Detecting Rhino Middens with Computer Vision and Active Learning: Preliminary Results

Lucia Gordon,[*1†] **Samuel Collier,**[*1‡] **Nikhil Behari,**[2] **Elizabeth Bondi-Kelly,**[2] **Peter Boucher,**[3] **Catherine Ressijac,**[3] **Andrew Davies,**[3] **Milind Tambe**[1]

[1] Center for Research on Computation and Society, Harvard University
[2] Massachusetts Institute of Technology
[3] Department of Organismic and Evolutionary Biology, Harvard University

Abstract

Wild rhinos are threatened with extinction as a result of human activities, in particular, poaching. Monitoring rhinos in the wild has proven difficult, which limits the effectiveness of conservation efforts. The presence of rhino middens, which are communal defecation sites, in a landscape can give information on a rhino population inhabiting the area. Despite the potential insights that can be gained into rhino population distributions and habitat use, rhino middens have yet to be mapped on large scales. We build classifiers to detect middens from remotely sensed thermal and RGB imagery using computer vision and active learning techniques. We present initial results, training and testing our classifiers on a novel dataset, resulting in a maximum accuracy of 0.76 ± 0.05 for the passive fused model. The long-term goal of this work is to detect rhino middens with as few labeled images as possible to save ecologists time.

Introduction

Rhinos are one of the most targeted animals by illegal poachers in Africa; on average at least one rhino is killed every day, putting the species at risk of extinction (SaveTheRhino 2022). The vast majority of African rhino poaching deaths occur in South Africa, where the rhino population in the country's largest national park, Kruger, has declined by 59% since 2013 (SaveTheRhino 2022). Understanding the distribution of rhino populations and how their habitat use changes in response to poaching is crucial to implementing effective conservation measures (Johnson and Gillingham 2005). Rhinos are elusive as well as dangerous to observe directly (Linklater, Mayer, and Swaisgood 2013), but their populations can be tracked by locating their communal feces piles, called middens (see Figure 1). These communal middens are used for social communication through smell by white and black rhinos (Marneweck, Jürgens, and Shrader 2017). The importance of middens in rhino populations suggests they likely influence and reflect movement and distribution patterns. A greater understanding of midden distributions will therefore give insights into rhino territorial ranges and habitat use, which can then inform anti-poaching strategies and conservation efforts. Detecting rhino populations in this way is also more practically efficient, as the static middens require less frequent re-mapping than methods identifying individual animals. Moreover, it is far less invasive than placing GPS trackers on rhinos.

Figure 1: A white rhino midden next to a road in iMfolozi (Marneweck 2013)

The Davies lab in Harvard's Department of Organismic and Evolutionary Biology has captured thermal and RGB imagery of a site in Kruger National Park using a drone. Bondi, Ressijac, and Boucher 2021 made a first attempt at detecting rhino middens in this imagery by exploiting the fact that rhino middens are often warm relative to the surrounding ground. Following this principle, they used a thresholding methodology to try to identify images that are more likely to contain middens. We extend this work by using deep learning to more robustly identify rhino middens in this same remotely sensed, unpublished dataset. Deep learning has often been used to analyze large volumes of remotely sensed data, e.g. for wildfires (Bouguettaya et al. 2022). The key idea behind deep learning is that repeated simple, non-linear transformations of an input allow for abstract and complex representations of features of the input that are learned, not assigned. In the case of convolutional neural networks, alternating sequences of convolutional and pooling layers form filters that extract features in an image while constraining the dimensionality of the features to be computationally tractable. These feature-extracting layers are often then connected to one or more fully-connected layers to perform the actual image classification. The whole

[*]These authors contributed equally.
[†]luciagordon@g.harvard.edu
[‡]scollier1@g.harvard.edu

network is trainable using backpropagation.

Several studies have fused thermal and RGB data in an attempt to create better-performing deep learning models. Alexander et al. 2022 utilized thermal and RGB fusion in a deep learning method for detecting cracks in civil infrastructure. They found that their fused model outperformed an RGB-only model. Bakalos et al. 2021 fed fused thermal and RGB imagery along with channel state information from WiFi into a bidirectional LSTM to detect abnormal activity in critical water infrastructure. Their combination of three data modalities outperformed fused thermal and RGB imagery or channel state information alone. Speth et al. 2022 harnessed the power of fused thermal and RGB imagery collected from drones to detect and locate people in disaster zones. They tested two methods for fusion. In the first, they performed late fusion, in which separate models are trained for the different modalities and the predictions are merged at the end. In the second, they used a neural network architecture that accepts 4 channels in order to train a single model on the fused data, which they call early fusion. They obtained better performance for the early fusion model than for the late fusion model.

