

Computational ethology for primate
sociality: a novel paradigm for
computer-vision-based analysis of animal
behaviour



Daniel Peter Schofield
Wolfson College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Trinity 2022

Acknowledgements

I'd like to thank my family for all their unwavering love, support and advice throughout this journey. To my partner Lydia Ng whose constant love and patience has carried me through, and who has come to share my love for the Bossou chimpanzees. In no particular order, to my friends at ICEA, who were willing to brave extreme cold weather therapy with me in the Cherwell; to Kat, who I started this journey with, and all my colleagues in the lab who have been great friends and offered support and advice, and have made the lab such a fun and productive space; to all my friends for always being on the other end of the phone and teaming up to pull through challenging times together. To all the Oxford and Kyoto University students and collaborators who helped with research and archive work. I would also like to express gratitude to the DAS and others at the university for their support - Dr Clive Sherlock for helping me develop resilience through Adaptation Practice, Dr Jean Debarros for his mentoring and Kathryn Pearson, who helped me get this thesis down on paper.

I would also like to express my gratitude to research collaborators who have helped me grasp such a broad array of challenging and interdisciplinary skills. For Arran Davis, Alex Mielke, Josh Firth and Greg Albery for help and advice with coding and statistics. To colleagues at Kyoto university - Prof Tetsuro Matsuzawa, Prof Satoshi Hirata, Dr Misato Hayashi, for supporting research in Japan and the amazing footage that I worked on for my DPhil. I am also indebted to all of the researchers and field assistants who sacrificed so much in the field to collect this data, and to the endlessly fascinating chimpanzees of Bossou. To the Visual Geometry Group (VGG) - Prof Andrew Zisserman, Arsha Nagrani, Max Bain, and the other engineers for all their support for my project and whose cutting edge techniques are transforming many fields of research. To my father, Ben Schofield for his ideas and expertise on cloud computing, and Sau Ming Ng who provided technical help and advice for linux programming for processing and cleaning the Bossou data. I would also like to thank my undergraduate supervisor Prof Bill McGrew, whose mentorship over the years has helped sow the seeds of my research and lay the foundations of this academic path.

I also would like to thank the Clarendon fund, Boise Trust, Kyoto University, Wolfson College, Oxford and the Anthropology department for funding my research,

and contributing to fees and living expenses throughout my DPhil - I would not have been able to conduct this research without their support.

Finally I would like to thank my supervisors; Prof Dora Biro for all her support, invaluable feedback, and keen critical eye, which have significantly contributed to my growth as a scientist. To Prof Susana Carvalho, who gave me the opportunity to conduct this research under her tuition. Her energy and dedication, care and kindness have kept me going when times were tough. I am grateful for all her support, guidance and creativity on this journey and for being there for me when I needed it most. I have been fortunate to have two supervisors as inspiring as these, without whom I would not have had the confidence to push the boundaries of research. I would also like to thank my examiners, Dr Alex Alvergne and Prof Andrea Migliano for investing their time in thoroughly reading my work and for offering such valuable feedback and engaging discussions during the viva.

Abstract

Research in the biological and wildlife sciences is increasingly reliant on video data for measuring animal behaviour, however large-scale analysis is often limited by the time and resources it takes to process video archives. Computer vision holds serious potential to unlock these datasets to analyse behaviour at an unprecedented level of scale, depth and reliability, however thus far a framework for processing and analysing behaviour from large-scale video datasets is lacking. This thesis attempts to solve this problem by developing the theory and methods for capturing long-term sociality of animal populations from longitudinal video archives, laying the foundations for an emerging field; computational ethology of animals in the wild. It makes several key contributions by a) establishing the first unified longitudinal video dataset of wild chimpanzee stone tool use across a 30 year period, and building a framework for collaborative research using cloud-technology b) developing a set of computational tools to allow for processing of large volumes of video data for automated individual identification and behaviour recognition c) applying these automated methods to validate use for social network analysis and d) measuring the social dynamics and behaviour of a group of wild chimpanzees living in the forest of Bossou, Guinea, West Africa.

In Chapter 1 I introduce the theoretical and historical context for the thesis, and outline the novel methodological framework for using computer vision to measure animal social behaviour in video. In Chapter 2 I introduce the methodology for processing and managing a longitudinal video archive, and future directions for a new framework for collaborative research workflows in the wildlife sciences using cloud technology. In Chapter 3 I lay the foundations of this framework for analysing behaviour and unlocking video datasets, using deep learning and face recognition. In Chapter 4 I evaluate the robustness of the method for modelling long-term sociality and social networks at Bossou and test whether life history variables predict individual-level sociality patterns. In Chapter 5 I introduce the final component to this framework for measuring long-term animal behaviour, through audiovisual behavioural recognition of chimpanzee nut-cracking. In my final chapter (6) I discuss the main contributions, limitations and future directions for research. Overall this thesis integrates a diverse range of interdisciplinary methods and concepts from primatology, ethology, engineering, and computer vision, to build the foundations for further exploration of cognition, ecology and evolution in wild animals using automated methods.

Contents

List of Figures	xi
List of Abbreviations	xiii
1 Introduction	1
1.1 Animal behaviour	2
1.1.1 Ethology	2
1.1.2 Measuring behaviour	3
1.1.3 Behaviour from a time-depth perspective	4
1.2 A brief history of primatology	4
1.2.1 The influence of Japanese primatology	4
1.2.2 Chimpanzee research in Africa	5
1.3 The field-site of Bossou	6
1.3.1 History of the field site	6
1.3.2 Location and ecology	7
1.3.3 Study population	7
1.3.4 Tool use	8
1.3.5 Field experiments	9
1.3.6 Opportunities for time-depth at Bossou	10
1.4 The emerging field of computational ethology	12
1.4.1 Deep learning and computer vision	13
1.4.2 Computational ethology	15
1.4.3 Social network analysis	15
1.4.4 Computational ethology for primate sociality: A novel frame- work	17
1.5 Thesis outline	18
1.5.1 Aims	18
1.5.2 Research questions	19
1.5.3 Thesis structure	19
Empirical work	21

2	The Bossou archive	23
2.1	Introduction	23
2.1.1	Video data collection in Primatology	23
2.2	Methods	25
2.2.1	Bossou filming protocol	25
2.2.2	Sorting and cataloguing	28
2.2.3	Programming and processing	30
2.2.4	Digitisation	30
2.2.5	File and directory structure	32
2.2.6	Storage and access	32
2.3	Discussion	33
2.3.1	Processing	34
2.3.2	Cloud computing	35
2.3.3	Towards reproducibility in computational ethology	35
3	Chimpanzee face recognition from videos in the wild using deep learning	37
4	Automated face recognition using deep neural networks produces robust primate social networks and sociality measures	59
5	Automated audiovisual behavior recognition in wild primates	107
6	Discussion	127
6.1	Summary	127
6.2	Key contributions	129
6.2.1	Data preservation and access	129
6.2.2	Deep learning	130
6.2.3	Social behaviour across the lifespan	131
6.2.4	Capturing behaviour	133
6.3	Future directions	134
6.3.1	The future of AI for ethology	134
6.3.2	Biological applications	137
6.4	Conclusion	143
Appendices		
A	Cumulative culture in non-humans: overlooked findings from Japanese monkeys?	147
B	Count, Crop and Recognise: Fine-Grained Recognition in the Wild	159

List of Figures

1.1	The geographic region of Bossou, Guinea	7
2.1	Screenshots of video archive footage collected at the outdoor laboratory	26
2.2	Video tapes from the archive during the initial sort and cataloging process.	27
2.3	Video labelling process	29
2.4	Label template for archive tapes	30
2.5	The digitisation station set up in a room on the fourth floor of the PRI, Inuyama.	31
2.6	File directory structure of the Bossou archive.	33
2.7	Schematic diagram of archive infrastructure developed in 2018 . . .	34

List of Abbreviations

AI		Artificial intelligence
CV		Computer vision
PRI		Primate Research Institute
GUID		Global Unique Identifier
AWS		Amazon Web Services
ANN		Artificial Neural Network
GPU		Graphical Processing Unit
CNN		Convolutional Neural Network
SNA		Social Network Analysis
NBDA		Network Based Diffusion Analysis
CC		Cumulative culture

1

Introduction

Contents

1.1	Animal behaviour	2
1.1.1	Ethology	2
1.1.2	Measuring behaviour	3
1.1.3	Behaviour from a time-depth perspective	4
1.2	A brief history of primatology	4
1.2.1	The influence of Japanese primatology	4
1.2.2	Chimpanzee research in Africa	5
1.3	The field-site of Bossou	6
1.3.1	History of the field site	6
1.3.2	Location and ecology	7
1.3.3	Study population	7
1.3.4	Tool use	8
1.3.5	Field experiments	9
1.3.6	Opportunities for time-depth at Bossou	10
1.4	The emerging field of computational ethology	12
1.4.1	Deep learning and computer vision	13
1.4.2	Computational ethology	15
1.4.3	Social network analysis	15
1.4.4	Computational ethology for primate sociality: A novel framework	17
1.5	Thesis outline	18
1.5.1	Aims	18
1.5.2	Research questions	19
1.5.3	Thesis structure	19

1.1 Animal behaviour

1.1.1 Ethology

Behaviour can be understood as the actions and reactions of organisms to changes in their environment (Bateson and Martin 2021) and an adaptive mechanism which enables individuals and groups to respond to ecological pressures faster than genetic evolution through natural selection. Evolutionary theory on natural selection and biological inheritance provides a useful framework for understanding behaviour, and subsequently many aspects of the genetic, biological and social evolution of organisms. The study of animal behaviour lies at the intersection of cognitive sciences, neuroscience and ecology, thus is critical for many fields of research in the biological sciences.

The field of ethology, defined as the biological study of behaviour, has a rich history. At its core, it seeks to explain the proximate and ultimate causes of animal behaviour by identifying the various genetic, ecological, developmental and social drivers of variation in wild animals (Bateson and Laland 2013). The ethologist Tinbergen famously outlined four questions for which behaviour can be modelled, namely causal/mechanistic, ontogenetic, functional, and evolutionary explanations (Tinbergen 1963). Tinbergen's framework emphasised that a theory of natural selection and the environmental context is needed to explain the evolution of organisms' behaviour holistically, stressing the need for observation of animals in their natural environments, as well as experimental investigation of behaviour through controlled laboratory work.

Lab-based research in neuroscience and cognitive sciences tends to focus on mechanistic lines of enquiry. However, given that behaviour represents the output of the brain, and the interface between an organism's genotype, phenotype and the environment, evolutionary theory and consideration of the natural conditions under which animals have evolved, are necessary to understand the complex whole (Kappeler 2010). For example, we cannot reduce the function of the brain solely to

a mechanistic view of the structure and activity of neurons (Pereira et al. 2020). Ethology provides a framework for understanding biological evolution in species; studies of animals in the wild have been key for examining how selective pressures have shaped the evolution of organisms. Ethological work has informed us of many aspects of cognition, ecology and evolution: how animals adapt and respond to changes in their environment (Hockings, McLennan, et al. 2015b; Strum et al. 2015a) the evolution of reproductive strategies, life history and health (Jones 2011; Alvergne et al. 2010), the consequences of social living (Rubenstein and Abbot 2017), theory of mind (Byrne and Bates 2006), cooperation, culture and collective intelligence (Allen 2019; Couzin 2009; Whiten, Harrison, et al. 2022; Migliano and Vinicius 2022). The study of animal behaviour using comparative frameworks has wider implications; for example, for wildlife conservation (Curio 1996; Brakes et al. 2021), modelling human evolution (Jolly 1972; Carvalho, Biro, Cunha, et al. 2012a; Washburn and DeVore 1961) and even the behaviour of machine intelligence (Rahwan et al. 2019).

1.1.2 Measuring behaviour

Capturing behaviour requires identifying patterns and parsing the continuous stream of information as an animal moves dynamically in its environment. Data collection is especially difficult in naturalistic settings, as it often entails continuous monitoring of animals that may be shy or hard to detect, in challenging terrain. Field research is frequently followed by painstaking coding of behaviour from collected data by human researchers. Over the years, methods for measuring behaviour through observational and experimental methods have changed considerably; increasingly, ethology has transitioned from descriptive reporting to a physical science (Brown and de Bivort 2018). Novel technologies, such as remote trackers and data loggers, have opened the way for ‘reality mining’, whereby interactions of entire groups and populations are recorded, enabling data to be collected over large temporal and geographic scales (Krause et al. 2013). As the study of behaviour has become increasingly quantitative, technological advances have led to the accumulation of larger and richer behavioural datasets for researchers to analyse.

