for each frame can be computed based only on information from the previous frame. Because DIPLOMAT begins the trace at an anchor frame that is generally neither the first nor last frame, the dynamic programming implementation is performed in both directions from the anchor, with no loss of generality. A naive implementation of this recurrence would suffer from extravagant run time due to the large number of pairwise positions computed in the recurrence; DIPLOMAT avoids this by computing on a sparse matrix, in which only non-negligible input (P) and trace (V) probabilities are retained and computed.

2.1.4 Skeletal constraints

. Body parts within an animal tend to maintain a particular distance, which can be used to update probability estimates. DIPLOMAT accounts for these constraints by developing an annulus distribution (see Figure 1, part 3.C) for the pairwise distances between all skeletal pairs, then incorporating skeletal distances as an additional factor of $V()$.

2.1.5 Mutually-independent traces

. When individual animals are co-located in a video for an extended stretch of time, a naive implementation of the above Viterbi algorithms can lead to identity collapse, in which two or more individuals end up being places on the same positions because the $P_{i,b}(x, y)$ dominate all others. To counteract this, DIPLOMAT implements a mutually-independent trace ("MIT") modification to the basic maximum probability (Viterbi) trace. After $V()$ values have been computed for all body parts and all individuals, a competition phase in performed: for each position (x, y) and body part b, the individual k with greatest $V_{i,b,k}(x, y)$ is identified and all other $V_{i,b,\cdot}(x, y)$ (where $\cdot \neq k$) are set to zero. This is appropriate because all future paths leading from that position will be dominated by individual k, and it forces other individuals to be assigned to other high-scoring positions.

2.1.6 Segmenting and stitching

. To improve memory usage and parallelize algorithm execution, DIPLOMAT selects more than one single anchor frame, by identifying relatively evenly-spaced frames with high reliability of animal separation. The video is broken into segments, with these anchors as boundary points. The above Viterbi pass is computed for each resulting segment, then a final processing step stitches the segments together, using the anchored frames to align animal identities across segments using relative distances between body centers, according to the Hungarian algorithm.

2.1.7 Occluded state

. The use of Viterbi and skeleton-based probabilities ensures that each body part has non-zero probability in each frame. This separates the DIPLOMAT algorithm from stand-alone CNN inferences, which show near-zero probabilities when body parts are not identified, such as during occlusion. DIPLOMAT treats these cases of low CNN probability by allowing auxiliary "occluded states". This allows DIPLOMAT to continue tracking the probable position of body parts even when the part can't be directly observed.

3 Results

To validate DIPLOMAT, we applied the software to a publicly available, standardized dataset of "Mouse triplet" videos (MABe3[6]). We compared performance of only automated components of DeepLabCut (DLC) and its DIPLOMAT extension, as well as SLEAP against its own DIPLOMAT extension, and observed a significant decrease in the number of identity swaps for the sum of individual body parts (Wilcoxon sign-rank test: DLC comparison: p = 0.0079; SLEAP comparison: p = 3.9 x 10-4) as well as mean body position (DLC: p = 0.015; SLEAP: p = 0.0034). Results are summarized in Figure 2.

4

Additionally, DIPLOMAT decreased false negatives (missing body parts) in body part detection, while sometimes increasing the number of false positives relative to the human-curated gold standard (due to DIPLOMAT's handling of occlusion).

5 Discussion

DIPLOMAT substantially reduces body-swapping in automated multi-animal tracking, and provides a graphical user tool to easily identify and cure errors in automated tracking. In ongoing work we are applying DIPLOMAT to tracking dyads of rats as well as degus for studies of rodent social neuroscience and behavior. The software is open-source, tested across operating systems, and is currently available for down-

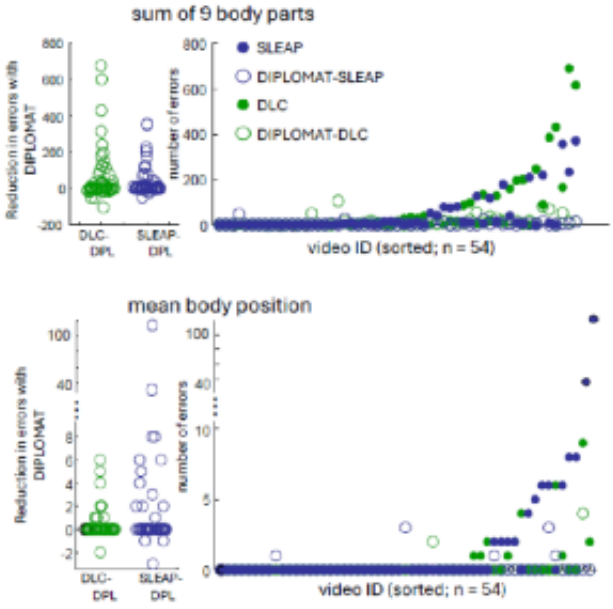

Figure 2: **Body-swaps across 54 "mouse triplet" videos.** Top panels are all swaps across 9 body parts, bottom panels are mean position of body. Left: errors in DLC (green) and SLEAP (blue) minus corresponding DIPLOMAT errors. Right: distribution across all videos, ordered by those with the most stand-alone DLC or SLEAP errors. Errors in DIPLOMAT (open circles in right panels) were statistically lower..

load at `https://diplomattrack.org/`, with a manuscript pending.

Acknowledgments

We gratefully acknowledge the high-performance computing (HPC) resources and expert administration supported by the University of Arizona TRIF, UITS, and Research, Innovation, and Impact (RII) and maintained by the UArizona Research Technologies department.

References

[1] T. D. Pereira, N. Tabris, A. Matsliah, D. M. Turner, J. Li, S. Ravindranath, E. S. Papadoyannis, E. Normand, D. S. Deutsch, Z. Y. Wang *et al.*, "Sleap: A deep learning system for multi-animal pose tracking," *Nature methods*, vol. 19, no. 4, pp. 486–495, 2022.

[2] J. Lauer, M. Zhou, S. Ye, W. Menegas, S. Schneider, T. Nath, M. M. Rahman, V. Di Santo, D. Soberanes, G. Feng *et al.*, "Multi-animal pose estimation, identification and tracking with deeplabcut," *Nature Methods*, vol. 19, no. 4, pp. 496–504, 2022.

[3] Z. Chen, R. Zhang, H.-S. Fang, Y. E. Zhang, A. Bal, H. Zhou, R. R. Rock, N. Padilla-Coreano, L. R. Keyes, H. Zhu *et al.*, "Alphatracker: a multi-animal tracking and behavioral analysis tool," *Frontiers in Behavioral Neuroscience*, vol. 17, p. 1111908, 2023.

[4] F. Romero-Ferrero, M. G. Bergomi, R. C. Hinz, F. J. Heras, and G. G. De Polavieja, "Idtracker. ai: tracking all individuals in small or large collectives of unmarked animals," *Nature methods*, vol. 16, no. 2, pp. 179–182, 2019.

[5] T. Walter and I. D. Couzin, "Trex, a fast multi-animal tracking system with markerless identification, and 2d estimation of posture and visual fields," *Elife*, vol. 10, p. e64000, 2021.

[6] J. J. Sun, M. Marks, A. W. Ulmer, D. Chakraborty, B. Geuther, E. Hayes, H. Jia, V. Kumar, S. Oleszko, Z. Partridge *et al.*, "Mabe22: A multi-species multi-task benchmark for learned representations of behavior," in *International Conference on Machine Learning*. PMLR, 2023, pp. 32 936–32 990.