Space, Time, and Counts: Improved Human vs Animal Detection in Thermal Infrared Drone Videos for Prevention of Wildlife Poaching

ANIKA PURI[1], ELIZABETH BONDI[2]
[1]HORACE GREELEY HIGH SCHOOL, CHAPPAQUA, NY, USA
[2]HARVARD UNIVERSITY, CAMBRIDGE, MA, USA

Wildlife poaching of endangered species such as elephants and rhinoceroses in Africa and Asia for illegal trading has become a biodiversity crisis, which has also been highlighted by the United Nations Sustainable Development Goal SDG15 of halting biodiversity loss. Recently, unoccupied aerial vehicles (UAVs) equipped with heat-sensing infrared cameras (and coupled with computer vision software) have been deployed to help park rangers monitor protected areas at night when illegal wildlife poaching typically occurs and protected areas are closed. In order to maximize the area covered within a fixed flight time and battery constraints, the UAVs usually fly at an altitude of approximately 400 ft. This results in small animal/human sizes in the captured thermal images, and consequently leads to poor detection accuracy of as low as 20% for humans. In this research, we study the spatio-temporal nature of the video data, i.e., the difference in movement pattern of animals and humans over time, such as their turning radius, speed, etc., to determine whether these features have promise in improving classification. When tested using thermal infrared video dataset called BIRDSAI [6], collected from four national parks in Africa, our method was able to use movement patterns to detect humans with 81.8% accuracy. We similarly complement this space-time model with a new animal/human count model that leverages the herd nature of animal behavior in further improving human/animal detection accuracy to 90.9%. In both cases, we aim to illustrate that these additional features could improve deep learning-based algorithms for the task of identifying human activity in thermal infrared UAV videos in order to prevent wildlife poaching.

1 INTRODUCTION

The World Wildlife Fund estimates that as many as 35,000 elephants are killed every year due to poaching, and unless this crisis is addressed urgently, elephants will become extinct by 2040 [21]. In order to protect wildlife from poaching [16], park rangers conduct patrols throughout protected areas. However, a single national park can be as large as 100,000 sq km, meaning it is almost impossible to cover with limited resources. Recently, the deployment of unoccupied aerial vehicles (UAVs) with on-board video cameras has facilitated more frequent monitoring of national parks for signs of wildlife poaching [10, 13, 14]. Since most poaching occurs at night, thermal infrared cameras mounted on these UAVs have been deployed to detect animal and human activity in real time in national parks in Africa and Asia [9]. In order to maximize the coverage area, the UAVs usually fly at an altitude of approx. 400 ft [1]. This results in small animal/human sizes in the captured thermal videos. Because of this, it is often very difficult for even human experts to recognize humans in these videos, leading to recognition errors.

To help alleviate this problem, recent research efforts have focused on computer vision-driven automatic animal and human detection methods. A comprehensive survey of various poaching detection technologies including UAVs and detection and their pros and cons is given in [10]. As automated object detection techniques rely on labeled ground truth data, Kellenberger et al. [11] presented a method to make this process more efficient. Lygouras et al. [14] discussed

Author's address: Anika Puri[1], Elizabeth Bondi[2]
[1]Horace Greeley High School, Chappaqua, NY, USA
[2]Harvard University, Cambridge, MA, USA
.

