# Nature Conservation Drones for Automatic Localization and Counting of Animals

Jan C. van Gemert[1](✉), Camiel R. Verschoor[2,3], Pascal Mettes[1], Kitso Epema[2], Lian Pin Koh[4], and Serge Wich[5,6]

[1] Intelligent Systems Lab Amsterdam,
University of Amsterdam, Amsterdam, The Netherlands
{j.c.vanGemert,p.s.m.Mettes}@uva.nl

[2] Dutch Unmanned Aerial Solutions, Amsterdam, The Netherlands
{camielVerschoor,kitso.Epema,lianpinkoh,sergewich}@gmail.com
http://dutchuas.nl/

[3] IDI Snowmobile, Amsterdam, The Netherlands
http://idisnow.com

[4] Applied Ecology and Conservation Group,
University of Adelaide, Adelaide, Australia

[5] Institute for Biodiversity and Ecosystem Dynamics,
University of Amsterdam, Amsterdam, The Netherlands

[6] School of Natural Sciences and Psychology,
Liverpool John Moores University, Liverpool, UK

**Abstract.** This paper is concerned with nature conservation by automatically monitoring animal distribution and animal abundance. Typically, such conservation tasks are performed manually on foot or after an aerial recording from a manned aircraft. Such manual approaches are expensive, slow and labor intensive. In this paper, we investigate the combination of small unmanned aerial vehicles (UAVs or "drones") with automatic object recognition techniques as a viable solution to manual animal surveying. Since no controlled data is available, we record our own animal conservation dataset with a quadcopter drone. We evaluate two nature conservation tasks: (i) animal detection (ii) animal counting using three state-of-the-art generic object recognition methods that are particularly well-suited for on-board detection. Results show that object detection techniques for human-scale photographs do not directly translate to a drone perspective, but that light-weight automatic object detection techniques are promising for nature conservation tasks.

**Keywords:** Nature conservation · Micro UAVs · Object detection

## 1 Introduction

Accurate monitoring of the distribution and abundance of animal species over time is a key ingredient to successful nature conservation [3,4]. Successful conservation also requires data on possible threats to animals. Such threats can be

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part I, LNCS 8925, pp. 255–270, 2015.
DOI: 10.1007/978-3-319-16178-5_17

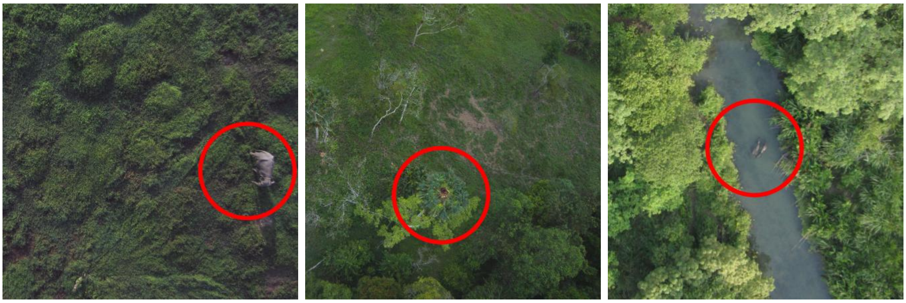


**Fig. 1.** Animal conservation images taken from a drone. From left to right: an elephant, an orangutan nest, and rhinos.

largely divided into habitat loss, disease and poaching. For some iconic species like the rhino, the elephant, and the tiger, poaching has reached proportions that places them at a high risk for local extinctions or even total extinction for some (sub)species as in the case of elephants [2,32].

Animal monitoring approaches typically involve both direct animal counts and indirect counting of animal signs such as nests, dung, and calls. Conventional ground surveys on foot can be time-consuming, costly, and nearly impossible to achieve in remote areas. For example, ground surveys of orangutan populations (Pongo spp.) in Sumatra, Indonesia can cost up to $250,000 for a three-year survey cycle. Due to this high cost, surveys are not conducted at the frequency required for proper statistical analysis of population trends. Furthermore, there remain many remote forested areas that have never been surveyed. Aerial surveys can overcome some of these constraints, although they have their own set of limitations, including the high cost of buying or renting small planes or helicopters, the lack of availability in remote areas, and the risks involved with flying low over landscapes in which landing is difficult, such as forests. There is thus a need for alternative methods for animal surveys.

Conservation workers have started using small unmanned aerial vehicles (UAVs, or "conservation drones") both for determining animal abundance and to obtain data on their threats [18,20]. Conservation drones are relatively inexpensive and easy to build, which makes drones accessible and affordable for many research teams in developing countries. These drones can fly fully autonomous missions to obtain high-resolution still images and videos. Recent studies have shown that the images from such drones can be used to detect not only large animal species (e.g. orangutans, elephants, rhinos, whales) and animal tracks (e.g. orang-utan nests, chimpanzee nests, turtle tracks), but also threats to animals (e.g. signs of human activity [16,21,26,35]). See Figure 1 for some examples of conservation images taken from a drone. Currently, most drone systems record data on board, which are then downloaded for manual visual inspection once the drone has landed. For animal abundance surveys, the amount of recorded data quickly grows to thousands of photos and hundreds of hours of video. Manually sieving through these data in search of animals is labor-intensive and inherently slow.

