Deep Learning for Accurate Population Counting in Aerial Imagery

Matt Epperson, James Rotenberg, Eric Lo, Sebastian Afshari & Brian Kim

1 Abstract

Understanding rainforest habitats is made difficult by accessibility, density of foliage, and coverage; yet it is vital for ecologist and conservationists to understand the variety and diversity of these ecosystems. In this paper we demonstrate a low cost pipeline for obtaining high resolution aerial imagery using a small unmanned aerial vehicle (UAV). Second we show how advances in deep learning enable processing the imagery to obtain highly accurate population counts of trees that is extendable to multiple species. Specifically we demonstrate the ability to detect 98% of visible Cohune Palm trees and 96% of deciduous trees who had visibly lost their foliage in a 8.73km^2 of rainforest. By using these trees, we identify key habitat characteristics of this forest not realized previously.

2 Introduction

Tropical rainforests worldwide are negatively impacted from a variety of human-caused threats (Wright, 2010) including deforestation and climate change. However, our ability to study these rainforests can be impeded by problems ranging from their physical inaccessibility to cloud-covered remote sensing data (i.e. Landsat imagery). One solution to these problems is the use of Unmanned Aerial Vehicles (UAV). UAVs have been used successfully in conservation (Wich, 2015), and for determining complex forest characteristics such as biomass from Structure from Motion (SfM) imagery (Ota et al., 2015).

Automated tree detection and crown delineations is an active field of study in both the satellite and UAVs domains using passive mediums such as high resolution imagery (Pouliot 2002) and with hyperspectral cameras (Zhang 2012). Recently Light Detection and Ranging (LIDAR) has been growing in popularity for its ability to discriminate in three dimensional space (Zhen 2016). Both hyperspectral and LIDAR are costly and drastically reduce the range of UAV due to heavy weight making them an unattractive choice for many conservationists. On the other hand small flying wing UAVs that are nearly end to end autonomous are increasingly being equipped with high resolution cameras with flight durations upwards of one hour. These platforms present researchers without significant budgets the ability to acquire huge amounts of very high resolution data.

An intelligent method for analyzing this data is needed that is able to handle both the density and diversity of trees present in rainforest canopies. Deep convolutional neural networks (CNN) have been extremely successful in classification tasks with fine-grained classes (Krizhevsky 2012), (Szegedy 2015). For example a deep CNN was shown to make medical diagnosis between malignant and benign skin cancer on par with seasoned dermatologists (Esteva 2016). There has been significant research on adapting CNNs for the challenging task

of object detection and recognition, i.e. the task of both localizing and object within an image as well as classifying it (Redmon 2016), (Liu 2016).

In this paper we present three contributions. First we describe a low cost UAV setup that allowed us to obtain high resolution georegistered orthomosaics with a spatial resolution of 4.93 cm/pixel at our main site, and . Second we describe the process of training a state of the art deep convolutional neural network (CNN) for detecting both Cohune Palm trees as well as Deciduous trees. We present results showing extremely accurate with greater than 96% detection rate and low false positives. We demonstrate the ability of this detector framework to work effectively on other dataset under different lighting conditions and flight parameters. Finally we draw conclusions from this data on the harpy eagle….

3 Convolutional Neural Networks

In 2012 Alex Krizhevsky proposed a learning algorithm that used layers stacked layers of 2D convolution filters to form a deep neural networks are able to learn complex representations of objects. The method is loosely inspired by how the human visual cortex system operates. His work is largely considered a silver bullet that helped computer vision researchers overcome a barrier in classification accuracy that they had been running up against. Since then CNNs have become the state of the art method for computer vision tasks.

CNNs work by convolutioning 2D filters across an image which has the effect of preserving the spatial structure of the objects in the image. Multiple convolution layers as well as non-linear layers such as max pooling and sigmoid activation are combined together in order to form models that can learn complex hierarchical representations of objects. Backpropagation is then used to learn the correct filter weights given sufficient training data which is usually quite large. CNNs have been extended to not only classify the image as a whole, but to localize individual objects within the image which is of particular interest in the paper.

