This WACV workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

KABR: In-Situ Dataset for Kenyan Animal Behavior Recognition from Drone Videos

Maksim Kholiavchenko[*] Rensselaer Polytechnic Institute
Jenna Kline The Ohio State Univeristy
Michelle Ramirez The Ohio State Univeristy
Sam Stevens The Ohio State Univeristy
Alec Sheets The Ohio State Univeristy
Reshma Babu The Ohio State Univeristy
Namrata Banerji The Ohio State Univeristy
Elizabeth Campolongo The Ohio State Univeristy
Matthew Thompson The Ohio State Univeristy
Nina Van Tiel ETH Zurich
Jackson Miliko Mpala Research Centre
Eduardo Bessa University of Brasilia
Isla Duporge Princeton University
Tanya Berger-Wolf The Ohio State Univeristy
Daniel Rubenstein Princeton University
Charles Stewart Rensselaer Polytechnic Institute

Abstract

We present a novel dataset for animal behavior recognition collected in-situ using video from drones flown over the Mpala Research Centre in Kenya. Videos from DJI Mavic 2S drones flown in January 2023 were acquired at 5.4K resolution in accordance with IACUC protocols, and processed to detect and track each animal in the frames. An image subregion centered on each animal was extracted and combined in sequence to form a "mini-scene". Behaviors were then manually labeled for each frame of each mini-scene by a team of annotators overseen by an expert behavioral ecologist. The resulting labeled mini-scenes form our resulting behavior dataset, consisting of more than 10 hours of annotated videos of reticulated giraffes, plains zebras, and Grevy's zebras, and encompassing seven types of animal behavior and an additional category for occlusions. Benchmark results for state-of-the-art behavioral recognition architectures show labeling accuracy of 61.9% for macro-average (per class), and 86.7% for micro-average (per instance). Our dataset complements recent larger, more diverse animal behavior sets and smaller, more specialized ones by being collected in-situ and from drones, both important considerations for the future of animal behavior research. The dataset can be accessed at https://dirtmaxim.github.io/kabr.

1. Introduction

Behavior, in the context of animal studies, is broadly defined as the way an animal acts or reacts in response to certain stimuli or situations. It encapsulates a wide range of activities and interactions that take place in an animal's life. Understanding animal behavior is vital not only for ecological and conservation reasons [1], but also because it provides insights into how different species adapt to their environment, how they communicate, and how they socialize [2]. This knowledge can have implications for a variety of fields, from wildlife management and conservation to agriculture and veterinary medicine.

Studying animal behavior in natural habitats, while clearly important, is extremely challenging. Just finding animals and getting in a position to observe their behaviors in an unobscured and clear way is often quite difficult. Traditionally, two methods are used to observe animal behaviors: *focal sampling* [3] records the behavior of a selected individual for a fixed period of time, while *scan sampling* records the behaviors of multiple individuals within a time interval as the observer gradually sweeps their line of sight through a field of view. These methods capture only a small fraction of the actual behaviors. These twin challenges — *limited access and limited observations* — can potentially be addressed through a combination of an aerial-based (such as drone) video capture to reach and record more animals, and automatic, computer vision-based behavior analysis to find each animal and determine its behavior.

Crucial to the development of modern computer vision technologies for animal behavior studies is the construction of well-curated datasets. Several large-scale datasets have been proposed recently for studying animal behavior recognition [4, 5]. These are generally sourced from online platforms like YouTube, allowing for the collection of a wide

[*]kholim@rpi.edu

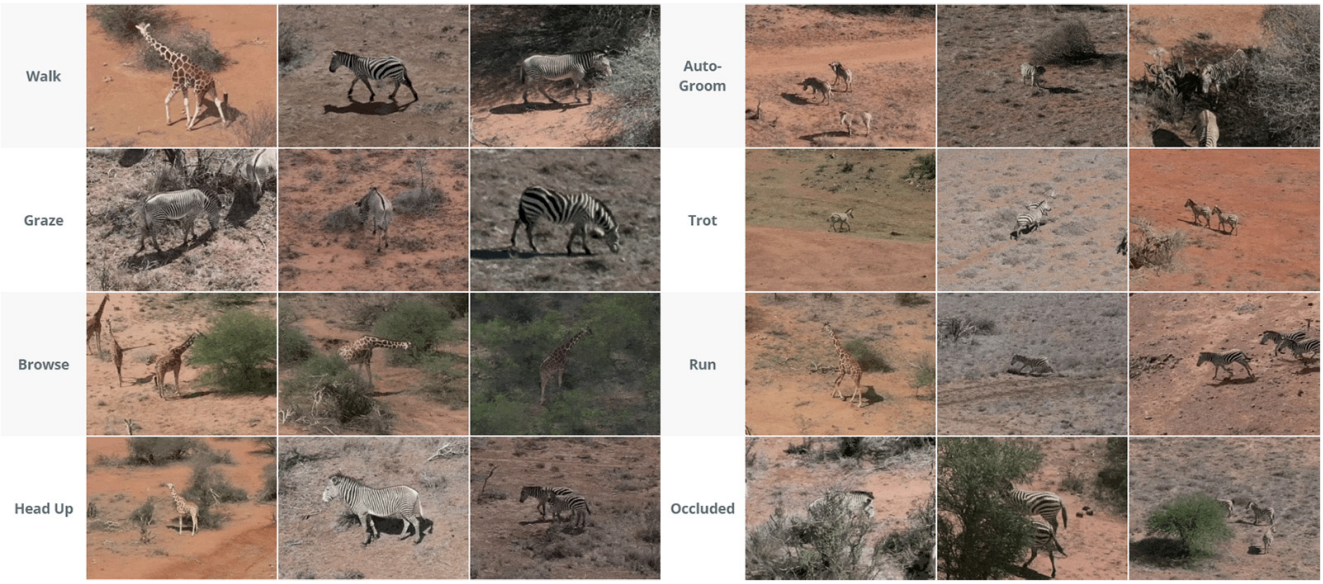

Figure 1. Examples of the behavior of giraffes, plains zebras, and Grevy's zebras from our dataset. It includes a total of eight distinct categories: "Walk", "Graze", "Browse", "Head-Up", "Auto-Groom", "Trot", "Run", and "Occluded".

range of species and behaviors. Complementing this work, there is a clear need for behavior recognition datasets that are collected in-situ and therefore form a more natural representation of behaviors. If drones are to become an important source of animal behavior information there is an equally important need to have experimental datasets that represent the properties of studying behaviors from drone video.