There is therefore a strong need to automate the detection of relevant objects on the still or video images. Recent efforts combining human labeling and automatic recognition seem promising [5], but this field is still in its infancy.

Another need for automated detection of objects comes from anti-poaching efforts. Ideally, drones would do on-board object detection and then only send the relevant images (i.e. those with a high probability of a positive identification) of the object of interest, e.g. *human*, *rhino*, *fire*, down to the ground station for a manual visual inspection by the rangers in order to take appropriate actions. This paper examines automatic object detection algorithms as a solution towards detecting animals and humans from images obtained from drones. The aim is to assess the potential of computer vision for surveys of animal abundance and anti-poaching efforts.

The field of computer vision has matured enough to automatically find objects in images with reasonable accuracy. Such methods are typically designed for and evaluated on general purpose objects by employing photographs from the Internet [22]. Thus, such methods are tuned towards human photographs, taken from a height of 1-2 meters with human-scale objects. Such objects can safely be assumed to consist of observable parts [11] or to be found by object-saliency methods (so called "object proposals"), tuned to human scale [1,15,17,31]. Yet, for drone imagery taken in high altitude (10-100m) the objects of interest are relatively small, questioning the suitability of current methods that use individual parts or object-saliency. Moreover, drone images are taken from above which results in a skewed vantage point, which changes influential surface and scene properties [8] when compared to normal human pictures. It can therefore not be taken for granted that current object detection methods for human-centered imagery find a one-to-one application in conservation drones.

In this paper we evaluate how current object detection techniques as developed for human-centered imagery scale to drone-centered nature conservation tasks. Because current object recognition methods make heavy use of object proposals, we first evaluate whether such proposals are capable of detecting animals in drone imagery. Next, we evaluate three light-weight object detection methods on two nature conservation tasks: i) animal detection in single images; ii) animal counting in video. We evaluate these two tasks on a novel fully annotated animal dataset recorded with a drone. The dataset consists of 18,356 frames containing 30 distinct animals. This work stems from a collaboration between conservation biologists, aerospace engineers and computer vision researchers. To facilitate a structured and repeatable evaluation we will make the dataset, annotations, code, and all our results publicly available.

## 2 Related Work

### 2.1 Drones for Nature Conservation

Drones are air vehicles that do not carry a pilot on board and are capable of flying autonomously. These vehicles follow flight plans based on GPS coordinates which are usually programmed before flight, but can also be changed during the flight. There are many types of drones, ranging in weight from a few grams to thousands of kilograms, varying in size from a few millimeters to tenths of meters with configurations according to normal aircrafts, helicopters, multi-rotors and flapping wings. The uses of these drones vary from the military to private consumers. The type of drones that are specifically useful for nature conservation are the modified model aircraft and multi-rotor. They are both affordable and easily converted into a drone with highly affordable open source autopilots like Ardupilot [20] or Paparazzi [13]. The modified model aircrafts yield long flying times and larger forward speed to cover more ground. In contrast, multi-rotor drones yield great control of the position and orientation of the camera as well as vertical take off and landing capabilities. Combine this with the birds eye view for the camera and these drones are perfect for conservation. Here we focus on the rotor-type drone.

Drones are currently already employed for conservation [20] for terrain mapping and classification of forest types [14,16,18,21,35]. These are examples of uses where no real time data analysis is needed [5]. For the protection of animals against poaching, real-time analysis is critical. This is recognized by the Wildlife Conservation UAV Challenge [38], which focuses on the on-board processing of data to find rhinos and humans. This is an international challenge (with almost 90 teams from all over the world) to create a cheap drone to help protect the rhino. The techniques to find animals and humans in real-time with limited computing power are not ready for real-world applications yet, validating our research on this topic.

### 2.2 Automatic Object Detection

The current state-of-the-art in automatic object detection is based on large convolutional neural networks [15,29]. Such “deep-learning” networks discriminatively learn image features from the bottom-up and proved hugely successful on global image classification [22]. The success of convolutional networks on full images spills over to the related task of object detection where in addition to the object class name, a detection bounding box around the object is required. The bounding box is obtained by elegant object-saliency methods that output a small set of only a few thousand bounding boxes that have a high likelihood to contain any type of object [1,31]. Such class-independent object-proposals serve as input to the convolutional network, yielding state-of-the-art accuracy. Such accuracy, however, relies heavily on modern computer hardware such as top of the line CPUs and massively parallel GPU implementations. The recognition times using modern hardware are reported as 53 sec/image by R-CNN [15] on a CPU and 13 sec/image on the GPU whereas OverFeat [29] operates at 2 sec/image on an heavy-weight GPU. These hardware requirements are not feasible in a light-weight drone, where every gram of weight reduces flight times. Since fast response time is essential for animal protection, convolutional networks are as of yet computationally too demanding for timely results on a drone.

