ECOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

MASTER THESIS

Combining UAV-imagery and machine learning for wildlife conservation

Author:

Nicolas Rey

Supervisors:

Dr. Stéphane Joost
Prof. Dr. Devis Tuia

Laboratory of Geographic Information Systems (LASIG)

June 2016

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Abstract

Semi-arid savannas are endangered by changes in the fragile equilibrium between rainfalls, fires and grazing pressure exerted by wildlife or cattle. To avoid bush encroachment and the decline of perennial grass, land managers must pay attention to keep the amount of cattle and wildlife in balance with the grass availability. In large farms and conservation parks, to estimate the animal populations is therefore an important management aspect.

Traditional methods of animal census – such as transect counts from a helicopter, or mark / recapture – are too expensive and laborious to be conducted on a regular basis. In this context unmanned aerial vehicles (UAVs) appear as an interesting tool for animals detection. They can be easily deployed, for lower cost and an increased safety. The drawback is that it is difficult to visually interpret the large number of very high resolution (VHR) images that they acquire. The recent advances in machine learning techniques could allow to automate the detection of animals in these aerial images.

This project aims to implementing such algorithms in order to investigate the feasibility and potential benefits of combining machine learning and UAVs for animals detection. This study uses an image dataset acquired in the Kuzikus Wildlife Reserve in Namibia and a ground truth acquired through crowd-sourcing. The machine learning techniques involved include Bags of visual Words, exemplar SVMs and active learning. The promising results show that recall rates in the range of 60 to 80% are possible, if a low precision (5 to 20%) is accepted. The study also discusses parameters related to the data acquisition, such as the image resolution and the time of the day when the images are acquired.

Résumé

Les savanes semi-arides sont menacées par des changements dans le fragile équilibre entre les pluies, les feux de brousse et la pression pastorale exercée par le bétail et les herbivores sauvages. Afin d'éviter l'avancement des broussailles ligneuses et le déclin des herbes pérennes, les éleveurs et gardiens de parcs doivent être attentifs à maintenir un nombre d'animaux en adéquation avec le fourrage disponible. Ainsi, estimer les populations d'herbivores des grandes fermes et parcs naturels est une étape importante dans la gestion des savanes semi-arides.

Les méthodes traditionnelles pour le comptage des animaux – telles que les comptages par transectes ou par marquage et recapture – sont trop chères et trop laborieuses pour être utilisées de façon régulière. Dans ce contexte, les véhicules aériens sans pilotes (UAV) semblent être un outil intéressant pour la détection et le comptage des animaux. Ils sont faciles à déployer, moins onéreux et assurent une meilleure sécurité. L'inconvénient est qu'il est difficile d'interpréter manuellement le grand nombre d'images à très haute résolution (VHR) produites par les UAVs. Les avancées récentes en apprentissage machine pourraient permettre d'automatiser la reconnaissance d'animaux dans les images aériennes.

Ce projet a pour but d'implémenter un tel système afin d'étudier la faisabilité et les bénéfices de l'utilisation conjointe d'imagerie par UAVs et d'apprentissage machine pour la détection d'animaux. Elle se base sur des images aériennes acquises dans la Kuzikus Wildlife Reserve et sur une réalité-terrain obtenue par crowd-sourcing. Les méthodes d'apprentissage machine employées dans cette étude sont notamment les suivantes : bag of visual words (BoVW), exemplar SVMs, apprentissage actif. Les résultats encourageants montrent que les méthodes implémentées permettent d'obtenir un taux de rappel entre 60 et 80%, pour autant qu'une précision relativement faible soit acceptée (de l'ordre de 5 à 20%). Cette étude discute également des paramètres liés à l'acquisition des images, comme la résolution des images et l'heure à laquelle elles sont acquises.

Acknowledgment

I would like to thank Prof. Devis Tuia who first introduced me to the world of remote sensing and computer vision. His enthusiastic way of sharing his passion and the special attention he pays for team-building generate a very positive environment of work around him.

My gratitude extends to the whole Multi-modal remote sensing group of the University of Zürich who hosted me during my project. Special thanks to Dr. Michele Volpi for the help with ESVM, to Diego Marcos Gonzalez for the discussions about rotation-invariant words, and to Shivangi Srivastava and Benjamin Kellenberger for showing me how it is like to start a PhD. They are a great team and making research among them is a privilege!

I would like to thank Dr. Stéphane Joost for encouraging me to better study the context of my work and for the advice on methodology.

Thanks to his personal experience in Kuzikus, Matthew Parkan gave me a lot of precious information and insights about the SAVMAP project. He also helped me in many occasions for more than a year and deserves many thanks.

I am also grateful to Dr. Friedrich Reinhard, manager of the Kuzikus Wildlife Reserve, for his practical explanations about conservation in Namibian wildlife reserves and for giving a sense of reality to the whole project.

Finally my gratitude goes to the SAVMAP consortium who provided the image datasets, and to MicroMappers (and the crowd) who created the ground truth.

Contents

1. Introduction

Semi-arid savannas are specific ecosystems that develop under hot semi-arid climates. They experience very hot summers and mild to warm winters. They have a short wet season, but do not receive sufficient rainfalls to develop into tropical savanna. Semi-arid savannas can be found in the Sahel, southern and eastern Africa, south-western U.S.A. and parts of India and Australia.

(Trodd & Dougill, 1998) recognizes three driving forces that shape semi-arid savannas: rainfall, fire and grazing. Periodic rainfalls are followed by a rapid increase in grass cover and production of biomass, while fires episodically eliminate the entire vegetation.

In contrast to these short-term effects, grazing - the third driving force – can lead to long-term ecological changes. Where wildlife is supplanted by cattle and sheep, the fragile equilibrium between vegetation and grazing pressure is often modified. As a result, the grass species composition changes and woody vegetation becomes predominant over large areas (Walker, Ludwig, Holling, & Peterman, 1981), (Dugill, 1995). Known as bush encroachment, this degradation of the ecosystem is recognized as a major problem in numerous semi-arid savannas with pastoral activities.

African wildlife reserves and conservation parks suffer from yet another issue. As the market of ivory developed in western countries and recently in Asian countries, large fractions of the elephant and rhinoceros populations were killed by poachers. Until now, governments have not been able to eradicate poaching and the conflicts between park rangers and poachers have led to several human deaths, in addition to threatening the survival of the concerned species.

In order to achieve a sustainable management, land owners, farmers and conservation parks need to take these issues into acount and adapt their method. As new technologies have emerged in the field of remote sensing, the use of space- and airborne images for the purpose of conservation has shown promising results. For example, land cover changes (Trodd & Dougill, 1998), (Ringrose, Vanderpost, & Matheson, 1996) and animal monitoring [5] is made possible over larger scale with less effort. In the recent years, unmanned aerial vehicles (UAVs) and software for image analysis have become easier to use and commercially available at lower prices. Conservation parks and land owners have started to show interest in integrating them in their toolkit.

This report is structured as follows: section 1 gives further explanations about the conservation challenges formulated above; section 2 indicates how space and aerial imagery is or could be used, and formulates the research questions and objectives of the study; section 3 provides a literature review of techniques for animals census; section 4 describes the methodology used in this study for automated animal detection; section 5 describes the dataset and the conservation reserve from where it originates; section 6 explains the machine learning methods involved in this study; section 7, 8 and 9 give the setup of the experiments, the results and the discussion respectively; and section 10 concludes and indicates the perspectives for further works.

1.1 Carrying capacity

The amount of wildlife that a parcel or a park can sustain (referred to as “carrying capacity”) heavily depends on the grass production, which serves as food for wildlife, cattle and sheep. A distinction is made between annual grass and perennial grass. The former are more affected by meteorological conditions, while the latter show better resilience to drought and offer a more stable source of grass. They form the backbone of the food system and are

favored by the land managers (Reinhard, 2016). The third type of vegetation encountered in the semi-arid savanna is woody vegetation, which includes bushes and shrubs. They do not serve as food for the large herbivores and are therefore undesirable.

In natural conditions, the wildlife population is regulated not only by the food availability, but also by their predators. However when wildlife was progressively replaced by cattle and sheep in the late 1800s, the fragile equilibrium of many semi-arid savannas was modified (Walker et al., 1981). Protected from predators, cattle and sheep populations grew until exceeding the carrying capacity, and then decreased to a lower number. During this process, the vegetal species composition was affected: perennial grasses declined at the benefit of annual grasses and woody vegetation. The carrying capacity was hence reduced and the grass growth is now more prone to drought.

We now understand that balancing the number of browsing animals and the grass availability is a crucial management step. If the number of animals is above the carrying capacity, land managers take actions to control the herbivore populations. They can sell or relocate animals, or allow hunting to decrease the population in a controlled manner (Reinhard, 2016).

In order to evaluate the current grass availability and estimate its future evolution, land managers must monitor the grass biomass and its growing stage. According to the land managers of Kuzikus Wildlife Reserve there is space for improvement on this point:

“Ideally, grass biomass and composition is estimated at the end of the growing (rain) season (January - May) and feeding regimes are adapted accordingly for the rest of the non-growing (dry) season (May - December). Such vegetation analysis, although the basis of a data-driven farming strategy, is still conducted by only a very small minority of farmers. Most rely on "experience" and hope.” (Reinhard, 2016)

Once the carrying capacity has been estimated, the next step is to determine the number of animals on the site and verify that this number can be sustained. Several methods have been used to estimate animals’ populations.

1.2 Bush encroachment

The Desert Research Foundation of Namibia uses following definition of bush encroachment:

“the invasion and/or thickening of aggressive, undesired woody species resulting in an imbalance of the grass:bush ratio, a decrease of biodiversity and a decrease in carrying capacity”(Seely & Montgomery, 2009).

Bush encroachment has become a major issue in Namibia. According to the same foundation, “over 26 million hectares of woodland savannas have suffered loss of productivity and carrying capacity by at least 100%”. In Kuzikus, bush encroachment affects around 10% of the park and the land managers pay great attention to this issue (Reinhard, 2016).

Bush encroachment is related to all driving-forces described above. Fires periodically burn the entire vegetation and grasses recover more rapidly than bushes and shrubs (Roques, O’Connor, & Watkinson, 2001).

The hydrology of the ecosystem is also of great importance regarding bush encroachment. Grasses only take their water from the top soil layers, and their presence reduces run-off and erosion, and increases infiltration in the same top layers. In contrast, the

roots of bushes extend deeper and have access to the subsoil water reserves. (Walker et al., 1981) explains that woody plants and trees influence the subsoil water content by intercepting rainfall and leading it more quickly to the deep soil layer through preferential flow channels along their stems and root systems. In the end, this favors the development of shrubs: “If the reduction in infiltration is not too drastic, then, as the amount of grass declines, less water will be taken up from the top soil by plants, an so more will penetrate to the subsoil”.

Browsing influences bush encroachment through its effect on fire and the hydrology: A low browsing and trampling pressure increases the frequency of fire (Roques et al., 2001), hence preventing bush encroachment. In overgrazed areas, water tends to percolate more quickly to the deep soil layers since it is less used by grasses in the top layers. This favors the bush encroachment (Reinhard, 2016).

1.3 Poaching

While several species suffer from illegal hunting, the rhinoceros are particularly vulnerable. Africa is home of two rhinoceros species: the white rhinoceros and the black rhinoceros. From 500'000 black rhinoceros in the 1900s, their population went down to 65'000 in 1970, due to the destruction of their natural habitats and excessive hunting. It reached its lowest number in 1993, with only 2'300 rhinoceros surviving on the continent. Efforts brought by governments and wildlife protection association helped reverse the tendency and save the species (“Poaching Statistics,” n.d.).

Today, the population of black rhinos is back to 5'000 – 5'500 individuals, and there are around 20'000 white rhinos on the continent. But both species are still endangered and the increase in poaching since 2008 is alarming. One reason is that the price of ivory has increased a lot as ivory products found a market in Asia (“Rhino Poaching Statistics,” 2015).

"With a kilo of rhino horn selling for around $60,000 (£35,000), a big specimen can fetch $250,000," Mr Breare (Ol Pejeta's chief commercial officer) explained in an interview to the BBC (Wall, n.d.).

The number of rhinoceros killed by poachers is rising accordingly. Figure 1 shows statistics of recent years for all rhinos (black and white) in South Africa. According to (“Rhino Poaching Statistics,” 2015) Namibia does not regularly report the number of kills. The different news outlets they cite report 103 deaths between 2005 and 2014, 25 for 2014, and 80 for 2015.

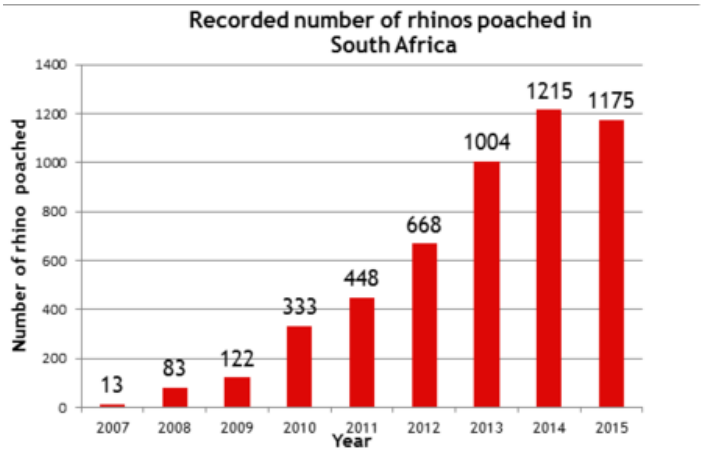

Figure 1 : Recorded number of rhinos poached in South Africa [15]

2. Using space- and airborne imagery for conservation in semi-arid savanna

Aerial or space born imagery can provide useful information regarding the issues described above. In the following, their use for carrying capacity estimation is briefly presented. Then the focus is set on animal monitoring, and the objectives of this study are formulated.

2.1 Estimation of the carrying capacity

By studying the land cover, one can estimate the carrying capacity of an area and assess bush encroachment. Depending on the platform used for image acquisition (satellite, airplane or unmaned aerial vehicle (UAV)), the problem can be studied at different spatial and temporal scales.

An estimate of the carrying capacity can be obtained by mapping the area to land cover classes, and assigning a specific carrying capacity to each class.

Typically, aerial images are classified in such classes as “bare soil”, “grassland” or “forest” and the surface area covered by each class is then computed. Acquisition in specific wavelengths, such as the near infra-red (NIR), allows assessing not only the presence of vegetation, but also its chlorophyll content, which gives precious indication about the plants health.

These methods are nowadays well established. However the cost of image acquisition depends on the desired spatial resolution: freely available satellite images typically have a resolution of 30m, while commercial satellites images have a resolution up to 20 cm. The legislation of many countries does not allow for a higher resolution, so that the limitation is more legal than technical.

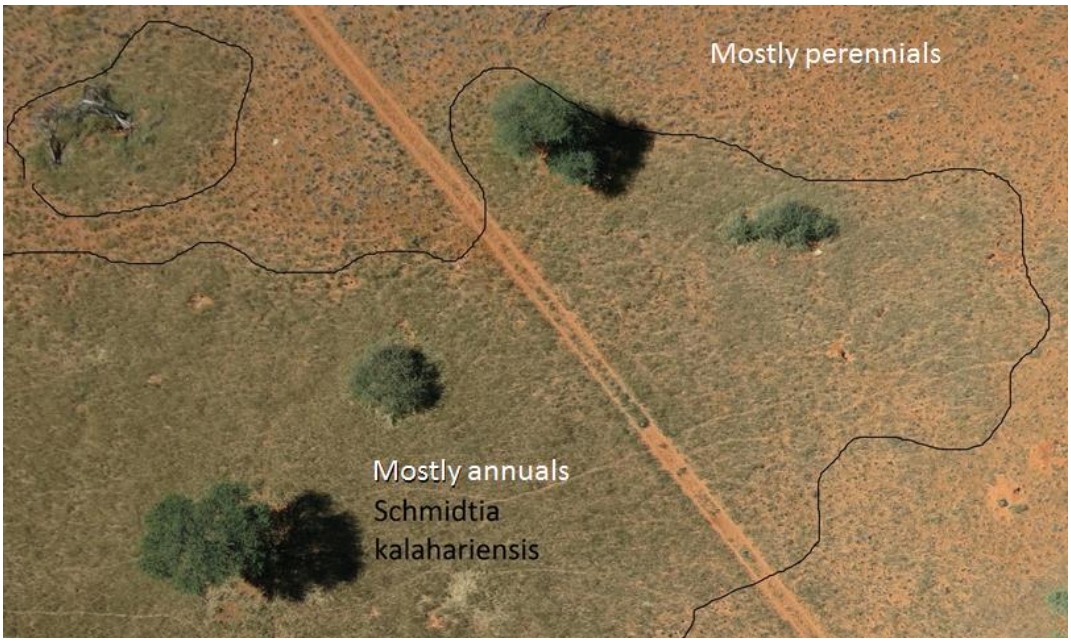

Figure 2 : Annual and perennial grasses viewed from a UAV. Example provided by Kuzikuz Wildlife Conservation Park.

On freely available, 30 m resolution images, a single pixel usually contains several landcover types: in the Namibian savanna for example, it can cover a patch of grass, some shrubs and several acacias as well. Here airborne imagery would be useful to refine our

understanding of the satellite images: individual trees and precise amount of grass and shrubs can be identified in the very high resolution (VHR) images, as shown in

Figure 2. A correlation between the classified VHR images and the satellite images could be used to acquire a precise ground truth for the satellite images, where new landcover classes corresponding to precise information about the food availability for grazing animals could be defined. Using VHR images, few calibration flights could then allow to estimate the carrying capacity based on freely available space born images with an improved precision.

