This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Large-Scale Ecological Analyses of Animals in the Wild using Computer Vision

Mikayla Timm Subhransu Maji Todd Fuller
University of Massachusetts Amherst
{mtimm,smaji}@cs.umass.edu, tkfuller@eco.umass.edu

Abstract

Camera traps are increasingly being deployed by ecologists and citizen-scientists as a cost-effective way of obtaining large amounts of animal images in the wild. In order to analyze this data, the images are labeled manually by ecologists, where they identify species of animals and more fine-grained details, such as animal sex or age, or even individual animal identities. However, with the number of camera trap images quickly outgrowing the capacity of the labelers, ecologists are unable to keep up with the wealth of data they are obtaining. Using computer vision, we can automatically generate labels for new camera trap images at the rate that they are being obtained, allowing ecologists to uncover ecological and biological information at a scale previously not possible. In this paper, we explore computer vision approaches for species identification in camera trap images and for individual jaguar identification, both of which show promising results. We make this novel dataset publicly available for future research directions and further exploration.

1. Introduction

Ecologists monitor animal population densities and migration patterns in the wild by manually identifying animal species and individual animal identities in camera trap image datasets, which often contain thousands or even millions of images. These images are quite difficult and time-consuming to sift through by hand. Many of the images the ecologists collect are just of moving grass or plants that happened to activate the motion sensors on the camera traps. The images actually containing animals are also challenging to scrutinize for the following reasons: the camera sometimes only catching part of the animal's body, occlusion due to the animal hiding in grass or behind a tree, differences in lighting between day/night, and the fact that many of these cameras vary in resolution. As a result, identification is time-intensive and error-prone, and with more and more ecologists and citizen scientists deploying camera traps as a means of collecting data, the scale of these camera trap datasets is quickly becoming too large for human labelers to handle.

We develop a computer vision pipeline for estimating animal counts in these images. Deep learning has shown impressive results on a number of visual recognition tasks ranging from image classification, semantic segmentation, object detection, to fine-grained recognition [3, 5, 2, 4], but often the performance is limited by the amount of training data available. One of our contributions is a large dataset of camera trap images taken over the period of one year across locations in Costa Rica. This dataset consists of over 12,000 images, each of which has been manually labeled by ecologists with the presence or absence of top 10 most-frequent animal species in the region. We first develop a computer vision pipeline to recognize the presence of different animal species by adapting deep networks trained on massive image datasets such as ImageNet [1] and iNaturalist 2017 [7]. The iNaturalist 2017 dataset mined from iNaturalist.org provides a rich resource of millions of animal and plant images taken in the wild with community verified species labels. These images are of much higher quality (as they are taken by photographers) compared to the camera trap images. Additionally, iNaturalist images are usually focused directly on the animal of interest, usually with the animal nicely centered in the image, with good lighting, and without any occlusions. Adapting deep networks pre-trained on this dataset offers slight improvements over a standard ImageNet pre-trained model on the same task.

We also explore a solution to the individual jaguar identification problem, an important step in estimating population counts and behavior of these animals. The rosettes on jaguars' coats are distinct for each jaguar, similar to a human fingerprint, and are used by ecologists in individual jaguar identification. Our approach finds a set of SIFT feature matches across image pairs using RANSAC and uses the number of consistent matches as an indicator of similarity. Our experiments on a dataset of jaguar images we collected by labeling the camera trap images, as well as collections of photos of jaguars in zoos across the country, show that this approach works remarkably well for identification. The combination of the classifier and the identification step has the potential to significantly reduce the manual effort required to analyze these images.

2. Experiments

2.1. Datasets

Below we describe the two datasets we use for species identification in camera trap images, and individual jaguar identification.

Costa Rica Camera Trap Dataset. The camera trap dataset was obtained from faculty and researchers at UMass Amherst and Universidad Nacional de Costa Rica, taken in Costa Rica over the last couple of years. The full camera trap dataset contains 300,000 images of wildlife taken from motion-activated camera traps in the wild. A portion of this dataset is currently labeled (12,000 images), but only around 2,000 were labeled at the time of training our models in this paper. There are 12 classes, consisting of the 10 most common animals found in the camera trap locations, "no animal", and "other animal". There was a large imbalance in the classes, however, with over half of the labeled images consisting of "no animal". For this reason, we utilize only 250 images of each class during training to help fix the class imbalance problem. The distribution of the species in the current labeled data benchmark is shown in Table 1. An example image of each species is shown in Figure 1. This dataset will be made publicly available at mtimm100.github.io/cameraTrap.html.