Next to convolutional networks, other competitive object detection methods are based on the bag-of-words (BOW) model [31,33,34] or its Fisher vector

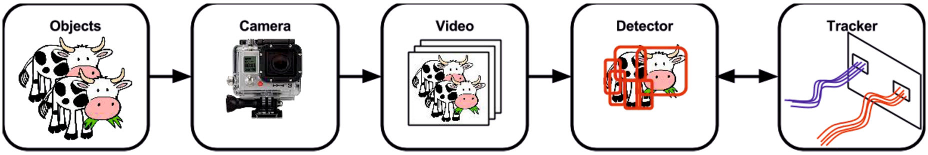


**Fig. 2.** Conservation evaluation pipeline. Animals are recorded on video or individual images. The animals are automatically detected, yielding a bounding box per animal per image. Individual detections are stitched together by tracking shared features to obtain an automatic estimate on the number of animals.

incarnation [6,28]. Such methods start with a limited set of object-proposals to reduce the search space. Each proposal is represented with a histogram of prototype counts of local features, e.g. sampled from interest points [9]. Larger prototype vocabularies typically yield best results, resulting in features sizes of over 170,000 [31] for BOW or over 300,000 for the Fisher vector [28] per bounding box. On a mobile drone the internal memory is limited making such large memory requirements of the BOW variants prohibitive.

Both the bag-of-words methods and the convolutional neural networks heavily rely on high quality object proposals. These proposals are tuned to a human scale, and we will first evaluate the suitability of object proposals for drone imagery. Successful low-memory and CPU-friendly object detection methods do not use object proposals but simple and fast image features combined with a classifier cascade. Such a cascade rejects obvious non-matching candidates early on; only allotting more computation time to promising candidates. An example of such an successful approach is the seminal Viola and Jones boosting method [36] used in embedded face detection algorithms for consumer cameras, phones and tablets. Other versions of a classifier cascade have been applied to a range of object detection methods with impressive speed-up results. The popular Deformable Part-based Model (DPM) of Felzenswalb et al. [11] models an object as a constellation of parts and a classifier cascade in combination with a coarse-to-fine search has successfully reduced computation time to 0.2 sec/image [10,27]. Similarly, the examplar SVM approach for object detection [25] can be sped up to 0.9 sec/image [23]. Cascade approaches are fast while retaining a reasonable accuracy and are thus most suitable for on a drone. Therefore we will focus our evaluation on the DPM and exemplar-based SVM methods.

## 3 Evaluating Nature Conservation

We evaluate two tasks for automatic nature conservation drones. i) animal detection and ii) animal counting. Automatically detecting animals gives insight in animal locations, which over time will reveal herd patterns and popular gathering places. Knowledge about the animal location will give the conservation worker valuable information about where to take anti-poaching measures. The second

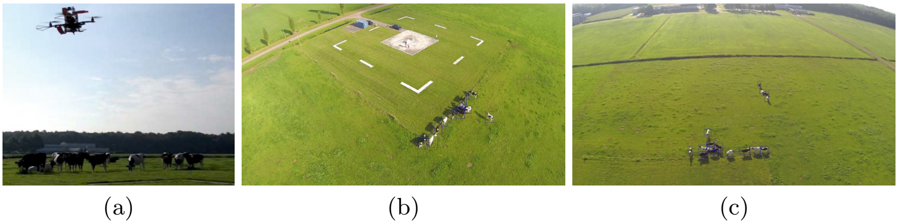


**Fig. 3.** Recording the dataset. (a): the Pelican quadcoptor drone used to record the dataset (b): an example image from the train-set (c): an example image from the test-set. Note the slanted vantage point and tiny animals.

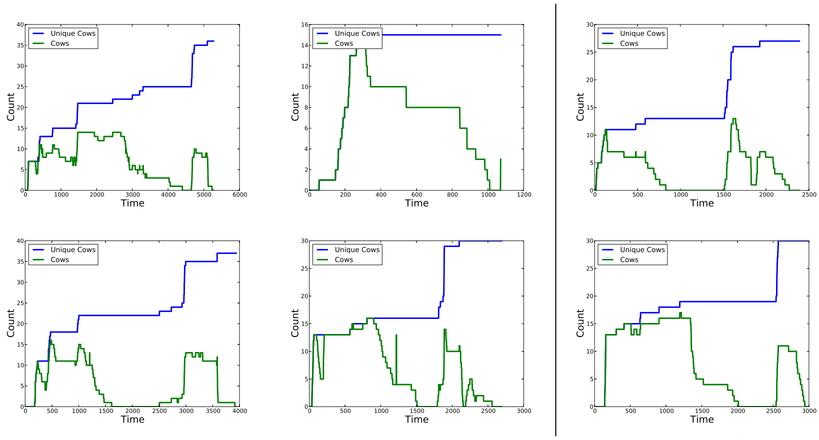
task, counting animals, will give abundance data over time, coupled to detection regions. This data gives the conservator a sense of the health of the animal population, and where and how many animals disappear, which warrants further investigation. The pipeline for the two evaluation tasks is visualized in Figure 2.

