Baboons on the Move: Enhancing Understanding of Collective Decision Making through Automated Motion Detection from Aerial Drone Footage

Christopher L. Crutchfield, Jake Sutton, Anh Ngo, Emmanuel Zadorian, Gabrielle Hourany, Dylan Nelson, Alvin Wang, Fiona McHenry-Crutchfield, Deborah Forster, Shirley C. Strum, Ryan Kastner, Curt Schurgers

Engineers for Exploration, University of California, San Diego

Introduction:

Collective and distributed decision-making has long been a topic of interest in animal research since it is a complex process in many nonhuman animal species. Long-lived social mammals that interact within societies have much in common with humans. Within these particular societies, individuals and their connections within their social network have a critical impact on group-level behavior. This is particularly true of nonhuman primates. In this paper, we examine tracking of baboon troop movements using a combination of human observers and computer vision techniques to aid in the study of the group-level behaviors that impact troop movement and collective decision-making.

To understand group-decision and the context thereof, one would ideally be able to continuously (1) monitor identified individuals and their activities, (2) track the relational dynamics and social networks, (3) monitor group-level behavior and (4) monitor the environment.

In order to find and analyze the moment when a decision is made—for example, what direction a troop will head—it is necessary to track the entire group of baboons individually. The decision is a complex negotiation that can originate from both local dynamics and relational history between individuals. These interactions between individuals may have more impact than other independent variables. Since it is difficult to know all of the variables involved in a decision, it is necessary to understand the moment in which a decision is made.

Currently, researchers in the field take notes about observations. Unfortunately, these notes do not provide the full context as a single researcher can only see a portion of the entire sleeping site. GPS radio collars have been used to augment the notes taken by field researchers, but they only allow for observing partial group membership and thus require attempting to fill in gaps left by having only partial results. The use of drones would allow for monitoring of individuals and small groups, but also the troop as a whole in a way that is not achievable with GPS collars or field observations alone.

Aerial drone footage can help to fill in some of these gaps, as it can provide a complete view of the entire site. As there may be hours worth of drone footage to review, it is impractical to do it by hand. Instead, computer vision techniques can be used to reduce the amount of video that must be reviewed. By annotating the beginning of a video, an individual identification can be maintained through the majority of the video through automated means. This ensures that the context of a decision is not lost.

Automated computer vision techniques, however, are challenged by the movement and distortion of drone-mounted cameras. As these cameras have six degrees of freedom (DoF)—the freedom of movement a non-fixed camera has in three-dimensional space—it becomes necessary to compensate for the potential movement of said camera. Additionally, individual identification currently requires a significant effort in the field as experienced field researchers must annotate individual baboon IDs in as close to real-time as possible. This motivates the minimization of the time it takes to process the drone footage.

In this paper, we report on methodological and computational developments that show promise towards solving these problems, with the primary goal of being able to use the processed footage to identify the moment of decision. Drone footage is less invasive than other methods (e.g. radio collars) and allows researchers to more easily view the entire group. This footage also provides additional context that collars cannot, as a collective decision may originate from an individual’s agenda.

Background:

Birds and honeybees have previously been used to study collective decision-making, but nonhuman primates are of particular interest to study because of their cognitive and behavioral similarities to humans (Strum, 2012). With the development and popularization of the global positioning system (GPS), researchers were better able to investigate leadership, decision-making, and model troop movement in groups of animals (Strandburg-Peshkin, Farine, Couzin, & Crofoot, 2015), leading to findings of group movement governed by a majority rule. When GPS data was overlaid on a detailed environment map, topography was also found to influence group movement (Strandburg-Peshkin, Farine, Couzin, & Crofoot, 2015).

The Uaso Ngiro Baboon Project (UNBP) encompasses the most complete baboon socio-ecological field research for close to half a century, with a more focused study on troop movement since 1994. While this project has many of the above elements, it does not currently capture the moment the collective decision is made. The troop’s first major, collective decision of the day—movement from the sleeping site—has the potential to provide this missing insight. As baboons’ preferred sleeping sites are rocky outcrops (see Figure 1), the initial movement is more constrained compared to decisions later in the day, therefore allowing the use of aerial imagery to be more practical.

