Small Object Bird Detection in Infrared Drone Videos Using Mask R-CNN Deep Learning

Yasmin M. Kassim¹, Michael E. Byrne², Cristy Burch³, Kevin Mote³, Jason Hardin³, David R. Larsen², Kannappan Palaniappan¹

¹Department of Electrical Engineering and Computer Science, University of Missouri-Columbia, Columbia, MO 65201,USA

² School of Natural Resources, University of Missouri-Columbia, Columbia, MO 65201,USA

³ Texas Parks and Wildlife Department, Austin, TX, 78744, USA

Abstract

A reliable method to estimate population sizes of wild turkeys (Meleagris gallopavo) using unmanned aerial vehicles and thermal video imaging data collected at several field sites in Texas is described. Automating the data processing of airborne survey videos provides a fast and reproducible way to count wild turkeys for wildlife management and conservation. A deep learning semantic segmentation pipeline is developed to detect and count roosting Rio-Grande wild turkeys (M.g. intermedia) which appear as small faint objects in drone-based thermal IR videos. The proposed approach to detect roosting turkeys that appear as small objects, relies on Mask R-CNN, a deep architecture semantic segmentation algorithm. This is followed by a post-processing data association and filtering (DAF) process for counting the number of roosting birds. DAF was used to eliminate false positives like rocks and other small bright objects, which often have noisy detections across temporally adjacent video frames, that can be filtered using appearance association and distance-based gating across time. Transfer learning was used to train the Mask R-CNN network by initializing using ImageNet weights. Drone-based thermal IR videos are extremely challenging due to the complexity of the natural environment including weather effects, occlusion of birds, terrain, trees, complex tree shapes, rocks, water and thermal inversion. The transect videos were collected at night at several times and altitudes to optimize data collection opportunities without disturbing the roosting turkeys. Preliminary performance evaluation using 280 video frames is promising.

Introduction

The ability to obtain accurate population estimates is vital for effective wildlife management, however cryptic behavior and high mobility make direct observation of many wildlife species difficult. Thus, despite the importance of wild turkey (Meleagris gallopavo) as a game species in the U.S. no widely-accepted methodology exists to estimate wild turkey population densities in an unbiased manner over large spatial scales. The development of a method to directly survey wild turkey populations that could be implemented by state agencies responsible for managing populations would provide an important tool to track population changes over time and provide information necessary to manage the resource effectively. Wild turkeys are a diurnal species that roosts in tree canopies at night. Taking advantage of this behavior, unmanned aerial vehicles (UAV's) with thermal imaging capabilities could be a useful tool to survey turkeys over large areas, provided roosting turkeys can be detected at night by on-board cameras.

During winter of 2019, we performed a pilot study to assess



Figure 1. Mask R-CNN for turkey detection in infrared drone videos

the ability of UAV-mounted thermal imaging to detect roosting Rio Grande wild turkeys (M.g. intermedia) at four field sites in north-central Texas. At each site wild turkeys were captured using walk-in traps [1], and one to three turkeys were fitted with UHF/VHF GPS back-pack transmitters [2] programmed to collect GPS locations at regular intervals throughout he night. This allowed researchers in the field to remotely download the GPS location of a roosting wild turkey in near-real time. UAV's were flown over the known location of roosting GPS-tagged turkeys while recording thermal image video. Flights consisted of 200 m transects that intersected the focal turkey's location. We conducted flights each night one hour after sunset, at midnight, and one hour before sunrise. Eight transects were flown at each time at various altitudes and speeds. Turkeys were clearly visible in all flights, regardless of altitude, speed or time of night, and the UAV did not flush birds from roosts.

