Automated species identification for camera trapping in the Iona Skeleton Coast Trans-Frontier Conservation Area

Paul Berry

Thesis presented in partial fulfilment of the requirements for the degree of Master of Natural Resources Management at the Namibia University of Science and Technology

Supervisor: Dr Morgan Hauptfleisch

Co-Supervisor: Dr Nichola Knox

April 2020

Declaration

I, Paul Edgar Berry, hereby declare that the work contained in this thesis is my own original work and that I have not previously in its entirety or in part submitted it at any university or higher education institution for the award of a degree.

Signature:  Date: 6 April 2020

Retention and Use of Thesis

I, Paul Edgar Berry, being a candidate for the degree of Master of Natural Resource Management, accept the requirements of the Namibia University of Science and Technology relating to the retention and use of theses deposited in the Library and Information Services.

In terms of these conditions, I agree that the original of my thesis deposited in the Library and Information Services will be accessible for purposes of study and research, in accordance with the normal conditions established by the Librarian for the care, loan or reproduction of theses.

Signature:  Date: 6 April 2020

Contents

1 Introduction **1**

1.1 The Iona Skeleton Coast Trans-Frontier Conservation Area 1

1.2 Camera trapping 3

1.3 Computer vision 4

1.4 Problem statement 6

1.5 Aim and objectives 6

1.6 Thesis outline 6

2 Methods **8**

2.1 Study areas 8

2.1.1 The Iona Skeleton Coast Trans-Frontier Conservation Area 8

2.1.2 Etosha Heights private game reserve 11

2.2 Computer vision 12

2.2.1 Constraints 12

2.2.2 Computer vision models 12

2.2.3 Process overview 13

2.2.4 Sampling 13

2.2.5 Manual assessment 16

2.2.6 Training 16

2.2.7 Inference 19

2.2.8 Evaluation 20

3 Results **21**

3.1 Inference 21

3.2 Object detection 22

3.2.1 Humans and vehicles 22

3.2.2 Animals 27

3.2.3 Other 27

3.3 Classifying large mammals by species 28

3.3.1 Size of training set 28

3.3.2 Performance by species 29

3.3.3 Performance by site 29

3.3.4 Performance by day and night 29

3.4 Diel patterns 32

4 Discussion **35**

4.1 Performance 35

4.1.1 Humans and vehicles 36

4.1.2 Animals 37

4.1.3 Factors influencing performance 40

4.2 Implementation 42

4.3 Utility and application 42

4.4 Strengths and weaknesses 44

4.5 Evaluation of the study 45

4.6 Suggestions for future work 46

4.7 Conclusion 47

5 Guidelines for the TFCA **49**

5.1 Potential study areas 49

5.2 Study design and methodology 50

5.3 Types of studies 51

5.3.1 Exploratory studies and general monitoring 51

5.3.2 Species inventories and distributions 52

5.3.3 Movement studies 53

5.3.4 Population estimates 54

5.3.5 Behavioural studies 56

5.3.6 Studies on rare, nocturnal and shy species 56

5.4 Camera trap equipment 57

5.5 Camera placement 58

5.6 Camera settings 59

5.7 Service intervals 59

5.8 Data management 60

5.9 Improving computer vision results 61

5.9.1 Expanding utility 61

5.9.2 Increasing training data 61

5.9.3 Technological advances 62

References **62**

A Glossary **76**

B Online resources used **78**

List of Figures

2.1 Overview of the two study areas: 1) The Iona Skeleton Coast TFCA comprises the Iona National Park in Angola, the Skeleton Coast Park in Namibia, as well as several Communal Conservancies. 2) The Etosha Heights private reserve lies along the southern boundary of the Etosha National Park, Namibia. 9

2.2 The area along the Kunene River, which forms the national border, is of particular interest to wildlife monitoring, given the anticipated movement of game from Namibia to Angola. 10

2.3 A map of the Etosha Heights private game reserve, showing the location of the camera trap sites used in this study. 11

2.4 Diagram summarising the sampling, manual assessment, training and inferences processes used in this study. 14

2.5 Example images from each of the four camera trap sites used in this study. . . . 15

2.6 Examples of images used to train the image classifier. 18

3.1 An example of zebra and giraffe identified in a photograph from the Bergpos camera trap site. 23

3.2 An example of impala identified in a photograph taken at the Bergwater camera trap site. 24

3.3 An example of giraffe identified in a night-time photograph taken at the Fence West 2 camera trap site. 25

3.4 An example of oryx identified in a photograph taken at the Fence near Mopanipos camera trap site. 26

3.5 Number of giraffe photographs for each hour of the day across all sites, as established by observation (ground truth) and prediction (inference). 32

3.6 Number of impala photographs for each hour of the day across all sites, as established by observation (ground truth) and prediction (inference). 33

3.7 Number of oryx photographs for each hour of the day across all sites, as established by observation (ground truth) and prediction (inference). 33

3.8 Number of zebra photographs for each hour of the day across all sites, as established by observation (ground truth) and prediction (inference). 34

List of Tables

2.1 Camera trap makes and models, image resolutions, aspect ratios for the four camera trap sites sampled at the Etosha Heights private reserve. 13

3.1 The number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), recall, precision and accuracy for human-related and animal classes as distinguished by the object detection model. 28

3.2 The number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), recall, precision and accuracy for four large mammal species generated on the basis of three training sets. Set A comprised just over 1 000 images per species, set B comprised 100 original images per class, extended to 1 000 by data augmentation, and set C comprised set A extended to 10 000 images per class by data augmentation. 28

3.3 The number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), recall, precision and accuracy for four large mammal species. . . 29

3.4 The number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), recall, precision and accuracy for four large mammal species by camera trap site. 30

3.5 The number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), recall, precision and accuracy for four large mammal species by site type. 31

3.6 The number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), recall, precision and accuracy for four large mammal species by day and night. 31

List of Abbreviations and Acronyms

ANN Artificial Neural Network

CBNRM Community-Based Natural Resource Management

CCF Cheetah Conservation Fund

CSV Comma Separated Value

CPU Central Processing Unit

CNN Convolutional Neural Network

DDR3 SDRAM Double Data Rate 3 Synchronous Dynamic Random-Access Memory

GB Gigabyte

GHz Gigahertz

GPU Graphics Processing Unit

kB Kilobyte

MB Megabyte

pixel picture element

NACSO Namibian Association of CBNRM Support Organisations

SADC Southern African Development Community

TFCA Trans-Frontier Conservation Area

Acknowledgements

Dr Morgan Hauptfleisch and Dr Nicky Knox supervised this study. I thank them for their guidance and advice during this project, and for supporting me in my goals.

The SCIONA project, aimed at co-designing conservation technologies for the Iona Skeleton Coast Trans-Frontier Conservation Area, and funded by the European Commission (EuropeAid/156423/DD/ACT/Multi), provided financial support for this study.

The staff at Etosha Heights private game reserve in northern Namibia are thanked for the setup and maintenance of camera traps and the collection of photos used in this study.

Dr Willie Brink, Stellenbosch University, and Deon Joubert, Innoventix, South Africa, gave valuable advice on some of the computer vision aspects of the project. I am indebted to them for the time they spent on this.

Dr Hermann Swart, Stellenbosch University, encouraged me to pursue this project and provided feedback on the project proposal.

Jesse Green provided me with a daily dose of humour on Telegram which added some much needed colour to many days of otherwise black-and-white text and numbers.

Robert Hering, University of Potsdam, assisted with the manual classification of photos according to species.

I am grateful to my parents-in-law, Anne and Lothar, for taking care of the little ones on countless occasions so that I could focus on this work.

Lastly, I thank my wife Sylvia for her understanding, patience and support during the course of this project.

Dedication

In memory of Mom, Dad and Mark. I shall cherish for the rest of my life the precious memories of our times together in Etosha, Waterberg and the Namib.

Abstract

The Iona Skeleton Coast Trans-Frontier Conservation Area (TFCA), straddling the border between Angola and Namibia, has suffered through decades of civil war and poaching. While this history has been detrimental to the community of large mammals in the TFCA, data collected on the mammal populations are insufficient to enable effective management. Survey methods such as aerial counts and community-based monitoring have various shortcomings. Therefore camera trapping, which has become important in surveying wildlife worldwide, could become an essential monitoring tool also for the TFCA. However, camera traps tend to capture large numbers of images over short periods of time. The cost and time involved in the manual analysis of such voluminous datasets are the major limiting factors in camera trapping.

Deep learning-based computer vision methods proposed to date to address this problem were found unsuitable for application to camera trapping in the TFCA, being computationally too expensive, requiring specialised hardware and large training datasets, focusing on only one species per photograph or relying on static backgrounds between sequential images. On the other hand, the method developed in this study requires only an entry-level computer and relatively few training data while handling multi-species photos with changing backgrounds. It is able to detect and distinguish between humans, vehicles and four large mammal species of importance in the TFCA, namely giraffe, impala, oryx and zebra.

Trained on images sourced from the web and applied to 4 000 camera trap photos, the system yielded a recall rate of 85.7% in detecting human-related object classes and 59.1% in detecting the presence of animals in camera trap photos. Its precision in detecting animals was 100% while its precision in distinguishing between the four large mammal species was 96.8%. Furthermore, frequency distributions of photographs inferred by computer roughly correlated to published diel activity levels for each of the four mammal species investigated. The method did not prove useful for the monitoring of rare species, however.

Based on the results, the method could be used to filter for photos containing human-related objects as well as animals, and to label or pre-label photos by species. This may make it

useful to monitor anthropogenic disturbance, aid in compiling species inventories, document animal migration, map species distributions and pick out images of species for which population densities are to be estimated. Further work would be needed to test the reliability of computer vision inference as an index of activity levels as well as to develop the ability to monitor rare species successfully.

Conceptual and technical aspects of using camera traps in combination with the proposed computer vision method are discussed for application in the Iona Skeleton Coast TFCA. However, the utility of the method has the potential to extend far beyond the TFCA and could be applied to a wide range of conservation projects.

Chapter 1

Introduction

1.1 The Iona Skeleton Coast Trans-Frontier Conservation Area

Protected areas play a crucial role in the conservation of biodiversity (Bruner et al., 2001; Brooks et al., 2004; Rodrigues et al., 2004; Jenkins and Joppa, 2009; Coetzee, Gaston and Chown, 2014; Gray et al., 2016). However, many areas in need of protection extend beyond political borders (Udvardy, 1975; Olson and Dinerstein, 1998, 2002; Dinerstein et al., 2017). This makes international cooperation as set out in the Convention of Biological Diversity (United Nations, 1992) essential for biodiversity conservation. A regional manifestation of this idea are Trans-Frontier Conservation Areas (TFCAs), large ecological regions that straddle the boundaries of two or more countries and encompass protected and multiple use areas (SADC, 1999). They form an important aspect of cooperation between member states of the Southern African Development Community, which includes among its fundamental objectives the sustainable use of natural resources and the protection of the environment (SADC, 2014).

The Iona Skeleton Coast Trans-Frontier Conservation Area is one of seven formally declared TFCAs in the SADC region (SADC, 2018). The TFCA includes the Namibe Partial Reserve and the Iona National Park in Angola, as well as the Skeleton Coast Park, and several adjoining Communal Conservancies in Namibia. Covering nearly 50 000 km^2 (SADC, 2018), the TFCA includes a large portion of the Kaokoveld Desert ecoregion (Spriggs, 2018), one of 142 terrestrial regions prioritised for biodiversity conservation globally (Olson and Dinerstein, 2002). The region is characterised by a high level of plant endemism (Craven, 2009), several endemic arthropod species (Cloudsley-Thompson, 1990) and a high diversity of amphibians and reptiles (Ceríaco et al., 2016). Of particular concern, however, are the large mammal populations in the Angolan part of the TFCA. While data relating to their conservation are scant, it is generally

known that the effects of the Angolan civil war—which spanned nearly three decades—have been detrimental (Huntley, 2017).

Before the war, the South-West Arid Biome of Angola, covered to a large extent by the Iona National Park, was inhabited by a diverse range of large mammals, including aardwolf *Proteles cristatus*, brown hyaena *Hyaena brunnea*, bat-eared fox *Otocyon megalotis*, Cape fox *Vulpes chama*, Hartmann's mountain zebra *Equus zebra* and Burchell's plains zebra *Equus quagga burchelli*, black rhinoceros *Diceros bicornis*, Damara dik-dik *Madoqua kirkii*, black-faced impala *Aepyceros melampus petersi*, springbok *Antidorcas marsupialis* and oryx *Oryx gazella* (Huntley, 1974). Considerable portions of the ranges of some of these species fall within the TFCA: the mountain zebra is confined to western Namibia and south-western Angola, the black-faced impala to northern Namibia, and the Damara dik-dik population of the southern African subregion is confined to northern Namibia and south-western Angola (Skinner and Chimimba, 2005).

The Angolan civil war (1975–2002) left protected areas in the country, including Iona, vulnerable to poaching (Beja et al., 2019). The consequences for the large mammal populations of the Park were devastating. The black rhino population was destroyed and the same is assumed for the lion population (Morais et al., 2018). The plains zebra was considered locally extinct by 1992, mountain zebra were almost extirpated and the oryx population declined (Beja et al., 2019).

Following the end of the civil war, a wildlife aerial survey of Iona (Kolberg and Kilian, 2003) showed the presence of ostrich *Struthio camelus*, Hartmann's zebra, springbok and oryx. In addition, one leopard *Panthera pardus* and one cheetah *Acinonyx jubatus* were seen, and some observations of jackal *Canis mesomelas* and vulture nests were made. The Cheetah Conservation Fund (CCF) confirmed the presence of cheetah in Iona in 2010 (CCF, 2010). A second aerial survey was done in 2017. It indicated a decrease of ostrich, oryx and springbok numbers since 2003, presumably due to poaching, and showed no sign of predator or scavenger species (Hauptfleisch and Brown, 2017).

While the available evidence points to a decrease of wildlife populations on the Angolan side of the TFCA during the last decades, an increase has been observed on the Namibian side. Although poaching also significantly impacted wildlife populations in Namibian communal areas up to the mid 1990s (NACSO, 2017), the introduction of Community-Based Natural Resource Management (CBNRM) marked a turn around. In 1996, the Government of Namibia enacted legislation (Nature Conservation Amendment Act, 1996) that gives traditional communities the power to create conservancies with which to manage wildlife and tourism on their communal land (Jones, 2010; Naidoo et al., 2016). The establishment of such Communal Conservancies has since lead to a decrease in poaching and recoveries of wildlife populations in these areas

(Weaver and Petersen, 2008).

Given this background, wildlife monitoring in the TFCA is important for a number of reasons. Firstly, there is a paucity of data on populations and their distributions, particularly on the Angolan side (Huntley, 1974; Kolberg and Kilian, 2003; Hauptfleisch and Brown, 2017). Secondly, trophy hunting and meat harvesting in the Communal Conservancies on the Namibian side need to occur at a sustainable level (Weaver and Petersen, 2008). Thirdly, little is known about the migration of species across the Kunene River and knowledge about this may inform the possible re-stocking of wildlife to Angola from Namibia (Kuedikuenda and Xavier, 2009).

However, the TFCA is large and encompasses areas that are remote and difficult to access. This makes wildlife monitoring a challenge. In addition, different monitoring methods have various shortcomings. Aerial censuses are very expensive and in many cases unaffordable (Lindeque and Lindeque, 1997; Saltz et al., 2004). They have therefore been done only twice in the TFCA (Kolberg and Kilian, 2003; Hauptfleisch and Brown, 2017). Also, aerial surveys are ineffective when population densities are very low, animals are highly mobile and not adequately visible from the air (Lindeque and Lindeque, 1997). In the case of carnivores, aerial surveys are effective only in relatively sparsely vegetated habitat (Gese, 2001), whereas the vegetation in the dry riverbeds of the TFCA provide opportunities for predators to hide. Furthermore, Reilly and van Hensbergen (2002) showed a decline in the number of aerial observations with time, which they ascribe to among other factors, observer fatigue. Another method, Community-Based Monitoring, requires ongoing training as well as technical and material support, and is dependent on the motivation of local participants; aggregating data in sparsely inhabited areas is challenging, financial resources are limited, illiteracy and the generation of fake data are further hurdles (Stuart-Hill et al., 2005). In addition, the findings of Saltz et al. (2004) suggest that population sizes cannot be reliably estimated using conventional ground-based techniques.

1.2 Camera trapping

Camera traps are remotely activated cameras that are either set to take photographs at certain time intervals or are triggered by movement, which is typically detected by an infra-red light sensor (Swann, Kawanishi and Palmer, 2011). Camera traps are used throughout the world to monitor wildlife populations (O'Connell and Nichols, 2011). They have enabled efficient studies of medium-sized and large mammals (Stein, Fuller and Marker, 2008) with regard to their distribution, abundance and behaviour (Burton et al., 2015). They have been instrumental in documenting species new to science, or that occur in areas were they were previously not known

to exist, or thought be locally extinct (Swann and Perkins, 2014). Camera traps have also been instrumental in estimating populations of species where individuals can be recognised based on individual markings (Karanth and Nichols, 1998). Methods have recently been proposed to extend this to unmarked species, though these are subject to model assumptions (Rowcliffe et al., 2008; Chandler and Royle, 2013; Lucas et al., 2015; Howe et al., 2017; Nakashima, Fukasawa and Samejima, 2018). The widespread adoption of camera traps can be attributed to the fact that they are versatile in their application, cost-effective and relatively non-invasive (Swann and Perkins, 2014).

Given the problems associated with the alternative survey methods mentioned above, camera trapping could well become an important complementary monitoring method for the TFCA. Camera traps have already been used to compile species lists for mammals in central Angola (Taylor et al., 2018) and the Luengue-Luiana and Mavinga National Parks, as well as to estimate leopard density (Funston et al., 2017). They could potentially be instrumental in documenting species presence, compiling species lists, establishing species ranges for the TFCA, and estimating population sizes of individually marked species, such as cheetah (Marker, Fabiano and Nghikembua, 2008), leopard (Funston et al., 2017), possibly spotted hyaena, giraffe (Halloran, Murdoch and Becker, 2015) and zebra (Lahiri et al., 2011). Also, in contrast to other methods, camera trapping provides object records, so that observations can be independently verified and analysed (Caravaggi et al., 2017).

However, camera trap projects tend to amass large numbers of photographs within relatively short periods of time. For instance, 3.2 million images were collected from 2010–2013 in the Snapshot Serengeti project (Swanson et al., 2015). The costly and time-consuming processing of large volumes of photographs has become the limiting factor of this monitoring method (Harris et al., 2010). One solution involving manual analysis has been the use of crowd-sourcing (Swanson et al., 2015). Nevertheless, whether done in-house or outsourced to the crowd, annotating large sets of images requires large amounts of human labour. Recent developments in computer vision promise an alternative.

1.3 Computer vision

Computer vision refers to the ability of computers to perform tasks similar to those of the human visual system. The field of computer vision has seen significant advances in recent years, particularly since an artificial intelligence approach referred to as *deep learning* was applied to the field in 2012 (Krizhevsky, Sutskever and Hinton, 2012).

Artificial intelligence describes the ability of machines (digital computers) to perform tasks commonly associated with humans, such as to learn, to understand, to generalise and to reason (Copeland, 2019). A branch of artificial intelligence called machine learning focuses on the automated detection of patterns in data (Shalev-Shwartz and Ben-David, 2014). One approach to machine learning is to let computers learn from experience and make sense of the world through a hierarchy of concepts—building complicated concepts from simpler ones (Goodfellow, Bengio and Courville, 2016). A widespread architecture enabling this is the Artificial Neural Network (ANN), inspired by biological brains (Hassabis et al., 2017). The basic building block of an ANN, the artificial neuron, loosely models a biological neuron. Each neuron receives inputs, processes them through a non-linear activation function and generates an output (Aggarwal, 2018). In ANNs, neurons are arranged into layers, with the neurons from one layer being connected to neurons in a succeeding layer. Each layer abstracts information from preceding layers. To perform complex tasks (as is required for instance in computer vision) an ANN consist of many layers. This depth of layers in ANNs has led to the term *deep learning* (Goodfellow, Bengio and Courville, 2016).

Recently, approaches to the visual analysis of camera trap photos using deep learning convolutional neural networks (CNNs) have been introduced (Norouzzadeh et al., 2018; Schneider, Taylor and Kremer, 2018; Nguyen et al., 2017; Yousif et al., 2019), but various limitations make them unsuitable for application to the TFCA. Norouzzadeh et al. (2018), using *image classification*, assign a single label to an entire photograph and so only consider photographs containing a single species. In contrast, Schneider, Taylor and Kremer (2018) use *object detection*, a method able to detect and label multiple objects in an image, to recognise several species per photograph. However, to prepare training sets for this approach involves the laborious drawing of bounding boxes around, and labelling of, objects of interest in hundreds or thousands of images. Also, specialised computer hardware (a high-end Graphics Processing Unit) or cloud computing is needed for training object detection models due to the computational expense involved. Nguyen et al. (2017) propose a two-step process, wildlife detection and wildlife identification, employing CNN-based image classification models. However, the computational expense of their approach is also high and newer models have become established since. Yousif et al. (2019) reduce computational expense by isolating animals from image backgrounds using background subtraction. This approach, however, requires that backgrounds remain static between consecutive photographs, which is not necessarily the case when time intervals between photographs are large. There is thus the need for an automated method for the analysis of camera trap photographs which is computationally inexpensive and requires relatively little manual work, while being able to deal with multiple species and changing image backgrounds.