JaguarID Dataset. The JaguarID dataset contains 176 images of 16 jaguars taken from camera traps in the wild, camera traps in zoos, and supplemented with some images from Flickr. These images range from low to high quality, some in color and some in grayscale, and varying levels of motion blur.

2.2. Results

Species Identification. To train our species classifier, transfer learning is utilized through fine-tuning two pre-trained CNNs, one trained on ImageNet and another trained on iNaturalist 2017. Both models' architectures are built off of InceptionV3 from Szegedy et al. [6], altering the final layers to match the number of classes for our problem.

The labeled camera trap data was split into a training and validation set (80% training, 20% validation), and we performed data augmentation on each set (random crops, horizontal flips, etc.) during the training process. For fine-tuning, each model was trained for 100 epochs with a batch size of 32 using stochastic gradient descent with Nesterov momentum (value of 0.9 selected for our experiments). Our implementation is based on PyTorch.

The final validation set accuracies after fine-tuning models on the camera trap image data are shown in Figure 2. We see that the iNaturalist pre-trained model provided a better initializations, leading to slightly better performance in the first 20 epochs, but both models converged to roughly the same accuracy. We obtained around 75% accuracy with the training set of size 2,000 images, but we plan incorporate the larger dataset in the next training iteration.

Individual Jaguar Identification. The JaguarID task focuses on identifying individual jaguars in images. As mentioned earlier, jaguars' rosettes (patterns on the skin) are uniquely identifiable, so the ability to find matches of groups of rosettes across two images of jaguars is necessary for this problem. Our method for matching jaguar image pairs first requires detecting corners from gradient information, which are scale, translation, and rotation invariant. This is important, as jaguars can contort their bodies in strange poses. Next, we match the corners robustly using SIFT descriptors, then use RANSAC to find consistent groups of matches (inliers) according to an affine transformation, as illustrated in Figure 3. The number of inliers between two jaguar images provides us with a similarity score. We used a leave-one-out procedure for evaluation. For each query image, we find the closest matching image by computing similarity scores across the remaining dataset and predicting its label. Our approach obtains 91.5% accuracy, where chance accuracy is 6.25%.

3. Discussion and Future Work

For species identification, further data pre-processing should be done to eliminate all of the "no animal" images before doing species classification. We aim to do this by training a CNN to first classify each image as "Animal" or "No Animal", then feed the "Animal" images through the species classifier. This could help remedy the class imbalance problem. Additional improvements to our model could incorporate object detection and pose estimation.

By providing ecologists with tools that use computer vision to accurately and efficiently classify camera trap images, they will be better equipped to deal with the vast amounts of data becoming available. This will enable significant, large-scale ecological analyses to be conducted. We plan to continue this work to improve the classifiers, extend individual ID to other animals, and package everything into a user-friendly tool that can be easily used by ecologists and citizen-scientists for future research.

Special thanks: to our collaborators in the UMass Department of Environmental Conservation and the Instituto Internacional en Conservación y Manejo de Vida Silvestre, Universidad Nacional, Heredia, Costa Rica, for providing us with the Costa Rica Camera Trap dataset; for the permission and cooperation of Santa Rosa National Park, Guanacaste Conservation Area, Costa Rica; and to our collaborators from Lee Richardson Zoo, Palm Beach Zoo, The Living Desert, Turtle Back Zoo, Sacramento Zoo, and San Diego Zoo Global for providing us with the JaguarID images.

None	Collared Peccary	Puma	Ocelot	Jaguar	Human	Tapir	Agouti	Currasow	White-tailed Deer	Crested Guan	Other
7351	282	75	21	149	52	47	65	1465	2036	36	684

Table 1. Distribution of the species in the camera trap dataset.