1.1.3 Behaviour from a time-depth perspective

A major limitation of ethological fieldwork has been the short-term nature of data collection; however, advances in technology and automation open up possibilities for longitudinal research. Time-depth is critical to investigate how individual behaviour changes across the lifespan, and, at the population level, how ecology and social behaviour interact over large scales to drive selection of certain behavioural traits (Clutton-Brock and Sheldon 2010a; Kappeler and Watts 2012). Such studies of behaviour require long-term and repeated observation of identified individuals. Over several decades, the field has produced longitudinal datasets with serious potential to analyse how cognition, ecology and evolution are related to social behaviour. Although other species have had a longer and more extensive history of observation in the field (Clutton-Brock 2012a), given their close biological relatedness, rich social behaviour and variation between species, primates are useful comparative models for understanding human evolution. Due to their long life spans and marked behavioural variation over time, long-term field sites of primates are essential and offer useful potential for integrating time-depth perspectives of animal behaviour. In the following text, I briefly introduce the history of primatological fieldwork, the potential of the chimpanzee field site of Bossou, the challenges of managing long-term video archives, and the promise of computational ethology as a solution to unlocking these untapped resources.

1.2 A brief history of primatology

1.2.1 The influence of Japanese primatology

Whilst Konrad Lorenz and Niko Tinbergen developed the field of ethology during the 1930s (Burkhardt 1999), early pioneers of primate behaviour conducted experimental work in the laboratory (Köhler 1925; Yerkes and Yerkes 1929) and observations in the field (Nissen 1931; Carpenter 1935). However, it was the work of Kinji Imanishi from Kyoto University which played a critical role in shaping field research of primates (Kappeler and Watts 2012). Influenced by the Kyoto School of

Philosophy, Imanishi and his lineage of students emphasised the interconnectedness of animals in their ecosystems as a collective whole, drawing no boundaries between humans and other animals (Imanishi 1941; Imanishi and Asquith 2013). Whereas contemporary Western scientists typically rejected the complex and emotional social lives of animals, Imanishi sought to understand the evolutionary origins of human behaviour through other species. This openness led to novel discoveries in free-ranging macaques at the Koshima islet, including cultural transmission, individual relationships, dominance hierarchies and kinship networks. Research at Koshima made several methodological contributions - notably, long-term observation, individual identification, and provisioning - which laid the foundations for modern primatology, a fact largely overlooked by researchers in the West (de Waal 2003).

Japan's central role in primatology continued with the naturalistic study of monkeys at Koshima, expanding over a 7-year period to 19 field sites of macaques (Matsuzawa and McGrew 2008). After a decade studying Japanese monkeys, the Kyoto team, including student Junichiro Itani, led expeditions to study gorillas in Uganda (1958) and chimpanzees at Mahale, Tanzania (1965). In the same period, Jane Goodall established research at Gombe National Reserve, Tanzania (1960) in the hope of uncovering the evolutionary roots of human behaviour (Goodall 1963).

1.2.2 Chimpanzee research in Africa

Early work in the field was opportunistic and descriptive, but led to key discoveries. Chimpanzee society was found to be patrilineal, and to consist of 'social units' or communities (Goodall 1965). Furthermore, individuals displayed complex social relationships (Nishida 1968), had linear dominance hierarchies and demonstrated aggression against other communities (Goodall 1986a). As more field sites became established, cross comparisons revealed regional diversity in behaviour, such as social customs in grooming (McGrew and Tutin 1978), suggesting chimpanzees had cultures akin to those in human society. The field has transitioned from ethnographic descriptions to complex multivariate analyses and modelling (Whiten, Ayala, et al. 2017). Large-scale data collection and cross-site comparisons have

revealed remarkable behavioural diversity between sites, without obvious genetic or ecological explanations, thus inferring cultural variation (Whiten, Goodall, et al. 1999; Tutin et al. 2001; Boesch et al. 2020).

1.3 The field-site of Bossou

Given chimpanzees' intelligence, high sociality, and extensive reliance on technology, they are essential for modelling our own evolutionary history, and are ideal candidates to explore aspects of technological and cultural evolution in non-human species in particular (McGrew 2010). Since the early expeditions in the 1960s, data on wild chimpanzees has been collected at 120 field sites (McGrew 2017), several of which have amassed over 50 years of observations (Nishida, Zamma, et al. 2010; Nakamura et al. 2015; Goodall 1986b). However, given the restricted access to these resources, and the methodological challenge of processing these data, thus far few studies have attempted to synthesize longitudinal datasets to model behaviour over many generations.

1.3.1 History of the field site

Of the core 6 long-term field sites, Bossou in southeastern Guinea offers a unique opportunity to measure time-depth and behaviour change (e.g see Appendix A) (Schofield, McGrew, et al. 2018) due to its long research history, behavioural repertoire of the chimpanzee community, and systematic video data collection. Early surveys of Bossou were conducted in the 1960s by Dutch researchers, who documented a community of Western chimpanzees (*Pan troglodytes verus*) and conducted field experiments on predatory responses (Yamakoshi 2011; Kortlandt 1967). Field work, including behavioural and ecological surveys, was later established by Yukimaru Sugiyama of Kyoto University (1976-1986). Sugiyama censused the population, making important discoveries, such as the high rate of grooming between females compared to other populations, as well as several tool use behaviours that were unique to the community (Yamakoshi and Sugiyama 1995; Sugiyama and Koman 1979). From 1986, Tetsuro Matsuzawa joined Sugiyama to conduct systematic

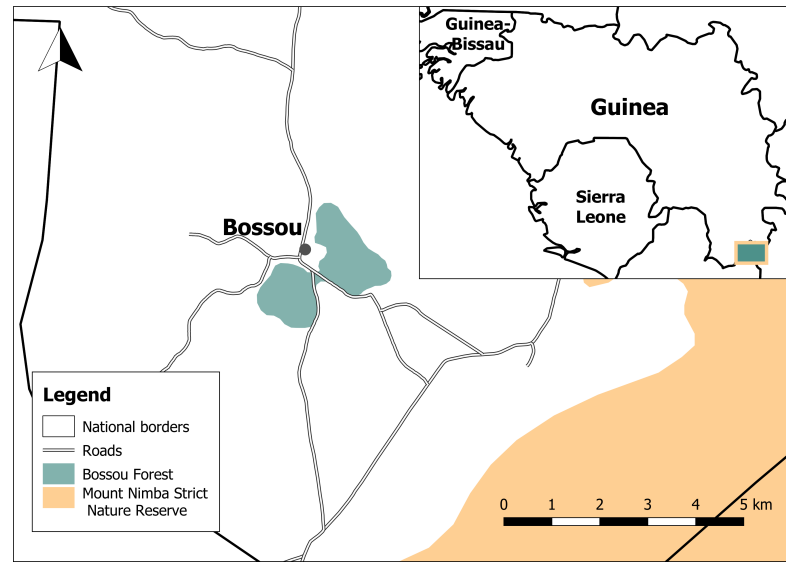


Figure 1.1: The geographic region of Bossou, Guinea (image credit: Katarina Almeida-Warren)

research, and an international team of researchers has returned annually ever since. For over four decades, the field site has been key in providing information about wild chimpanzee subsistence, demography, ecology, cognition, social behaviour and tool use, and in pioneering novel field experiments (Matsuzawa, Humle, et al. 2011).

1.3.2 Location and ecology

The field site is located close to Guinea's borders with Liberia (4km) and Côte d'Ivoire (8km)(Figure 1.1). The forest is home to a small population of chimpanzees (min. 7 - max. 22), occupying a compact area of primary and secondary forest, with a home range of 7-15km². The local area of Bossou is under strong pressure from overpopulation by human settlement, and degradation of primary rainforest.

1.3.3 Study population

The population was naturally habituated due to the close relationship with the Manon people in local villages, who consider the chimpanzees sacred; this is believed to have sustained the peaceful coexistence over the years. The population of Bossou has remained relatively stable across time, at around 20 individuals. See Table S1 in Chapter 3 (Schofield, Nagrani, et al. 2019) for basic information about the individuals

identified and used for analysis in this thesis. The population declined due to the disease epidemic in 2003 (Humle 2011c). Since then, numbers have gradually decreased to 7 individuals, with the population currently under threat of extinction.

1.3.4 Tool use

Bossou chimpanzees have a diverse repertoire of tool use behaviours unique to the community, including leaf sponging, algae fishing, army ant dipping and nut-cracking (Sugiyama 1995a; Sugiyama 1995b; Humle 2011b; Sugiyama and Koman 1979; Matsuzawa 1994). This repertoire is especially diverse given the small population size, and comparison with other long-term field sites has indicated that 35% of the tools are unique to the Bossou community (Humle 2011b). A focus of research has been nut-cracking behaviour. At Bossou, chimpanzees use a stone hammer and anvil in a complex sequence of actions (Carvalho, Biro, McGrew, et al. 2009b); this behaviour has been observed at other field sites, but is not widespread, mainly being found in West Africa. It is absent in East and Central Africa, for example, despite the availability of oil palm nut-bearing species, stones, wooden clubs and tree roots across many sites (McGrew, Ham, et al. 1997). Nevertheless, chimpanzees at a reserve in Cameroon (*P.t. ellioti*) have been observed using stone tools (Dutton and Chapman 2015), and detailed ecological surveys are needed to assess the opportunities for chimpanzees to exploit such resources at other sites (Carvalho, Matsuzawa, et al. 2013; Koops, McGrew, et al. 2013). A unique feature of nut-cracking at Bossou is the use of composite tools, including movable hammers and anvils, to crack open the hard exterior of oil palm nuts (*Elaeis guineensis*) and access the edible kernel (Carvalho, Biro, McGrew, et al. 2009a). Nut-cracking and oil palms form a key aspect of year-round nutrition, but usually peak during the wet season (June-August), and the dry season (November-December), when fruit availability is low (Humle 2011a; Yamakoshi 1998; Almeida-Warren et al. 2021). This is unlike other sites, such as Taï, where chimpanzees also use wooden hammers and substrates predominantly to process different species of nuts that are only seasonally available (e.g. *Coula edulis* or *Panda oleosa*).

1.3.5 Field experiments

Influenced by laboratory work on ape cognition and field methods in Japan, Tetsuro Matsuzawa pioneered a quasi-experimental research design by provisioning nuts and stones in the forest clearing at the top of a hill at the core of the home range (Matsuzawa 2011a; Matsuzawa 1994). Stone tools were numbered, and the material properties measured, so that individual tool use preference and performance could be recorded systematically across field seasons.

Social learning and culture

The unique field experimental approach and close observation over the years yielded important discoveries into ontogeny, social learning and cultural transmission. Tetsuro Matsuzawa coined the term ‘master by apprenticeship’ as a model of learning (Matsuzawa, Biro, et al. 2008a; Inoue-Nakamura and Matsuzawa 1997). At a young age, infants are exposed to nut-cracking by their mother, and social tolerance by other individuals and close observation facilitates them to acquire the skill. Several individuals never master the technique, however, highlighting the complexity of the task and indicating a ‘critical window’ of learning (Inoue-Nakamura and Matsuzawa 1997). Provisioning of novel nut species also revealed transmission through the group, with juvenile chimpanzees closely observing nut cracking by older conspecifics (Biro, Inoue-Nakamura, et al. 2003).

Subtle variation

For nut-cracking, there is evidence of variability between groups, and between nearby sites, indicating that the behaviour is culturally mediated (Luncz, Svensson, et al. 2017). Yeale chimpanzees in the Nimba mountains neighbouring Bossou crack oil palm nuts and Coula nuts (*Coula edulis*) on fixed substrates (rocky outcrops, tree roots) (Humble and Matsuzawa 2004), whereas Seringbara chimpanzees do not, despite the availability of these resources (Hockings 2007). Recent field experiments also support the idea that nut cracking is a challenging skill to learn, as the chimpanzees in the Nimba mountains do not spontaneously crack nuts when

provisioned with stones and nuts (Koops, Soumah, et al. 2022). Opportunities for horizontal and cultural transmission (e.g exploitation of novel species of nut) between groups are limited to female emigration from natal groups. However, females appear to conform to the behaviours of their new communities, and adult chimpanzees are conservative about adopting new behaviours, which may be a factor in maintaining cultural divergence between neighbouring groups (Biro, Inoue-Nakamura, et al. 2003; Luncz and Boesch 2015; Luncz, Mundry, et al. 2012). Recent evidence points towards a variation in the material culture between communities, and the efficiency of wooden vs stone tool users (Luncz, Sirianni, et al. 2018).