2.2 Animals monitoring

Balancing the carrying capacity with herbivores populations also needs an estimate of their number. As described in the following, the different methods that are traditionally used are time-consuming and do not scale very well to large areas. As the acquisition of very-high resolution images becomes easier and less expensive, conservation parks and farmers are showing a growing interest in these technologies and envisage including them in their toolkit.

While the resolution of space borne sensors does not allow to recognize individual animals, airplanes and UAVs can produce images with very high resolution (typically less than 20cm) where animals are recognizable and can be counted. UAVs are also seen as a promising tool against poaching. Patrolling UAVs would make it possible to locate injured animals and rescue them before they die. Used as a surveillance system, UAVs could help to understand poachers' modus operandi, allow quick interventions and discourage poachers from committing illegal acts.

2.3 Methods for animal counting

There are numerous methods for wildlife estimation. First the most common techniques are presented according to a classification based on the statistical methods involved. Then, the advantages and drawbacks are discussed from a practical point of view.

From a statistical point of view, wildlife estimation can be subdivided into the following groups:

Complete counts are attempts to count the entire population. This can be achieved by a drive approach in enclosed areas, where people form a line and cross the whole area, counting all animals passing through the line. Capturing the entire population is also a form of complete count, which is possible for small populations of certain species only. These methods require large manpower and are expensive and laborious. In general they are more suitable for cattle and are seldom used in the context of wildlife conservation.

Incomplete counts are methods where only part of the population is counted, and the total population is extrapolated by the use of statistical methods. For example, a total count can be achieved over a small area, and with the assumption that the density of animals does not vary much in space, the total population can be estimated.

Transect count is another example of incomplete count: a fixed route (transect) is selected and a person walks along the transect and notes the distance to and the direction of all animals that he observes. This method is very popular and extensive work has been done on the statistical methods to retrieve the total population, as presented in (Burnham, Anderson, & Laake, 1982).

Finally, aerial counts can most often be regarded as incomplete counts similar to transect counts or drive count. They allow covering a larger area but should rely on the same statistical approach, as a complete count is usually not achievable, even from the sky.

Indirect counts use any signs of the presence of the animals to estimate their population. It can be easier to count nests than birds, or scats than deers. These methods can only be effective if there exists a strong relation between the number of signs and the population size, and if this relation can be easily derived.

Mark-recapture analysis is based on a different idea and has been extensively used for game animals. A subset of the population is captured and marked, and then released. In a second time, a group of animals is again captured and the fraction of marked animals in this group is used to determine the population size P, with following formula:

$$P = \frac{C\ M_1}{M_2}$$

Where M_1 is the number of animals marked and released the first time, M_2 is the number of marked animals in the second catch, and C is the total number of captured animals in the second catch

An assumption of this method is that every individual is captured with the same probability, meaning that animals do not learn to avoid traps after their first capture.

Camera traps can be used in the same framework. Instead of capturing and marking the animals, this technique involves a motion detector that triggers a camera whenever an animal passes by. A similar statistical approach can be employed if each individual can be identified. More recently, (Rowcliffe, Field, Turvey, & Carbone, 2008) proposed to use a model of the rate of contact between animals and cameras, so that recognizing each individual is not necessary. In this manner, camera traps can be used outside the framework of mark-recapture techniques.

From a statistical point of view, two sources of errors are generally recognized for all these methods (except for a theoretical complete count) (Caughley, 1974): first, due to the randomness in the spatial distribution of animals, the count has a variance – two consecutive counts do not yield the same result. Second, not all animals can be counted or capture with similar probability. A fraction of them is hidden by the vegetation, cannot be distinguished from the background environment, or simply escape to the attention of the crew. In the case of mark-recapture techniques, the assumption that animals are captured with a same, constant probability does not hold in all cases. This results in biased counts where the mean over repeated counts usually underestimates the actual population.

Let’s now consider the different methods from a practical point of view and compare their advantages and drawbacks.

Drive counts require a lot of man power and are labor intensive. In addition, they scare the wildlife and may damage the habitats, with heavy consequences for the population.

Transect count is also time-consuming and its success depends on the choice of the transect. It has to be representative for the whole area in terms of habitat. Depending on the terrain and vegetation cover, it may be difficult to follow a predefined path and find the same path during a later survey. The result of a transect count also depends on the skills of the person involved, leading to biased comparisons if different people are involved. On the other hand, transect count is probably the least expensive technique and the most simple in terms of logistic.

Both techniques can be used from a helicopter. This allows covering larger areas and operating in areas that are not accessible from the ground, such as swamps or mountainous regions. But this is achieved at the expense of a higher price, a more complex logistic and higher safety risks. A competent pilot and a copilot trained to count are needed. In general the helicopter increases the disturbance to the wildlife.

Camera traps techniques are influenced by the location and calibration of the cameras. Usually they produce a very large number of pictures that must be collected and analyzed in a time-consuming phase. Depending on the environment, humidity may cause malfunction of the camera, which is also limited by battery life-time. However, this technique has the advantage of being less intrusive, and gives the best chances to observe animals without disturbing them.

2.4 Using aerial imagery for animal counts

In this respect, aerial imagery with UAVs provides a new way of estimating wildlife populations. In short, a UAV can automatically fly in transects over a predefined area, and its embedded camera takes pictures perpendicular to the ground. The frame rate is such that there is an overlap between each image, ensuring that the whole area along the transect is captured. Then, animals can be counted on the pictures.

Even though the statistical approach is similar to the well-studied transect counts, the use of UAVs can change many practical aspects of the task. Compared to helicopters, UAVs require a much shorter training to pilot them and represent a lower financial investment. Their use is safer, since the pilot stays on the ground and away from potential conflicts with poachers. It is also more flexible than helicopters, because the logistic is easier. Finally, the disturbances to the wildlife and the environment are drastically reduced.

One major drawback is the short autonomy of the batteries that limits the area covered by a single flight. This issue will probably become less important in the coming years, as technological improvements in this regard are very likely. Another drawback is that UAVs are prone to accidents, due to technical failures or improper use.

2.5 Automated detection

The previous section highlighted the advantages of using UAVs for animal monitoring. This section introduces the challenging task that comes after data acquisition: extracting useful information from the huge amount of images produced by surveying UAVs.

Traditionally, images are visually interpreted one by one by a human observer who identifies animals and other objects of interest. This is an exhausting and time-consuming task that must be repeated for every new data acquisition. During their experimental study on detection of animals and poachers, (Mulero-Pázmány, Stolper, van Essen, Negro, & Sassen, 2014) reported that “On average, an observer with a computer needed around 45 minutes to process a 500 pictures flight, which is the usual number of pictures taken per flight.” Another drawback of this method is that the accuracy of detection depends on the human observer’s skills and the time spent at the task, making it difficult to compare results from different studies.

The field of computer vision provides many machine learning techniques for objects detection. Applied to aerial animal surveys, these techniques can be used to automatically localize and count animals in the images. If the same precision is achieved compared to

human observers, automatic detection drastically reduces the time spent to make sense of the aerial images, thus greatly improving the overall benefits of using UAVs.

While the computer vision community was primarily focused on the detection of hand-written digits, pedestrians, human faces or vehicles in natural images, recent datasets include up to several hundreds of classes of objects found in indoors and outdoors scenes. In remote sensing, machine learning has been extensively used for aerial and satellite image classification, typically to produce landcover and landuse maps. The task of object detection remains less common.

A successful animal detection usually requires a very high resolution, which excludes free satellite images. Regarding the spectral properties of the cameras, RGB cameras have been predominant because other wavelengths are less suited for visual interpretation. However, studies based on thermal images have been done, as mentioned in the literature review (section 3).

2.6 Research questions and objectives

In this context, building a system for automatic detection of animals in the Namibian savanna appears as a challenging and promising task, at the intersection of data-driven wildlife monitoring and anti-poaching efforts.

The advantages of aerial images and their applications in Namibian wildlife reserves have been discussed. It appears that the use of UAV imagery could provide precious information to estimate wildlife population and mitigate poaching. However the benefits of UAV imagery will only be substantial if a system for automatic detection of animals can be developed.

In light of this, the present study will attempt to answer the following questions:

- Is it possible to build a performant system for animal detection with current computer vision techniques?
- What precision and recall rates can be expected?
- Regarding the data acquisition, are simple RGB cameras suited for this task?
- What recommendations can be made about the flight parameters, such as the height of flight above ground and the time of the day?

The objective of this study is to answer these questions by developing and using an automatic system for detection of animals in the Namibian savanna, and based on the results of this system.

As mentioned, UAVs could be employed for other tasks than animal detection as well. However, to limit the extent of this project, it has been decided to focus on animal detection.

3. Literature review

This literature review provides examples where the methods described in section 2 for animal census have been used in practice.

In 1997 an important gorilla census in Bwindi Impenetrable National Park was conducted. The method used and reported by (McNeilage, Plumptre, Brock-Doyle, & Vedder, 2001) is a good example of both total and indirect count. Six teams of at least 3 to 4 people crossed the 331 square kilometers park in a methodic manner, so that no more than 500 to 700m separated adjacent paths. They reported all signs of gorillas, such as nests and dungs, and with these indices the experienced team members were able to determine the number of individuals in each group, and to distinguish groups based on their sex and age composition. The park was searched in such dense manner that the authors believe that very few groups, if any, were missed or counted twice. Their results indicated that the population amounted to 292 gorillas. The study also provided much side information such as the number and composition of the groups and the disturbances by human activities.

(Silver et al., 2004) used camera traps to conduct a large survey of jaguars over five sites in Belize and Bolivia in 2003. They deployed a total of 160 cameras for about 60 days. They identified each individual jaguar thanks to the pattern of its fur and there were between 7 and 11 individuals per site. This allowed them computing population densities through capture / recapture analysis. According to the author this was the first successful measurement of jaguar densities. But they deplored that this method was very expensive due to the high cost of the cameras and the high requirement in trained field assistance. In several cases accessing the sites required to open new trails, with the risk of facilitating illegal hunting and logging.

Regarding aerial counts from airplanes, using line transects, many studies discuss the statistical approaches to account for visibility bias and to model the detection probability as a function of the distance to the animal ((Caughley, 1974), (Quang & Becker, 1997)). (Pollock & Kendall, 1987) provides an interesting comparison of several methods to deal with this issue.

(Marsh & Sinclair, 1989) explain in details their survey procedure for the census of dugongs in northern Australia. The airplane flew at around 140m above sea at a speed of 185 km/h. Two observers sat on each side of the aircraft and surveyed a 200m wide strip. The results of both team members sitting on a same side could be confronted in order to apply a form of mark / recapture analysis to model the visibility.

(Linchant, Lisein, Semeki, Lejeune, & Vermeulen, 2015) provide a recent literature review on wildlife monitoring using UAVs. They distinguish three types of animals for which UAVs have been used: birds such as gulls (Grenzdörffer, 2013) and geese (Chabot & Bird, 2012), marine mammals such as dugongs ((Hodgson, Kelly, & Peel, 2013), (Maire, Mejias, & Hodgson, 2014)), and large terrestrial mammals such as elephants and orangutans (Koh & Wich, 2012) , rhinoceroses (Mulero-Pázmány et al., 2014), and deers (Israel, 2011).

These attempts were based on various sensor types, including RGB and thermal cameras. Most often, small fixed-wings UAVs were used. The success of these studies largely depends on the environment (open terrain or dense forest, fields, beaches) and the contrast between animals and the background. The behavior of the animals (living in flocks or herds, staying in open terrain or below shelters) also is of great importance (Linchant et al., 2015).

Regarding poaching, UAVs have already been used in some conservation parks. In the Province of KwaZulu-Natal in South Africa, Air Shepherd has deployed in several parks a system combining UAVs with long flight autonomy and thermal cameras. According to this organization, the number of rhinoceros and elephant kills was reduced by 60% over a two-year period after implementation of their system. Their effort are now directed toward Kruger National Park, which is home to around 65% of the worldwide rhinoceros population ("Where We Fly," n.d.). The press has reported similar efforts in other national parks, such as in Ol Pejeta Conservancy, Kenya (Wall, n.d.).

(Mulero-Pázmány et al., 2014) have conducted experiments to assess the use of remotely piloted UAVs to monitor poaching activities. Using three different sensors (RGB pictures, RGB videos and thermal videos) they surveyed rhinoceroses, fences and people mimicking poachers. The acquired images were reviewed by human observers and provided encouraging detection rates. Three different approaches to integrate UAVs in anti-poaching work are then proposed and technical aspects are discussed. Unfortunately, the number of such studies is still very limited in the literature.

Even though many studies exist on the use of UAVs for wildlife monitoring, only few have implemented an automatic detection. In their review, (Linchant et al., 2015) explain that most attempts concerned birds detection ((Grenzdörffer, 2013), (Chabot & Bird, 2012)).

Interesting results were obtained by (Maire et al., 2014) for the automatic detection of dugongs. Thanks to data augmentation and hard negative mining, they could train a convolutional neural network that gave promising results.

Finally, it can be mentioned that WIPSEA, a company based in France, offers commercial solutions for automatic detection of animals ("Wipsea," n.d.). According to the description on their website, in many cases they still need to adapt existing software (both the algorithms and the user interface) to the special requirements of a new task. Their solutions also integrate the detections to a geographical information system to make spatial analysis and produce maps.

4. Methodology

The detection system will be developed and tested on a dataset from Kuzikuz Wildife Reserve in Namibia. Instead of modeling the visual appearance of animals, the chosen approach relies on machine learning methods that integrate a large amount of data and learn the visual traits of the animals from the data.

4.1 Detection: the pipeline

In general, a system for automated detection based on machine learning can be subdivided into the following steps:

- Data acquisition
- Ground truth acquisition
- Objects proposals (segmentation):
- Features extraction
- Classification

The present study focuses on the last three steps, as the data and ground truth were provided by the SAVMAP consortium and MicroMappers. Post-processing of the ground truth was however necessary and is briefly described. Figure 3 presents the general pipeline.

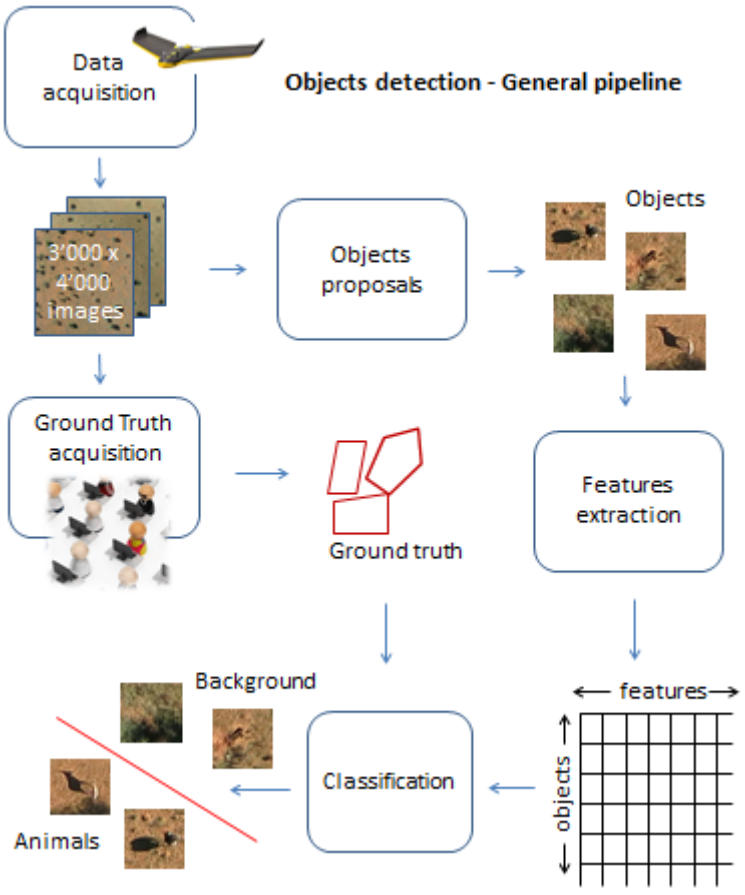

Figure 3 : General pipeline for objects detection

Data acquisition refers to the use of a certain platform and sensors to acquire the photographs. In this case, a fixed-wing UAV and an RGB camera were used.

Ground truth acquisition is the process of defining objects of interest in the images used to train and test the model. In other words, the regions of the images containing objects of interest must be annotated, usually by drawing a polygon on the regions and labeling them with the name of the object. The set of these annotations is called ground truth. Because it is a time-consuming task, in the present case crowd sourcing was used.

Objects proposals, also referred to as segmentation, is the process of recognizing interesting parts of the images that may correspond to the real-world objects to be detected. The output of this block is a set of objects that ideally has the following characteristics:

- The objects boundaries are precisely defined
- For each instance of the real-world objects of interest in the images, there is a corresponding segmented object, so that no real-world object is missed at this stage.

In this study, a simple method based on color and gradient intensity is used.

Feature extraction defines a set of features (also called attributes or descriptors) that are used to recognize objects of the same class and distinguish them from objects of different classes. The features can be any type of categorical or numerical variable; however most of the classification methods preferably work with real values. Once a set of features is defined, the features are computed for each object and the concatenation of these values forms a vector that describes the object.

In this study, color histograms and bags of visual words are used as features.