3.1 Deep Learning Framework

Two state of the art networks for object detection and recognition, Faster-RCNN and YOLOv2, were examined for use. Faster-RCNN creates region proposals by using a a region proposal network (RPN) that shares convolutional layers with a deep convolutional network (Ren, 2016). YOLOv2 instead uses a single network that simultaneously predicts location and class (Redmon 2016). Both implementations have comparable results on the standard datasets as well as open source software implementations. Faster-RCNN is implemented using the Caffe Framework; open source GPU accelerated framework for training and testing deep learning algorithms from Berkley (Jia 2014). YOLOv2 uses Darknet a custom GPU accelerated framework specifically created for YOLO and YOLOv2, but with extensions to other deep learning techniques such as recurrent neural networks (Redmon, 2016).

We chose to use the darknet framework and YOLOv2 for its ability to easily train on higher resolution images which is of particular interest in preserving geospatial features from the UAV imagery. The time required to get YOLOv2 working with a custom dataset was also much lower than FASTER-RCNN whose implementation was difficult to work with.

4 Methods

4.1 Study Area

We collected aerial survey data over two adjacent protected areas located in the Maya Mountains in the Toledo District of southern Belize (Figure 1). The first was a 467-hectare (1,153-acre) site administered by the Belize Foundation for Research and Environmental Education (BFREE; 16.5°N, 88.6°W). The second site was the 39,270-hectare (97,039-acre) Bladen Nature Reserve (BNR), a government reserve, under the highest level of protected status in Belize (Forestry Department, Government of Belize). The two sites are part of the 607,028 hectares (1.5 million acres) Maya Mountains protected area system, often considered one of the largest remaining unspoiled, mixed tropical forest ecosystems remaining in Central America (Brewer and Webb 2002, Olivet and Asquith 2004, Dourson 2013). Elevation and precipitation in the Maya Mountains range from 80 to 1000 meters with rainfall averages between 2500 and 3000 millimeters (80-100 inches) of rain per year. The area has distinct wet and dry seasons, of which 89% of the rainfall occurs between May and December (Brokaw and Lloyd-Evans 1987). Both the BFREE and BNR areas have been classified with only coarse-scale, generalized habitat categories including: mainly tall and low tropical evergreen forest (low forest is referred to as “Broken Ridge” in Belize), and a variety of disturbed habitats including early secondary forest, riparian edge, seasonally inundated forests, and cacao agroforest at BFREE; and tall to medium, tropical evergreen forests as well as smaller amounts of mixed low forest with some disturbed riparian edge near the Bladen River for the BNR (S. Brewer, personal communication).

4.2 Flight System

The imagery was collected using a 2 meter wingspan fixed wing airplane, the Voltanex Ranger, shown in figure 1. The Ranger is fitted with an ardupilot autopilot system, which allows for autonomous control over long range flights. This flight system was selected due to its rugged design, payload capacity, and long range. The tough body and 60 km range are requirements for jungle flight operations, where landing areas are rare and rough.

Figure 1: Voltanex Ranger, Last Flight

Data was collected over 3 days with 6 flights total. The paths of all flights are shown in figure 2. All flights began from the same location, where there is ample room for takeoff and landing. All flights were conducted with both visible camera and near infrared cameras, which limited range due to weight but lowered the number of required flights. This was necessary because of the short flight window due to weather.