1.4 Problem statement

Wildlife monitoring in the vast and remote Iona Skeleton Coast Trans-Frontier Conservation Area has been insufficient to date. Therefore, camera traps promise to be a valuable monitoring tool for the TFCA given their versatility and remote operation, but the cost of manually analysing large numbers photographs is a major limitation. This calls for a practical method to automatically analyse image content. Methods developed so far are computationally too expensive, require too much manual pre-processing, focus on single-species photographs or rely on image backgrounds that are static between consecutive images. A method suited to the constraints and limitations of camera trapping in the TFCA was lacking. There is further a need for the systematic use of camera traps as a monitoring tool to replace the ad-hoc nature of camera trap use in wildlife monitoring often practised in Namibia and elsewhere.

1.5 Aim and objectives

The aim of the study was to investigate the potential of camera traps in combination with computer vision to monitor the presence of key wildlife species in the TFCA.

The objectives were:

- to develop a computationally inexpensive computer vision method for the automatic recognition of multiple species in camera trap images;
- to apply the method to a pilot study to determine the efficacy of monitoring the presence of large wildlife species in a wildlife management area in which a camera trap grid is already in place;
- to develop a set of guidelines for camera trap-based wildlife monitoring in the Iona Skeleton Coast TFCA, based on the findings of the above two objectives and current literature.

1.6 Thesis outline

This thesis is structured as follows. This Chapter serves as an introduction to the study, giving some background on the TFCA, its history, the importance of wildlife monitoring in this context, outlining the current state of camera trapping and computer vision, as well as stating the study aim and objectives. Chapter 2 describes the study areas as well as the methods relating to the

computer vision approach developed in this study in detecting animals and recognising species. Chapter 3 gives examples of the capabilities and limitations of the computer vision method and reports on its performance in terms of *accuracy*, *precision* and *recall* metrics. Chapter 4 discusses, on the basis of the aforementioned results, the performance, implementation, utility and application, as well as the strengths and weaknesses of the computer vision method, and gives suggestions for future work. Chapter 5 provides guidelines and recommendations aimed at increasing the effectiveness of camera trapping in combination with automated visual processing for the Iona Skeleton Coast TFCA and similar conservation areas.

Chapter 2

Methods

2.1 Study areas

Two study areas were considered in this project. The study was done for the Iona Skeleton Coast Trans-Frontier Conservation Area. Therefore its extent, topography, rainfall, community of large mammals and population distributions are taken into account. However, as camera traps had not yet been deployed in the TFCA at the time of this study, the Etosha Heights private game reserve served as a source of photographs on the basis of which to develop and test the automated recognition system. The community of large mammals in Etosha included the most prevalent species known to occur in the TFCA.

2.1.1 The Iona Skeleton Coast Trans-Frontier Conservation Area

The Iona Skeleton Coast Trans-Frontier Conservation Area (TFCA) is situated along the coastal region of south-western Angola and north-western Namibia (Figure 2.1). The TFCA stretches from 11.72°E to 13.96°E and from 15.21°S to 21.19°S, covering a total area of 47 698 km^2 (SADC, 2018). The TFCA includes the Namibe Partial Reserve and the Iona National Park in Angola, as well as the Skeleton Coast Park and a number of adjoining Communal Conservancies in Namibia.

The area primarily being considered for wildlife monitoring lies in proximity of the Kunene River which forms the border between the two countries (Figure 2.2). On the Angolan side of the Kunene, the southernmost part of the Iona National Park extends c. 175 km inland (c. 10 km upriver of the Epupa Falls) up to the boundary of the Namibe and Cunene Provinces of Angola. On the Namibian side of the Kunene lie the northernmost part of the Skeleton Coast

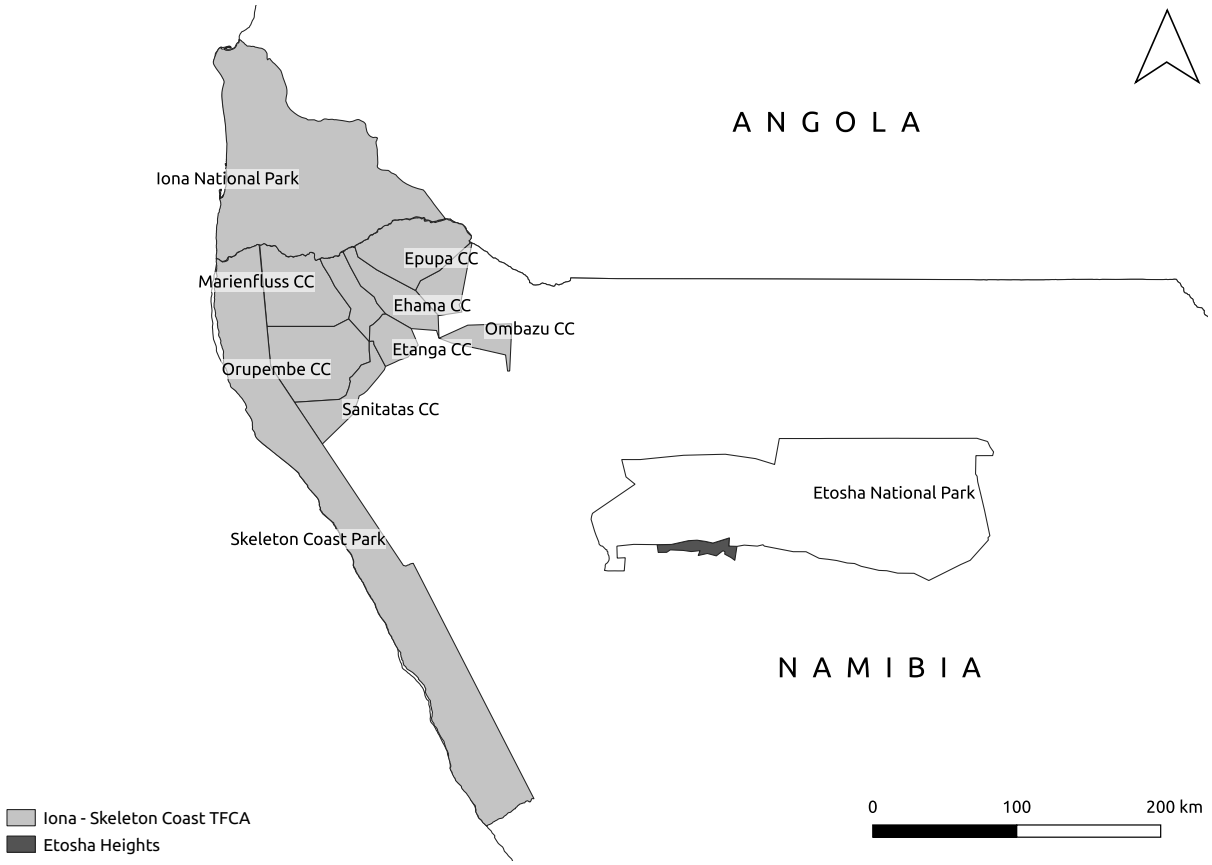


Figure 2.1: Overview of the two study areas: 1) The Iona Skeleton Coast TFCA comprises the Iona National Park in Angola, the Skeleton Coast Park in Namibia, as well as several Communal Conservancies. 2) The Etosha Heights private reserve lies along the southern boundary of the Etosha National Park, Namibia.

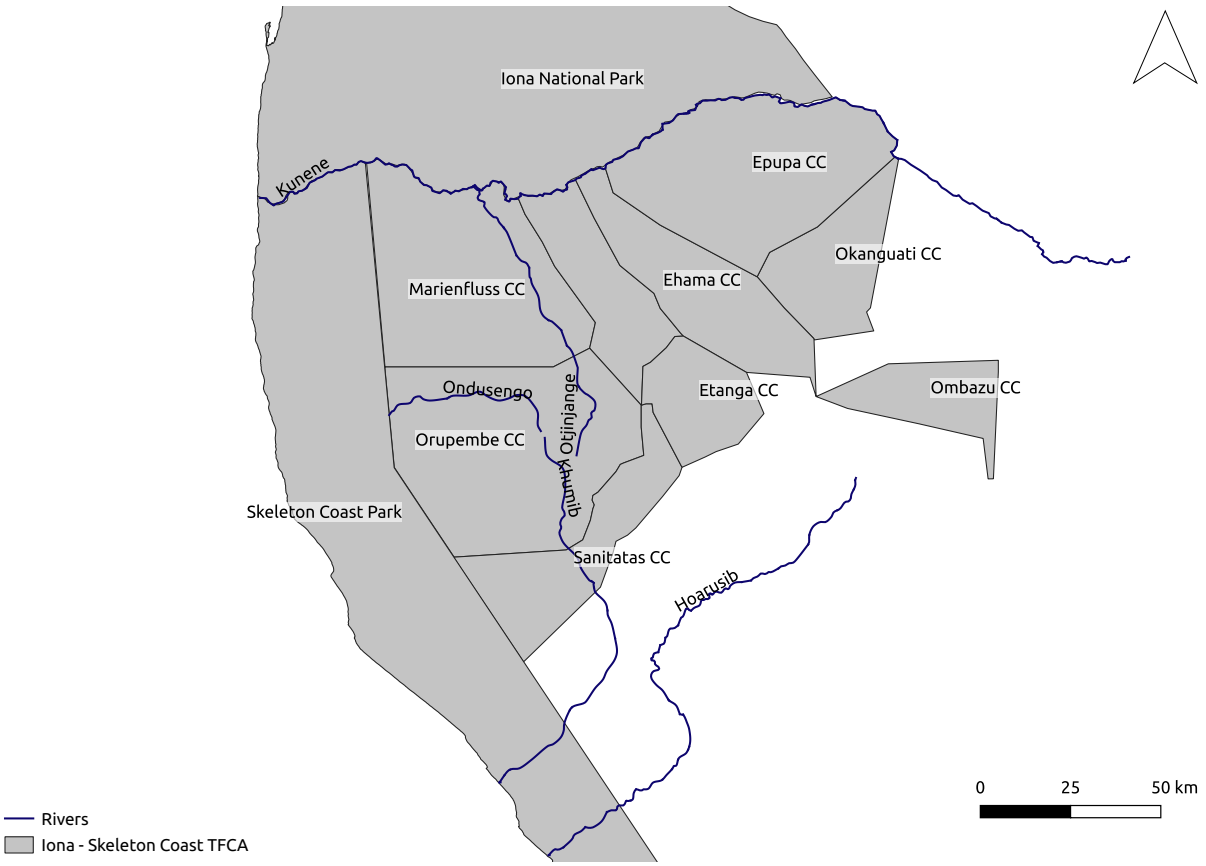


Figure 2.2: The area along the Kunene River, which forms the national border, is of particular interest to wildlife monitoring, given the anticipated movement of game from Namibia to Angola.

Park as well as several Communal Conservancies, the latter extending c. 190 km inland from the coast.

Rainfall tends towards zero at the coast and increases sharply towards the east (Huntley, 1974), the easternmost part of Iona National Park receiving a mean annual rainfall of around 300 mm (Dean, 2001). The stretch of River from the Kunene mouth until c. 50 km inland is extremely arid and dominated by gravel plains to the north and sand dunes to the south. Further east, terrain on both sides of the river becomes mountainous. Aerial photography shows numerous valleys and ravines containing stands of vegetation (Google Maps, 2018).

The predominant large mammal species known to inhabit the Iona National Park are oryx *Oryx gazella*, springbok *Antidorcas marsupialis* and Hartmann's mountain zebra *Equus zebra hartmannae* and are mostly confined to the plains in the southern part of the Park (Hauptfleisch and Brown, 2017). These are also among the most well represented mammal species recorded in game counts for north-western Namibia (NACSO, 2017).

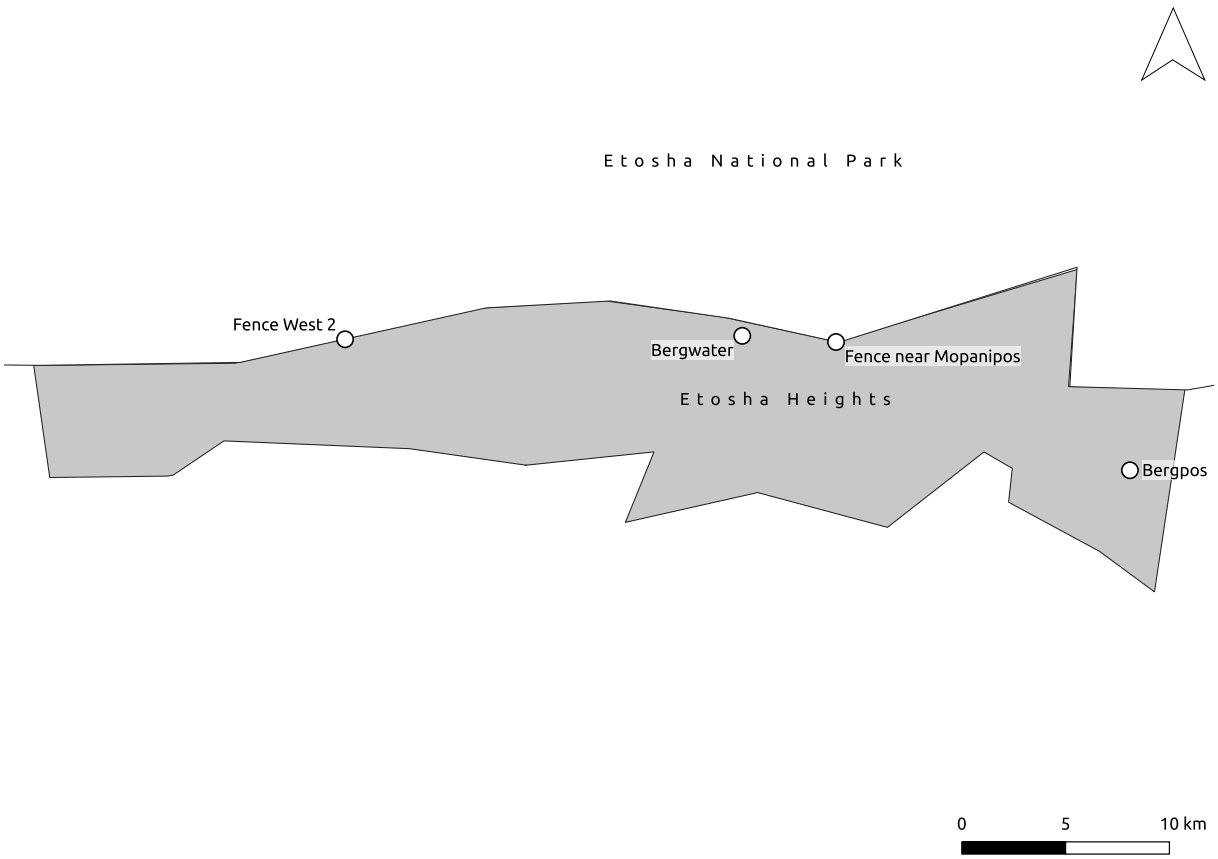


Figure 2.3: A map of the Etosha Heights private game reserve, showing the location of the camera trap sites used in this study.

2.1.2 Etosha Heights private game reserve

Situated in northern Namibia, the Etosha Heights private game reserve lies adjacent to the southern boundary of the Etosha National Park. The reserve is centred on 15.08°E and 19.24°S and covers an area of 495 km^2.

The area features two major geological substrates, Otavi dolomite and Etosha calcrete. Vegetation at Etosha Heights comprises plains of grassland, shrubland and low woodland ("Etosha plains"), while slightly higher ground is covered with open tree and shrub savanna with interspersed patches of woodland ("Etosha mixed low trees"), whereas dolomite hills are vegetated with lower and taller trees, shrubs and grasses (Mendelsohn, El Obeid and Roberts, 2000).

The reserve contains several water points for wildlife and supports a diverse community of large mammals, including, but not restricted to, the large mammal species commonly found in the TFCA. A map of the Etosha Heights game reserve showing the location of the four camera trap sites used in this study is given in Figure 2.3. The camera trap sites are described in section 2.2.4.

2.2 Computer vision

2.2.1 Constraints

The state-of-the-art methods for the computer vision tasks relevant to this study—object detection and image classification—are deep learning-based (Redmon and Farhadi, 2018; Szegedy et al., 2015). Deep learning neural networks typically require specialised, high-performance hardware, usually in the form of a high-end Graphics Processing Unit (Goodfellow, Bengio and Courville, 2016). Budgetary and organisational constraints of the research project funding this study did not allow the procurement of a specialised GPU computer, however. Also, funds for the management of the TFCA are expected to be limited in future. Therefore, an entry-level laptop (2015 model) equipped with a dual-core 2 GHz CPU (Intel® Core™ i3-5005U) and 8 GB of memory (DDR3 SDRAM) was used for all work (including the training of an image classification model as well as inference using object detection and image classification models) as described below. Another constraint was the limited availability of, and capacity to prepare, training images. Given these constraints, a solution was sought which is computationally lightweight and required relatively small training datasets.

2.2.2 Computer vision models

The automated recognition of species in camera trap photos entailed the use of two computer vision models used in sequence:

1. An object detector was used to localise objects of interest in photographs and to distinguish between humans, vehicles and animals. The decision of using an object detector was made based on preliminary investigations into using only an image classifier to label camera trap photos by species. This approach showed little promise, particularly when applying image classification to sites it had not been trained on, matching the findings of Beery, van Horn and Perona (2018). A lightweight, high-performance object detector, YOLO v3, (Redmon and Farhadi, 2018)) was chosen. After experimenting with different input resolutions, the default resolution of 416×416 pixels for input images was chosen, as it seemed to provide the best trade-off between computation time and performance. At this resolution, computation speed on the hardware used in this study was approximately 2 seconds per photograph.

2. An image classifier was used to identify by species the animals localised by the object

detector. The Inception v3 image classification model was selected given its relatively high performance and speed (Bianco et al., 2018). The default resolution of 299 × 299 pixels for input images was chosen. Computation speed for the image classifier was also about 2 seconds per photograph.

2.2.3 Process overview

An overview of the process followed in this study is given in Figure 2.4. Four main tasks were performed, namely, 1) sampling of camera trap photos, 2) manual assessment of species in camera trap photos, 3) training of the image classifier and 4) inference on the sample of camera trap photos. The following sections provide descriptions of these tasks.

2.2.4 Sampling

In selecting camera trap sites in the Etosha Heights reserve, a distinction was made between waterholes and game trails. This was considered important, as higher animal densities were expected at waterhole sites than at game trail sites, potentially influencing species detection rates and identification success.

Two waterhole sites, namely "Bergpos", "Bergwater", and two sites at fence-crossing game trails, namely, "Fence near Mopanipos" and "Fence West 2" were selected (Figures 2.3 and 2.5). A sequence of 1 000 photographs was sampled from each site, resulting in a total sample size of 4 000 images. Sequential images were used so that results could be aggregated over time periods.

Three camera models were used across the four sites, differing in image resolutions and aspects ratios (Table 2.1). Day-time images were captured in colour while night-time images were captured in greyscale due to the use of infra-red flash. The cameras were motion-triggered by infra-red sensor.

Table 2.1: Camera trap makes and models, image resolutions, aspect ratios for the four camera trap sites sampled at the Etosha Heights private reserve.

