A Novel Wildlife Poaching Detection Solution using Spatio-temporal Data with Dynamic Time Warping

Anika Puri
Horace Greeley High School, Chappaqua, NY

***Abstract*—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 recognized by a United Nations Sustainable Development Goal of halting biodiversity loss. The demand for ivory has decimated the elephant population. The main issue associated with attempts to tackle this crisis is the vast area of the wildlife national parks. Recently, unmanned aerial vehicles (UAVs) equipped with heat-sensing infrared cameras have been deployed to help park rangers survey large areas of national parks at night when >80% of poaching occurs. In order to maximize surveillance area (with fixed flight time and battery constraints) while avoiding detection, UAVs need to fly at high altitudes (>400ft). This results in very few pixels associated with the objects of interest, i.e., humans and animals, in these thermal infrared videos. Current state-of-art methods utilize shape-based object detection techniques to identify animals/humans in these thermal images, with detection accuracy of only 20%. In order to battle this inaccuracy, this research presents a novel solution that exploits video data's spatio-temporal nature (differences in movement pattern of animals/humans over time, i.e. turning radius, speed) to derive unique time series. These spatio-temporal series are then classified as human or animal by leveraging dynamic time warping metrics with K-Nearest Neighbor Clustering. When tested on a real-life thermal infrared videos dataset (BIRDSAI), collected in partnership with four African national parks, this method was able to detect human activity with 94% sensitivity enabling real-time inferencing. Furthermore, this solution has the potential to eliminate the need for high-resolution high-cost $4,800 nighttime-thermal cameras with commodity thermal cameras (<$250), as demonstrated by a design prototype.**

I. Introduction

Wildlife conservation is one of the most important sustainability goals for our environment. Every year over 30,000 species are driven to extinction [1] due to human activity. The population of elephants — the largest animal currently walking the earth has declined by 70 percent in the last 40 years, in large part because of the illegal ivory trade, which is the biggest driver of elephant poaching [2]. In fact, 20,000 elephants are killed every year to feed this trade—which is equivalent to one death every 26 minutes. The World Wildlife Fund for Nature WWF recently warned that unless this biodiversity crisis is addressed urgently, elephants will become extinct within two decades [2]. Numerous other animal species face similar biodiversity crisis levels. In the 1970s, there were about 70,000 black rhinos in Africa, and today fewer than 5,000 are left in the wild. These animals play a crucial role in preserving our planet's ecosystem, helping to maintain healthy habitats for many other species.

In order to protect wildlife from poaching [3], park rangers patrol wide swaths of national parks, however, a single national park can be as large as 100,000 sq km. Part of the issue in policing these large national parks is that the governments of nations where Africans elephants live often lack sufficient resources to protect and monitor elephant herds, which usually reside in remote and inaccessible habitats. To help the resource strapped national parks, recently, conservation programs such as Air Shepherd [4] have deployed unmanned aerial vehicles (UAVs) with video surveillance in order to protect wildlife from poaching in Africa and Asia. The UAV operators on the ground pre-program the drone flight path based on typical poaching hotspots or tips and animal density which is highly correlated with poaching activity. They monitor the live video stream, transmitted via radio waves, for any signs of poachers. Should anyone be spotted, the team manually takes control to follow the suspects, notify nearby park rangers, who are on patrol or in a van with the team, and guide them to the poachers. Monitoring these videos all night is difficult and painstaking task due to the quality of infrared video. With very few pixels associated with the objects of interest in the UAV videos (humans/aninals), and many objects that look similar to those of interest, hours of this painstaking live video monitoring task at night leads to human errors, missing poaching activity. In addition, any human activity during the night in these UAV thermal infrared videos is presumed by park rangers to be suspicious activity for poaching warranting further investigation.

