Towards Automatic Detection of Animals in Camera-Trap Images

Alexander Loos
Metadata, Audio-visual Systems
Fraunhofer IDMT
98693 Ilmenau, Germany
alexander.loos@idmt.fraunhofer.de

Christian Weigel
Metadata, Audio-visual Systems
Fraunhofer IDMT
98693 Ilmenau, Germany
christian.weigel@idmt.fraunhofer.de

Mona Koehler
Metadata, Audio-visual Systems
Fraunhofer IDMT
98693 Ilmenau, Germany
mona.koehler@idmt.fraunhofer.de

***Abstract*—In recent years the world's biodiversity is declining on an unprecedented scale. Many species are endangered and remaining populations need to be protected. To overcome this agitating issue, biologist started to use remote camera devices for wildlife monitoring and estimation of remaining population sizes. Unfortunately, the huge amount of data makes the necessary manual analysis extremely tedious and highly cost intensive. In this paper we re-train and apply two state-of-the-art deep-learning based object detectors to localize and classify Serengeti animals in camera-trap images. Furthermore, we thoroughly evaluate both algorithms on a self-established dataset and show that the combination of the results of both detectors can enhance overall mean average precision. In contrast to previous work our approach is not only capable of classifying the main species in images but can also detect them and therefore count the number of individuals which is in fact an important information for biologists, ecologists, and wildlife epidemiologists.**

I. Introduction

Due to the ongoing biodiversity crisis, many species are on the brink of extinction. The current biodiversity crisis is observed all over the planet [1]. Those agitating facts demonstrate the urgent need to intensify close surveillance of threatened species in order to protect the remaining populations. However, effectively protecting animals requires good knowledge of existing populations and fluctuations of population sizes over time. Unfortunately, it is a labor intensive task to estimate population sizes in the wild. Nowadays, noninvasive monitoring techniques which are based on automatic camera traps are frequently being used and the number of published studies that utilize autonomous recording devices is tremendously increasing [2]. Consequently, there is a high demand for automated algorithms which are able to assist biologists in their effort to analyze remotely gathered image and video recordings.

An interesting attempt to solve the problem of annotating huge amounts of camera trap footage is the citizen science project Snapshot Serengeti [3] where thousands of volunteers from the general public annotate millions of images. However, also citizen science projects like Snapshot Serengeti could benefit from the recent progress in the field of computer vision making vast amount of valuable information easily available for biologists and conservation researchers. For instance it takes 2-3 months for the citizen science community of Snapshot Serengeti to classify each 6-month batch of images. As digital cameras become better and cheaper, more and more conservation projects will use remote recording devices generating an amount of data which cannot be handled even with the huge volunteer community of projects like Snapshot Serengeti. Moreover, automatic analysis can help to increase user engagement in manual annotation intending to extend the length of a session a user keeps labeling photographs. As shown in [4] the number of blank images presented to a user influences the session length. Further experiments which also include the species and the number of individuals on an image have been considered by Zooniverse, the citizen science web platform which contains Snapshot Serengeti. Therefore the automatic detection of the number and the type of animals - even if not fully correct - can support the choice of pictures presented to users and thereby increase session length. Unfortunately, detection and recognition of animals in camera-trap images taken in uncontrolled environments place high demands on automatic procedures due to occlusion, poor illumination, and complex animal poses. Images showing the whole animal body under ideal conditions are rare. Instead, in most cases only parts of the animal body are visible, images are often over- or underexposed and animals are frequently too far away or to close to the camera. In this paper we will make a contribution to automize the process of tedious image annotation by re-training, applying, and combining two deep-learning based object detectors: YOLOv2 [5] and SSD [6]. We will show that state-of-the-art object detectors are capable of automatically localizing and classifying different species in images taken in uncontrolled environments. For evaluation we utilize the publicly available Snapshot Serengeti dataset [3] which to date consists of 1.9 million capture events containing 3.2 million images of 48 species annotated by voluntary members of the general public. We will show that both detectors have different drawbacks and the combination of YOLOv2 and SSD improves the overall detection performance. Although other attempts have been made to automatically classify images of the Snapshot Serengeti database our approach covers cases where previous methods fail. For instance the works of [7], [8] only concentrate on estimating the presence/absence of animal species in images, ignoring the fact that the number of animals in an image is of huge interest for biologist. Furthermore, their approaches cannot handle images were more than a single animal species is present. Therefore, our approach can be seen as complement to the works of [7], [8].