Active learning, as opposed to passive deep learning, is a distinct machine learning paradigm that seeks to achieve greater accuracy with fewer labeled training instances by allowing a machine learning algorithm to choose the training data from which it learns (Settles 2009). Active learning algorithms are generally distinguished by their query strategy, the method they use to evaluate how informative a given unlabeled sample is (Settles 2009). Some classic examples include the uncertainty in the label for a sample (Lewis and Gale 1994), as measured through probabilities or entropy, the disagreement between a committee of models on a certain sample, called Query-By-Committee (Seung, Opper, and Sompolinsky 1992), or by choosing the sample to label that will minimize the model's estimated generalization error (Roy and McCallum 2001). Active learning has been successfully applied to image classification tasks to avoid having to hand-label very large image datasets (Li and Guo 2013), including in remote sensing applications (Tuia et al. 2011) and for multi-spectral imagery (Haut et al. 2018). We compare the performance of passive and active deep learning methods for detecting rhino middens in thermal, RGB, and fused imagery.

Data Collection

The remote sensing data used for this project was collected by a UAV flown over a roughly 2 km x 2 km site in Kruger National Park back in January 2020. A Sony A6000 camera (24 mp) and a FLIR Tau-2 thermal camera (327,680 mp) were used to take images as a DJI M600 multicopter flew at an altitude of 100 m at a speed of 8 m/s over the region of study. The thermal infrared images taken were converted into false-color images with a fourth temperature band using ThermoViewer software and an automatic color scale. The imagery was rectified and mosaicked with the Terrasolid software suite. The thermal imagery was rectified and mosaicked at a resolution of 0.05 m and the RGB imagery at a resolution of 0.5 m (Bondi, Ressijac, and Boucher 2021).

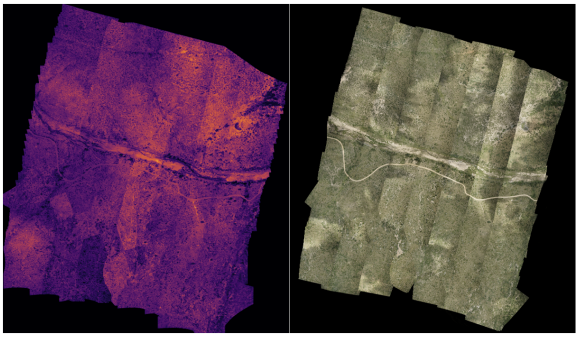

Figure 2: Thermal (left) and RGB (right) orthomosaics

Data Engineering

Several data engineering steps were necessary to create the datasets we ultimately used to train and test our models. It was first necessary to crop the background around the photographed region to ensure that the thermal and RGB orthomosaics covered the same land area in meters. Next, we performed additional cropping to remove as many background rows and columns as possible, with the same removal operations being done to both orthomosaics to preserve the geographic correspondence between them. Doing so, we obtained the orthomosaics shown in Figure 2. We then added back the necessary number of background rows and columns to ensure that an integral number of images could be cropped from the orthomosaics without cutting off any pixels given our chosen interval of 20 m (40 pixels for thermal and 400 for RGB) and stride of 5 m (10 pixels for thermal and 100 for RGB).

We were given the x and y coordinates of 79 rhino middens in the region of interest identified by experts in the field. We mapped these middens onto the thermal orthomosaic, yielding a 2D-matrix of the same size. Each value in the matrix was assigned to be 1 if the corresponding pixel in the thermal orthomosaic was the center of a midden and 0 otherwise. This midden matrix was cropped simultaneously with the thermal orthomosaic such that each individual image had a corresponding midden matrix. Each of these images was assigned the label 0 if its midden matrix had all zeros and otherwise was assigned the label 1. By construction, the cropping of the RGB orthomosaic yielded the same number of individual images, each of which corresponded to one of the previously labeled thermal images. Thus, we labeled each RGB image with 0 (1) if its corresponding thermal image had the label 0 (1). The initial cropping procedure yielded 132 images with middens and 14,316 empty images. We then removed any empty images with all zeros, leaving 10,781 empty images, which means that our dataset contains 10,913 images, 1.21% of which have a midden. In order to prepare the images to be fed into a pretrained VGG16 model, we converted all of them to PNGs with a DPI of 60.7 so that the arrays corresponding to the images would be of shape (224,224,3). The dataset of fused images was created by using the blend function in the PIL class with an alpha value of 0.5 to blend the 10,913 pairs of thermal and RGB images. This particular fusion method was selected so that the re-

sulting fused images would be in a format compatible with the VGG16 architecture. Each of the fused images was then labeled according to the pair of images from which it was generated. These were then converted back into arrays as was done for the thermal and RGB images.