Site	Camera Make and Model	Image Resolution	Aspect Ratio
Bergpos	Spypoint BF8	3264 × 2448	4:3
Bergwater	Moultrie M-880	3840 × 2160	16:9
Fence near Mopanipos	Spypoint FORCE 11D	3840 × 2880	4:3
Fence West 2	Spypoint FORCE 11D	2560 × 1920	4:3

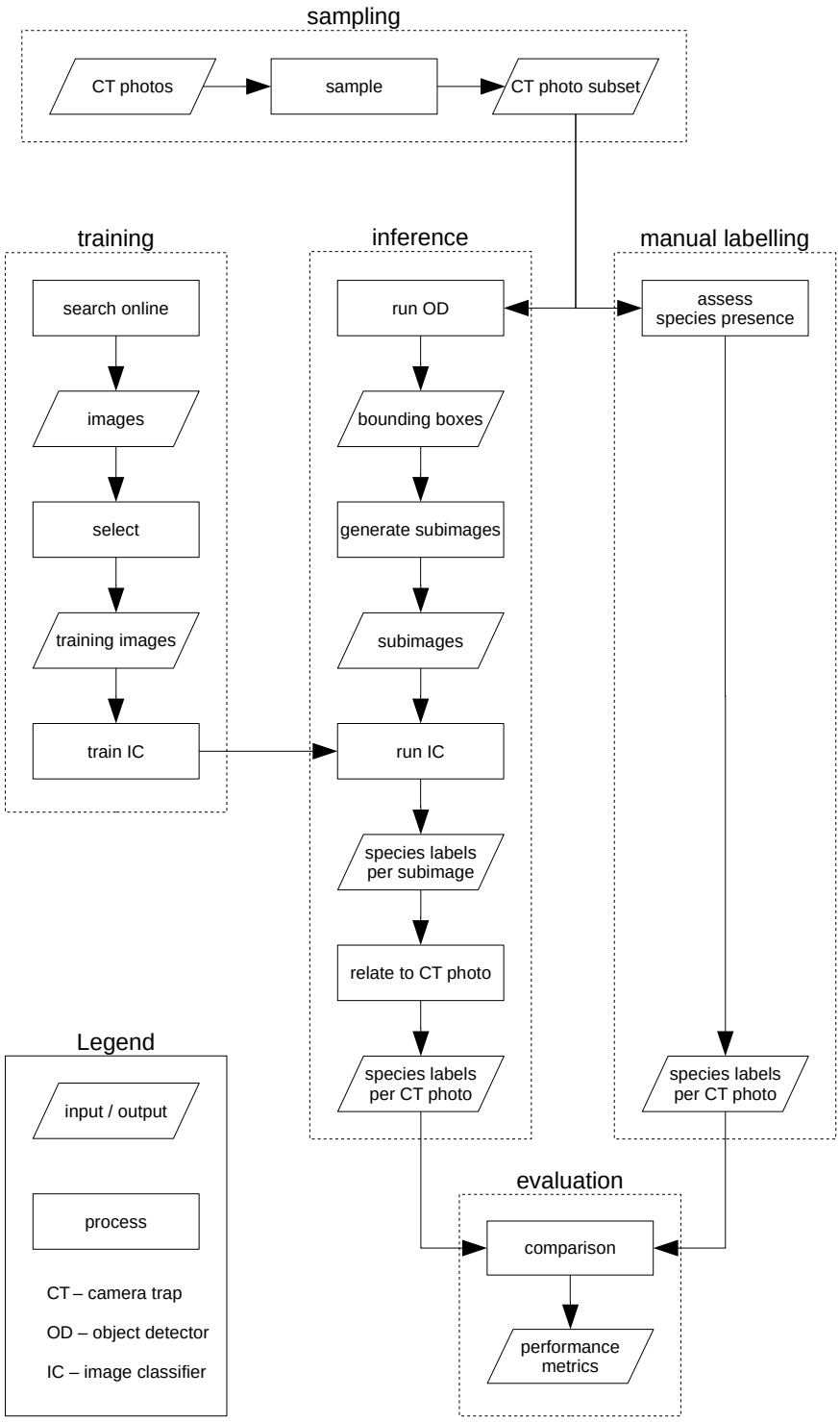


Figure 2.4: Diagram summarising the sampling, manual assessment, training and inferences processes used in this study.

(a) Bergpos

(b) Bergwater

(c) Fence near Mopanipos

(d) Fence West 2

Figure 2.5: Example images from each of the four camera trap sites used in this study.

2.2.5 Manual assessment

Species presence was manually assessed for each of the 4 000 camera trap photos sampled. Following this assessment, six object categories were defined, namely humans, vehicles, giraffe, impala, oryx and zebra. Humans and vehicles were of interest as indicators of anthropogenic disturbance. Springbok, oryx and zebra are the predominant large mammal species recorded in the TFCA (Hauptfleisch and Brown, 2017). However, as springbok were not contained in the sample of camera trap photos, impala were used as a proxy. Giraffe were included given the focus on giraffe conservation in the area (https://giraffeconservation.org/). Carnivorans (black-backed jackal *Canis mesomelas*, brown hyaena *Hyaena brunnea* and spotted hyaena *Crocuta crocuta*) were poorly represented in the sample. Most of the carnivore images were taken at night, with the subjects poorly lit and blurred. Given this, carnivores were not detected by the computer vision system. Ostrich *Struthio camelus* would have been another terrestrial vertebrate of interest in the TFCA, but were not contained in the images sampled from Etosha Heights.

2.2.6 Training

Object detector

Although inference with the YOLO object detector was feasible on the computer hardware used, training on this hardware proved to be too slow. Another challenge associated with training an object detection model is that it requires bounding boxes to be defined for objects found in the training set—a labour-intensive task. Thus a YOLO (version 3) object detection model which had already been trained on the Open Images dataset (Kuznetsova et al., 2018) and was available for download (see Appendix B for URL) was used to localise objects within the photographs.

Image classifier

To distinguish between the four large mammal species on which the study was focused, an image classifier needed to be retrained for this purpose. For a given set of training images and training parameters, retraining is essentially a once-off process.

Training a convolutional neural network (CNN) from scratch requires large labelled datasets and high computational power. In practice therefore, instead of training from scratch, a CNN pre-trained on a related task is repurposed for the task at hand, a technique known as transfer learning (Yosinski et al., 2014). Here, an Inception v3 model (Szegedy et al., 2015) trained on

the ImageNet dataset (Deng et al., 2009) was used as a feature extractor. A softmax classifier was trained on a suitable dataset (see below) to map the output of the feature extractor to the mammal species relevant to the study.

To train the image classifier used in this study, the aim was to prepare a set of training images with at least 1 000 images for each of the following large mammal species: giraffe *Giraffa camelopardalis*, oryx *Oryx gazella*, impala *Aepyceros melampus* and plains zebra *Equus quagga burchelli*. The visual differences between common impala and black-faced impala, as well as between mountain zebra and plains zebra were considered negligible for the purposes of image classification. Using conventional search engines, a few thousand images were collected online for each of these four species. From these collected images, images suitable for training were manually selected using the following criteria: 1) Only the species corresponding to the object class in question is present in the image and, 2) the animal is situated in the foreground, occupying at least a third of the image width and a third of the image height. As the images were of low resolution, they could be displayed as large "thumbnails" in a file browser without losing much visual information. This way, images could be rapidly browsed and unsuitable ones marked for deletion. This resulted in 1 462 images of giraffe, 1 223 of impala, 1 513 of oryx, and 1 093 of zebra, giving a total of 5 291 training images for the four species. The images, being presented in "thumbnail" size by the search engines, were relatively small. Images ranged in resolution from 183×122 pixels to 474×761 pixels. These images were resized (resampled) to 299×299 pixels for input to the image classifier. Training images ranged in file size from 3 kB to 139 kB. The training set was therefore compact, totalling only 111 MB of information. Examples of training images are given in Figure 2.6.

To investigate the effect of training set size on performance, the image classifier was trained on each of three sets of images. The first set (A) comprised all training images and amounted to more than 1 000 images per species (see above). The second set (B) comprised subsets of 100 randomly selected images per species. These were then extended to 1 000 per species using data augmentation techniques, involving a random combination of horizontal flipping, cropping, scaling or brightness variation, and the introduction of Gaussian noise. The third set (C) of training images comprised the original complete set (A), extended however to 10 000 images per species using the same data augmentation techniques.

The image set used for training was randomly split into 80% training set (used to adjust weights in the neural network), 10% validation set (used to avoid over-fitting by ensuring that the accuracy over the training data is matched to the accuracy over the validation data) and 10% test set (used for testing the final solution to confirm the predictive power of the network). The

(a) giraffe Giraffa camelopardalis © *Bernard Dupont / CC BY-SA 2.0*

(b) oryx Oryx gazella © *Charles J Sharp / CC BY-SA 4.0*

(c) black-faced impala Aepyceros melampus petersi © *Charles J Sharp / CC BY-SA 4.0*

(d) Burchell's zebra Equus quagga burchelli © *Gusjer / CC BY-2.0*

Figure 2.6: Examples of images used to train the image classifier.

following default values were used: a learning rate of 0.01 and training and validation batch sizes of 100 images each.

The Tensorflow (version 1.13.1) open source machine learning platform was used in combination with a freely available Python script for image classifier retraining (see Appendix B for URL of script). Some 4 000 training iterations were performed on the model at which point accuracy and cross-entropy did not further improve. Retraining the image classifier took approximately one hour on the entry-level laptop used (see section 2.2.1).

2.2.7 Inference

Step 1: Object detection

The object detection model was run using the inference module contained in the OpenCV library (Bradski, 2000). This library was specifically chosen for its speed. Being nine times faster than its contester (Nayak, 2018), it made object detection on a CPU viable for this study. Input images were resized (resampled) to 416×416 pixels. This resolution has been regarded as being a good trade-off between performance and processing time (Redmon and Farhadi, 2018), though different resolutions (multiples of 32) would be possible. Output confidence thresholds were set at 0.25, 0.05 and 0.00. These low confidence thresholds were chosen given that the YOLO v3 object detection model generates comparatively few false positives (Redmon and Farhadi, 2018). The non-maximum suppression threshold was kept at 0.4 in all cases (Nayak, 2018). The object detector, applied to the Etosha Heights camera trap sample, localised objects in the photographs, outputting corresponding bounding box coordinates for these objects.

The bounding boxes varied greatly in their aspect ratios, particularly for those cases in which only parts of an object had been captured in the photograph. The boxes were thus reshaped to correspond to the aspect ratios of the parent image (4:3 or 16:9) while retaining the centre of the bounding box. The sub-images defined by the resulting boxes thus did not have to undergo excessive distortion to fulfil the requirement of square input images for the image classifier. The sub-images were saved into different folders, depending on whether they were classified as human, vehicle, or animal. All of these processes were automated by Python script.

Step 2: Image classification

The retrained image classifier (see section 2.2.6) was used to distinguish between the four large mammal species of interest to this study. The sub-images of animals generated in Step 1 (object detection, see above) were input to the retrained image classification model (refer to section 2.2.6) at an input resolution of 299×299 pixels (Abadi et al., 2017). Upon classification, the sub-images were automatically moved into different folders, each folder corresponding to a species. As a side-effect, this grouping of object classes and species by folder facilitates relatively easy manual verification.

Lastly, the classified sub-images were related back to the original camera trap photographs, resulting in an automated labelling of the photographs by object class (humans, vehicles and large mammal species). The file names of the camera trap images, together with the labels,

were stored in a CSV file. These tasks were also automated by Python script.

2.2.8 Evaluation

For each of the 4 000 camera trap photographs in the sample, the computer inferred presence or absence of an object class was compared to the manually observed presence or absence of that object class. In this way, the computer inference for each photograph and object class was established as true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

The following performance metrics were calculated:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

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

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

Accuracy indicates the fraction of correctly labelled photographs (including true negatives) with respect to an object class. Recall, sometimes also referred to as Sensitivity, indicates the fraction of photographs actually containing relevant objects that are found. Precision expresses the fraction of photographs said to contain relevant objects that actually do.

Calculation of these metrics was done by main object class (human, vehicle, animal), by species (giraffe, impala, oryx and zebra), by camera trap site and by day versus night. Species other than these four were excluded in the evaluation of species recognition performance.

For each of the four study species, the number of photographs was aggregated by hour of the day. This was done for both manually labelled and automatically labelled photographs to investigate possible correlations with known activity levels.

Chapter 3

Results

3.1 Inference

Example photographs from each of the four the camera trap sites investigated are shown in Figures 3.1, 3.2, 3.3 and 3.4, together with the inferences (predictions) on species presence made by the computer vision system. Some of the abilities and shortcomings of the method are thereby illustrated.

Figure 3.1 shows a photograph taken at the Bergpos camera trap site. Two of the five zebra contained in the photograph were detected, as was the only giraffe, which was backlit. The three zebra not detected illustrate that the object detector in some cases fails to detect relevant objects that are obvious to the eye. Although not all animals in the photograph were detected, both species present were. They were also correctly classified, resulting in the correct labelling of the photograph by species. The water reservoir and drinking trough seen in the photograph were detected in some of the photographs from the Bergpos site. In those cases however, they were invariably classified as inanimate objects by the object detector.

Figure 3.2 shows a photograph taken at the Bergwater camera trap site. Four impala are present in the photograph of which two were detected and correctly classified. Detection of two of the four individuals was sufficient to ensure inference in terms of species presence in the photograph was correct.

Figure 3.3 shows a photograph taken at night at the Fence West 2 camera trap site. The giraffe on the right was automatically detected while the giraffe on the left, apparent to the eye, was not. Detecting one animal of the species present in the photograph and classifying it correctly was, however, sufficient to label the photograph correctly.

Figure 3.4 shows a photograph taken at the Fence near Mopanipos camera trap site. Backlit photographs such as this can make detection and classification more challenging, although the method did succeed in this case. This illustrates that the method in some cases is able to work in poorly exposed, low contrast photographs.

3.2 Object detection

Object detection, the first step in the inference process, yielded 2 968 objects at the default detection threshold of 0.25, and 3 371 objects at detection thresholds of 0.05 and 0.00 in the 4 000 camera trap photographs analysed in this study. Given that the sensitivity of the object detector did not increase when the detection threshold was set below 0.05, the dataset was further analysed with the threshold set at 0.05.

The object detector placed the 3 371 objects detected in 4 000 photographs into nine different classes out of a possible 600 found in the OpenImages dataset. For the purposes of this study, these nine classes were related to four broad categories as follows:

1. Three classes (“Person”, “Human face”, “Personal care”, the latter including sunglasses) were related to humans,
2. two classes (“Vehicle” and “Wheel”) were related to vehicles,
3. one class (“Animal”) was related to animals and
4. three miscellaneous classes were related to some inanimate objects.

The classification of photographs into positives and negatives for each of three categories (”Human”, ”Vehicle” and ”Animal”), together with the performance metrics, is summarised in Table 3.1.

3.2.1 Humans and vehicles

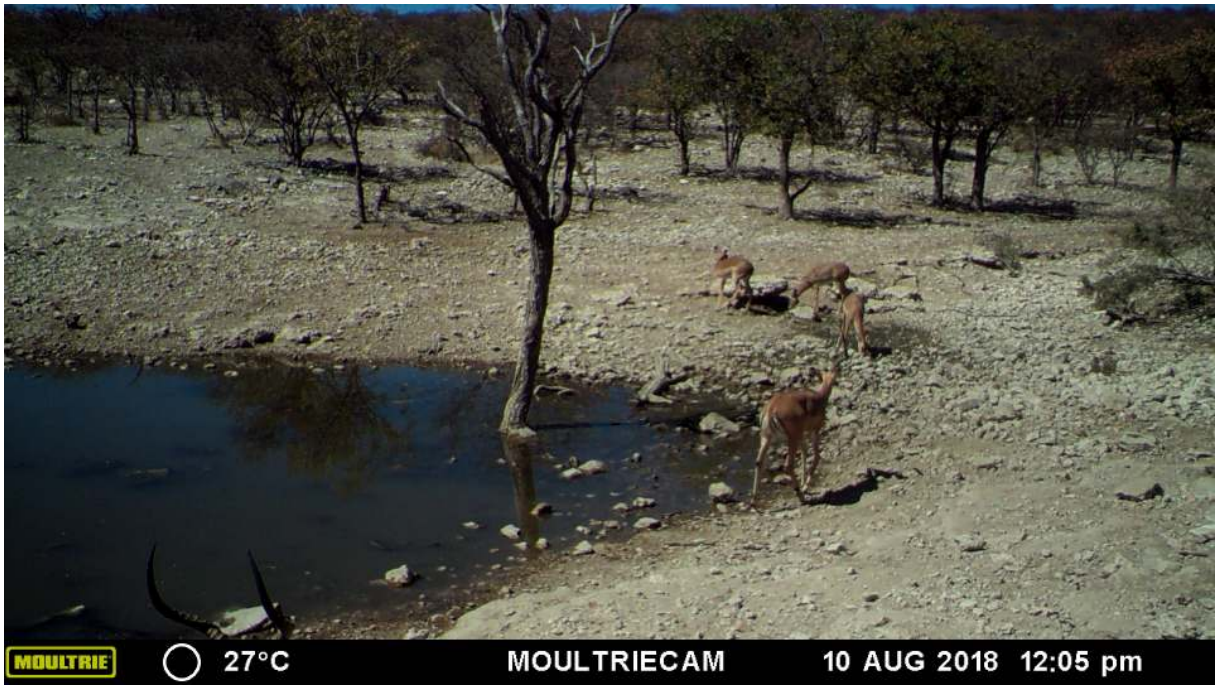
The presence of humans and vehicles in the camera trap photographs was due to staff servicing camera traps. Humans were correctly detected in 8 out of 11 photographs in which they occurred, remaining undetected in 3 photographs. They were falsely detected in 48 photographs. The high proportion of true positives, low proportion of false negatives and high proportion of false positives led to a moderately high recall rate (72.7%), but low precision (14.3%). In many

(a) entire photograph

(b) zebra localised and identified

(c) giraffe localised and identified

Figure 3.1: An example of zebra and giraffe identified in a photograph from the Bergpos camera trap site.



(a) entire photograph

(b) impala localised and identified

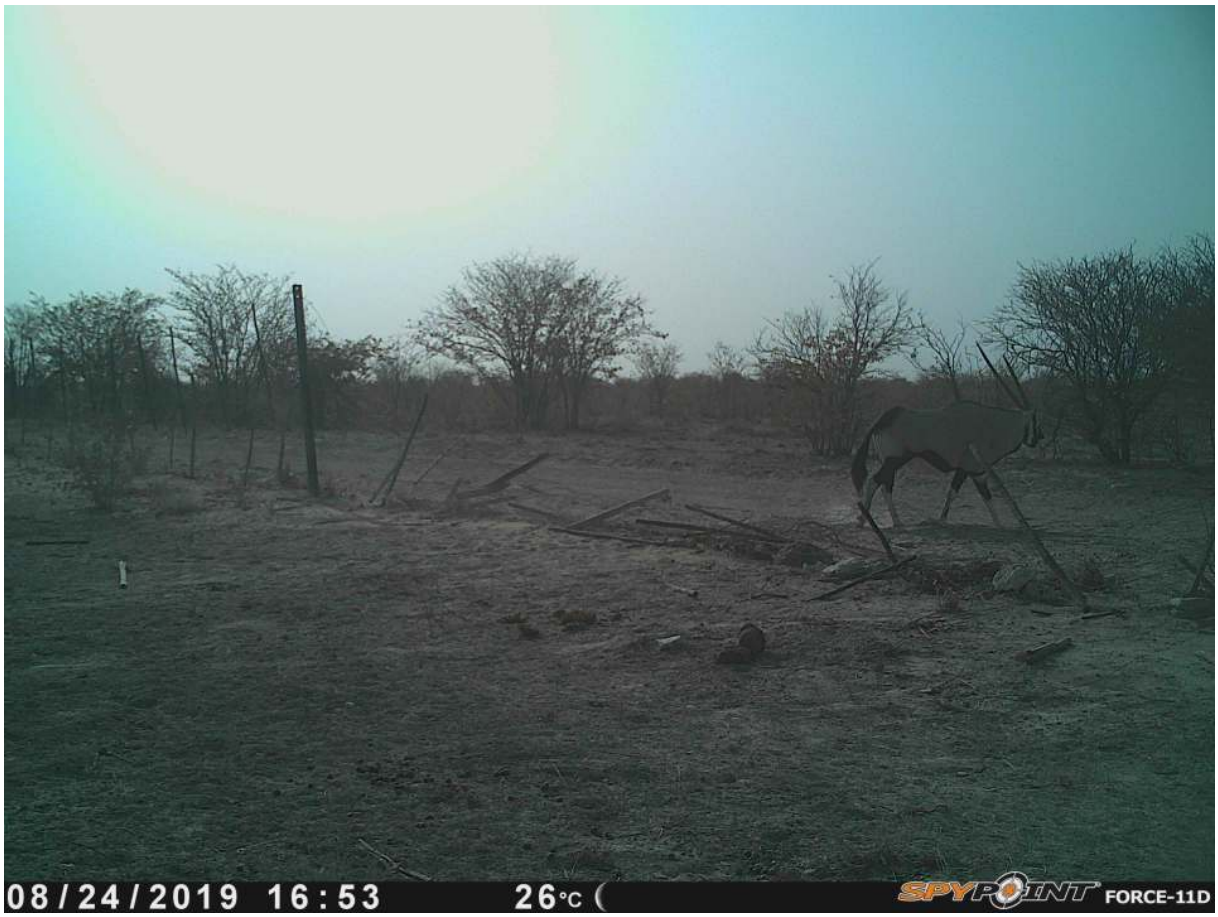
(c) impala localised and identified

Figure 3.2: An example of impala identified in a photograph taken at the Bergwater camera trap site.

(a) entire photograph

(b) giraffe localised and identified

Figure 3.3: An example of giraffe identified in a night-time photograph taken at the Fence West 2 camera trap site.



(a) entire photograph

(b) oryx localised and identified

Figure 3.4: An example of oryx identified in a photograph taken at the Fence near Mopanipos camera trap site.

cases, these false positives were images of the front or rear of animals, the silhouettes of which more closely resemble a human silhouette than lateral views of animals. Vehicles were correctly detected in 22 out of 24 photographs in which they occurred, remaining undetected in 2 photographs. They were falsely detected in 28 photographs. The high proportion of true positives, low proportion of false negatives and high proportion of false positives led to a high recall rate (91.7%), but moderate precision (44.0%). The large proportion of true negatives (photographs correctly labelled as not containing the object class in question) for humans and vehicles resulted in high accuracies for both these object classes (98.7% and 99.3%, respectively).