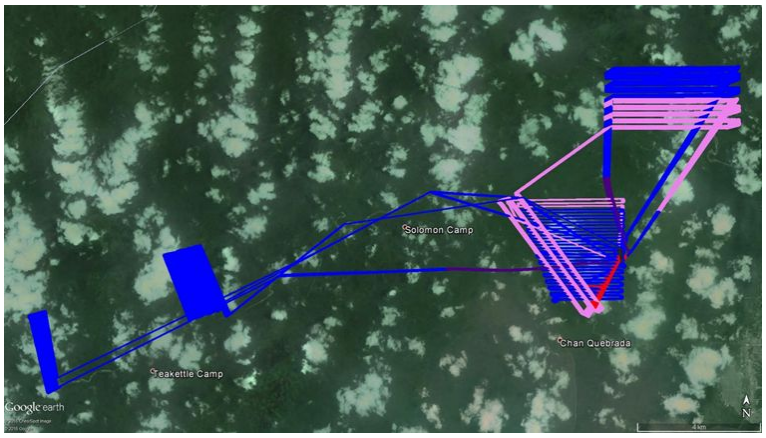

Figure 2: Flight Paths

To access other areas of interest in the region, future flight operations will benefit from upgrades to the system. The areas of the long range surveys were impacted due to range limits. This problem could be fixed by a new plane capable of longer range, or a system capable of takeoffs and landings in smaller areas, which would allow for more potential takeoff and landing locations deeper in the BNR.

4.3 Dataset

The raw data collected from the UAV flights consists of very high resolution images that have been orthorectified and stitched together into a single image using a process called structure from motion (Dandois and Ellis, 2010). We trained our CNN based object detector on pieces of this large orthomosaic which we will refer to as tiles. Each tile represents approximately 100m^2 of rainforest. The tiles were shown to volunteers who were asked to label Cohune Palm trees by the drawing bounding boxes around each visible palm tree. A demo video and images were provided which described how the boxes should be created. An example of labeled palm trees using the web tool is shown in figure 3. The labelers collectively spent 10.5 hours annotating the data resulting in a total of 9330 bounding boxes. This was divided into a training set and testing set with 75% and 25% of the boxes respectively.

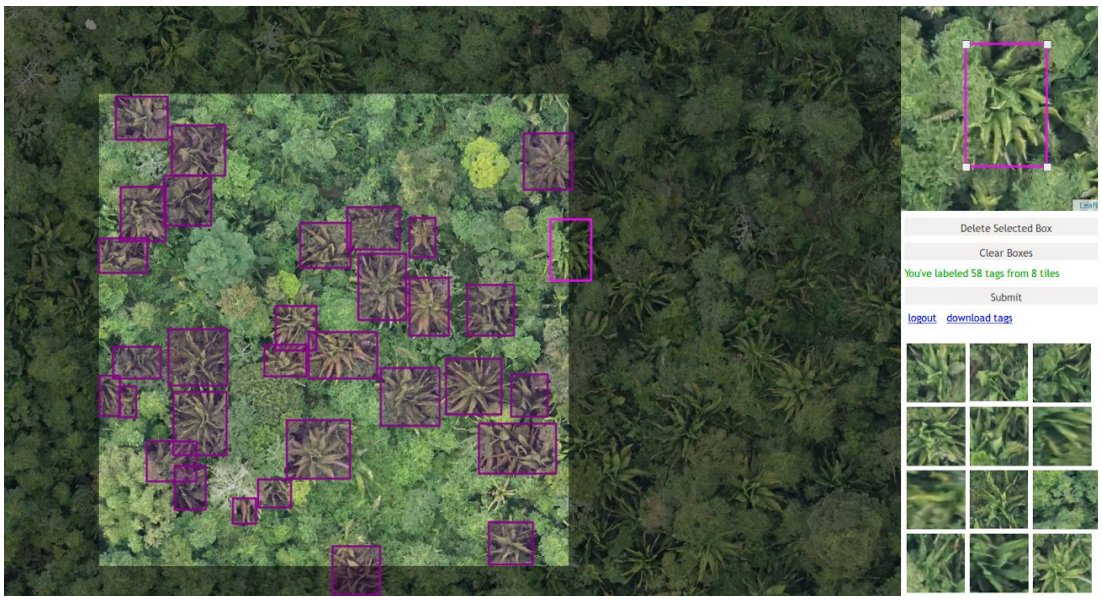

Figure 3: Labeling Web Tool

For the deciduous trees the total number of tree was much fewer and the labeling was accomplished in approximately 6 hours. We found that the interclass variation was much higher with deciduous trees which put a greater strain on the labelers. As a result the labeled data varied much more. For example in figure 4 are two examples of deciduous trees where