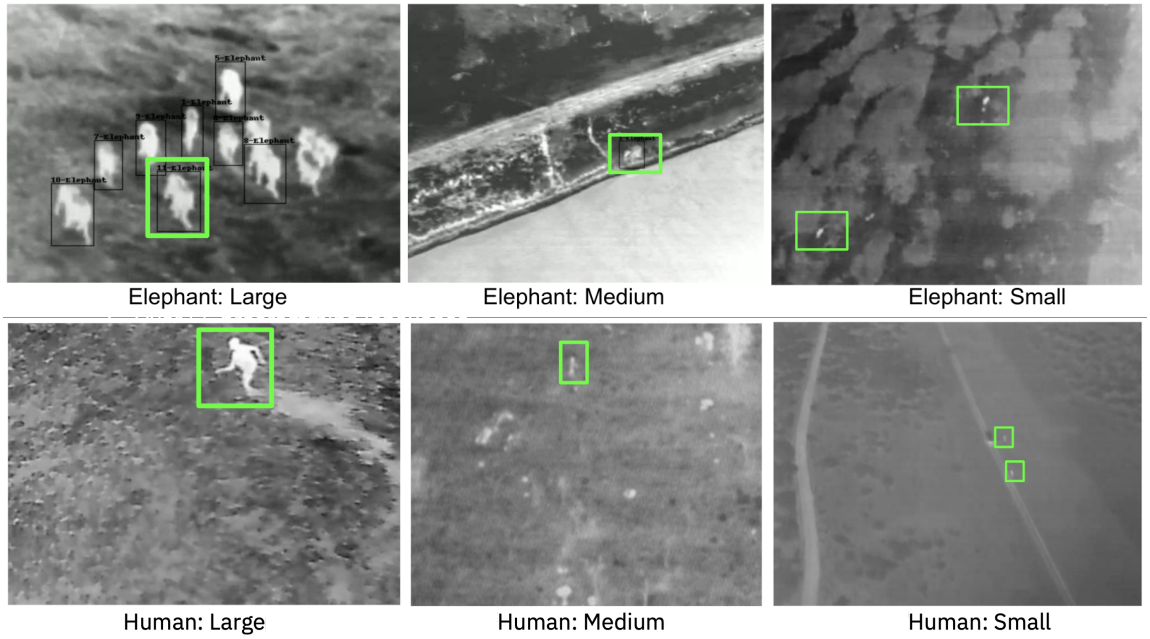

Fig. 1. Decreasing size of Elephant and Human images with increasing UAV height, captured in thermal infrared video data collected from four national parks in Africa [6]

techniques for human detection for search and rescue and animal detection with computer vision and UAVs. However, these techniques make use of daytime videos.

We instead focused on locating animals and humans in *nighttime thermal infrared videos* (e.g., Fig. 1) in the BIRDSAI dataset [6]. Notice that these image frames are grayscale, with few pixels on objects of interest, i.e., humans/animals. In addition, many other objects in the video frames look similar to the objects of interest. There are also scale variations, background clutter due to thermal reflections, large camera rotations, and motion blur. Hannaford et al. developed EyeSpy [8], an application that was used by Air Shepherd in practice [1] for detecting moving objects based on edge detection. However, several limitations prevent widespread use of this tool, such as the need for: subject matter experts for monitoring to provide parameters like edge detection thresholds, sizes, altitude, and camera look angle throughout the UAV flight. Bondi et al. [4, 5] addressed the limitations of the traditional feature engineering-driven computer vision techniques. Their method uses recent deep learning methods to localize and classify the objects of interest in the images. This method, although limited to object recognition techniques in static video frames, is currently the state-of-the-art for human detection in thermal infrared videos captured by UAVs in national parks for wildlife conservation. However, due to the small size of the objects, such techniques result in poor classification accuracy of as low as 17-21% [6] for detecting humans when the UAV is at a higher altitude.

Fortunately, we have access to videos which capture the spatio-temporal movements of objects of interest, and ecological studies show that the movement patterns of animals differ from those of humans with respect to speed, turning patterns, etc. [7, 17, 20]. Furthermore, it is well known that several animal species such as elephants exhibit herd behavior, i.e., they are typically found in groups [18]. Together, we believe the two components could improve the