In order to be able to better understand all of these complex factors, and building on the comprehensive history of baboons in the field within "Darwin’s monkey: Why baboons can’t become human" by Dr. Shirley Strum (2012), we established an interdisciplinary effort that includes engineering students within the Computer Science and Engineering, the Electrical and Computer Engineering, Mathematics, and Data Science departments at University of California, San Diego, and field primatologists. While field methods in tracking have benefited over the years from technology augmented tools, there are still gaps in the capabilities of these tools. Here we discuss the progress of an on-going effort to bridge one of these gaps by augmenting field observations from the ground with aerial footage from drones processed using computer vision.

Related Works:

This paper has similar goals to the paper by Haalck et al; we want to be able to detect moving subjects even when given an unstable camera and low pixel density of those individual subjects. While their goal was to study, follow, and map the paths of individuals, our objective, rather than follow an object already in motion, was to study the group dynamic of our subjects and how they collectively decide before their movement is smooth.

This makes the method of reducing noise employed by Haalck et al difficult for us to use. While Haalck et al are able to eliminate almost all noise, they do so by employing statistical models that require the animals to have smooth motion (Hallck, Mangan, Webb, & Risse, 2020). Where we differ is in the number of animals and kinds of movements that we are able to pick up. As our baboons will typically stop or change direction on a whim, our data does not fit this smooth motion requirement. This necessitates implementing a different means of following the decision-making progress.

This period of decision-making is relatively hard to study with current methods of manual field research, not only because of the sheer quantity of baboons involved but also because of the poor viewing angles of people taking notes on the ground. Any single researcher on the ground can see only a partial view of the big picture. With drone footage, not only can it be easier to skim through hours of irrelevant footage, as a computer can assist, but it can also be digitized and pieced together so that the group can be understood as completely as possible. The moment we are trying to define has multiple stages, including instances where smaller portions of the group or even individuals advocate for different directions, until they collectively decide on one final

direction. Field researchers currently are uncertain of how they collectively make that decision, so by identifying when this is happening, making it easier to visualize, and giving insight into things we can't visualize (such as baboon direction and future trajectory prediction) we hope to make understanding their process much more feasible.

Methods and Results:

In our testing, we used a DJI Mavic 2 Pro drone equipped with an L1D-20C RGB, 4K camera. We found that flying at an altitude of 50 meters allows for baboons to have a frame size of about 25 by 25 pixels, which provides for sufficient detail for the animals to be resolved by humans and our algorithm. In the initial test footage we acquired, it appeared that the baboons habituated to the drone within a day or two.

We encountered some drawbacks from this method of data collection. First of all, even though the drone is very stable, it still moves slightly, having rotated 2.64° and 0.5 meters during a minute-long video. This is can be accounted for with our computer algorithm.

Another drawback that we have yet to fully solve is the relatively short flight time of the most commercial drones. As this flight time is often limited to under 30 minutes, it will be necessary to fly a second drone to fill in the gaps.

The algorithm that produces this result happens in multiple distinct steps, as can be seen in Figure 1.

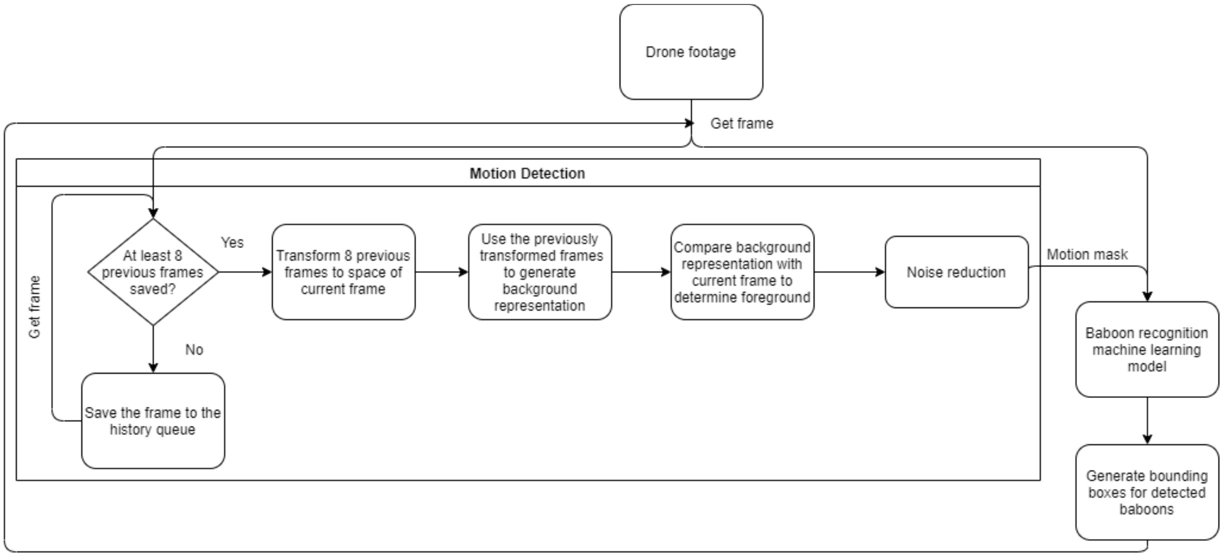
Drone footage
Get frame
Motion Detection
Get frame
At least 8 previous frames saved?
Yes
No
Transform 8 previous frames to space of current frame
Use the previously transformed frames to generate background representation
Compare background representation with current frame to determine foreground
Noise reduction
Motion mask
Save the frame to the history queue
Baboon recognition machine learning model
Generate bounding boxes for detected baboons