Classification is the process of assigning a class label to each of the objects, based on the value taken by the features. The idea is that in the n-dimensional feature space, where n is the number of features, objects form clusters according to their class. In supervised classification, a set of objects with known class label (ground truth) is used to train a classifier. The classifier learns the boundaries between classes in the feature space, and is then able to predict the class of any new, unlabeled objects.

The classifier chosen in the present work is a support vector machine (SVM). The use of exemplar SVMs and Hard negatives mining is also explored.

For each of these blocks there exist a great number of different methods and in many cases it is unclear which method performs best. Here, not only the expected performance but also the interest of the author has guided the choice of the methods.

The methods used in this study are implemented in Matlab (*Matlab*, 2015). The Matlab Image processing toolbox (*Matlab Image Processing Toolbox*, 2015) has been extensively used. Besides, the LIBSVM library (Chang & Lin, 2011) is used for the SVM classification. The implementation of Exemplar SVMs and Hard negatives mining are inspired by the code provided by (Malisiewicz, Gupta, & Efros, 2011).

5. Study site and data set

The animal detection system proposed in this study is based on a dataset of aerial images from Kuzikus Wildlife Reserve. The images were acquired during two campaigns conducted by the SAVMAP consortium in May 2014 and May 2015. In order to build a ground truth, MicroMappers made a crowd-sourcing campaign to let volunteers identify and tag animals in the RGB images of 2014.

5.1 Kuzikus Wildlife Reserve

Kuzikus is located on the edge of the Kalahari in Namibia. The Kalahari is a semi-arid sandy savanna that extends over Botswana, South Africa and Namibia and is home of a large variety of animals, including many large mammal species.

From the beginning of 20th century and until 1980 Kuzikus was a cattle and sheep farm. While the region is still largely dominated by this activity, Kuzikus has been progressively restored into a wildlife reserve since 1964. Todays it is a private, state- acknowledged nature reserve that combines habitats conservation and wildlife protection, and demonstrates that tourism, education and research can provide and alternative and sustainable income for several families. The reserve now offers lodges for tourists and several scientific studies are being led.

The reserve extends over 103 km2 (10'300 ha) and is the home of rich and abundant wildlife:

“The vast diversity of free-living wildlife (most is conservation - dependent in IUCN red list) is the major attraction of Kuzikus: there are over 3000 individuals from more than 20 larger animal species such as the Common Eland (*Taurotragus oryx)*, the Greater Kudu (*Tragelaphus strepsiceros)*, the Gemsbok (*Oryx gazella)*, the Hartebeest (*Alcelaphus buselaphus)*, Gnu (*Connochaetes gnou and C. taurinus),* the Blesbok (*Damaliscus albifrons)*, the Springbok (*Antidorcas marsupialis*), the Steenbok (*Raphicerus campestris)*, the Common Duicker (*Sylvicapra grimmia)***,** the Impala (*Aepyceros melampus)*, the Burchell’s Zebra (*Equus quagga burchellii*), the Ostrich (*Struthio camelus australis* and the Giraffe (*Giraffa camelopardalis giraffa)*.”(“Kuzikus - Wildlife Reserve Namibia,” n.d.)

Around 200 bird species, 44 mammals, 50 reptiles and 100 insects species have been observed. Scorpions and spiders also add up to this great diversity (Kuzikus Wildlife Reserve, n.d.).

As part of a breeding program, the iconic and most endangered Black Rhinoceros (Diceros bicornis) was reintroduced in Kuzikus. From a global population of around 400'000 rhinos in the beginning of 19th century, extending over all savannas of Africa, only 2'000 individuals were left in 1994. Poaching is still a predominant threat, as described in the following section.

While the region is most famous for its fauna, the flora is also unique and gets a considerable attention from the land managers, since it is at the basis of the food chains.

Only 6 tree species are found in Kuzikus, with the Camelthorn (*Acacia aerioloba*) being the most represented. The vegetation mainly consists of grass, herbs and shrubs.

5.2 Dataset

5.2.1 Images

The images used in this study were acquired with a UAV and an RGB camera. The main features of the image acquisition system are presented in Table 1.

Table 1

Source	SAVMAP Consortium (“SAVMAP,” n.d.)
Product type	RGB images with dimensions 3’000 x 4’000 pixels
Platform	eBee – light UAV commercialized by Sensefly (“eBee: senseFly SA,” n.d.)
Sensor	RGB camera: Canon PowerShot S110
Spatial resolution	4-8 cm
Spectral resolution	3 large bands in the red, green and blue domains
Radiometric resolution	24 bits

Other cameras have also been used, mounted on the same eBee.

Table 2 indicates the number of flights made with each type of sensors and the total number of images acquired during these flights, for 2014 and 2015:

Table 2

	May 2014		May 2015	
	# images	# flights	# images	# flights
RGB	9’734	55	4’838	25
NIR	4’006	24	1’993	16
RE	539	3	0	0
MS	427	40	1’363	7
TIR	0	0	31’368	9

The present study only uses the RGB images. The practical reason is that there exists no ground truth for the other types of images. Indeed, identifying animals in false-color images would be a very challenging and tedious task for the crowd.

Since RGB cameras are less expensive and more commonly used, demonstrating that automated detection is possible without relying on more sophisticated captors would be an interesting result.

Figure 4 presents a map of the areas where RGB acquisition was made in May 2014.

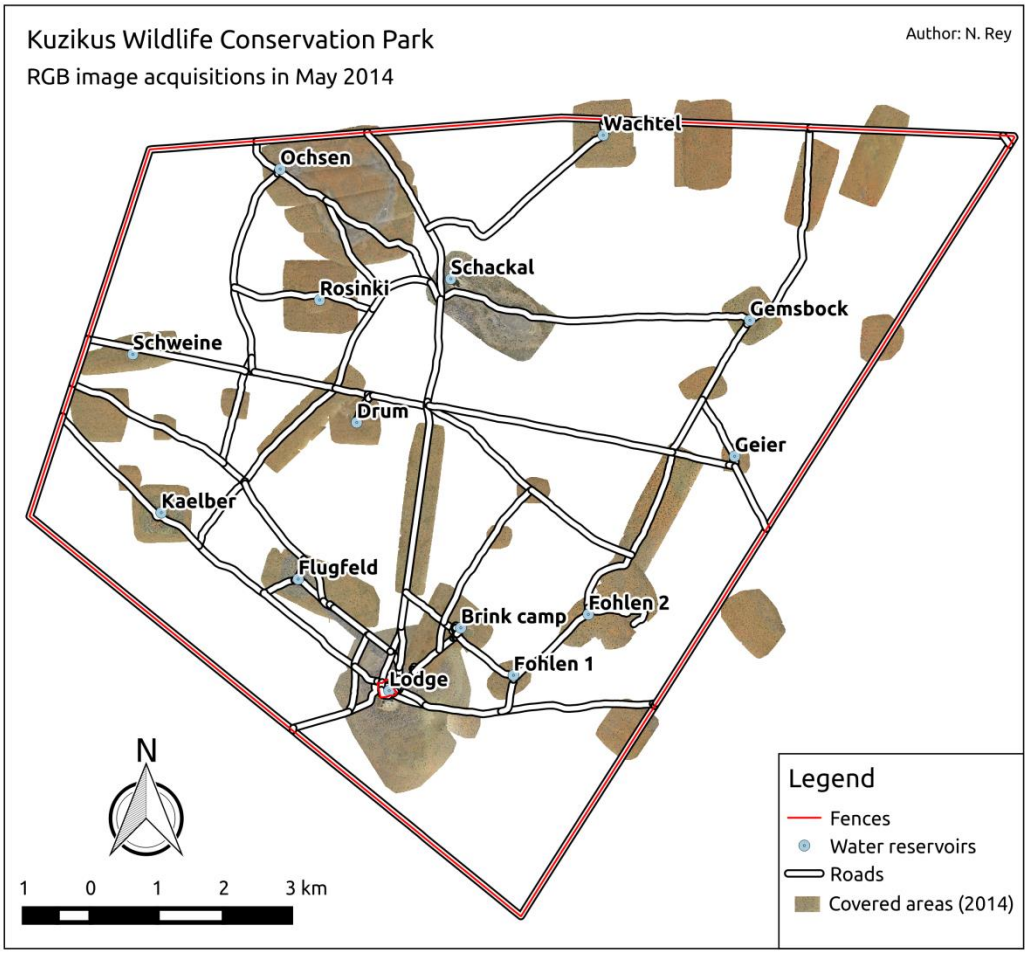

Figure 4 : Map of Kuzikus Wildlife Conservation Park and areas covered by the 2014 RGB dataset

5.2.2 Crowd-sourced Ground Truth

MicroMappers has delivered a crowd-sourced ground truth for the 2014 dataset, in which animals have been tagged by volunteers. A total of 6'500 RGB images (of size 3'000 x 4'000 pixels) have been analyzed by the crowd and each image was shown to at least 3 volunteers. The task was to draw a polygon around each animal found in the images, without distinction between species. Signs of animal presence such as Aardwolves' holes or termite mounds should not be reported.

7'474 polygons were drawn by the crowd in a total of 654 images from 5 different flights. After a specific merging of overlapping polygons and removal of the unconfirmed ones (i.e. objects tagged by one single volunteer, see section 6.1 Ground Truth post-processing for details), the number of tagged animals is 976. It should be mentioned that the number of unique individuals is less, since the same animal could be observed in several consecutive, overlapping images. In this case the animal is viewed under a different angle and often a different pose.

6. Methods

6.1 Ground Truth post-processing

The ground truth delivered by MicroMappers, obtained through a crowdsourcing campaign, is the set of all polygons drawn by the crowd. This means that several polygons are usually overlapping on an animal, since the same image was shown to several volunteers. Each user drew a different polygon with more or less precision.

On the other side, there are many locations tagged by a single volunteer, resulting in lonely polygons. These are likely to be erroneous annotations on objects that are difficult to identify, or the result of volunteers who misunderstood the task or worked with little care.

In order to obtain a final ground truth, erroneous polygons must be deleted and overlapping ones must be merged in a way that produces a precise delineation of the animals. For this purpose, the following procedure has been implemented, that looks at each pile of overlapping polygons separately:

Only the pixels covered by at least n tags are considered. In this case, n was set to 2. This allows discarding erroneous polygons. Then, for each pixel, a confidence c is computed as the ratio between the number of different tags covering that pixel, and the total number of tags in the pile the pixel belongs to. Note that two polygons that do not overlap belong to the same pile if there are connected through other overlapping polygons. All pixels with a confidence c below a given threshold are discarded. In this case, c was set to 0.5.

Finally, the remaining pixels are converted to polygons, and the length of the major axis of the polygons is used to discard small polygons. Here, the threshold was set to 20 pixels. Figure 5 illustrates this procedure.

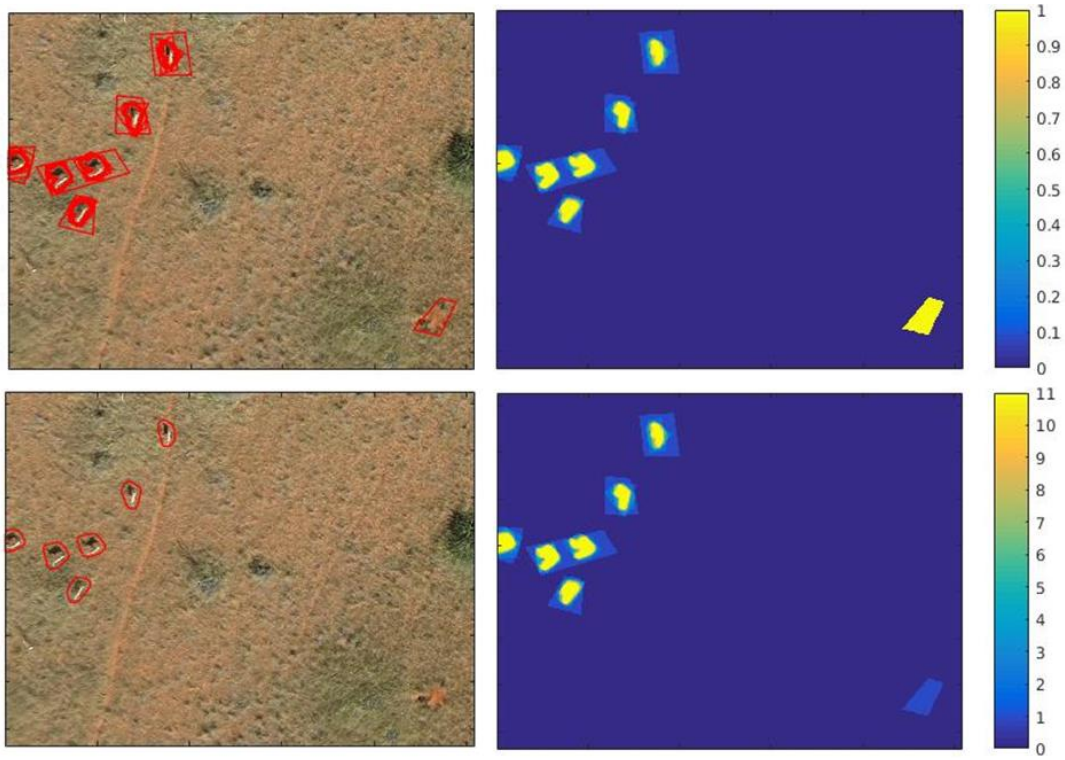

Figure 5 : Post-processing of the tags to build a ground truth. Top left: collection of polygons drawn by the volunteers. Top right: confidence map. Note that the erroneous polygon at the bottom right corner obtains a high confidence. Bottom right: number of overlapping polygons. The erroneous polygon is discarded. Bottom left: final polygons used as ground truth

6.2 Objects proposals

6.2.1 Support

In a detection task, a first step is to define the spatial support of the detection. The spatial support defines how objects of the real world will be represented by pixels or groups of pixels. The detection can be based on:

- **Entire objects:** for example full trees including the trunk and the canopy
- **Regions:** parts of objects, for example canopies without the trunks. A region is generally defined by a homogeneous visual aspect (color, texture). An object of the real world is made out of a variable number of adjacent regions.
- **Patches:** groups of adjacent pixels belonging to the objects. Here the boundary of a region is not defined. Instead, all pixels around a center of interest are considered. The size and the shape of the patches is fixed.
- **Pixels:** individual pixels are considered without taking into account their surroundings.

The advantage of using small spatial units such as pixels or patches is that it avoids the challenge of computing regions boundaries. But the drawback is that making sense of the detections is more difficult: the same real-world object can produce several detections that must be understood as a single object. Also, the number of pixels or patches that can be considered in the image is by orders of magnitude larger than the number of regions or objects.

In this work, patches will be used as spatial support. This choice is driven by the fact that animals of interest all have a relatively similar size, which allows defining a single patch size. Also, it allows considering the texture of the objects while avoiding boundaries segmentation.

6.2.2 Threshold-based objects proposals

Because the images are large (around 3'000 x 4'000 pixels) compared to the size of animals (around 100 x 100 pixels), it would not be efficient to consider all possible patches in the image. Instead, a few hundreds of pixels that are likely to be close to animals are extracted, and the patch surrounding each of these proposed pixels is considered. In contradiction with the definitions above, these patches are called *objects proposals* in the remainder of this text, or proposals for short, even though the spatial support is a patch.

Two approaches can be used separately or combined in order to extract the objects proposals:

Because most of the (real-world) objects, including standing animals, cast a shadow on the ground, the first method aims to find shadows. The image is represented in the HSV color space and binarized based on the *value* band and a heuristic threshold. The centroid of each connected region is retrieved, except for the regions smaller than a minimal area.

Making objects proposals based on the shadows is not sufficient because laying animals or animals located in the shade of a tree do not cast a distinctive shadow. Therefore, another approach is used: a sobel filter is applied to the image in order to locate edges. In the same manner, the image is thresholded based on its sobel value and the centroid of connected regions larger than a minimal area are retrieved. This method proved to be very efficient because many animals have a white fur that produces a significant contrast and sharp

edges. It appears that computing the sobel on the blue band gives the best results: in this band the contrast is high around white and black animals, but low on bare ground and vegetation, which are dominated by red and green colors.

If used in combination, those two methods may produce proposals that fall very close to each other. To remove redundant proposals, a buffer is computed around each proposed pixel, and the centroids of the so formed regions are extracted, where closely located proposals are now merged into a single region.

The choice of the threshold is a trade-off between number of retrieved animals and total number of proposals. An optimal threshold would give a proposal inside each of the tagged regions, while keeping the total number of proposals as low as possible. Indeed, tagged animals that are not retrieved [1] in this step cannot be detected in the following and will reduce the recall rate. Therefore it is important not to miss too many tags at this stage. On the other hand, the number of false positives is expect to increase with the total number of proposals, as the absolute number of misclassification is likely to become greater.

6.3 Features extraction

6.3.1 Histogram of colors

The first type of features used in this study is the histogram of colors. These features are simple to define and compute, yet surprisingly efficient in some applications.

These features are computed over a region centered on the object. The histogram of each band (red, green and blue) is computed over this region and the bin counts are used as features. The bin counts of the three histograms are simply concatenated to form the feature vector.

In this study, histograms were computed over a region of about 25 x 25 pixels for a ground sampling distance (GSD) of 8 cm, and the size of the region was adapted when another GSD was used. Note that it is a good practice to divide the bin counts by the size of the region, so that histograms are comparable even if the size of the region is not constant for all objects.

The histograms were defined with 10 bins, yielding 30 features.

6.3.2 Bag of visual words

The Bag of Word (BOW) is a model used in natural language processing and information retrieval. It describes a text document by the frequency of words without considering their position in the text, thus disregarding any grammatical structure.

