Towards Automated Visual Monitoring of Individual Gorillas in the Wild

Clemens-Alexander Brust[1], Tilo Burghardt[2], Milou Groenenberg[3,4], Christoph Käding[1,5], Hjalmar S. Kühl[6,7], Marie L. Manguette[3,6], Joachim Denzler[1,5,7]

[1]Computer Vision Group, Friedrich Schiller University Jena, Germany
[2]Dept. of Computer Science, University of Bristol, United Kingdom
[3]Mbeli Bai Study, Wildlife Conservation Society-Congo Program, Republic of Congo
[4]Wildlife Conservation Society, Global Conservation Program, USA
[5]Michael Stifel Center Jena, Germany
[6]Dept. of Primatology, Max Planck Institute for Evolutionary Anthropology, Germany
[7]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Germany

Abstract

In this paper we report on the context and evaluation of a system for an automatic interpretation of sightings of individual western lowland gorillas (Gorilla gorilla gorilla) as captured in facial field photography in the wild. This effort aligns with a growing need for effective and integrated monitoring approaches for assessing the status of biodiversity at high spatio-temporal scales. Manual field photography and the utilisation of autonomous camera traps have already transformed the way ecological surveys are conducted. In principle, many environments can now be monitored continuously, and with a higher spatio-temporal resolution than ever before. Yet, the manual effort required to process photographic data to derive relevant information delimits any large scale application of this methodology.

*The described system applies existing computer vision techniques including deep convolutional neural networks to cover the tasks of detection and localisation, as well as individual identification of gorillas in a practically relevant setup. We evaluate the approach on a relatively large and challenging data corpus of 12,765 field images of 147 individual gorillas with image-level labels (*i.e. *missing bounding boxes) photographed at Mbeli Bai at the Nouabal-Ndoki National Park, Republic of Congo. Results indicate a facial detection rate of 90.8% AP and an individual identification accuracy for ranking within the Top 5 set of 80.3%. We conclude that, whilst keeping the human in the loop is critical, this result is practically relevant as it exemplifies model transferability and has the potential to assist manual identification efforts. We argue further that there is significant need towards integrating computer vision deeper into ecological sampling methodologies and field practice to move the discipline forward and open up new research horizons.*

Figure 1. **Automated Facial Identification of a Wild Gorilla.** Visual data acquisition in the field often captures sufficient information to establish encounters with individual gorillas. However, relevant information is locked within the pixel patterns measured, usually requiring expert knowledge and time-consuming efforts for identification. Computer vision can help to extract gorilla identities by performing automated species detection, followed by individual facial identification. We show that standard deep learning models combined with a traditional SVM classifier can be used for this task. To assist encounter processing, predictions can be presented graphically with known population information as shown.

1. Introduction

Current ecological information concerning global change points towards an evolving and severe biodiversity crisis [81]. In order to evaluate the effectiveness of conservation interventions accurate monitoring tools are needed for assessing the status of animal populations, species or entire ecological communities at sufficiently high spatio-temporal resolution. The utilisation and interpretation of field photography and inexpensive autonomous cameras [55, 65] can often provide detailed information about species presence, abundance or population dynamics.

2473-9944/17 $31.00 © 2017 IEEE
DOI 10.1109/ICCVW.2017.333

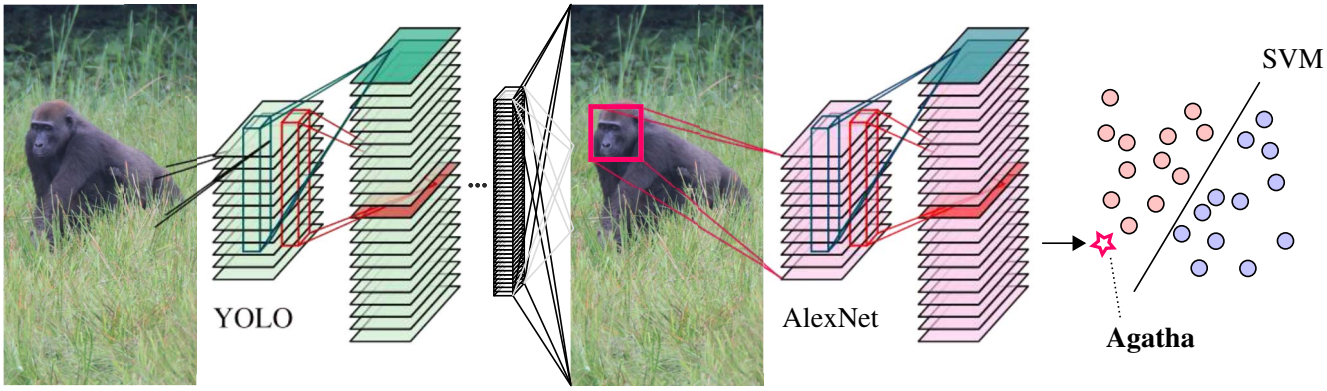

Figure 2. **Overview of Computational Identification Pipeline.** Given field imagery, face detection is performed using a fine-tuned YOLO model [60] resulting in a sequence of candidate regions of interest within each image. Each candidate region is then processed up to the *pool5* layer of the BVLC AlexNet Model [28] for feature extraction. Finally, a linear SVM [11] trained on facial reference images of the gorilla population at hand performs classification of the extracted features to yield a ranked list of individual identification proposals.

In fact, these new methodologies have been transforming the way ecological surveys are conducted [36]. In addition, once images are interpreted, statistical tools [25] applied to visual sighting data can be used to estimate abundance in a study area. However, the manual effort required to conduct such studies currently limits their application [12]. The processing of the number of images or footage collected with even only a few devices quickly exceeds any capacity available. Thus, at least partly automated strategies to assist the image interpretation process are required (see Fig. 2). However, such systems are still not well integrated into daily monitoring practices. As a consequence, keeping biodiversity assessments up-to-date in a near-to-real time manner analogous to the remote sensing of landcover change is currently not possible, although much needed.

The aim of this paper is to briefly discuss the status and limitations of field monitoring particularly within the context of great apes, and to motivate computerized visual processing. Based on that reflection, we describe a facial identification system tested on wild western lowland gorillas. We evaluate the system composed of both deep learning-based and tradition machine learning components (see Fig. 2) and trained towards the task of automatic interpretation of individual gorilla sightings as captured in facial field photography in the wild.