Recent progress in deep-learning and machine-learning methods have revolutionized the field of computer vision and automated object detection [6] with applications in almost every field. Researchers have leveraged these advances in computer vision technology for wildlife conservation with automated visual surveillance for human presence with deep-learning driven object detection methods [7]. Since most poaching occurs at night, high resolution thermal infrared cameras mounted on these UAVs have been deployed to detect animal and human activity in national parks in Africa and Asia [4]. Although the resolution and cost of thermal infrared cameras have improved recently, they continue to lag their visual light cameras counterparts by an order of magnitude both in maximum resolution as well as cost, due to the complexities of the thermal sensor technologies. In order to maximize the surveillance area and avoid detection, the UAVs usually fly at an altitude of approximately 400 ft or above [4]. This results in small animal/human sizes in the captured thermal videos. Due to this, it is often very difficult for even human experts to recognize poachers in these videos, leading to recognition errors. To help alleviate this problem,

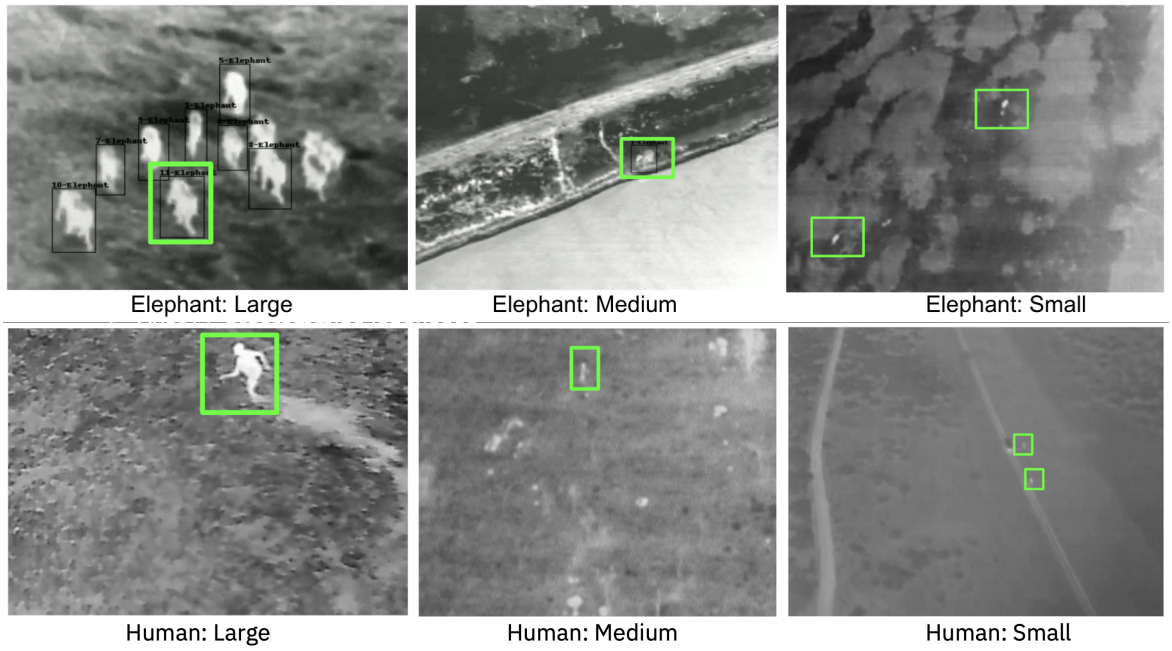

Fig. 1. Decreasing size of Elephant and Poacher images with increasing UAV height, captured in thermal infrared video data [5]

recent research efforts have focused on computer vision-driven automatic animal and human detection methods.

Hannaford et al. developed EyeSpy [8], an application that was used by Air Shepherd in practice [4] 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. To make best use of this tool, the UAV crew either needed to restrict the way the UAV flies by keeping the flight altitude and camera look angle almost the same throughout the mission, or have the expert monitoring personnel manually adjust the parameters from time to time as the settings change. Bondi et al. [9] addressed the limitations of the traditional feature engineering driven computer vision techniques by leveraging deep learning driven object detection technique using convolutional neural networks [6]. Their method treats each frame of the video as an image, and tries 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.

Although significant progress has been made for automated object detection with deep-learning technology [10], Figure 1 illustrates the difficulty associated with identifying animals/humans objects in night time thermal infrared videos, collected by Air Shepherd UAVs in African National Parks, that are part of the BIRDSAI dataset [5]. 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. Due to the small size of the objects, current state-of-the-art techniques that are limited to object recognition in static video frames, result in poor classification accuracy of as low as 20% for detecting humans [5].