### 3.1 Recorded Dataset

Wildlife is notoriously hard to record in a controlled setting. To approximate a realistic conservation task, we used a quadcopter drone to record a dataset of domesticated farm animals. In figure 3(b,c) we show examples of the recordings. While the exact type of animal (cow), the lack of camouflage in the open fields and the presence of man-made structures is not common in the wild, this recording nevertheless retains important properties that match a realistic conservation scenario. Such realistic properties include the use of a quadcopter drone which is often used in the wild because of its maneuverable and its ability to take off from dense areas. Moreover, this type of drone gives us the opportunity to record under a wide variation of positions, heights, and orientations of the camera. Therefore, the recording setup matches closely as experienced in the wild. Furthermore, the animals are smaller and of a similar size and build as many conservation animals like the rhino or the elephant. The dataset provides an excellent first opportunity to evaluate nature conservation drone algorithms.

The dataset was recorded by the Ascending Technologies Pelican (quadcopter) with a mounted GoPro HERO 3: Black Edition action camera. In Figure 3(a) we show the drone in action. We manufactured a 3D printed custom-made mount to attach the camera to the drone. The mount is filled with foam to counter vibration of the camera during flight. The camera recorded videos at a quality of 1080p (1920 x 1080 pixels) having a medium field of view (55° vertical and 94.4° horizontal) with 60 frames per second.

We performed two separate flights to obtain a set for training and a disjoint set for testing. We manually annotated all animals in the dataset with vatic [37]. We removed large portions of the videos that do not contain any animals, which resulted in 6 videos obtained from the two seperate flights. We use the first 4 videos from the first flight for training, and the latter 2 videos from the second



Training videos

Testing videos

**Fig. 4.** Number of animals per frame in each video, the green line shows the number of unique animals per frame, the blue line the cumulative total. The 2 left columns represent the training videos, the right column the test videos.

flight for testing. In total there are 12,673 frames in the training set and 5,683 frames in the test set. There are 30 unique animals present in the dataset. In figure 4 we visualize the appearance and disappearance of animals during the flight.

## 4 Object Detection Methods Suitable for Drones

### 4.1 Deformable Part-based Model

The deformable part-based model (DPM) of Felzenszwalb et al. [11] is a popular and successful object detection instantiation of the pictorial structure representation [12] where an object is modeled as a flexible constellation of parts. In addition to the gray-valued DPM, we also evaluate the color-DPM [19], where color information is added to the features.

In the DPM [11], an object consists of a collection of parts connected in a star-structure to a coarse root node. The parts and root node are represented by HOG features and the quality of a star model is the score of the root features at a given location plus the sum of the maximum over the part placements minus a deformation cost based on the deviation of the parts from its ideal location.

The DPM is trained on bounding box annotations around the whole object; where object-part locations are not labeled. To discriminatively train models in such a semi-supervised setup a latent SVM is used. From labeled bounding boxes $\{x_1, x_2, \ldots, x_n\}$ where each box has a class label $y_i$ being either $+1$ or $-1$ a latent SVM allows training when the part-locations are unknown. A latent SVM scores an example $x$ as

$$f_\beta(x) = \max_{z \in Z(x)} \beta \cdot \Phi(x, z), \tag{1}$$

where $\Phi(x, z)$ is a feature vector and the set $Z(x)$ has all possible latent variables (object configurations) and $\beta$ is a vector of model parameters trained by minimizing the SVM hinge loss

$$L(\beta) = \frac{1}{2}||\beta||^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i f_\beta(x_i)), \tag{2}$$

where $C$ is the regularization parameter. See [11] for details.

A significant speed-up for the DPM can be obtained by a part-based cascade [10]. This cascade is based on an ordering of the parts hypotheses and prunes low scoring hypotheses, allowing the bulk of the computation time to be spent on promising candidates. Besides the cascade, another speed-up can be obtained by a hierarchical coarse-to-fine feature matching approach [27]. The speed-up is based on the observation that the computational cost of the DPM is dominated by the cost of matching each part to the image. The number of part matches can be significantly reduced by using a coarse-to-fine inference strategy. The combination of the cascade and the coarse-to-fine matching yields a speed-up of one up to two orders of magnitude resulting in detection rates of 0.2 sec/image [27]. Such speeds are acceptable for nature conservation applications on low-cost drone hardware.

### 4.2 Exemplar SVM

The ensemble of exemplar SVM object detection approach by Malisiewicz et al. [25] trains a parametric SVM for each positive exemplar in the dataset. This reaps the benefits of a non-parametric nearest neighbor search with training a parametric model. The parametric SVM can effectively deal with negative samples whereas the non-parametric approach retains the link between positively labeled training exemplars which allows knowledge transfer such as pose, geometry or layout from an exemplar to a new object. This approach is conceptually simple and yields good performance.

An exemplar SVM aims to separate each training example $x_E$ from other examples in the set $N_E$ that do not containing the object class by learning a weight vector $w_E$ by optimizing

$$||w||^2 + C_1 h(w^T x_E + b) + C_2 \sum_{x \in N_E} h(-w^T x - b), \tag{3}$$

where $h(x) = \max(0, 1 - x)$ is the hinge loss and $C_1$ and $C_2$ are regularization parameters. Each individual exemplar SVM is trained on a unique set of negative examples which makes the output scores of the SVM's not necessarily comparable. To calibrate each SVM, a sigmoid function is fitted on hold-out data, resulting in comparable SVM output between 0 and 1. All exemplar models are applied on a test image by means of a sliding window, where exemplar co-occurences are used to obtain a detection.