This model has been extended to computer vision, where the method is called “Bag of Visual Words” (BoVW) by analogy. This is the second type of features used in this study.

A collection of patches, called visual words, is used to describe the image. The features at a given location of interest are then defined by the frequency of occurrences of the different visual words in the surrounding region.

[1] Here a tag is said to be retrieved if at least one proposal falls into the tag.

The Bag of Visual Words is an efficient model to extract features over images without hand-crafting them, and achieved state-of-the-art performance in many situations. In the very recent years however, convolutional neural networks have become very competitive and outperformed the BoVW in several occasions.

The method

To define the visual words, a few thousands of patches are randomly chosen in the data set. These patches may be square or circular, with a width or diameter of around 20 pixels. They include the three bands red, green and blue. An unsupervised clustering is performed on these patches according to visual similarity, and the center of each cluster is called a "visual word". The collection of these words forms the *vocabulary*.

To extract features over a given image, each pixel is first assigned to its closest visual word. Here is how this assignment is done: a patch centered on the pixel and of the same size and shape as the words is considered. The visual similarity between this patch and each of the words is computed and the most similar word is retained.

Finally to obtain the value of the features at a given location (i.e. of a given object), the region surrounding that location is considered and the number of occurrences of each visual word is counted to build a histogram. Each bin of the histogram corresponds to a visual word, and the bin counts indicate how many of the surrounding pixels were assigned to each of the visual words.

Each bin of the histogram is used as a feature to describe the object or location of interest. Hence the number of features equals the number of visual words.

Algorithm

1. Extract a few thousands patches at random in the data set
2. Perform k-means clustering of these patches. Each cluster center is a visual word
3. For each pixel in the image:
4. Consider the patch centered on this pixel
5. Find the closest visual word and assign its ID to the pixel
7. For each object in the image:
8. Consider a region around the object center
9. Count the occurrences of each visual word in this region and build a histogram

Visual similarity

A visual similarity between patches needs to be computed in order to define the visual words (by clustering) and assign patches to the closest word. A very simple, yet efficient way to define this similarity is simply to take the absolute value of the difference between the two patches, pixel by pixel, and sum these differences over the whole patch:

$$S(p_1, p_2) = \sum_{i=0}^{n} |p_{1,i} - p_{2,i}|$$

Where n is the number of pixels in a patch.

However, this similarity measure is not rotation-invariant. Indeed, two patches that differ only by a rotation will not necessarily obtain a high similarity. This is not a desired property in this case, because in aerial images the orientation of the image does not hold valuable information. In contrast to natural images where the sky is generally at the top of the image

and the wheels of a car are below that car, in nadir aerial images the orientation is solely due to the direction of flight, which should not affect the interpretation of the scene.

Thus, to define a rotation-invariant similarity S_{RI}, we decided to compute the similarity S with several relative angles between the patches and retain the similarity that is highest:

$$S_{RI}(p_1, p_2) = S(p_1, p_{2,\alpha})$$

Where $p_{2,\alpha}$ is the patch p_2 after rotation by an angle α, and α is the angle that maximizes S. Note that all patches are circular, in order to allow rotations without changing their shape. This method increases the computation time significantly. In practice, a fixed number of rotations must be chosen, and the computation time increases linearly with the number of rotations.

Parameters

There are four parameters that considerably affect the features extraction using a bag of visual words. Finding the optimal values for these parameters is not straightforward.

Number of visual words

This parameter must be adapted to the heterogeneity of the data set. In natural images, the number of possible objects, colors and shape is almost unlimited, so that using a very large number of visual words may be necessary. Experiments have been done with as many as several thousands of visual words. However in a dataset of aerial images originating from the same location, the images are more homogeneous and a few hundred words is usually sufficient to describe the dataset.

Note that increasing the number of visual words does not necessarily lead to improved results. If this number is too large, the feature vector will contain many zeros, which may be a problem depending on the classification algorithm that is used. Moreover if the similarity between visual words is high, two patches that are very similar may be assigned to different words. As a result, some of the features could be very correlated.

In the following displays of maps of words, each color corresponds to a distinct word. The colors used to depict each word are chosen at random, so that similarity between colors does not imply similarity between words. Words cannot be directly compared between experiments, but the frequency and disposition of words can be analyzed.

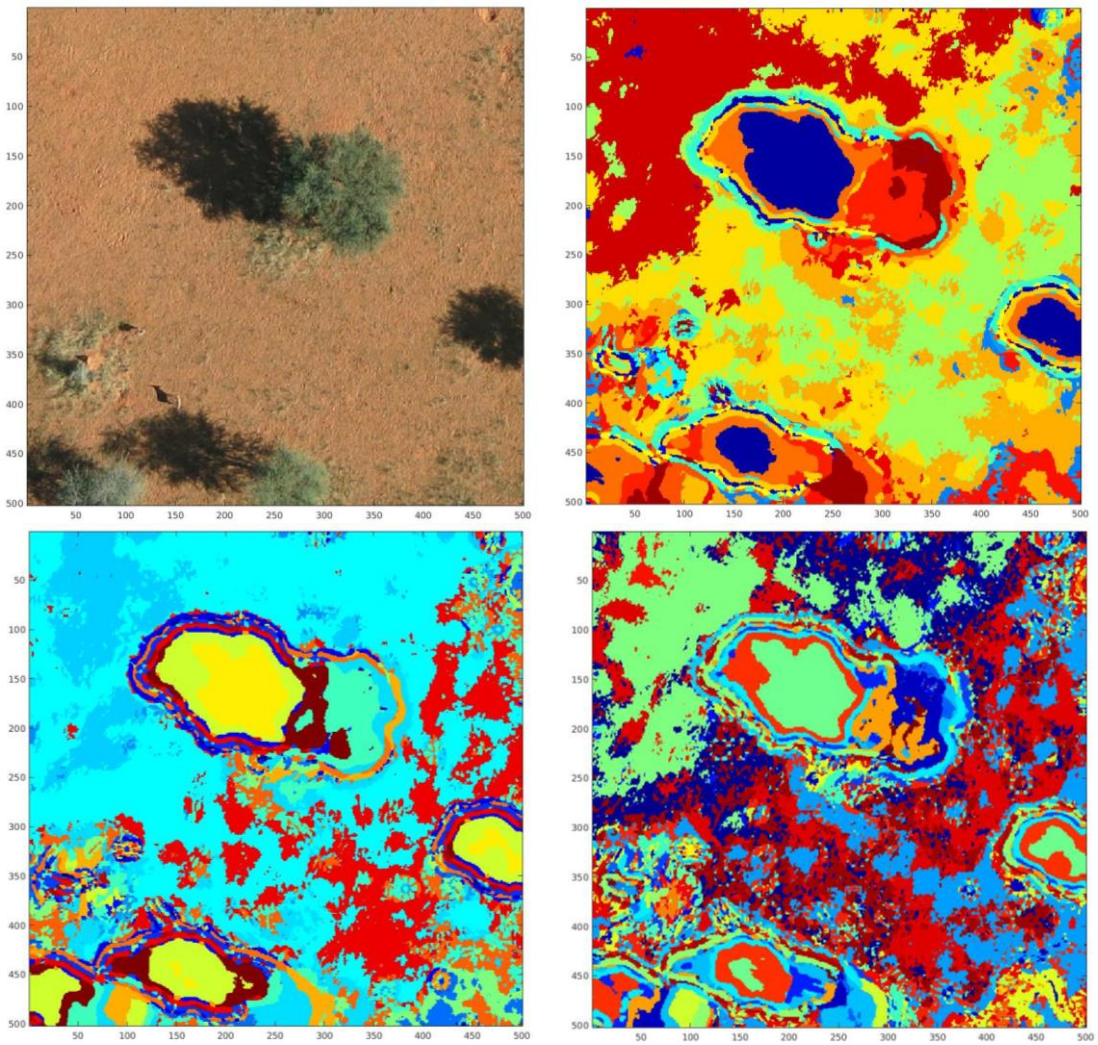

Figure 6 : Map of words obtained with different number of visual words. From top to bottom and left to right: original image, 60 words, 100 words and 300 words.

Patch diameter

The diameter of the circular patches defines the level of details that are retained. A small patch diameter results in more heterogeneity and a salt-and-pepper effect. The optimal diameter depends on the size of the objects of interest, and on the spatial resolution of the image.

Diameters between 17 and 30 pixels have been found to produce good results. Interestingly, doubling the spatial resolution does not mean that the patch diameter should be doubled.

Number of rotations

Allowing more rotations improves the results, but at the cost of computation time. Using 8 angles (with 45° difference between each) is a reasonable choice. Figure 7 shows the clusters without rotations, with 8 and with 16 rotation angles:

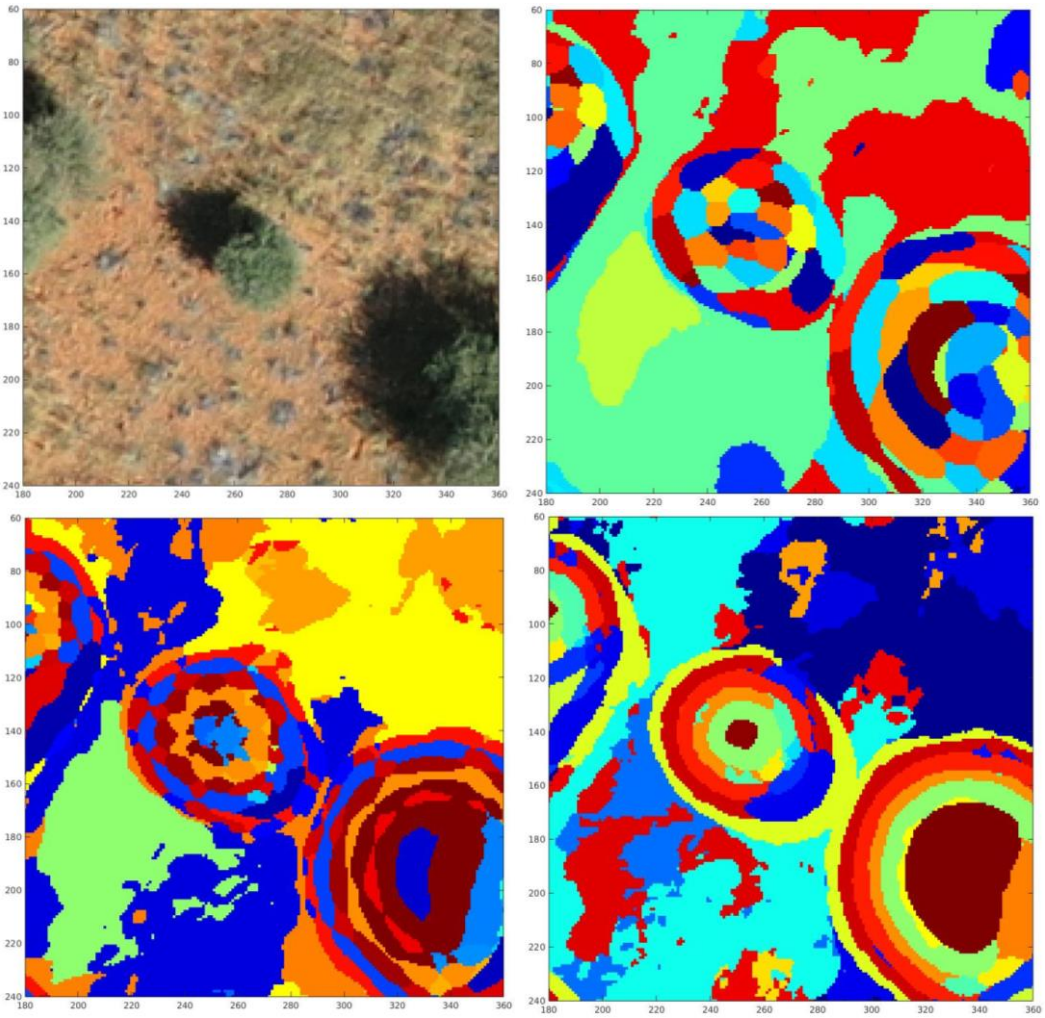

Figure 7 : Map of words obtained with different number of rotations. From top to bottom and left to right: original image, no rotation, 8 rotations, 16 rotations.

When no rotation is made, different sides of objects that are circular in appearance, such as the tree in this example, are assigned to different clusters. With 8 rotations, the situation is improved but the 8 directions can be recognized on the cluster image, meaning that the model is not completely rotation invariant. With 16 angles, the cluster image shows regular, circular shapes.

Note that the effect of rotations is more pronounced when using a large patch diameter, such as in this example.

Spatial resolution

In general, a higher spatial resolution allows to recognize more details and shapes and is therefore an advantage. However, to decrease the computation time the resolution should be decreased if a lower level of details is sufficient.

The spatial resolution also affects the other parameters:

- Because the image will be more heterogeneous with a higher resolution, the number of visual words should increase with the resolution.

- To some extent the diameter of the patches should be adapted to the resolution, so that one patch covers a meaningful part of the objects of interest in the image.

6.3.3 Normalization of the features

In order to treat all the features equally and to be able to compare them, the classifier needs that the features have the same distribution of values. Therefore, the feature vectors were replaced by their z-score, obtained by subtracting their mean and dividing them by their standard deviation.

When exemplar SVMs were used, one further normalization was done. The feature vectors were divided by their L_2-norm, so that they have a unit length. The reasons for this additional step are explained in section 6.4.4.

6.4 Classification

6.4.1 Definitions of terms

Classification is the process of assigning an object (or sample) to a class (or category) based on known properties of the object, called features (or descriptors). In binary classification there exist only two classes, while multi-class classification refers to the case where there are more than two possible classes. In this study, the aim is to detect animals without making further distinction between them, so that it will be a binary classification with the classes “background” and “animals”. The former contains all non-animal objects, and will also be referred to as the negative class. The latter is the class of interest, also called positive class.

Let’s now define the possible cases of correct classifications and errors. Considering that both the ground truth and the prediction can take the value 0 (class background) or 1 (class animals), there are four possible combinations, as defined in Figure 8:

		Ground Truth	
		Positive	Negative
Prediction	Positive	True positive (TP) *Correct detection*	False positive (FP) *False detection*
	Negative	False negative (FN) *Missed*	True negative (TN) *Correct rejection*

Figure 8 : Nomenclature based on the ground truth and predicted values. Green cells correspond to correct classifications. The colors for false negative and false positive should remind the reader that in this task, missing animals is worse than making false detections.

From this, a number of indices can be derived that better express the quality of the classification.

The *False positive rate* (FPR) is defined as the ratio of false positives over the total number of negatives:

$$FPR = \frac{FP}{FP + TN}$$

The *Recall rate*, or the fraction of retrieved positives, is given by:

$$Recall = \frac{TP}{TP + FN}$$

Finally the *Precision* is concerned only with positive predictions:

$$Precision = \frac{TP}{TP + FP}$$

Two types of graphs will be drawn with these indices: ROC curves that plot the recall as a function of the false positive rate, and precision-recall curves that plot the recall as a function of the precision. Note that once a classifier is trained, it is still possible to choose the minimal score for an object to be classified as animal. With a lower threshold value, more objects are classified as animals and the recall increases at the cost of false detections or precision, and vice versa for higher thresholds. Each value of threshold corresponds to one point on the ROC curve or on the precision-recall curve.

The difference between both types of curves is that the precision-recall curve does not depend on the true negatives (TN), while the ROC curve depends on all four cells of Figure 8. In the case of a dataset that contains many more negatives than positives, the number of true negatives can be very high and the false positive rate is not meaningful any longer. Instead, precision keeps all its meaning with imbalanced datasets. Therefore, precision-recall curves will be used in the case of imbalanced dataset. ROC curves will be used in order to make some results comparable to those of (Ofli et al., 2016).

Note that a classification should have a low false positive rate and a high recall, so that its ROC curve should reach as close as possible to the top left corner of the plot. In contrast, a classification should have a high precision. In the case of precision-recall curves, it is hence desired that the curves extend as much as possible towards the top right corner.

6.4.2 Characteristics of the classification task in this study

The classification problem addressed in this study is purely binary, meaning that there are only two classes involved. Even if some techniques easily extend to multi-class classification, binary classification is usually simpler. The *animal* class will also be referred to as the positive class, and the *background* class as the negative class.

It is common to find a high visual heterogeneity in the background class. A more specific feature of this dataset is that the positive class is also very heterogeneous, as shown in Figure 9. Most of the animals have a light fur but there are also darker, brown individuals, and the ostriches are grey or black. The variations in shape are also important. The presence of a shadow next to the animal is frequent but not necessary.

Figure 9: Visual heterogeneity among animals

Another particular aspect is that animals are very rare in the dataset and occupy only a tiny fraction of the images: we are looking for needles in a haystack. The ratio of positive to

negative samples affects the behavior of the classifier, and this ratio is especially low in this case.

Finally, in this task, the recall rate is thought to be more important than the precision. Indeed, if the precision is low, the user can visually check the detection and delete the false positives. Even if the precision is as low as 10%, the system would not be useless: reviewing the detections would still be much quickly done than visually interpreting the entire images of the whole dataset. In contrast, a good recall is essential for the system, as animals that are missed cannot be easily detected by another mean.

6.4.3 Support vector machine

Support vector machines classifiers find a linear boundary in the feature space that separates the two classes. Once this boundary is defined, any new data can be classified by looking at which side of the boundary it falls on. The following presents the framework and the mathematics behind SVMs.