Our work represents an initial stride toward addressing these needs. By introducing a novel dataset collected from drone videos in the natural habitats of Kenyan wildlife, we aim to enrich the current pool of resources available for the study of animal behavior. This dataset, specifically designed to reflect in-situ scenarios, is a pioneering effort to bring the nuances of real-world animal behavior to the forefront of this field of study.

This paper presents a novel dataset for animal behavior recognition collected in-situ from drone videos. Specifically focused on Kenyan wildlife, it contains behaviors of giraffes, plains zebras, and Grevy's zebras. The methodology is extensible to other species and environments. The current dataset includes a total of eight categories that describe various animal behaviors. Examples of selected behaviors are shown in Fig. 1. We make several significant contributions to the study of animal behavior recognition:

1. We introduce a novel technique for building a dataset for behavior recognition from drone videos. See Fig. 2. We detect and track each individual animal in each high-resolution video, and link the results into tracklets. For each tracklet, we create a separate video, called a *mini-scene*, by extracting a sub-image centered on each detection in a video frame. This allows us to compensate for the movement of the drone and provides a stable and zoomed-in representation of the animal. This also preserves fine-grained details of animal behavior, such as auto-grooming.

2. We present a new dataset for animal behavior recognition collected *in-situ* and from *drones*, focused specifically on Kenyan wildlife. The dataset, referred to as Kenyan Animal Behavior Recognition (KABR), comprises over 10 hours of annotated mini-scenes and provides a natural view of animal behavior in the wild, resulting in 54.2 GB of annotated image sequences in the Charades [6] format.

3. We present baseline behavior recognition results using several state-of-the-art, deep learning models for video classification. These show approximately 62% classification rate, indicating the challenge of the KABR dataset. This serves as a starting point for future research.

Our contributions provide a valuable resource for researchers studying animal behavior and ecology, particularly in the context of wildlife conservation efforts in Kenya. By accurately categorizing and analyzing animal behaviors, we can better understand their natural patterns and inform conservation strategies to protect endangered animals.

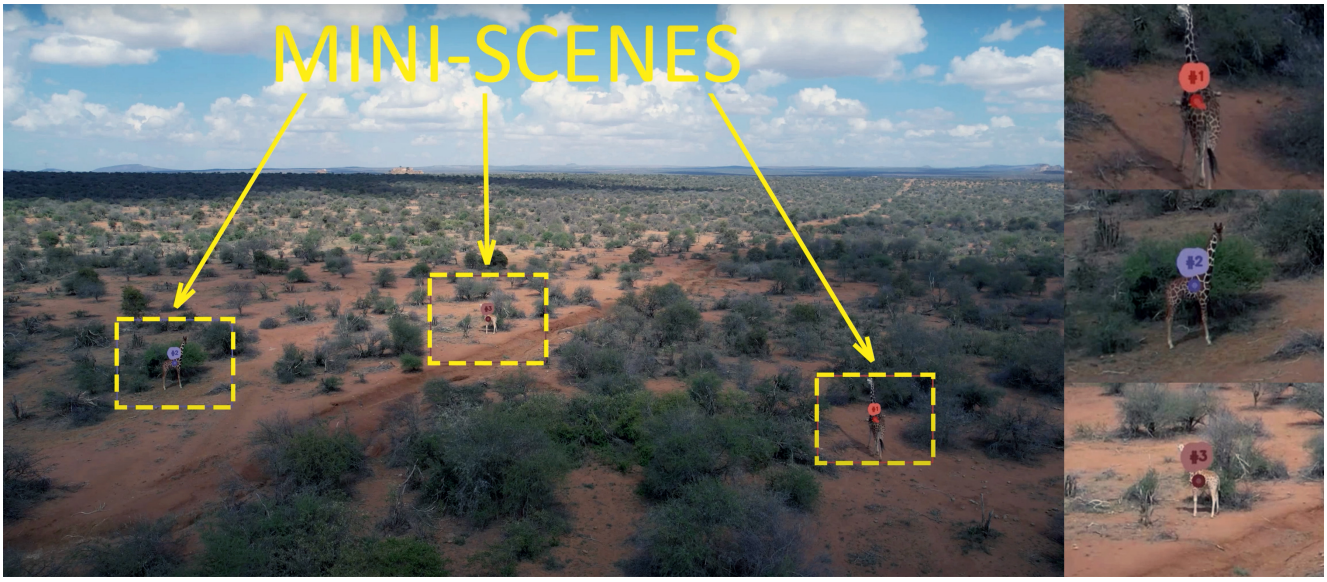

Figure 2. A mini-scene is a sub-image cropped from the drone video footage, centered on and surrounding a single animal. Mini-scenes simulate the camera as well-aligned with each individual animal in the frame, compensating for the movement of the drone, and ignoring everything in the large field of view but the animal's immediate surroundings. The KABR dataset consists of mini-scenes and their frame-by-frame behavior annotation.