In summary, our contributions are as follows:

1) For the first time we re-train state-of-the-art deep-learning based object detectors to automatically detect and classify different animal species in the Snapshot Serengeti database.
2) We thoroughly evaluate and compare the results of two object detectors: YOLOv2 [5] and SSD [6] and consequently
3) we show that both object detectors tend to make different kinds of errors and thus can be successfully combined to improve detection performance.
4) We not only evaluate the proposed system using standard evaluation metrics for object detection but we also use the output of the system to count the number of individual animals and compare our automatic approach to human annotations.

The rest of the paper is organized as follows: In Section II we first give an introduction to the Snapshot Serengeti project and then briefly review the state of the art in object detection, particularly when applied to the field of visual animal biometrics. The dataset we used in our experiments as well as the experimental setup is described in Section III. Also, the performance of both detectors as well as their combination is discussed for animal detection and animal counting. Section IV concludes and summarizes the paper and future directions of research are indicated.

ISBN 978-90-827970-1-5

II. Background and Related Work

The emerging and highly interdisciplinary field of Animal Biometrics aims at developing and applying approaches to automatically detect, represent, and interpret phenotypic appearance of various animal species [9]. Such algorithms can be used to detect animals in audiovisual footage, recognize particular behaviors, classify different species, or even identify individuals.

A variety of different approaches for animal detection [10], [11], species classification [7], [8], [12] or even identification [13], [14] of individuals in wildlife footage were presented in the recent past. The closest works to ours are [7] and [8] which harness deep learning to classify animals in camera trapped footage. Both approaches apply very deep convolutional neural networks (CNNs) to automatically identify animal species in the publicly available Snapshot Serengeti dataset [3]. In [7], Gomez et al. compare the performance of several CNN architectures when applied to the special task of species classification. Especially, it was studied how different deep architectures can cope with the main challenges of image footage gathered in the wild: unbalanced samples, empty frames, incomplete animal images, and objects too far from focal distance. Norouzzadeh et al. [8] on the other hand follow a two-step approach: First, they trained a CNN to distinguish images that contain animals from images that do not. They then trained a different network to actually classify the species.

Two main drawbacks of both approaches can be identified which we will address in this paper. First, both methods only classify the main species present in the images. However, none of these approaches is going down to the individual level and count the number of animals. Actually, this is an important information for biologists when it comes to estimating occurrence or visitation rates of certain locations. Thus, volunteers of the Serengeti Project not only have to annotate the presence or absence of a certain animal species but they also have to annotate the number of individuals present in the images. Secondly, both works do not cover the case where more then one species is present in the images. We will address both drawbacks by re-training deep learning based object detectors which have the potential to overcome above mentioned limitations. To the best of our knowledge this is the first published case study which analyzes the performance of deep learning based object detectors on the Snapshot Serengeti dataset.

III. Experiments and Results

A. Description of Datasets

All images and metadata we used in our experiments were retrieved from the Snapshot Serengeti project data of season 1-8. The whole set contains about 1.9 million subjects (triggered camera trap shots of one or three pictures). We describe selected subsets and additional annotations in the following sections.

Animal Detection: In order to train and evaluate object detectors one usually needs strongly annotated data, i.e. tight bounding regions around each object instance. We therefore created a Zooniverse project [15] where 3234 volunteers drew rectangles and classified the animal's level of occlusion. Since this was the first experiment of this kind of labeling and because of the huge effort it requires, we restricted ourselves to the following six most common species in the Snapshot Serengeti dataset: elephant, grant gazelle, thomson's gazelle, giraffe, ostrich, and zebra. Overall 108.298 classifications had been provided for 10.000 images. Since multiple volunteers provided metadata for a single image we needed to cleanup the annotations, in particular the bounding box coordinates. We did a non-maximum suppression for the rectangle regions and used the most common agreement among users for metadata. Yet, in order to ensure a clean final dataset we manually checked each image again and removed wrong bounding regions. Furthermore, in addition to these images, we annotated data ourselves using a custom annotation tool. Figure 1 shows example images for all six species in the dataset. Poor illumination, occlusion, complex body poses, animals too close or too far away from the camera, and unbalanced data place high demands on automatic object detectors. Moreover, we included two species which are hard even for humans to distinguish: Thomson's gazelles and Grant gazelles. In summary we gathered 17585 images containing 33437 annotated boxes. We randomly split the dataset such that approximately 80% of all images were used for training and 20% were held out for testing and evaluation. However, it has to be taken into account that camera traps often shoot a short sequence of images when triggered (usually three pictures). In order to warrant a fair evaluation we ensured that near duplicates, i.e. images taken at a single event, are not split among training and test set. Additionally, we randomly hold out 2000 images from the training set for validation. We plan to publish our dataset in order to stimulate further research of this important topic.