Figure 4: Deciduous trees used for Training

Figure 5: BFREE Orthomosaic

4.4 Training and Tuning the Network

Training large convolutional neural networks is a very computational task which is often performed on Graphic Processing Units or GPUs. We trained the networks using a NVIDIA GTX 1080 graphics card using a pre trained model provided on the darknet website. This model was originally trained on the VOC dataset (Everingham, M. 2015). We trained for approximately 10,000 iterations on our dataset which took roughly 36 hours. This was validated by a test set that made up 25% of the total data.

One important tuning parameter is the size of the image. Images typically aren't processed with their original size as its computationally impractical; however, YOLOv2 allows the user to easily select the resized image. We found that since we inherently had very high resolution imagery we could significantly increase performance by increasing the input resolution to the network.

5 Results

Two ways of calculating the performance our network were considered. The simplest approach to calculate the number of trees correctly found would have been to first discard any boxes whose confidence prediction doesn't meet a certain threshold and then compare the center of the remaining boxes with the human boxes. If that center is within the human box we could consider it a success. Now consider the case where the predicted box is very small in

comparison to the ground truth and perhaps located in one of the corners. It would be clear to any human evaluator that this was not a good prediction.

We instead evaluated efficacy by using the intersection over union (IOU); a more holistic representation of how well the predicted boxes matches training data. IOU is the intersection of the darknet created bounding box with the human created bounding box divided by the union of the darknet bounding box and the human bounding box. Since the network produces many bounding boxes we need to associate the generated boxes with the human created bounding boxes. This is done by greedily searching for the darknet box that gives the highest IOU given a human box. If the IOU meets a certain threshold we conclude that the tree was found. Two examples of IOU calculations can be seen in figure 9. If all possible IOUs are zero or below a threshold then we count the human box as a false negative. Otherwise we count the darknet box as having successfully found the tree. Any darknet boxes left over after this process completes are considered false positives. We chose an IOU threshold of .3 as our standard, but report results using several IOU thresholds. While this may seem rather low the idea is to simply match the boxes perfectly, but to detect if there is actually a palm tree at that location. We found empirically from examining the results that the majority of box pairs with an IOU of .3 were true detections. In the two tables below we report results from running darknet on the palm tree and deciduous tree test set while varying the IOU threshold required to consider a tree found.

IOU Threshold	Total	False Negatives	Recall	Average IOU
.300	2984	89	98.20%	59.90%
.325	2984	110	97.66%	59.90%
.350	2984	135	96.64%	59.90%
.375	2984	174	95.32%	59.90%
.400	2984	220	93.76%	59.90%
.450	2984	368	88.74%	59.90%
.500	2984	619	80.22%	59.90%

IOU Threshold	Total	False Negatives	Recall	Average IOU
.300	236	228	97.03%	58.68%
.325	236	226	95.76%	58.68%
.350	236	219	92.80%	58.68%
.375	236	214	90.68%	58.68%
.400	236	207	87.71%	58.68%
.450	236	195	82.63%	58.68%

.500	236	179	75.85%	58.68%

In figures 6, 7, and 8 we show heat maps of both sets of trees. This is done by creating shape files that contains points consisting of the latitude, longitude pair. These shapefiles are then import into a geographic information system called QGIS. QGIS gives us the ability to turn the shapefiles into heatmaps and overlay them on the aerial imagery.