Given a dataset of *n* objects belonging to either the positive or the negative class, and described by *D* features, the SVM classifier finds a boundary in the *D*-dimensional feature space that best separates the two classes. As shown on the 2D example of Figure 10, several lines that do not make any classification errors on the training set could be drawn and serve as boundary. But the green line, for instance, seems to be a risky choice because a negative object is located very close to it. In contrast, the black line keeps all objects as far away as possible, and is therefore more likely to predict correct classes for new objects from a test set. Therefore the SVM will try to find the line that maximizes the distance between the boundary and the closest objects, also called the *margin*.

Figure 10 : Illustration of a 2D feature space, with objects of the positive class and negative class, depicted as red + and blue – respectively. Several lines can separate the two classes (left) but the best solution is the line that results in the maximal *margin* (right).

Let's consider a vector *w* perpendicular to the boundary. An object is located on the positive side of the boundary if the following holds:

$$\vec{w}\,\vec{x} + b \geq 0$$

where $\vec{x}$ is the feature vector of the object and *b* is a constant term called bias. The dot product takes the projection of the feature vector on $\vec{w}$, and if the latter is a unit vector, the bias indicates where along $\vec{w}$ the boundary is located.

The function $s = \vec{w}\,\vec{x} + b$ is called the score function and takes positive values for objects situated on the positive side, and negative values for objects situated on the negative side.

A desired property for the score function is that objects located exactly one the “gutters” take the value -1 or +1 and other objects further away from the boundary take values below -1 or above +1, as shown on Figure 11.

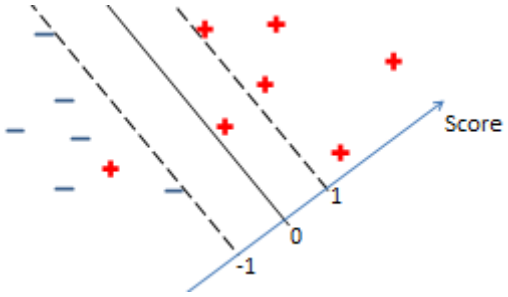

Figure 11 : The score is a measure of the distance to the margin. On the gutters, it takes the value +1 or -1. In this example, one positive object is misclassified and another one falls very close to the margin, so that its score is below 1.

With y equals to -1 for negative objects and +1 for positive objects, these conditions become:

$y\ (\vec{w}\ \vec{x} + b) = 1$ for objects on the gutters

$y\ (\vec{w}\ \vec{x} + b) > 1$ for objects further away

An expression for the margin (or the distance between the gutters) is derived as follows. Consider a point x_+ located on the positive gutter, and a point x_- located on the negative gutter. The width of the margin *M* is the projection of the vector $(\overrightarrow{x_+} - \overrightarrow{x_-})$ on a unit vector perpendicular to the boundary:

$$M = (\overrightarrow{x_+} - \overrightarrow{x_-}) \cdot \frac{\vec{w}}{\|\vec{w}\|}$$

By substitution of $\overrightarrow{x_+}$ and $\overrightarrow{x_-}$ using the previous equation, we find:

$$M = ((1 - b) - (1 + b)) \cdot \frac{1}{\|\vec{w}\|}$$

$$M = \frac{2}{\|\vec{w}\|}$$

Hence maximizing *M* is equivalent to minimizing $\|\vec{w}\|$ or, as often presented for mathematical convenience, $\frac{1}{2}\|\vec{w}\|^2$.

Finding the best boundary is hence formulated as optimizing $\frac{1}{2}\|\vec{w}\|^2$ under the constraint $y\ (\vec{w}\ \vec{x} + b) \geq 1$

This can be solved with the Lagrangian approach for constrained optimization:

$$L = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{n} \alpha_i\ [y\ (\vec{w}\ \vec{x} + b) - 1]$$

where *n* is the number of training objects and the α_i are lagrangian multipliers.

In this expression, the first term is a regularization loss that tends to keep the vector *w* small. Minimizing its L_2-norm encourages the solution to assign small values to all the w_i, thus making use of all the input features x_i. This is a desired characteristic, because ignoring some features could lead to overfitting the training dataset.

The second term is a data loss. The terms of the sum are positive for objects that are correctly classified and outside the margin.

6.4.4 Exemplar SVMs

Exemplar SVM (ESVM) is a specific use of SVM classifiers introduced in 2011 by (Malisiewicz et al., 2011). The idea is to train a different SVM for each positive object (called *exemplars*) of the training set. Hence each of these exemplar SVMs learns to perform a much simpler task: to distinguish objects highly similar to its exemplar only. To classify a new object, each ESVM produces one prediction, and the final class for the unknown object is decided by combining these individual predictions.

The problem to be optimized becomes (Malisiewicz et al., 2011):

$$L = \|w\|^2 + C_p \, \mathrm{h}\big(1 - (w^T x_p + b)\big) + C_n \sum_{x \in N_E} \mathrm{h}(1 + w^T x + b)$$

where N_E denotes the set of negative objects, x is the feature vector of a negative object x_p the feature vector of the positive object and $h(x) = \max(0, x)$ is a hinge loss function. C_p and C_n are two parameters that allow weighting each of the three terms: the regularization, the cost induced by a misclassification of the positive exemplar, and the cost induced by misclassification of the negative objects.

In their very recent contribution, (Kobayashi, 2015) demonstrate that ESVM can be formulated as one-class SVM by centering the data on the exemplar. This allows to eliminate one parameter:

$$L = \|w\|^2 - \rho + C \sum_{x \in N_E} \mathrm{h}(\rho - w^T (x - x_p))$$

where the margin is now $\frac{\rho}{\|w\|^2}$

(Kobayashi, 2015) also explains that this unique parameter *C* is bounded between $\frac{1}{N}$ and 1. It controls the number of support vectors and, since all the support vectors are from the negative class, it can be set to 1.

In order to avoid the exhaustive search of the best parameters C_p and C_n, it was decided to follow this appealing method. Because the code of (Kobayashi, 2015) was not available, we started from the freely available library of (Malisiewicz et al., 2011) and made the following adjustments to implement the method of (Kobayashi, 2015):

- Normalize the all the feature vectors to a unit L_2-norm [2]
- Center the data by subtracting x_p to the negative samples
- Train a one-class SVM with linear kernel
- Set the cost parameters to 1

[2] The features were already centered to have a zero mean and a unit variance, as explained in section 6.3.3.

This implementation was successfully tested on a toy example before being applied to the real dataset.

The last step is to combine the scores given by the ensemble of exemplar SVMs. Each exemplar assigns a different score to each object and the higher the score, the higher the visual similarity between the object and the exemplar.

A first issue is that in general, the scores cannot be directly compared because they do not have the same distribution. In other words, if two exemplars give the same score to an object, these scores may not translate to the same class probabilities, because one object may tend to assign scores close to zero and with little variance, while the variance in the scores given by the other exemplar may be higher.

There are several methods to deal with this issue. A common solution is the following: for each exemplar, a separate set of objects is predicted (i.e. the score of these objects is computed). A sigmoid is fit to the distribution of these scores. At test time, the scores given by the exemplars are passed through the sigmoids of the exemplars. After that the scores are comparable between exemplars.

A relevant question is which set of objects can be used to fit the sigmoids. This set must contain positive and negative objects. Conveniently, to fit the sigmoid of any exemplar $x_{p,i}$ the other exemplars of the training set can be used, because they were not involved in the training of $x_{p,i}$. Since the training set contains many negatives, a fraction of them can be kept aside (i.e. not used for training) and used to fit the sigmoid.

(Kobayashi, 2015) pretend that the scores given by their exemplars can be directly compared, because they normalized the feature vectors of all objects to a unit L_2-norm. In this way, the feature space is bounded (all objects live on sphere of unit radius) and this should ensure that scores are comparable.

In this study, after few trials and based on geometrical considerations, it was decided to normalize the scores by dividing them by the width of the margin.

6.4.5 Hard negative mining

Due to the strong imbalance between positive and negative objects in the dataset, a huge number of negatives are available for training. But using all of these negatives is not possible, due to computational limitations (both in time and memory). The simplest solution is then to select a subset of negatives at random.

However all negative objects are not equally useful. The Support Vector Machine only uses a fraction of the provided data to draw its separating hyperplane: the so-called support vectors, which are located close to the boundary between the two classes. Negative examples that have been misclassified as positive are also useful examples because they induce a misclassification cost and tend to attract the hyperplane toward them.

The aim of “Hard negative mining” is to train the model using the most useful negative examples: the “hard” negatives that are located close to the positive examples. In this way, the boundary drawn by the SVM will move closer to the positive examples, and the number of false positives is expected to decrease.

Hard negative mining is done in an iterative manner: the classifier is first trained with a random subset of negatives, and this classifier is used to identify a first set of hard examples.

The score of all negatives of the training set is computed by the classifier and they are assigned to a class. All negatives that are misclassified are regarded as hard negatives.

Figure 12 : Toy example of hard negative mining. Dark blue minus signs represent negative samples in the training set, while the light blue ones are not used for training. a) The first iteration is based on a random subset of negatives. b) It allows finding two hard negatives (red minus signs) that are added to the training set. c) After the second iteration, the boundary has moved closer to the positive sample (red plus sign)

Then, at each iteration, the subset of negatives that is used for training is updated by adding the hard negatives found at the previous iteration. The model is retrained with this updated set, the score of all negatives is recomputed with the updated classifier and a new set of hard negatives is identified.

In order to implement negative mining, one has to choose the initial size of the negative set (the number of negatives taken at random) and the number of hard negatives added at each iteration. Starting with a large subset allows reaching good results in less iteration, but the computation time for each iteration is longer, since the number of examples is larger. On the other hand, a small initial subset may require more iterations to converge.

6.4.6 Active learning

One issue related to ESVM is the enhanced influence of errors in the ground truth. There are two types of errors:

- False positives: objects labeled (in the ground truth) as animal while their correct class would be background
- False negatives: objects labeled (in the ground truth) as background while their correct class would be animal

A straightforward way of solving the problem of false positives is to ask an expert to visually inspect each positives and confirm or infirm them. For the data set used in the present study, it means that 976 objects must be inspected. This approach was used and it took approximately 30 minutes to review all positive examples, among which there were 23 false positives.

The problem of false negatives is more complex. For this dataset using the same approach would require to visually inspect 403’859 negative samples – a task that cannot be

decently asked to anyone. To tackle the issue, a specific form of active learning has been implemented.

First, let's consider active learning in its most common acceptance: similarly to hard negative mining, active learning deals with the problem of building an efficient training set. While hard negative mining makes an intelligent selection among the existing training samples, active learning search for new training samples (outside the training set) that would be particularly helpful to the model. In order to include them in the training set, the model asks the user to give them a label.

This approach is particularly efficient in situations where a limited amount of labeled samples exist, and the acquisition of a larger ground truth is expensive. This is the case in some remote sensing applications where the ground truth can only be acquired through terrain campaigns. It is also the case for the present study: since the positive class is very rare, finding new positive samples requires inspecting a huge amount of images.

Active learning is usually used to extend the training set by including new areas that were not covered by the ground truth. In this study, the ground truth already (supposedly) covers the whole data set, but it contains false negatives. The aim of active learning is hence to correct these errors of the ground truth.

The idea is that false negatives will be located close to some of the positive exemplars in the feature space and therefore these positive exemplars will give them a high score. Most of the false negatives should therefore be found in the subset of negatives that obtained high scores.

At each training iteration of each exemplar, the 9 negative objects that obtain the highest score are displayed and the user can inspect them, and decide if they are real negative samples. The user has the following options:

Animal: the object is removed from the negatives and is used as a new positive exemplar that will be trained.

I cannot say: the sample is removed from the negatives, but will not be used as an additional positive exemplar. The user chooses this option if he does not want the model to avoid retrieving similar samples, but neither wants to encourage the model to look for them.

Background: the object is kept as a negative sample.

The system keeps a list of all objects that have been assessed by the user, so that it avoids showing the same object again. Most of the "difficult" objects are shown to the user during the training of the first exemplars. If the user keeps on working, the additional time spent will become less productive, since most of the ground truth errors (i.e animals considered as negatives) are rapidly displayed and discarded by the user.

7. Experiments

This section describes the goals and experimental setups of the different experiments that were conducted.

7.1 Objects proposals

The aim of this experiment is to compare the two methods (based on edge detection and shadow detection respectively) and see if using them together brings any benefit.

The algorithms described in section 6.2 are run on the full image dataset for the three cases: shadow detection, edge detection, and both combined. The resolution of the images was reduced by a factor of two, yielding a GSD of around 8 cm.

The number of retrieved tags and the mean number of proposals per image are reported as functions of the threshold.

7.2 Features

The aim of this series of experiments is to compare the performance of the histogram of words and the histogram of colors, and to analyze how the different parameters of the HOW affect the discriminative power of the features.

To this end a classification using a linear SVM and a balanced dataset is performed. The hypothesis is made that features performing well with SVM will also perform well with ESVM, so that it should be possible to rely on the results obtained with this simple setup, even when moving to ESVM.

All available animal objects are used, and the same number of negative objects is used to ensure a 1:1 class ratio. The training sets comprise 1324 objects, and the test sets have 568 objects. For each experiment five replicates are done to mitigate the effect of random selection of the negatives. The cost parameter for the linear SVM is optimized with a 5-fold cross validation.

The results are reported as ROC curves averaged over the 5 replicates.

The first experiment compares the discriminative power of the HOW, the HOC and the concatenation of both. The following experiments compare the classification results obtained with three different image resolutions, with and without rotation-invariant words, and with 100 and 300 words. Table 2 gives the parameter values for each experiment.

Table 2 : Experimental setup

	GSD	Features	# words	Rotation-invariance of words
Type of features	8 cm	HOC, HOW, HOC+HOW	100	Yes
Resolution	8, 12, 16 cm	HOC + HOW	100	NO
Rotation-invariance	8 cm	HOW, HOC+HOW	100	Yes / NO
Number of words	8 cm	HOC+HOW	100, 300	No

7.3 Classification in an imbalanced dataset

The aim of these experiments is to assess to quality of classification using a simple implementation of exemplar SVMs in the case of a very imbalanced dataset, and to analyze how hard negative mining and active learning can improve the performance.

For these experiments all the negatives are included to the training and test set. The training set comprises 662 exemplars (positive objects) and 403'859 negatives, giving a ratio of 1:610. The test set has 284 exemplars (positive objects) and 160'384 negatives, yielding a ratio of 1:564. The sets were determined by ranking the images by number of animals they contain, and then assigning images to the train and test set in turn. In this manner, all objects of a same image are always assigned to the same set, and both sets are believed to contain a similar number of large herds, medium herds, isolated animals, etc.

The first experiment compares the classification results obtained without hard negative mining, with one single iteration of negative mining, and with as many iterations as needed for the cache to become stable. Three different values were used for the initial number of negatives in the cache.

The next experiment deals with active learning. To assess the benefit of using this method to correct errors in the ground truth, exemplars have been trained using the form of active learning described in section 6.4.6. After one hour of work, active learning was switched off. The detected false negatives were added to the remaining exemplars, and all of them were trained with the improved negative set. Results obtained with these actively trained exemplars were confronted with those obtained without active learning.

7.4 Influence of the time of the day

This experiment aims to find the time period where detection is easiest. It compares the classification results of subsets of images taken at different times of the day. Based on the available images, three periods of time where defined: the first from 09:13 until 09:28 ("morning" subset), the second from 13:08 to 13:00 ("midday" subset) and the last from 15:00 to 15:10 ("afternoon" subset).

After examination it appeared that the afternoon subset contains only few animals, and half of them are ostriches. Because ostriches are visually very different to other animals, comparing subsets with such inequalities seems a poor idea. For this reason, the afternoon subset was not used. For a fair comparison of the other subsets, and because the morning subset contained no ostriches at all, the few ostriches of the midday subset were removed. Table 3 details the number of images and animals (quadrupeds only) in each subset, after the division into a training and a test set for each time period.

Table 3 : Composition of the subsets of different time periods

	Training		Test	
	# images	# quadrupeds	# images	# quadrupeds
Morning	89	261	36	119
Midday	120	176	48	82

To ensure a fair comparison, only 176 exemplars of the morning training set were used. In this manner both training sets had the same number of exemplars.

8. Results

8.1 Objects proposals

Each of the two methods presented in section 6.2 has been tested separately with a range of different threshold values. Then both methods were simultaneously applied with a grid search to find suitable parameters. This was made to determine the benefit of using both methods combined.

Figure 13 shows how the ratio of retrieved tags and the total number of proposals evolve with the threshold value, for each method. When combining both methods, the maximal percentage of tags retrieved is 88%, with a mean number of proposals per image of 490.