Figure 1. Proposed pipeline of the baboon detection algorithm

At this point in time, much of Figure 1 is currently implemented. The algorithm can successfully segment motion given a couple of constraints. (1) The background of the video must be sufficiently featureful. This means that the background must have regions of sufficiently different contrast levels to be able to select unique, common features between frames. (2) The drone camera must be relatively stable so that when the frames are wrapped, image artifacts are not a significant issue.

Motion Detection:

In order to detect motion, it is first necessary to generate a representation of the background. We do so by implementing the following steps.

Frame transformation - In order to align the previous eight frames to the space of the current frame, it is necessary to generate a transformation matrix that can be used to warp the previous frame's space to that of the current frame. If we define the previous frame, f_{t-i} and the current frame, f_t, an ORB feature detector (Rublee, Rabaud, Konolige, & Bradski, 2011) can be used to find like features between these two frames. The n most similar features are then chosen to estimate a transformation matrix from f_{t-i} to f_t, through the use of the RANSAC algorithm. This transformation matrix is then applied to wrap the previous frames into the space of the current frame.

Intersection - Once the previous eight frames are warped to match the current frame, the previous eight frames are then intersected as described in "An Efficient Approach for Object Detection and Tracking of Objects in a Video with Variable Background," equation (8) (Ray & Chakraborty, 2017). Instead of using the formula listed for quantizing (see equation 7) the frames, we use equation (1) listed below instead.

$$H_i^Q = \{h_l^Q : h_l^Q = q_j * 40, \text{for } q_{j-1} < h_l^N \leq q_j\} \tag{1}$$

Quantization here compresses the image so that all pixel values are between 1 and 40. This effectively thresholds what pixels are considered to be the same. Originally, Ray & Chakraborty (2017) had compressed the values to be between 1 and 10. Our change here allows for increased sensitivity to contrast changes. H_i is defined as the i^{th} historical transformed frame produced by applying the transformation matrix to f_{t-i}. h_l^Q represents the quantized value of the pixel at the l^{th} location. Finally, h_l^N is defined as $h_l^N \in \frac{H_i}{2^8-1}$.

The goal of the intersections is to remove parts of the image that are changing between the frames. The intersection operation leaves pixels that have changed between the two frames being compared with a value of 0 intensity.

Union - Once we have intersected each pair of eight frames, the seven intersected frames are then combined by a union operation as listed in "An Efficient Approach for Object Detection and Tracking of Objects in a Video with Variable Background, equation (9) (Ray & Chakraborty, 2017).

The goal of the union operation is to fill the gaps created by the intersection operation with the true background. By filling in the pixel values with 0 intensity with values from other frames, we estimate the actual background.

Foreground Selection - Foreground selection is also implemented as defined in the same paper by Ray & Chakraborty. For this operation, section III, part C until equation (14) (Ray & Chakraborty, 2017). We deviate after this equation in an attempt to reduce noise.

The goal of foreground selection is to choose candidate pixels for the moving foreground. Current pixels selected include potential noise and thus we must compensate for this.