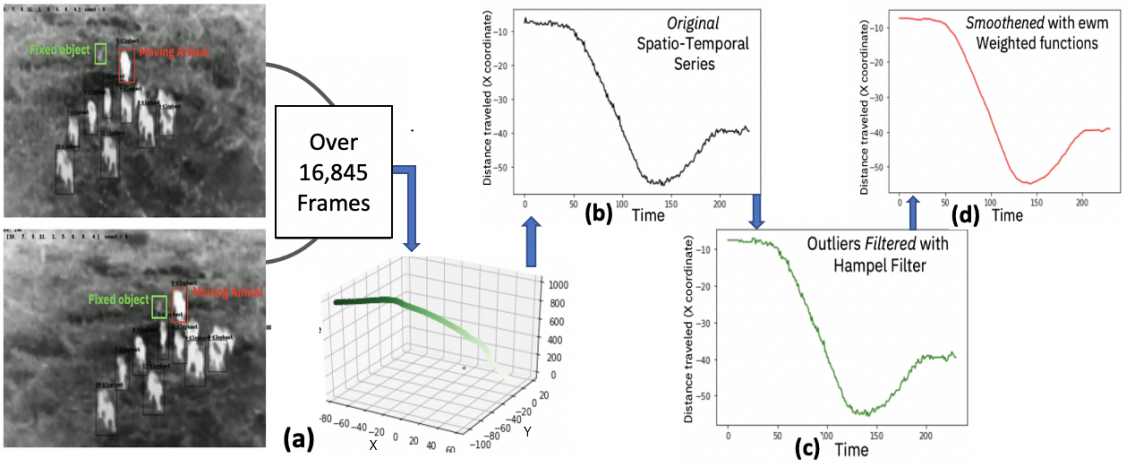

Fig. 2. (a) Extracting spatio-temporal movement series from TIR videos (b) Original spatio-temporal movement (c) Outliers filtered w/ Hampel Filter (d) Smoothening w/ *ewm* weighted functions.

shortcomings of existing methods to detect humans and animals, and thereby help address the critical United Nations Sustainable Development Goal SDG15 of halting the biodiversity loss [19].

2 METHODOLOGY

We will first derive unique time series representing object movement for each object of interest (i.e., humans/animals), and then train the classification model. Note that this process depends on given object of interest labels, which could be seeded from ground truth (as is the case in this proof-of-concept), or from an object detection algorithm.

2.1 Extracting spatio-temporal series training data

We started with the animal and human ground truth labels in the BIRDSAI dataset [6]. To extract spatio-temporal movement given a moving UAV, we added fixed objects for reference. The fixed object selected manually was either a tree or a water's edge, which are prominent in TIR videos. We used Discriminative Correlation Filter with Channel and Spatial Reliability (DCF-CSR) also known as CSRT tracker, to track the fixed object due to its higher precision and performance in tracking objects of interest in thermal infrared videos [2]. We then manually confirmed that these fixed objects were tracked correctly. For every instance of a human/animal, we tracked several fixed objects to maintain the relative movement data of the object of interest in case one of the fixed objects went out of video frame. The objects of interest were initialized in the first frame using the ground truth labels in the BIRDSAI dataset and were then dynamically tracked with a CSRT tracker along with the fixed object tracking. We use this data to derive the spatio-temporal time series of animal/human movement by computing the relative distance between the tracking coordinates of the fixed object and the object of interest. This process is repeated for both humans and animals, and yields high quality raw data of spatial movement over time, captured as a time series.

2.2 Outlier removal and time-series smoothing

Some jitter due to the imperfection of the fixed object tracking algorithm can naturally creep into the movement patterns. Since it is physically impossible for either humans or animals to move tens of meters within a fraction of a second (wrt the fixed object), we can remove these outliers in the spatio-temporal time series with a Hampel filter [12], which detects points with significant deviation from the sliding window median and replaces the detected outliers with the median. A threshold of 3 standard deviations and a windows size of 5 was used to filter these abrupt unnatural movements. Fig. 2(a)-(b) shows a representative spatio-temporal series, and the filtered series with outliers removed is shown in Fig. 2(c). Time series can be further smoothed to remove the unnatural jittery patterns seen in Fig. 2(c). We used the exponential weighted functions from pandas *ewm* library in python [15], to smooth out the movement patterns while ensuring we maintain their key movement features. The smoothed representative spatio-temporal movement pattern is shown in Fig. 2(d).