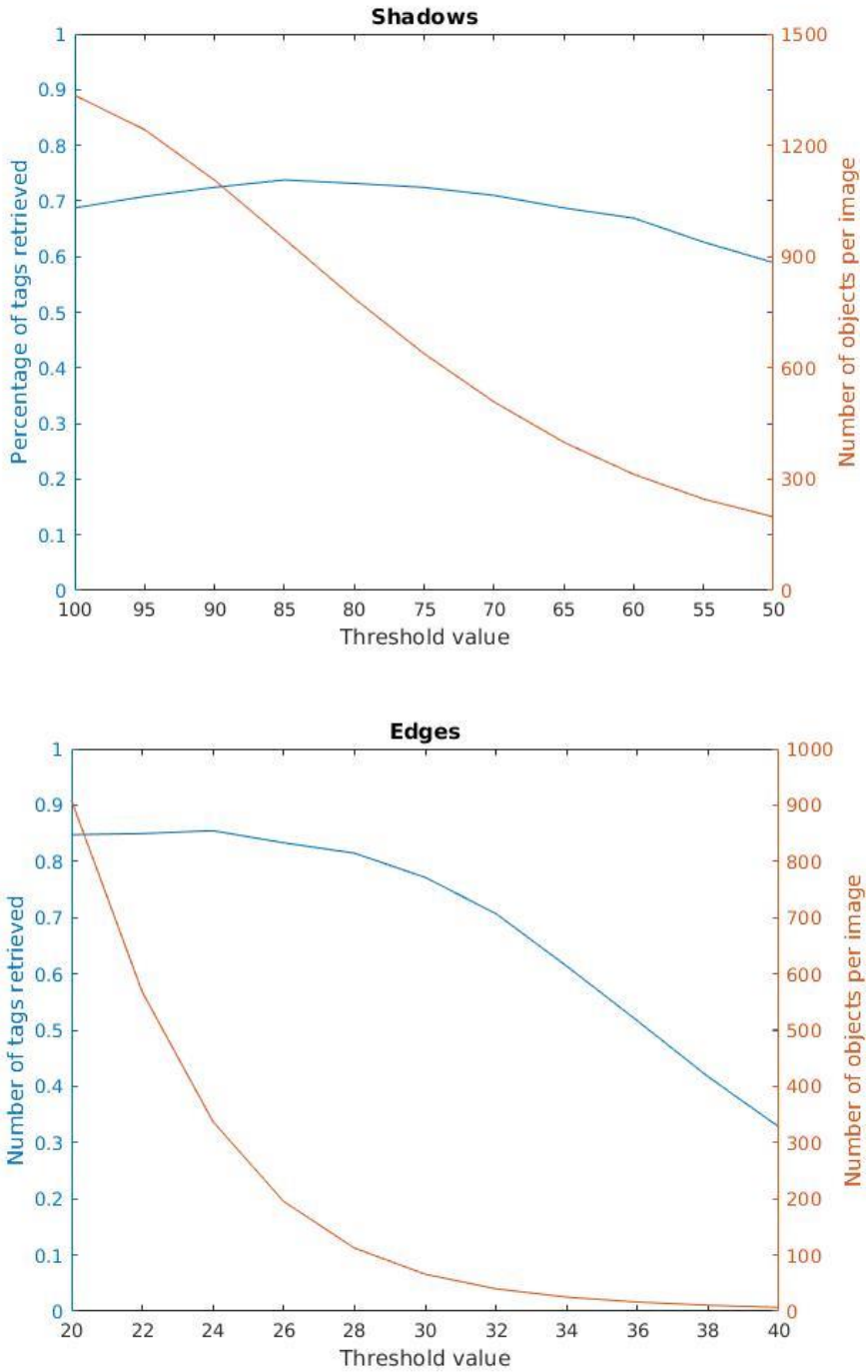

Figure 13 : Performance of the Shadow and Edge methods for objects proposal: Number of tags retrieved and mean number of objects per image, as a function of the selected threshold.

In this experiment the spatial resolution of the images was reduced by a factor of two, yielding a GSD of around 8 cm.

The other parameters have been chosen heuristically:

- Minimal area:	2	pixels
- Buffer size:	5	pixels
- Size of sobel filter	5x5	pixels

Note that these parameters must be adapted to the resolution of the image.

8.2 Features

The results regarding the types of features and the parameters for the histogram of visual words are presented as ROC curves. The training and the test were performed on balanced data sets, i.e. containing the same number of animals and background objects.

8.2.1 Feature type

The ROC curves of Figure 14 present the classifications obtained with two types of features (histogram of colors and histogram of visual words) as well as the concatenation of both feature types.

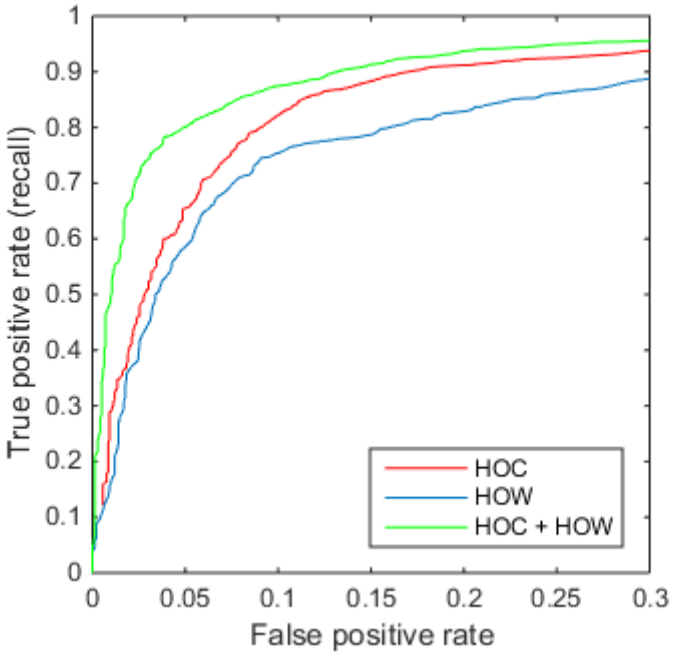

Figure 14 : ROC curves of the classification based on histogram of colors (red line), histogram of visual words (blue line), and both feature types combined (green line)

8.2.2 Effect of the resolution

Figure 15 presents the classification results for different rescaling of the images. Note that the resolution of the images was always reduced by at least a factor of two. The reason is that the computational cost for histograms of visual words increases very rapidly with the resolution and using a GSD lower than 8 cm would only make sense with a better implementation than the Matlab code written for this study.

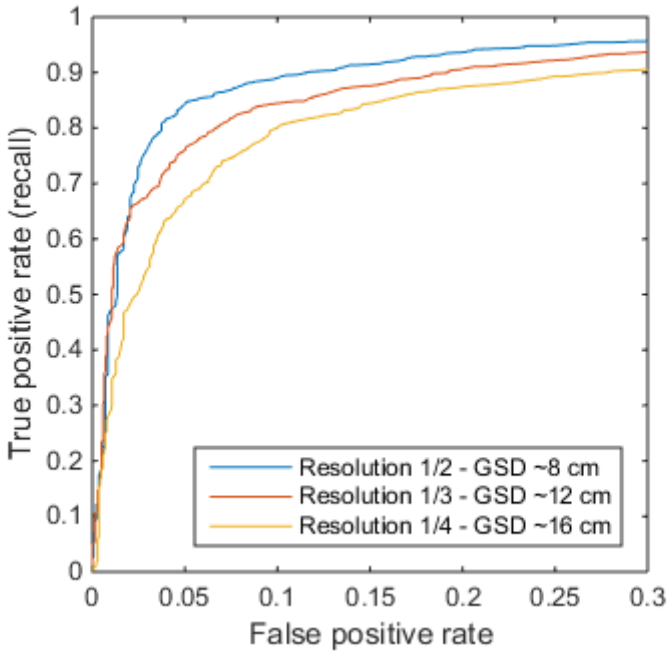

Figure 15 : ROC curves of the classification at a half (blue line), a third (red line) and a forth (red line) of the original image resolution. The ground sampling distance is approximatively 8, 12 and 16 cm respectively

8.2.3 Effect of the rotation invariance

Figure 16 confronts the traditional bag of visual words with its rotation-invariant version. For the latter, 16 rotations of the visual words are used. The choice of 16 rotations is supported by Figure 7.

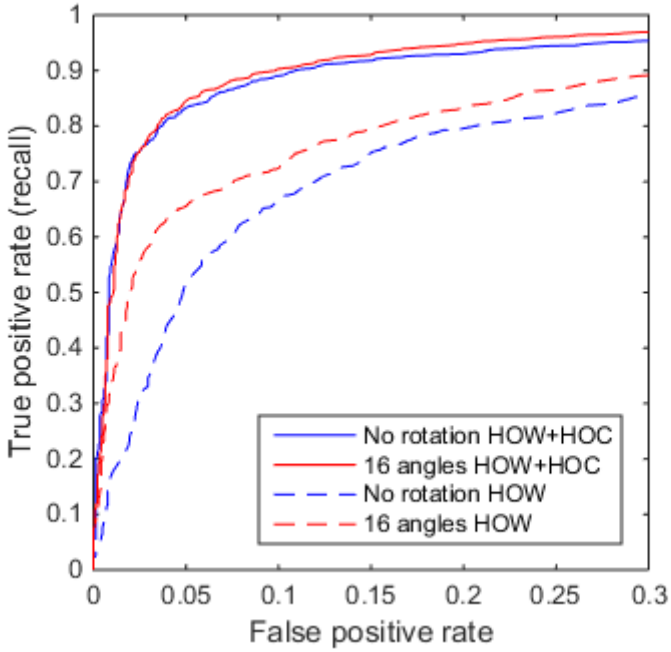

Figure 16 : ROC curves of the classification with (red lines) and without (blue lines) rotation of the visual words, when using histogram of visual words alone (dashed lines) and together with histograms of color (solid lines).

8.2.4 Effect of the number of words

Figure 17 shows the influence of the number of visual words on the classification results.

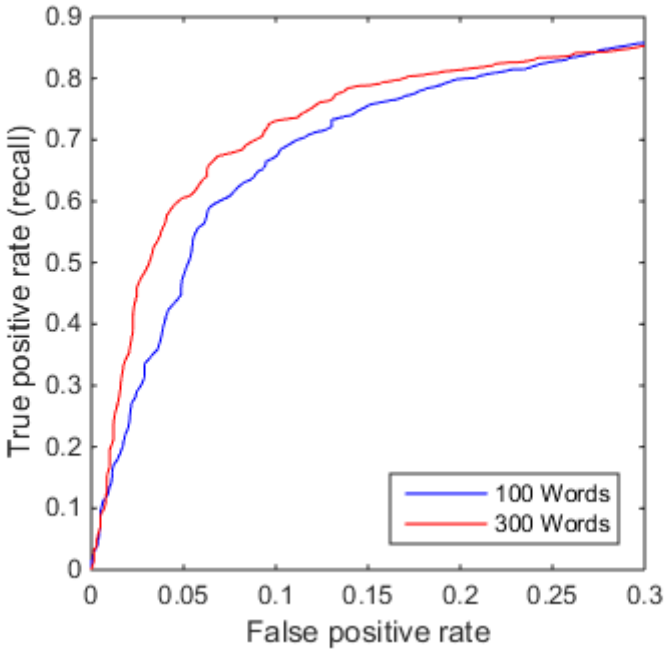

Figure 17 : ROC curves of the classification with 100 words (blue) and 300 words (red)

8.3 Exemplar SVM

8.3.1 Simple ESVM and hard negatives mining

Figure 18 shows the precision-recall curve obtained with and without hard negative mining. For all curves, training started with 5'000 negatives. The red curve is obtained by making one single update of the cache. The orange curve is obtained when HNM is iterated until all hard negatives are included in the cache.

It should be mentioned that because the animals class is strongly underrepresented, a random classifier would obtain such a small precision (in the range of 10^{-3}) that it cannot be represented on the graph.

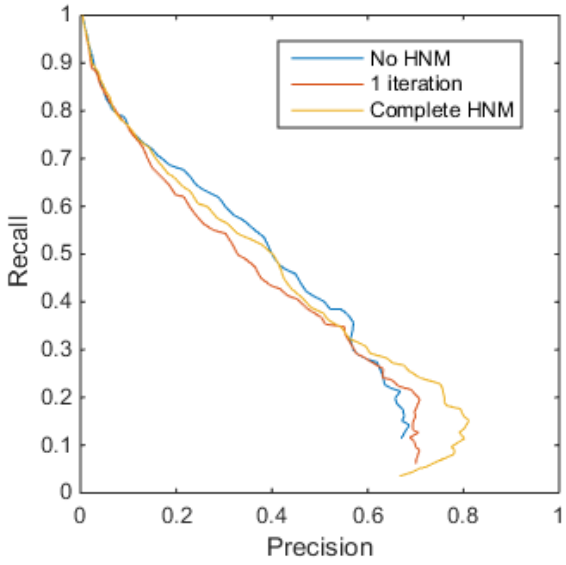

Figure 18 : Influence of hard negative mining (HNM) on the precision-recall curve, when starting with 5'000 negative objects.

The next figure better displays the effect of HNM on the precision and on the recall, by plotting them separately against the threshold on the objects' score. Objects having a score higher than the threshold are classified as animals, and those with a lower score as background.

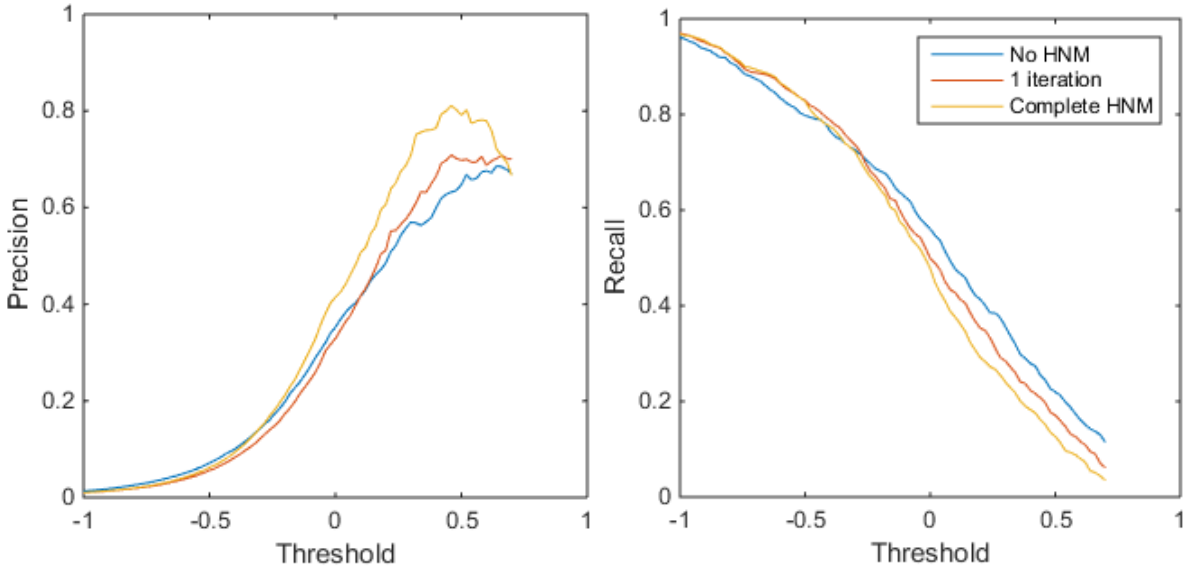

Figure 19 : Influence of hard negative mining on the precision (left) and the recall (right).

Figure 20 shows the precision-recall curves obtained when starting with a higher number of negative examples in the cache (60'000 and 120'000) and can be compared with Figure 18, where this number was much lower (5'000).

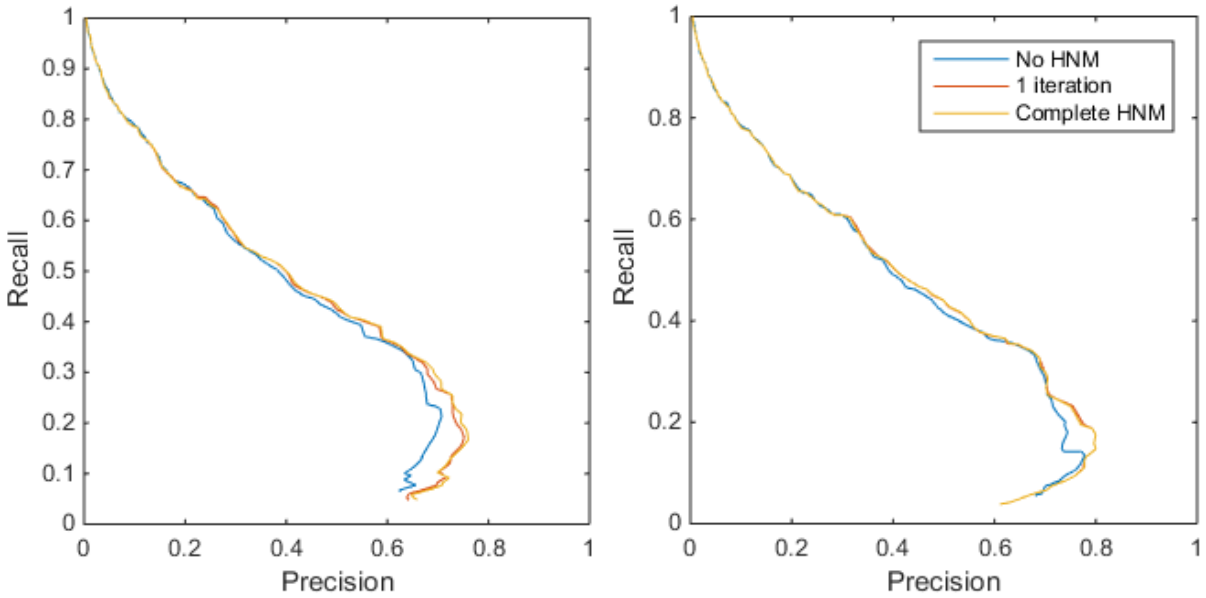

Figure 20 : Precision-recall curves obtained when starting with a high number of negative objects: 60'000 (left) and 120'000 (right).

The next figure displays a few detections along with the exemplars that gave them the highest score (i.e. the most similar exemplars).