2. Related Work

Action classification and action detection are two different tasks in the field of behavior recognition [7]. While both tasks involve analyzing and understanding actions, they differ in their objectives and methodologies. The objective of action classification [8–10] is to assign a single category to a given video, indicating the action being performed in the scene. It aims to identify the overall action without specifying the temporal extent or location of the action. Action detection [11] aims to not only recognize the action category but also detect and localize the temporal extent of the action within a video. We use our concept of mini-scenes to bridge the gap between action detection and recognition.

Action recognition datasets, such as Charades [6] and UCF [4, 12, 13] have been crucial in advancing the field of behavior recognition. However, these datasets are mainly focused on human actions, and may not be suitable for studying animal behavior.

Animal Kingdom [5] and MammalNet [14] are both prominent large-scale datasets for animal behavior recognition. These datasets offer comprehensive collections of annotated video footage featuring a wide range of animal species over 50 and 539 hours, respectively. These datasets primarily rely on videos sourced from online platforms such as YouTube and therefore lack the in-situ aspect of data collection where observations occur directly in animals' natural habitats. APT-36K [15], also sourced from YouTube videos, further pushes to bridge the gap between behavior recognition and animal detection, with a collection of 80 video clips for each of the 30 species represented. In our paper, we contribute to bridging this gap by introducing a novel in-situ dataset specifically centered around Kenyan wildlife.

Prior research has explored the potential of drone videos in addressing challenges related to animal behavior recognition. Notably, Koger et al. [16] introduced a deep learning method focused on reconstructing landscapes from drone videos, enabling the recognition of animal body postures and the ecological context in which they reside. In contrast to the proposed approach, our method is focused on recognizing animal behavior at the individual level rather than understanding the relationship between animals and their landscapes. Additionally, the authors of [17] employed drones to study spatial positioning within groups of feral horses, while [18] used drones to track sharks, unveiling their movement patterns. Furthermore, drone technology was harnessed by [19] for wildlife detection. These diverse applications underscore the potential of drone videos in advancing our understanding of animal behaviors and ecological dynamics.

Several other substantial datasets have been meticulously assembled with a strong focus on recognizing animals, estimating their poses from images [20, 21], or generating new views of images with animals [22]. For instance, the iNaturalist dataset [23] contains over 859,000 images of more than 5,000 different types of plants and animals. Similarly, the iWildCam [24] dataset contains 263,528 images from

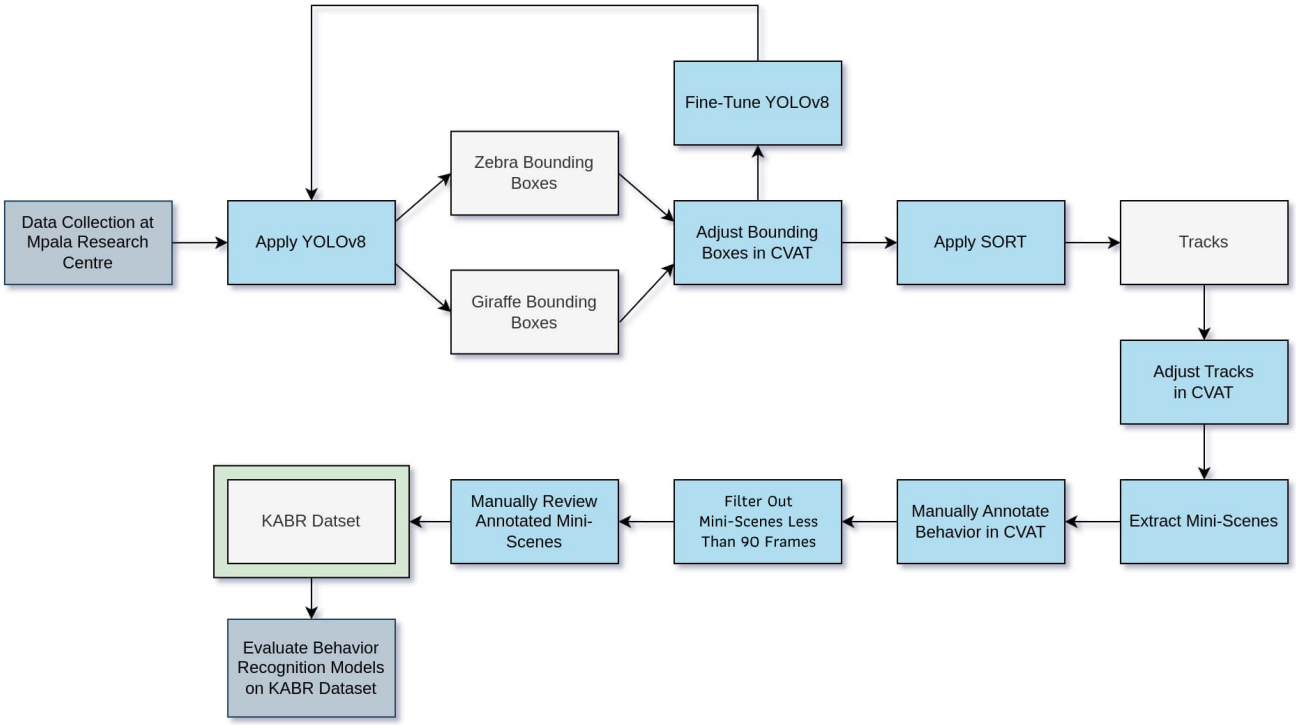

Figure 3. Overview of the pipeline for KABR dataset preparation.

323 locations of camera traps. These datasets provide a plethora of sample images, but they are designed to classify species and count individual animals in images rather than study their behavior.