Computer Vision Methodology

Given the rarity of midden images in our dataset, we employed transfer learning with a VGG16 model pretrained on the ImageNet dataset. VGG16 was selected as it is powerful but still quite fast to train, and it has been used previously in remote sensing applications (Bouguettaya et al. 2022). We froze all the parameters in the model and recreated the classifier to have three sets of linear, ReLU, and dropout layers with the linear layers having 512, 256, and 128 out features, respectively. We added a final linear layer with a single out feature before concluding the architecture with a sigmoid function so that the output of our model would represent the probability that an image contains a midden. All the parameters in our classifier were trainable. We used a batch size of 16, the Binary Cross Entropy loss function in PyTorch along with an Adam optimizer with a learning rate of 0.0001. We performed several transformations on the images before they were fed into the neural network. First, we subtracted the minimum pixel value of each image such that each image had a minimum of 0. We then divided each image by its maximum pixel value such that each image had a maximum of 1. Next, we normalized the pixel values so that the three bands would have means of 0.485, 0.456, and 0.406, respectively and standard deviations of 0.229, 0.224, and 0.225, respectively. Finally, we reshaped each image from (224,224,3) to (3,224,224).

Preliminary Passive Learning Results

Our passive learning models began with all the data labeled. We split this labeled data into training and testing groups as follows. First, we added 80% of the midden images to the training set and left 20% for the test set. We added the same number of empty images to the test set, yielding a balanced test set of 52 images. Of the remaining empty images, we took for training the same number of images as there were midden images in the training set so that our training set was also balanced. We performed 5 trials for the thermal, RGB, and fused models, and each trial had a different assignment of images to the training and test sets. For each trial, we trained the models for 20 epochs and recorded the accuracy on the test set at the end of each epoch as shown in Figure 3 for the thermal model, Figure 4 for the RGB model, and Figure 5 for the fused model. We used a threshold of 0.5 for classification. We also report the mean and standard deviation of the accuracy, precision, recall, and F1-score across the 5 trials at the end of the 20 epochs in Table 1, where the F1-score is the harmonic mean of the precision and recall. We observe that the fused model achieves the best performance on all metrics, outperforming the individual thermal and RGB models and pointing to the advantages of exploiting multimodality in image classification.

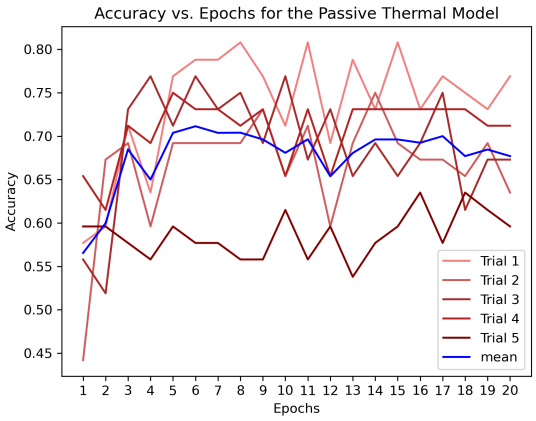

Figure 3: Accuracy for the passive thermal model across 5 trials when training for 20 epochs

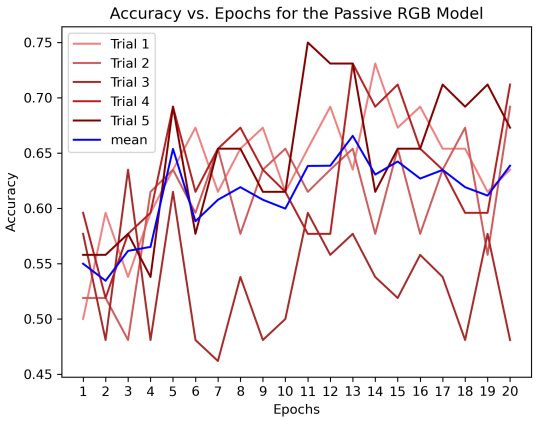

Figure 4: Accuracy for the passive RGB model across 5 trials when training for 20 epochs