Paper Structure. The remainder of this paper is structured as follows: first, Section 2 will review the current state-of-the-art in ecological field monitoring and its limitations particularly with regard to great ape research. Then, Section 3 will briefly discuss relevant related work from the literature for identification and detection tasks. This will be expanded into a detailed review of the most related prior work on chimpanzee facial identification in Section 4. Section 5 will then introduce the acquisition scenario and data used for the case study on gorillas. Based on this, Section 6 will discuss in detail the computational models used, whereas Section 7 will report on results. Finally, Section 8 will draw conclusions and argue that there is significant further gain to be had in fully integrating computational vision into ecological sampling methodologies, evolving visual species and population models, as well as adjusting actual day-to-day field practice.

2. Monitoring in Ecology Today

Motivation and Task. A key element for any ecological or conservation-related work is precise information about species distribution, density and abundance. For instance, ecologists may be interested in species interactions for which they need to know how the density of one species influences the occurrence of another. Or, wildlife managers, conservation researchers, and biodiversity policy makers want to understand whether the protective interventions they have implemented influence species abundance in a positive way or not [54]. All of this urgently requires effective monitoring techniques that provide accurate empirical data from which informed conservation decisions can be made at appropriate spatial and temporal scales and in a timely manner [54]. Due to chronic limitations in financial and human capacity [27, 49], such methods should ideally be inexpensive, logistically feasible, and easily applicable.

Current Survey Methodologies. Over the last decades a broad spectrum of survey methods has been developed, many of them based on human observers. The most well-known techniques include plot sampling [35], terrestrial or aerial strip transect [35], line and point transect distance sampling [8] or capture-mark-recapture methods [1]. The developments of theoretical foundations, field applications and statistical procedures for data analysis have produced robust estimation methods, which have found very wide application across numerous animal taxa, ecosystems and

regions. More recently genetic survey methods, mainly based on capture-mark-recapture techniques have extended the portfolio of available methods [2, 66].

The advent of digital audio-visual sensors has opened up new ways for species monitoring, in particular regarding the temporal resolution with which biodiversity information can be collected. Analogous to the near-real-time acquisition of satellite based remote sensing data, digital audio-visual sensors allow theoretically the continuous assessment of the status of biodiversity in an area. However, this is currently prevented by the methodological gap between data acquisition and processing, which prohibits both applications across large scales and provisioning of information in near real-time to the user. Successful attempts to address this methodological gap include for instance the inclusion of citizen scientists into data processing, which can speed up the processing of camera trap images and footage considerably [74].

Monitoring of Great Apes. The monitoring of critically endangered African great apes is particularly challenging and complex due to their remote and inaccessible locations (see Fig. 3), their elusive nature, and the spatio-temporal variability of their density [35].

The most commonly applied procedure is the counting of ape sleeping nests along line transects [52, 73, 78]. As it requires highly variable parameters, such as the rate of nest production and decay, when converting ape nest density into individual ape density, it frequently only provides imprecise or even biased estimates [37, 40, 50, 83]. More recently, promising results have been obtained by non-invasive genetic mark-recapture studies, demonstrating exceedingly precise estimates compared to traditional survey methods, [2, 66, 20], unambiguous differentiation between species [2], and no observer or site-specific biases [2, 20]. However, such studies require high levels of expertise and may thus prove infeasible or prohibitively expensive [35, 20].

Great Ape Surveys using Visual Capture. A more widely accessible and long-term economical alternative is the rigorous application of visual data acquisition, particularly remote camera trapping in combination with distance sampling [25] or capture-recapture models [35, 58, 67]. Remote camera trapping can effectively record all apes in a given region [35]. It effectively bypasses common sources of error in traditional survey methods [35, 58, 23] and is not restricted to a singular species nor research question.

It has great potential to provide unique and valuable data on the impacts of conservation threats [71], socio-demographics [23], behavioural plasticity [32, 55, 3], disease mapping and screening [9], species interactions [23], habitat use [53, 24], feeding ecology [24, 51, 59], activity patterns [32, 55], and ranging patterns [48].

However, camera trap methodologies have only been used sporadically in population assessments of great apes due to the difficulty of consistent individual identification and prohibitive costs and man-power requirements of large-scale data processing [35].

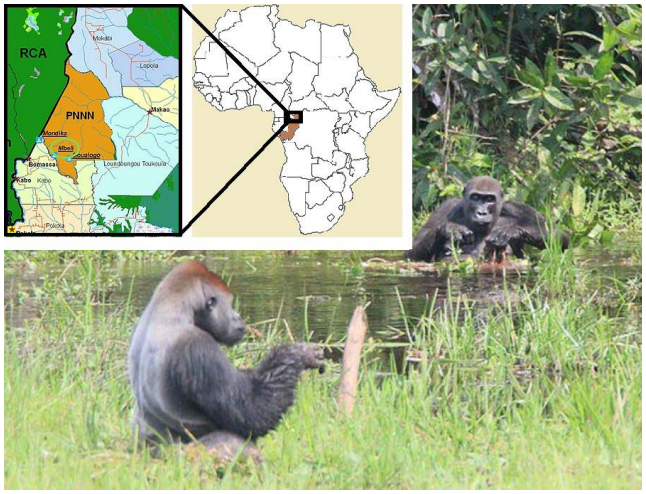

Figure 3. **Study Site: Western Lowland Gorillas in their Natural Habitat.** Mbeli Bai is a large (12.9ha) forest clearing located in the Nouabalé-Ndoki National Park ($2^\circ 15' 5'' \text{N}$, $16^\circ 24' 7'' \text{E}$), in the Republic of Congo. Remote and inaccessible locations often complicate monitoring efforts of habitats and populations. Visual capture over distance or via camera traps addresses this problem to some extent. *(images: Mbeli Bai Study)*

Recent progress in the emerging field of animal biometrics [36], in particular in ape facial recognition [47, 12, 46, 16] promises to overcome some of these obstacles and make broad-scale, real-world applications a realistic prospect.