Noise reduction - In order to reduce noise, we first perform a morphological opening operation as defined below in equation (2)

$$A_{opened} = A_{foreground} \circ B_{opening} = \left(A_{foreground} \ominus B_{opening}\right) \oplus B_{opening} \tag{2}$$

where $\ominus$ and $\oplus$ represent erosion and dilation respectively. $A_{foreground}$ refers to the mask generated by the foreground selection step and $B_{opening}$ is an ellipse kernel of size 6x6 pixels. This is a very lossy operation, as such, it is necessary to dilate this mask to make the elements of the remaining motion mask significantly larger. This is done by the following operation.

$$A_{dilated} = A_{opened} \oplus B_{dilation} \tag{3}$$

where $B_{dilation}$ refers to another ellipse kernel with a size of 30x30 pixels. It is next necessary to combine the $A_{foreground}$ mask with $A_{dilated}$ mask to a less noisy mask. We do this by performing a boolean and on the two masks.

$$A_{reduced} = A_{dilated} \wedge A_{foreground} \tag{4}$$

Since the motion represented in $A_{dilated}$ has a much larger radius, it is expected that the regions covered by $A_{dilated}$ will encompass those of true motion in $A_{foreground}$.

Connecting the blobs - The remaining mask may not have connected blobs. We address this by performing the following operation (Ray & Chakraborty, 2017).

$$A_{motion} = (A_{reduced} \oplus B_{kernel}) \ominus B_{kernel}$$

where B_{kernel} is an ellipse kernel of size 12x12 pixels. A_{motion} represents the motion mask output by the motion detection part of our pipeline.

The left image of Figure 2 displays a frame from one of the videos that were used to produce test results, while the right image is the same rock formation from the ground to provide a sense of scale. This is one of the main sleeping sites of the baboons observed by UNBP.

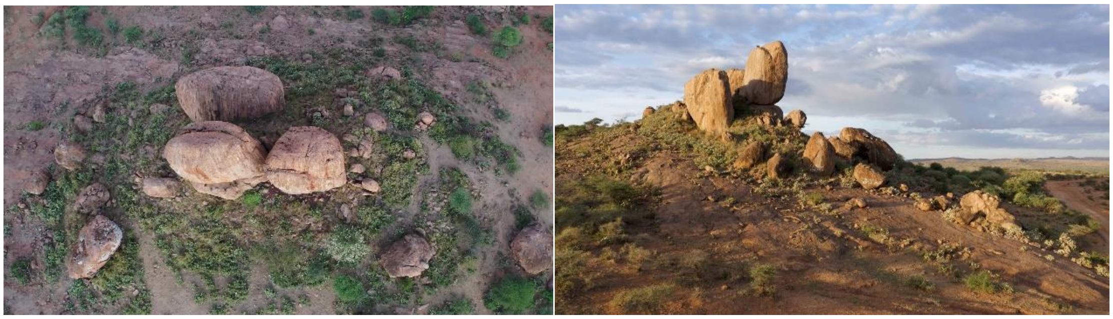

Figure 2. White Rock Sleeping Site. Right-Top View. Left – ground view

Figure 3 provides an example of flagged movement pixels as generated by the above algorithm. This image provides a small cop of a video taken from the same height as Figure 2. As can be seen, these moving baboons are correctly flagged as areas of interest.

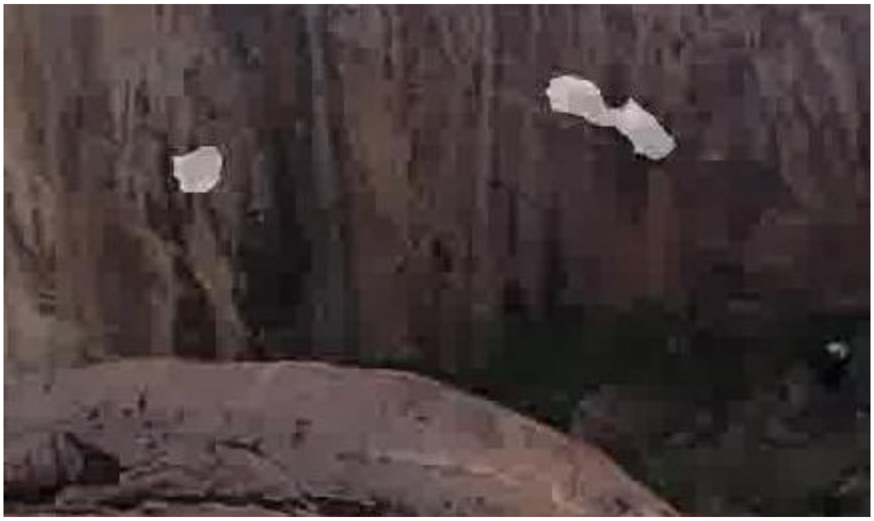

Figure 3. Baboon movement detected and flagged

Baboon Recognition Machine Learning Model

We currently do not have this part of the pipeline implemented. This stage of the pipeline will allow us to detect non-moving baboons and filter out additional noise.