Qualitatively, turkeys were more clearly discernible at lower altitudes, with the trade-off being that a narrower field of view is surveyed as altitude decreases. The application of UAVs to survey wildlife populations has seen considerable attention in recent years [3–8] however, our work is unique, fully automated,



Figure 2. Manual annotation of bird masks with expert assistance for training shown in red color and then converted to binary masks visualized in different colors to indicate multiple objects (birds). The left part represents zoomed in area from the right part

robust, and oriented towards detecting and counting turkeys from thermal video footage. Although more research is needed, our results suggest that UAV-mounted thermal imaging and our automated process could be useful tools to survey turkeys over large spatial scales. The motivation behind developing our automated approach is the tedious, time consuming and error prone work of manually counting small objects in IR videos, especially for surveys which can cover a wide area and include multiple videos. Developing a computerized algorithm is the optimal solution for many such problems including detection and tracking vehicles in aerial Wide Area Motion Imagery (WAMI) [9–11], as well as detecting and tracking small, thin objects in biomedical applications with very heterogeneous backgrounds [12–17], both of which motivated and guided our approach for counting birds in drone-based thermal IR videos.

Counting small bird objects in drone-based thermal IR videos is extremely challenging. Some of the challenges include small object size of wild turkeys, lack of salient visual features in infrared video, a turkey's body may not always be fully visible due to thermal insulation or occlusion from branches, the bright head of turkeys in the thermal band is similar to small warm background objects on the ground (e.g. stones and rocks), several turkeys may be close to each other, and the complexity of the natural environment. To overcome these difficulties, we developed a computerized approach that adapted Mask R-CNN [18], a stateof-the-art deep learning network for detection and semantic segmentation in computer vision, with a post processing algorithm to filter out false detections which we refer to as Data Association and Filtering algorithm (DAF). Our automated method is characterized by accurate and consistent detections of turkeys in video frames across time with inconsistent detections of other small objects robustly filtered out.

Computerized Bird Counting Using IR Video

Our automated algorithm consists of two part: first, a well known semantic deep learning network called Mask R-CNN, shown in Figure 1, that has been adapted to detect and identify wild turkeys in IR thermal videos. Second, a post processing al-



Figure 3. Convergence during training, sample training patches for Mask *R*-CNN with total number of anchors = 352,935 of which 64 used for training, positive anchors : 19 (shown above), negative anchors : 45, and neutral anchors: 352,871 (not used in the training), the objects encircled by two bounding boxes, the dotted one represents boxes before refinement (ROI proposals) and the solid ones are boxes after refinement (final classification and localization predictions)



Figure 4. Positive and negative anchors for one bird

gorithm DAF to eliminate spurious results that have been misidentified by our first network. Figure 7 describes in a flowchart all the steps of our computerized approach, the gray elements represent the DAF algorithm. The following two sections describes the two parts in detail.

Adapting Mask R-CNN for Wild Turkey Detection

Recently, several powerful networks for identifying objects have been introduced in the literature [19,20,20–23], however, all those networks provide a bounding box around the object that includes some background pixels. Instance segmentation is a detection that delineate the boundaries of the detected objects, besides a regressed bounding box. Mask R-CNN [18] is the state of the art network in instance segmentation. Our first task is to detect and segment wild turkeys. We've adapted Mask R-CNN [18] to produce reliable detection for our thermal videos. The network consists of several modules:

- 1. Backbone: Our adapted model uses Residual Network with 50 layers, ResNet50 [24], as a convolutional neural network (CNN) for feature extraction.
- Region Proposal Network (RPN): this network obtains region proposals predefined by anchors with different scales and ratios using sliding window operation over the last convolutional feature map layer. This process solves the memory and time limitations of the selective search algorithms



Figure 5. Image augmentation during training process for one bird proposal. The top part contains the positive proposal with corresponding mask, the bottom part contains two augmented proposals and their corresponding masks



Figure 6. Confusion matrix before and after DAF, TN is written as NA because we don't use this measure since it is not suitable for precision and recall (most of the image pixels are TN)

in predecessor detection networks. RPN selects the positive proposals using a parameter called Intersection over Union (IoU), see Equation 1, where P_{Bx} is the predicted bounding box and GT_{Bx} is the Ground Truth bounding box. IoU computes area of overlap versus area of intersection between P_{Bx} and GT_{Bx} . If IoU of the proposal compared to the ground truth above a predefined value, it is considered as a positive candidate, and if IoU is less than that value, it is considered as a negative candidate. Parameter settings and predefined values are discussed in the experimental results and discussion section. Figure 3 shows positive anchors for birds for one of the frames in the training process and Figure 4 describes the process of applying positive and negative anchors in different scales and ratios for one bird.