3. Automated Detection and Identification

Over the past years, computer vision researchers have developed a multitude of algorithms and techniques applicable to the automated interpretation of field imagery in general, and facial gorilla recognition in particular. The following review section briefly introduces some milestone concepts and most recent approaches directly relevant to the task at hand.

Deep Neural Architectures for Object Detection. An early implementation of object detection using deep learning is Region-based CNN (R-CNN) [18], improving previous sliding window-based approaches by a large margin. R-CNN employs an unsupervised method to generate proposals for regions of interest. These proposals are processed by a CNN for feature extraction. Support Vector Machines (SVM), one for each class, are used to score the extracted features[1]. A threshold is applied to the SVM outputs to extract detections from the large number of proposed regions.

[1]Note, that we explicitly build on this hybrid classification strategy later for the design of the individual identification component.

For facial detection we build on the YOLO [60] framework instead, another deep learning-based object detector that is trained using an end-to-end approach. It produces all detections in a single pass (hence the name "You Only Look Once") and requires only minimal post-processing. The output encoding separates detection and classification completely, leading to robust detections even on possibly unknown classes.

There are numerous alternatives to YOLO, including Single-Shot Detection (SSD) [42], a state-of-the-art approach that uses only a single pass per image similar to YOLO, delivering the detections using a more complex output encoding. It can be more accurate than YOLO as a result of a series of improvements, including assumptions about the aspect ratio distribution of bounding boxes as well as predictions on different scales.

A number of improvements for YOLO are released as YOLOv2 [61] including aspect ratio priors for bounding boxes, more fine-grained feature maps using pass-through layers to increase resolution and making the network aware of multiple scales by resizing it during training. This is possible because the network only contains convolutional and pooling layers – similar to how fully convolutional networks [44] exploit this restriction to segment arbitrary image sizes.

In our work we use YOLO because the implementation extends previous work based on YOLO [16].

Fine-grained Recognition. The distinction of fine-grained categories has been studied deeply in the past [4, 19, 15, 86, 64] where applications range from fashion style recognition [30] or cars [33] to more biodiversity-driven scenarios like recognition of flowers [38], birds [82, 80], dogs [29] or moths [63]. One of the most recent and promising developments is the guidance of attention to identify meaningful parts of objects [43, 68] refined by advanced pooling approaches [41, 69].

In contrast to purely fine-grained category recognition, we are interested in the identification of individual animals [16, 26] (or instances) of a single species (or category) rather than highly similar categories. Whilst animal biometrics [36] may operate on a wide variety of entities to achieve identification, our technical focus of the review will be solely on techniques applicable to facial identification.

Facial Identification of Humans. Facial identification of individual humans is a fundamental and traditional application in computer vision. One of the earliest approaches is based on Principal Component Analysis (PCA) and related projections as proposed by Turk and Pentland in 1991 [77]. This 'Eigenface' technique forms a base method and has often been extended and adjusted, e.g. by He *et al*. using more advanced projection techniques [22].

Various other approaches were developed during the last 15 years [84, 85, 70] until recently the usage of end-to-end learning via deep neural architectures provided new, so far unseen performance levels. As one of the initial works, Deepface proposed by Taigman *et al*. [75] uses 4M face images to train and establish an identification model. The VGG-faces architecture by Parkhi *et al*. continued that trend of large sample utilisation employing 2.6M face images for training [56]. While these advanced approaches show promising results in human facial identification, they are not necessarily transferable to great apes, where general models were shown to work more reliably [16].

Facial Identification of Chimpanzees. Loos *et al*. [45, 46] proposed the first pipeline for identification of chimpanzees. Only images showing near-frontal chimpanzee faces serve as suitable input to the technique to guarantee feature comparability. Initially, an affine transformation is applied using facial keypoints. After cropping and scaling the aligned and normalised facial region, the resulting image is described by extended Local Ternary Patterns [76] extracted on spatially divided Gabor magnitude maps. The dimensionality of the obtained features is then reduced by locality preserving projections [21]. Finally, individual classification is performed using a sparse representation [85].

Another system for chimpanzee identification, including attribute estimation for gender and age, was proposed by Freytag *et al*. [16]. Since this system serves as the base template for our gorilla identification approach it is now reviewed in detail.

4. Deep Facial Chimpanzee Identification

In [16], Freytag *et al*. present a system, which significantly outperforms previous results in chimpanzee identification. Instead of using hand-crafted features, they identify individual chimpanzees by utilizing deep-learned image representations. The paper investigates the efficacy of different CNN architectures (in particular BVLC AlexNet [34] and VGGfaces [56]) as well as parameterizations (*i.e*. fine-tuning) and reports on the effects on identification performance. Furthermore, various feature processing steps like bilinear pooling [41], LOGM [79, 10] transformation and normalization are evaluated. The authors show that post-processed *pool5* features obtained from a standard AlexNet classified by a linear SVM outperform a fine-tuned AlexNet as well as an off-the-shelf VGG-faces network.

For their experiments, Freytag *et al*. prepare an extended version of the chimpanzee dataset by Loos and Ernst [46], covering 24 individuals on 2109 images. The images of this C-Zoo dataset are cropped faces. This is opposed to our scenario of images from the wild where faces first have to be detected and localised. They also evaluate their methods on a second dataset of cropped faces called the C-Tai dataset. It

contains 4,377 images of 78 individuals, where the quality difference between images is much larger.

The primary use case for the work of Freytag *et al*. is identification, where they achieve an accuracy of 92% on the C-Zoo dataset. However, the proposed CNN-and-SVM approach is also used to estimate an individual's gender with a very high accuracy of up to 98%. The supplementary material also investigates age group estimation and age regression using Gaussian processes. By identification in combination with database look-ups, the gender is estimated with an accuracy of 97%, close to the 98% achieved when estimating gender directly and actually higher than the 92% identification accuracy.

Freytag *et al*. offer their dataset[2] as well as a working implementation of their identification pipeline[3] on GitHub. The package contains identification, age and gender estimation components, but also a face detection component. This detector is based on the YOLO detection framework of Redmon *et al*. [60]. We use this detector and the general system design as a starting point for our work. While the detector was designed for and works reliably on chimpanzees, it has very low recall for gorilla faces. Since our dataset (see Section 5) does not provide labeled faces, our work depends strongly on a reliable face detection system applicable to in-habitat gorilla imagery. We build this detector by annotating a small subset of our dataset with bounding boxes and then fine-tuning the detector supplied by Freytag *et al*. Before we discuss our system in detail, we describe the image corpus we work on.