To address this accuracy gap, recently, Puri et al. [11] leveraged studies that show that the movement patterns of elephants differ significantly from those of humans with respect to speed, turning patterns, etc. [12] [13]. They proposed a novel poaching detection solution that exploited this difference in animal and human movement patterns in UAV thermal infrared video data to significantly increase the accuracy of identifying human/poacher activity in wildlife national parks up to 80%. Although this approach achieved significant accuracy improvements over previous object detection base methods, unfortunately, it relies on manually extracting the features from the time series in terms of the number of turns per unit time, the curvature/radius of each turn etc. Manual feature engineering is well known to be time consuming as well as potentially limiting in terms of further accuracy gains. In this paper, we extend this approach to overcome the manual feature extraction limitation by leveraging automated Dynamic Time Warping clustering and demonstrate accuracy up to 94% on real life BIRDSAI data [5]. This novel automated time warping clustering approach also enables real-time high-accuracy detection of human activity during nigh-time for wildlife conservation.

II. Proposed Research

In this paper, a detection model driven by the unique multivariate time series for each movement pattern for the object of interest (i.e., humans/animals) is developed, that is able to automatically classify a given spatio-temporal series as that of a human or an animal. These spatio-temporal series have some unique characteristics. They are multivariate in nature

due to movement over time in both horizontal and vertical dimensions. In addition, they are also of unequal length as they are extracted from real infrared videos (ground truth data from BIRDSAI dataset [5]). Recent progress in machine learning methods for time-series have enabled classification of unequal, multivariate time series [14] [15]. Methods such as dynamic time warping (DTW) [16] are being deployed to find patterns of interest in complex time series data. New techniques combining DTW driven distance metrics with various clustering methods are enabling wider adoption of machine learning for real-world time series data [17]. The proposed novel solution trains a human/animal activity detection model with the unique human/animal spatio-temporal series by leveraging a combination of K-Nearest Neighbor clustering method along with dynamic time warping driven similarity/distance metric.

Testing of this trained model on real life night time infrared videos in BIRDSAI dataset yields human activity detection sensitivities of 94.4%. *To the best of our knowledge, this is the first method that utilizes the animal and human movement patterns for achieving over 94% poacher detection accuracy in infrared thermal wildlife video data enabling real time inferencing*. Since the proposed method leverages spatio-temporal patterns to detect human activity in nighttime infrared videos, it eliminates the reliance on the resolution of objects in image that required high-resolution thermal cameras costing over $4,800. This enables the use of much cheaper commodity thermal infrared cameras costing less than $250 [18] validated by a design prototype, reducing the overall potential cost of the drone solutions for wildlife conservation by almost 45% (from approx. $10,000 to $5,500). This proposed cost-effective solution can remove a critical bottleneck for resource strapped national parks in Africa and Asia to significantly scale the deployment of UAVs for wildlife conservation.

III. Methodology

In this section, the overall methodology for curating the human/animal spatio-temporal movement data is summarized [11] and then training of the proposed dynamic time warping driven KNN poacher detection model along with the real-time inferencing method is discussed in detail.

A. Extracting spatio-temporal series training data

The Benchmarking Infrared Dataset for Surveillance with Aerial Intelligence (BIRDSAI) is the long-wave thermal infrared (TIR) dataset [5] containing real-life nighttime videos of animals and humans in Southern Africa, taken from UAVs flying at various heights. This dataset includes TIR videos of humans and animals with several challenging scenarios like scale variations, background clutter due to thermal reflections, large camera rotations, and motion blur. The spatio-temporal data is carefully extracted from the videos in BIRDSAI by automatically tracking the movement of the object of interest with respect to a fixed object (selected manually for training data curation). Typically, tree, water's edge, or bushes were selected as fixed objects, which are prominent in TIR videos. Discriminative Correlation Filter with Channel and Spatial Reliability (DCF-CSR) also known as CSRT tracker, was used due to its higher precision and performance in tracking objects of interest in thermal infrared videos [19]. For every instance of a human/animal, several fixed objects were tracked, to maintain the relative movement data of human/elephant, in case one of the fixed objects went out of video frame. For training data, elephants were the main focus for two reasons: (1) they are under a heavy threat from poachers, and (2) the BIRDSAI real-life TIR video data comprised mostly of elephants, ensuring a good amount of training for the machine learning model. This process yields high quality raw data of spatial movement of both humans and elephants over time, captured as time series as shown in Figure 2(a).