1.3.6 Opportunities for time-depth at Bossou

Technological evolution

The discovery of percussive technology and a lithic material culture at Bossou has been crucial for providing insights into the evolutionary origins of stone tool technology in our hominin ancestors (Carvalho and McGrew 2012). Analysis of nut-cracking at Bossou has shown that different resources are exploited in the form of raw materials (harder stones for different species of nuts) and specific stones (hammer-and-anvil) are selected and preferred for different techniques of nut-cracking; over time, this food processing activity creates discernible wear patterns on tools (Carvalho, Cunha, et al. 2008; Carvalho, Biro, McGrew, et al. 2009b). Excavations of nut cracking sites in West Africa show that this behaviour has persisted in some communities for at least 4,300 years (Mercader, Panger, et al. 2002; Mercader, Barton, et al. 2007). This longevity stimulates some interesting questions: Is there detectable individual-level variation in nut-cracking technique? Can beneficial improvements be adopted and maintained, and how do social dynamics, ecology and repeated exposure to learning opportunities drive behavioural variation over time? The complexity and time-depth of this tool use makes it a good candidate as a behaviour which might be subject to cultural evolutionary processes, and a parsimonious model for hominin technological evolution.

Bossou in the Anthropocene: a driver of change?

In addition to these interesting avenues for exploring tool use across generations, long-term research has opened up possibilities to study how the population's sociality and behaviour respond to changes to their environment. Human activities threaten animal cultures in the wild (Kühl et al. 2019), but also introduce novel problems which are exploited by behaviourally flexible species, such as great apes (Gruber, Luncz, et al. 2019). Increasing habitat fragmentation and encroachment by local human populations at Bossou has changed the ecological landscape and behaviour of chimpanzees, driving the prevalence of behaviours such as crop raiding (Hockings, McLennan, et al. 2015a).

Human interference, such as habituation and provisioning of primates by researchers, can affect cultural traditions in wild populations. For example, introducing novel food resources led to innovative behaviours, such as the food processing traditions (Schofield, McGrew, et al. 2018) and hot-spring bathing (Suzuki 1965) of Japanese macaques. Such interventions have measurable outcomes for reproduction and survival, affecting population growth and inter-birth intervals (Strum et al. 2015b; Watanabe et al. 1992), disease transmission (Sarabian and MacIntosh 2015) and even conferring beneficial effects such thermoregulation and stress-relief (Takeshita et al. 2018). Habitat encroachment at Bossou has impacted the population dynamics, leading to changes in age, structure and size of the community due to disease epidemics and the dispersal of individuals (Humble 2011c). Thus, this raises further interesting questions concerning demographics, life history, social dynamics and ecology. How does social behaviour at the individual and group level change with the loss of key individuals? How do aspects of changing demography, such as ageing population structure, affect social interactions, social transmission across generations, and development of behaviours over time?

Longitudinal video records

Opportunities to answer these questions and measure behaviour change over generations in wild populations has been limited, as historical ethnographic reports

often lack enough detail to measure subtle changes in behaviour comprehensively (see Appendix A). Use of video information is a key method in primatology, as the data can be reanalysed, however encoding this information can be extremely time-consuming. Systematic data collection using video cameras has produced a long-term archive of chimpanzee behaviour at Bossou, however traditional manual methods of analysing video present a bottleneck for unlocking its full potential. In Chapter 2 I highlight the methodology and protocol followed to curate 30 years of video data, which enables fine scale analysis of behaviour over several generations. In the following section I introduce the challenges of manual research of long-term datasets such as Bossou, and introduce the novel framework of computational ethology.

1.4 The emerging field of computational ethology

Challenges of manual research

Once the infrastructure for video archives has been developed, and the data systematically processed and organised, video datasets offer unique opportunities to model behaviour and cultural evolution of wild populations. Harnessing the full capability of these resources requires a series of methodological breakthroughs, as well as careful consideration of how multiple disciplines can be used together effectively to develop a new research paradigm. Typically, researchers have relied on manual coding to painstakingly process video datasets (Zuberbühler and Wittig 2011; Marchant et al. 1995). Despite open-source software projects which make ethological video coding more efficient (Friard and Gamba 2016), this work can be time-consuming and work-intensive, reducing the depth and breadth of analyses or limiting research to short-term field seasons. Multiple annotators can be used to accelerate the process, with inter-observer reliability using statistical tests to ensure consistency. However, as the size of video archives increases, manual annotation projects with multiple individuals can become complex and unsustainable, especially for comparative research of multiple species or populations. Outsourced solutions such as (Amazon Web Services Mechanical Turk offer paid labour for annotation tasks, or crowd-sourcing through citizen science (Dickinson et al. 2010) (which

leverages members of the public to process data online) can have issues with accuracy and bias (Northcutt et al. 2021), and also cannot be used for tasks in science for which expert researchers are required. Although innovating new frameworks for automating the measurement of behaviour from video requires a significant time investment for development and validation, in future new automated frameworks could impact the scale and depth of animal data that can be analysed.

1.4.1 Deep learning and computer vision

The field of machine learning attempts to develop computer programs which are able to ‘learn’ from experience, using data to make predictions or complete a specific task, without each step being manually coded by a human programmer (Samuel 1962; Howard and Gugger 2020). A major challenge in this field has been developing algorithms which do not rely on the representation of the data they are given - simple predictive models such as logistic regression rely on feeding a model a defined set of features (or ‘variables’) which correlate with and successfully predict various outcomes (Goodfellow et al. 2016; Lemon et al. 2003). Traditionally, many artificial intelligence problems have been solved using this ‘hand-crafted’ approach which relies on humans to transform raw data into representations, to for example, determine which treatment to give a patient given certain health metrics from medical records (Lemon et al. 2003). However, for many complex tasks which involve recognition using complex multi-media information (audio, text, images etc), it is often unclear which features are best to extract, as handcrafted features are often vulnerable to complex variation in real world settings. For example, for a computer vision task such as detecting cats in images, a useful predictive feature may be to manually encode the geometric shape of cat ears using pixel values; however, this shape is easily distorted by changes in pose, lighting, and perspective. A powerful alternative approach is to use an algorithm to discover the best performing set of features which represent the problem, otherwise known as ‘Representation learning’(Bengio et al. 2013).

Deep Neural Networks

Such flexible approaches to machine learning were envisioned early in the field with the conception of Artificial Neural Networks (ANNs), which are loosely inspired by neurons in the brain. Early on it was demonstrated that information could pass through a network to make simple predictions (McCulloch and Pitts 1943); however, the popularity of ANNs fluctuated over the years, in part due to their complexity and the lack of computational resources. Increasing availability of affordable Graphical Processing Units (GPUs), greater processing power and rich open source video datasets, has led to more practical and powerful techniques - now known as ‘deep learning’ or ‘deep neural networks’ (Trask 2019). Modern ANNs utilise more layers and greater computation to perform complex tasks such as object detection (predicting where an object is in an image) and classification (predicting what the object is) from images (LeCun et al. 2015; Parkhi et al. 2015).

Ultimately, deep learning is more effective and flexible as it can learn a task by using many processing layers to represent abstract phenomena as a nested hierarchy of simpler concepts (e.g breaking down an images from edges identified by contrast of adjacent pixels, to collections of edges and eventually whole detectable objects) (Goodfellow et al. 2016). Rather than humans hand selecting important features for prediction in a task, deep learning models use millions of parameters to automatically select those which improve the predictive abilities of the system. The growth of the broader field of machine learning, along with increasing availability of large image datasets (Deng et al. 2009), has led to significant advances in computer vision. Applied to animal behaviour, this can be utilised to identify different species or individual animals moving across video frames (Weinstein 2018; Mathis et al. 2018). Deep learning is revolutionising numerous fields of research, such as biology (Jumper et al. 2021; Christin et al. 2019), medicine (Ching et al. 2018), natural language processing (Young et al. 2018) and computer vision (Voulodimos et al. 2018).

1.4.2 Computational ethology

The field of ‘computational ethology’ is emerging (Anderson and Perona 2014; Mobbs et al. 2021), which applies computer vision to detect and recognise animals in images for behavioural analysis. This approach uses machine learning algorithms to automatically process video to produce frame-level information, such as a database of animal movements across time and space, which can be used for statistical and behavioural analysis. Thus far, research in this space is predominantly for lab-based studies in neuroscience (Mathis et al. 2018), and application to naturalistic field data has been limited partly due to the complexity and accessibility of the datasets. In their natural environments animals move dynamically and are often hard to detect, due to variable lighting, vegetation, and occlusion; thus, experts are often required for annotation. Advances in the field of deep learning and computer vision have typically been developed using constrained/artificial video datasets of humans. Nevertheless, many of the tools - such as individual tracking and recognition (Cao et al. 2018), pose estimation, objection recognition (Ragusa et al. 2020), gaze direction (Abdelrahman et al. 2022; Marin-Jimenez et al. 2019) and behaviour recognition (Carreira and Zisserman 2017)- have been developed to a high level of performance for humans, and can be adapted to capture behaviour in other species. Thus far, however, little research has integrated multiple models together as a framework to both automatically identify individuals and model behaviour in unconstrained, real-world settings, otherwise known as ‘in the wild’.

1.4.3 Social network analysis

In addition to new methodological techniques for processing video information, in recent years there have been advancements in statistical analysis for measuring behaviour. Social relationships of animals can be quantified as a way of describing the structure of the group (Hinde 1976b; Carpenter 1935). Networks are made up of a series of individuals, represented as ‘nodes’, and connections between them defined as interactions or associations such as proximity or behaviour. Modelling social relationships in this way has the advantage of quantifying individual-level

social behaviour, as well as structure at the group-level, to examine, for example, how groups respond collectively to environmental pressures (Shizuka and Johnson 2019). With the increase in statistical and computational resources, social network analysis has become more practical (Farine and Whitehead 2015). Individual interactions used for generating social network measures can be automatically computed, which has enabled fine-grained analyses of the fitness consequences of social position (Lehmann et al. 2016, the interaction of ecology, group structure and individual behaviour (Albery, Kirkpatrick, et al. 2021). Recent increases in tracking technology have allowed the high-density automated sampling of animals in the wild to collect social network data on a large scale using animal-borne passive transponders and biologgers (Krause et al. 2013; Strandburg-Peshkin et al. 2015). Although these are relatively inexpensive, fitting devices can be logistically difficult and, as such, they are not feasible for many species (such as chimpanzees). Recent use of camera-traps highlights that robust social networks can be collected using camera trap data (McCarthy et al. 2019), but video coding is extremely time-consuming, limiting the scope of analyses. Collection of video data from field research and conservation has led to the accumulation of large-scale video archives, which have huge potential to tackle cultural evolutionary processes over longitudinal time-scales (Clutton-Brock and Sheldon 2010b), thus new methods for efficiently quantifying social interaction from footage is much needed. One area of promise for modelling cultural evolution is dynamic network-based diffusion analysis (NBDA) (Hoppitt 2017; Canteloup et al. 2021) - this method can quantify the time series of acquisition of traits, such as disease or behaviours. Currently, application of NBDA is lacking - although automated tracking devices are excellent at capturing social associations and proximity based data, recording behaviour is still reliant on manual approaches. Tracking fine-scale behavioural change, and matching it up to this social association data is currently limiting the application of social network analysis for cultural evolution (Wild et al. 2019), and previous approaches have either been restricted to single behaviours or short timescales.

1.4.4 Computational ethology for primate sociality: A novel framework

The advancement of deep learning has high potential to implement an automated framework for quantifying wild chimpanzee behaviour over longitudinal datasets. Since information can be extracted from videos and images, the movement and behaviour of animals across time and space can be tracked and analysed. Together, the combination of new methodological approaches to quantifying animal behaviour using social networks, the availability of large-scale video datasets for retrospective analysis, and automated computational processing have opened up new opportunities to measure sociality and behaviour naturalistically from video footage of wild chimpanzees.

In order to capture social behaviour from video, there are several core tasks that an automated system needs to solve:

- Individual identification - individuals must be detected, tracked and recognised continuously and across years despite changes to physical appearance (with age, development etc.).
- Social structure - social network analysis can measure individual-level patterns of association and social bonds and group level structure. Social networks can be used to examine how demographic life history or environmental variables affect social patterns, and how behaviour is socially transmitted in the population.
- Behavioural recognition - individual social behaviour using individual recognition and network analysis can be layered with behaviour recognition, to examine social transmission, adoption and modification of behavioural traits over time.

1.5 Thesis outline

In order to measure behavioural processes and sociality change over multiple generations, longitudinal data is critical (Clutton-Brock 2012b; Kappeler and Watts 2012; Clutton-Brock and Sheldon 2010b). With a burgeoning field of research utilising time-depth frameworks for exploring socio-ecology (Shizuka and Johnson 2019; Shizuka and Johnson 2020) collective intelligence (Biro, Sasaki, et al. 2016), and cultural evolution in non-humans (Whiten 2019), new tools for the naturalistic study for social behaviour at scale and efficiency is timely and much needed. This project is pioneering in its application of novel methodologies to analyse a 30 year video dataset spanning multiple generations.

The main goal of this thesis is to develop a new paradigm for the automated analysis of behaviour from video, through the development of visual AI to measure chimpanzee social dynamics and nut-cracking from a longitudinal perspective. Given that extracting information from large video archives poses numerous challenges, several core methodological aims are outlined below.