5. Acquisition Scenario and Dataset

Study Site. Mbeli Bai is a large (12.9ha) forest clearing located in the Nouabalé-Ndoki National Park (2°15′5″N, 16°24′7″E), in the Republic of Congo (see Fig. 3 for a map). The clearing is comprised of waterbodies and swampy soils that support (semi-)aquatic herbaceous vegetation dominated by species in the Cyperaceae, Hydrocharitaceae, and Gramineae families (see [72, 5, 57] for a full description of the study site).

[2] https://github.com/cvjena/chimpanzee_faces
[3] https://github.com/cvjena/analyzing-chimpanzees

Figure 4. **Examples from the Data Corpus.** 20 representative images from the set of 12,765 images filmed at Mbeli Bai. The photographic database represents a wide variety in lighting conditions, distances to focal object, angles and viewpoints.

Study Population. The visiting population of western lowland gorillas (Gorilla gorilla gorilla) to Mbeli Bai has been consistently followed in the period between February 1995 and March 2017 and detailed demographic records exist for 482 individuals. The photographic dataset comprises a subset of this population and includes a total of 147 individuals, all observed between 2012 and 2017, representing 12 sex-age classes (34 adults silverbacks (>20yo), 50 adults females (>10yo), 18 young silverbacks (14-18yo), 5 black backs (11-14yo), 5 sub-adult females (7.5-10yo), 7 sub-adult males(7.5-11yo), 4 sub-adults of unknown age class (7.5-10yo), 2 juvenile females (4-7.5yo), 4 juvenile males (4-7.5yo), 5 infant females (0-4yo), 10 infant males (0-4yo), 3 infants of unknown sex-class (0-4yo)). A total of 129 individuals constituted of members of twenty different social groups (mean 6.45 individuals per group, $\pm$ 4.6, 1-19, *i.e*. not all individuals of each group were included in the dataset), and 18 individuals are lone silverbacks.

Observation and Ground Truth Identification Methods. Gorillas were observed from a viewing platform at the edge of the bai with the use of telescopes (16-48×60) and binoculars (10×25). Individual identification was based on characteristics including the nose print, the colour of pelage and the size and shape of the brow-ridge, the crest, and the ears [57]. During each visit, a minimum of one photograph was taken of every single individual with an EF600mm with EF2x lens extender. A total of 12,765 pictures were selected and annotated with per-image information of a single individual identity, with a mean rate of 86.8 pictures/individual (see Fig. 4). The photographs cover a wide variety in lighting conditions, distances to focal object, angles and viewpoints. The risk of inter-observer identification differences was reduced to a minimum by ensuring independent identification by at least two experienced observers at data collection stage (on the platform) as well as at the subsequent stage of annotation.

6. Computational Methodology

Overview. Our identification pipeline consists of two sequential components (see Fig. 2): first, a detector based on the YOLO model [60] detects and locates gorilla faces in images. In a second step, each candidate face region is processed up to the *pool5* layer of the BVLC AlexNet Model [28] for feature extraction, before a linear SVM [11] component trained on facial reference images of the gorilla pop-

ulation performing classification of the extracted features to yield a ranked list of identification proposals.

6.1. Gorilla Face Detector

To construct the detector, we import the architecture and parameters supplied by Freytag *et al*. as part of the online supplementary material to [16], originally trained using Darknet[4], into the CN24 deep learning framework [7]. Whilst this detector operates reliably on images of chimpanzees, there are substantial differences in appearance between chimpanzees and gorillas. As a result, transferability of detection efficacy is poor.

We improve on this shortcoming by fine-tuning. A small subset of 2,500 images from the data corpus is annotated with bounding boxes marking the faces and up to 2,000 of them are subsequently used to optimize the model (500 withheld for validation). The YOLO model is trained using the CN24 [7] framework for 3,500 gradient steps of the Adam optimization algorithm [31]. The resulting detector reliably locates gorilla faces (see Figure 5) in images of various lighting conditions, resolutions and aspect ratios. All further processing is done on a per-face basis. Faces are extracted using the detected bounding boxes and resized to uniform dimensions.

6.2. Individual Facial Identification

Following the procedure proposed in [16], we identify individual gorillas by performing feature extraction via a deep architecture before performing classification using a linear SVM [11]. Even with large amounts of data, a linear SVM can be effectively trained.

For feature extraction, we employ the BVLC AlexNet Model [28]. Faces are resized to the appropriate input dimensions (224x224 pixels) and preprocessed, including ImageNet mean image subtraction and channel swapping. By doing this, we avoid retraining a feature extractor and use the reference weights instead. In [16], the authors show that this generic model trained on ImageNet [13] is superior to a fine-tuned model as well as a face-specific network [56].

The extracted features are used directly by the linear SVM. We choose to avoid post-processing steps like bilinear pooling [41] and LOGM transform [79, 10], because the additional computation time does not warrant the insignificant increase in identification accuracy [16].

7. Experimental Results

7.1. Gorilla Face Detection

Performance results for the gorilla face detector are shown in Figure 5 and Table 1. In order to quantify the value of annotating additional bounding boxes, we compare different amounts of annotated training images (from none to 2,000) on a separate validation set of 500 images. To assess overall detection performance, we report average precision (AP) as well as precision and recall figures following the method of [14]. The blue curve in Figure 5 and the leftmost column in Table 1 quantify the poor transferability of detection efficacy from facial detectors trained on chimpanzees applied to gorillas without fine-tuning.