3.2.2 Animals

Animals were correctly detected in 1 944 out of 3 290 photographs in which they occurred, remaining undetected in 1 346 photographs. Notably, there were no false positive detections of animals. The moderate proportions of true positives and false negatives and the lack of false positives led to a moderate recall rate (59.1%), but perfect precision (100%). The sizeable proportion of false negatives for animals resulted in an accuracy of only 66.3% for this object class. The OpenImages dataset (Kuznetsova et al., 2018) contained at least 34 object classes related to animals. Of these, 29 were mammal species and the rest were generalised groupings such as "Mammal", "Carnivore" and "Animal". Among the mammal species were some also contained in the Etosha Heights dataset, notably "Giraffe" and "Zebra". However, all detected animals were merely classed as "Animal". As described in the Methods chapter, an image classifier was thus used to distinguish between the four large mammal species of interest.

3.2.3 Other

A total of 30 objects were assigned to three miscellaneous classes, the majority of these (28) correctly (true positives) in that they did not belong to any of the other three categories, being of the water storage tank at Bergpos waterhole, and the minority (2) incorrectly (false positives, low quality nighttime images of the front and rear of a rhino).

Table 3.1: The number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), recall, precision and accuracy for human-related and animal classes as distinguished by the object detection model.

Classes	TP	TN	FP	FN	Precision	Recall	Accuracy
human	8	3941	48	3	0.143	0.727	0.987
vehicle	22	3948	28	2	0.440	0.917	0.993
animal	1944	710	0	1346	1.000	0.591	0.663

Table 3.2: The number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), recall, precision and accuracy for four large mammal species generated on the basis of three training sets. Set A comprised just over 1 000 images per species, set B comprised 100 original images per class, extended to 1 000 by data augmentation, and set C comprised set A extended to 10 000 images per class by data augmentation.

Training set	TP	TN	FP	FN	Precision	Recall	Accuracy
A	1112	14002	37	849	0.968	0.567	0.945
B	463	13153	886	1498	0.343	0.236	0.851
C	1091	13989	50	870	0.956	0.556	0.943

3.3 Classifying large mammals by species

3.3.1 Size of training set

The overall performance of the image classifier varied greatly depending on the size of the training set. When the image classifier had been trained on set A (comprising all original training images, more than 1 000 per class and 5 291 in total), moderate recall (56.7%), high precision (96.8%) and high accuracy (94.5%) were achieved in labelling photographs by species. When the image classifier had been trained on set B (comprising 100 original images per class, extended to 1 000 by data augmentation), low recall (23.6%), low precision (34.3%) and relatively low accuracy (85.1%) were achieved. When the image classifier had been trained on set C (set A extended to 10 000 images per class by data augmentation), slightly lower recall (55.6%), precision (95.6%) and accuracy (94.3%) were achieved than with training set A. The performance results based on the three training sets are summarised in Table 3.2.

Further analysis was done on the basis of the image classifier having been trained on set A, as this had achieved the best performance.

Table 3.3: The number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), recall, precision and accuracy for four large mammal species.

Species	TP	TN	FP	FN	Precision	Recall	Accuracy
giraffe	133	3714	14	139	0.905	0.489	0.962
impala	69	3859	6	66	0.920	0.511	0.982
oryx	595	3108	17	280	0.972	0.680	0.926
zebra	315	3321	0	364	1.000	0.464	0.909
overall	1112	14002	37	849	0.968	0.567	0.945

3.3.2 Performance by species

With training set A producing the highest overall metrics for the classification, both accuracy and precision for all four species were above 90% (Table 3.3) while the recall rates ranged from 46.4 to 68.0%. Precision was highest for zebra (100%) and lowest for giraffe (90.5%); recall was highest for oryx and lowest for giraffe. Taking into account both recall and precision, the best overall performance was achieved for oryx.

3.3.3 Performance by site

Overall recall rates were moderate, ranging from 44.5% to 61.7% between sites (Table 3.4). Thus for all sites, species were located in roughly half of the photographs in which the species were actually present. Overall precision was high, ranging from 91.2% to 98.1% between sites. Overall accuracy ranged from 91.1% to 97.3% between sites.

The two sites located at point features (i.e. waterholes) had an overall recall rate of 56.3%, a precision of 97.4% and an accuracy of 92.6%. The two sites located along linear features (i.e. game trails crossing fence lines) had an overall recall rate of 51.5%, precision of 91.2% and accuracy of 97.3% (see Table 3.5).

3.3.4 Performance by day and night

Overall recall rates were substantially higher during the day (64.4%) than at night (30.6%, Table 3.6). Similarly, overall precision was marginally higher during the day (97.4%) than at night (91.2%). This difference is mainly attributed to the difference found in oryx; precision in zebra remained 100% during both day and night. Accuracy barely differed between day (94.5%) and night (94.4%) which is attributable to the high proportion of true negatives in both cases.

Table 3.4: The number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), recall, precision and accuracy for four large mammal species by camera trap site.

(a) Bergpos

Classes	TP	TN	FP	FN	Precision	Recall	Accuracy
giraffe	11	957	1	31	0.917	0.262	0.968
impala	28	960	1	11	0.966	0.718	0.988
oryx	92	800	7	101	0.929	0.477	0.892
zebra	52	863	0	85	1.000	0.380	0.915
overall	183	3580	9	228	0.953	0.445	0.941

(b) Bergwater

Classes	TP	TN	FP	FN	Precision	Recall	Accuracy
giraffe	71	862	9	58	0.887	0.550	0.933
impala	12	939	2	47	0.857	0.203	0.951
oryx	308	592	0	100	1.000	0.755	0.900
zebra	163	698	0	139	1.000	0.540	0.861
overall	554	3091	11	344	0.981	0.617	0.911

(c) Fence West 2

Classes	TP	TN	FP	FN	Precision	Recall	Accuracy
giraffe	22	954	1	23	0.957	0.489	0.976
impala	29	961	2	8	0.935	0.784	0.990
oryx	136	814	4	46	0.971	0.747	0.950
zebra	85	812	0	103	1.000	0.452	0.897
overall	272	3541	7	180	0.975	0.602	0.953

(d) Fence near Mopanipos

Classes	TP	TN	FP	FN	Precision	Recall	Accuracy
giraffe	29	941	3	27	0.906	0.518	0.970
impala	0	999	1	0	0.000	nan	0.999
oryx	59	902	6	33	0.908	0.641	0.961
zebra	15	948	0	37	1.000	0.288	0.963
overall	103	3790	10	97	0.912	0.515	0.973

Table 3.5: The number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), recall, precision and accuracy for four large mammal species by site type.

(a) Camera trap sites situated at point features

Classes	TP	TN	FP	FN	Precision	Recall	Accuracy
giraffe	82	1819	10	89	0.891	0.480	0.951
impala	40	1899	3	58	0.930	0.408	0.970
oryx	400	1392	7	201	0.983	0.666	0.896
zebra	215	1561	0	224	1.000	0.490	0.888
overall	737	6671	20	572	0.974	0.563	0.926

(b) Camera trap sites situated at linear features

Classes	TP	TN	FP	FN	Precision	Recall	Accuracy
giraffe	29	941	3	27	0.906	0.518	0.970
impala	0	999	1	0	0.000	nan	0.999
oryx	59	902	6	33	0.908	0.641	0.961
zebra	15	948	0	37	1.000	0.288	0.963
overall	103	3790	10	97	0.912	0.515	0.973

Table 3.6: The number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), recall, precision and accuracy for four large mammal species by day and night.

(a) day-time

Classes	TP	TN	FP	FN	Precision	Recall	Accuracy
giraffe	106	2344	13	109	0.891	0.493	0.953
impala	67	2437	3	65	0.957	0.508	0.974
oryx	579	1779	11	203	0.981	0.740	0.917
zebra	223	2188	0	161	1.000	0.581	0.937
overall	975	8748	27	538	0.973	0.644	0.945

(b) night-time

Classes	TP	TN	FP	FN	Precision	Recall	Accuracy
giraffe	27	1370	1	30	0.964	0.474	0.978
impala	2	1422	3	1	0.400	0.667	0.997
oryx	16	1329	6	77	0.727	0.172	0.942
zebra	92	1133	0	203	1.000	0.312	0.858
overall	137	5254	10	311	0.932	0.306	0.944

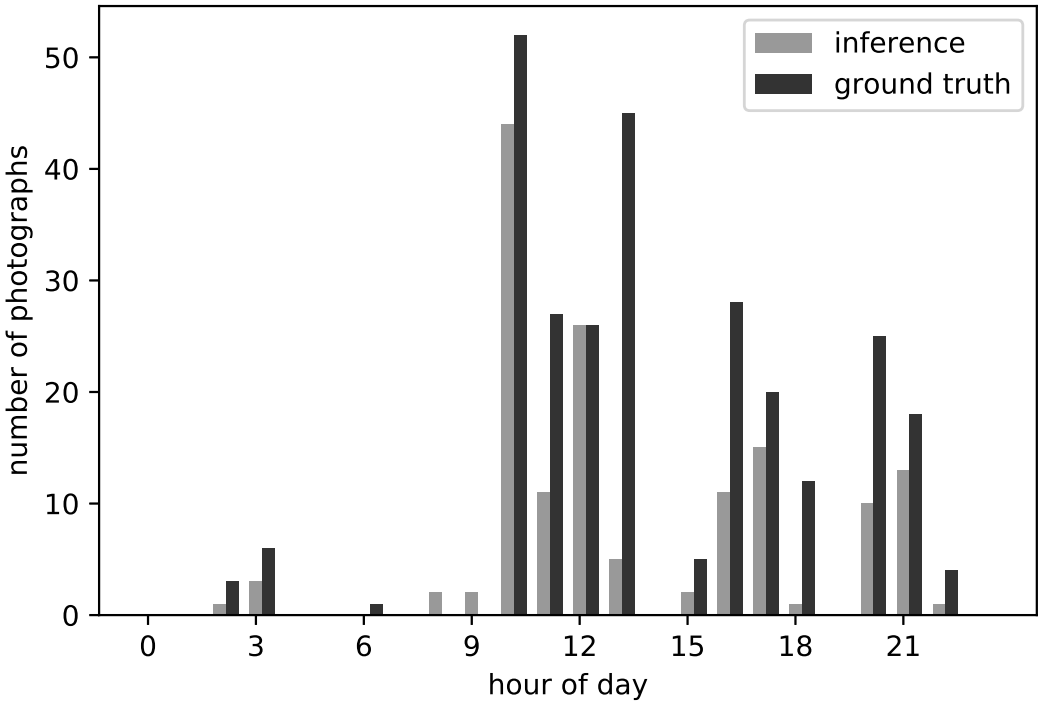


Figure 3.5: Number of giraffe photographs for each hour of the day across all sites, as established by observation (ground truth) and prediction (inference).

3.4 Diel patterns

Figures 3.5, 3.6, 3.7 and 3.8 show the number of photographs by hour of the day predicted (by computer) as well as observed (manually) to contain each of the four large mammal species.

For all four species, prediction (inference) roughly tracks observation (ground truth). For giraffe (observed in 272 photographs, predicted in 147, see Table 3.3), there is a peak around 10:00 which trails down to around 21:00, with few predictions as well as observations outside of this time. For impala (observed in 135 photographs, predicted in 75), the photographs are mostly limited to day-time, with a peak around noon. A similar pattern is evident in oryx (observed in 875 photographs, predicted in 612), whereas for zebra (observed in 679 photographs, predicted in 315) several frequency clusters are spread around the clock.

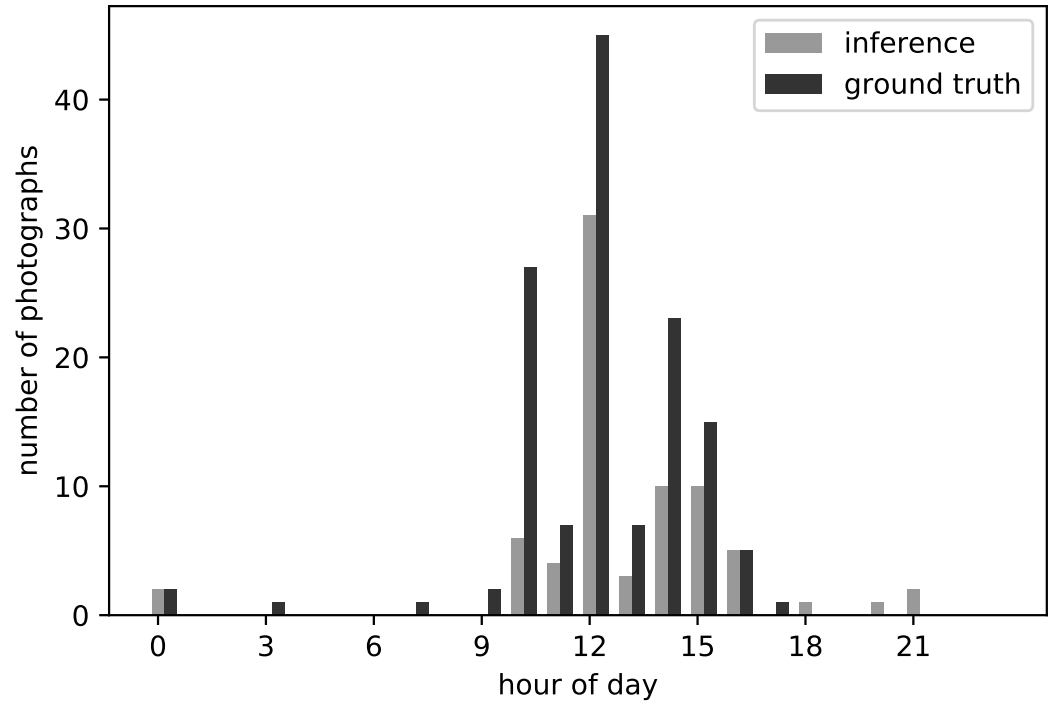


Figure 3.6: Number of impala photographs for each hour of the day across all sites, as established by observation (ground truth) and prediction (inference).

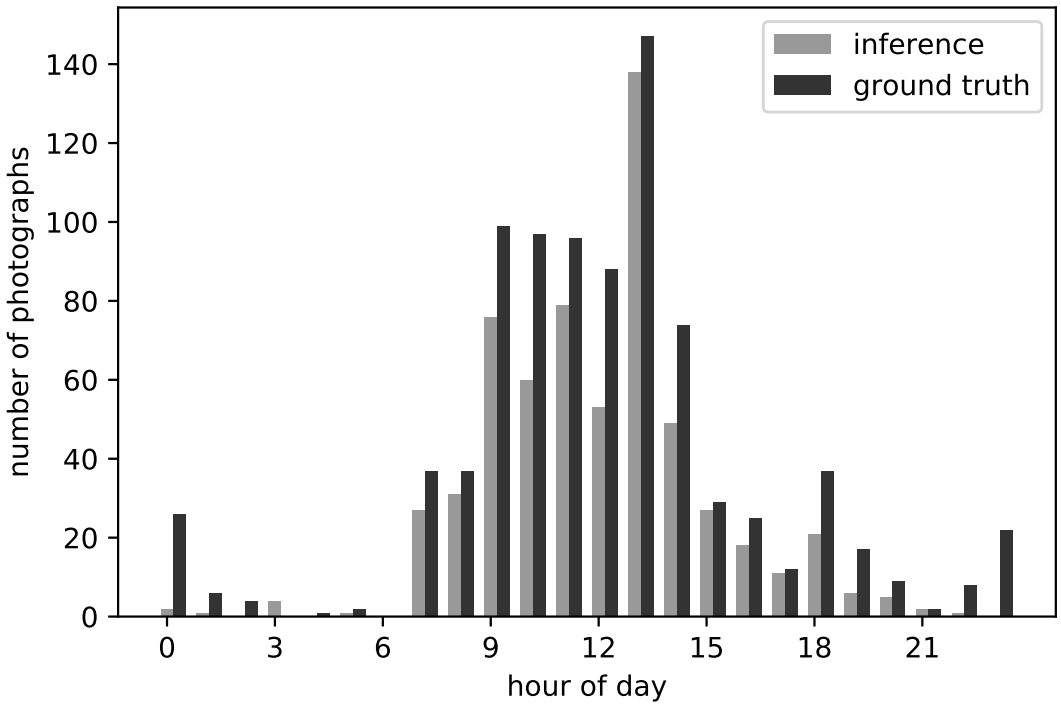


Figure 3.7: Number of oryx photographs for each hour of the day across all sites, as established by observation (ground truth) and prediction (inference).

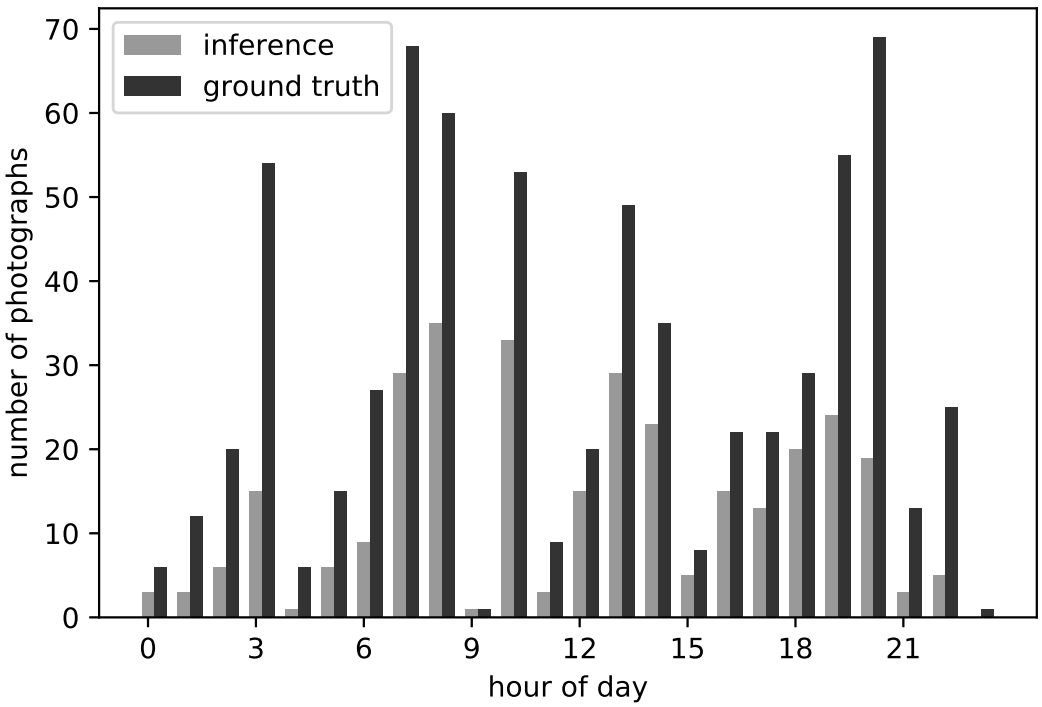


Figure 3.8: Number of zebra photographs for each hour of the day across all sites, as established by observation (ground truth) and prediction (inference).

Chapter 4

Discussion

Camera trapping could play an important role in monitoring wildlife in the Iona Skeleton Coast Trans-Frontier Conservation Area. However, the large numbers of photographs produced by camera traps (Swanson et al., 2015) present a major problem in terms of time and cost spent on their analysis (Harris et al., 2010). This calls for automation. Deep learning-based approaches typically require significant amounts of computer processing power as well as manual preparatory work (Chetlur et al., 2014). These requirements arguably place the automated visual analysis of photos beyond the reach of many camera trap projects, including the one for which this study was done. The lack of a suitable method for automated image analysis led to the development of a new method in this study. The method detects humans, vehicles, and animals and distinguishes between four large mammal species. It is computationally lightweight and requires relatively few training data. It makes use of modular components, which are easily repurposed for the task at hand. Moreover, the results contrast with the findings of Beery, van Horn and Perona (2018) that trained classifiers do not generalise well to new locations, as in this study the system was trained on images from one source (online search engine results) and successfully applied to images from another (photos from a camera trap project).

4.1 Performance

Different performance metrics quantify different attributes of machine learning algorithms and the evaluation of algorithms on the basis of different measures is a matter of continued debate (Sokolova, Japkowicz and Szpakowicz, 2006).

Accuracy is perhaps the most intuitive and hence widely used metric (Sokolova, Japkowicz and Szpakowicz, 2006). It is simply the fraction of correct predictions (Goodfellow, Bengio and

Courville, 2016). The accuracy metric is skewed when there is an imbalance between positives and negatives and therefore higher accuracy does not necessarily indicate better performance of an algorithm (Sokolova, Japkowicz and Szpakowicz, 2006). In this study for instance, accuracy is inflated for those object classes for which there are a high proportion of true negatives (humans, vehicles, giraffe, impala, oryx and zebra, see Tables 3.1 and 3.3).