To speed-up exemplar SVMs, Li et al. [23] use each exemplar as a weak classifier in a boosting approach. Boosting builds a strong classifier with a linear combination of weak classifiers. These weak classifiers are iteratively selected to optimize the mis-classification rate. The iterative approach of boosting performs feature selection, using only the best $T$ weak classifiers. Feature selection drastically reduces the number of used exemplars where [23] need only 500 exemplars for state of the art performance. In addition to the feature selection, Li et al. [23] propose efficient feature-sharing across image pyramid scales resulting in a detection speed of 0.9 sec/image, which is similarly acceptable for on a drone.

# 5 Animal Counting

For counting animals, the algorithm needs to keep track of each unique animal to make sure a single animal is only counted once. This is a challenging task for several reasons. The animal detection algorithm may fail missing the animal completely or only in some frames in a video. It is also possible that the detection algorithm detects an animal where there is none which will inflate the amount of counted animals. Furthermore, because of the maneuverability of drone and possible pre-programmed flying paths, the same animal may appear and disappear from the drone camera completely.

The detection algorithms process every frame. To determine if two detections in subsequent frames belong to the same unique animal, the object detections have to be stitched together over time. We aim to track detections over multiple frames, even when one or a few detections are is missing. Thus, instead of basing our method on detection tracking which may be noisy, we track salient points which are more stable. For point tracking we use the KLT tracker [24] which uses optical flow to track sparse interest points for $L$ frames, where $L$ is a parameter. To determine whether two subsequent detection bounding boxes A and B belong to the same unique animal we use the intersection over union measure $\frac{A \cap B}{A \cup B} > 0.5$ of the set of point tracks through $A$ and through $B$ [7].

Note that animal counting is different from a spatio-temporal localization task, as e.g. proposed in [17,30]. In a localization task the objective is to carefully and completely identify what, when and where each object is in a video. In animal counting, the exact position is not as relevant to the conservation worker since she is interested in how many unique animals are correctly found.

# 6 Experiments

## 6.1 Experiment 1: Evaluating Proposal Quality

Object proposals significantly reduce the object detection search space by locating a modest collection of bounding boxes that with a high likelihood contain any type of object. We focus on the selective search object proposal algorithm of Uijlings et al. [31]because of its high-quality proposals. Selective search generates object proposals from an initial set of super-pixels obtained through an over-segmentation and merges super-pixels using a range of color, shape and other super-pixel similarity measures. We evaluate both the *fast* and *quality* settings of selective search [31].

**Table 1.** Statistics of object proposals on the drone dataset. Overlap between proposals $A$ and ground-truth boxes $B$ is measured as $\frac{A \cap B}{A \cup B}$. ABO (Average Best Overlap) is the best overlapping proposal per frame, averaged over all frames. Recall is measured as the fraction of ground truth boxes that have an overlap greater than 0.5.

| **Setting** | **ABO** | **Recall** | **Proposals / frame** | **Time / frame** |
|---|---|---|---|---|
| Fast | 0.635 | 0.873 | ca. 18,369 | ca. 31 sec. |
| Quality | 0.740 | 0.976 | ca. 64,547 | ca. 140 sec. |

In Table 1 we show an overview of the results achieved on a set of sampled video frames. As can clearly be seen in the Table above, the average best overlap scores and recall scores yielded on the dataset of this work do not meet the results reported in [31]. For the *fast* setting of selective search, only 87% of the cows can be found in the set of proposals, while the average best overlap (ABO) is a mere 63.5%. In contrast, on the Pascal VOC 2007, the same setting yields a mean ABO of 80.4%, while nearly all the objects can be found in the proposals. Similarly, the *quality* setting only yields an average best overlap of 74%. Besides a low detection and overlap rate, there is also a strong increase in the number of generated object proposals and the proposal generation time per frame. This is best shown for the *quality* setting, where it takes nearly two and a half minutes per frame to generate proposals resulting in more than 64,000 proposals.

As we are interested in lightweight solutions for drone imagery, the evaluation time of the selective search algorithm poses serious practical problems. Not only will it take at least 30 seconds to generate a set of proposals with an acceptable recall rate, after that, features have to be extracted and a classifier applied on tens of thousands of proposals. Thus, the proposals do not significantly reduce the search time and we conclude that object proposal-based detection systems are from a computational standpoint not suited for the problem at hand.

### 6.2 Experiment 2: Animal Detection

For the second experiment, the three high-speed object detection methods: DPM, color-DPM and exemplar SVM are evaluated on our dataset. In order to generate a final model for a specific object class, all three methods use a hard negative mining procedure. As such, a diverse set of negatives is required. To meet this requirement, the set-up of the Pascal VOC challenge is mimicked. More specifically, the images of the *cow* class in the Pascal VOC are replaced by randomly sampled images from the train and test video for resp. the train and test set. In this setup, the train images from the other 19 classes can be used for discriminative learning. These classes include people, animals (*cats, dogs, horses, sheep*),