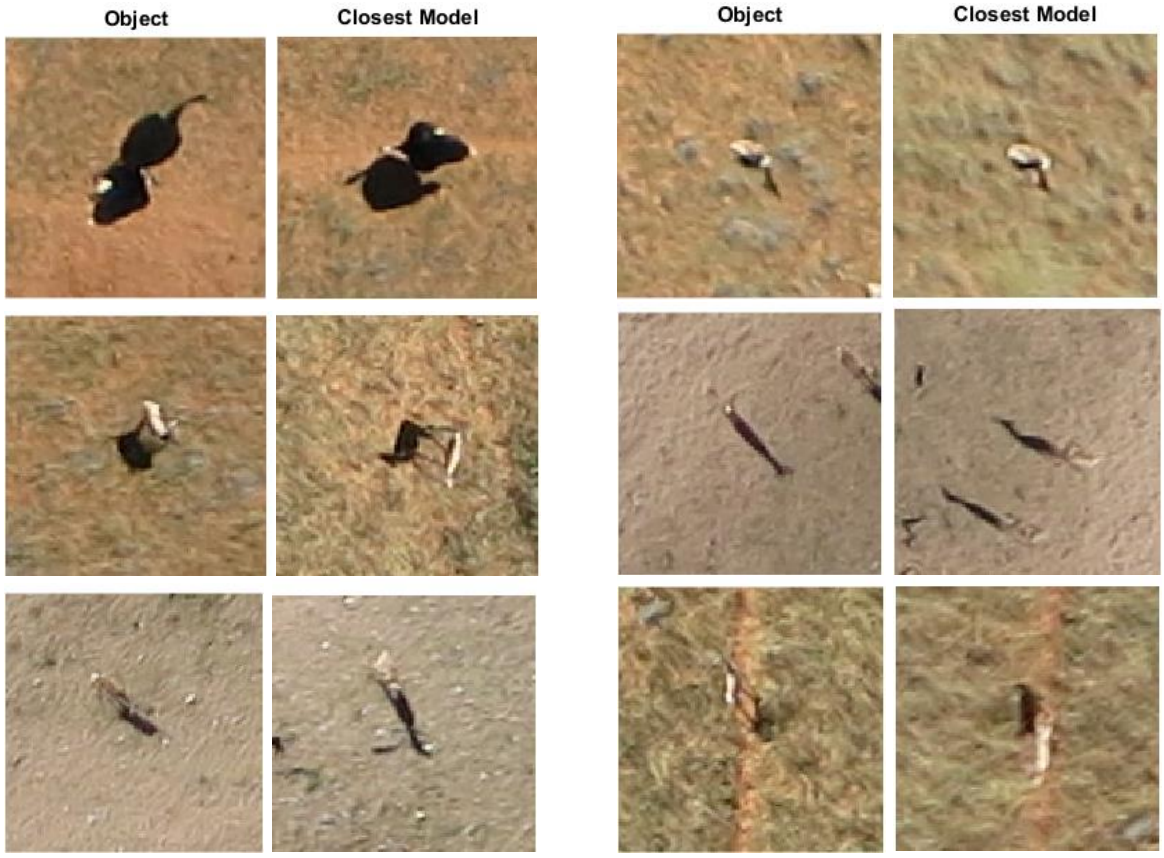

Figure 21 : Detected animals (left columns) and most similar exemplars (right columns)

8.3.2 Active learning

Learning started with 574 exemplars and 403'859 negatives in the training set. After one hour of interactive learning, 55 animals had been found among the negatives and added as new positive exemplars. Another 52 negatives had been assessed as "I am not sure" and removed from the negatives pool. A total of 1678 small tiles (presenting one object) were visually assessed, or 0.42% of all tiles. The frequency of animals found among the negatives decreased rapidly over time. During this hour of active learning, 120 exemplars were trained. The remaining 509 exemplars (454 + 55 additional ones) were trained without active learning.

Figure 22 : Precision-recall curve of classification after one hour of active learning (red) and without active learning (blue).Figure 22 compares the precision-recall curves obtained with and without active learning.

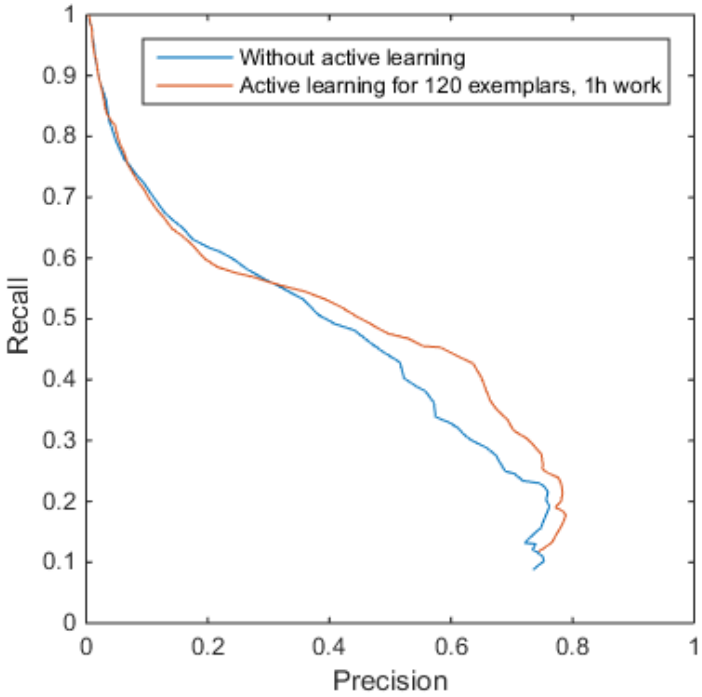

Figure 22 : Precision-recall curve of classification after one hour of active learning (red) and without active learning (blue).

8.4 Influence of the time of the day

The histogram of the three bands red, green and blue at the three periods are shown in Figure 23. Figure 24 presents the precision-recall curves. To ease the interpretation, the precision and the recall are also displayed separately, as function of the threshold. There are three sets of exemplars (those from the morning images, those from the midday images, and both combined) that can be used to predict two test sets (morning and midday), yielding 6 possible combinations – each line of the graphs correspond to one combination.

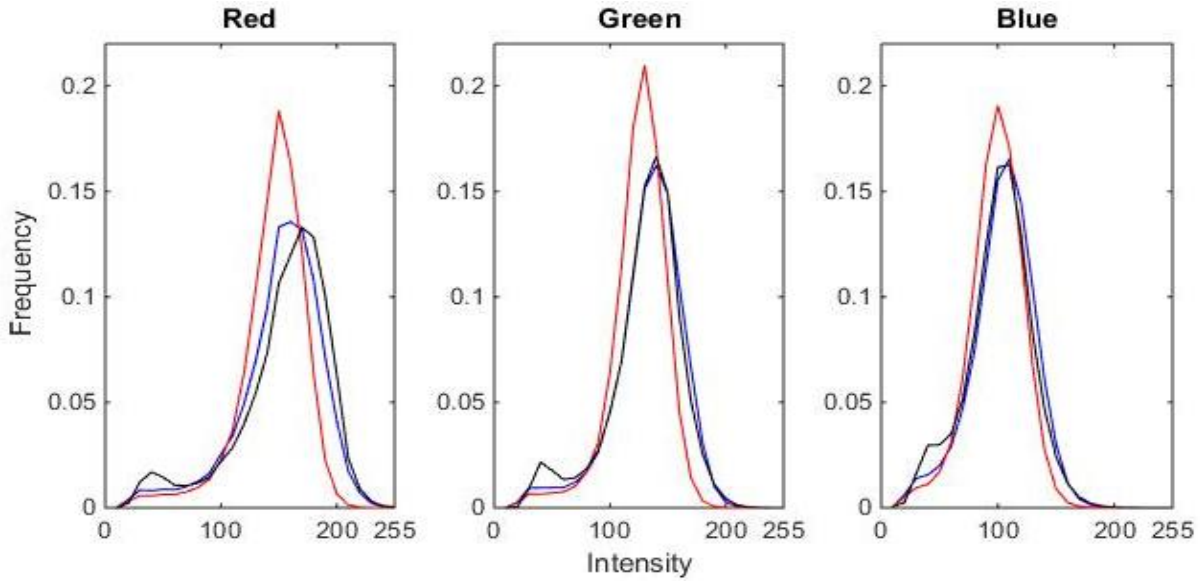

Figure 23 : Histogram of the red, green and blue bands at three moment of the day: in the morning from 09:13 to 09:28 (blue curve), from 13:08 to 13:30 (red line), and in the afternoon from 15:00 to 15:10 (black). The histograms are a mean over all the images taken during the respective time periods.

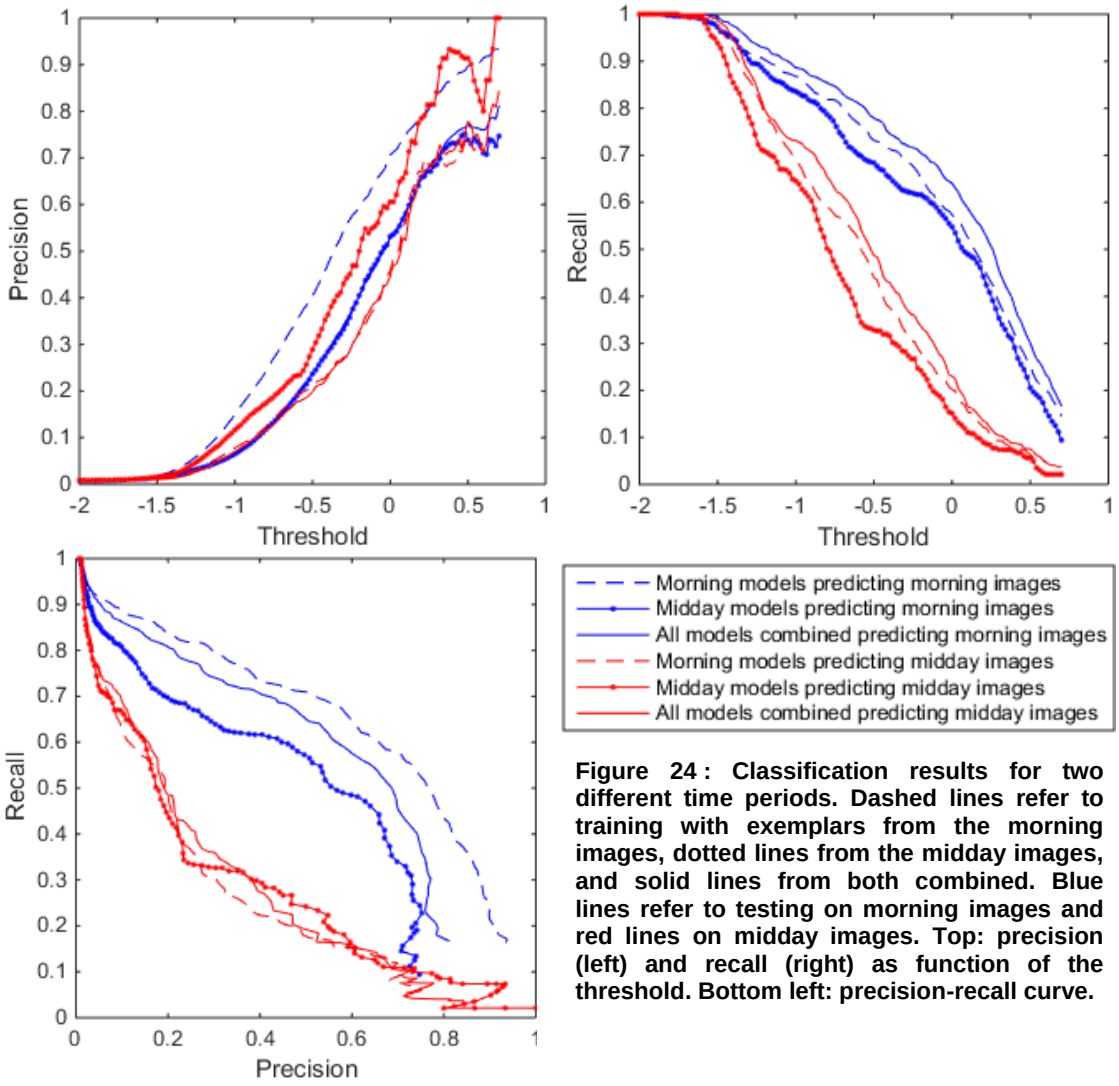

Figure 24 : Classification results for two different time periods. Dashed lines refer to training with exemplars from the morning images, dotted lines from the midday images, and solid lines from both combined. Blue lines refer to testing on morning images and red lines on midday images. Top: precision (left) and recall (right) as function of the threshold. Bottom left: precision-recall curve.

9. Discussion

9.1 Objects proposals

The method based on edge detection is able to retrieve more tagged animals than the method based on shadows, and at the same time it gives a smaller total number of objects. Figure 25 shows a situation where the edge detection was successful, while the shadow detection failed because the animal is in the shade of a tree. Such situations are not rare and can explain why the shadow detection does not reach the retrieval rate of the edge detection. Furthermore the edge detection proposes a lower total number of objects because most many trees and bushes have diffused borders and shadows that do not respond much to the sobel filters.

Figure 25 : Situation where the edge detection successfully defined an object on the tagged animals, while the shadow detection failed. Left : original image. Center: mask of shadow detection. The tree and the animals are merged in a single blob and the object is defined at its center, away from the animals. Right: mask of edge detections. The diffuse border of the tree have not fired above the threshold. Two blobs correspond to the animals, one on their white bodies, another one on the sharp edge of their shadow.

When both methods are combined, the maximal number of retrieved animals increases by 2% with respect to the edge detection alone, at the cost of 140 additional objects per image on average. The benefit is therefore questionable. Nevertheless, it was decided to use both methods together because the additional objects should not be especially difficult to classify as background, if they have this characteristic that they lack sharp borders. In addition, the shadow detection might be necessary to retrieve ostriches, because of their overall dark color. Missing the few members of this species would be sad.

Retrieving 88% of the tagged animals is a satisfactory result, when considering that the ground truth contains a certain number of false positives (i.e. background objects tagged as animals). To analyze the quality of the proposed objects, it would be interesting to quantify the cases where several proposals were made on the same animal, and the cases where a single proposal is made for two adjacent animals.

The complete set of objects, obtained by combining both the edge and shadow detection, comprises 976 objects of the class animal, and 403’859 objects of the class background.

9.2 Features

Before describing the figures, it should be reminded that in this experiment the dataset includes the same number of animal and background objects, which strongly differ from the natural case where animals are very rare compared to background objects. Thus, keeping in sight the final goal of detection in a very imbalanced dataset, the attention should be primarily focused on the left side of the ROC curves, corresponding to low false positive rates. Indeed in the naturally imbalanced dataset, a false positive rate of 0.1 would already correspond to several thousands of false detections

At first sight the histogram of color performs surprisingly well in comparison to the much more elaborated and time-consuming histogram of visual words (Figure 14). This indicates that colors hold the major part of the information, while the shapes and structures that could be captured by visual words are less important in this task.

However, on the very left side of the graph the advantage of the HOC is smaller. There, combining both methods brings a huge improvement. For a false positive rate of 0.03 for instance, the recall is 0.75 for combined histograms while it is only 0.50 for HOC alone and 0.45 for HOW alone.

A GSD of around 16 cm is not able to compete with higher resolutions. But, interestingly, the benefit of using a GSD of 8 cm over a GSD of 12 cm only appears at a recall of 0.65. This indicate that two thirds of the animals do not need such high resolution to be distinguished from background objects, but the last third of the animals becomes more distinguishable when the resolution is increased to 8 cm. Finally, it is noteworthy that increasing the resolution is very expensive in terms of computational time, because the relation between the resolution and the size of the patches to be compared is quadratic.

As for the rotation-invariant HOW, it brings a real improvement when the HOW is used alone. However this advantage becomes much less obvious when the histogram of colors is added to the feature vector, and completely vanishes for very low false positive rates. The conclusion is that the rotation-invariant HOW may be appealing in theory, but does not bring a clear improvement and induces unnecessary computational burdens.

As expected, using more words improves the classification. The benefit is highest for recalls between 35 and 60. In this range, using 300 words instead of 100 can improve the recall by up to 15%.

These experiments show that fine-tuning the parameters for the histogram of visual words is not easy. With the Matlab code implemented during this study, computing the HOW for the whole dataset of 654 images takes from 12 to 72 hours, making a fine search of the parameter space very tedious.

Nevertheless this first analysis indicates that it is better to use the histogram of color and the histogram of visual words in combination. Then, using a higher number of words should be preferred over rotation-invariance. Finally, using a higher image resolution is also beneficial but at the expense of high computational cost

But before trying to optimize the HOW, it would probably be better to try adding other features such as a histogram of gradients and features that describe the texture of the image (such as the variance or the entropy).

9.3 Exemplar SVM

9.3.1 Simple ESVM and Hard negative mining

Let's first consider the blue curve of Figure 18, which corresponds to the simplest use of exemplar SVMs. This precision-recall curve indicates that, for instance, there exists a threshold value that yields a recall of 70% for a precision of 16%. This precision may seem quite low in comparison to the results generally reported for classification tasks. However, considering that animals are very rare in the dataset, this is already satisfactory. In practice, a human observer could visually review the detections and eliminate the false positives. Displaying 18 tiles at a time on a screen (each tile representing one detection), on average he would find 3 animals per screen – many more than the volunteers can ever hope during a crowd-sourcing campaign!

Regarding hard negative mining (HNM), Figure 18 shows disappointing results. In this case HNM has decreased the performance of the system.