2.3 Training human/animal spatio-temporal model

These curated spatio-temporal series derived for numerous instances of both humans and animals serve as the training data for our classification model. We extract a number of features such as speed and duration, number of turns, and radius for each of this time series. We used the k-nearest neighbor (KNN) algorithm for building a classification model. We match a newly observed time series to the nearest known time series instance and check the known instance's class - if it is a human, then the newly observed instance is also likely to be a human. KNN is simple, easy to interpret, and works well on small amounts of training data.

2.4 Herd Count Model

We next developed a *count model* from the BIRDSAI data based on the observation that animals tend to be grouped in herds. As shown in Fig. 3, this model clearly demonstrates that as the number of objects of interest in a frame increases, the probability of the object of interest being a human decreases, while the probability of being an animal increases. We further enhanced the accuracy of our spatio-temporal KNN model by adding an additional feature, "object count," that captures this herd behavior in the corresponding video segment.

3 RESULTS

BIRDSAI captures the ground truth for over 160,000 thermal infrared video frames. Curation of time series data from the BIRDSAI dataset using our Methodology (Section 2.1) resulted in approx. 18,700 thermal infrared video frames with human/animal movement data. For animal training data, we focused on elephants because they are under a heavy threat from poachers, and the BIRDSAI real-world TIR video data comprised mostly of animals, ensuring a good amount of training for machine learning. There were more frames with elephants (16,845 frames) than with humans (1,853 frames), as many videos simply did not observe humans. Since spatio-temporal sequences had to be long enough to ensure meaningful data could be collected on movement patterns, these frames were curated into 41 spatio-temporal movement series with an average of 450 frames per series. Out of these 41 movement series, 32 series were for elephants - the animal in the majority of videos, and 9 series were for humans. We used 30 times series (75%) for training our KNN model as discussed in Section 2.3. The remaining 11 time series (25%) were reserved for model testing and were excluded from training. When the trained model was used to classify the movement as human or animal in this test set, *our spatio-temporal KNN model achieved an accuracy of 81.8%.*

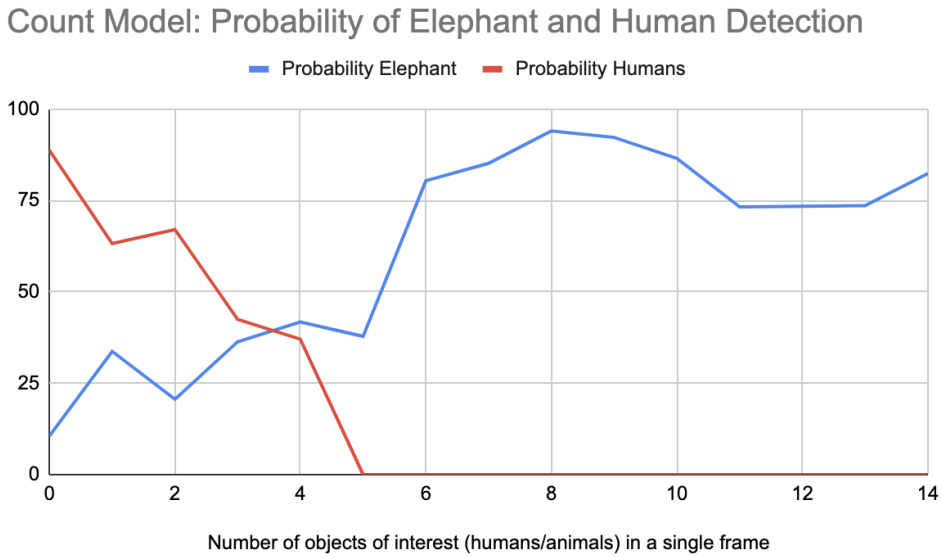

Fig. 3. Herd Count Model Results.