1) Outlier Removal and time-series smoothing: It is well known that even state-of-the-art tracking algorithms are not perfect, and therefore may give rise to some outliers in the spatio-temporal data. For example, it is physically impossible for either humans or animals to move tens of meters within a fraction of a second. In addition, such outliers can interfere with downstream smoothing. To address this issue, outliers are removed with a Hampel filter [20], 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. Figure 2(b) shows a representative original spatio-temporal series and the filtered series with outliers removed is shown in Figure 2(c). Time series of movement patterns can be further smoothed to remove unnatural jittery patterns of Figure 2(c). The exponential weighted functions from pandas ewm library in python were used, to smooth out the movement patterns while preserving the overall nature of key movement features. The smoothed representative spatio-temporal movement pattern is shown in Figure 2(d).

B. Training human/animal spatio-temporal model

These curated spatio-temporal series derived for numerous instances of both humans and elephants in BIRSDAI thermal infrared videos serve as the training data for the human/animal classification model. For the implementation, k-nearest neighbor (KNN) clustering algorithm *KNeighbor*s in the *TimeSeriesClassifiers* function of the recently released *tslearn* library [15] was used. KNN algorithm works by measuring "distance" between points. The goal is to measure distance in order to find similarities in movement: speed, time, distance, turning patterns etc. The "distance" between two time-series can be measured with a euclidean distance metric, or a dynamic time warping metric. For euclidean distance metric, series have to be exactly the same for them to have a relationship. In contrast, Dynamic Time Warping metric can take speed, time, and patterns of the series of unequal lengths into account. KNN algorithm has many advantages such as its simplicity, easy of interpretability, and its ability to work well on small amount of training data with multi-category classification problems. Although data had 1000s of frames in which human/animal movement is present, it is curated into 100s of spatio-temporal

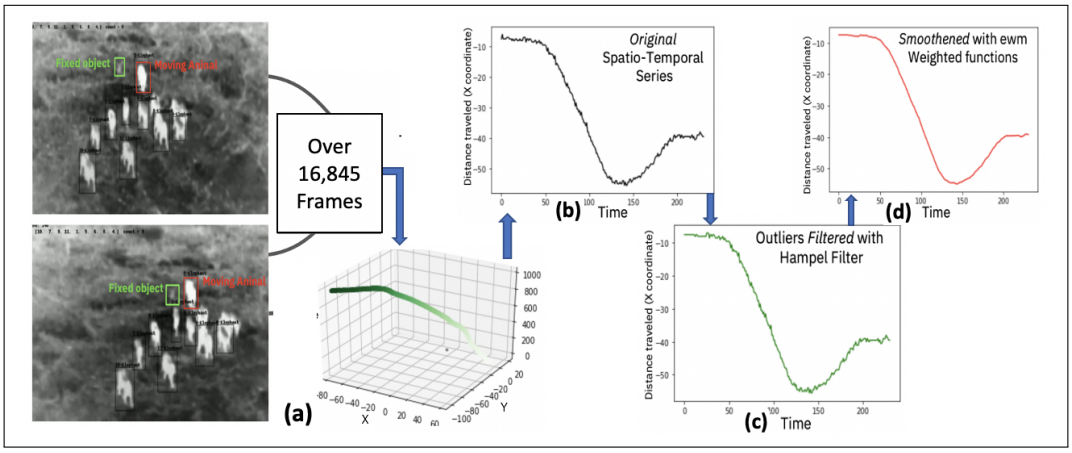

Fig. 2. (a) Extracting spatio-temporal movement series from TIR videos (b) Original spatio-temporal movement (b) Outliers filtered w/ Hampel Filter (c) Smoothening w/ *ewm* weighted functions.

series training data points. Since these time-series are multivariate and are unequal in length, use of euclidean metric to measure neighboring distance was not desirable. Therefore, dynamic time warping (DTW) algorithm was leveraged for computing the neighbor distances, and used the standard DTW formulation [16] [15]. The algorithm first finds the optimal warping path and then optimizes the cost function.