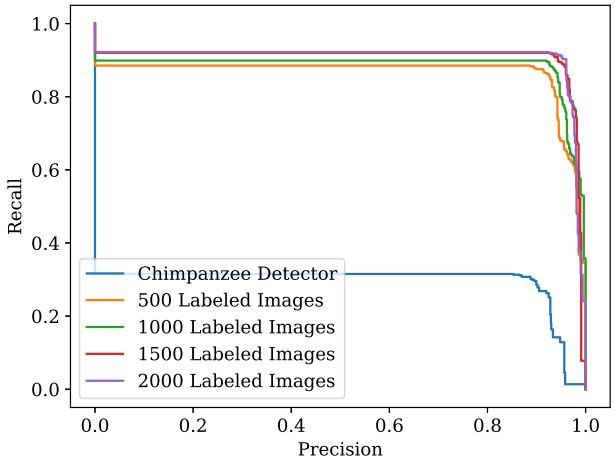

Figure 5. **Gorilla Face Detector Performance.** Precision-Recall plot on a validation set of 500 labeled images after training on between 500 and 2000 annotated samples, respectively. The plot also quantifies (poor) transferability of detection efficacy from facial detectors trained on chimpanzees *(blue)* applied to gorillas.

Additionally, since our dataset assumes at least one gorilla face per image, we report the percentage of images of the whole dataset where there is no detection, as that may indicate a missed face. Note that there are some instances in the dataset where the individual is facing away from the camera. We also report the percentage of images where there is more than one detection. Not all of those are false positives, since in many images there are multiple individual faces present. A typical situation is an adult with an infant on their back. To provide further insight into failure cases, Figures 6 and 7 show randomly selected cases where no or more than one detection was made. A large

Table 1. **Detection Results.** A chimpanzee detector is compared with a fine-tuned model after annotating between 500 and 2000 images with bounding boxes.

# Training Images	-	500	1000	1500	2000
Validation set					
AP (%)	29.5	86.6	88.4	90.6	**90.8**
Precision (%)	83.1	85.6	89.4	88.9	90.1
Recall (%)	31.5	88.7	90.3	92.0	92.2
Whole dataset					
No Detection (%)	59.8	0.7	0.8	0.3	0.4
1 Detection (%)	39.1	92.3	93.7	93.9	**95.4**
$>$ 1 Detection (%)	1.1	7.0	5.4	5.9	4.1

[4] https://github.com/pjreddie/darknet

number of false negatives and false positives are a result of the dataset assuming exactly one individual per image, but actually containing images violating this assumption.

7.2. Individual Gorilla Facial Identification

Results are shown in Table 2 and further details are given in Fig. 8. For evaluating the individual identification performance of the system, the whole dataset is divided into five random folds for cross validation. Bounding boxes are supplied by the detector as proposed in Section 6.1. Since we can only verify the validity of detections in the 500 annotated validation images, errors reported for identification are cumulative with detector failures.

For each image, the only known identity is the individual used in the single supplied annotation. Usually this applies to the animal the image is focused on. If the detector predicts exactly one bounding box, we can assume that it is the appropriate face. However, for more than one prediction, we have to decide which bounding box is the correct one for the given individual. We propose two heuristics: selecting the face with the highest detection score and selecting the bounding box with the largest area. For each heuristic we report again accuracy, precision and recall at operation on the data corpus. In addition, we report top 5 accuracy in Table 2 , that is where an identification is considered correct as long as the correct individual produced a score amongst the top 5 highest ranking SVM scores. Top N rankings are of practical importance particularly in a semi-automated setting where the system presents the best N matches to the human user. Fig. 8 quantifies this aspect further and details accuracy vs. membership in Top N ranked set.

8. Discussion and Future Work

8.1. Transfer of Design and Models

The results presented exemplify that deep learning pipelines constructed for a biometric entity, species and setup (e.g. chimpanzee identification on bounding box labelled face images [16]) open up new possibilities to transfer both system design and parameterisation parts across to similar species and application scenarios (e.g. gorilla facial identification without bounding box information).

Figure 6. **Example images where more than one individual was detected.** Red boxes are cases where subsequent identification failed.

Figure 7. **Example images where no detections were made.** Individuals may be occluded or facing away from the camera, preventing face detection.

In particular, we showed that 1) fine-tuning of a chimpanzee-trained YOLO model using a small, random sample set from the target domain allows to establish a well-performing gorilla face detection model ($AP = 90.8\%$), and that 2) a subsequent facial identification using an approach successful in chimpanzees yields useful outputs when trained on gorillas and tested on a large field data corpus (top 5 accuracy at 80.3% for $12k$+ images of 147 individuals). Immediate next commissioning steps now include cycles of use and testing by field practitioners via the built graphical tools.

8.2. Integration of Active Learning Capabilities

Despite the benefits of system transferability, a key limitation in initialising and maintaining computational animal biometrics systems is the fact that labeling is costly and time consuming (see Section 2), yet traditionally fundamental to injecting new or correcting information into models.

Table 2. **Identification Results.** Using the different fine-tuning steps of the detector as well as the original chimpanzee model, we evaluate identification performance. We compare two heuristics for selecting the individual associated to an identity of an image.

# Training Images	-	500	1000	1500	2000
Highest Score					
Accuracy (%)	20	59	60.3	61	61.7
Precision (%)	48.6	55.5	57.9	58	59.5
Recall (%)	20.3	59.9	61.7	62.3	62.5
Top-5 Accuracy (%)	27.6	77.6	78.9	79.3	79.6
Largest Box					
Accuracy (%)	20	58.8	61.3	62	**62.4**
Precision (%)	48.7	55.5	58.7	59	60.2
Recall (%)	20.3	59.9	62.5	63.2	63.1
Top-5 Accuracy (%)	27.8	77.4	79.5	80.1	80.3

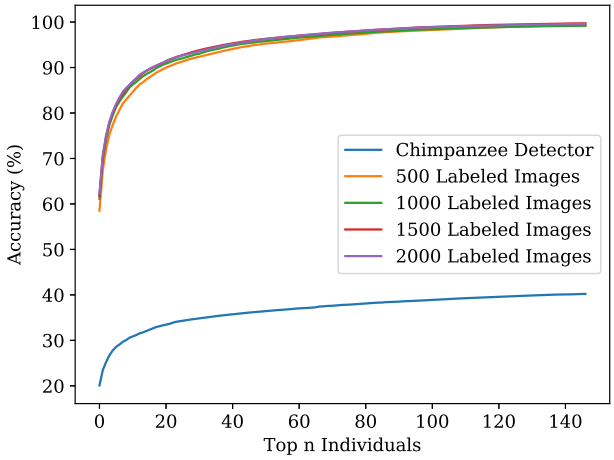

Figure 8. **Quantification of Identification Performance.** Accuracy vs. membership in top N ranked set using largest boxes for training and five-fold cross validation.