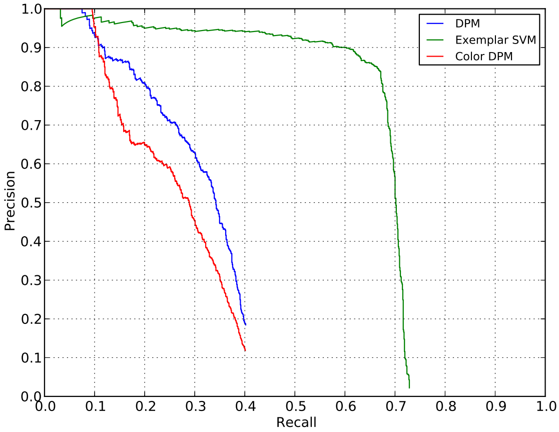


**Fig. 5.** Precision-recall curves for all three methods on the test images of the dataset

vehicles (*aeroplanes, bikes, boats, busses, cars, motorbikes, trains*), and indoor objects (*bottles, chairs, dining tables, potted plants, sofa's, tv's*).

The trained models are applied to all frames in the test set, containing a total of 1,227 ground truth bounding boxes. The result of this evaluation is a list of bounding box detections, ranked by their respective confidence value which is evaluated by precision and recall values.

In Figure 5, the precision-recall curves are shown for exemplar SVM, DPM, and color-DPM. As can be deduced from the Figure, exemplar SVM significantly outperforms both other methods in terms of precision (after a recall of 0.15) and recall. This holds similarly for the Average Precision scores; 0.66 for exemplar SVM, compared to 0.30 for DPM and 0.26 for color-DPM. These results are surprising when compared with reported results on standard object detection datasets. For example, Khan et al. [19] indicate that color-DPM is prefered over standard (grayscale) DPM, while DPM in turn reports better results than exemplar SVM [11].

A particular interesting aspect of the curves in Figure 5 is that both DPM models reach a final recall of roughly 0.4, while exemplar SVM reaches a final recall of roughly 0.72. A primary reason for this discrepancy lies in the total number of detected objects for the methods. While there are in total 1,227 positive instances of the object to be discovered, DPM detects 2,673 and color-DPM detects 4,156 bounding boxes. Exemplar SVM on the other hand, report a total of 40,654 bounding boxes. With such a high number of object detections, the final recall is bound to be higher.

The high number of object detections do however not explain the high precision of Exemplar SVM. The answer to this is two-fold. First, the use of a joint global and part-based model in DPM does not work favorably given the small scale of the animals in the drone images. As shown in Figure 6, individual

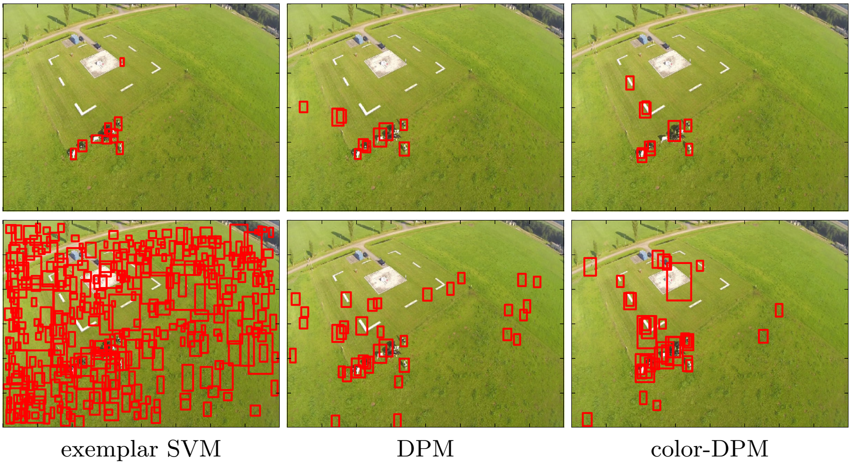


**Fig. 6.** Examples of detected bounding boxes for a test frame. The top row show the 10 highest ranked detections per method, while the bottom row shows all positive detections.

animals are generally tiny, due to the high altitude of the drones. When an animal is then visible e.g. in a window of 25 by 25 pixels, there is not enough gradient information for reliable global and part-based models. A second reason for the high results of exemplar SVM lies in the dataset. Since this evaluation is aimed at detecting cows, there is limited discrepancy between the instances to discover during training and testing. As we are in a practical scenario also interested in a limited range of wildlife animals, the use of animal exemplars for detection is beneficial.

In Figure 6, qualitative detection results are shown for a single test frame. For the top ranked detections (top row), the results of all methods look rather promising. When looking at all the positively detected bounding boxes however, the results become cluttered. For exemplar SVM, it is even unknown what the promising locations of the cows are. Nevertheless, the methods are capable of highly ranking the correct image locations, but also tend to fire on high contrast corner areas (such as the white lines or humans in Figure 6) and cluttered locations. Similar to the results of the first experiment, the results of this experiment indicate that results yielded on human-scale images are not be directly applicable to drone imagery.