For camera-trap imagery it is almost guaranteed to have drastic data balancing problems. This issue can be a challenge for automatic classification procedures since an unbalanced dataset might draw machine learning algorithms towards the classes with the most samples. However, we chose to keep this inequality in our dataset since it reflects the distribution in real-world conditions.

Animal Counting: The annotators of the Snapshot Serengeti project are not only asked to classify the species but also to count the number of animals in each image. Up to ten animals are explicitly annotated. For more then ten individuals within a single image just two ranges are given: From 11 to 50 and 50+. In addition to localizing animals we want to use the detector results to count the number of animals and compare the results to human annotations. As shown in [4], this information can help to pre-filter images before presented to the annotators in order to increase session length and annotator engagement. It can also be used as initial annotation result and thereby reduce the number of human annotations until consensus for a subject is reached. In our experiments we concentrated on images with up to 5 animals for two reasons: First, for the six animal species of interest, this covers over 86% of the total pictures in the database. Secondly, for some species it is hard to get images with more than five animals in the database, e.g. for ostriches or giraffes. We applied the task of animal counting to two sets, a balanced and an unbalanced dataset. We crawled the MongoDB database of the Snapshot Serengeti project, searching for images containing one of the six animal species of interest. We restricted ourselves to images that had been labeled as finished with consensus and where at least ten people agreed on their decision. For the balanced dataset we chose pictures where the number of images are nearly equally distributed over the number of animals in a picture. The balanced dataset consists of 15660 images. Note, however, that for some species it was hard to find images that contain lots of animals (e.g. ostrich) while for other species (e.g. zebra) most pictures contain flocks of animals.

Species		**Number of individuals in image balanced (b)/ unbalanced (u)**				
		1	**2**	**3**	**4**	**5**
Elephant	b	146	146	146	146	146
	u	3607	1138	541	276	146
Grant gazelle	b	130	130	130	74	46
	u	994	239	130	74	46
Thomson's gazelle	b	1120	1120	1120	1120	1120
	u	8962	4444	2588	1677	1120
Giraffe	b	142	142	142	42	17
	u	4310	533	142	42	17
Ostrich	b	121	5	5	1	0
	u	121	5	5	1	0
Zebra	b	1473	1589	1589	1749	1803
	u	13655	7128	4611	2811	1803

TABLE I: Number of pictures in the balanced and unbalanced dataset.

We also created an unbalanced dataset which better reflects the likelihoods of the number of animals per species per image as captured over 8 seasons of the Snapshot Serengeti project. In total the entire unbalanced dataset contains 61166 images. Details for both datasets can be found in Table I.

ISBN 978-90-827970-1-5 © EURASIP 2018

(a) Elephant (b) Grant gazelle (c) Thomson's gazelle (d) Giraffe (e) Ostrich (f) Zebra

Fig. 1: Example images of our dataset consisting of six common serengeti species.

In both, the balanced and the unbalanced dataset, we kept the distribution among the 6 species according to their distribution in the whole Zooniverse database. We also ensured that none of those images had been used for training.
We furthermore created a set of 300 empty images to evaluate system's capability to distinguish empty images from images which contain animals.