$$IoU = \frac{P_{Bx} \cap GT_{Bx}}{P_{Bx} \cup GT_{Bx}} \tag{1}$$

3. Multi-class classifier, regressor and binary classifier: after producing the bounding boxes with the corresponding scores, Fast R-CNN works as a classifier to decide whether each region of interest (ROI) contains an object or not and a regressor to predict bounding box coordinates. In parallel to that process, a fully connected network is added to achieve pixel level segmentation to generate object masks in pixelslevel commonly known by instance segmentation.

Data Association and Filtering Algorithm (DAF)

DAF filters out false positives by using multi-objects data association across time. The gray elements in Figure 7 represent our DAF process. For every detection, the algorithm will check whether there is a detection in the next video frame within a square generated from the bounding box coordinates provided



Figure 7. Flowchart for the inference stage of our computerized approach, the gray elements represent our DAF algorithm steps

by Mask R-CNN detection. If there is a corresponding detection, the algorithm keeps this detection as there is a large possibility it is a bird, however, if the detection does not exist, the algorithm eliminates it. Our DAF process takes advantage of that fact that roosting wild turkeys do not fly at night because there is minimal change in a bird's location in subsequent frames. Figure 8 describes the process of our filtering process in one sequence of consecutive frames. Blue bounding boxes are the true detections while red bounding boxes are false positive detections. It is obvious that DAF was able to eliminate 6 out of 7 false positives, the top part contains frames sequence before DAF, whereas, the bottom part includes the same sequence after DAF. For clarification, Figure 9 shows the DAF process on only two consecutive frames. DAF eliminates one false positive detection since it has no correspondence in the next frame. The algorithm keeps only the true objects if they have corresponding detections such as the one bounded by the blue square.

Evaluation

We evaluate our bird detection results based on annotated ground-truth, which represent individual wild turkey masks. To



Figure 8. We present here one sequence of six frames (frame 63 - 68). The top row is the sequence after Mask R-CNN detection and before applying DAF algorithm, the bottom row is after DAF. Our expert recognizes the detections bounded by the blue rectangle as a turkey, detections bounded in red as false positives. It is clear that the top sequence has 7 false positive detections, while the bottom sequence after applying DAF has only one false positive with consistent detection for the bird across all frames.



Table 1: Evaluation of Mask R-CNN detection before and after DAF using three evaluation methods: recall, precision and F1 measure

	Mask R-CNN	Mask R-CNN after DAF
Recall	97.27 %	94.00%
Precision	74.58 %	86.55 %
F1	84.43 %	90.12 %

It is also called sensitivity and is given by

$$Recall = \frac{TP}{TP + FN},\tag{2}$$

Precision is an evaluation metric that indicates how well our algorithm eliminates false detections.

$$Precision = \frac{TP}{TP + FP},\tag{3}$$

F1 combines those two measures to provide a better estimation of overall performance.

$$F1 = 2 \times \frac{Recall \times Precision}{Recall + Precision},$$
(4)

Implementation Details and Data Augmentation

We've applied our computerized approach to one of the IR videos that has been sampled to 280 total frames. The dimension of each frame is 480×720 . The expert annotated 37 frames manually to train the deep learning detection network Mask R-CNN. We've utilized the Firefly annotation tool [25]¹ to annotate birds by polygons, followed by creating binary masks, Figure 2 shows the process of annotation with corresponding masks, 25 frames out of 37 are used for training, the rest (12 frames) are used for validation. To train and adapt Mask R-CNN for birds segmentation, several parameters have been set such as: detection max instances= 100, learning momentum=0.9, learning rate=0.001, num of classes= 2 (either bird or background), num of epochs=10,

Figure 9. Eliminating spurious identification of our computerized approach using our post processing process DAF. One false positive detection has been eliminated since it has no correspondence in the following frame. White squares represent the area in which DAF searches inside for a corresponding detection. Blue square represents the true positive that has been kept or identified as birds after DAF

evaluate our results, we compute True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN):