Some works have proposed targeted solutions for recognizing the behavior of certain animals. These solutions are often based on specific characteristics of the animal's behavior, which may not apply to other species. For instance, a study may focus on recognizing the behavior of primates [25–27], pigs [28–30], goats [31], cows [32, 33], meerkats [34], dogs [35], cats [36], or mice [37–40]. Though these specialized solutions are useful for studying particular animal behaviors, they are typically smaller and may not generalize well to other species or contexts. Therefore, it is important to consider the scope and limitations of these targeted approaches when using them to study animal behavior.

In contrast, our dataset offers a distinctive, valuable contribution to the field of animal behavior recognition, as it focuses specifically on in-situ drone videos of Kenyan wildlife. Our innovative approach provides numerous benefits over traditional video analysis methods and supplies a valuable resource for researchers studying animal behavior and ecology, particularly within the critical context of wildlife conservation efforts in Kenya.

3. Dataset

3.1. Data Collection

The drone videos used in our dataset were collected by our research team at Mpala Research Centre, Kenya. The data collection period spanned from January 6, 2023, to January 21, 2023. During this time, our team conducted multiple expeditions to different locations within the research center's vicinity. The drone flights were strategically planned to capture the behaviors of giraffes, plains zebras, and Grevy's zebras. These species were selected based on their ecological importance and conservation status in the region.

The dataset consists of 1,139,893 individual frames: 488,638 featuring Grevy's zebras, 492,507 of plains zebras, and 158,748 frames featuring giraffes. In total, there are 14,764 distinct sets of behaviors. To ensure high-quality footage, our team utilized DJI Mavic 2S drones equipped with advanced camera capabilities. The videos were recorded in 5.4K resolution at a speed of 29.97 frames per second, providing a smooth and accurate representation of the animals' behaviors. The drones were flown at varying altitudes and distances from the animals to capture a diverse range of perspectives. The distances maintained during the flights ranged from 10 meters to 50 meters away from the animals, depending on the specific circumstances and safety

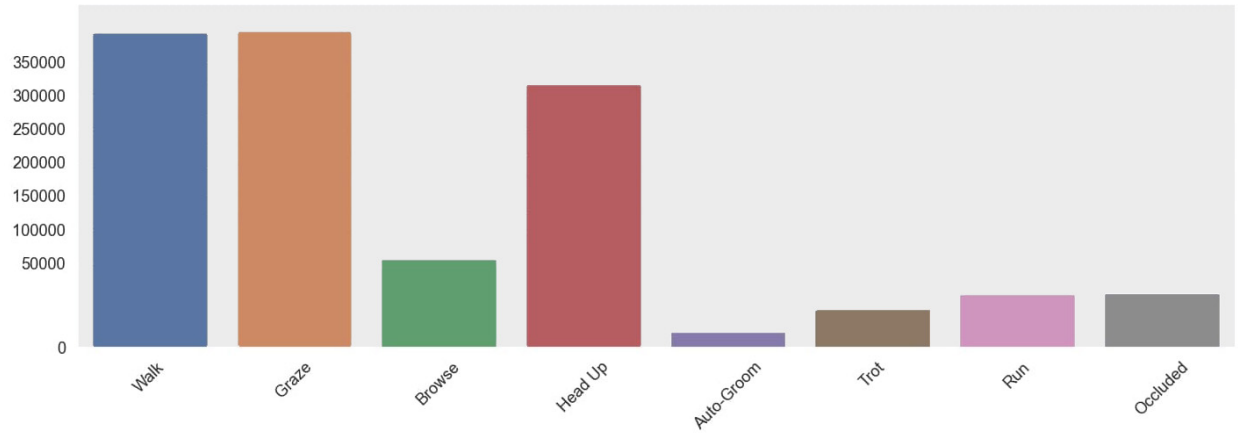

Figure 4. The distribution of classes in the KABR dataset.

regulations. The diversity in recording distances allows us to observe behaviors at different scales, and will eventually allow us to consider social dynamics within animal groups.

During the flights, the pilot carefully maneuvered the drones to capture the behaviors of the animals. The pilot employed a variety of flight paths, including vertical ascents and descents, circular orbits, and linear trajectories, depending on the specific behavior being recorded. The maneuvers were executed with precision and consideration for the animals' well-being, maintaining a safe and non-intrusive distance.

3.2. Ethical Considerations

Two important categories of ethical considerations were addressed in our work. First, no humans appeared in the videos, and all participants were faculty, students, or employees of the Mpala Research Centre. Second, our research was conducted under the authority of a Nacosti Research License. This license confirms our adherence to the regulations in place and allows us to collect drone footage of animals in their natural habitats. We followed a data collection protocol that strictly complies with the guidelines set forth by the Institutional Animal Care and Use Committee (IACUC). These guidelines are designed to ensure the ethical and humane treatment of animals involved in research activities. We also followed the guidelines laid out in [41]. One particular instance of this is that we consistently approached the animals from downwind, allowing the noise to dissipate before reaching the animals.

3.3. Data Curation — Mini-Scenes

The raw drone video data typically contains multiple animals in each frame with each animal occupying a small fraction of the image. In our case, the maximum number of animals visible in the frame at one time is 13. Attempting to directly analyze these to extract behavior is impractical. Instead, we extract mini-scenes, which are sub-videos of the full-resolution video, each of which is centered on an animal as it moves through the scene, and cropped to the animal and its immediate surroundings. The use of mini-scenes allows us to compensate for much of the movement of the drone and provides a stable, zoomed-in representation of the animal's behavior. This approach allows for accurate tracking of individual animals within a group. We anticipate that in future work this will be particularly useful for studying social dynamics among animals.

To implement our mini-scenes approach, we utilized YOLOv8 [42] to detect the animals in each frame and the SORT [43] tracking algorithm to follow their movement. We then extract a window of size 400 pixels wide and 300 tall, values determined empirically based on the characteristics of the animals observed and the surrounding environment, and properties of the drone.