1.5.1 Aims

1. Develop the strategy for the preservation, access and analysis of archive video footage and a reproducible framework for analysing video data collected in the field.
2. Develop new computational methods to sample longitudinal data (individual identification, social networks and behaviour recognition)
3. Analyse the social dynamics and structure of the Bossou community over multiple years, and evaluate the robustness of deep learning for measuring behaviour of wild chimpanzees.

1.5.2 Research questions

1. Can chimpanzee behaviour and identity be reliably recognised automatically from raw video footage using Deep Convolutional Neural networks (i.e. Deep Learning)?
2. Can face recognition technology be used to produce robust chimpanzee social networks, and how does varying the sampling affect the results?
3. How does wild chimpanzee social structure change over time, and how do age and sex affect individual social network positions at Bossou?

1.5.3 Thesis structure

In the following empirical chapters I investigate how application of deep learning can be used as part of a novel framework of computational ethology to give insight into social structure and individual-level association patterns from long-term video datasets. In collaboration with the Engineering Department at the University of Oxford, I develop and test the application of deep learning for automating behavioural analysis, and use this system to measure chimpanzee sociality over multiple years.

Chapter 3 outlines the key first step for developing individual identification, to track individuals over time and generate social networks. In Chapter 4, using this method on a larger sample of the archive, I perform a series of analyses to validate and optimise data collection, to investigate how sampling rates affect social networks, and to examine individual and population level change in social patterns over time, and how this relates to key life history parameters such as age and sex. Chapter 6 demonstrates how the individual identification can be layered with automated behaviour recognition of chimpanzee nut-cracking. Finally, I conclude with a discussion of the main findings and key contributions of the thesis, and highlight how this framework can be used to measure sociality, cultural evolution in non-human species from video archives. I further explore the limitations of such work, along with avenues for future research.

This framework utilises a diverse and inter-disciplinary tool set of methods from machine learning and computer vision and animal behaviour to enable detailed investigations of animal behaviour and sociality on a large scale and in unprecedented levels of detail. This research opens the way for more in-depth investigations of behaviour, ecology and evolution, to reconstruct the cultural history of our closest living relatives and our own evolution.

Empirical work

2

The Bossou archive

Contents

2.1	Introduction	23
2.1.1	Video data collection in Primatology	23
2.2	Methods	25
2.2.1	Bossou filming protocol	25
2.2.2	Sorting and cataloguing	28
2.2.3	Programming and processing	30
2.2.4	Digitisation	30
2.2.5	File and directory structure	32
2.2.6	Storage and access	32
2.3	Discussion	33
2.3.1	Processing	34
2.3.2	Cloud computing	35
2.3.3	Towards reproducibility in computational ethology	35

2.1 Introduction

2.1.1 Video data collection in Primatology

In the 60 years since research was pioneered at Japanese macaque field sites, there has been a shift in field methods in primatology (Setchell and Curtis 2011). Early pioneers such as Jane Goodall took camera equipment to supplement their handwritten field notes and document their discoveries. Over the years, research

methods such as habituation, close observation and filming have enabled audiovisual multimedia to be collected for long-term field sites (Plooij et al. 2015; Clutton-Brock 2012a). Impressive examples include an ethogram of 1400 chimpanzee behaviours at the community and individual level (Nishida, Zamma, et al. 2010), large-scale camera trap grids such as the PanAf project, or long-term field sites such as Gombe. Reliance on this video data for behavioural studies has become a staple of the field, enabling researchers to retrospectively code behavioural events accurately and in detail for statistical analysis (Martin and Bateson 2007). However although a wealth of video footage now exists, the processing of that content is still a largely manual ‘artisan’ process by highly skilled researchers. The selection, real-time-review and manual labelling of clips is time-consuming and subjective, restricting the sample size for reproducible results. The new digital approach seeks to systematize this process: preserve critical footage for future research, respect the rights of archive owners, leverage the power of computer vision algorithms trained by experts, and track activity for authenticated researchers accessing content around the world through the public cloud to improve research efficiency, reduce costs and encourage collaboration.

The field has transitioned from natural history and opportunistic descriptions documented in notebooks, to rigorous and systematic research utilising advances in technology, with rich multimedia information which can be analysed and shared (McGrew 2017). Handheld recording equipment has steadily reduced in size and cost, with increased memory capacity and image quality, allowing for flexible deployment in the field (Rowe and Myers 2011). From the 1980s, handheld video cameras were increasingly used at long-term sites such as Bossou, Gombe and Tai (Matsuzawa 1994; Wittig 2018). Use of video is now pervasive in primatology, and facilitates fine-scale video coding of behaviours which was previously only feasible for captive animals in controlled laboratory environments. It has been key to discoveries in multiple domains; for example, gestural communication (Hobaiter and Byrne 2011), tool use (Musgrave et al. 2020), social learning (Lonsdorf et al. 2004), food sharing (Gilby 2006) and locomotion (Carvalho, Biro, Cunha, et al. 2012b). Increasing use of remote sensing camera traps are generating large scale image

datasets, capturing behaviour of unhabituated animals across large geographic and temporal ranges (Swanson et al. 2015; Musgrave et al. 2020; Kühl et al. 2019). Although these accumulating resources can be harnessed for primate behaviour research and conservation, methods for management, preservation, collaborative access, and analysis of such video archives are still undeveloped in the field, leaving many of these resources untapped.

Research at Bossou is unique due to its reliance on field experiments and video data collection. Tetsuro Matsuzawa utilised his research of captive chimpanzees in the laboratory (Matsuzawa 2003; Matsuzawa 1985) to develop a field experimental approach at Bossou, by conducting systematic and intensive video data collection of naturalistic behaviour at the ‘outdoor laboratory’ (Sakura and Matsuzawa 1991; Matsuzawa 1994). The annual recording led to the accumulation of a long-term video archive of over 1200 hours stored at the Primate Research Institute (PRI) in Japan, yet this resource had not been systematically organised or digitised. This video archive dataset details the complex percussive behaviour of nut-cracking - a key adaptive strategy for the Bossou chimpanzees - and is particularly relevant for answering questions on chimpanzee culture, tool use and ontogeny (see introduction section 1.3.4). Data collected and analysed for this thesis is composed of a sample of experimental nut-cracking sessions extracted from 30 years of records, focusing on the individuals present in the chimpanzee population between 1988 and 2019. In this chapter I outline the methodology for the preservation of content with controlled accessibility to encourage academic collaboration and how to leverage recent commercial advances in cloud based video analysis for animal behaviour.

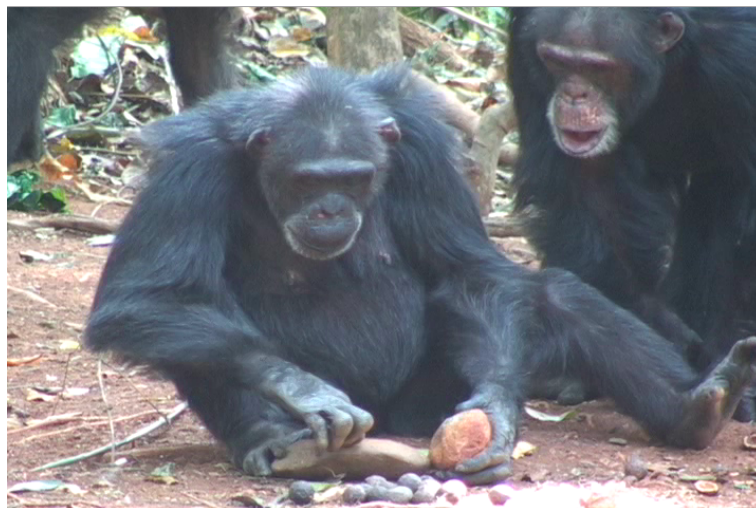
2.2 Methods

2.2.1 Bossou filming protocol

The ‘Bureau’, or ‘outdoor lab’ is located at a forest clearing on Mount Gban (07°39’N; 008° 30’W) and was first established in 1991. Use of tripods for shot stability, handheld Sony video cameras (e.g Sony Digital Handycam, DCR-VX 1000; Sony Digital Handycam, DCR-PC 110), and clear visibility from multiple camera



(a) Middle distance shot of several individuals at the outdoor laboratory cracking nuts and eating palm fruit in 1997



(b) Zoom shot of Yo cracking nuts using a stone hammer and anvil while observed by Foaf at the outdoor laboratory in 2008.

Figure 2.1: Screenshots of video archive footage collected at the outdoor laboratory

angles allowed for stable images to capture behaviour in a high level of detail. Circa 20-30h of observations were conducted per year during the dry season (December - February). In December 2009, another location was established ('Salon') on Mount Guein, a short distance from the KUPRI-IREB base research station (Matsuzawa 2011b), where data has now been collected more frequently since 2012. A further 217 tapes contain footage collected in other locations in the village, or focal follows in the forest, which were digitised and catalogued but not analysed in this thesis. 1312



Figure 2.2: Video tapes from the archive during the initial sort and cataloging process.

tapes have been digitised, collected from yearly expeditions and systematic filming from focal follows and observation at the Bossou outdoor laboratories spanning 1988-2019. The archive predominantly comprises analogue tapes stored at the PRI.

Individual researchers have previously accessed the archive in Japan for projects, using the tapes to manually playback and annotate selected video clips to support their research. However tapes have a limited life and wear out. Older tape technologies may no longer be accessible as playback machinery becomes ‘end-of-life’. A key goal of this research project was to organise and digitise the dataset to secure it for the future and make it accessible to other research teams with the explicit permission of the institute, as well develop a general framework for preserving, processing and maintaining large video archives to enable new avenues of research in animal behaviour.

2.2.2 Sorting and cataloguing

The initial objective was to locate, organise and catalogue the existing tapes. Three trips were conducted to Japan, one month in February 2017 and July 2017, and 2 months in March-April 2018; sorting and cataloguing tapes, digitising, as well as implementing protocols to ensure a reproducible process, and so the origins of the content could be tracked.

The first goal was to organise tape content by year. Tapes were sorted and roughly categorised based on labelled dates (Figure 2.2). The archive comprises several main data formats; ‘Hi8’, ‘MiniDV’, ‘Betacam’, ‘VHS’. Annual filming by researchers generated tape sets, sequential in time and often using the same camera. Tapes had been stored in rooms adjacent to the chimpanzee enclosures at the PRI Language and Intelligence section; due to the high levels of humidity, some had sustained damage and had to be professionally cleaned ($n = 72$). Labels were often written in Japanese, with the added complexity that each original tape set (the ‘master’ set) collected in the field had a backup set (‘copy’ set). Tapes were often missing from one or both sets, and there was varying quality due to damage or inherent quality/resolution of the tape format. Tapes that were recorded in sequence from the same camera were often labelled in numerical order, and retained the same handwriting, tape type or label style and could thus be easily grouped together. For each tape and corresponding digital file, a randomly generated 128-bit long string of numbers and lowercase letters was used as a Global Unique Identifier (GUID) (Figure 2.4). This same approach is used by large-scale systems such as Youtube - GUIDs are easily randomly generated using a function, online generator or spreadsheet plugin. Individual research file naming conventions are often idiosyncratic and difficult to query in databases - randomly generated GUIDs solve this problem, making content easily traceable and forming a simple and standardised system for cataloguing content. After a rough sort, tapes that followed in sequence were grouped and identified by a ‘tape set’ GUID. Given that multiple cameras were used at Bossou, often several ‘tape sets’ were recorded in parallel. Original content (‘master tapes’) were separated from duplicate copies and prioritised for digitization.

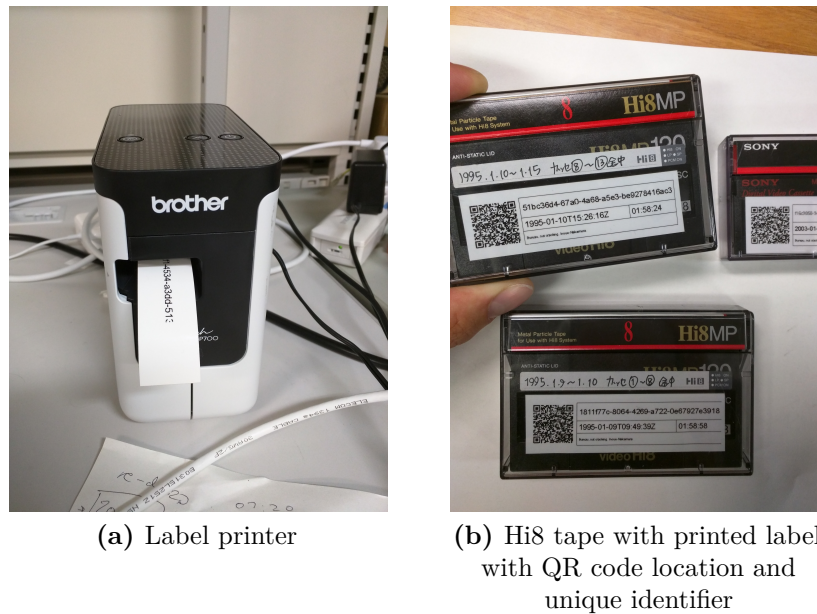


Figure 2.3: Video labelling process

A separate GUID was made for each tape, however each copy was traceable to its original master tape. It was always intended to register this ‘local’ GUID in a proper DOI registry to enable citation of video in the same way that datacite and crossref make research objects easy to find, cite, link, assess, and reuse.