Active learning tries to reduce this effort by selecting valuable samples *first* and thereby better utilise human expert annotation. An incorporation of state-of-the-art active learning frameworks like WALI [39] is therefore considered as immediate future work. A fusion of our presented system with this approach promises a system which can quickly adapt towards novel domains while requiring minimal expert interaction during the initial adaptation phase.

8.3. Integration into Monitoring Practice

Building effective detection and identification frameworks are only first steps towards integrating these computational tools into field practitioners day-to-day work. Whilst speeding up the processing of incoming photographic datasets and allowing for quicker identification of encountered individuals may be the main immediate purpose for using visual animal biometric systems, the availability of independent filtering and validation procedures for accuracy, misclassifications and completeness of encounters provides a further tool for building and maintaining socio-demographic datasets. In particular, the integration of spatially-explicit data from camera trap monitoring with capture-recapture or distance sampling approaches via animal biometric systems may provide an opportunity to generate important and conservation-relevant information on population status, trends and socio-ecology for Mbeli Bai as well as in other settings.

8.4. New Research Horizons

An integration of automated identification software promises to open up a realm of new applications in long-term biological research such as the Mbeli Bai Study, both as stand-alone tools, as well as in integrated combination with camera trap monitoring regimes.

Basic Identity Maintenance and Update. Automated identification software could assist the manual identification process. Particularly, since group stability of western gorillas is low, individuals transfer between groups regularly, groups are formed or dissolved, and groups or individuals may immigrate in, or emigrate out of the bai population [72, 57, 62]. When unknown individuals appear in the bai, it can be challenging to establish with certainty if they are truly new to the population or if they were already known to prior research teams [5].

Spatio-Temporal Coverage. Gorillas only spend little of their time in forest clearings [57], which leads to large gaps in observations and therefore dates of life history milestones. In addition, animal transfers between groups can currently only be estimated with an accuracy that ranges from just a few days up to years [5, 57, 6]. Automatic monitoring with camera traps in the wider geographical area surrounding Mbeli Bai could help to improve the accuracy of the socio-demographic data [23], and improve our understanding the extent and variation of dispersal patterns and group dynamics [2, 23, 17].

Socio-Ecological Insights. Automated monitoring from integrated camera trapping could answer research questions that go far beyond the information that is bound by the limitations of static snap-shots in space and time. The long-term data on the demographics and group structure of dozens of known gorilla units from Mbeli Bai could form a unique and novel combination with spatially-explicit capture-recapture data that, when combined, could tackle a myriad of socio-ecological questions (e.g. ranging patterns, habitat use, seasonal activity patterns, disease screening) as well as population estimates for the wider area, potentially at a park-wide or landscape level [35, 53, 23, 24].

Acknowledgements

This research was partly supported by grant DE 735/10-1 of the German Research Foundation (DFG). We thank the Ministry of Forest Economy and Environment and the Ministry of Scientific Research in the Republic of Congo for permission to work in the Nouabalé-Ndoki National Park. We are grateful to the Wildlife Conservation Society's Congo Program for crucial logistical and administrative support. We are indebted to all research assistants who contributed to the datasets of the Mbeli Bai Study, in particular, Jana Robeyst, Davy Ekouoth, Barbara Hendus, and Vidrige Kandza. We are grateful for the financial support provided by the funders of the study. The contents of this publication are the sole responsibility of its authors and can in no way be taken to reflect the views of the funders.

References

[1] S. C. Amstrup, T. L. McDonald, and B. F. Manly. *Handbook of capture-recapture analysis*. Princeton University Press, 2010.

[2] M. Arandjelovic, J. Head, H. Kuehl, C. Boesch, M. M. Robbins, F. Maisels, and L. Vigilant. Effective non-invasive genetic monitoring of multiple wild western gorilla groups. *Biological Conservation*, 143(7):1780–1791, 2010.

[3] K. M. Boyer-Ontl and J. D. Pruetz. Giving the forest eyes: The benefits of using camera traps to study unhabituated chimpanzees (pan troglodytes verus) in southeastern senegal. *International Journal of Primatology*, 35(5):881, 2014.

[4] S. Branson, G. Van Horn, S. Belongie, and P. Perona. Improved bird species categorization using pose normalized deep convolutional nets. In *British Machine Vision Conference (BMVC)*, 2014.

[5] T. Breuer. *Male reproductive success in wild western gorillas (Gorilla gorilla)*. PhD thesis, University of Leipzig, Leipzig, 2008.

[6] T. Breuer, M. B.-N. Hockemba, C. Olejniczak, R. J. Parnell, and E. J. Stokes. Physical maturation, life-history classes and age estimates of free-ranging western gorillasinsights from mbeli bai, republic of congo. *American Journal of Primatology*, 71(2):106–119, 2009.

[7] C.-A. Brust, S. Sickert, M. Simon, E. Rodner, and J. Denzler. Convolutional patch networks with spatial prior for road detection and urban scene understanding. In *Conference on Computer Vision Theory and Applications (VISAPP)*, 2015.

[8] S. T. Buckland, D. R. Anderson, K. P. Burnham, J. L. Laake, D. L. Borchers, and L. Thomas. *Introduction to distance sampling estimating abundance of biological populations.* Oxford University Press., 2001.

[9] D. Caillaud, F. Levréro, R. Cristescu, S. Gatti, M. Dewas, M. Douadi, A. Gautier-Hion, M. Raymond, and N. Ménard. Gorilla susceptibility to ebola virus: the cost of sociality. *Current Biology*, 16(13):R489–R491, 2006.

[10] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Free-form region description with second-order pooling. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(6):1177–1189, 2015.

[11] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *Transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[12] A.-S. Crunchant, M. Egerer, A. Loos, T. Burghardt, K. Zuberbühler, K. Corogenes, V. Leinert, L. Kulik, and H. S. Kühl. Automated face detection for occurrence and occupancy estimation in chimpanzees. *American journal of primatology*, 79(3):1–12, 2017.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, June 2010.