We have developed a set of tools to facilitate the data annotation process. One of the tools we used extensively was the interpolation tool, which filled in any missing detections within a track, thereby improving the overall tracking quality. The tool uses a linear interpolation algorithm that estimates an animal's location based on its previous movements, helping fill in gaps where automatic detection may have failed. Our data processing pipeline is illustrated in Fig. 3. We considered a mini-scene to be inappropriate if it did not satisfy the length criterion. If the total length of the behaviors in a mini-scene was less than 90 frames, we filter it out.

The mini-scenes we extracted using our pipeline are a

Method	All	Giraffes	Plains Zebras	Grevy's Zebras
I3D (16x5)	53.41	61.82	58.75	46.73
SlowFast (16x5, 4x5)	52.92	61.15	60.60	47.42
X3D (16x5)	61.9	65.1	63.11	51.16

Table 1. The results of the I3D, SlowFast, and X3D models on our dataset. I3D and X3D were trained with 16 input frames with a sampling rate of 5. For SlowFast, the Slow branch was trained with 16 input frames with a sampling rate of 5, and the Fast branch was trained with 4 input frames with a sampling rate of 5. The results reflect the macro (per class) average metric.

crucial component of the manual annotation process for behavior recognition. These mini-scenes provide a zoomed-in and stable view of individual animals' behavior, making it easier for human annotators to accurately identify and label their behavior.

3.4. Behaviors and Annotation

Our dataset contains a total of eight behavior categories, including "Walk", "Graze", "Browse", "Head Up", "Auto-Groom", "Trot", "Run", and "Occluded" as determined by our expert behavioral ecologist looking at the properties of the videos. These include three locomotion behaviors, "Walk", "Trot" and "Run", each representing a different gait. "Run" could have been split into canter and gallop, but these were too infrequent and indistinguishable. Two of the other behaviors refer to eating: "Graze" refers to the behavior of an animal when they are eating grass or other vegetation, while "Browse" describes the behavior of animals feeding on trees and bushes. For the remaining categories, "Head Up" refers to the behavior of an animal when it lifts its head to look around or observe its surroundings, typically, these are different types of vigilance, and "Auto-Groom" describes the behavior of animals when they groom themselves, which can include licking, scratching, or rubbing their bodies. Finally, the category of "Occluded" is used when the animal is not fully visible in the video footage. This can occur due to obstructions such as trees or other animals blocking the view, or due to technical limitations of the camera or drone.

To ensure accurate behavior annotation in our dataset, we employed a team of 10 individuals, all of whom were trained in the process. The team was led by an experienced expert behavioral ecologist who oversaw the annotation process. We utilized CVAT [44], a powerful tool for collaborative video annotation, to enable the team to work together remotely and efficiently. Once the initial annotations were complete, we took an additional step to ensure quality control by having all videos manually reviewed by a designated annotator. Finally, we utilized an automatic filtering process to split the annotated videos into convenient training iterations based on their resulting length. This ensured that the training data was properly organized and could be effectively used in the development of deep learning models. Overall, our comprehensive annotation process and quality control measures ensure that our dataset is accurate, reliable, and suitable for a wide range of research applications.

3.5. Class Distribution

Our dataset exhibits a long-tailed distribution, signifying a considerable disparity in the count of samples across the categories. This is expected since certain behaviors are considerably more frequent in animals' natural settings compared to other behaviors. The distribution of classes is shown in Fig. 4. Similar imbalances occur in recent larger datasets [5, 14, 15] scraped from YouTube.

3.6. Data Split

We provide a train-test split of the mini-scenes for evaluation purposes, with 75% for train and 25% for testing. No mini-scene was divided by the split. The splits ensured a stratified representation of giraffes, plains zebras, and Grevy's zebras.

4. Experiments

To comprehensively assess the performance of different models on our dataset, we conduct evaluations using three well-known architectures: I3D [45], SlowFast [46], and X3D [47]. The results are summarized in Tab. 1, where we report the Top-1 accuracy scores for all species, giraffes, plains zebras, and Grevy's zebras.

The model was trained for 120 epochs. During training, we use a batch size of 5. To improve the model's performance and reduce the risk of overfitting, we apply data augmentation techniques during training. Specifically, we use flip augmentation to randomly mirror the input frames horizontally, and color augmentations to randomly modify the brightness, contrast, and saturation of the input frames.

To address the issue of long-tailed distribution, we employ the EQL [48] loss function. The proposed loss function selectively ignores gradients for frequent categories, enabling the learning of rare categories during network parameter updates.

The confusion matrix depicted in Fig. 6 demonstrates the performance of the X3D model. The model performs