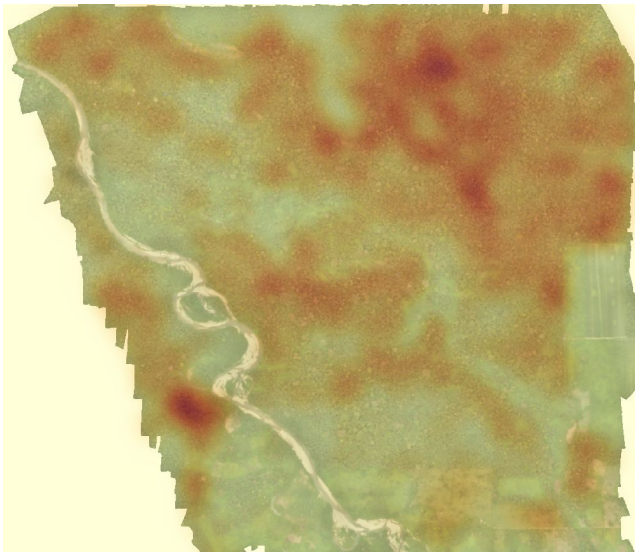

Figure 6: Cohune Palm Tree Heat Map

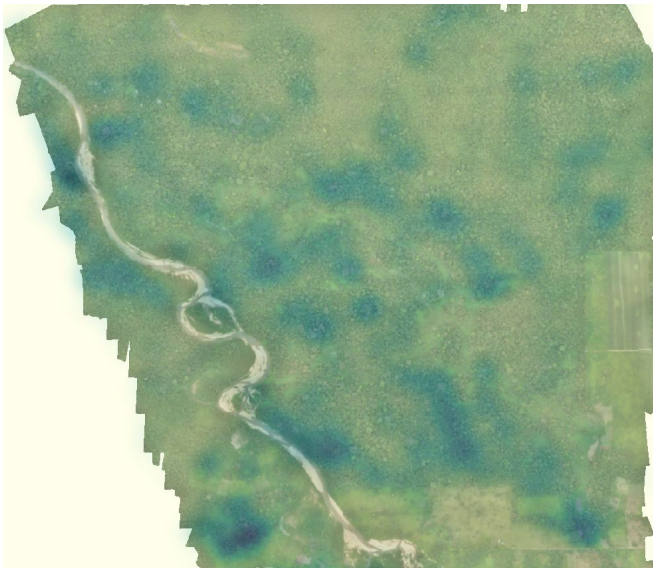

Figure 7: Deciduous Tree Heatmap

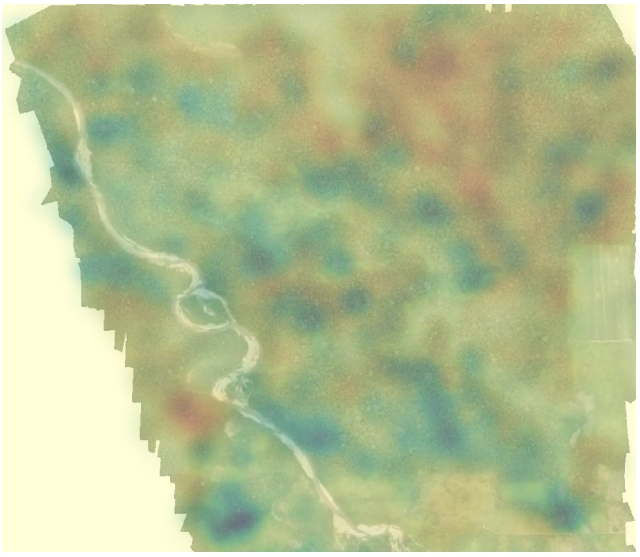

Figure 8: Overlay of both Cohune Palm Trees and Deciduous Trees Heatmap

6 Discussion

One issue that contributed to this threshold was the variation in how humans labeled data. As you can see in figure **9** the human labeler on the right created boxes that did not encompass the entire tree while the labeler on the left did. This variation in ground truth led to a lower than desired IOU results as the network was trained using inconsistent ground truth data.

Figure 9: Example IOU with blue corresponding to the human box and red the darknet created box.