B. Experimental Setup

In order to detect and classify different animal species in real-world footage, we adapt, re-train, and combine two of the best performing state-of-the art object detectors: YOLOv2 [5] and SSD [6].
For YOLOv2 we employ the freely available Tensorflow implementation of [16]. To prevent overfitting, we did not train from scratch but fine-tuned our network from an existing checkpoint which was trained on the Microsoft Common Objects in Context (MS COCO) dataset [17]. The output layer of YOLOv2 was modified to deal with our six classes instead of the 80 object categories of the MS COCO dataset. The finetuning process was done running backpropagation using the RMSprop optimizer [18] with its default parameters. We used a learning rate of $1e{-}5$, a batch size of 16 and trained for 70 epochs. Apart from that we left all other hyperparameters as recommended by the authors. As input dimension we chose 608×608 since according to [5] this configuration achieved the best results on the MS COCO dataset at a reasonable processing time.
For SSD we used the Tensorflow re-implementation of the original Caffe code by [19]. In particular we utilized the VGG-based CNN network architecture [20] of SSD with an input dimension of 300×300. Again, we first modfied the output layer of SSD to handle the six species in our dataset and then finetuned from an existing checkpoint provided by the authors. The original SSD object detection architecture was trained on both the PASCAL VOC 2007 and 2012 training set [21] as well as the MS COCO dataset [17]. As recommended by [19], we used ADAM [22] as optimizer. We started with a learning rate of 0.001 and applied an exponential learning rate decay factor of 0.94. Additionally, we applied a weight decay factor of 0.0005. We trained with a batch size of 32 for approximately 500 epochs.
In order to prevent both object detectors from overfitting we applied various data augmentation techniques. After cropping the annotated ground truth region we horizontally flipped the input image and randomly distorted brightness, saturation, hue and contrast of each channel. To address the fact that often camera traps are not perfectly horizontal, we additionally applied random rotations of $\pm10^{\circ}$. Furthermore, to enhance the systems robustness against truncation of animals we randomly cropped each ground-truth bounding box such that at least $\frac{2}{3}$ of the original image region is preserved.
In order to set the detection threshold of both object detectors we first trained both networks using the paradigm explained above. We then applied both detectors on a validation set consisting of 2000 images which were neither used for testing nor training. We then analyzed the resulting ROC-curves by varying the detection threshold of YOLOv2 as well as SSD and finally picked the thresholds which maximized the F_1-score of the respective detector. By following this guideline we obtained a detection threshold of 0.391 and 0.269 for YOLOv2 and SSD, respectively. We kept these hyperparameters constant for all experiments conducted in this paper.

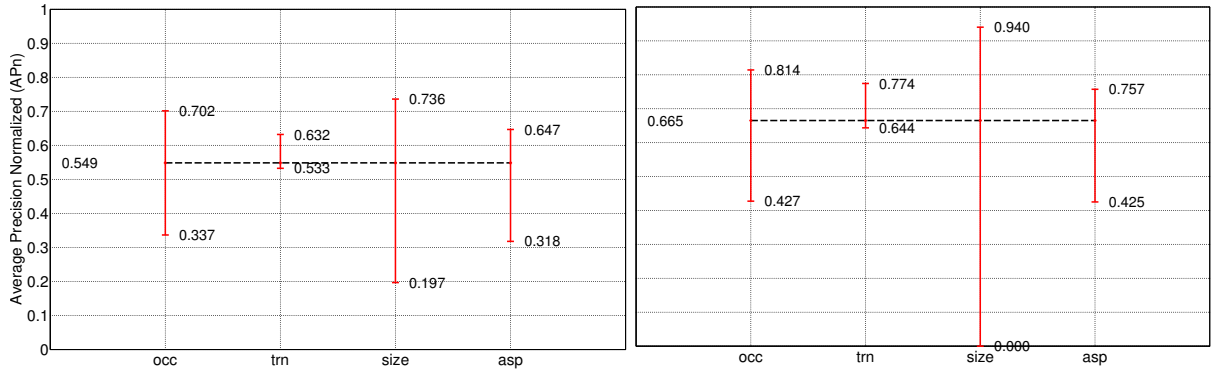
As we will show and analyze in the subsequent section, both object detectors have their strengths and weaknesses. Thus, it makes sense to combine their results to increase robustness and accuracy of the final system. In order to fuse the detections of YOLOv2 and SSD we implemented and applied a method known as *Dynamic Belief Fusion* proposed by Lee et al. in [23] which dynamically assigns probabilities to detection hypothesis based on precision-recall-curves calculated on a validation set for every species. A joint probability assignment of each detection by YOLOv2 and SSD is subsequently determined by the Dempster-Shafer combination rule [24].