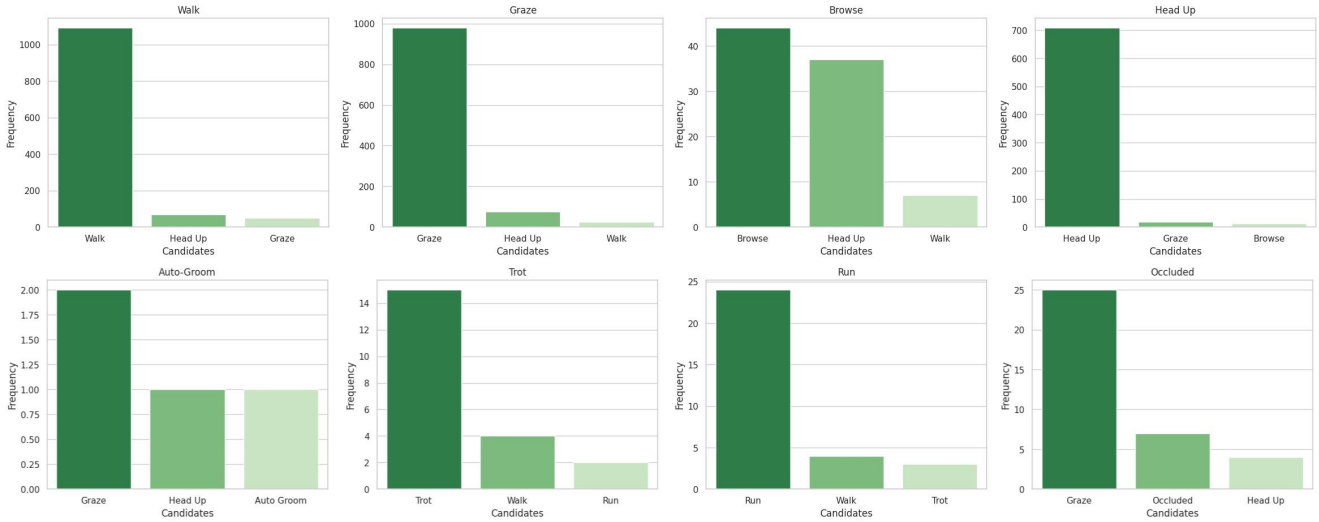

Figure 5. A bar plot representation of the three most frequently predicted classes from the X3D model for each category within the KABR dataset.

Figure 6. This confusion matrix showcases the performance of the X3D model, which has been determined as the top-performing model on the KABR dataset based on our evaluation.

quite well for the most frequent behaviors in our dataset: "Walk", "Graze", "Browse", and "Head Up". The imbalance seen here is reflected in the difference between the macro and micro average Top-1 accuracy scores. The macro (per class) average, reported in Tab. 1 peaks at 61.9%, but the micro (per instance) average is 86.7%. The model also demonstrates good performance for "Trot" and "Run" despite fewer instances in the dataset for these categories. Interestingly, "Trot" is most frequently confused with the other locomotion behaviors, "Walk" and "Run". The same applies to the "Run" behavior.

Further insight can be gained from Fig. 5 which shows the three most frequently occurring predictions made by the X3D model for each category. This illustrates again that, in most cases, the correct category showcases a dominant frequency, noticeably higher than the frequencies of the second and third most common predictions. This highlights the ability of the model to learn from the KABR dataset to predict the correct behavior.

This also highlights some interesting challenge cases: "Browse" (a giraffe behavior) is frequently confused with "Head Up", which is quite intuitive. "Auto-Grooming", a very rare behavior in the KABR dataset, is often misclassified with similar looking behaviors, "Graze" and "Head Up". Finally, the "Occluded" category is often confused with "Grazing", most likely due to the subjectivity of what constitutes an occlusion when looking at video from Mpala's ecosystems. Interestingly, within the "Occluded" category, the model has a tendency to factor in surrounding elements like shrubs and trees (visible in the on-line videos). We anticipate addressing these issues through further data collection and analysis, as enabled by the pipeline developed here.

5. Discussion

The benchmark results using state-of-the-art video classification algorithms indicate that the dataset is both interesting and challenging. Though it is necessarily smaller than recent Animal Kingdom and MammalNet datasets and captures a more focused set of behaviors, it represents an important step in the evolution of animal behavior data col-

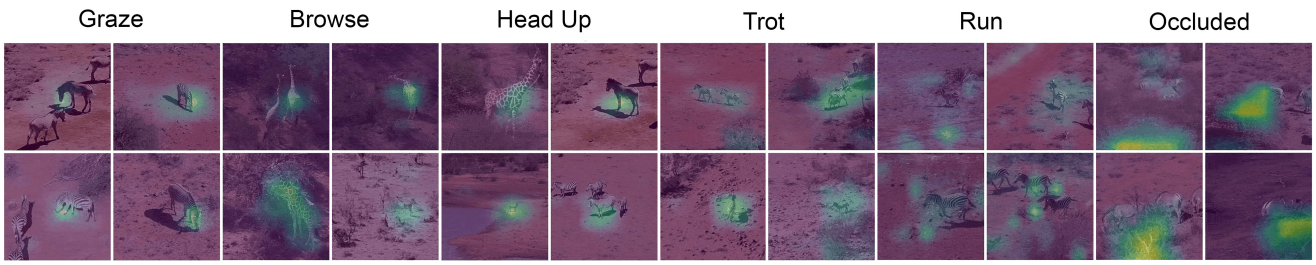

Figure 7. Grad-CAM visualization for different behaviors in the dataset.

lection and analysis because the videos were collected in-situ and from drones. As such it is closer to, and more representative of, how behavioral analysis can be carried out in the field in the future. One limitation of the dataset as it currently exists is that some rare behaviors are either captured infrequently or not at all. The complete set of tools for KABR that we have developed and shared openly form a powerful framework to support searching for examples of these behaviors.

The mini-scenes approach provides a means of rapidly processing high-resolution videos into a form that can be analyzed for individual behaviors. The next step would be to augment the behavior classification approaches to facilitate anomaly detection. An interesting question is the potential integration of KABR with MammalNet or Animal Kingdom for exactly this purpose.

The proposed pipeline has several important advantages. By applying detection and tracking algorithms, we can extract zoomed-in footage that is stabilized on the animal of interest. Consequently, the animal remains consistently centered in the frame throughout the mini-scene, enhancing the accuracy of subsequent analysis. This is unlike typical action recognition where the animal could be moving across a fixed frame. Consequently, if an object moves from one side of the frame to the opposite side, the resulting bounding box may fail to accurately reflect the object's actual position. In contrast, our approach avoids this issue by maintaining the animal of interest at the center of the frame throughout the extracted mini-scene, allowing for more precise localization of the moving object over a longer period of time.