Recall is the fraction of true cases that were detected (Powers, 2011). As recall is penalised by false negatives, it is a suitable measure when the detection of rare events is important (Goodfellow, Bengio and Courville, 2016). For instance in this study, few photographs contained humans and vehicles. Given that the presence of humans or vehicles can be important information in terms of recording anthropogenic disturbance, missing them could be a costly mistake. Recall is therefore thought to be a good indicator of performance for the human and vehicle classes in this study.

Precision is the fraction of detections that are true cases (Powers, 2011). Precision, which is penalised by false positives, is a useful metric when false positives are undesirable (Goodfellow, Bengio and Courville, 2016). In this study, there are a large number of photographs containing animals. False positives in this case would be undesirable because the large number of positives would need to be checked manually to see if they truly contain animals or not. Precision is therefore thought to be a good indicator of performance for the animal class in this study.

4.1.1 Humans and vehicles

Given the small number of photographs containing humans and vehicles, and the fact that vehicles can be seen as an indicator of human presence, these two categories are combined below.

The high recall rate, reflecting proportionally few false negatives, indicates that human-related objects were detected in the large majority of photographs in which they actually occurred, suggesting that the method is very sensitive to detecting human presence. The high proportion of false positives for these two categories, however, means that photographs labelled as containing these categories would need to be checked manually. Manual validation should be reasonably unproblematic as long as the absolute number of photographs to be checked remains relatively small. In this study, 76 out of 4000 photographs (less than 2%) were falsely labelled as containing humans and vehicles. To reduce the number of photographs to be inspected manually, those predicted as containing humans and vehicles and taken during known times of camera servicing could be automatically excluded from the set of photographs which need to be checked manually.

An approach such as this would be congruent with the suggestions of Harris et al. (2010) to streamline and standardise the processing of camera trap data.

The results for the human and vehicles classes need to be considered with caution however, as sample sizes for both classes were small; humans were present in only 11 photographs and vehicles in 24. Also, humans and vehicles captured in the photographs were related to the servicing of camera traps. Some of the photographs thus contained persons close to the cameras, possibly influencing detection and classification rates, though vehicles were generally further away.

4.1.2 Animals

The performance achieved in this study cannot be compared directly to other studies because image sets differed. However, the focus of this study was to develop and test an automated method for a specific use case rather than benchmark the performance of the method against others using a standard dataset. Performance can be expected to be influenced by a number of factors, as discussed in section 4.1.3. The study by Schneider, Taylor and Kremer (2018), discussed below, exemplifies how performance can differ between image sets. With this in mind, performance of the method is related to four recent studies (section 1.3).

Accuracy in detecting images that contain animals is lower in this study (66.3%) than in other studies. While this may be due to simpler models, fewer training data and different datasets, accuracy in this study also may have been negatively affected by the relatively small proportion of true negatives (17.8%) in the sample. Norouzzadeh et al. (2018) achieved 96.6% on a set of photos half of which contained animals and half of which did not, while Nguyen et al. (2017) also achieved an accuracy of 96.6% based on a slightly imbalanced dataset (31% true negatives).

Although animals were detected in only 59.1% of photographs in which they occurred (recall), the visit of a species to a camera trap site in many cases resulted in several photographs being taken. Though not empirically quantified in this study, the probability that the species is detected in at least one photograph should increase as the number of photographs taken per visit increases. Establishing empirical values for the detection probability per species visit would be desirable and is something to consider in future work.

Notably, precision in detecting animals was 100%. In other words, although animals were not detected in every photograph in which they occurred, if the system claimed there was an animal present, this was correct. Such perfect precision has value in applications where one would want little doubt that positives are true.

In comparison, Yousif et al. (2019) report a recall rate of 73.1% and a precision of 83.6% in distinguish between background, animals / humans by using background subtraction on datasets that contained between 3 and 10 images per trigger event. In the preliminary investigations to this study, background subtraction was briefly considered, albeit in a more naive way than Yousif et al. (2019), and it was found that the backgrounds changed considerably in the Etosha Heights image set. Given that in many cases minutes or hours passed between consecutive photographs in the image set, lighting changed and shadows moved, as did the picture frame (cameras were evidently not fixed to structures sufficiently rigid to prevent camera movement). This method was therefore not further investigated.

In terms of distinguishing between species, the accuracy attained in this study (94.5%) was similar to those in other studies. Norouzzadeh et al. (2018) report an accuracy of 94.9% in identifying animal species in the Snapshot Serengeti dataset (Swanson et al., 2015). Importantly though, their accuracy is based 48 species as opposed to only four in this study. Performance can be expected to drop as the number of classes increases. This is seen, for instance, in the study done by Nguyen et al. (2017): in distinguishing between animal species they achieved an accuracy of 90.4% for three species and 84.4% for six species. How performance can vary not just between models, but also between datasets, is illustrated by Schneider, Taylor and Kremer (2018). They obtained accuracies of 93.0% and 76.7% on the Reconyx Camera Trap and Gold Standard Snapshot Serengeti (Swanson et al., 2015) datasets, respectively, using the Faster R-CNN (Ren et al., 2017) object detection model and 73.0% and 40.3%, respectively, using the YOLO (version 2) object detection model (Redmon et al., 2016). To discount the deleterious effect of small objects on performance, Schneider, Taylor and Kremer (2018) removed from the dataset bounding boxes containing less than 750 pixels.

Performance of the recognition system with regard to the four mammal species varied to some degree between species as well as between sites (Table 3.3 and 3.4), illustrating performance over a range of conditions. Several confounding factors could have led to these differences. For one, the number of true occurrences (the sum of the predicted true positives and false negatives) of a species in the sample of photographs varied greatly between impala (135 photographs) and oryx (875 photographs), thus the recall rate (which is based on true positives and false negatives) is probably more representative for oryx than it is for impala. Similarly, the number of positives, on which the precision metric is based, varied between species. Other confounding factors could be the distance between camera and animal, the proportion of photographs taken during daylight, the type of camera, camera setup, and camera settings, and features specific to a site (e.g. area captured, terrain and vegetation density). Given the numerous confounding factors, comparisons between species and sites cannot be conclusive. Firmer conclusions may

be drawn only from larger sample sizes taken from a wider range of sites. Large projects such as Snapshot Serengeti (Swanson et al., 2015) may be more suitable in this regard.

Even if one bears in mind the possible influence of the above-mentioned confounding factors, there are still some results which stand out. For one, the perfect precision of the image classifier in recognising zebra may indicate that their distinct, striped pelage is a visual characteristic which may favour image classification performance. This is in agreement with the work of Miao et al. (2019) which suggests that the visual features that humans use to identify species are in some cases similar to those used by CNNs. Although the highest recall rate was obtained for oryx, there is significant variation between sites as well as between day and night, suggesting that confounding factors play a substantial role. Further, there is some indication that camera traps situated at waterholes may result in better performance (recall 56.3%, precision 97.4%) than camera traps situated along game trails (recall 51.5%, precision 91.2%, Table 3.5), although these differences are relatively small. The better performance at waterholes than along game trails might be due higher animal density, and therefore more animals being captured per photograph at the former, which would result in a higher likelihood of a species being detected. Lastly, the recognition system performed better during day-time than during night-time. The difference between day and night is substantial in terms of recall (64.4% and 30.6%, respectively, see Table 3.6), while precision was also somewhat higher during the day (97.3%) than at night (93.2%).

The diel patterns illustrated in section 3.4 suggest two things: Firstly, the predicted number of photographs per hour and species is a rough index of the observed number of photographs per hour and species. Secondly, the numbers of predicted and observed photographs for each species and hour of the day broadly match published activity levels, as discussed in the following.

Most of the photographs of giraffe were taken during daylight hours, with the frequency decreasing after nightfall and reaching very low levels during the early hours of the morning. Correspondingly, giraffe are mainly diurnal (Skinner and Chimimba, 2005). While they do show some night time activity (Skinner and Chimimba, 2005), giraffes lie down for part of the night; the lying frequency increases (meaning movement decreases) through the night and peaks (movement is lowest) in the early hours of the morning (Estes, 1991). Most records of impala are from day-time and only few from night-time. This corresponds to impala being primarily diurnal (Estes, 1991). While showing some night time activity (Skinner and Chimimba, 2005), they spend most of the night lying down (Estes, 1991). Almost all photographs of oryx were taken between daybreak until just past midnight, with most photographs restricted to the daylight hours. Oryx have been observed to become active shortly after daybreak, exhibiting polyphasic activity until c. 02:00, after which they lie down to rest (Walther, 1978). For zebra, photographs

are spread more evenly throughout the day and night, but do show a number of frequency clusters throughout the daily cycle. Mountain zebra have three diel grazing periods, early, middle and late in the day, with some feeding concluded to occur also at night (Estes, 1991) and drink at any time of day where undisturbed (Skinner and Chimimba, 2005). Plains zebra are mostly active during the day, but also exhibit some night time activity (Estes, 1991).

Sample sizes would need to be increased to see if diel patterns emerge more clearly. Also, there was no differentiating between behavioural modes (drinking, feeding, moving). Photograph frequency matched known activity levels less clearly for zebra than for the other species. Results may have be confounded given that both mountain and plains zebra were lumped together in the analysis.

Rowcliffe et al. (2014) proposed a method by which activity levels for a number of mammal species can be reliably estimated based on the frequency of photographs. For ethological studies on the species considered in this study, it may thus be worthwhile to further explore the merit of using camera traps, and the automated species recognition method developed in this study, in quantifying activity levels. This may be conservation-relevant, given that Caravaggi et al. (2017) point out that a shift in activity patterns could indicate environmental changes.

Also, the aggregation of results by time period could be further explored. What can be done by aggregating results by hour of the day to reveal diel patterns, could be done by day, season, and year, to investigate seasonal or inter-annual trends and correlations with potential drivers such as rainfall and the availability of water and food (Young et al., 2020). Also, the relationship between photographs taken at waterhole sites and the degree of water dependence of different species could be investigated (Hayward and Hayward, 2012). Any patterns that emerge would need to be treated with caution however, given that other confounding factors may influence the number of photographs taken.

4.1.3 Factors influencing performance

Computer vision performance can be expected to vary considerably between—as well as within—studies, given different mammal communities, sites, camera equipment and settings. Some of these factors are further discussed below.

The mammal community of a given study may influence recognition performance as different species have different degrees of visual distinction (Miao et al., 2019) and exhibit different behaviours (Harmsen et al., 2010), both in terms of time spent within the capture area of the camera and in terms of their diel activity patterns. The study focused on only four large mammal

species, each of which is visually very different from the others. The high precision rates found in this study is presumably due to these pronounced visual distinctions. During preliminary investigations (the results of which were not quantified), there was a tendency for the system to more readily mistake one similar looking species for another (e.g. impala for kudu and vice versa). These mistakes seemed to occur more often when animals were photographed in low-light or backlit conditions.

The distance between camera and subject is likely to be another factor influencing recognition performance. The object detection model seemed to detect animals in the foreground at a higher rate than animals in the background, presumably due to object size (Huang et al., 2017).

Image quality can vary between camera makes and models (Rovero et al., 2013) and low image quality has been shown to have a detrimental effect on the performance of neural networks (Grm et al., 2018). In this study, differences in image quality between photographs from the different sites are apparent, presumably attributable to different camera models and settings. This is clearly visible when sub-images generated in the object detection process are magnified, see for example Figure 3.2b. Textures then appear smudged, presumably due to a relatively high degree of "lossy" compression (Wallace, 1992). This loss of visual information can be expected to have a negative effect on the recognition system. Although it is not apparent that the method performed more poorly for the sites with lower quality images, this would need to be investigated more rigorously over a greater number of sites. Interestingly, Snapshot Serengeti photos (Swanson et al., 2015) are of relatively low resolution, but have high image quality (few noticeable image compression artefacts) compared to the photos used in this study. High image quality is probably more important than image resolution, given that the models accept only low-resolution images as input. Low light levels may be another factor affecting image performance. Images from the sample that were taken under low-light conditions were blurred in many cases.

Lastly, training set size is known to significantly influence the performance of neural networks in classifying images (Foody, McCulloch and Yates, 1995). The results (section 3.3.1) show that using only 100 training images per object class, even when these are extended to 1 000 images by data augmentation (training set B), was insufficient for the task at hand. Given that images suitable for training can be in short supply and preparing a training set may be tedious (Schneider, Taylor and Kremer, 2018), it would have been useful if a small training set such as this would have yielded an acceptable level of performance. While this was not the case, using just over 1 000 training images per class (training set A) resulted in acceptable performance, whereas extending the training set to 10 000 images per class by data augmentation (training set C) showed no additional benefit. Nonetheless, more careful consideration of augmentation

techniques, with a possible focus on geometric augmentation (Taylor and Nitschke, 2017) may be worth investigating.

In comparison to the to 5 291 training images (set A) in this study, the training set used by Norouzzadeh et al. (2018) was massive, containing 1.4 million images; the accuracies obtained in classifying 48 species were correspondingly high. Nguyen et al. (2017) reported using 35 629 images used for training and 8 907 for validation, also achieving high accuracies, but for a smaller number of species than Norouzzadeh et al. (2018). Schneider, Taylor and Kremer (2018), on the other hand, used small training datasets sourced from 946 images of 20 species and 4 432 images of 48 species, attributing the relatively poor performance of the YOLO object detector to limited training data.

4.2 Implementation

Neural network implementations are in most cases based on graphics processing units (GPUs), specialised hardware originally designed for graphics applications (Goodfellow, Bengio and Courville, 2016). Due to financial constraints in this study, the necessity of a (costly, high-end) GPU was avoided by choosing lightweight approaches and components (section 2.2). Only an entry-level personal computer was therefore required. All software used was open-source and freely available. The YOLO v3 object detector (Redmon and Farhadi, 2018) was available pre-trained and thus ready to use. The Inception v3 image classifier (Szegedy et al., 2015) was retrained, by transfer learning, with relatively little computational effort, on the four large mammal species of interest. Code examples provided for the deep learning framework (Abadi et al., 2017) and computer vision library (Bradski, 2000) could be used—with some adaptation necessary to suit the task at hand (see Appendix B for URLs of the most important resources used in this study). What was developed in this study can be seen as a proof-of-concept, however, and further work would be required to produce a user-friendly software package that would eliminate the need for any coding by the end user.

4.3 Utility and application

Camera traps have been applied to a variety of tasks, including compiling species inventories (Burton et al., 2015), mapping species distributions, documenting species movement (Ford, Clevenger and Bennett, 2009), estimating population densities—although these are subject to important assumptions and limitations (Karanth, Nichols and Kumar, 2011; O'Brien, 2011),

collecting behavioural data (Caravaggi et al., 2017), monitoring anthropogenic disturbance to species (Carter et al., 2012), and detecting rare and elusive species, which are in many cases nocturnal or crepuscular (Burton et al., 2015). Further, Meek et al. (2014) mention the importance of exploratory studies as a precursor to focused ones in which specific research or management questions are addressed.

The computer vision method proposed here has the potential to assist to some degree in most of these tasks, by filtering out photographs that contain animals, detecting the presence of humans and vehicles, classifying the predominant large mammals by species, or a combination of these. The suitability of the method depends on the objectives of the particular study in question and on its permissible tolerances of false predictions.

For many applications, it could serve at least as a preliminary classification of photographs, which then could be verified and corrected manually. Such manual correction and validation would arguably be a far lesser task than classifying from scratch.

False negatives can be compensated for to a large degree simply by analysing more photos, which might mean setting cameras to record multiple frames per trigger event or having the cameras record over longer periods of time. It is worth emphasising that false negatives merely lower an already imperfect detection rate. Camera detection rates are always less than 1 (Kéry, 2011), and can vary greatly between camera models (Hughson, Darby and Dungan, 2010). The larger the number of photographs containing a certain species, the higher the probability that the species will be detected in at least one of them.

Some inference errors could be revealed by screening photograph *labels* for occurrences which are irregular or unexpected for a specific locality, time of day or season. Also, the following technique could be used to find animals classified falsely by species: Because the system groups sub-images of the different object classes by folder, e.g. sub-images inferred to contain oryx are grouped into one folder, sub-images inferred to contain zebra in another, and so on, errors can easily be picked up by eye when the sub-images are displayed as an array of small preview images ("thumbnails"). This works because the objects of interest, having been localised by the object detector, fill the frames of the sub-images (see the sub-images displayed in Figures 3.1 through 3.4). Thus the objects are still clearly recognisable when the sub-images are displayed as thumbnails. Objects that were not detected by the object detector will of course not be included in the sub-images, making this technique unsuitable for finding non-detections.

The detection of rare species remains a challenge, however. Until improvements at the computer vision level are achieved, modifications to sampling must be undertaken at the camera trap level.

To increase the chance of capturing rare animals, camera traps could be placed at attractants, e.g. bait lures in the case of carnivores, and sampling effort could be increased. If the species in question is nocturnal, the use of incandescent flash could be considered, though it has been shown that doing so can alter animal behaviour (Wegge, Pokheral and Jnawali, 2004).

Potential applications of the method are further discussed, with special reference to the Iona Skeleton Coast TFCA, in the following chapter, which gives guidelines and recommendations for the use of camera trapping and computer vision for the Iona Skeleton Coast TFCA.

4.4 Strengths and weaknesses

The computer vision method is sufficiently light on computation resources to run on entry-level hardware. In contrast, Norouzzadeh et al. (2018) used an ensemble of nine image classification architectures, some of which are computationally considerably more expensive than the Inception v3 model used in this study (Bianco et al., 2018). Similarly, the computational complexity of the VGG-16 model (Nguyen et al., 2017) used was found to be higher than of the Inception v3 model used in this study, while the accuracies of VGG-16, AlexNet, and ResNet50 models were all found to be lower than that of Inception v3 (Bianco et al., 2018).

This computational economy means processing speed is relatively high even on a CPU, with the once-off process of retraining the image classifier having taken approximately 1 hour (section 2.2.6), and inference per photograph by the object detection and image classification pipeline having taken about 4 seconds per photograph (section 2.2.2). By extrapolation, an entry-level computer could process around 900 photos per hour. In comparison, the Snapshot Serengeti project (Swanson et al., 2015), using a minimum of 200 cameras, collected 3.2 million images over three years. This works out to approximately 120 photographs taken per hour, suggesting that the computer vision method could be suitable even for large-scale projects.

Relatively little manual work is required for the preparation of training data. It was demonstrated that a relatively small set of training data (just over 1 000 images per object class) delivered useful results. Furthermore, images can be sourced online using search engines, and are therefore by and large already correctly classified. In contrast, training object detectors as Schneider, Taylor and Kremer (2018) did, requires the laborious annotation of training images with bounding boxes.

Using existing building blocks, the computer vision method is relatively easy to implement. This, together with the above-mentioned lightweight nature, both in terms of processing speed

and training data required, potentially places it within reach of camera trap projects that would neither have the resources nor the expertise to otherwise use computer vision for analysing images.

Furthermore, it has been shown that the method has high sensitivity in detecting human-related objects, perfect precision in detecting animals and high precision in distinguishing between four large mammal species. However, some performance aspects also count towards the weaknesses of the system. The object detector in many cases fails to detect animals that are obvious to the eye; examples of this can be found in Figures 3.2a and 3.3a. This makes the method unsuitable for applications in which individual animals in a photograph needs to be detected and classified, for instance for counting purposes as explored by Norouzzadeh et al. (2018). The method also has difficulty in properly classifying similar looking species, such as springbok and impala. A greater number of training images, as well as training images sourced from the project to which it is to be applied, might improve this limitation.

In terms of classifying the animals that have been detected, the method offers a closed solution to an open problem: the image classifier is able assign each input image only to one of the classes it has been trained on. So in this study, all animal detections are fed to the image classifier and labelled as either a giraffe, impala, oryx, or zebra. This is perhaps not detrimental to the recognition rate of those four species as long as other species are rare. However, those other species will not be identified as such, which is why the method in its current form is not suitable for finding rare species. Another reason that the method is not suitable for monitoring rare species is that they are in many cases nocturnal. Low-light conditions lead to low shutter speeds which results in blurred images of moving objects. Also, night-time images are grey-scale, so that colour information is not available for the computer vision models to process. This limitation provides scope for future work, however (section 4.6).

4.5 Evaluation of the study

Given that no camera trap photographs were available from the TFCA, a substitute sample had to be sourced from elsewhere (Etosha Heights). However, performance results are known to differ between datasets, see for instance Schneider, Taylor and Kremer (2018), so that the performance results obtained may not be sufficiently representative of those that would be obtained from photos from the TFCA.

The choice of species on which the study focused is based on those known to predominate in the TFCA, namely, springbok (proxied by impala), oryx and zebra (Hauptfleisch and Brown, 2017)

and giraffe for which there was a conservation focus in the area (https://giraffeconservation.org/). Ostrich may have been an additional terrestrial vertebrate to consider, but was not sampled. The results of exploratory studies in the TFCA may suggest a focus on a different set of species, especially given the diversity previously documented by Huntley (1974), see section 1.1.