C. Results

Animal Detection: After training and determination of the hyperparameters of both object detectors we conducted several experiments which we will describe in this section. As proposed in the guidelines for PASCAL VOC [21], a widely used evaluation measure to estimate the performance of object detectors is Mean Average Precision (mAP), where a true positive detection is registered for any detector hypothesis with an Intersection over Union (IoU) of at least 0.5. However, since our dataset is highly unbalanced, we report performance statistics using the normalized version of mAP as proposed in [25]. Furthermore, to better understand the performance and drawbacks of YOLOv2 and SSD in more detail, it is important to evaluate the detector's sensitivity to different types of object characteristics. In this paper we are especially interested in the following criteria: the degree of ***occlusion (occ)***, ***truncation (trn)***, ***object size (size)***, and ***aspect ratio (asp)***. For evaluation of these object characteristics we utilized the detection analysis tool of [25].
Secondly, we thoroughly analyze false positive detections of both detectors in order to be able to compare their main drawbacks. False positive detections can be divided into four main categories: ***Localization errors (loc)*** occur when an object of the target class is detected with a misaligned box. Confusion with ***similar objects (sim)*** happens when an object was correctly detected but confused with a similar object category. Note that for our six animal species we only consider Grant gazelles and Thomson's gazelles to be similar. Confusion with ***other objects (oth)*** on the other hand describe false positive detections that have an IoU of least a 0.1 with a non similar object category. All remaining false positive detections are categorized as confusion with ***background (bg)***.

Figure 2 shows the average normalized precision (AP_n) as well as the impact and sensitivity of the four object characteristics for YOLOv2 (a) and SSD (b). The AP_n of SSD with 0.67 is significantly higher than for YOLOv2 with 0.55. However, it is obvious that SSD is more sensitive to the size of an object, i.e. SSD is able to detect large objects better than YOLOv2 while on the other hand YOLOv2 is able to detect small objects better than SSD. This observation led us to the conclusion that combining the object hypotheses of both detectors should enhance the performance of the overall system.
Table II summarizes the results of both detectors as well as the proposed combination scheme for each animal class in the dataset. It can be seen that for every single species in our dataset the combination of both detectors surpasses the results of each detector alone. Table III shows the distribution of the top false-positive detections for each species.

ISBN 978-90-827970-1-5 © EURASIP 2018



(a) YOLOv2 (b) SSD

Fig. 2: Summary of sensitivity and impact of object characteristics of (a) YOLOv2 and (b) SSD.

AP_n	YOLOv2	SSD	*Combination*
Elephant	0.55	0.67	**0.70**
Grant gazelle	0.53	0.63	**0.70**
Thomson's gazelle	0.38	0.58	**0.64**
Giraffe	0.70	0.78	**0.84**
Ostrich	0.64	0.66	**0.75**
Zebra	0.48	0.67	**0.74**
All	0.55	0.67	**0.73**

TABLE II: Detection results of YOLOv2 and SSD as well as the combination of both detectors.

False Positives	loc	sim	oth	bg
Elephant	80%	–	6%	14%
Grant gazelle	38%	37%	7%	18%
Thomson's gazelle	57%	26%	–	17%
Giraffe	82.06%	–	15.38%	2.56%
Ostrich	12.5%	–	25%	62.5%
Zebra	54%	–	7%	39%

TABLE III: Detection results of YOLOv2 and SSD as well as the combination of both detectors.

As expected, a significant amount of false positive detections for Grant gazelle and Thomson's gazelle are caused by confusion with one another. Thus, depending on the usecase it might be beneficial to combine both species into a single class. Interestingly, for elephant and giraffe the amount of false-positive detections due to localization error is extremely high compared to the other classes. This is due to the fact that elephants often tend to be very close to the camera and are therefore often truncated or occluded by conspecifics. Thus, although the detector is able to detect parts of the animal it often misses other parts of the same animal. For giraffes on the other hand, the ground-truth regions contains a significant amount of background due to thin limbs and neck. Thus, apparently the detector is able to detect the main body of giraffes but often misses head and limbs which in turn leads to a IoU < 0.5. Figure 3 shows typical false-potitive detections for both classes.