Another important future step is using the mini-scenes approach to analyze complex social behaviors, such as dominance, aggression, mating, and grooming. Behaviors can be analyzed in isolation within each mini-scene, in the overlap between the bounding regions of mini-scenes, and in a graphical representation of a neighborhood of mini-scenes.

A final justification of the efficacy of the mini-scenes approach can be seen in a Grad-CAM analysis [49] of the mini-scene classification activation, as shown in Fig. 7. This demonstrates that the neural network indeed prioritizes the region covered by the animal in the center of the frame and even the body part. In the case of the Occluded category, where the animal is not visible within the frame, the network shifts its attention to focus on other objects present. In the case of Run, the background changes very rapidly, especially in the region that is being newly occluded in each frame as the animal moves. This allows the network to identify it as Run.

6. Conclusion

This paper has presented a new in-situ dataset for animal behavior recognition from drone videos, with a focus on Kenyan wildlife, including giraffes, plains zebras, and Grevy's zebras. We introduced a novel technique for building this dataset, which compensates for the movement of the drone and allows us to capture fine-grained details of animal behavior. Our dataset contains eight categories that describe various animal behaviors, providing a comprehensive view of animal behavior in their natural habitat. Our baseline solution demonstrates the effectiveness of our dataset for training conventional deep-learning models for video classification. Our contributions provide a valuable resource for researchers studying animal behavior and ecology, particularly in the context of wildlife conservation efforts in Kenya. Our work represents an important step forward in the field of animal behavior recognition and provides a solid foundation for future research in this area.

7. Acknowledgments

This material is based upon work supported by the National Science Foundation under Award No. 2118240 and Award No. 2112606. The data was gathered at the Mpala Research Centre in Kenya, in accordance with Research License No. NACOSTI/P/22/18214. The data collection protocol adhered strictly to the guidelines set forth by the Institutional Animal Care and Use Committee under permission No. IACUC 1835F.

References

[1] Alison L Greggor, Daniel T Blumstein, Bob Wong, and Oded Berger-Tal. Using animal behavior in conservation management: a series of systematic reviews and maps, 2019. 1

[2] Charles T Snowdon. Animal signals, music and emotional well-being. *Animals*, 11(9):2670, 2021. 1

[3] Jeanne Altmann. Observational study of behavior: sampling methods. *Behaviour*, 49(3-4):227–266, 1974. 1

[4] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arxiv 2012. *arXiv preprint arXiv:1212.0402*, 2012. 1, 3

[5] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19023–19034, 2022. 1, 3, 6

[6] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016. 2, 3

[7] Liangliang Cao, YingLi Tian, Zicheng Liu, Benjamin Yao, Zhengyou Zhang, and Thomas S Huang. Action detection using multiple spatial-temporal interest point features. In *2010 IEEE International Conference on Multimedia and Expo*, pages 340–345. IEEE, 2010. 3

[8] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3

[9] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 3

[10] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 3

[11] Ang Li, Meghana Thotakuri, David A Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*, 2020. 3

[12] Jingen Liu, Jiebo Luo, and Mubarak Shah. Recognizing realistic actions from videos "in the wild". In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1996–2003. IEEE, 2009. 3

[13] Kishore K Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *Machine vision and applications*, 24(5):971–981, 2013. 3

[14] Jun Chen, Ming Hu, Darren J Coker, Michael L Berumen, Blair Costelloe, Sara Beery, Anna Rohrbach, and Mohamed Elhoseiny. Mammalnet: A large-scale video benchmark for mammal recognition and behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13052–13061, 2023. 3, 6

[15] Yuxiang Yang, Junjie Yang, Yufei Xu, Jing Zhang, Long Lan, and Dacheng Tao. Apt-36k: A large-scale benchmark for animal pose estimation and tracking. *Advances in Neural Information Processing Systems*, 35:17301–17313, 2022. 3, 6

[16] Benjamin Koger, Adwait Deshpande, Jeffrey T Kerby, Jacob M Graving, Blair R Costelloe, and Iain D Couzin. Quantifying the movement, behaviour and environmental context of group-living animals using drones and computer vision. *Journal of Animal Ecology*, 2023. 3

[17] Sota Inoue, Shinya Yamamoto, Monamie Ringhofer, Renata S Mendonça, Carlos Pereira, and Satoshi Hirata. Spatial positioning of individuals in a group of feral horses: A case study using drone technology. *Mammal Research*, 64:249–259, 2019. 3

[18] Vincent Raoult, Louise Tosetto, and Jane E Williamson. Drone-based high-resolution tracking of aquatic vertebrates. *Drones*, 2(4):37, 2018. 3

[19] Evangeline Corcoran, Megan Winsen, Ashlee Sudholz, and Grant Hamilton. Automated detection of wildlife using drones: Synthesis, opportunities and constraints. *Methods in Ecology and Evolution*, 12(6):1103–1114, 2021. 3

[20] Hemal Naik, Alex Hoi Hang Chan, Junran Yang, Mathilde Delacoux, Iain D Couzin, Fumihiro Kano, and Máté Nagy. 3d-pop-an automated annotation approach to facilitate markerless 2d-3d tracking of freely moving birds with marker-based motion capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21274–21284, 2023. 3

[21] Hongmin Shao, Jingyu Pu, and Jiong Mu. Pig-posture recognition based on computer vision: Dataset and exploration. *Animals*, 11(5):1295, 2021. 3

[22] Simon Giebenhain, Urs Waldmann, Ole Johannsen, and Bastian Goldluecke. Neural puppeteer: Keypoint-based neural rendering of dynamic shapes. In *Proceedings of the Asian Conference on Computer Vision*, pages 2830–2847, 2022. 3