- TP: means a bird is present in GT and detected by the automated algorithm (true detection).
- FP: means a bird not presents in a GT is detected by the automated algorithm (false detection).
- FN: means a bird is presents in a GT and not detected by the automated algorithm.
- TN: means a bird is not presents in a GT and is not detected by the automated algorithm (we don't have number for TN here since it is not a valid case)

Our evaluation considers the standard performance metrics Recall, Precision, and F1 measure for evaluating our detection results and comparing them to the expert gold standard. Recall is a statistical measure used to assess how well the birds are detected.

¹http://www.firefly.cs.missouri.edu.



Figure 10. Three examples that show three cases for our automatic detection. For all examples, the first column are the raw images, that second column are the ground truth, and the last column are the raw images superimposed by our detections. The first example in the first row has perfect detection for all birds. The second row has only one repetitive detection over one bird that has been solved by taking the non-maximum suppression for the two detections. The third row has three detections, however two of them are examples of false positives that usually occur because of the similarity in terms of size and thermal appearance between birds and other objects such as stones. Those two false positive objects are eliminated by our DAF algorithm, figure 8 provides more details about the DAF process for the whole sequence.

steps per epochs=109, 2000 valid proposals [('BG', 1626), (Bird', 374)], grid of anchors covers the full image across different scales [8, 16, 32, 64, 128] with positive anchors use IoU > 0.7, negative anchors have IoU < 0.3, neutral are excluded from the training and finally the network has been trained with transfer learning from ImageNet weights.

The data has been augmented using horizontal and vertical flipping, affine transformation, brightness change of the original value, and blurring. Two methods are randomly selected to augment each positive proposal. See Figure 5 for two augmented images that are taken during training process.

Experimental Results and Discussion

After we utilized 37 frames for training and validation, our mask R-CNN learned model is ready for testing the rest frames. 243 frames have been tested using the trained model for bird detection, followed by DAF for filtering out false predictions. Table 1 shows the evaluation results for three standard metrics Recall,

Precision, and F1 measure. We note DAF succeeds in raising precision by 12% with only 3% loss in recall. This is also clear in Figure 6 that shows the confusion matrix for TP, FP, FN, and TN, it is noticeable that FP has been hugely decreased after DAF with small decrease in TP. F1 measure is equal to 90% which means that our computerized approach with DAF algorithm almost successfully discriminates between FG (birds) and BG. Figure 10 shows three examples in three rows, the first example (first row) represents perfect detection, the second example (second row) includes only one false positive detection that has been eliminated by non maximum suppression since it is a repetitive detection, and finally the third example (third row) provides an example about filtering out two false predictions, bounded by two yellow ovals, using our DAF algorithm.

Conclusions

In this work, we've developed a computerized approach to identify and count roosting wild turkeys in thermal video footage. The automated algorithm consists of two parts: Mask R-CNN for bird detection and DAF algorithm to filter out false predictions. The performance is very promising, and gives us a strong motivation to handle and work with other videos covering different areas with different altitudes. For future work, we will focus on evaluating dependence on altitude, speed, time of night, etc. on count accuracy. We want total count per unit area vs number of birds per frame in video which is the input to geospatial analysis.

Acknowledgments

This work was partially supported by awards from U.S. Army Research Laboratory W911NF-18-2-0285 and Army Research Office DURIP W911NF1910181 and the Texas Parks and Wildlife Department (TPWD Contract No. 515973). We thank TPWD personnel who assisted in capturing wild turkeys and collecting field data, as well as landowners that provided access to their ranches for field work. Dr. Byrne is supported inpart by the USDA National Institute of Food and Agriculture, McIntire Stennis project 1015915. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the U.S. Government or agency thereof. YMK was partially supported by an HCED Government of Iraq doctoral scholarship.