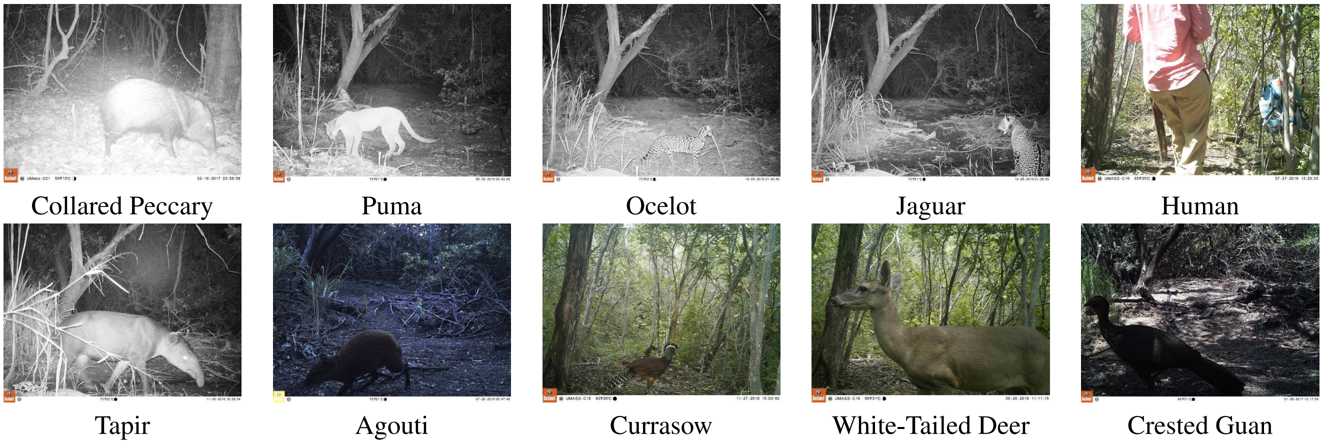
Collared Peccary
Puma
Ocelot
Jaguar
Human
Tapir
Agouti
Currasow
White-Tailed Deer
Crested Guan

Figure 1. Species ID: Example image of each species in the Camera Trap dataset.

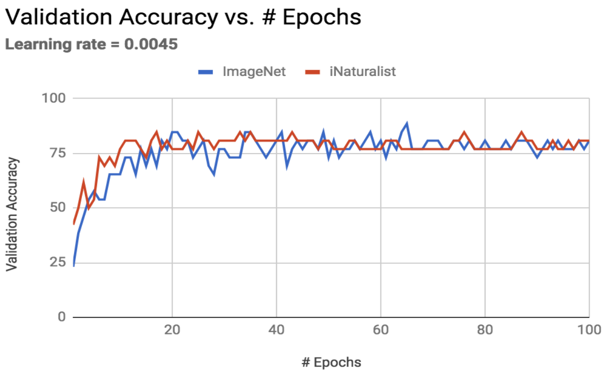

Figure 2. Species ID: Validation accuracies for two networks finetuned on Camera Trap images. Blue: ImageNet-pretrained model. Red: iNaturalist-pretrained model.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, 2009. 1

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE CVPR*, 2014. 1

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[4] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear CNN models for fine-grained visual recognition. In *ICCV*, 2015. 1

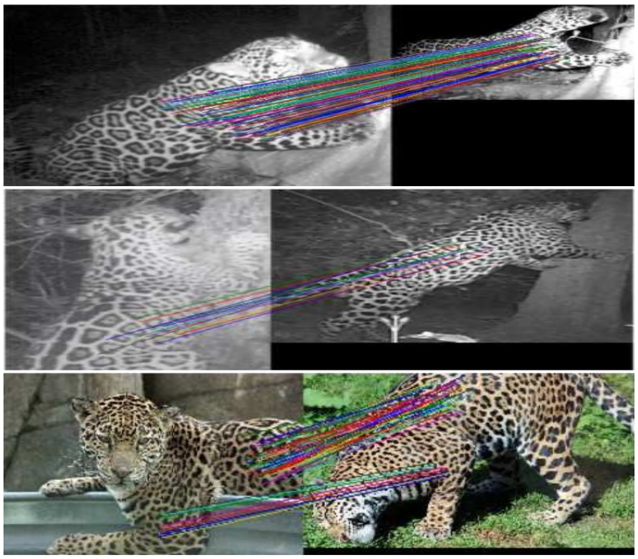

Figure 3. Jaguar ID: Matched SIFT descriptors across images using RANSAC to estimate an affine transformation. Top two rows shows images from the Camera Trap dataset, while the last row shows images taken in a Zoo. The matching works across a wide range of view point and pose changes.

[5] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE CVPR*, 2015. 1

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *IEEE CVPR*, 2016. 2

[7] G. Van Horn, O. Mac Aodha, Y. Song, A. Shepard, H. Adam, P. Perona, and S. Belongie. The inaturalist challenge 2017 dataset. In *IEEE CVPR*, 2018. 1