Each tape was labelled using a Brother TD2020ZU1 electronic printer with a QR code which linked the tape’s location to the PRI in case of removal (Figure 2.3), basic information on the date and runtime, along with a GUID (Figure 2.4). To save time and reduce human error, labels were automatically generated and batch printed using ‘Brother P Touch Editor’ software connected to the archive spreadsheet database. Every tape and every file was traceable using its GUID - if any digital files were corrupted, or digitisation errors were found then a complete audit trail of that tape could be reconstructed. In addition to generating a GUID, key metadata was recorded including the tape sequence number and label, start time metadata and tape type.

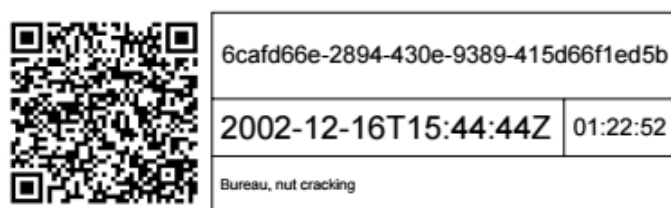


Figure 2.4: Label template for archive tapes. Top: 128 bit Global Unique Identifier. Middle: ISO 8601 format date/timestamp. Middle right: tape duration in hh:mm:ss. Bottom: notes on content or filming location. Left: QR code used for PRI location.

2.2.3 Programming and processing

2.2.4 Digitisation

Once a set of tapes had been sorted, labelled and catalogued, each set was digitised sequentially. A basic protocol was followed that if the original master tape was missing or of poor quality, it's duplicate/copy file was used, retaining the GUID for the master tape. VHS and Hi8 Tapes were captured onto memory cards using an 'IO data' digital capture device (GV-HDREC/E) connected to a Sony VHS /Hi8 player. For mini-DV tapes, mini-DV players and a firewire adapter captured content directly onto laptops and desktop computers (Figure 2.5). Videos were digitised using a firewire adapter cable and the open source software 'Vidi'. Once tapes were digitised, information was catalogued in the spreadsheet database, and files transferred to a local hard drive for further processing. Digitisation took the length of the tape, making the process time intensive, and thus could be only processed in real time. To speed up the process, multiple devices were used in parallel by Kyoto University students and research assistants and devices were left to run overnight.

The shift to digital cameras has transformed how much data can be stored and analysed from the field. Large, high quality videos of behaviour can now be captured, with reliable time/date metadata information, and often location data, to provide rich information on the context of the footage which useful for data analysis (e.g time and location). In addition to the organisational challenges of maintaining large video archives, archival footage filmed in the pre-digital era have numerous challenges: analogue tapes are vulnerable to damage and errors in the digitisation process, and



Figure 2.5: The digitisation station set up in a room on the third floor of the PRI, Inuyama.

the output files can be in unusual file formats, and the resulting footage can be lower in quality or lacking crucial metadata, which may need additional processing to extract. Video files need to be cleaned and compressed, which is essential for the reproducibility and access of video resources in primatology, so as to have consistent file structure and to provide as much context for the footage as possible.

Given the scale of the dataset (both in the size and number of files), the open source operating system GNU Linux was used for scripting (using the command line/bash terminal) to efficiently batch process and manipulate video files for metadata extraction, file checking, moving and file renaming (Haddock and Dunn 2011). The software package FFmpeg (a multimedia framework for processing

video and audio files) was used to extract metadata, combine/cut videos, compress files and convert video file types.

2.2.5 File and directory structure

Once digitised, files were named using the GUID labelled on the physical tape, and then stored within a directory named after the tape set GUID (Figure 2.6). Any metadata (such as mini-DV ‘.dv’ file extensions, which contain frame-level time/date information) was extracted into text files before the videos were batch compressed using FFMPEG and converted to MP4 file extensions. Some mini-DV tapes had previously been digitised using outdated Apple editing software (Imovie 2009). The algorithm used to automatically split video files into clips performed poorly due to visual glitches and timestamp errors on the original tapes, producing many small video files which broke up the continuous sequence of shots, making coding and observation of the behavioural events challenging. To ensure there was a single video file corresponding to tape, any files that had been previously digitised and split into clips were concatenated into a single file, matching the GUID of the physical tape. This makes locating content simple, clean and effective. Linux Bash scripts were used to organise cluttered files into different folders, and check that missing tapes were found and transferred to the central hard drive. Files that were digitised from CD-ROM produced VOB files, Hi8 produced MP4, and DV tapes produced dv files. All of these had to be digitised and converted to MP4 to reduce the file size and maximise compatibility for multiple devices.

2.2.6 Storage and access

Data was initially processed and stored on 1TB USB hard drives, before being transferred to a Synology Network Attached Storage device at the PRI. A mirror device was established at the Primate Models for Behavioural Evolution Lab in Oxford, which could be accessed remotely online. Data was synced between the two devices for local processing in Oxford. Amazon Web Services (AWS) provides a series of low cost cloud storage solutions (see ‘cloud computing’ section 2.3.2 below)

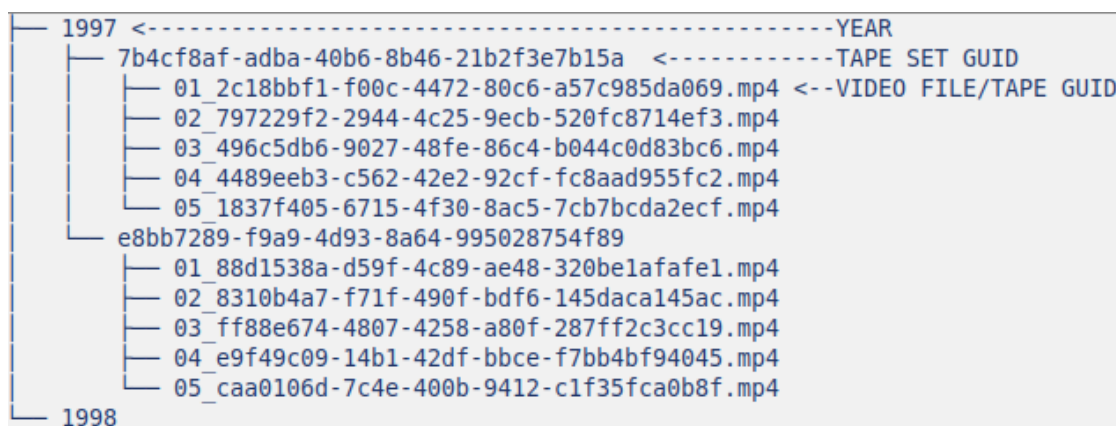


Figure 2.6: File directory structure of the Bossou archive. Each year contains directory named after 'Tape set' GUID (an identifier for the tapes filmed in sequence on the same camera). Within each Tape set directory are video files corresponding to the original tape GUIDs in sequential order.

; AWS 'Glacier' was used as an emergency backup (see section of future directions below) in case of local disaster, theft or corruption of video files (Figure 2.7).

To date 1312 manual tapes have been digitised: 406 were duplicate tapes (back ups or copies), and 904 represented 'original' content which can be analysed. The average number of hours per year is 31.5, and the average size of each video is 55 minutes. In the archive this comprises 1840 files, 11.44TB, totalling 1200 hours of unique content spanning 1988-2018.

2.3 Discussion

In this chapter I have outlined a workflow for organisation, digitisation and processing of a large video archive of chimpanzee behaviour for analysis. This resource is unique to primatology, and remains the longest and clearest record of chimpanzee stone tool use and field experiments; it provides a platform for research into ape tool use, culture and behaviour. Although the dataset is now functional for research projects, here I outline remaining tasks, future directions and the wider implications of the strategy for methods in primatology.

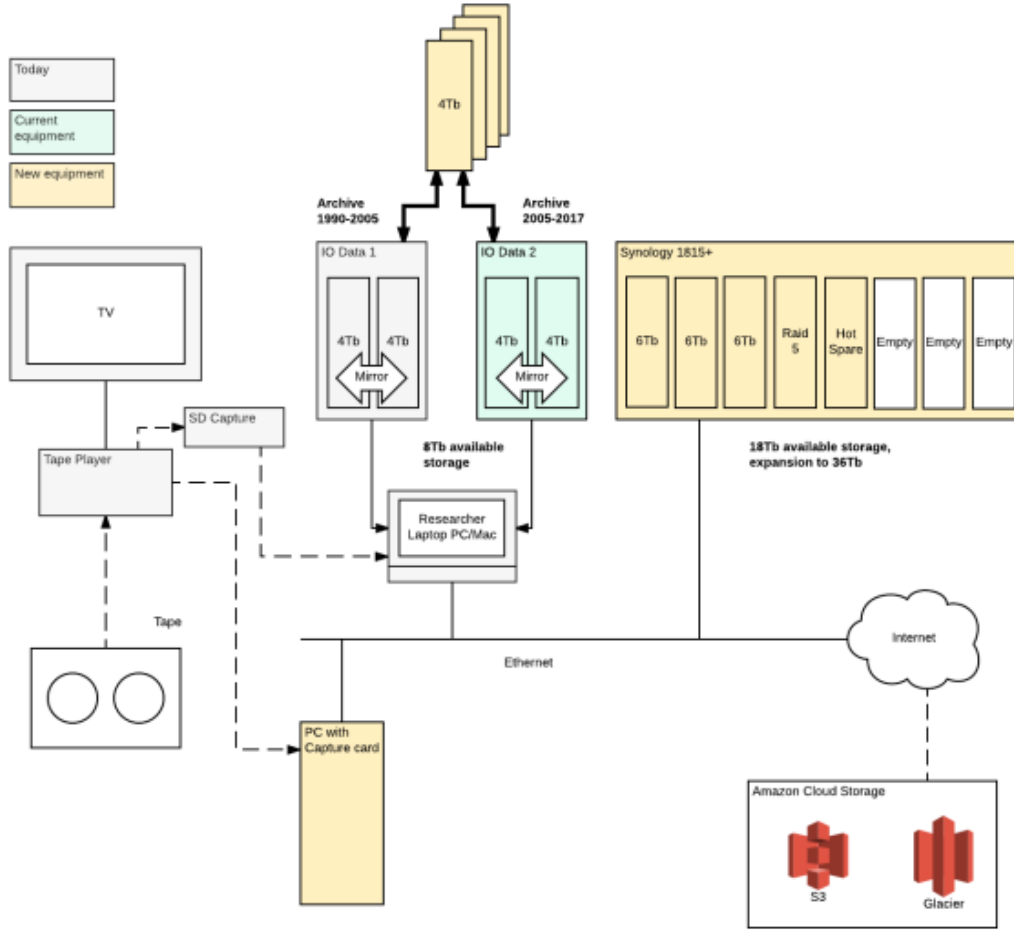


Figure 2.7: Schematic diagram of archive infrastructure developed in 2018.

2.3.1 Processing

A significant challenge remains to manually annotate the time and date information for videos which lack metadata. For tapes which retain this information (e.g digital or miniDV) time/date metadata needs to be extracted automatically using FFmpeg. Due to the nature of filming on analogue tapes, continuous video sequences often have sudden ‘jumps’ in time where the camera was switched on or off. Shot detection can be used to demarcate the points where the scene/time/date changes. Automated algorithms can be applied to automatically detect these points (e.g. PySceneDetect, Castellano 2020) - this was performed effectively on a sample of data (see Chapter 4 of this thesis), but processing for the entire archive dataset remains to be done.

Currently, the archive is in a spreadsheet database; transitioning to a unified standardised/structured relational database would help protect the metadata and improve the long-term preservation and accessibility of video files.

2.3.2 Cloud computing

Several ongoing projects are now utilising the digitised archive via USB hard drives for local video coding, and master copies of the dataset are stored on the network attached storage devices in Kyoto and Oxford. Technology companies such as Google and AWS designed ‘cloud computing’ services to maximise the efficiency and support logistics of their commercial systems; these include computer storage and processing which is remotely accessible online for consumers. In recent years, these services have increased the availability and accessibility of large computer power and storage at increasing cost effectiveness. As opposed to researchers owning and maintaining data centers locally, storing data in the cloud has several advantages, including protection from security threats and data loss, as well as improving the accessibility. This ultimately enables data to be accessed remotely from anywhere, facilitating secure and global collaborative access. Currently the Bossou archive is backed up using AWS ‘Glacier’ for low cost monthly storage (\$67 per month) supported by Kyoto University which offers protection, but with limited access for users, and thus plans are being developed for transition to a more accessible web-based system for researchers (see below).