References

- B. D. Davis, "A funnel trap for rio grande turkey," in *Conference of the Southeastern Association of Game and Fish Agencies*, 1994, p. 109–116.
- [2] J.D. Guthrie, M. E. Byrne, J.B. Hardin, C.O. Kochanny, K.L. Skow, R.T. Snelgrove, M. J. Butler, M. J. Peterson, M. J. Chamberlain, and B.A. Collier, "Evaluation of a global positioning system backpack transmitter for wild turkey research," *Journal of Wildlife Management*, pp. 539–547, 2011.
- [3] A.C. Seymour, J. Dale, M. Hammill, P.N. Halpin, and D.W. Johnston, "Automated detection and enumeration of marine wildlife using unmanned aircraft systems (uas) and thermal imagery," *Scientific Reports*, vol. 7, pp. 45127, 2017.
- [4] E. Corcoran, S. Denman, J. Hanger, B. Wilson, and G. Hamilton, "Automated detection of koalas using low-level aerial surveillance and machine learning," *Scientific Reports*, vol. 9, no. 1, pp. 3208, 2019.
- [5] J.C. Hodgson, S.M. Baylis, R. Mott, A. Herrod, and R.H. Clarke, "Precision wildlife monitoring using unmanned aerial vehicles," *Scientific Reports*, 2016.
- [6] J.C. Hodgson, R. Mott, S.M. Baylis, T.T. Pham, S. Wotherspoon, A.D. Kilpatrick, R.R. Segaran, I. Reid, A. Terauds, and L.P. Koh, "Drones count wildlife more accurately and precisely than humans," *Methods in Ecology and Evolution*, pp. 1160–1167, 2018.
- [7] C. Louis-Philippe, J. Théau, and P. Ménard, "Visible and thermal infrared remote sensing for the detection of white-tailed deer using and unmanned aerial system," *Wildlife Society Bulletin*, pp. 181– 191, 2016.
- [8] J.A. Barasona, M. Mulero-Pázmány, P. Acevedo, J.J. Negro, M.J. Torres, C. Gortázar, and J. Vicente, "Unmanned aircraft systems for studying spatial abundance of ungulates: relevance to spatial epidemiology," *Plos One*, 2014.
- [9] N. Al-Shakarji, F. Bunyak, H. Aliakbarpour, G. Seetharaman, and K. Palaniappan, "Multi-cue vehicle detection for semantic video compression in georegistered aerial videos," in *Computer Vision*

and Pattern Recognition Workshops, 2019, pp. 56-65.

- [10] K. Palaniappan, R. Rao, and G. Seetharaman, "Wide-area persistent airborne video: Architecture and challenges," in *Distributed Video Sensor Networks: Research Challenges and Future Directions*, pp. 349–371. 2011.
- [11] R. Pelapur, S. Candemir, F. Bunyak, M. Poostchi, G. Seetharaman, and K. Palaniappan, "Persistent target tracking using likelihood fusion in wide-area and full motion video sequences," in *International Conference on Information Fusion*, 2012, pp. 2420–2427.
- [12] Y. M. Kassim, O. V. Glinskii, V. V. Glinsky, V. H. Huxley, G. Guidoboni, and K. Palaniappan, "Deep u-net regression and hand-crafted feature fusion for accurate blood vessel segmentation," in *Int. Conf. on Image Processing*, 2019, pp. 1445–1449.
- [13] Y. M. Kassim, N. Al-Shakarji, E. Asante, A. Chandrasekhar, and K. Palaniappan, "Dissecting branchiomotor neuron circuits in zebrafish - toward high-throughput automated analysis of jaw movements," in *International Symposium on Biomedical Imaging*, 2018.
- [14] Y. M. Kassim, R. J. Maude, and K. Palaniappan, "Sensitivity of cross-trained deep cnns for retinal vessel extraction," in *Int. Conf.* on Engineering in Medicine and Biology Society, 2018, pp. 2736– 2739.
- [15] Y. M. Kassim and K. Palaniappan, "Extracting retinal vascular networks using deep learning architecture," in *Int. Conf. on Bioinformatics and Biomedicine*, 2017.
- [16] Y. M. Kassim, V. B. S. Prasath, O. V. Glinskii, V. V. Glinsky, V. H. Huxley, and K. Palaniappan, "Microvasculature segmentation of arterioles using deep cnn," in *Int. Conf. on Image Processing*, 2017.
- [17] F. Bunyak, K. Palaniappan, S. K. Nath, T. I. Baskin, and G. Dong, "Quantitative cell motility for in vitro wound healing using level set-based active contour tracking," in *International Symposium on Biomedical Imaging*, pp. 1040–1043. 2006.
- [18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *International Conference on Computer Vision*, 2017, pp. 2980– 2988.
- [19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in Advances in Neural Information Processing Systems, 2015, pp. 91–99.
- [20] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition*, 2014, pp. 580–587.
- [21] R. Girshick, "Fast R-CNN," in International Conference on Computer Vision, 2015, pp. 1440–1448.
- [22] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Computer Vision and Pattern Recognition*, 2017, pp. 7263–7271.
- [23] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," preprint arXiv:1804.02767, 2018.
- [24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [25] U. Sampathkumar, V. B. S. Prasath, S. Meena, and K. Palaniappan, "Cloud-based interactive image segmentation using the FireFly visualization tool," in *IEEE Int. Conf. on Applied Imagery Pattern Recognition*, Oct 2016.