C. Real-time poacher detection/inferencing

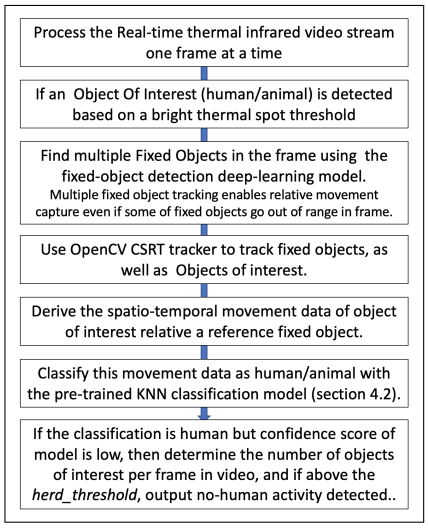

Fig. 3. Inference methodology.

The workflow to detect human activity in real-time thermal infrared video feed during UAVs flight is given in Figure 3. To classify a movement pattern as human/animal in real-time, identification and tracking of fixed objects along with objects of interest is required. For this purpose, a dataset was curated with 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 [21]. 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 [21] (2 Google Tensor Processing Units). The AutoML object recognition computer vision model thus derived can detect several fixed objects with high precision (93.3%), based on the prior manually-verified tracking labels. This AutoML model could allow for relative movement pattern extraction in real time.

IV. Results and Discussion

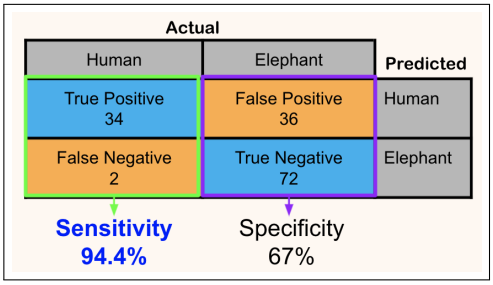

Fig. 4. DTW-KNN Model Results.

The Benchmarking Infrared Dataset for Surveillance with Aerial Intelligence (BIRDSAI) [5] contains over 162,000 frames of nighttime thermal videos labeled small, medium, large representing the surveillance footage with UAVs at various heights in Southern Africa National Parks. In these videos, any human activity during the night is presumed by park rangers to be suspicious poaching activity warranting further investigation. Since these videos captured real-life surveillance footage, there were lot more frames with elephants (16,845 frames) than with humans/poachers (1,853 frames). Curation of this Real-life thermal infrared video data using the proposed methodology discussed in Section III-A), including rotating the spatio temporal series in 4 random directions resulted in 516 spatio-temporal movement series, out of which 408 series were for elephants - the animal in majority of images, and 108 series were for humans. Among this data, 372 times series were used for training the dynamic time warping driven KNN model as discussed in Section III-B (300 time series for elephant movements, and 72 time series for human

Fig. 5. Design prototype

movement). In order to make the model more robust, The remaining 144 time series (i.e., 30% of the overall data) were excluded from training set, i.e, 108 for elephant movements and 36 for human movements were reserved for testing.

Since detecting human activity is of highest importance for prevention of wildlife poaching, sensitivity of human detection was prioritized in the model training. After training the proposed model with methodology in III-B, the dynamic time warping KNN model was tested with these 144 spatio-temporal series (testset) it has never seen before. The DTW-KNN model was able to automatically classify all the human spatio-temporal series correctly, except one, i.e., a **94.4% sensitivity for human activity detection**. Although prioritized lower, the spatio-temporal model also correctly classified 72 out of 108 elephant movement series as well, yielding an specificity of 67%. These results are shown in Figure 4.

A. Hardware Design Prototype with integrated software

Currently Air Shepherd UAVs in operations [4] deploy an expensive thermal infrared FLIR camera [22] with a resolution of 640 x 512 pixels and a field of vision of 32^o costing over $4,850. This yields very high resolution of over 327,000 pixels, which is required for shape based object detection methods currently in use. Simple geometry calculations convey that a 5ft human object of interest will map to 15 pixels in the video frame of the commercial FLIR thermal camera. In contrast, a commodity Camera "Seek Thermal Compact" [18] available for less than $250 will yield 31,000 pixels for the video frame. With this commodity camera, the 5ft human object of interest will map to 4 pixels in its video frame. Since the proposed DTW-KNN model achieves over 94% accuracy for human activity detection in nigh-time thermal infrared videos based on spatio-temporal movement patterns alone, it significantly minimizes the need for higher resolution images of objects of interests, i.e., humans and animals. This enables the use of a relatively lower resolution and low-cost thermal cameras such as SEEK Thermal Compact Camera [18].