It's also important to note that all labelers were not trained in tree identification. While not necessarily a problem for palm trees which are rather distinct from other tree species it became a problem with deciduous trees where the difference in annotation throughout the dataset variedly largely.

Both of these issues could be resolved by giving labelers better training prior to created the ground truth datasets. We would expect that this along with creating larger datasets would help improve the CNNs performance.

7 Conclusion

In this paper we demonstrated a pipeline for capturing high resolution aerial imagery in high density rainforests using UAVs. We then pull useful population distribution information of Cohune Palm Trees and all Deciduous Trees by applying state of the art convolutional neural networks as an object detection and recognition system. We are able to achieve very good results for recall and competitive IOU scores. This method while applied to finding trees in South American rainforests could easily be applied to other applications and environments. It could even be applied to tracking and counting populations of animals.

8 Literature Cited

Brewer, S. W., & Webb, M. A. (2002). A seasonal evergreen forest in Belize: unusually high tree species richness for northern Central America. *Botanical Journal of the Linnean Society*, *138*(3), 275–296.

Brokaw, N., & Lloyd-Evans, T. (1987). *Report and Recommendations of the Manomet Bird Observatory—Missouri Botanical Garden Investigation of the Upper Bladen Branch Watershed, Maya Mountains, Belize, March 1997*.

Dandois, Jonathan P., and Erle C. Ellis. "Remote sensing of vegetation structure using computer vision." *Remote Sensing* 2.4 (2010): 1157-1176.

Dourson, D. C. (2013). *Biodiversity of the Maya mountains with a focus on the Bladen nature reserve, Belize, Central America*. U.S.: Goatslug Publications.

Esteva, Andre, et al. "Dermatologist-level classification of skin cancer with deep neural networks." *Nature* 542.7639 (2017): 115-118.

Everingham, Mark, et al. "The pascal visual object classes challenge: A retrospective." *International Journal of Computer Vision* 111.1 (2015): 98-136.

Jia, Yangqing, et al. "Caffe: Convolutional architecture for fast feature embedding." *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

Liu, Wei, et al. "SSD: Single shot multibox detector." *European Conference on Computer Vision*. Springer International Publishing, 2016.

Olivet, C. R., & Asquith, N. (2004). Ecosystem Profile: Northern Region of the Mesoamerica Biodiversity Hotspot, Belize, Guatemala, and Mexico. *Conservation International, Mexico and Central American Program. Critical Ecosystems Partnership Fund Report*.

Ota, T., Ogawa, M., Shimizu, K., Kajisa, T., Mizoue, N., Yoshida, S., … Ket, N. (2015). Aboveground Biomass Estimation Using Structure from Motion Approach with Aerial Photographs in a Seasonal Tropical Forest. *Forests*, *6*(11), 3882–3898. https://doi.org/10.3390/f6113882

Pouliot, D. A., et al. "Automated tree crown detection and delineation in high-resolution digital camera imagery of coniferous forest regeneration." *Remote Sensing of Environment* 82.2 (2002): 322-334.

Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

Wich, S. A. (2015). Drones and conservation. *DRONES AND AERIAL OBSERVATION: NEW TECHNOLOGIES FOR PROPERTY RIGHTS, HUMAN RIGHTS, AND GLOBAL DEVELOPMENT A PRIMER*, 63–70. (**Incomplete reference - need to fill in)

Wright, S. J. (2010). The future of tropical forests: Future tropical forests. *Annals of the New York Academy of Sciences*, *1195*(1), 1–27. https://doi.org/10.1111/j.1749-6632.2010.05455.x

Zhen, Zhen, Lindi J. Quackenbush, and Lianjun Zhang. "Trends in automatic individual tree crown detection and delineation—Evolution of LiDAR data." *Remote Sensing* 8.4 (2016): 333.

Zhang, Caiyun, and Fang Qiu. "Mapping individual tree species in an urban forest using airborne lidar data and hyperspectral imagery." *Photogrammetric Engineering & Remote Sensing* 78.10 (2012): 1079-1087.