Animal Counting: As described in the previous section the combination of SSD and YOLOv2 achieves the best detection results. We therefore use this combined detector to find out how good it performs at the task of animal counting in the balanced and unbalanced datasets. We count an image as correctly classified, if the number of detected animals matches the number of individuals annotated by the crowd. As expected, the animal count accuracy drops with the number of animals present in the image as shown for the balanced data set in Fig 4a. This basically means, that in about 45% of the Snapshot Serengeti database we can achieve a rate of

(a) Elephant (b) Giraffe

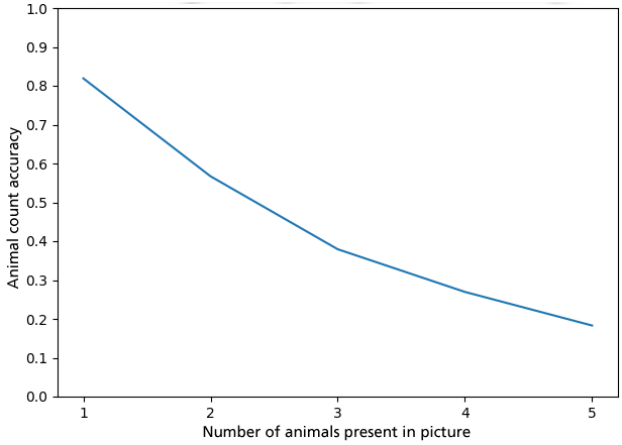
Fig. 3: Typical Examples of top-ranked false-positive detections for classes *Elephant* and *Giraffe*.

correctly counted animals of 82%. In order to give a better intuition on how reliable the animal counting works we plotted Fig. 4b. It shows the count accuracy for sets of images with increasing number of animals (i.e. images containing one, images containing one and two animals, etc.) for both, the balanced dataset and the unbalanced, more realistic, dataset, respectively. In addition, the percentage of overall pictures from the Snaphot Serengeti database that would be covered by the respective achieved accuracy is shown. For instance, for 75% of images that contain animals in the Serengeti database (1-3) a 70% (realistic unbalanced distribution) accuracy can be reached.
For the task of blank image detection (about 70% of the whole Serengeti database) the combined detector achieves results of 92% accuracy.

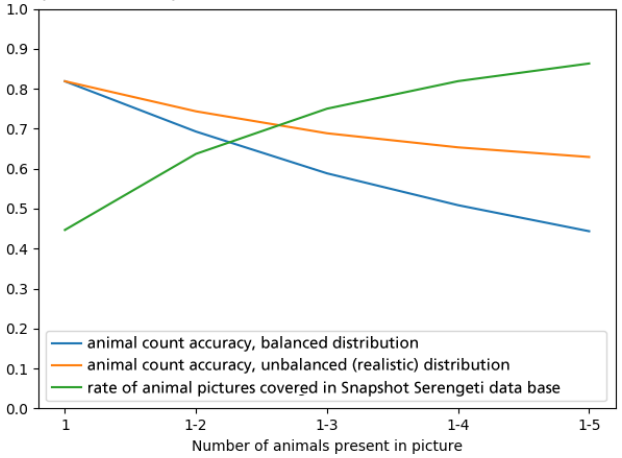
IV. Conclusion and Future Work

In this paper we successfully applied state-of-the-art deep-learning based object detectors to the field of automatic animal detection and counting in camera trap images. For that purpose we first created a ground-truth annotation task at the citizen science platform *Zooniverse* and then re-configured and re-trained two of the best performing object detectors, YOLOv2 [5] and SSD [6]. We thoroughly evaluated both algorithms on our dataset and found that both object detectors make different type of errors. Thus, combining the results of both detectors lead to a more robust and accurate detector. In contrast to previous work in the field of automatic animal monitoring, our approach is not only able to classify the main species present in visual footage but is also capable of localizing animals, classifying them and consequently count the number of animals in an image which is an important information for biologist, ecologists and gamekeepers. Thus, our approach in combination with previous work in the field of automatic visual animal biometrics has the potential to contribute significantly to wildlife research and ecological science in general.

ISBN 978-90-827970-1-5 © EURASIP 2018



(a) Animal count accuracy in the balanced data set tested on images containing 1 - 5 animals.



(b) Animal count accuracy vs. Snapshot Serengeti data base coverage.

Fig. 4: Plots of the animal count accuracy for the balanced as well as the unbalanced dataset.

Future work comprises application of more object detectors as well as the extension of our dataset by including more species. However, manually generating ground-truth information used for training by annotating thousands of images for each species can not only be tedious and time consuming but also error prone and expensive. Thus, approaches for weakly supervised learning in the context of automatic object detection such as the works by Teh et al. [26] or Li et al. [27] might be worth investigating in the near future.

