INTRODUCTION

Computer vision-based methods are being used increasingly as tools to assist with animal object recognition. The ability to identify individual animals from images enables population surveys through sight-seeing identification and forms the basis for demographic studies.

Detection involves the obvious steps of finding animals in images, determining their species, and placing bounding boxes around them, creating what we refer to as an annotation. But the problem is more complex than this, especially when large data volumes gathered by non-specialists are considered: there may be multiple (or no) animals from different species in an image; some annotations may have poor quality while others may show only parts of an animal due to occlusion, or occlusion by other animals or vegetation; and may be seen from a range of viewpoints and poses, only some of which show identifiable information. We propose a five stage detection pipeline (middle, top) to address these challenges and deliver a new dataset to evaluate its performance.

METHODS - IMAGE CLASSIFICATION

The purpose of the image classifier is to predict the existence of species of interest within an image. We structure the classifier to predict a multi-label, multi-class vector where the species flag is set to 1 if at least one animal of that species exists in the image and 0 otherwise. Our implementations are built using Theano and Lasagne.

METHODS - ANNOTATION LOCALIZATION

The annotation localization network design is based on the You Only Look Once (YOLO, version 1) network by [5]. The network’s goal is to perform bounding box regression (left) and species classification along all objects of interest, the result being a collection of sub-regions that can be cropped into a candidate list of object annotations. The predicted bounding boxes by the localization network have species label classifications. Since we are performing annotation classification anyway, we can treat these localizations simply as salient object detections.

METHODS - BACKGROUND SEGMENTATION

The annotation classification network architecture is very similar to the image classification component except that it performs a standard single-label, multi-class classification. We intentionally train a separate set of weights for the convolutional feature extractors in each component. The primary task of annotation classification is to correctly label the annotation’s species and the correct viewpoint together. Poor scoring boxes do not continue in the pipeline.

METHODS - AOI CLASSIFICATION

The novel task for the AoI classification network to solve is to predict an annotation area of interest (AoI). But, the problem is more complex than this, especially when large data volumes gathered by non-specialists are considered: there may be multiple (or no) animals from different species in an image; some annotations may have poor quality while others may show only parts of an animal due to occlusion, or occlusion by other animals or vegetation; and may be seen from a range of viewpoints and poses, only some of which show identifiable information. We propose a five stage detection pipeline (bottom) to address these challenges and develop a new dataset to evaluate its performance.

RESULTS

Our method is able to achieve a localization mAP of 81.67%, a species and viewpoint combination annotation classification accuracy of 94.28% and 87.11%, respectively, and an AoI accuracy of 72.75% across 6 animal species. Others may also introduce the Wildlife Image and Localization Dataset (WILD), which contains 5,784 images and 12,007 labeled annotations across 28 classification species and a variety of challenging, real-world detection scenarios.

The overall accuracy of species and viewpoint combination classifications is 61.71% over 42 distinct categories (right). The accuracy improves from this baseline when we take into account how viewpoint variance impacts identification (i.e. a ±45% degree shift in yaw is acceptable for giraffes and plains), which achieves a “fuzzy” accuracy of 87.11%. Further, AoI accuracy goes up to 93.33% if we treat false-negatives as the true errors of filtering.

WILD DATASET

We created a new ground-truthed dataset called WILD. WILD is comprised of photographs taken by biologists, wildlife rangers, citizen scientists [9], and conservationists, and captures detection scenarios that are uncommon in publicly-available computer vision datasets like PASCAL, ILSVRC, and COCO. In WILD, the data is organized into a list of object annotations that describe the image in the following order: species, species viewpoint, species attributes to mark viewpoints and AoI flags. The output AoIs, for example, could be fed as input data into an appearance-based identification system (confusion matrix to right).

CONCLUSIONS

We evaluated five detection components against WILD, a new dataset of real-world animal sightings that focuses on challenging detection scenarios. The end result of our proposed pipeline is a collection of novel annotations of interest (AOI) with species and viewpoint labels. The output AoIs, for example, could be fed as input data into an appearance-based identification system (confusion matrix to right).

The goal of our method is to increase the reliability and automation of animal census studies and to provide better ecological information to conservationists. Future work can be focused on improvements to WILD, improving the accuracy of AoI classification, and performing a comprehensive identification performance study.

REFERENCES


An Animal Detection Pipeline for Identification

Jason Parham*, Jonathan Crall, Daniel Rubenstein, Jason Holmberg, Tanya Berger-Wolf, and Charles Stewart


Dataset: http://lev.cs.rpi.edu/public/datasets/wild.tar.gz

We do not have fully-grounded truth for a quantitative evaluation of our network's background segmentation performance, as shown in the paper. We do provide examples, however, from the detection pipeline (middle, bottom).