2.3.3 Towards reproducibility in computational ethology

A significant goal for science is complete reproducibility - often videos cannot easily be shared due to their size or complexity, limiting collaboration, openness and validation of research. For video datasets in the wildlife sciences, there is great potential to transition to ‘cloud native’ online platforms. In the case of the Bossou archive, this could constitute a complete pipeline for video access, collaborative annotation, and model training/analysis, all within a cloud-based system online. All annotation and viewing of the video files would be through this portal, centralising

and coordinating annotation effort. Additional tools such as user authenticated access and database management (such as AWS ‘DynamoDB’) enable flexible and secure remote access and analysis by researchers. Furthermore, an annotation framework which utilises open source software (e.g VGG Image Annotator‘VIA’, Dutta et al. 2016) - see Chapters 4 and 5 of this thesis) can be developed for more specific tasks in ethology/behaviour research. The field of machine learning has made huge advances in parallel with massive expansion of accessible video datasets online. In primatology there is increased protectionism around data, as researchers and field assistants spend years often in hostile, inaccessible and high risk areas, making biological data extremely valuable compared to open source videos (e.g. on YouTube) which can be scraped online for training models. Moving to cloud native frameworks as described above would not only protect these valuable archives, but would also enable researchers to reproduce findings and collaborate efficiently and remotely online. Ultimately, transitioning to open video datasets whilst maintaining a set of controls to protect fair use of the data will impact substantially the scale and novelty of analyses that can be conducted in the behavioural and wildlife sciences. There is always a concern in academia that valuable research data collected at great cost and effort is exploited by other competing teams without consent or proper acknowledgment. This new approach models video ‘rights’ aligned with commercial media principles to protect content owners while advancing research in the field. The laborious ‘craft’ approach to video annotation is slow, costly and subjective, typically only a small fraction of available video footage can be analysed to support a project. Reproducibility is core to good science however with current video approaches it is economically infeasible to repeat analysis on a large scale or test assertions and difficult to calibrate and refine annotation skills. Adoption of this framework opens up the possibilities of automating the rerunning of analysis and comparing the annotation skills of practitioners to massively increase the sample size, accelerate the development of annotation experts to advance research.

3

Chimpanzee face recognition from videos in the wild using deep learning

Published in *Science Advances*

RESEARCH METHODS

Chimpanzee face recognition from videos in the wild using deep learning

Daniel Schofield^{1*†}, Arsha Nagrani^{2*†}, Andrew Zisserman², Misato Hayashi³, Tetsuro Matsuzawa³, Dora Biro⁴, Susana Carvalho^{1,5,6,7}

Video recording is now ubiquitous in the study of animal behavior, but its analysis on a large scale is prohibited by the time and resources needed to manually process large volumes of data. We present a deep convolutional neural network (CNN) approach that provides a fully automated pipeline for face detection, tracking, and recognition of wild chimpanzees from long-term video records. In a 14-year dataset yielding 10 million face images from 23 individuals over 50 hours of footage, we obtained an overall accuracy of 92.5% for identity recognition and 96.2% for sex recognition. Using the identified faces, we generated co-occurrence matrices to trace changes in the social network structure of an aging population. The tools we developed enable easy processing and annotation of video datasets, including those from other species. Such automated analysis unveils the future potential of large-scale longitudinal video archives to address fundamental questions in behavior and conservation.

INTRODUCTION

Video data have become indispensable in the retrospective analysis and monitoring of wild animal species' presence, abundance, distribution, and behavior (1, 2). The accumulation of decades' worth of large video databases and archives has immense potential for answering biological questions that require longitudinal data (3). However, exploiting video data is currently severely limited by the amount of human effort required to manually process it, as well as the training and expertise necessary to accurately code such information. Citizen science platforms have allowed large-scale processing of databases such as camera trap images (4); however, ad hoc volunteer coders working independently typically only tag at the species level and cannot solve tasks such as recognizing individual identities. Here, we provide a fully automated computational approach to data collection from animals using the latest advances in artificial intelligence to detect, track, and recognize individual chimpanzees (*Pan troglodytes verus*) from a longitudinal archive. Automating the process of individual identification could represent a step change in our use of large image databases from the wild to open up vast amounts of data available for ethologists to analyze behavior for research and conservation in the wildlife sciences.

Deep learning

The field of machine learning uses algorithms that enable computer systems to solve tasks without being hand programmed, relying instead on learning from examples. With increasing computational power and the availability of large datasets, "deep learning" techniques have been developed that have brought breakthroughs in a number of different fields, including speech recognition, natural language processing, and computer vision (5, 6). Deep learning involves training computational

models composed of multiple processing layers that learn representations of data with many levels of abstraction, enabling the performance of complex tasks. A particularly effective technique in computer vision is the training of deep convolutional neural network (henceforth CNN) architectures (6) to perform fine-grained recognition of different categories of objects and animals (7), including automated image and video processing techniques for facial recognition in humans (8, 9), outperforming humans in both speed and accuracy.

Advances in the field of computer vision have led to the realization among wildlife scientists of the potential of automated computational methods to monitor wildlife. In particular, the emerging field of animal biometrics has adopted computer vision models for pattern recognition to identify numerous species through phenotypic appearance (10–12). Similar methods, developed for individual recognition of human faces, have been applied to nonhuman primates, including lemurs (13), macaques (14), gorillas (15), and chimpanzees (16–19). A substantial hurdle is developing models that are robust enough to perform on highly challenging datasets, such as those with low resolution and poor visibility. Since most of these previous face recognition methods applied to primates are limited by the size of training datasets, they are mostly shallow methods (small in the number of trainable parameters, hence not using deep learning), using cropped images of frontal faces or datasets from the controlled conditions of animals in captivity. While these methods have made valuable contributions, they are not robust to the inevitable variation in lighting conditions, image quality, pose, occlusions, and motion blur that characterize "wild" footage. Generating datasets to train robust recognition models has, so far, been restricting progress, as manually cropping and labeling faces from images are time consuming and limit the applicability of these methods to scale. While obtaining such datasets is possible for human faces (9) (for example, from multimedia sources or crowdsourcing services such as Amazon Mechanical Turk), obtaining large, labeled datasets of, for example, nonhuman primate faces is an extremely labor-intensive task and can typically only be done by expert researchers who are experienced in recognizing the individuals in question. Here, we attempt to solve this problem by providing a set of tools and an automated framework to help researchers more efficiently annotate large datasets, using wild chimpanzees as a case study to illustrate its potential and suggest new avenues of research.

Copyright © 2019
The Authors, some
rights reserved;
exclusive licensee
American Association
for the Advancement
of Science. No claim to
original U.S. Government
Works. Distributed
under a Creative
Commons Attribution
NonCommercial
License 4.0 (CC BY-NC).

¹Primate Models for Behavioural Evolution Lab, Institute of Cognitive and Evolutionary Anthropology, University of Oxford, Oxford, UK. ²Visual Geometry Group, Department of Engineering Science, University of Oxford, Oxford, UK. ³Primate Research Institute, Kyoto University, Inuyama, Japan. ⁴Department of Zoology, University of Oxford, Oxford, UK. ⁵Gorongosa National Park, Sofala, Mozambique. ⁶Interdisciplinary Center for Archaeology and Evolution of Human Behaviour (ICAEHB), Universidade do Algarve, Faro, Portugal. ⁷Centre for Functional Ecology—Science for People & the Planet, Universidade de Coimbra, Coimbra, Portugal.

*Corresponding author. Email: daniel.schofield@anthro.ox.ac.uk (D.S.); arsha@robots.ox.ac.uk (A.N.)

†These authors contributed equally to this work.

RESULTS

We developed an automated pipeline that can individually identify and track wild apes in raw video footage and demonstrate its use on a dataset spanning 14 years of a longitudinal video archive of chimpanzees (*P. troglodytes verus*) from Bossou, Guinea (20). Data used were collected in the Bossou forest, southeastern Guinea, West Africa, a long-term chimpanzee field site established by Kyoto University in 1976 (21). Bossou is home to an “outdoor” laboratory: a natural forest clearing (7 m by 20 m) located in the core of the Bossou chimpanzees’ home range (07°39’N; 008°30’W) where raw materials for tool use—stones and nuts—are provisioned, and the same group has been recorded since 1988. The use of standardized video recording over many field seasons has led to the accumulation of over 30 years of video data, providing unique opportunities to analyze chimpanzee behavior over multiple generations (22). Our framework consists of detection and tracking of individuals through the video (localization in space and time) as well as sex and identity recognition (Fig. 1 and movie S1). Both the detection and tracking stage and the sex and identity recognition stage use a deep CNN model.

We applied this pipeline to ca. 50 hours of footage featuring 23 individuals, resulting in a total of 10 million face detections (Figs. 2 and 3) and more than 20,000 face tracks (see Fig. 1A and Materials and Methods). The training set for the face recognition model consisted of 15,274 face tracks taken from four different years (2000, 2004, 2008, and 2012) within the full dataset, belonging to 23 different chimpanzees of the Bossou community, ranging in estimated age from newborn to 57 years (table S1). A proportion of face tracks were held out to test the model’s performance in each year, as well as to provide an all-years overall accuracy (table S2). Our chimpanzee face detector achieved an average precision of 81% (fig. S1), and our recognition model performed well on extreme poses and profile faces

typical of videos recorded in the wild (Fig. 2B, table S3, and movie S1), achieving an overall recognition accuracy of 92.47% for identity and 96.16% for sex. We tested both frame-level accuracy, wherein our model is applied to detections in every frame to obtain predictions, and track-level accuracy, which averages the predictions for each face track. Using track-level labels compared with frame-level labels provided a large accuracy boost (table S3), demonstrating the superiority of our video-based method to frame-level approaches. We note that these results include faces from all viewpoints (frontal, profile, and extreme profile); if only frontal faces were used, then the identity recognition accuracy improves to 95.07% and the sex recognition accuracy to 97.36% (table S3).

Generalizability across years

To further investigate the generalizability of the model, we next tested it on footage from two additional years not used in our training: an interpolated year that fell within the period used in training (2006) and an extrapolated year (2013) that fell outside it. For the interpolated year, identity recognition accuracy was 91.81% and sex recognition accuracy was 95.35%; for the extrapolated year, 91.37 and 99.83%, respectively. These accuracies were obtained despite some individuals in our dataset undergoing significant changes in appearance with age, such as the maturation of infants to adults (Fig. 2C). This suggests that our system may exhibit some degree of robustness to age-related changes in our population. However, as our system has not been specifically designed for age invariance, future work should test how performance is affected by the duration of the gap between the training and test sets on a dataset featuring more individuals and spanning a greater number of years. As is often the case with multi-class classification, classification error is only very rarely uniformly distributed for all classes. To understand where the recognition

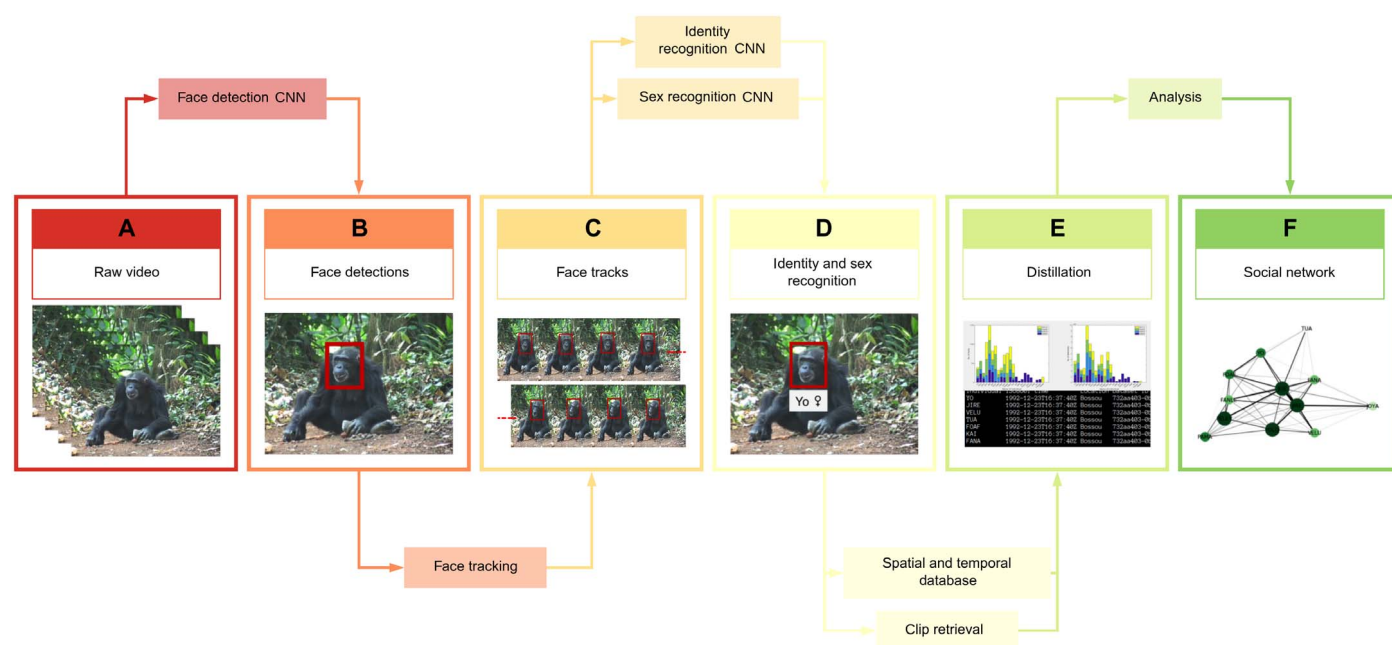


Fig. 1. Fully unified pipeline for wild chimpanzee face tracking and recognition from raw video footage. The pipeline consists of the following stages: (A) Frames are extracted from raw video. (B) Detection of faces is performed using a deep CNN single-shot detector (SSD) model. (C) Face tracking, which is implemented using a Kanade-Lucas-Tomasi (KLT) tracker (25) to group detections into face tracks. (D) Facial identity and sex recognition, which are achieved through the training of deep CNN models. (E) The system only requires the raw video as input and produces labeled face tracks and metadata as temporal and spatial information. (F) This output from the pipeline can then be used to support, for example, social network analysis. (Photo credit: Kyoto University, Primate Research Institute)