Generate Bounding Boxes for Detected Baboons

After we implement the machine learning model, we will use its output to produce bounding boxes that define where the baboons are.

Discussion:

The five main constraints for acquiring the footage are (1) the video must capture the entire baboon sleeping site, approximately 100 meters in length, in order to observe troop-level decision making, (2) the resolution of the video must be large enough so that each baboon is sufficiently large such that it can be resolved so that its individual movement can be registered and tracked, (3) the footage needs to be relatively stable, since it uses motion detection as a primary means of tracking, (4) the footage must be obtained in a way that does not disrupt the baboons' normal behavior, and (5) the footage needs to be sufficiently long to ensure that all necessary information is collected to understand the group-level decision.

Since the DJI MAVIC 2 PRO has about a 25 to 30 minute flight time, it is necessary for us to compensate for this flight time as it may not be sufficient to collect the necessary data. As a result, we will investigate the idea of flying a second drone near the end of the flight time of the first. We will design a hand-off concept to allow the information from the first video to flow into the second video.

In order to be able to fully understand the collective-decision made by the baboon troop, it is necessary to be able to identify individuals within the group. Our current plan for doing so is to pre-label the first frame of the video with individual identification. It will then be possible to propagate these individual identifiers through the video using the pipeline discussed in methods. If the tracking of the individuals begins to drift, it will be necessary to have the researchers relabel at the point where drift occurred, allowing a course-correct to happen mid-processing.

Conclusion:

As we evolve the method proposed here, we hope to be able to better understand how baboons make collective decisions. Understanding how nonhuman primates make group-level decisions has the opportunity to help inform our knowledge of how other primates, including humans, make decisions. This knowledge allows us to improve the group-level decisions that we make, as well as the ones we instruct our technology to make.

As we continue to create more adaptive technology that integrates deeper into everyday life, mixed groups of humans and other intelligent agents—such as a fleet of autonomous vehicles mixed with human-operated vehicles in traffic—are expected to perform collaborative tasks. The understanding of emergent, group-level behavior is invaluable. Understanding how different entities within a group vary in skill, knowledge, and social influence decisions will be important to our future as a species.

Ethical Statement:

The baboons are wild and except for an intervention to translocate them in 1984, we do not interact, touch or interfere with them. The Uaso Ngiro Baboon Project (previously the Gilgil Baboon Project) has been studying these baboons since 1972, long before formal ethical statements were required. We continue to act in accordance with ethical best practices.

Acknowledgments:

The authors would like to thank: Neil Thomas for providing test data, Eric Lo for providing additional guidance on the algorithm, Allison Busa for initial experiments on machine learning processes, Charmaine Beluso for reviewing annotation tools used in testing, Charles Kinzel for his work improving performance, Davin Tjong for his work testing other avenues not discussed in this paper. The authors would also like to thank Corina Most for her useful feedback on the project and the other field staff for their support.

This work was funded in part by the REU Site Engineers for Exploration, supported by the NSF under Grant No. 1852403. Additional support was provided by the UCSD ECE and CSE departments and the UCSD Qualcomm Institute.

References

1. Hallck, L., Mangan, M., Webb, B., & Risse, B. (2020). Towards image-based animal tracking in natural environments using a freely moving camera. *Journal of Neuroscience Methods, 330*.

2. Ray, K. S., & Chakraborty, S. (2017, 5 11). An Efficient Approach for Object Detection and Tracking of Objects in a Video with Variable Background.

3. Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. *International Conference on Computer Vision*, (pp. 2564-2571). Barcelona, Spain.

4. Strandburg-Peshkin, A., Farine, D., Couzin, I., & Crofoot, M. (2015). Shared decision-making drives collective movement in wild baboons. *Science, 348*, 1359-1361.

5. Strum, S. (2012). Darwin’s monkey: Why baboons can’t become human. *Yearbook of Physical Anthropology, 55*, 3-23.