Adding the herd model features to our spatio-temporal model increases the accuracy of human/animal prediction to 90.9%. Furthermore, for human prediction, we achieved a precision score of 1.0, as detecting any suspicious human activity with high precision is critical in a wildlife conservation scenario. For the human activity prediction, recall was 0.5, yielding an F1-score of 0.67. For prediction of the animal movements, our model achieved a precision score of 0.9 and a recall of 1.0 - corresponding to an F1-score of 0.95.

4 FUTURE WORK

While combining these features may be helpful for processing a full video after a flight, park rangers likely wish to use any automatic detection method in real time. To detect a human in a real-time thermal infrared video feed during a UAV flight using this methodology, we propose the workflow given in Fig. 4. We need to identify and track fixed objects along with objects of interest. For this purpose, we captured over 1000 images of trees and bushes as training data, which are almost universally present in every video frame, and trained a fixed object detection model with Google AutoML Vision [3]. This capability uses transfer learning and neural architecture search to tune the final layers of neural network and customizes the trained model for the given labeled data. It took approximately 2 hrs of training time with AutoML's default compute resources [3] (2 Google Tensor Processing Units). The AutoML object recognition computer vision model thus derived can detect several fixed objects with high precision (83.3%), based on the prior manually-verified tracking labels. This AutoML model could allow for relative movement pattern extraction in real time. We are currently incorporating this model with our spatio-temporal KNN model for real time inferencing.

5 CONCLUSION

Wildlife poaching has become a biodiversity crisis which has also been highlighted by the United Nations Sustainable Development Goal SDG15 of halting biodiversity loss. Progress along this goal, especially with respect to poaching, can

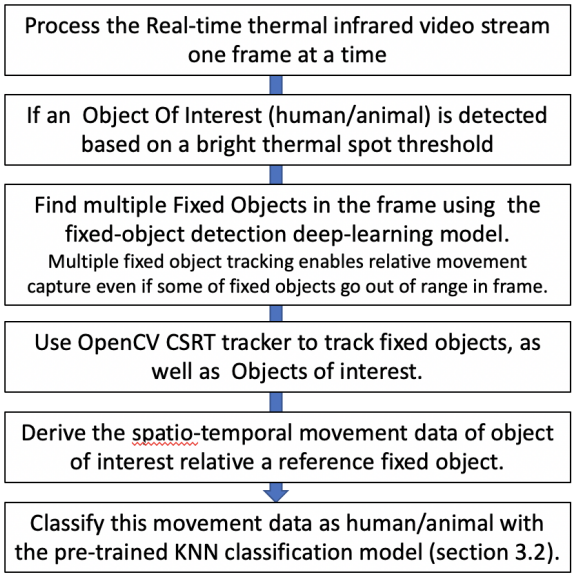

Fig. 4. Our real-time inference methodology.

only be achieved through a combination of multiple methods. We propose a potential technical method, though we encourage potential users to consider whether it is necessary to use real-time surveillance in their particular setting and to consult their communities. We also encourage those interested to contact us for data and models, which are not fully publicly available to maintain animal safety. When this type of monitoring is deemed necessary, our proposed "Space, Time, and Counts" method can improve human vs animal detection accuracy in thermal infrared drone videos for preventing wildlife poaching. We look forward to expanding our promising 90.9% accuracy result in the future.

ACKNOWLEDGMENTS

This work was supported by Harvard CRCS.

REFERENCES

[1] airshepherd 2021. Airshepherd: The lindbergh foundation. http://airshepherd.org.

[2] Abdulla AlSaadi AlMansoori, Issacniwas Swamidoss, Slim Sayadi, and Abdulrahman Almarzooqi. 2020. Analysis of different tracking algorithms applied on thermal infrared imagery for maritime surveillance systems. In *Artificial Intelligence and Machine Learning in Defense Applications II*, Vol. 11543. International Society for Optics and Photonics, 1154308.