### 6.3 Experiment 3: Animal Counting

In the third experiment we evaluate the quality of animal counting based on frame-based detections in combination with point tracks obtained by a KLT tracker [24]. To avoid counting the same animal multiple times for several frames, the frame-based object detections are either stitched to an existing group of detections or are seen as a new unique group of detections; where ideally each

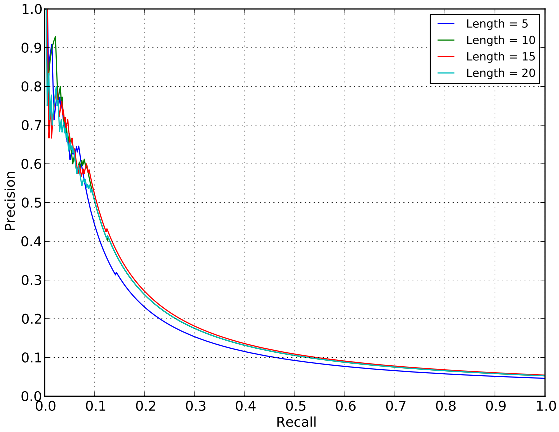


**Fig. 7.** Precision-recall curves for different point track sizes on the ground truth where bounding boxes are sampled every 5 frames from the annotations.

group of detections represents a single unique animal. The KLT algorithm generates point tracks of length $L$ and we chose the values $L \in \{5, 10, 15, 20\}$ as a range of stable values.

Counting is evaluated by precision-recall curves, albeit with special considerations. The recall is defined as all unique animals. The precision is computed based on the correctness of a stitched group of detections. For counting, we consider a count correct only if it adheres to these three strict rules: i) there are no multiple animals within a single track; ii) all individual detections are animals and iii) the found animal is unique and has not been counted before. These strict criteria allow us to draw a precision-recall curve where we sort stitched frames based on the number of aggregated frames.

We first evaluate the quality of the tracking algorithm. To this end, we simulate a perfect detector by sampling ground truth boxes for every 5 frames in the video. The average precision is 0.216 and in Figure 7 we show the precision-recall curve. A track-length of $L = 15$ is performing best, although the difference is minimal.

Next, we evaluate the counting results using the automatically generated bounding box detection, using an empirically set threshold of -0.8 to discard false positives. In Figure 8 we show the corresponding precision-recall curves. Compared to the results of Figure 7, the curves are lower, while also not all animals are found. Similarly, the average precision score is lower, with a score of 0.193.

Based on the above yielded results, animal counting turns out to be a challenging problem. Even on ground truth detections, there is plenty of room for improvement, which we will focus on in future research.

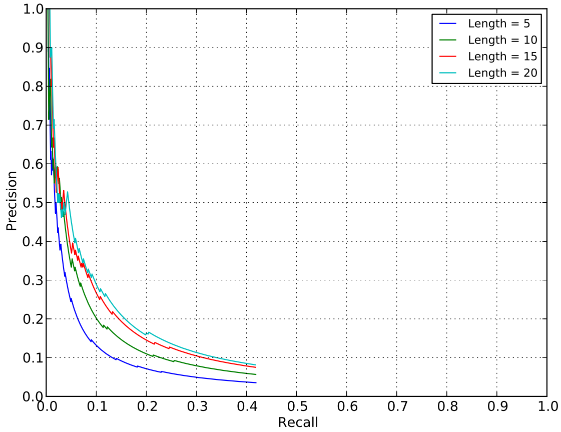


**Fig. 8.** Precision-recall curves for different point track sizes on the DPM detections with a threshold of -0.8.

## 7 Conclusion

We investigate if current automatic object detection methods as designed for human-centered objects are suitable for nature conservation on a drone, where objects are typically much smaller and observed from above. We define two task: i) animal detection and ii) animal counting. These tasks are important for monitoring animal distribution and animal abundance as typically required for successful nature conservation. To evaluate these tasks, we manually recorded and annotated a new dataset with a quad-copter drone. The animal detection task is benchmarked with three light-weight object detection algorithms that are suitable for on-board implementation since their potential detection speed is less than 1 second per image. We base the animal counting task on the detection task for which we define a suitable evaluation protocol. Results on counting show that this task is difficult, and as such an interesting research question. Results for object detection show that the performance of object detection methods as evaluated on images of human objects does not directly translate to drone imagery. According to the literature, the color-DPM should outperform the standard DPM, which in turn should outperform exemplar SVM. Our results are the exact opposite of this ordering. Nevertheless, detection results are promising, showing that automatic animal conservation with drones is a fruitful combination of biology, aircraft engineering and computer vision.

## References

1. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(11), 2189–2202 (2012)