[15] A. Freytag, E. Rodner, T. Darrell, and J. Denzler. Exemplar-specific patch features for fine-grained recognition. In *German Conference on Pattern Recognition (GCPR)*, pages 144–156, 2014.

[16] A. Freytag, E. Rodner, M. Simon, A. Loos, H. S. Khl, and J. Denzler. Chimpanzee faces in the wild: Log-euclidean cnns for predicting identities and attributes of primates. In *German Conference on Pattern Recognition (GCPR)*, pages 51–63, 2016.

[17] N. Galvis, A. Link, and A. Di Fiore. A novel use of camera traps to study demography and life history in wild animals: a case study of spider monkeys (ateles belzebuth). *International Journal of Primatology*, 35(5):908, 2014.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[19] C. Göring, E. Rodner, A. Freytag, and J. Denzler. Nonparametric part transfer for fine-grained recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2489–2496, 2014.

[20] K. Guschanski, L. Vigilant, A. McNeilage, M. Gray, E. Kagoda, and M. M. Robbins. Counting elusive animals: comparing field and genetic census of the entire mountain gorilla population of bwindi impenetrable national park, uganda. *Biological Conservation*, 142(2):290–300, 2009.

[21] X. He and P. Niyog. Locality preserving projections. In *Neural Information Processing Systems (NIPS)*, volume 16, page 153, 2004.

[22] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using laplacianfaces. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 27(3):328–340, 2005.

[23] J. S. Head, C. Boesch, M. M. Robbins, L. I. Rabanal, L. Makaga, and H. S. Kühl. Effective sociodemographic population assessment of elusive species in ecology and conservation management. *Ecology and Evolution*, 3(9):2903–2916, 2013.

[24] J. S. Head, M. M. Robbins, R. Mundry, L. Makaga, and C. Boesch. Remote video-camera traps measure habitat use and competitive exclusion among sympatric chimpanzee, gorilla and elephant in loango national park, gabon. *Journal of Tropical Ecology*, 28(6):571–583, 2012.

[25] E. J. Howe, S. T. Buckland, M.-L. Després-Einspenner, and H. S. Kühl. Distance sampling with camera traps. *Methods in Ecology and Evolution*, 2017.

[26] B. Hughes and T. Burghardt. Automated visual fin identification of individual great white sharks. *International Journal of Computer Vision (IJCV)*, 122(3):542–557, 2017.

[27] S. K. Jacobson and M. D. Duff. Training idiot savants: the lack of human dimensions in conservation biology. *Conservation Biology*, 12(2):263–267, 1998.

[28] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Conference on Multimedia (MM)*, pages 675–678. ACM, 2014.

[29] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei. Novel dataset for fine-grained image categorization. In *Con-*

ference on Computer Vision and Pattern Recognition (CVPR-WS), Colorado Springs, CO, June 2011.

[30] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg. Hipster wars: Discovering elements of fashion styles. In *European Conference on Computer Vision (ECCV)*, pages 472–488. Springer, 2014.

[31] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations (ICLR)*, 2014.

[32] M. Klailova, C. Casanova, P. Henschel, P. Lee, F. Rovero, and A. Todd. Non-human predator interactions with wild great apes in africa and the use of camera traps to study their dynamics. *Folia Primatologica*, 83(3-6):312–328, 2012.

[33] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *Conference on Computer Vision and Patter Recognition (CVPR-WS)*, pages 554–561, 2013.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Neural Information Processing Systems (NIPS)*, pages 1097–1105. Curran Associates, Inc., 2012.

[35] H. Kühl. *Best practice guidelines for the surveys and monitoring of great ape populations*. Number 36. IUCN, 2008.

[36] H. S. Kühl and T. Burghardt. Animal biometrics: quantifying and detecting phenotypic appearance. *Trends in Ecology and Evolution*, 28(7):432–441, 2013.

[37] H. S. Kühl, A. Todd, C. Boesch, and P. D. Walsh. Manipulating decay time for efficient large-mammal density estimation: gorillas and dung height. *Ecological Applications*, 17(8):2403–2414, 2007.

[38] N. Kumar, P. Belhumeur, A. Biswas, D. Jacobs, W. Kress, I. Lopez, and J. Soares. Leafsnap: A computer vision system for automatic plant species identification. In *European Conference on Computer Vision (ECCV)*, pages 502–516. Springer, 2012.

[39] C. Kding, E. Rodner, A. Freytag, and J. Denzler. Watch, ask, learn, and improve: A lifelong learning cycle for visual recognition. In *European Symposium on Artificial Neural Networks (ESANN)*, 2016.

[40] S. Laing, S. Buckland, R. Burn, D. Lambie, and A. Amphlett. Dung and nest surveys: estimating decay rates. *Journal of Applied Ecology*, 40(6):1102–1111, 2003.

[41] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1449–1457, 2015.

[42] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, 2016.

[43] X. Liu, T. Xia, J. Wang, and Y. Lin. Fully convolutional attention localization networks: Efficient attention localization for fine-grained recognition. *arXiv preprint arXiv:1603.06765*, 2016.

[44] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[45] A. Loos. Identification of great apes using gabor features and locality preserving projections. In *Workshop on Multimedia Analysis for Ecological Data*, pages 19–24. ACM, 2012.

[46] A. Loos and A. Ernst. An automated chimpanzee identification system using face detection and recognition. *EURASIP Journal on Image and Video Processing*, 2013(1):1–17, 2013.

[47] A. Loos, M. Pfitzer, and L. Aporius. Identification of great apes using face recognition. In *Signal Processing Conference, 2011 19th European*, pages 922–926. IEEE, 2011.

[48] L. Maffei, A. J. Noss, E. Cuéllar, and D. I. Rumiz. Ocelot (felis pardalis) population densities, activity, and ranging behaviour in the dry forests of eastern bolivia: data from camera trapping. *Journal of Tropical Ecology*, 21(3):349–353, 2005.

[49] E. McDonald-Madden, P. W. Baxter, and H. P. Possingham. Making robust decisions for conservation with restricted money and knowledge. *Journal of Applied Ecology*, 45(6):1630–1638, 2008.

[50] P. T. Mehlman and D. M. Doran. Influencing western gorilla nest construction at mondika research center. *International Journal of Primatology*, 23(6):1257–1285, 2002.