An iterative analysis of performance using subsamples to obtain confidence intervals was not done in this study. However, given that the performance was analysed on a substitute dataset (from Etosha Heights), such an iterative analysis may have been of limited value in indicating the performance to be expected from TFCA camera trap photos. Further, the analysis was based on a limited sample size (4 000 photos) obtained from only four camera trap sites. A larger sample size, taken from more sites, may have yielded more representative performance results.

Although care was taken to be accurate in determining the ground truth of species presence in photographs, ground truth was not independently verified in this study.

Lastly, details available on camera traps (specifications, setup and settings) with which the photographs used in this study were captured was limited, although the importance of reporting these details has been emphasized, for example by Burton et al. (2015).

4.6 Suggestions for future work

Performance of machine learning algorithms is best when the training and test data are identically distributed (Goodfellow, Bengio and Courville, 2016). Such identical distribution is not given when data sources are heterogeneous (Swaminathan et al., 2017). Training the image classifier on photos obtained from the camera trap project on which it is to be applied would therefore be desirable. There are two caveats to this, however. Firstly, often camera trap image sets will have very imbalanced classes, so that it might be difficult to find suitable training images for rarely captured species (Nguyen et al., 2017; Norouzzadeh et al., 2018). Secondly, the quasi static background of camera trap photos could form a potential confounding factor. Given that correlations between some camera trap sites and some species are to be expected, the model could learn to associate a certain background with certain species (Miao et al., 2019), instead of abstracting the visual features of the species themselves. There was some indication of this occurring in the preliminary investigations into this study when the image classification model had been trained on complete camera trap photos (as opposed to sub-images isolated by the object detection model).

In an effort to improve recall rates of animals, the relatively high resolution camera trap photographs (5 to 11 megapixel, Table 2.1) could be divided into sub-images and each sub-image input to the object detector. Also, the input resolution of the object detector could be increased for better performance. Both measures would result in longer processing times, however, and may at some point require the use of a GPU computer.

A more rigorous investigation into background subtraction may be worthwhile (Yousif et al., 2019). Although backgrounds are generally dynamic in the sample analysed, in some cases, site visits of animals did lead to sequences of images being taken in intervals of a few seconds or minutes, resulting in little background change. In such cases, background subtraction may provide an additional way of locating mobile objects in an image.

The sample size for human-related objects in this study was small. With a larger sample size, it might prove useful to train the image classifier on persons and animals to test whether increased precision can be achieved.

In addition to outputting a label for an image, the image classifier also outputs an estimated probability of the label being correct. This probability value was not taken into account in this study, as it did not always appear to be a reliable indicator. However in future work, the merit of flagging classifications that fall below a certain probability threshold could be further investigated. These flagged classifications could then be inspected manually.

The compilation of standardised, publicly accessible training sets for a range of mammal species could be considered, given the portability of the image classifier that was demonstrated in this study (it was trained on images from one source and applied to images of another). Researchers could then select ready-made training sets on which to train the image classifier for identifying species relevant to their projects. Small-sized training images as used in this study (see section 2.2.6) would facilitate the portability of training sets.

Lastly, a software implementation suitable for end-users, as done for instance by Yousif et al. (2019) would enhance the usability of the method.

4.7 Conclusion

The method presented in this study could lower the barrier of entry of using computer vision in camera trap projects because no special computer hardware is needed, the system requires limited training data and is easily repurposed for new species and sites.

In summary, the method performed as follows: 1) Animals were detected in 59.1% of photographs in which they occurred (recall) and occurred in 100% of photographs in which they were detected (precision). 2) Precision in distinguishing between four large mammal species was 96.8%. 3) Humans and vehicles together were detected in 85.7% of photographs in which they occurred (recall). With this level of performance, the method could have practical utility for a variety of applications involving the detection of human and animal presence as well as the classification of large mammals by species. It may also offer an efficient way of pre-labelling photographs. The labels could then be verified and corrected manually, should better performance be desired.

These attributes could make the computer vision method developed and tested in this study applicable not only to the Iona Skeleton Coast Trans-Frontier Conservation Area, but subject to being trained on the relevant species of the area to which it is to be applied, to other conservation areas and projects.

Chapter 5

Guidelines for the TFCA

This chapter provides a set of guidelines and recommendations for the Iona Skeleton Coast Trans-Frontier Conservation Area aimed at increasing the benefit of using camera traps in combination with the computer vision method described in this work. These guidelines and recommendations are based on the literature reviewed, observations made and insight gained during this study. Aspects covered include potential study areas, study design and methods, types of studies, camera trap equipment, camera placement, camera settings, service intervals, data management, training data and computer vision technology. The scope of this work precludes a complete and in-depth coverage of the topic. Nonetheless this chapter is thought to be used as a basis on which to build a camera-trapping framework for the Iona Skeleton Coast TFCA, as well as other areas to which it may be suited.

5.1 Potential study areas

Situated in the very sparsely inhabited regions of south-western Angola and north-western Namibia, the Iona Skeleton Coast TFCA is both vast and remote. Areas in which to do camera trapping therefore need to be narrowed down. Large parts of the TFCA are extremely arid, offering limited water and food sources for wildlife, resulting in low population densities. Camera trapping, at least initially, should thus rather concentrate on areas where higher mammal densities are to be expected.

An area of primary interest would thus be along the Kunene river. As a perennial river, it provides a permanent source of water, supporting dense stands of vegetation in otherwise arid surroundings. Also, it forms the national border of two countries with different political histories which have affected conservation measures. Higher game densities are expected in Namibia than

in Angola (see section 1.1). Migration over traversable stretches of the Kunene River is therefore thought to occur. The stretch of the river within the TFCA spans nearly 200 km, ranging from c. 10 km upriver (16.98°S, 13.35°E) of the Epupa Falls to the river mouth (17.25°S, 11.75°E). Potential crossing points for large game would need to be identified in collaboration with local communities and tourism operators, and accessible areas would need to be identified at which it would be possible to service camera traps at regular intervals. Collaboration could be sought with tourist lodges along the river, such Serra Cafema (17.21°S,12.20°E), Okahirongo River Camp (17.21°S, 12.42°E,) and the Epupa Falls Lodge (17.00°S, 13.24°E).

Ephemeral rivers such as the Khumib, Hoarusib and Hoanib would be further localities at which to place camera traps. Riverine habitats such as these provide critical resources for ungulate populations of the Namib (Kok and Nel, 1996). A large, vegetated floodplain in the Hoanib river east of the dune belt (19.41°S, 12.92°E) is particularly apparent in aerial photography (Google Maps, 2018). Springs such as those found near the Hoanib river (19.45°S, 12.82°E and 19.40°S, 12.89°E) would be further attractants for wildlife.

In communal areas, camera traps could complement community-based studies; refer to section 1.1 and Stuart-Hill et al. (2005). While site selection should be determined primarily by study design (section 5.2), local knowledge, where available, will be valuable not only in narrowing down areas, but also in choosing suitable camera trap sites (Rovero, Tobler and Sanderson, 2010). Local stakeholders—park staff, local communities and tourism operators—should therefore be included in this process.

Although the results of this study suggest slightly better performance for sites at water sources than at game trails, the computer vision method is considered suitable for being applied to both types of sites.

5.2 Study design and methodology

As mentioned in the Discussion (section 4.3), camera traps can be used to answer a variety of ecological and conservation-relevant questions. However, for camera trap projects to be effective, clear study objectives and careful study design (Meek et al., 2014; Swann and Perkins, 2014) are required. The chosen sampling strategy depends on the ecological questions that need to be answered (Hamel et al., 2013) and the relationship between the sampling method and the underlying ecological processes needs to be defined (Burton et al., 2015).

Depending on the objectives, sampling design can be random or targeted, and the interpretation

of survey results depends on this (Burton et al., 2015). The even spacing of camera traps, covering all habitat types of interest, enables more rigorous statistical analysis, such as occupancy analysis (Rovero, Tobler and Sanderson, 2010). And while randomised camera trap placement may be appropriate for studying entire communities (Swann and Perkins, 2014), trapping success can be maximised by placing cameras along trails or water sources (Rovero, Tobler and Sanderson, 2010).

Furthermore, methodological details relevant to detection rates, such as the choice and number of camera trap sites and their spatial arrangement, survey duration and sampling effort, camera make and model and settings, to name some, should be reported. Meek et al. (2014) provide a more complete set of methodological aspects to take into account.

5.3 Types of studies

Given the few data that have been collected on the mammal community of the TFCA, its recent formal declaration and the fact that management plans still need to be developed, it is recommended to start off with camera trapping studies which are easiest to accomplish in terms of design, have the least number of limitations and require the fewest assumptions. Then, as management objectives develop and research questions arise, progress can be made towards more focused and involved studies. The recommendation is thus to consider exploratory studies as a starting point, which could easily transform to species inventories for a given area, out of which can result species distribution maps and the documentation of migration. As monitoring progresses, the focus could move on to estimating population densities of the predominant large mammal species. This is more readily achievable for species in which individuals are uniquely recognisable, for instance by their coat patterns. Population densities for unmarked species could also be attempted, but estimations must be done with caution, as a number of assumptions need to be met. Other possible studies include the investigation of activity patterns, and the monitoring for human disturbance as well as for the presence of carnivores. The different kinds of studies are discussed in more detail below, as well as to what extent the computer vision method proposed in this thesis could be applied in each case.

5.3.1 Exploratory studies and general monitoring

Camera trapping could be of particular interest for remote areas in which no other regular or constant monitoring methods are feasible. Camera traps can operate autonomously for several

weeks, enabling continuous observation in remote areas which are difficult to access (Trolliet et al., 2014). Even in cases where the use of camera traps does not constitute a scientifically rigorous approach, cameras could serve as eyes on the ground to monitor for unexpected or irregular events. The information gained thereby might be valuable from a management perspective. Also, camera trap photographs could be used to construct time-lapse photography for monitoring vegetation status, disturbance events and environmental changes over certain time periods (Brown et al., 2016).

Exploratory studies, the object of which might be to get a first impression of species occurrences for a particular area, can be an important precursor to more focused camera trap studies (Meek et al., 2014). Given the little information that has been collected on the large mammal community of the TFCA, such a study may be an important preliminary step in the development of a camera trap monitoring programme for the area.

The survey reported on by Hauptfleisch and Brown (2017) recorded the presence of human settlements and domestic animals in Iona. Camera traps could be used to investigate interactions between wildlife and humans, as well as potential shifts in temporal activity patterns (Carter et al., 2012).

The computer vision method could be useful in this regard, given its sensitivity of detecting humans and vehicles and its precision in detecting animals. A trade-off can be expected between the number of species to automatically classify and the precision in doing so (Nguyen et al., 2017).

5.3.2 Species inventories and distributions

An exploratory study could be closely linked to establishing species inventories. The aim of a species inventory is to compile a list of all species within a certain taxon and area; it can be a useful indicator of ecosystem health when compared to a regional species list (Rovero, Tobler and Sanderson, 2010). As is the case with an exploratory study, there is flexibility in the spatial arrangement of camera traps and there is no time limit when collecting data for species inventories (Rovero, Tobler and Sanderson, 2010). Mammal inventories show diversity at a specific site, allow comparisons between sites, provide the basis for species distribution maps and can indicate the human impact on animal activities (Tobler et al., 2008). Camera traps have documented species in areas in which they were thought to be locally extinct, or in areas in which they were not known to occur (Swann and Perkins, 2014).

As species inventories allow flexibility in study design, it would be advantageous to concentrate on places species are most likely to visit, such as water sources, game trails or crossings, while

ensuring that all relevant major habitat types are taken into account (Tobler et al., 2008). Survey effort, often defined as the number of camera traps multiplied by the number of sampling days (Rovero, Tobler and Sanderson, 2010), is the most important factor determining the number of species recorded (Tobler et al., 2008). Although camera trap surveys have been shown to be an efficient and accurate method for inventorying medium and large terrestrial mammals (Tobler et al., 2008), with 57% to 86% of species detected with a survey effort of 1035 to 3400 camera trap days in some studies (Rovero, Tobler and Sanderson, 2010), substantial survey effort may be required to detect some species (Tobler et al., 2008). Large trap effort, however, does not guarantee that all species are detected (Rovero, Tobler and Sanderson, 2010).

The detection of animals in photographs, as well as a preliminary classification of those animals into one of the four predominant mammal species can be done by computer, following the method proposed in this study. This would need to be followed by manual processing to correct for false classifications between the four species as well as for the false classification of any other species as one of the four. Alternatively, manual processing could be limited to only looking into unexpected results and doing spot checks.

5.3.3 Movement studies

Camera traps could be used to document migration of wildlife across the Kunene River, as they have been used to document movement in other studies.

Ford, Clevenger and Bennett (2009) found camera traps effective in monitoring the movement of several carnivore and ungulate species over wildlife crossing structures across Canadian highways. Comparing camera-trapping to spoor recording on track-pads, they found that detection rates for camera traps were higher in some of the ungulate species (elk *Cervus elaphus* and deer *Odocoileus* sp.) and lower in some of the carnivore species (coyotes *Canis latrans* and grizzly bears *Ursus arctos*) studied, though this difference was not categorical for all carnivores and ungulates recorded. Spoor recording could be attempted at crossing points along the Kunene and the data obtained compared to those recorded by camera trap.

Tape and Gustine (2014) demonstrated camera traps to be effective in documenting migration in caribou *Rangifer tarandus*, deploying 14 camera traps along a 100 km transect in the Alaskan Arctic. Counting individuals in photographs, they saw a northward increase in median herd size as spring progressed. In contrast, given the low population densities of ungulates in the TFCA (Hauptfleisch and Brown, 2017), only single animals or small herds can be expected to cross the Kunene. Camera traps would need to be placed on the river bank so as to capture

actual crossing events. While the computer vision method could aid in detecting and identifying species captured, manual assessment would subsequently be needed to determine the direction of animal movement.

Kolowski and Forrester (2017) showed a significant increase in capture rates of cameras placed on game trails compared to random placements. Camera traps have been set on paths and trails in studies done on low density populations, for instance jaguars *Panthera onca* (Tobler et al., 2008) and tigers *Panthera tigris* (Karanth, Nichols and Kumar, 2011). The increased capture rate on game trails in these studies suggest that, where possible, exact crossing points would need to be established along the Kunene.

The performance attained in this study on automatically detecting and classifying wildlife along game paths suggest this to be a viable application of the computer vision method (section 3.5). Training may need to be done on a different set of species, however, depending on which cross the river.

5.3.4 Population estimates

Population density estimates have been done both for populations in which individuals are uniquely identifiable as well as for unmarked populations.

Marked populations

Capture-recapture models are often used to estimate the population density of species in which animals are individually recognisable. Three assumptions must be met, namely that the population is closed to birth, death, immigration and emigration, that there is no loss of markings during the study, and that variation in detection probability are identified and modelled (O'Brien, 2011). Capture-recapture must be limited to a few months to adhere to the assumption of population closure (Rovero, Tobler and Sanderson, 2010). Each individual must have some probability of being captured, meaning the entire area of interest needs to be covered without distances between cameras being so large that an individual is likely not to be detected during the sampling period (Karanth, Nichols and Kumar, 2011). Furthermore, this type of study places demands on the amount of equipment needed, as each camera trap site requires two opposing cameras to photograph both sides of an individual (Karanth, Nichols and Kumar, 2011).

The computer vision method could be used to filter for individually marked species on which

to then apply capture-recapture models to provide populations estimates (Burton et al., 2015). Giraffe and zebra populations potentially could be estimated this way, as photo-identification of individuals has been done elsewhere on giraffe by Halloran, Murdoch and Becker (2015) and on zebra by Lahiri et al. (2011). Zero et al. (2013) even used camera traps to estimate efficiently and precisely population densities of Grevy's zebra *Equus grevyi* using the Random Encounter Model (Rowcliffe et al., 2008), which does not require uniquely marked individuals.

Unmarked populations

Estimating population density for unmarked species is a major challenge for camera trap surveys (Burton et al., 2015). Several methods have been proposed, but are subject to model assumptions. These include the Random Encounter Model (Rowcliffe et al., 2008), the Generalised Random Encounter Model (Lucas et al., 2015), the Random Encounter and Staying Model (Nakashima, Fukasawa and Samejima, 2018), spatially explicit models (Chandler and Royle, 2013), distance sampling for camera traps (Howe et al., 2017), and instantaneous sampling and time-to-event methods (Moeller, 2017),

In many cases indices are used to estimate populations of unmarked species. Any number that is thought to vary directly with population size can serve as an index (O'Brien, 2011). The number of photographs of a species per unit time could in principle provide an index of species abundance, but this assumption may not be valid as other factors could also influence the number of photographs taken, such as behaviour, detection, attraction to or repulsion from the camera trap, and others (Swann and Perkins, 2014). Unsuitable sampling design can thus result in biased estimations of populations (Hamel et al., 2013). Sollmann et al. (2013) illustrate bias in relative abundance indices (RAIs) toward the more detectable species and species with larger home ranges. They also show that variations in trap setup biased RAIs, and that changes in detection rates over time did not vary with actual population trends. Indices of relative abundance should therefore be avoided unless there is no reasonable alternative (O'Brien, 2011). When trap rates serve as an index of abundance, the relationship between these two variables must be calibrated with independently inferred density estimates (Rovero, Tobler and Sanderson, 2010), with calibration needed across time, sites, and species (Burton et al., 2015). Nonetheless, O'Brien (2011) concedes that indices that track population changes over time at one site may be justified in some cases.

5.3.5 Behavioural studies

Camera trap photographs, typically labelled with date and time stamps, can be used to reveal how activity patterns differ between species or how changes in activity relate to the impact of humans as well as other species (Rovero, Tobler and Sanderson, 2010). Caravaggi et al. (2017) advise to test camera trap-based inferences by comparing with and calibrating against more established methods.

Activity has been reliably estimated, subject to the assumption that the entire population is active during part of the activity cycle (Rowcliffe et al., 2014). The timing of resource use between sympatric species can be investigated, see Hayward and Hayward (2012) as well as behavioural responses (e.g. of herbivores) to species (e.g. of predators) that have reintroduced or extended their range (Davies et al., 2016).

As has been shown in this study, there are indications that the frequencies of photographs correspond to some degree to published activity patterns for the four mammal species investigated (section 4.1.2). Increasing sample sizes and differentiating between site-specific behaviours (drinking at waterholes and moving on game paths) may shed further light on this relationship.

5.3.6 Studies on rare, nocturnal and shy species

Camera traps can overcome the problem of recording data on rare species, which are often nocturnal and avoid humans (Swann and Perkins, 2014). Sampling effort may need to be high, with Burton et al. (2015) giving a general recommendation of more than 1000 trap days for rare species.

The computer vision method in its present form has not demonstrated effectivity in detecting rare species. As rare species in many cases are night-active, very few instances of them were recorded in the sample of camera trap photos analysed in this study, and none of them were detected by the object detector. Larger sample sizes would therefore be required, as would an additional catch-all object class for animals not belonging to one of the four mammal species dealt with in this study. The use of incandescent flash, while potentially affecting detection rates of animals by the camera traps (Wegge, Pokheral and Jnawali, 2004), may increase the detection rates of captured animal images by the computer vision method.

5.4 Camera trap equipment

A wide variety of camera traps is available on the market and new models are continuously introduced (Rovero, Tobler and Sanderson, 2010). Whether a certain camera is suitable depends on the research objective of the study (Trolliet et al., 2014). Some of the most pertinent aspects to consider in selecting camera trap equipment are discussed below.

Cost, while important, should not be the only factor to consider when choosing camera traps (Rovero, Tobler and Sanderson, 2010). There are three aspects to take into account in evaluating the cost-effectiveness of camera trapping, namely the purchase price of the camera and batteries, the cost of field trips to replace the batteries and storage media, and the survey duration (Rovero, Tobler and Sanderson, 2010).

Battery life is influenced by the power consumption of the camera, and the movement detection rate; some brands provide solar cells to extend battery life (Trolliet et al., 2014). Limitations in battery life or storage capacity that make more frequent maintenance visits necessary will drive up the operational expenditure of the project (Rovero, Tobler and Sanderson, 2010). Better quality batteries and cameras which extend battery life will be cost saving in the long run, particularly if field trips for camera trap maintenance are expensive (Rovero, Tobler and Sanderson, 2010).

Image quality is affected among other things, by image resolution and sensor size (Trolliet et al., 2014). Higher resolutions do not necessarily result in better image quality as for a given sensor size, a higher resolution means smaller pixel size and therefore less light sensitivity per pixel (Nakamura, 2006). Though not quantified, differences in the quality of photographs between camera traps were apparent also in this study. Furthermore, Snapshot Serengeti photos (Swanson et al., 2015), though of lower resolution (3 megapixel) than the photos sampled for this study (5 to 11 megapixel, Table 2.1), appeared to have higher image quality and less compression artefacts. For computer vision purposes, low image compression should be favoured over high image resolution: Both object detection and image classification models require low resolution images as input (416×416 and 299×299 pixels were chosen in this study, respectively).