Author Biography

Yasmin M. Kassim is a PhD candidate in the Electrical and Computer Engineering department at University of Missouri-Columbia. She received her BS and MS in Computer Engineering from the University of Technology, Baghdad, Iraq (2006). She is a lecturer in the Computer Engineering department at Al-Nahrain university, Iraq. She will graduates in May 2020. Her work has focused on applying machine learning algorithms and image processing analysis for biomedical and aerial imagery. She will join Lister Hill National Center for Biomedical Communications (LHNCBC) which is a research and development division of the National Library of Medicine in NIH beginning from June 2020.

Dr. Michael E. Byrne is an assistant professor of wildlife ecology in the School of Natural Resources at the University of Missouri-Columbia. His focus is primarily on animal movement ecology, habitat use and selection, and population ecology, with the aim of answering basic questions in ecology as well as informing conservation and species management. His research is primarily based on empirical field studies of free-ranging vertebrates in terrestrial, marine and aquatic ecosystems, and often involves the use of telemetry to collect data on the movements and behaviors of individuals. He has conducted research on a wide-range of species, from shortfin mako sharks to gray wolves in the Chernobyl Exclusion Zone, and has been actively engaged in wild turkey research since 2008.

Kannappan Palaniappan is a Professor of Electrical Engi-Dr. neering and Computer Science at the Univ. of Missouri-Columbia. He directs the Center for Computational Imaging and VisAnalysis (CIVA) focusing on research at the synergistic intersection of image and video analysis, deep learning, computer vision, high performance computing, artificial intelligence and machine learning to understand, quantify and model physical processes with applications to biomedical, space and defense imaging. His research has been funded by the DoD, NIH, NASA, NSF and others. His current interests include video object tracking, 3D computer vision, multicore image processing, neuroimaging and microscopy image analysis. He has received several notable awards including the NASA Public Service Medal for pioneering contributions to petabyte-sized scientific visualization, the ASEE Air Force Summer Faculty Fellowships, the Boeing Welliver Summer Faculty Fellowship, and the National Academy of Sciences Jefferson Science Fellowship. His team has been recognized with best paper awards at the IEEE Applied Imagery Pattern Recognition Workshop (2018), CVPR Automatic Traffic Surveillance Workshop (2016), finalist for a best student paper award at the IEEE Engineering in Medicine and Biology Society conference (2016) and winner CVPR Change Detection Challenge Workshop (2014). He is an Associate Editor for IEEE Transactions on Image Processing the flagship journal in the field.



JOIN US AT THE NEXT EI!

IS&T International Symposium on Electronic Imaging SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!







- SHORT COURSES EXHIBITS DEMONSTRATION SESSION PLENARY TALKS •
- INTERACTIVE PAPER SESSION SPECIAL EVENTS TECHNICAL SESSIONS •



www.electronicimaging.org