Figure 19 reveals that HNM led to a decrease in recall and an increase in precision. This has the following explanation: as hard negatives are found, the boundary between the classes moves closer to the exemplar, thus requiring a higher similarity with the exemplar for other objects to be classified as animals. This naturally decreases the recall and increases the precision. Unfortunately, the reduction in recall was in general stronger than the increase in precision.

When training starts with a higher number of negatives in the cache (Figure 20), the effect of HNM becomes negligible. The performance is marginally increased. The reason could be that the initial 60'000 or 120'000 negatives already contained sufficiently hard negatives. This idea is supported by the way that objects were defined: the edge detection and the shadow detection produced many objects that are visually close to the animals. In this sense, the set of negative objects may include a sufficient proportion of hard objects, so that it is not necessarily needed to go and look for them.

Considering this, the use of HNM is not recommended for this dataset.

Figure 21 displays a few detections and their closest exemplar. Very often, the exemplar shares many characteristic with the detection: similar species, similar background (note that in one instance even the presence of a trail is shared by the detection and its exemplar), and sometimes a similar posture (as in the top right example). Here, the advantage of exemplar SVMs over standard SVM becomes clear: if the exemplars were annotated with such attributes, these could be transferred to the detections, apparently with a good accuracy. In particular, the species is an attribute that would be very relevant.

In other words, this indicates that a species-specific detection can be envisioned. If it may not be possible to distinguish every species, at least it would be feasible to classify animals into groups of similar species. To further investigate this, a ground truth with information about the species would be required.

9.3.2 Active learning

The precision-recall curves indicate that for this dataset, active learning does not enhance the predictive ability of the ensemble of SVMs if recall rates above 55% are wanted. It is only beneficial if a high precision is desired. Thus at first sight it seems not very interesting in the context of this study.

However if we consider that 55 additional animals were found in the *training* set, it appears that this form of active learning is a neat way to improve a ground truth. It can be recommended in the case where additional detections in the training set is precious information, i.e. the training set is not only used to train a classifier, but further analyses are made on this training set as well.

9.4 Influence of the time of the day

Looking first at the histogram of colors, the afternoon subset is characterized by a peak in the low intensities. This clearly indicates that the images taken at this time of the day contain a larger fraction of shadow. The morning subset does not show a clear peak, but still has a little higher number of low intensities than the midday subset. Because shadows only occupy a small fraction of the image, these small differences are already interesting.

For the rest, the morning and afternoon histograms are quite similar. The midday histogram is spikier and, surprisingly, its maximum corresponds to a slightly smaller intensity. It also falls more quickly on the right side. This is probably due to internal calibration of the camera or to the JPEG compression. To further investigate these histograms, they should be recomputed with images in RAW format instead of JPEG.

Considering now the precision-recall curve (bottom left graph of Figure 24), it appears that it was much easier to classify the morning test set than the midday one, regardless of the exemplars involved.

The top right graph indicates that the morning subset of exemplars gives a better recall, even when classifying the midday test set. However combining both training set provides even better results.

Regarding precision, another behavior is observed. There, the best results are obtained when the training and the test subsets are from the same time period. Then, training with additional exemplars from the other time period does not affect the precision.

These comments lead to the following statements:

1) Exemplars from the morning subset are more efficient at retrieving animals than those of the midday subset.
2) In general exemplars produce more false positives when used to classify images taken at another time of the day.
3) Animals of the morning test set are easier to retrieve than animals of the midday test set.

However, it is difficult to know if these results reflect a true effect of the hour of the day. They could also be due to differences in the background environment, since the images of the two subsets were not taken at the same location. If these statements need to be confirmed, a new data acquisition campaign should be purposely designed so that a stronger conclusion could be obtained.

10. Conclusions and perspectives

This study depicted the main issues for the conservation of semi-arid savannas, and showed that estimating the population of herbivores is an important concern for farmers and managers of conservation reserves. The traditional methods for animal census are too expensive and laborious to be used regularly and to serve as a basis for data driven management. In this context, introducing UAVs as a new tool for land managers, especially for animal census and the mitigation of poaching, appears as a promising solution.

The literature review showed that UAVs have already been used for animal detection, but only few attempts were made to automate the detection in the pictures. Usually a visual interpretation is made by a human. If this is possible for punctual estimations of populations, and when the dataset is small, it strongly discourages the use of UAVs for frequent census.

The detection system implemented in this study was based on two simple methods for objects proposals and showed that an edge detector provides better proposals than a shadow detector.

The Bag of Visual Words was used as a feature extractor. Tuning the parameters was not an easy task, because classifying the images into visual words is time consuming, and assessing the quality of a particular set of parameters requires to use the obtained features in a classification task. In the end, it turned out that the much simpler histogram of colors performs better than the histogram of words. However, using both types of features together gave the best results, meaning that the HOW still adds discriminative information.

The influence of the image resolution was also analyzed at this stage. A ground sampling distance (GSD) of 16 cm proved to be too high, but the advantage of using a GSD of 8 cm over 12 cm becomes less clear. The computational burden linked with very high resolutions inhibited the use of the original, 4 cm resolution. This means that the UAVs could fly at higher altitude than they did (around 130 m).

In order to address the particular characteristic of this dataset (high intra-class heterogeneity and strong underrepresentation of the class *animals*), exemplar SVMs have been used for classification. Following the recent work of (Kobayashi, 2015), the exemplars were trained as one-class SVMs and fitting the cost parameter was avoided. After rescaling the scores produced by the ensemble of exemplar SVMs, it appeared that the simple rule of retaining the maximal score gave satisfactory results.

The advantage of exemplar SVM became clear as it was observed that the detected animals match very closely with the exemplars that give them the highest score. In particular, this means that the species of the detected animals could be retrieved, if the ground truth contained information about the species.

Hard negative mining did not allow to improve the results, but active learning was successfully used to increase the precision and retrieve a significant number of animals from the training set, that were wrongly labeled as background.

The dataset was split into two subsets according to the time of the day when the pictures were acquired. The analysis indicated that the classification of the images acquired in the morning was much easier. Furthermore, the exemplars from the morning subset gave better results than those of the midday subset, whatever subset was used for testing. A strong limitation of this analysis lays on the fact that the flights done in the morning and at midday were done in different locations, so that it may be biased by different types of land covers.

This study demonstrated that automated detection of large herbivores in semi-arid savanna is possible, even with simple RGB cameras. But at this stage the intervention of a human observer to verify the detections and discard the false positives is still needed. It is likely that better results can be obtained by using better, or simply more features, and fine-tuning some of the parameters of the model.

Before envisioning UAVs for regular and fully automated counts of animals, it is needed to consider each of the steps presented in Figure 26. This study delt only with information extraction, while a considerable amount of work is also needed before and after that. For data collection, UAVs with a longer autonomy of flight would be required. Longer and more frequent flights mean even more pictures, which should be properly organized and stored thanks to a database management system. Once animals are detected in the images, the following step is to estimate the population density, with the help of appropriate statistical methods. Here the issue of double-counting animals in consecutive images will need to be addressed, and more generally the statistical framework that allows computing population densities from incomplete counts should be integrated. The maps that are produced at this step must be meaningful to the land managers, so that they can easily use them to improve their management practices. A training course to help them changing their working method and operating the whole system may be required.

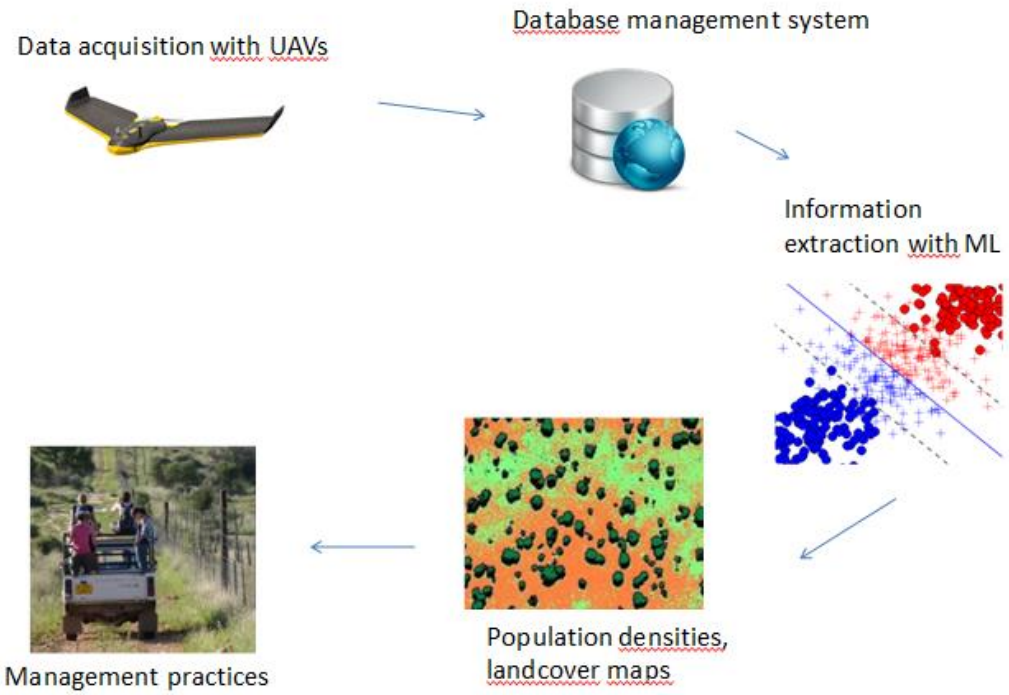

Figure 26 : Steps to be considered, from data acquisition to improved management practices

In conclusion, if the rapid advances of machine learning keep the same pace in the coming years, detecting animals in the images will not be a bottleneck. The challenges will be on the side of data acquisition and management, and about integrating the information extracted from the images into the management practices.

References

Burnham, K. P., Anderson, D. R., & Laake, J. L. (1982). Estimation of Density from Line Transect Sampling of Biological Populations. *Biometrical Journal*, *24*(3), 256–256. http://doi.org/10.1002/bimj.4710240306

Caughley, G. (1974). Bias in Aerial Survey. *The Journal of Wildlife Management*, *38*(4), 921–933. http://doi.org/10.2307/3800067

Chabot, D., & Bird, D. M. (2012). Evaluation of an off-the-shelf Unmanned Aircraft System for Surveying Flocks of Geese. *Waterbirds*, *35*(1), 170–174. http://doi.org/10.1675/063.035.0119

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.*, *2*(3), 27:1–27:27. http://doi.org/10.1145/1961189.1961199

Dugill, A. (1995). Land Degradation and Grazing in the Kalahari: New Analysis and Alternative Perspectives. *ODI*. Retrieved from https://www.odi.org/publications/4474-livestock-rangeland-ecology-land-degredation-kalahari

eBee: senseFly SA. Retrieved May 12, 2016, from https://www.sensefly.com/drones/ebee.html

Grenzdörffer, G. J. (2013). UAS-based automatic bird count of a common gull colony. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *1*, 169–174. http://doi.org/10.5194/isprsarchives-XL-1-W2-169-2013

Hodgson, A., Kelly, N., & Peel, D. (2013). Unmanned Aerial Vehicles (UAVs) for Surveying Marine Fauna: A Dugong Case Study. *PLOS ONE*, *8*(11), e79556. http://doi.org/10.1371/journal.pone.0079556

Israel, M. (2011). A UAV-based Roe Deer Fawn Detection System. *International Archives of Photogrammetry and Remote Sensing, Vol XXXVIII-1/C22*, 1–5.

Kobayashi, Takumi. 2015. "Three Viewpoints Toward Exemplar SVM." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2765–73. http://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Kobayashi_Three_Viewpoints_Toward_2015_CVPR_paper.html.

Koh, L., & Wich, S. (2012). Dawn of drone ecology: low-cost autonomous aerial vehicles for conservation. *Mongabay*. Retrieved from https://digital.library.adelaide.edu.au/dspace/handle/2440/84717

Kuzikus - Wildlife Reserve Namibia. Retrieved May 11, 2016, from http://www.kuzikus-namibia.de/xe_wildlife_reserve.html

Kuzikus Wildlife Reserve. Visible Biodiversity on Kuzikus. Retrieved from http://www.kuzikus-namibia.de/_wildschutzgebiet/Kuzikus_Species_List.pdf

Linchant, J., Lisein, J., Semeki, J., Lejeune, P., & Vermeulen, C. (2015). Are unmanned aircraft systems (UASs) the future of wildlife monitoring? A review of accomplishments and challenges. *Mammal Review*, *45*(4), 239–252. http://doi.org/10.1111/mam.12046

Maire, F., L. Mejias, and A. Hodgson. 2014. "A Convolutional Neural Network for Automatic Analysis of Aerial Imagery." In 2014 International Conference on Digital Image Computing: Techniques and Applications (DlCTA), 1–8. doi:10.1109/DICTA.2014.7008084.

Malisiewicz, T., Gupta, A., & Efros, A. A. (2011). Ensemble of exemplar-SVMs for object detection and beyond. In *2011 International Conference on Computer Vision* (pp. 89–96). http://doi.org/10.1109/ICCV.2011.6126229

Marsh, H., & Sinclair, D. F. (1989). Correcting for Visibility Bias in Strip Transect Aerial Surveys of Aquatic Fauna. *The Journal of Wildlife Management*, *53*(4), 1017–1024. http://doi.org/10.2307/3809604

Matlab. (2015). (Version R2015b). Natick, Massachusetts, United States: The MathWorks, Inc.

Matlab Image Processing Toolbox. (2015). (Version R2015b). Natick, Massachusetts, United States: The MathWorks, Inc. Retrieved from http://www.mathworks.com/products/image/

McNeilage, A., Plumptre, A. J., Brock-Doyle, A., & Vedder, A. (2001). Bwindi Impenetrable National Park, Uganda: gorilla census 1997. *Oryx*, *35*(1), 39–47. http://doi.org/10.1046/j.1365-3008.2001.00154.x

Mulero-Pázmány, M., Stolper, R., van Essen, L. D., Negro, J. J., & Sassen, T. (2014). Remotely Piloted Aircraft Systems as a Rhinoceros Anti-Poaching Tool in Africa. *PLoS ONE*, *9*(1), e83873. http://doi.org/10.1371/journal.pone.0083873

Ofli, F., Meier, P., Imran, M., Castillo, C., Tuia, D., Rey, N., … Joost, S. (2016). Combining Human Computing and Machine Learning to Make Sense of Big (Aerial) Data for Disaster Response. *Big Data*, *4*(1), 47–59. http://doi.org/10.1089/big.2014.0064

Poaching Statistics. Retrieved February 19, 2016, from https://www.savetherhino.org/rhino_info/poaching_statistics

Pollock, K. H., & Kendall, W. L. (1987). Visibility Bias in Aerial Surveys: A Review of Estimation Procedures. *The Journal of Wildlife Management*, *51*(2), 502–510. http://doi.org/10.2307/3801040

Quang, P. X., & Becker, E. F. (1997). Combining Line Transect and Double Count Sampling Techniques for Aerial Surveys. *Journal of Agricultural, Biological, and Environmental Statistics*, *2*(2), 230–242. http://doi.org/10.2307/1400405

Reinhard, F. (2016, March). Personal communication.

Rhino Poaching Statistics. (2015, January 17). Retrieved from http://www.poachingfacts.com/poaching-statistics/rhino-poaching-statistics/

Ringrose, S., Vanderpost, C., & Matheson, W. (1996). The use of integrated remotely sensed and GIS data to determine causes of vegetation cover change in southern Botswana. *Applied Geography*, *3*(16), 225–242.

Roques, K. g., O'Connor, T. g., & Watkinson, A. r. (2001). Dynamics of shrub encroachment in an African savanna: relative influences of fire, herbivory, rainfall and density dependence. *Journal of Applied Ecology*, *38*(2), 268–280. http://doi.org/10.1046/j.1365-2664.2001.00567.x

Rowcliffe, J. M., Field, J., Turvey, S. T., & Carbone, C. (2008). Estimating animal density using camera traps without the need for individual recognition. *Journal of Applied Ecology*, *45*(4), 1228–1236. http://doi.org/10.1111/j.1365-2664.2008.01473.x

SAVMAP. Retrieved June 16, 2016, from http://lasig.epfl.ch/savmap

Seely, M., & Montgomery, S. (2009). *Proud of our deserts: Combating desertification. An NGO perspective on a National Programme to Combat Desertification*. Windhoek: Desert Research Foundation of Namibia. Retrieved from http://www.drfn.info/docs/napcod/napcod_book_small.pdf

Silver, S. C., Ostro, L. E. T., Marsh, L. K., Maffei, L., Noss, A. J., Kelly, M. J., … Ayala, G. (2004). The use of camera traps for estimating jaguar Panthera onca abundance and density using capture/recapture analysis. *Oryx*, *38*(2), 148–154. http://doi.org/10.1017/S0030605304000286

Trodd, N. M., & Dougill, A. J. (1998). Monitoring vegetation dynamics in semi-arid African rangelands: Use and limitations of Earth observation data to characterize vegetation

structure. *Applied Geography*, *18*(4), 315–330. http://doi.org/10.1016/S0143-6228(98)00024-1

Walker, B. H., Ludwig, D., Holling, C. S., & Peterman, R. M. (1981). Stability of Semi-Arid Savanna Grazing Systems. *Journal of Ecology*, *69*(2), 473–498. http://doi.org/10.2307/2259679

Wall, M. Can drones help tackle Africa's wildlife poaching crisis? Retrieved February 19, 2016, from http://www.bbc.com/news/business-28132521

Where We Fly. Retrieved May 11, 2016, from http://airshepherd.org/where-we-fly/

Wipsea. Retrieved May 15, 2016, from http://www.wipsea.com/