[3] Ekaba Bisong. 2019. Google AutoML: cloud vision. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Springer, 581–598.

[4] Elizabeth Bondi, Fei Fang, Mark Hamilton, Debarun Kar, Donnabell Dmello, Jongmoo Choi, Robert Hannaford, Arvind Iyer, Lucas Joppa, Milind Tambe, et al. 2018. Spot poachers in action: Augmenting conservation drones with automatic detection in near real time. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[5] Elizabeth Bondi, Fei Fang, Mark Hamilton, Debarun Kar, Donnabell Dmello, Venil Noronha, Jongmoo Choi, Robert Hannaford, Arvind Iyer, Lucas Joppa, et al. 2019. Automatic Detection of Poachers and Wildlife with UAVs. *Artificial Intelligence and Conservation* 77 (2019).

[6] Elizabeth Bondi, Raghav Jain, Palash Aggrawal, Saket Anand, Robert Hannaford, Ashish Kapoor, Jim Piavis, Shital Shah, Lucas Joppa, Bistra Dilkina, et al. 2020. Birdsai: A dataset for detection and tracking in aerial thermal infrared videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1747–1756.

[7] Henrik J de Knegt, Jasper AJ Eikelboom, Frank van Langevelde, W François Spruyt, and Herbert HT Prins. 2021. Timely poacher detection and localization using sentinel animal movement. *Scientific reports* 11, 1 (2021), 1–11.
[8] Robert Hannaford. 2017. personal communication.
[9] Jesús Jiménez López and Margarita Mulero-Pázmány. 2019. Drones for conservation in protected areas: present and future. *Drones* 3, 1 (2019), 10.
[10] Jacob Kamminga, Eyuel Ayele, Nirvana Meratnia, and Paul Havinga. 2018. Poaching detection technologies—a survey. *Sensors* 18, 5 (2018), 1474.
[11] Benjamin Kellenberger, Diego Marcos, and Devis Tuia. 2019. When a Few Clicks Make All the Difference: Improving Weakly-Supervised Wildlife Detection in UAV Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.
[12] Hancong Liu, Sirish Shah, and Wei Jiang. 2004. On-line outlier detection and data cleaning. *Computers & chemical engineering* 28, 9 (2004), 1635–1647.
[13] Jesús Jiménez López and Margarita Mulero-Pázmány. 2019. Drones for Conservation in Protected Areas: Present and Future. Drones 3, 1.
[14] Eleftherios Lygouras, Nicholas Santavas, Anastasios Taitzoglou, Konstantinos Tarchanidis, Athanasios Mitropoulos, and Antonios Gasteratos. 2019. Unsupervised human detection with an embedded vision system on a fully autonomous uav for search and rescue operations. *Sensors* 19, 16 (2019), 3542.
[15] Wes McKinney et al. 2011. pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing* 14, 9 (2011), 1–9.
[16] Stephen F Pires and William D Moreto. 2016. The illegal wildlife trade.
[17] Lei Ren and John R Hutchinson. 2008. The three-dimensional locomotor dynamics of African (Loxodonta africana) and Asian (Elephas maximus) elephants reveal a smooth gait transition at moderate speed. *Journal of the Royal Society Interface* 5, 19 (2008), 195–211.
[18] David JT Sumpter. 2010. *Collective animal behavior*. Princeton University Press.
[19] unsdg 2021. United Nations Department of Economic and Social Affairs Sustainable Development Goals. https://sdgs.un.org/goals/goal15.
[20] Rory P Wilson, Iwan W Griffiths, Michael GL Mills, Chris Carbone, John W Wilson, and David M Scantlebury. 2015. Mass enhances speed but diminishes turn capacity in terrestrial pursuit predators. *Elife* 4 (2015), e06487.
[21] WWF 2019. WWF Says African Elephants Will Be Extinct by 2040 If We Don't Act Right Away, Newsweek.