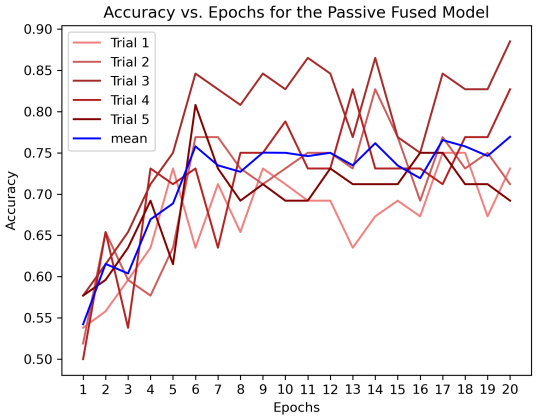

Figure 5: Accuracy for the passive fused model across 5 trials when training for 20 epochs

	Thermal	RGB	Fused
Accuracy	0.67 ± 0.06	0.64 ± 0.05	0.76 ± 0.05
Precision	0.69 ± 0.09	0.64 ± 0.04	0.77 ± 0.05
Recall	0.64 ± 0.06	0.64 ± 0.11	0.75 ± 0.10
F1-score	0.66 ± 0.06	0.64 ± 0.07	0.75 ± 0.06

Table 1: Passive learning statistics

Active Learning Methodology

Training our network in this passive way required 10,000+ images to be labeled. While our ecologist collaborators were able to provide the midden locations for this site of interest, they would also like to map middens in other sites they have photographed. Such a problem is infeasible with passive deep learning techniques, motivating an active learning strategy. It is not uncommon for unlabeled data to massively outpace labeling efforts in social impact domains. Active learning aims to reduce the number of images that need to be labeled to reach a desired level of performance by only asking a user to label the images that are most useful to the model's learning. From the set of unlabeled images, a small subset is chosen according to some criteria and sent to an expert for hand-labeling. This small batch of labeled images is used to begin training the model. Some criteria is then used to select the next batch of images to be labeled. This process continues until some stopping criteria is met, e.g. some budget for hand-labeling is reached. The trained model can then be run on all of the images to classify the entire dataset.

Our proposed algorithm (see Algorithm 1) uses a selection criteria specific to our domain. Due to the drastic class imbalance in our data, we must prioritize feeding the model balanced batches of images. We sort the raw, single-channel thermal images by their maximum pixel value, leveraging the fact that middens often show up as hotspots. Those images with the highest maximum pixel value, which are the ones most likely to contain middens, are provided labels. The model then trains on this labeled set and classifies a small set of unlabeled images, which have the next highest maximum pixel values. Each image is assigned a score, which is the sum of the model's prediction and the image's maximum pixel value on a normalized 0-1 scale. A subset of those with the highest score are the next to be labeled. For active learning we use the same dataset and perform the same train-test split as for passive learning except that we do not balance the training set. We report the accuracy of the system as the number of labeled images increases in Figure 6. We plan to improve our active learning results by using the histograms of the images to incorporate more information into the querying method than solely the maximum pixel value of each image.

Algorithm 1: Proposed Algorithm for Active Learning

Input: List A of the normalized maximum pixel values of the unlabeled training images U with labels L
Variables: Batch size b, number of images to predict on i

while images labeled $<$ labeling budget **do**
  Select the i images in U corresponding to the indices with the maximum values in A
  Run the thermal model on those images
  Calculate the scores for each of the images as the sum of its model output and maximum pixel value
  Select the b images out of the set of those predicted on that have the highest score
  Keep all the midden images in that set, but if less than half are middens, replace an empty image with one of the midden images in the list until the set of images is balanced
  Append this balanced set of images and labels to the training loader
  Remove the indices corresponding to the original b images from A, U, and L
  Train the model for 2 epochs on the training loader
end while

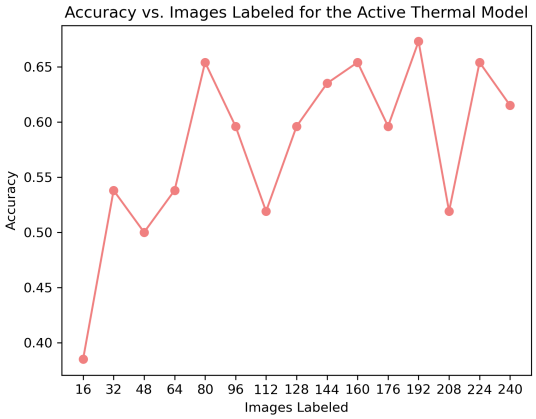

Figure 6: Accuracy for the thermal active learning model with up to 240 images labeled