References

[1] J. Vie, C. Hilton-Taylor, S. Stuart, I.-T. W. C. Union, and I. S. S. Commission, "Wildlife in a changing world: An analysis of the 2008 iucn red list of threatened species," *IUCN*, 2009.

[2] J. M. Rowcliffe and C. Carbone, "Surveys using Camera Traps: Are We Looking to a Brighter Future?," *Animal Conservation*, vol. 11, no. 3, pp. 185–186, 2008.

[3] Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer1, "Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna," *Scientific Data*, vol. 2, 2015.

[4] Alex Bowyer, Veronica Maidel, Chris Lintott, Ali Swanson, and Grant Miller, "This image intentionally left blank: Mundane images increase citizen science participation," in *Human Computation and Crowdsourcing: Works in Progress and Demonstrations. An Adjunct to the Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing*, 2015.

[5] Joseph Redmon and Ali Farhadi, "Yolo9000: Better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016.

[6] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg, "SSD: single shot multibox detector," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, 2016, pp. 21–37.

[7] Alexander Gomez Villa, Augusto Salazar, and Francisco Vargas, "Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks," *Ecological Informatics*, vol. 41, pp. 24 – 32, 2017.

[8] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Ali Swanson, Craig Packer, and Jeff Clune, "Automatically identifying wild animals in camera trap images with deep learning," *CoRR*, vol. abs/1703.05830, 2017.

[9] Hjalmar S. Kühl and Tilo Burghardt, "Animal Biometrics: Quantifying and Detecting Phenotypic Appearance," *Trends in Ecology & Evolution*, vol. 28, no. 7, pp. 432–441, Mar. 2013.

[10] Karina Figueroa, Antonio Camarena-Ibarrola, Jonathan García, and Héctor Tejeda Villela, *Fast Automatic Detection of Wildlife in Images from Trap Cameras*, pp. 940–947, Springer International Publishing, Cham, 2014.

[11] Zhi Zhang, Zhihai He, Guitao Cao, and Wenming Cao, "Animal detection from highly cluttered natural scenes using spatiotemporal object region proposals and patch verification," vol. 18, pp. 1–1, 10 2016.

[12] Xiaoyuan Yu, Jiangping Wang, Roland Kays, Patrick A. Jansen, Tianjiang Wang, and Thomas Huang, "Automated identification of animal species in camera trap images," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, pp. 52, Sep 2013.

[13] Tilo Burghardt, *Visual Animal Biometrics - Automatic Detection and Individual Identification by Coat Pattern*, Phd thesis, University of Bristol, 2008.

[14] Alexander Loos and Andreas Ernst, "An Automated Chimpanzee Identification System Using Face Detection and Recognition," *EURASIP Journal on Image and Video Processing: Special Issue on Animal Behaviour Understanding in Image Sequences*, vol. 2013, no. 49, 2013.

[15] Zooniverse, "Computer vision serengeti," https://www.zooniverse.org/projects/alexbfree/computer-vision-serengeti, 2017.

[16] Trinh Hoang Trieu, "Darkflow," GitHub repository, https://github.com/thtrieu/darkflow, 2016.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, "Microsoft coco: Common objects in context," in *13th European Conference on Computer Vision (ECCV)*, 2014.

[18] Tijmen Tieleman and Geoffrey Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," COURSERA: Neural Networks for Machine Learning, 2012.

[19] Paul Balancap, "SSD tensorflow," GitHub repository, https://github.com/balancap/SSD-Tensorflow, 2016.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[21] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, June 2010.

[22] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[23] Hyungtae Lee, Heesung Kwon, Ryan M. Robinson, William D. Nothwang, and Amar M. Marathe, "Dynamic belief fusion for object detection," in *IEEE Winter Conference on Applied Computer Vision (WACV)*, 2016.

[24] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *The Annals of Mathematical Staticstics*, vol. 38(2), pp. 325 – 339, 1967.

[25] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai, "Diagnosing error in object detectors," in *European Conference on Computer Vision (ECCV)*, 2012.

[26] Eu Wern Teh, Mrigank Rochan, and Yang Wang, "Attention networks for weakly supervised object localization," in *British Machine Vision Conference (BMVC)*, 2016.

[27] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang, "Weakly supervised object localization with progressive domain adaptation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

ISBN 978-90-827970-1-5