Deep pose estimation of animals in videos using multiple calibrated views

Background: In the next few years, deep learning will transform the way biologists collect data from video. In biomechanics, we look for patterns in the ways humans, animals, and plants move, extract principles and mechanisms governing their motion, and use these principles in the design of robots that fly, swim, and run. We identify interesting behaviors and record hundreds of repetitions using multiple synchronized high speed cameras. By manually extracting (or "digitizing") the position of several body locations, we reconstruct the 3D locations of these parts and fit models to the data to test our hypotheses. Digitizing long videos, however, is by far the slowest bottleneck in biomechanics research. Recent advances in object detection, segmentation, tracking, dense 3D reconstruction, pose estimation, markerless motion capture, and action recognition achieved using deep neural networks could profoundly increase rate at which we can learn about how animals move and translate these principles and discoveries into functional designs.

Description: With brief periods of human-in-the-loop guidance for selecting features of interest and generating small sets of training data, this system will use a combination of unsupervised and supervised learning to estimate the pose (in the 3D coordinate frame of the calibrated camera system) of an animal and its moving parts in all frames of a video dataset consisting of 10s-100s of videos. Upon project completion, the sponsor and students may submit a writeup of the method to WACV, the IEEE Winter Conference on Applications in Computer Vision.

Data available: High speed videos of bee and hummingbird flight. Bees: 150-200 trials of 1000-5000 frames each from two to four views. Ground truth locations for four features of interest (top of head, tip of abdomen, wing base, wing tip) are available for ~ 50 of these trials. Hummingbirds: 640 trials of 100-300 frames each from two views. Ground truth locations for four features of interest (bill tip, tail tip, wing base, wing tip) are available for all videos in one view. Note that these videos are all very similar in lighting, contrast, and other features, as is typically the case for video datasets in biomechanics.

Baseline method: Performance will be compared to Argus (http://argus.web.unc.edu/) or DLTdv5 (https://www.unc.edu/~thedrick/software1.html), which are commonly used to manually select points of interest on animals. Program includes basic template-based "autotrack" mode, but requires continuous human supervision because the tracker halts and waits for the user if the track is lost.

Desired/proposed pipeline: The program is initialized on a folder of videos and run "overnight" unsupervised. During this time, the program finds proposals for animals and their parts using spatial and temporal consistency of the parts within and across videos in the dataset. Animals are assumed to change position/move within a video and should be present in most videos. The model should allow animal parts to articulate as the animal moves through the scene. Upon returning, the user is prompted to (i) semantically label 20-40 part proposals of interest (e.g. center of head, body piece x, leg x, foot x, wing, wing tip, wing base), and (ii), using the user's knowledge of the animal, connect part proposals in a graph structure representing which groups of parts are rigidly connected and which groups of parts are allowed to move/rotate relative to one another. The user may also be prompted to manually label any additional points of interest

by clicking them in several frames/views of an example video. After user input, the system re-analyzes the videos and outputs 2D image positions ("tubelets") as well as the full 3D pose for each animal part over time in each video. Note that the team may choose a different pipeline if desired (such as having the user collect a small training dataset at the beginning of the pipeline).

Implementation and hardware: Flexible, but prefer implementation in Python or Matlab using TensorFlow, Keras with TensorFlow backend, or MXNet for deep learning components (but flexible if team feels a different choice is more feasible/easier). A computer with a GPU suitable for deep learning will be provided if campus computing resources or student computers are insufficient.

Criteria for success: We are flexible on details of the user experience/interface, but the system should require less human input time than a template search method that requires continuous human supervision. Performance should be measured by the average user interaction time with the program per video.

Deliverables: The sponsor expects the following deliverables:

1. A package or collection of code files ("the system") uploaded to a github repository containing a main function to be run by the user and supporting functions, class definitions, network models, etc.
2. A brief user manual, installation guide, and script for dependency installation (e.g. through an Anaconda environment)
3. Short report comparing effectiveness of the system to existing alternative solutions (e.g. speed per video and accuracy on a small test set). The report should also describe how the team arrived at the design choices made in the system architecture.

Contact: The sponsor (Marc Badger) expects to meet with the student team one hour per week.

Additional reading:

Elhayek et al. 2016. MARCOnI—ConvNet-Based MARker-Less Motion Capture in Outdoor and Indoor Scenes. *PAMI*. IEEE link.

Pavlakos1 et al. 2017. Harvesting Multiple Views for Marker-less 3D Human Pose Annotations. *CVPR*. arXiv link.

Song et al. 2017. Thin-Slicing Network: A Deep Structured Model for Pose Estimation in Videos. *CVPR*. arXiv link.

Bautista et al. 2017. Deep Unsupervised Similarity Learning using Partially Ordered Sets. *CVPR*. arXiv link.

Breslav et al. 2016. Discovering Useful Parts for Pose Estimation in Sparsely Annotated Datasets. *WACV*. arXiv link.