Conclusion

We have curated a novel dataset for rhino monitoring and presented preliminary results for a deep learning approach to rhino midden classification. We have also implemented an active learning methodology unique to our problem domain. As we do not want to release the locations of where rhinos are likely to be found, we make our code publicly available at https://github.com/colliers95/rhino-midden-detection but keep the dataset private. We also acknowledge that the most important ways to tackle poaching are through ranger patrols and education about the lack of medicinal properties of rhino horn, and we see our system as complementary to the efforts of local experts.

References

Alexander, Q.; Hoskere, V.; Narazaki, Y.; Maxwell, A.; and Spencer, B. 2022. Fusion of thermal and RGB images for automated deep learning based crack detection in civil infrastructure. *AI in Civil Engineering*, 1(3).

Bakalos, N.; Voulodimos, A.; Doulamis, N.; Doulamis, A.; Papasotiriou, K.; and Bimpas, M. 2021. Fusing RGB and Thermal Imagery with Channel State Information for Abnormal Activity Detection Using Multimodal Bidirectional LSTM. In Abie, H.; Ranise, S.; Verderame, L.; Cambiaso, E.; Ugarelli, R.; Giunta, G.; Praça, I.; and Battisti, F., eds., *Cyber-Physical Security for Critical Infrastructures Protection*, 77–86. Cham: Springer International Publishing. ISBN 978-3-030-69781-5.

Bondi, E.; Ressijac, C.; and Boucher, P. 2021. Preliminary Detection of Rhino Middens for Understanding Rhino Behavior. CVPR 2021 Workshop on Computer Vision for Animal Behavior Tracking and Modeling.

Bouguettaya, A.; Zarzour, H.; Taberkit, A. M.; and Kechida, A. 2022. A review on early wildfire detection from unmanned aerial vehicles using deep learning-based computer vision algorithms. *Signal Processing*, 190: 108309.

Haut, J. M.; Paoletti, M. E.; Plaza, J.; Li, J.; and Plaza, A. 2018. Active Learning With Convolutional Neural Networks for Hyperspectral Image Classification Using a New Bayesian Approach. *IEEE Transactions on Geoscience and Remote Sensing*, 56(11): 6440–6461.

Johnson, C. J.; and Gillingham, M. P. 2005. An evaluation of mapped species distribution models used for conservation planning. *Environmental Conservation*, 32(2): 117–128.

Lewis, D. D.; and Gale, W. A. 1994. A Sequential Algorithm for Training Text Classifiers. In Croft, B. W.; and van Rijsbergen, C. J., eds., *SIGIR '94*, 3–12. London: Springer London.

Li, X.; and Guo, Y. 2013. Adaptive Active Learning for Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Linklater, W. L.; Mayer, K.; and Swaisgood, R. R. 2013. Chemical signals of age, sex and identity in black rhinoceros. *Animal Behaviour*, 85(3): 671–677.

Marneweck, C. 2013. What is a Rhino Midden?

Marneweck, C.; Jürgens, A.; and Shrader, A. 2017. Dung odours signal sex, age, territorial and oestrous state in white rhinos. *Proc Biol Sci.*

Roy, N.; and McCallum, A. 2001. Toward Optimal Active Learning through Sampling Estimation of Error Reduction. In *Proc. 18th International Conf. on Machine Learning*, 441–448. Morgan Kaufmann, San Francisco, CA.

SaveTheRhino. 2022. Poaching Stats.

Settles, B. 2009. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Seung, H. S.; Opper, M.; and Sompolinsky, H. 1992. Query by Committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, 287–294. New York, NY, USA: Association for Computing Machinery. ISBN 089791497X.

Speth, S.; Gonçalves, A.; Rigault, B.; Suzuki, S.; Bouazizi, M.; Matsuo, Y.; and Prendinger, H. 2022. Deep learning with RGB and thermal images onboard a drone for monitoring operations. *Journal of Field Robotics*, 39(6): 840–868.

Tuia, D.; Volpi, M.; Copa, L.; Kanevski, M.; and Munoz-Mari, J. 2011. A Survey of Active Learning Algorithms for Supervised Remote Sensing Image Classification. *IEEE Journal of Selected Topics in Signal Processing*, 5(3): 606–617.