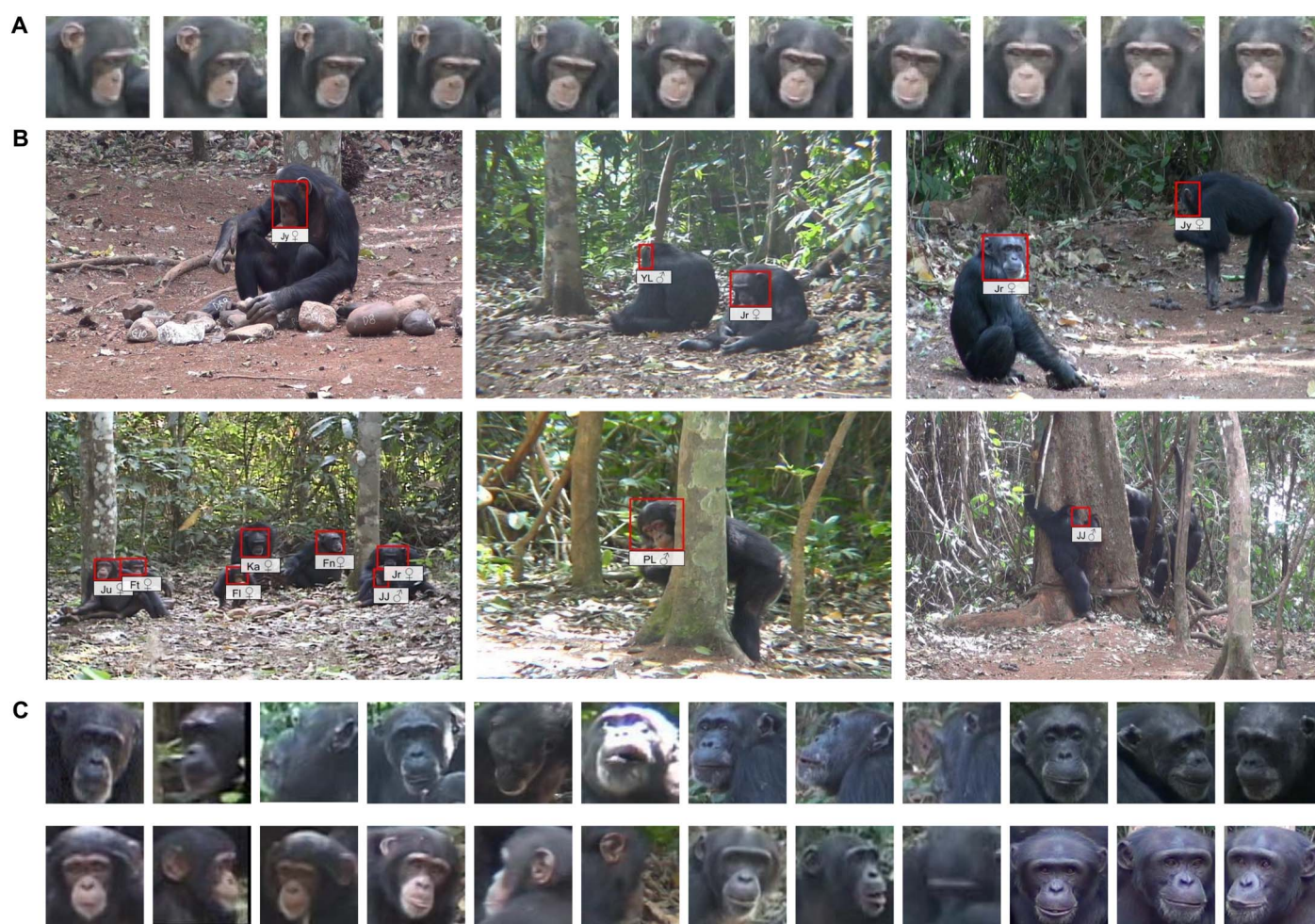


Fig. 2. Face recognition results demonstrating the CNN model's robustness to variations in pose, lighting, scale, and age over time. (A) Example of a correctly labeled face track. The first two faces (nonfrontal) were initially labeled incorrectly by the model but were corrected automatically by recognition of the other faces in the track, demonstrating the benefit of our face track aggregation approach. (B) Examples of chimpanzee face detections and recognition results in frames extracted from raw video. Note how the system has achieved invariance to scale and is able to perform identification despite extreme poses and occlusions from vegetation and other individuals. (C) Examples of correctly identified faces for two individuals. The individuals age 12 years from left to right (top row: from 41 to 53 years; bottom row: from 2 to 14 years). Note how the model can recognize extreme profiles, as well as faces with motion blur and lighting variations. (Photo credit: Kyoto University, Primate Research Institute)

model was erring, we created a confusion matrix of the individuals in the test set (table S4). We used frame-level predictions to assess the raw recognition power of the model (since track level labels are also affected by the relative length of tracks for different individuals). The model was more accurate at identifying certain individuals, and the lowest per class accuracies were for two infants in the dataset (Jy and FE).

Sex recognition in unknown individuals

Although the small number of individuals in our dataset is a limitation, we performed a preliminary study to test how well our sex recognition model generalizes to individuals outside of the training set. We randomly divided the corpus into 19 individuals (7 males and 12 females) for training and 4 individuals (2 males and 2 females) for testing. With this test split, we obtained a sex recognition accuracy of 87.4% (above chance performance, which is 50%) showing promise of the model's ability to generalize sex recognition to unseen individuals when using larger datasets. This is particularly relevant for longitudinal studies where, over time, new individuals are added

to populations via birth or immigration. Successful recognition of the sex of these previously unseen/unlabeled individuals would allow the automated tracking of natural demographic processes in wild populations. We further discuss how to extend our model to classify and add new individual identities in Materials and Methods.

Comparison to human performance

To test our model's performance against that of human observers, we conducted an experiment using expert and novice annotators. We selected 100 random face images from the test set and provided them to researchers and students with coding experience of Bossou chimpanzees to annotate using the VIA web browser interface (fig. S2) (17). Annotators only had access to cropped face images, the same input as the computer algorithm for this task. To assist annotators, each had access to a gallery of 50 images per individual from the training set, as well as a table with three examples of each individual for easy side-by-side comparisons (we show a screenshot of the interface in fig. S3). There were no time limits on the task. We classified human annotators

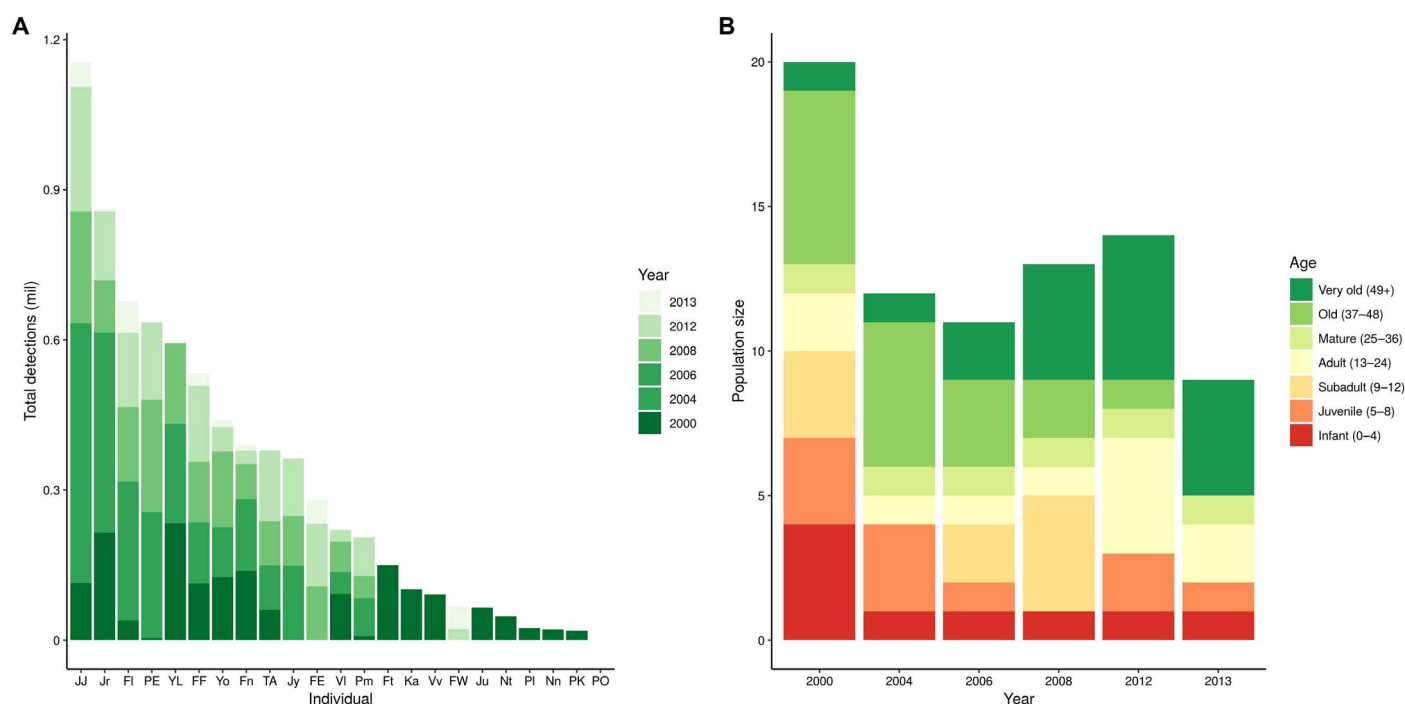


Fig. 3. Face detection and recognition results. (A) Histograms of detection numbers for individuals in the training and test years of the dataset (2000, 2004, 2006, 2008, 2012, and 2013). (B) Output of model for number of individuals detected in each year and proportion of individuals in different age categories based on existing estimates of individual ages.

into expert (prior experience with identifying 50% or more of the individuals in the dataset and over 50 hours of coding experience) and novice annotators with limited coding experience and familiarity with individual identities. On this frame-level identity classification task, expert human annotators ($n = 3$) scored an average of 42% ($42.00 \pm 24.33\%$), while novice annotators ($n = 3$) performed significantly worse ($20.67 \pm 11.24\%$), demonstrating the importance of familiarity with individuals for this task. It took experts an estimated 55 min and novices 130 min to complete the experiment. Our model achieved 84% [in 60 ms using a Titan X graphics processing unit (GPU) and 30 s on a standard central processing unit], outperforming even the expert human annotators not only in speed but also in accuracy. Further work should test a larger sample of human annotators and examine how additional cues and contextual information (e.g., full-video sequences or full-body images) affect performance.