Different camera makes and models can vary considerably in their sensitivity to detect and ability to record animals. The size of the detection zone of a camera trap (the area in which movement is detected) is key in determining detection rate (Rowcliffe et al., 2011). The detection zone can be wider or narrower than the field of view, depending on the camera model (Trolliet et al., 2014).

Movement and time-triggered cameras both have advantages and disadvantages. With movement-triggering, an absence of photos does not necessarily mean an absence of animals (Hamel et al., 2013). Camera sensors detect movement of objects that differ in temperature from their surroundings (Rovero, Tobler and Sanderson, 2010), making them less sensitive when ambient temperatures are close to animal body temperatures. Movement-triggering is therefore relatively ineffective in the case of reptiles (Welbourne, 2013) or when ambient temperatures reach the body temperatures of mammals and birds. Differences of less than 3 K between ambient and temperatures may lead to detection failure with infra-red sensors (Meek, Fleming and Ballard, 2012). For a range of endothermic species, this means that when the ambient temperature ranges between 31.5° C and 42.5° C, infra-red triggered camera trapping can be unreliable (Meek, Fleming and Ballard, 2012). In the use of time-triggered cameras, the non-independence of sequential pictures can be controlled for but animals can also be missed if time intervals are too large (Hamel et al., 2013). Notably, Hamel et al. (2013) found in their study that time-triggered cameras recorded more daily presences for a number of species than movement-triggered cameras. This study was done on endotherms in the Norwegian winter, conditions under which one would expect animals to emit a heat signal sufficient for detection by infra-red sensors (Soininen et al., 2015).

Light and sound from camera traps can disturb animals. Several species have been shown to exhibit behaviours (both attraction and repulsion) indicating they noticed camera traps (Meek et al., 2016). Two kinds of flash-light are commonly used in night-time photography with camera traps: infra-red and incandescent. Incandescent flash enables taking clearer, colour photographs at night, which is important for the identification of individuals (Trolliet et al., 2014). This may be particularly relevant to nocturnal marked species such as leopard, but can scare animals away (Wegge, Pokheral and Jnawali, 2004). Infra-red flash is suitable when animal disturbance is to be minimised, but delivers greyscale images of lower quality, which negatively affects the performance of deep learning computer vision models (Grm et al., 2018).

5.5 Camera placement

Cameras should be placed in a way as not to be triggered by moving vegetation such as branches or leaves blowing in the wind (Rovero, Tobler and Sanderson, 2010). Camera height should be adjusted to the species surveyed and reported in the methodology (Burton et al., 2015). Fixation of cameras should be fast and secure to prevent movement of the photograph frame. This would aid in making photographs suitable to background subtraction, which can be an efficient way to detect objects that move between consecutive photographs being taken.

For applying computer vision to photographs, cameras should be oriented so that they preferably do not point into the sun, as backlit photographs may lead to lower recall and precision (see section in Discussion). The seasonal variation of the path of the sun should be taken into account for this, too.

5.6 Camera settings

Camera settings are a variable to be controlled for when attempting comparisons within or between sites (but see section 5.3.4).

One of the main weaknesses of the computer vision method is the only moderate sensitivity of the object detector to animals. The number of photographs taken per species visit should therefore be increased where possible. Rovero, Tobler and Sanderson (2010) generally recommend to set the camera to high sensitivity for use in hot climates. They report typical delay times of 1 to 15 minutes between photographs. However for the computer vision method, the recommendation here—both for movement-triggered and time-triggered cameras—is a delay time of at most one minute, given that several photographs taken in quick succession may well increase the recall rate per visit. Shorter delay times might be considered when a computer vision approach using background subtraction (Yousif et al., 2019) is to be applied. Of course, the choice of delay time also depends on storage capacity and maintenance intervals. Settings pertaining to sensitivity and delay between consecutive triggers should be reported (Burton et al., 2015).

Cameras should be configured so that image quality is maximised (for instance by lowering image compression as well as image resolution, where possible, to an appropriate extent). The resolution of the images input to both the object detector (416×416 pixels) and the classifier (299×299 pixels) are relatively low, limiting the benefit of high-resolution photographs. However, the input resolution of the object detector could be increased to any multiple of 32, but doing so would increase computation time.

5.7 Service intervals

Limited capacities of both batteries as well as storage media can result in prolonged periods of non-recording, as could technical failure or external disturbances. The risk of data loss should be taken into account when determining service intervals (Caravaggi et al., 2017).

5.8 Data management

The importance of data backup to guard against loss or failure of storage media or accidental deletion of data needs to be emphasised. From experience however, it is advised that duplication of photographs beyond the backing up data should be avoided as this can lead to confusion, as well as duplication of work, during analysis. Duplication occurs, for instance, when photographs containing species of particular interest are copied into a separate folder. An alternative to this practice would be tagging photos with photo management software. Automating the tagging process on the basis of labels contained in text (CSV) files generated by the computer vision system could be explored. A variety of software for managing camera trap data is available, but the costs and benefits of this should be weighed up against using a simple spreadsheet (Rovero, Tobler and Sanderson, 2010).

Based on the experience gained during this study, the following practices are recommended for organising and sorting photos.

Firstly, the number of folder levels should be reduced to a manageable level; two folder levels are recommended within a study. The following folder structure is suggested for organising camera trap photographs:

- First folder level: Unique name for camera trap site
- Second folder level: year and month, in the format `yyyy-mm`

For example, for photos that were taken in July 2018 at a camera trap site called "Fence West 2" the folder structure would be: `Fence_West_2/2018-07/`

The time interval for the second folder level can be adjusted in accordance with the number of photographs contained. Folders containing more than a few thousand photos were found cumbersome to work with.

Secondly, photographs should have unique filenames which include both a camera trap identifier and a timestamp including year, month, day, hour, minute and second. Harris et al. (2010) suggest a similar approach in naming camera trap photographs. It is further recommended that names of camera trap sites should be descriptive, only one language should be used, and spelling should be consistent. A timestamp of the format `yyyy-mm-dd_hh-mm-ss` should be used, so that sorting files by filename will automatically also result in the chronological ordering of photographs per camera trap.

For example, a photograph taken at the Fence West 2 site on 1 July 2018 at 12:34:56 would be named:

`Fence_West_2_2018-07-01_12-34-56.JPG`

5.9 Improving computer vision results

There may be scope for improving computer vision performance, even while staying within the same basic framework of the method proposed in this study.

5.9.1 Expanding utility

If accordingly trained, the computer vision system is able to classify virtually any kind of object class. The more different the object class is to other classes, the higher performance can be expected (refer to section 4.1.3). A trade-off between the number of classes and the performance can be expected (Nguyen et al., 2017). Classes can be adjusted according to the relative abundance of species in the study area. For instance, a catch-all class for all species other than the four considered in this study may be of use. The training set would need to be representative of such a class.

The monitoring of rare and elusive species with the aid of computer vision is conceivable in principle. The image classifier would need to be trained on the set of elusive species known or suspected to occur in the area sampled. The sample size (the number of photographs) would need to be increased by increasing camera sensitivity, increasing sample effort, or both. The use of incandescent flash, while possibly influencing detection rate by the camera (Wegge, Pokheral and Jnawali, 2004), would increase computer-based animal detection in photographs as well as facilitate individual recognition in the case of marked populations (Karanth and Nichols, 1998).

5.9.2 Increasing training data

Accumulating a large number of training images can be challenging (Taylor and Nitschke, 2017). In this study, it was difficult to collect more than a few thousand images per species online, of which around 1 000 to 1 500 were deemed suitable for training. Therefore, images from the camera trap project in question could be included in the training set. This would 1) increase the size of the training set and 2) make it more similar to the set on which inference is to be done. Both measures can be expected to increase computer vision performance.

Care must be taken, however, to prevent unintended correlations in the training data, such as a particular species correlating to a particular site. For instance, Miao et al. (2019) found that when most training images of porcupines also contained palm plants, their presence contributed to CNNs inferring the presence of porcupines. Such correlations should be weaker though, if an object detector is used to locate animals in an image, as done in this study. This makes the detected object the main subject of the image (meaning less noise is introduced by the background) as well as causing greater variation in the background within a site.

5.9.3 Technological advances

It is recommended to keep up-to-date with the most recent developments in computer vision. Increases in performances have been great since using deep learning in 2012 (Krizhevsky, Sutskever and Hinton, 2012). Should this trend continue, then more powerful and efficient detectors and classifiers can be expected to be released in future. The modularity of the method means that the object detector could be replaced with a newer model independently of the image classifier and *vice versa*.

Better models, together with improved computational power of affordable recent computers, can be expected to positively influence performance in future, and so make the method increasingly feasible and applicable to a wider range of problems.

Lastly, the proportion of project funding allocated to hardware for computer vision should be reconsidered. High-performance hardware opens up the possibility of using more powerful models and higher input resolutions, which are likely to result in improved recognition performance. If computer vision proves effective in wildlife monitoring in the TFCA, the capital outlay for hardware (in the order of 1 000 to 2 000 USD) would be amortised quickly.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y. and Zheng, X., 2017. TensorFlow: A system for large-scale machine learning. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*. Savannah, GA: USENIX Association, pp.265–283.

Aggarwal, C.C., 2018. *Neural Networks and Deep Learning: A Textbook*. Cham: Springer International Publishing.

Beery, S., van Horn, G. and Perona, P., 2018. Recognition in Terra Incognita. *arXiv:1807.04975 [cs]*, pp.1–20. 1807.04975.

Beja, P., Vaz Pinto, P., Veríssimo, L., Bersacola, E., Fabiano, E., Palmeirim, J.M., Monadjem, A., Monterroso, P., Svensson, M.S. and Taylor, P.J., 2019. The mammals of Angola. In: B.J. Huntley, V. Russo, F. Lages and N. Ferrand, eds. *Biodiversity of Angola: Science & Conservation: A Modern Synthesis*. Cham: Springer International Publishing, pp.357–443. Available from: https://doi.org/10.1007/978-3-030-03083-4_15.

Bianco, S., Cadene, R., Celona, L. and Napoletano, P., 2018. Benchmark analysis of representative deep neural network architectures. *IEEE Access*, 6, pp.64270–64277. Available from: https://doi.org/10.1109/ACCESS.2018.2877890.

Bradski, G., 2000. The OpenCV Library. http://www.drdobbs.com/open-source/the-opencv-library/184404319 [Accessed 11 April 2019].

Brooks, T.M., Bakarr, M.I., Boucher, T., Fonseca, D., B, G.A., Hilton-Taylor, C., Hoekstra, J.M., Moritz, T., Olivieri, S., Parrish, J., Pressey, R.L., Rodrigues, A.S.L., Sechrest, W., Stattersfield, A., Strahm, W. and Stuart, S.N., 2004. Coverage provided by the global protected-area system: Is it enough? *BioScience*, 54(12), pp.1081–1091. Available from: https://doi.org/10.1641/0006-3568(2004)054[1081:CPBTGP]2.0.CO;2.

Brown, T.B., Hultine, K.R., Steltzer, H., Denny, E.G., Denslow, M.W., Granados, J., Henderson, S., Moore, D., Nagai, S., SanClements, M., Sánchez-Azofeifa, A., Sonnentag, O., Tazik, D. and Richardson, A.D., 2016. Using phenocams to monitor our changing Earth: Toward a global phenocam network. *Frontiers in Ecology and the Environment*, 14(2), pp.84–93. Available from: `https://doi.org/10.1002/fee.1222`.

Bruner, A.G., Gullison, R.E., Rice, R.E. and da Fonseca, G.A.B., 2001. Effectiveness of parks in protecting tropical biodiversity. *Science*, 291(5501), pp.125–128. Available from: `https://doi.org/10.1126/science.291.5501.125`.

Burton, A.C., Neilson, E., Moreira, D., Ladle, A., Steenweg, R., Fisher, J.T., Bayne, E. and Boutin, S., 2015. Wildlife camera trapping: A review and recommendations for linking surveys to ecological processes. *Journal of Applied Ecology*, 52(3), pp.675–685. Available from: `https://doi.org/10.1111/1365-2664.12432`.

Caravaggi, A., Banks, P.B., Burton, A.C., Finlay, C.M.V., Haswell, P.M., Hayward, M.W., Rowcliffe, M.J. and Wood, M.D., 2017. A review of camera trapping for conservation behaviour research. *Remote Sensing in Ecology and Conservation*, 3(3), pp.109–122. Available from: `https://doi.org/10.1002/rse2.48`.

Carter, N.H., Shrestha, B.K., Karki, J.B., Pradhan, N.M.B. and Liu, J., 2012. Coexistence between wildlife and humans at fine spatial scales. *Proceedings of the National Academy of Sciences*, 109(38), pp.15360–15365. Available from: `https://doi.org/10.1073/pnas.1210490109`.

CCF, 2010. After 30-year civil war, cheetah presence in Angola confirmed. https://cheetah.org/press-release/after-30-year-civil-war-cheetah-presence-in-angola-confirmed-press-release/ [Accessed 24 August 2018].

Ceríaco, L.M.P., Stanley, E.L., Kuhn, A.L., Marques, M.P., Vindum, J.V., Blackburn, D.C. and Bauer, A.M., 2016. Herpetological survey of Iona National Park and Namibe Regional Natural Park, with a synoptic list of the amphibians and reptiles of Namibe Province, southwestern Angola. *Proceedings of the Calafornia Academy of Sciences*, 63(2), pp.15–61.

Chandler, R.B. and Royle, J.A., 2013. Spatially explicit models for inference about density in unmarked or partially marked populations. *The Annals of Applied Statistics*, 7(2), pp.936–954. Available from: `https://doi.org/10.1214/12-AOAS610`.

Chetlur, S., Woolley, C., Vandermersch, P., Cohen, J., Tran, J., Catanzaro, B. and Shelhamer, E., 2014. cuDNN: Efficient primitives for deep learning. *arXiv:1410.0759 [cs]*. `1410.0759`.

Cloudsley-Thompson, J.L., 1990. Etosha and the Kaokoveld: Problems of conservation in Namibia. *Environmental Conservation*, 17(4), pp.351–354. Available from: https://doi.org/10.1017/S037689290003280X.

Coetzee, B.W.T., Gaston, K.J. and Chown, S.L., 2014. Local scale comparisons of biodiversity as a rest for global protected area ecological performance: A meta-analysis. *PLoS ONE*, 9(8), p.e105824. Available from: https://doi.org/10.1371/journal.pone.0105824.

Copeland, B., 2019. Artificial intelligence. https://www.britannica.com/technology/artificial-intelligence [Accessed 3 December 2019].

Craven, P., 2009. *Phytogeographic study of the Kaokoveld Centre of Endemism*. PhD Thesis. University of Stellenbosch.

Davies, A.B., Tambling, C.J., Kerley, G.I.H. and Asner, G.P., 2016. Limited spatial response to direct predation risk by African herbivores following predator reintroduction. *Ecology and Evolution*, 6(16), pp.5728–5748. Available from: https://doi.org/10.1002/ece3.2312.

Dean, W.R.J., 2001. Angola. In: L.D.C. Fishpool and M.I. Evans, eds. *Important bird areas in Africa and associated islands: Priority sites for conservation*. Cambridge: BirdLife International.

Deng, J., Dong, W., Socher, R., Li, L.J., Kai Li and Li Fei-Fei, 2009. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, FL: IEEE, pp.248–255. Available from: https://doi.org/10.1109/CVPR.2009.5206848.

Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N.D., Wikramanayake, E., Hahn, N., Palminteri, S., Hedao, P., Noss, R., Hansen, M., Locke, H., Ellis, E.C., Jones, B., Barber, C.V., Hayes, R., Kormos, C., Martin, V., Crist, E., Sechrest, W., Price, L., Baillie, J.E.M., Weeden, D., Suckling, K., Davis, C., Sizer, N., Moore, R., Thau, D., Birch, T., Potapov, P., Turubanova, S., Tyukavina, A., de Souza, N., Pintea, L., Brito, J.C., Llewellyn, O.A., Miller, A.G., Patzelt, A., Ghazanfar, S.A., Timberlake, J., Klöser, H., Shennan-Farpón, Y., Kindt, R., Lillesø, J.P.B., van Breugel, P., Graudal, L., Voge, M., Al-Shammari, K.F. and Saleem, M., 2017. An ecoregion-based approach to protecting half the terrestrial realm. *BioScience*, 67(6), pp.534–545. Available from: https://doi.org/10.1093/biosci/bix014.

Estes, R.D., 1991. *The behavior guide to African mammals: Including hoofed animals, carnivores, primates*. Halfway House, South Africa: Russel Friedman Books.

Foody, G., McCulloch, M.B. and Yates, W.B., 1995. The effect of training set size and composition on artificial neural network classification. *International Journal of Remote Sensing*, 16(9), pp.1707–1723. Available from: https://doi.org/10.1080/01431169508954507.

Ford, A.T., Clevenger, A.P. and Bennett, A., 2009. Comparison of Methods of Monitoring Wildlife Crossing-Structures on Highways. *Journal of Wildlife Management*, 73(7), pp.1213–1222. Available from: https://doi.org/10.2193/2008-387.

Funston, P., Henschel, P., Petracca, L., Maclennan, S., Whitesell, C., Fabiano, E. and Castro, I., 2017. *The distribution and status of lions and other large carnivores in Luengue-Luiana and Mavinga National Parks, Angola*. Kasane, Botswana: KAZA TFCA Secretariat.

Gese, E.M., 2001. Monitoring of terrestrial carnivore populations. In: J.L. Gittleman, S.M. Funk, D.W. MacDonald and R.K. Wayne, eds. *Carnivore Conservation*. Cambridge: Cambridge University Press & The Zoological Society of London, pp.372–396.

Goodfellow, I., Bengio, Y. and Courville, A., 2016. *Deep Learning*. MIT Press.

Gray, C.L., Hill, S.L.L., Newbold, T., Hudson, L.N., Börger, L., Contu, S., Hoskins, A.J., Ferrier, S., Purvis, A. and Scharlemann, J.P.W., 2016. Local biodiversity is higher inside than outside terrestrial protected areas worldwide. *Nature Communications*, 7(12306), pp.1–7. Available from: https://doi.org/10.1038/ncomms12306.

Grm, K., Štruc, V., Artiges, A., Caron, M. and Ekenel, H.K., 2018. Strengths and Weaknesses of Deep Learning Models for Face Recognition Against Image Degradations. *IET Biometrics*, 7(1), pp.81–89. 1710.01494, Available from: https://doi.org/10.1049/iet-bmt.2017.0083.

Halloran, K.M., Murdoch, J.D. and Becker, M.S., 2015. Applying computer-aided photo-identification to messy datasets: A case study of Thornicroft's giraffe (*Giraffa camelopardalis thornicrofti*). *African Journal of Ecology*, 53(2), pp.147–155. Available from: https://doi.org/10.1111/aje.12145.

Hamel, S., Killengreen, S.T., Henden, J.A., Eide, N.E., Roed-Eriksen, L., Ims, R.A. and Yoccoz, N.G., 2013. Towards good practice guidance in using camera-traps in ecology: Influence of sampling design on validity of ecological inferences. *Methods in Ecology and Evolution*, 4(2), pp.105–113. Available from: https://doi.org/10.1111/j.2041-210x.2012.00262.x.

Harmsen, B.J., Foster, R.J., Silver, S., Ostro, L. and Doncaster, C.P., 2010. Differential Use of Trails by Forest Mammals and the Implications for Camera-Trap Studies: A Case Study from Belize: Trail Use by Neotropical Forest Mammals. *Biotropica*, 42(1), pp.126–133. Available from: https://doi.org/10.1111/j.1744-7429.2009.00544.x.

Harris, G., Thompson, R., Childs, J.L. and Sanderson, J.G., 2010. Automatic storage and analysis of camera trap data. *The Bulletin of the Ecological Society of America*, 91(3), pp.352–360. Available from: `https://doi.org/10.1890/0012-9623-91.3.352`.

Hassabis, D., Kumaran, D., Summerfield, C. and Botvinick, M., 2017. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2), pp.245–258. Available from: `https://doi.org/10.1016/j.neuron.2017.06.011`.

Hauptfleisch, M. and Brown, C., 2017. *An aerial photographic wildlife survey of the Iona National Park, Angola November 2016 to February 2017*. Ministry of Environment, Angola.

Hayward, M.W. and Hayward, M.D., 2012. Waterhole use by African fauna. *South African Journal of Wildlife Research*, 42(2), pp.117–127. Available from: `https://doi.org/10.3957/056.042.0209`.

Howe, E.J., Buckland, S.T., Després-Einspenner, M.L. and Kühl, H.S., 2017. Distance sampling with camera traps. *Methods in Ecology and Evolution*, 8(11), pp.1558–1565. Available from: `https://doi.org/10.1111/2041-210X.12790`.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S. and Murphy, K., 2017. Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, pp.3296–3297. Available from: `https://doi.org/10.1109/CVPR.2017.351`.

Hughson, D.L., Darby, N.W. and Dungan, J.D., 2010. Comparison of motion-activated cameras for wildlife investigations. *California Fish and Game*, 96(2), pp.101–109.

Huntley, B.J., 1974. Outlines of wildlife conservation in Angola. *South African Journal of Wildlife Research*, 4(3), pp.157–166.