[51] S. Miura, M. Yasuda, and L. C. Ratnam. Who steals the fruits? monitoring frugivory of mammals in a tropical rainforest. *The Malayan Nature Journal (Malaysia)*, 1997.

[52] D. Morgan, C. Sanz, J. R. Onononga, and S. Strindberg. Ape abundance and habitat use in the goualougo triangle, republic of congo. *International Journal of Primatology*, 27(1):147–179, 2006.

[53] Y. Nakashima, Y. Iwata, C. Ando, C. Nze Nkoguee, E. Inoue, E.-F. O. Akomo, P. M. Nguema, T. D. Bineni, L. N. Banak, Y. Takenoshita, et al. Assessment of landscape-scale distribution of sympatric great apes in african rainforests: Concurrent use of nest and camera-trap surveys. *American Journal of Primatology*, 75(12):1220–1230, 2013.

[54] J. D. Nichols and B. K. Williams. Monitoring for conservation. *Trends in Ecology and Evolution*, 21(12):668–673, 2006.

[55] A. F. O' Connell, J. D. Nichols, and K. U. Karanth. *Camera traps in animal ecology: methods and analyses*. Springer Science & Business Media, 2010.

[56] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference (BMVC)*, 2015.

[57] R. J. Parnell. Group size and structure in western lowland gorillas (gorilla gorilla gorilla) at mbeli bai, republic of congo. *American Journal of Primatology*, 56(4):193–206, 2002.

[58] P. A. Pebsworth and M. LaFleur. Advancing primate research and conservation through the use of camera traps: introduction to the special issue. *International Journal of Primatology*, 35(5):825, 2014.

[59] S. Prasad, A. Pittet, and R. Sukumar. Who really ate the fruit? a novel approach to camera trapping for quantifying frugivory by ruminants. *Ecological Research*, 25(1):225–231, 2010.

[60] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640*, 2015.

[61] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016.

[62] A. M. Robbins, M. Gray, T. Breuer, M. Manguette, E. J. Stokes, P. Uwingeli, I. Mburanumwe, E. Kagoda, and M. M. Robbins. Mothers may shape the variations in social organization among gorillas. *Royal Society Open Science*, 3(10):160533, 2016.

[63] E. Rodner, M. Simon, G. Brehm, S. Pietsch, J. W. Wgele, and J. Denzler. Fine-grained recognition datasets for biodiversity analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR-WS)*, 2015.

[64] E. Rodner, M. Simon, B. Fisher, and J. Denzler. Fine-grained recognition in the noisy wild: Sensitivity analysis of convolutional neural networks approaches. In *British Machine Vision Conference (BMVC)*, 2016.

[65] J. M. Rowcliffe and C. Carbone. Surveys using camera traps: are we looking to a brighter future? *Animal Conservation*, 11(3):185–186, 2008.

[66] J. Roy, L. Vigilant, M. Gray, E. Wright, R. Kato, P. Kabano, A. Basabose, E. Tibenda, H. S. Kühl, and M. M. Robbins. Challenges in the use of genetic mark-recapture to estimate the population size of bwindi mountain gorillas (gorilla beringei beringei). *Biological Conservation*, 180:249–261, 2014.

[67] S. C. Silver, L. E. Ostro, L. K. Marsh, L. Maffei, A. J. Noss, M. J. Kelly, R. B. Wallace, H. Gómez, and G. Ayala. The use of camera traps for estimating jaguar panthera onca abundance and density using capture/recapture analysis. *Oryx*, 38(2):148–154, 2004.

[68] M. Simon and E. Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[69] M. Simon, E. Rodner, Y. Gao, T. Darrell, and J. Denzler. Generalized orderless pooling performs implicit salient matching. *arXiv preprint arXiv:1705.00487*, 2017.

[70] K. Simonyan, A. Vedaldi, and A. Zisserman. Learning local feature descriptors using convex optimisation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(8):1573–1585, 2014.

[71] R. Steinmetz, S. Srirattanaporn, J. Mor-Tip, and N. Seuaturien. Can community outreach alleviate poaching pressure and recover wildlife in south-east asian protected areas? *Journal of Applied Ecology*, 51(6):1469–1478, 2014.

[72] E. J. Stokes, R. J. Parnell, and C. Olejniczak. Female dispersal and reproductive success in wild western lowland gorillas (gorilla gorilla gorilla). *Behavioral Ecology and Sociobiology*, 54(4):329–339, 2003.

[73] E. J. Stokes, S. Strindberg, P. C. Bakabana, P. W. Elkan, F. C. Iyenguet, B. Madzoké, G. A. F. Malanda, B. S. Mowawa, C. Moukoumbou, F. K. Ouakabadio, et al. Monitoring great ape and elephant abundance at large spatial scales: measuring effectiveness of a conservation landscape. *PLOS ONE*, 5(4):e10294, 2010.

[74] A. Swanson, M. Kosmala, C. Lintott, R. Simpson, A. Smith, and C. Packer. Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific Data*, 2, 2015.

[75] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1708, 2014.

[76] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Transactions on Image Processing (TIP)*, 19(6):1635–1650, 2010.

[77] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–591, 1991.

[78] C. E. Tutin and M. Fernandez. Nationwide census of gorilla (gorilla g. gorilla) and chimpanzee (pan t. troglodytes) populations in gabon. *American Journal of Primatology*, 6(4):313–336, 1984.

[79] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(10):1713–1727, 2008.

[80] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2015.

[81] J.-C. Vié, C. Hilton-Taylor, and S. N. Stuart. *Wildlife in a changing world: an analysis of the 2008 IUCN Red List of threatened species*. IUCN, 2009.

[82] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, 2011.

[83] P. D. Walsh and L. J. White. Evaluating the steady state assumption: simulations of gorilla nest decay. *Ecological Applications*, 15(4):1342–1350, 2005.

[84] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 31(2):210–227, 2009.

[85] M. Yang and L. Zhang. Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. In *European Conference on Computer Vision (ECCV)*, pages 448–461. Springer, 2010.

[86] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based r-cnns for fine-grained category detection. In *European Conference on Computer Vision (ECCV)*, 2014.