Figure 5 shows the design prototype which is assembled from commodity hardware: (1) A commodity "Seek Thermal Compact Camera with 206x156 pixel resolution, costing $250 (2) A Raspberry Pi 4 costing $59 (3) A 9V rechargeable battery, costing $20. The experimental design prototype along with the accuracy of the human/animal detection model discussed above further demonstrate that the proposed solution can potentially save significant cost by enabling the use of a much cheaper commodity onboard UAV thermal camera ($4,850 vs $250) for real-time wildlife conservation efforts.

V. Conclusion and Future Work

Wildlife poaching of elephants has become a biodiversity crisis. In this research, a novel spatio-temporal model that significantly improves Human vs Animal Detection accuracy in thermal infrared drone videos for prevention of wildlife poaching was proposed. When tested on a real life night time infrared videos dataset, collected from four national parks in Africa, the proposed method was able to detect poachers with 94.4% sensitivity. To the best of our knowledge, this is the first method that utilizes the animal and human movement patterns for achieving over 94% human activity detection accuracy in infrared thermal wildlife video data enabling real-time inferencing. Since this solution eliminates the need for shape based object detection methods currently in use, one can now use commodity thermal cameras costing <$250, as opposed to commercial high-resolution nighttime-thermal cameras costing over $4,800.

Acknowledgments: This work was supported by Harvard Center for Research on Computation and Society.

References

[1] "The numbers are just horrendous: Almost 30,000 species face extinction because of human activity, TIME," July 2019.
[2] "World wildlife fund says african elephants will be extinct by 2040 if we don't act right away, Newsweek," November 2019.
[3] S. F. Pires *et al.*, "The illegal wildlife trade," 2016.
[4] "Airshepherd: The lindbergh foundation. http://airshepherd.org." 2021.
[5] E. Bondi *et al.*, "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*, 2020, pp. 1747–1756.
[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, pp. 436–444, 2015.
[7] J. Kamminga *et al.*, "Poaching detection technologies—a survey," *Sensors*, vol. 18, no. 5, p. 1474, 2018.
[8] R. Hannaford, personal communication.
[9] E. Bondi *et al.*, "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, no. 1, 2018.
[10] A. Borji *et al.*, "Salient object detection: A survey," *Computational visual media*, vol. 5, no. 2, pp. 117–150, 2019.
[11] A. Puri and E. Bondi, "Space, time, and counts: Improved human vs animal detection in thermal infrared drone videos for prevention of wildlife poaching." *ACM Knowledge Discovery and Data Mining (KDD) Conference, Fifth annual Fragile Earth Workshop*, August, 2021.
[12] R. P. Wilson *et al.*, "Mass enhances speed but diminishes turn capacity in terrestrial pursuit predators," *Elife*, vol. 4, p. e06487, 2015.
[13] H. J. de Knegt *et al.*, "Timely poacher detection and localization using sentinel animal movement," *Scientific reports*, vol. 11, pp. 1–11, 2021.
[14] M. Löning *et al.*, "sktime: A unified interface for machine learning with time series," *arXiv preprint arXiv:1909.07872*, 2019.
[15] R. Tavenard *et al.*, "Tslearn, a machine learning toolkit for time series data," *Journal of Machine Learning*, vol. 21, no. 118, pp. 1–6, 2020.
[16] D. J. Berndt *et al.*, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, vol. 10, no. 16, 1994, pp. 359–370.
[17] W. Pouw *et al.*, "Gesture networks: Introducing dynamic time warping and network analysis for the kinematic study of gesture ensembles," *Discourse Processes*, vol. 57, no. 4, pp. 301–319, 2020.
[18] "SEEK thermal imaging, www.thermal.com/compact-series.html."
[19] A. A. AlMansoori *et al.*, "Analysis of different tracking algorithms applied on thermal infrared imagery," in *AI and MLin Defense Applications II*, vol. 11543. SPIE, 2020, p. 1154308.
[20] H. Liu *et al.*, "On-line outlier detection and data cleaning," *Computers & chemical engg.*, vol. 28, pp. 1635–47, 2004.
[21] E. Bisong, "Google automl: cloud vision," in *Building Machine Learning and Deep Learning Models*. Springer, 2019, pp. 581–598.
[22] "FLIR 640 thermal camera, www.flir.com/products."