Huntley, B.J., 2017. *Wildlife at war in Angola: The rise and fall of an African Eden*. South Africa: Protea Book House.

Jenkins, C.N. and Joppa, L., 2009. Expansion of the global terrestrial protected area system. *Biological Conservation*, 142(10), pp.2166–2174. Available from: `https://doi.org/10.1016/j.biocon.2009.04.016`.

Jones, B., 2010. The evolution of Namibia's communal conservancies. In: F. Nelson, ed. *Community rights, conservation and contested land: The politics of natural resource governance in Africa*. London: Earthscan, pp.106–120.

Karanth, K.U. and Nichols, J.D., 1998. Estimation of tiger densities in India using photographic captures and recaptures. *Ecology*, 79(8), pp.2852–2862. Available from: https://doi.org/10.1890/0012-9658(1998)079[2852:EOTDII]2.0.CO;2.

Karanth, K.U., Nichols, J.D. and Kumar, N.S., 2011. Estimating tiger abundance from camera trap data: Field surveys and analytical issues. In: A.F. O'Connell, J.D. Nichols and K.U. Karanth, eds. *Camera Traps in Animal Ecology*. Tokyo: Springer Japan, pp.97–117. Available from: https://doi.org/10.1007/978-4-431-99495-4_7.

Kéry, M., 2011. Species richness and community dynamics: A conceptual framework. In: AF O'Connell, JD Nichols and KU Karanth, eds. *Camera Traps in Animal Ecology*. Tokyo: Springer Publishing, pp.207–232.

Kok, O.B. and Nel, J.a.J., 1996. The Kuiseb river as a linear oasis in the Namib desert. *African Journal of Ecology*, 34(1), pp.39–47. Available from: https://doi.org/10.1111/j.1365-2028.1996.tb00592.x.

Kolberg, H. and Kilian, W., 2003. *Report on an Aerial Survey of Iona National Park, Angola, 6 to 14 June 2003*. Ministry of Environment and Tourism, Namibia.

Kolowski, J.M. and Forrester, T.D., 2017. Camera trap placement and the potential for bias due to trails and other features. *PLOS ONE*, 12(10), p.e0186679. Available from: https://doi.org/10.1371/journal.pone.0186679.

Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), pp.84–90. Available from: https://doi.org/10.1145/3065386.

Kuedikuenda, S. and Xavier, M.N.G., 2009. *Framework report on Angola's biodiversity*. Ministry of Environment, Angola.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T. and Ferrari, V., 2018. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv:1811.00982 [cs]*. 1811.00982.

Lahiri, M., Tantipathananandh, C., Warungu, R., Rubenstein, D.I. and Berger-Wolf, T.Y., 2011. Biometric animal databases from field photographs: Identification of individual zebra in the wild. *Proceedings of the 1st ACM International Conference on Multimedia Retrieval - ICMR '11*. Trento, Italy: ACM Press, pp.1–8. Available from: https://doi.org/10.1145/1991996.1992002.

Lindeque, M. and Lindeque, P.M., 1997. Aerial sample counts of large game in northern Namibia. *Madoqua*, 19(2), pp.75–86.

Lucas, T.C.D., Moorcroft, E.A., Freeman, R., Rowcliffe, J.M. and Jones, K.E., 2015. A generalised random encounter model for estimating animal density with remote sensor data. *Methods in Ecology and Evolution*, 6(5), pp.500–509. Available from: `https://doi.org/10.1111/2041-210X.12346`.

Marker, L.L., Fabiano, E. and Nghikembua, M., 2008. The use of remote camera traps to estimate density of free-ranging cheetahs in north-central Namibia. *Cat News*, 49, pp.22–24.

Meek, P., Ballard, G., Fleming, P. and Falzon, G., 2016. Are we getting the full picture? Animal responses to camera traps and implications for predator studies. *Ecology and Evolution*, 6(10), pp.3216–3225. Available from: `https://doi.org/10.1002/ece3.2111`.

Meek, P., Fleming, P. and Ballard, G., 2012. *An introduction to camera trapping for wildlife surveys in Australia*. Canberra: Invasive Animals Cooperative Research Centre. OCLC: 902749684.

Meek, P.D., Ballard, G., Claridge, A., Kays, R., Moseby, K., O'Brien, T., O'Connell, A., Sanderson, J., Swann, D.E., Tobler, M. and Townsend, S., 2014. Recommended guiding principles for reporting on camera trapping research. *Biodiversity and Conservation*, 23(9), pp.2321–2343. Available from: `https://doi.org/10.1007/s10531-014-0712-8`.

Mendelsohn, J., El Obeid, S. and Roberts, C., 2000. *A profile of north-central Namibia*. Windhoek, Namibia: Gamsberg Macmillan Publishers.

Miao, Z., Gaynor, K.M., Wang, J., Liu, Z., Muellerklein, O., Norouzzadeh, M.S., McInturff, A., Bowie, R.C.K., Nathan, R., Yu, S.X. and Getz, W.M., 2019. Insights and approaches using deep learning to classify wildlife. *Scientific Reports*, 9(1), p.8137. Available from: `https://doi.org/10.1038/s41598-019-44565-w`.

Moeller, A.K., 2017. *New Methods to Estimate Abundance from Unmarked Populations Using Remote Camera Trap Data*. Master's Thesis. University of Montana.

Morais, J., Castanho, R.A., Pinto-Gomes, C. and Santos, P., 2018. Characteristics of Iona National Park's visitors: Planning for ecotourism and sustainable development in Angola. *Cogent Social Sciences*, 4(1), pp.1–15. Available from: `https://doi.org/10.1080/23311886.2018.1490235`.

NACSO, 2017. *Game counts in north-west Namibia*. Windhoek, Namibia: Namibian Association of Community Based Natural Resource Management (CBNRM) Support Organisations.

Naidoo, R., Weaver, L.C., Diggle, R.W., Matongo, G., Stuart-Hill, G. and Thouless, C., 2016. Complementary benefits of tourism and hunting to communal conservancies in Namibia. *Conservation Biology*, 30(3), pp.628–638. Available from: https://doi.org/10.1111/cobi.12643.

Nakamura, J., 2006. *Image sensors and signal processing for digital still cameras*. Boca Raton, FL: CRC Press.

Nakashima, Y., Fukasawa, K. and Samejima, H., 2018. Estimating animal density without individual recognition using information derivable exclusively from camera traps. *Journal of Applied Ecology*, 55(2), pp.735–744. Available from: https://doi.org/10.1111/1365-2664.13059.

Nature Conservation Amendment Act, 1996. Republic of Namibia.

Nayak, S., 2018. Deep learning based object detection using YOLOv3 with OpenCV (Python / C++). https://www.learnopencv.com/deep-learning-based-object-detection-using-yolov3-with-opencv-python-c/ [Accessed 30 January 2019].

Nguyen, H., Maclagan, S.J., Nguyen, T.D., Nguyen, T., Flemons, P., Andrews, K., Ritchie, E.G. and Phung, D., 2017. Animal recognition and identification with deep convolutional neural networks for automated wildlife monitoring. *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. Tokyo, Japan: IEEE, pp.40–49. Available from: https://doi.org/10.1109/DSAA.2017.31.

Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C. and Clune, J., 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25), pp.E5716–E5725. Available from: https://doi.org/10.1073/pnas.1719367115.

O'Brien, T.G., 2011. Abundance, density and relative abundance: A conceptual framework. In: A.F. O'Connell, J.D. Nichols and K.U. Karanth, eds. *Camera traps in animal ecology*. Tokyo, Japan: Springer, pp.71–96. Available from: https://doi.org/10.1007/978-4-431-99495-4_6.

O'Connell, A.F. and Nichols, J.D., eds., 2011. *Camera traps in animal ecology: Methods and analyses*. Tokyo: Springer.

Olson, D.M. and Dinerstein, E., 1998. The Global 200: A representation approach to conserving the Earth's most biologically valuable ecoregions. *Conservation Biology*, 12(3), pp.502–515. Available from: https://doi.org/10.1046/j.1523-1739.1998.012003502.x.

Olson, D.M. and Dinerstein, E., 2002. The Global 200: Priority ecoregions for global conservation. *Annals of the Missouri Botanical Garden*, 89(2), pp.199–224. Available from: https://doi.org/10.2307/3298564.

Powers, D., 2011. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), pp.37–63.

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A., 2016. You Only Look Once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, pp.779–788. Available from: https://doi.org/10.1109/CVPR.2016.91.

Redmon, J. and Farhadi, A., 2018. YOLOv3: An incremental improvement. *arXiv:1804.02767*.

Reilly, B.K. and van Hensbergen, H.J., 2002. Time bias and correction factors for helicopter-based total counts of large ungulates in bushveld. *South African Journal of Wildlife Research*, 32(2), pp.115–119.

Ren, S., He, K., Girshick, R. and Sun, J., 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), pp.1137–1149. Available from: https://doi.org/10.1109/TPAMI.2016.2577031.

Rodrigues, A.S.L., Andelman, S.J., Bakarr, M.I., Boitani, L., Brooks, T.M., Cowling, R.M., Fishpool, L.D.C., da Fonseca, G.A.B., Gaston, K.J., Hoffmann, M., Long, J.S., Marquet, P.A., Pilgrim, J.D., Pressey, R.L., Schipper, J., Sechrest, W., Stuart, S.N., Underhill, L.G., Waller, R.W., Watts, M.E.J. and Yan, X., 2004. Effectiveness of the global protected area network in representing species diversity. *Nature*, 428(6983), pp.640–643. Available from: https://doi.org/10.1038/nature02422.

Rovero, F., Tobler, M. and Sanderson, J., 2010. Camera trapping for inventorying terrestrial vertebrates. In: J. Eymann, J. Degreef, C. Häuser, J.C. Monje, Y. Samyn and D. van den Spiegel, eds. *Manual on field recording techniques and protocols for All Taxa Biodiversity Inventories and Monitoring*. The Belgian National Focal Point to the Global Taxonomy Initiative, vol. 8, pp.100–128.

Rovero, F., Zimmermann, F., Bersi, D. and Meek, P., 2013. "Which camera trap type and how many do I need?" A review of camera features and study designs for a range of wildlife research applications. *Hystrix, the Italian Journal of Mammalogy*, 24(2), pp.148–156.

Rowcliffe, J.M., Carbone, C., Jansen, P.A., Kays, R. and Kranstauber, B., 2011. Quantifying the sensitivity of camera traps: An adapted distance sampling approach. *Methods in Ecology*

and Evolution, 2(5), pp.464–476. Available from: https://doi.org/10.1111/j.2041-210X.2011.00094.x.

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

Rowcliffe, J.M., Kays, R., Kranstauber, B., Carbone, C. and Jansen, P.A., 2014. Quantifying levels of animal activity using camera trap data. *Methods in Ecology and Evolution*, 5(11), pp.1170–1179. Available from: https://doi.org/10.1111/2041-210X.12278.

SADC, 1999. Protocol on Wildlife Conservation and Law Enforcement.

SADC, 2014. Consolidated Text of the Treaty of the Southern African Development Community.

SADC, 2018. Angola and Namibia sign MoA for Iona-Skeleton Transfrontier Park. https://tfcaportal.org/angola-and-namibia-sign-moa-iona-skeleton-transfrontier-park [Accessed 24 August 2018].

SADC, 2018. Iona Skeleton Coast TFCA. https://tfcaportal.org/node/404 [Accessed 24 August 2018].

Saltz, D., Ward, D., Kapofi, I. and Karamata, J., 2004. Population estimation and harvesting potential for game in arid Namibia. *South African Journal of Wildlife Research*, 34(2), pp.153–161.

Schneider, S., Taylor, G.W. and Kremer, S.C., 2018. Deep learning object detection methods for ecological camera trap data. *arXiv:1803.10842 [cs]*. 1803.10842.

Shalev-Shwartz, S. and Ben-David, S., 2014. *Understanding Machine Learning*. New York: Cambridge University Press.

Skinner, J.D. and Chimimba, C.T., 2005. *The mammals of the southern African sub-region*. Cape Town: Cambridge University Press.

Soininen, E.M., Jensvoll, I., Killengreen, S.T. and Ims, R.A., 2015. Under the snow: A new camera trap opens the white box of subnivean ecology. *Remote Sensing in Ecology and Conservation*, 1(1), pp.29–38. Available from: https://doi.org/10.1002/rse2.2.

Sokolova, M., Japkowicz, N. and Szpakowicz, S., 2006. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In: D. Hutchison, T. Kanade, J. Kittler, J.M. Kleinberg, F. Mattern, J.C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M.Y. Vardi, G. Weikum, A. Sattar

and B.h. Kang, eds. *AI 2006: Advances in Artificial Intelligence*. Berlin: Springer, vol. 4304, pp.1015–1021. Available from: https://doi.org/10.1007/11941439_114.

Sollmann, R., Mohamed, A., Samejima, H. and Wilting, A., 2013. Risky business or simple solution – relative abundance indices from camera-trapping. *Biological Conservation*, 159, pp.405–412. Available from: https://doi.org/10.1016/j.biocon.2012.12.025.

Spriggs, A., 2018. Africa: Coastal Namibia and Angola. https://www.worldwildlife.org/ecoregions/at1310 [Accessed 28 September 2018].

Stein, A.B., Fuller, T.K. and Marker, L.L., 2008. Opportunistic use of camera traps to assess habitat-specific mammal and bird diversity in northcentral Namibia. *Biodiversity and Conservation*, 17(14), pp.3579–3587. Available from: https://doi.org/10.1007/s10531-008-9442-0.

Stuart-Hill, G., Diggle, R., Munali, B., Tagg, J. and Ward, D., 2005. The event book system: A community-based natural resource monitoring system from Namibia. *Biodiversity and Conservation*, 14(11), pp.2611–2631. Available from: https://doi.org/10.1007/s10531-005-8391-0.

Swaminathan, M., Yadav, P.K., Piloto, O., Sjöblom, T. and Cheong, I., 2017. A new distance measure for non-identical data with application to image classification. *Pattern Recognition*, 63, pp.384–396. 1610.09766, Available from: https://doi.org/10.1016/j.patcog.2016.10.018.

Swann, D.E., Kawanishi, K. and Palmer, J., 2011. Evaluating types and features of camera traps in ecological studies: A guide for researchers. In: A.F. O'Connell and J.D. Nichols, eds. *Camera traps in animal ecology: Methods and analyses*. Tokyo: Springer, pp.27–43.

Swann, D.E. and Perkins, N., 2014. Camera trapping for animal monitoring and management: A review of applications. In: P. Meek and P. Fleming, eds. *Camera Trapping Wildlife Management and Research*. Collingwood, Australia: CSIRO Publishing, pp.3–12.

Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A. and Packer, C., 2015. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific Data*, 2(150026), pp.1–14. Available from: https://doi.org/10.1038/sdata.2015.26.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., 2015. Rethinking the Inception architecture for computer vision. *arXiv:1512.00567 [cs]*. 1512.00567.

Tape, K.D. and Gustine, D.D., 2014. Capturing migration phenology of terrestrial wildlife using camera traps. *BioScience*, 64(2), pp.117–124. Available from: `https://doi.org/10.1093/biosci/bit018`.

Taylor, L. and Nitschke, G., 2017. Improving deep learning using generic data augmentation. *arXiv:1708.06020 [cs, stat]*. `1708.06020`.

Taylor, P.J., Neef, G., Keith, M., Weier, S., Monadjem, A. and Parker, D.M., 2018. Tapping into technology and the biodiversity informatics revolution: Updated terrestrial mammal list of Angola, with new records from the Okavango Basin. *ZooKeys*, 779, pp.51–88. Available from: `https://doi.org/10.3897/zookeys.779.25964`.

Tobler, M.W., Carrillo-Percastegui, S.E., Leite Pitman, R., Mares, R. and Powell, G., 2008. An evaluation of camera traps for inventorying large- and medium-sized terrestrial rainforest mammals. *Animal Conservation*, 11(3), pp.169–178. Available from: `https://doi.org/10.1111/j.1469-1795.2008.00169.x`.

Trolliet, F., Huynen, M.C., Vermeulen, C. and Hambuckers, A., 2014. Use of camera traps for wildlife studies. A review. *Biotechnology, Agronomy, Society and Environment*, 18(3), pp.446–454.

Udvardy, M.D., 1975. *A classification of the biogeographical provinces of the world*. Morges, Switzerland: International Union for Conservation of Nature and Natural Resources.

United Nations, 1992. Convention on Biological Diversity.

Wallace, G.K., 1992. The JPEG still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1), pp.18–31.

Walther, F.R., 1978. Behavioral observations on oryx antelope (*Oryx beisa*) invading Serengeti National Park, Tanzania. *Journal of Mammalogy*, 59(2), pp.243–260. Available from: `https://doi.org/10.2307/1379910`.

Weaver, C.L. and Petersen, T., 2008. Namibia Communal Area Conservancies. In: R.D. Baldus, G.R. Damm and K.U. Wollscheid, eds. *Best practices in sustainable hunting a guide to best practices from around the world*. Budakeszi, Hungary: International Council for Game and Wildlife Conservation, pp.48–52.

Wegge, P., Pokheral, C.P. and Jnawali, S.R., 2004. Effects of trapping effort and trap shyness on estimates of tiger abundance from camera trap studies. *Animal Conservation*, 7(3), pp.251–256. Available from: `https://doi.org/10.1017/S1367943004001441`.

Welbourne, D., 2013. A method for surveying diurnal terrestrial reptiles with passive infrared automatically triggered cameras. *PloS One*, 6(e18965), pp.247–250.

Yosinski, J., Clune, J., Bengio, Y. and Lipson, H., 2014. How transferable are features in deep neural networks? *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems*. Cambridge, MA: MIT Press, vol. 2, pp.3320–3328.

Young, C., Fritz, H., Smithwick, E.A. and Venter, J.A., 2020. The landscape-scale drivers of herbivore assemblage distribution on the central basalt plains of Kruger National Park. *Journal of Tropical Ecology*, 36(1), pp.13–28. Available from: `https://doi.org/10.1017/S0266467419000312`.

Yousif, H., Yuan, J., Kays, R. and He, Z., 2019. Animal Scanner: Software for classifying humans, animals, and empty frames in camera trap images. *Ecology and Evolution*, 9(4), pp.1578–1589. Available from: `https://doi.org/10.1002/ece3.4747`.

Zero, V.H., Sundaresan, S.R., O'Brien, T.G. and Kinnaird, M.F., 2013. Monitoring an endangered savannah ungulate, Grevy's zebra *Equus grevyi*: Choosing a method for estimating population densities. *Oryx*, 47(3), pp.410–419. Available from: `https://doi.org/10.1017/S0030605312000324`.

Appendix A

Glossary

Artificial intelligence is the manifestation in machines (computers) of cognitive functions commonly associated with the human mind, such as learning and problem solving.

Artificial neural networks are computer models inspired by biological brains. Such networks learn to perform tasks by being exposed to examples rather than by being programmed explicitly with rules. Through the learning process, they automatically generate rules from the examples that they process.

Computer vision is the ability of computers to gain a high-level understanding of the visual content of images or videos; it automates tasks commonly associated with human vision such as detecting objects and distinguishing between different kinds of objects.

Convolutional neural network refers to a class of deep (many-layered) artificial neural networks commonly used to analyse visual information.

Deep learning is a form of machine learning which is based on multi-layered (deep) artificial neural networks. It is used in a variety of fields such as computer vision, audio and speech recognition, and natural language processing.

Ground truth refers to information stemming from direct observation as opposed to information provided by inference; it is observation as opposed to prediction.

Image classification is the process of labelling an image according to its visual content. The image classification process takes a picture of an animal, for example, as input and outputs the label "animal", sometimes also outputting an estimated probability of the label being correct.

Inference is the process in which a machine learning algorithm makes a prediction.

Machine learning is the process of a computer to the acquire knowledge to perform a particular task without being explicitly programmed, knowledge acquisition relies on the detection of patterns instead. It is a subdomain of artificial intelligence and is used in a wide variety of fields, such as computer vision and speech recognition.

Object detection is the computer vision task of locating and identifying instances of semantic objects of a certain class (such as humans, cars or animals) in digital images. Object localisation is often defined by a rectangular box bounding the object.

Python is a popular general-purpose computer programming language known for its easy syntax and large standard library.

Training refers to the process of creating an machine learning algorithm through exposure of a machine learning system to a large set of examples (a training dataset).

Appendix B

Online resources used

YOLOv3 object detector trained on the OpenImages dataset:

`https://pjreddie.com/media/files/yolov3-openimages.weights`

Guide on how to retrain the image classifier:

`https://www.tensorflow.org/hub/tutorials/image_retraining`

Inception v3 feature vector:

`https://tfhub.dev/google/imagenet/inception\_v3/feature\_vector/3`

Python script for retraining the image classifier:

`https://raw.githubusercontent.com/tensorflow/hub/master/examples/image\_retraining/retrain.py`

Python script for using the image classifier (needs to be modified for the task at hand):

`https://raw.githubusercontent.com/tensorflow/tensorflow/master/tensorflow/examples/label_image/label_image.py`