[23] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 3

[24] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021. 3

[25] Xiaoxuan Ma, Stephan P Kaufhold, Jiajun Su, Wentao Zhu, Jack Terwilliger, Andres Meza, Yixin Zhu, Federico Rossano, and Yizhou Wang. Chimpact: A longitudinal dataset for understanding chimpanzee behaviors. *arXiv preprint arXiv:2310.16447*, 2023. 4

[26] Yujie Lei, Pengmei Dong, Yan Guan, Ying Xiang, Meng Xie, Jiong Mu, Yongzhao Wang, and Qingyong Ni. Postural behavior recognition of captive nocturnal animals based on deep learning: a case study of bengal slow loris. *Scientific Reports*, 12(1):7738, 2022. 4

[27] Max Bain, Arsha Nagrani, Daniel Schofield, Sophie Berdugo, Joana Bessa, Jake Owen, Kimberley J Hockings, Tetsuro Matsuzawa, Misato Hayashi, Dora Biro, et al. Automated audiovisual behavior recognition in wild primates. *Science Advances*, 7(46):eabi4883, 2021. 4

[28] Dan Li, Yifei Chen, Kaifeng Zhang, and Zhenbo Li. Mounting behaviour recognition for pigs based on deep learning. *Sensors*, 19(22):4924, 2019. 4

[29] Kaifeng Zhang, Dan Li, Jiayun Huang, and Yifei Chen. Automated video behavior recognition of pigs using two-stream convolutional networks. *Sensors*, 20(4):1085, 2020. 4

[30] Jake Cowton, Ilias Kyriazakis, and Jaume Bacardit. Automated individual pig localisation, tracking and behaviour metric extraction using deep learning. *IEEE Access*, 7:108049–108060, 2019. 4

[31] Min Jiang, Yuan Rao, Jingyao Zhang, and Yiming Shen. Automatic behavior recognition of group-housed goats using deep learning. *Computers and Electronics in Agriculture*, 177:105706, 2020. 4

[32] Chuong Nguyen, Dadong Wang, Karl Von Richter, Philip Valencia, Flavio AP Alvarenga, and Gregory Bishop-Hurley. Video-based cattle identification and action recognition. In *2021 Digital Image Computing: Techniques and Applications (DICTA)*, pages 01–05. IEEE, 2021. 4

[33] Ali Zia, Renuka Sharma, Reza Arablouei, Greg Bishop-Hurley, Jody McNally, Neil Bagnall, Vivien Rolland, Brano Kusy, Lars Petersson, and Aaron Ingham. Cvb: A video dataset of cattle visual behaviors. *arXiv preprint arXiv:2305.16555*, 2023. 4

[34] Mitchell Rogers, Gaël Gendron, David Arturo Soriano Valdez, Mihailo Azhar, Yang Chen, Shahrokh Heidari, Caleb Perelini, Padriac O'Leary, Kobe Knowles, Izak Tait, et al. Meerkat behaviour recognition dataset. *arXiv preprint arXiv:2306.11326*, 2023. 4

[35] Yumi Iwashita, Asamichi Takamine, Ryo Kurazume, and Michael S Ryoo. First-person animal activity recognition from egocentric videos. In *2014 22nd International Conference on Pattern Recognition*, pages 4310–4315. IEEE, 2014. 4

[36] Liqi Feng, Yaqin Zhao, Yichao Sun, Wenxuan Zhao, and Jiaxi Tang. Action recognition using a spatial-temporal network for wild felines. *Animals*, 11(2):485, 2021. 4

[37] Brian Q Geuther, Asaf Peer, Hao He, Gautam Sabnis, Vivek M Philip, and Vivek Kumar. Action detection using a neural network elucidates the genetics of mouse grooming behavior. *Elife*, 10:e63207, 2021. 4

[38] Hueihan Jhuang, Estibaliz Garrote, Xinlin Yu, Vinita Khilnani, Tomaso Poggio, Andrew D Steele, and Thomas Serre. Automated home-cage behavioural phenotyping of mice. *Nature communications*, 1(1):68, 2010. 4

[39] Xavier P Burgos-Artizzu, Piotr Dollár, Dayu Lin, David J Anderson, and Pietro Perona. Social behavior recognition in continuous video. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1322–1329. IEEE, 2012. 4

[40] Cristina Segalin, Jalani Williams, Tomomi Karigo, May Hui, Moriel Zelikowsky, Jennifer J Sun, Pietro Perona, David J Anderson, and Ann Kennedy. The mouse action recognition system (mars) software pipeline for automated analysis of social behaviors in mice. *Elife*, 10:e63720, 2021. 4

[41] Isla Duporge, Marcus P Spiegel, Eleanor R Thomson, Tatiana Chapman, Curt Lamberth, Caroline Pond, David W Macdonald, Tiejun Wang, and Holger Klinck. Determination of optimal flight altitude to minimise acoustic drone disturbance to wildlife using species audiograms. *Methods in Ecology and Evolution*, 12(11):2196–2207, 2021. 5

[42] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, January 2023. 5

[43] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 5

[44] Boris Sekachev, Nikita Manovich, Maxim Zhiltsov, Andrey Zhavoronkov, Dmitry Kalinin, Ben Hoff, TOsmanov, Dmitry Kruchinin, Artyom Zankevich, DmitriySidnev, Maksim Markelov, Johannes222, Mathis Chenuet, a andre, telenachos, Aleksandr Melnikov, Jijoong Kim, Liron Ilouz, Nikita Glazov, Priya4607, Rush Tehrani, Seungwon Jeong, Vladimir Skubriev, Sebastian Yonekura, vugia truong, zliang7, lizhming, and Tritin Truong. opencv/cvat: v1.1.0, August 2020. 6

[45] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[46] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 6

[47] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020. 6

[48] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11662–11671, 2020. 6

[49] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8