2. Bouché, P., Renaud, P.C., Lejeune, P., Vermeulen, C., Froment, J.M., Bangara, A., Fiongai, O., Abdoulaye, A., Abakar, R., Fay, M.: Has the final countdown to wildlife extinction in northern central african republic begun? African Journal of Ecology **48**, 994–1003 (2010)
3. Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L., Thomas, L.: Introduction to Distance Sampling. Oxford University Press (2001)
4. Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L., Thomas, L.: Advanced Distance Sampling: Estimating Abundance of Biological Populations. Oxford University Press (2004)
5. Chen, Y., Shioi, H., Montesinos, C.F., Koh, L.P., Wich, S., Krause, A.: Active detection via adaptive submodularity. In: Proceedings of The 31st International Conference on Machine Learning, pp. 55–63 (2014)
6. Cinbis, R.G., Verbeek, J., Schmid, C.: Segmentation driven object detection with fisher vectors. In: International Conference on Computer Vision (ICCV) (2013)
7. Everingham, M., Sivic, J., Zisserman, A.: Taking the bite out of automated naming of characters in tv video. Image and Vision Computing **27**(5), 545–559 (2009)
8. Everts, I., van Gemert, J.C., Gevers, T.: Per-patch descriptor selection using surface and scene properties. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 172–186. Springer, Heidelberg (2012)
9. Everts, I., van Gemert, J.C., Gevers, T.: Evaluation of color spatio-temporal interest points for human action recognition. IEEE Transactions on Image Processing **23**(4), 1569–1580 (2014)
10. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.: Cascade object detection with deformable part models. In: Computer Vision and Pattern Recognition (CVPR) (2010)
11. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(9), 1627–1645 (2010)
12. Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. IEEE Transactions on Computers **22**(1), 67–92 (1973)
13. Gati, B.: Open source autopilot for academic research - the paparazzi system. In: American Control Conference (ACC), pp. 1478–1481 (2013)
14. Getzin, S., Wiegand, K., Schöning, I.: Assessing biodiversity in forests using very high-resolution images and unmanned aerial vehicles. Methods in Ecology and Evolution **3**, 397–404 (2012)
15. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Computer Vision and Pattern Recognition (CVPR) (2014)
16. Hodgson, A., Kelly, N., Peel, D.: Unmanned aerial vehicles (uavs) for surveying marine fauna: a dugong case study. PloS One **8** (2013)
17. Jain, M., van Gemert, J.C., Bouthemy, P., Jegou, H., Snoek, C.: Action localization by tubelets from motion. In: Computer Vision and Pattern Recognition (CVPR) (2014)
18. Jones IV, G.P., Pearlstine, L.G., Percival, H.F.: An assessment of small unmanned aerial vehicles for wildlife research. Wildlife Society Bulletin **34**, 750–758 (2006)
19. Khan, F.S., Anwer, R.M., van de Weijer, J., Bagdanov, A., Vanrell, M., Lopez, A.M.: Color attributes for object detection. In: Computer Vision and Pattern Recognition (CVPR) (2012)
20. Koh, L.P., Wich, S.A.: Dawn of drone ecology: low-cost autonomous aerial vehicles for conservation. Tropical Conservation Science **5**(2), 121–132 (2012)

21. Koski, W.R., Allen, T., Ireland, D., Buck, G., Smith, P.R., Macrander, A.M., Halick, M.A., Rushing, C., Sliwa, D.J., McDonald, T.L.: Evaluation of an unmanned airborne system for monitoring marine mammals. Aquatic Mammals **35**(347) (2009)
22. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 1097–1105 (2012)
23. Li, H., Lin, Z., Brandt, J., Shen, X., Hua, G.: Efficient boosted exemplar-based face detection. In: Computer Vision and Pattern Recognition (CVPR) (2014)
24. Lucas, B.D., Kanade, T., et al.: An iterative image registration technique with an application to stereo vision. In: International Joint Conference on Artificial Intelligence, vol. 81 (1981)
25. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-svms for object detection and beyond. In: International Conference on Computer Vision (ICCV) (2011)
26. Mulero-Pázmány, M., Stolper, R., Essen, L.V., Negro, J.J., Sassen, T.: Remotely piloted aircraft systems as a rhinoceros anti-poaching tool in Africa. PloS One **9** (2014)
27. Pedersoli, M., Vedaldi, A., Gonzalez, J.: A coarse-to-fine approach for fast deformable object detection. In: Computer Vision and Pattern Recognition (CVPR) (2011)
28. van de Sande, K.E.A., Snoek, C.G.M., Smeulders, A.W.M.: Fisher and vlad with flair. In: Computer Vision and Pattern Recognition (CVPR) (2014)
29. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: integrated recognition, localization and detection using convolutional networks. In: International Conference on Learning Representations (2014)
30. Tian, Y., Sukthankar, R., Shah, M.: Spatiotemporal deformable part models for action detection. In: Computer Vision and Pattern Recognition (CVPR), pp. 2642–2649 (2013)
31. Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International Journal of Computer Vision **104**(2), 154–171 (2013)
32. UNEP: Elephants in the dust the african elephant crisis. a rapid response assessment (2013). www.grida.no, united Nations Environment Programme, GRID-Arendal
33. Van Gemert, J.C., Veenman, C.J., Geusebroek, J.M.: Episode-constrained cross-validation in video concept retrieval. IEEE Transactions on Multimedia **11**(4), 780–786 (2009)
34. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: International Conference on Computer Vision (ICCV) (2009)
35. Vermeulen, C., Lejeune, P., Lisein, J., Sawadogo, P., Bouché, P.: Unmanned aerial survey of elephants. PloS One **8** (2013)
36. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Computer Vision and Pattern Recognition (CVPR) (2001)
37. Vondrick, C., Patterson, D., Ramanan, D.: Efficiently scaling up crowdsourced video annotation. International Journal of Computer Vision, 1–21 (2012)
38. WCUAVC: Wildlife conservation uav challenge. http://www.wcuavc.com/