Social network analysis

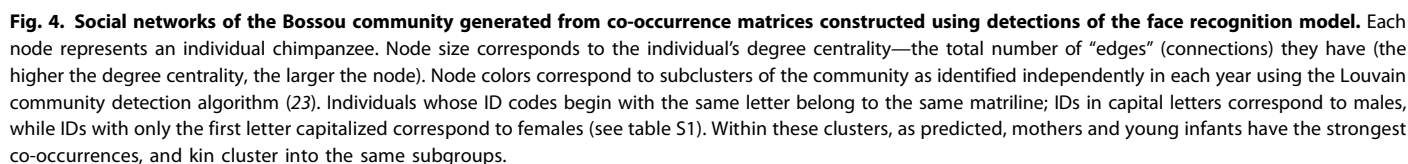
We used the output face detections from our pipeline to automatically generate adjacency matrices by recording co-occurrences of identified individuals in each video frame in our training dataset. Figure 4 shows the social networks sampled from four field seasons over 12 years for the Bossou community, from approximately 4 million co-occurrence events (see Materials and Methods). Subclusters of the community are visualized as defined using the Louvain community detection algorithm (23), using the density of connections within and between groups of nodes. These subclusters correctly identify mothers and young infants as those with the strongest co-occurrences, and kin cluster into the same subgroups. At the end of 2003, the size of the Bossou chimpanzee community declined drastically because of an epidemic, causing significant demographic changes and an increasingly aging population over subsequent years (24). By 2012, the isolation of some individuals from the

rest of the community becomes striking, with two of the oldest females (Yo and Velu, both over 50 years old) detected together in many of the videos but very rarely with the rest of the group.

DISCUSSION

Our model demonstrates the efficacy of using deep neural network architectures for a direct biological application: the detection, tracking, and recognition of individual animals in longitudinal video archives from the wild. Unlike previous automation attempts (17, 18), we operate on a very large scale, processing millions of faces. In turn, the scale of the dataset allows us to use state-of-the-art deep learning, avoiding the use of the older, less powerful classifiers. Our approach is also enriched by the use of a video-based, rather than frame-based, method, which improves accuracy by pooling multiple detections of the same individual before coming to a decision. We demonstrate that face recognition is possible on data at least 1 year beyond that supplied during the training phase, opening up the possibility of analyzing years that human coders may not have even seen themselves.

We do not constrain the video data in any way, as is done for other primate face recognition works [e.g., (13, 18)], by aligning face poses or selecting for age, resolution, or lighting. We do this to perform the task “in the wild” and ensure an end-to-end pipeline that will work on raw video with minimum preprocessing. Hence, the performance of our model is highly dependent on numerous factors, such as variation in image quality and pose. For example, model accuracy increases monotonically with image resolution (fig. S4), and testing only on frontals increases performance. On unconstrained faces, our model outperformed humans, highlighting the difficulty of the task. Humans’ poor performance is likely due to the specificity of the task: Normally, researchers who observe behavior in situ can rely on multiple additional



While our model was developed using a chimpanzee dataset, the extent of its generalizability to other species is an important question for its immediate value for research. We show some preliminary examples of our face detector (with no further modification) applied to other primate species in Fig. 5. Our detector, trained solely on chimpanzee faces, generalized well, and the tracking part of our pipeline is completely agnostic to the species to be tracked (25). Individual recognition will require a corpus annotated with identity labels; however, we release all software open source such that researchers can produce their own training sets using our automated framework. Such corpus may not have to be as large as the corpus that we use in this study; in supervised machine learning, features learned on large datasets are often directly useful in similar tasks, even those that are data poor. For instance, in the visual domain, features learnt on ImageNet (26) are routinely used as input representations in other computer vision tasks

The ultimate goal for using computational frameworks in wildlife science is to move beyond the use of visual images for the monitoring and censusing of populations to automated analyses of behaviors, quantifying social interactions and group dynamics. For example, sampling the sheer quantity of wild animals' complex social interactions for social network analysis typically represents a daunting methodological challenge (28). The use of animal-borne biologgers and passive transponders has automated data collection at high resolution for numerous species (29), but these technologies require capturing subjects, are expensive and labor intensive to install and maintain, their application may be location specific (e.g., depends on animals approaching a receiver in a fixed location), and the data recorded typically lack contextual visual information.

We show that by using our face detector, tracker, and recognition pipeline, we are able to automate the sampling of social networks over

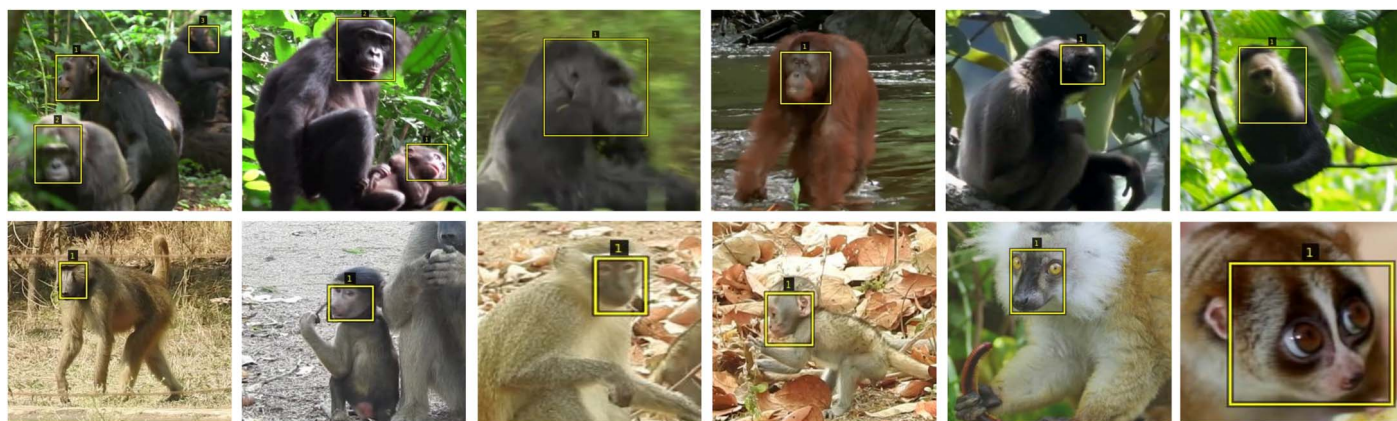


Fig. 5. Preliminary results from the face detector model tested on other primate species. Top row: *P. troglodytes schweinfurthii*, *Pan paniscus*, *Gorilla beringei*, *Pongo pygmaeus*, *Hylobates muelleri*, and *Cebus imitator*. Bottom row: *Papio ursinus* (x2), *Chlorocebus pygerythrus* (x2), *Eulemur macaco*, and *Nycticebus coucang*. Image sources: Chimpanzee: www.youtube.com/watch?v=c2u3NKXbGeo; Bonobo: www.youtube.com/watch?v=JF8v_HWvLc&t=9s; Gorilla: www.youtube.com/watch?v=wDECqJsiGqw&t=28s; Orangutan: www.youtube.com/watch?v=Gj2W5BHu-SI; Gibbon: www.youtube.com/watch?v=C6HucLWksVc; Capuchin: Lynn Lewis-Bevan (personal data); Baboon: Lucy Baehren (personal data); Vervet monkey: Lucy Baehren (personal data); Loris: www.youtube.com/watch?v=2Syd_BUbl5A&t=2s.

multiple years, providing high-resolution output on the spatio-temporal occurrence and co-occurrence of specific group members. This automated pipeline can aid conservation and behavioral analyses, allowing us to retrospectively analyze key events in the history of a wild community, for example, by quantifying how the decrease in population size and loss of key individuals in the community affect the network structure, with a decrease in the connectivity and average degree of the network (Fig. 4 and table S5). Traditional ethology has been reliant on human observation, but adopting a deep learning approach for the automation of individual recognition and tracking will improve the speed and amount of data processed and introduce a set of quantifiable algorithms with the potential to standardize behavioral analysis and, thus, allow for reproducibility across different studies (30, 31).

We have demonstrated that the current generation of deep architectures trained using annotations on the Bossou dataset can cope with the relatively unconstrained conditions in videos of chimpanzees in the forest outdoor laboratory and should translate well to the low-resolution and variable conditions that typify camera trap or archive footage from the wild. We use both a VGG-M architecture and a ResNet-50 architecture and find that all else kept constant; the performance is comparable (even though the ResNet architecture overfits more). A larger dataset could be required to take advantage of deeper CNNs [see (32) for trade-offs between deep models and dataset size].

A key driver for the advancement of the use of artificial intelligence systems for wildlife research and conservation will be the increasing availability of open-source datasets for multiple species. As more models in this domain are developed, future work should examine how multiple variables such as features of the training dataset, different neural network architectures, and benchmarks affect performance [see (33) for review and benchmarks of existing animal deep learning datasets]. Ultimately, this will help maximize the adoption and application of these systems to solve a wide range of different problems and allow researchers to gauge properties of the data for which these models should perform well.

There are some limitations to our study, notably the size of our dataset (in terms of individuals), which consisted of only 23 chimpanzees. The small population size of Bossou chimpanzees and their

lack of genetic admixture with neighboring groups (34) could indicate that our dataset is, likely, less phenotypically diverse than other chimpanzee groups. We note that the model found some individuals more distinctive than others, and errors in recognition (table S4) were likely due to facial similarity between closely related individuals within the population. We expect that as with human face recognition models (9), performance will increase as more individuals are added, and populations are combined from multiple field sites. Another issue is that with our tracking pipeline, individuals are not tracked if the head is completely turned or obscured, which may bias social network analysis based on co-occurrences. In relation to that, given that our network performs the task in isolation, further performance improvements could be achieved by incorporating contextual detail: For example, the identities of individuals in proximity may provide important information, as is the case with mother-infant pairs. Another direction is to move beyond faces, as for some species the face may not be the most discriminative part. Instead, whole-body detectors can be trained to discriminate bodies in much the same way as they are trained in this paper to discriminate faces. With these potential future improvements in mind, we hope that our automated pipeline and tools used for annotation will facilitate other research and generate larger datasets to improve accuracy and generalizability and drive the development of a multitude of new tools beyond recognizing faces, such as full-body tracking and behavioral recognition.

MATERIALS AND METHODS

Structure of the dataset

The dataset used to train our deep CNN features 23 wild chimpanzees (14 females and 9 males) at the long-term field site of Bossou, Guinea, recorded between 2000 and 2012. In the start year of our dataset (2000), there were 20 individuals: 10 adults (13+ years), 3 subadults (9 to 12 years), 3 juveniles (5 to 8 years), and 4 infants (0 to 4 years). Eight of these individuals had been present in the community since the Bossou field site was established in 1976, and hence, their ages are estimates only (see table S1). In subsequent years, three infants were born, and the overall population size decreased as individuals disappeared or died (see table S1 and Fig. 3B).

A random sample of videos totaling almost 50 hours were extracted from six field seasons (2000, 2004, 2006, 2008, 2012, and 2013) and used to train and test our CNN model (note that 2006 and 2013 were only used for testing; see below for details). All videos were taken at the outdoor laboratory with no preprocessing (movie S1). The videos were recorded at different times of the day and span a range of lighting conditions. Since the footage was shot in a natural forest clearing, often there was heavy occlusion of faces from vegetation. The individuals moved around and interacted freely with one another, and the camera panned and zoomed; hence, faces in the videos had large variations in size, with small faces for chimpanzees in the background, motion blur, and occlusion due to other individuals. Often faces appeared as extreme profiles (in some cases, only a single ear was visible), and there were visible changes to facial features as individuals age over time.

Model training

Our pipeline consists of two main components: (i) detection and tracking and (ii) facial recognition (Fig. 1). First, we trained a deep single-shot detector (SSD) model (35) with a VGG-16–based CNN architecture (36) to automatically localize chimpanzee faces in the raw footage (see fig. S1 for evaluation of the detector and a detailed description of this in the section on Face detection). We then implemented face tracking using a version of the Kanade-Lucas-Tomasi (KLT) tracking method to group faces belonging to the same individual across frames into a “face track” as a single unit (Fig. 2A). For the second stage, these face tracks were tagged with chimpanzee identities by a human coder using a custom-built, lightweight web-based annotation tool (37), thus creating a training dataset for a deep CNN recognition model (fig. S2). For both identity and sex recognition, the base network architecture used was a variant of the VGG-M architecture introduced by Chatfield *et al.* (38). The network was trained for the task of classification by minimizing the softmax log loss (cross-entropy) over 25 classes for identity recognition (23 individual identity classes corresponding to the number of individuals in the dataset, 1 class for negatives, and 1 class for false-positive tracks; defined in the section on Face recognition) and over 2 classes for the task of sex recognition (male and female). We provide a detailed description for each stage of the pipeline in the following sections.

Face detection

Our detection pipeline consisted of the following steps, allowing us to obtain chimpanzee face detections in every frame of raw video:

1) To train the detector, 3707 video frames were annotated with detection boxes. These frames were extracted every 10 s from a 2008 video. Annotation involved drawing a tight bounding box around each head using our web-based VIA annotation interface (29). This resulted in 5570 detection boxes. The statistics of the dataset can be seen in table S2.

2) These annotations were then used to train a chimpanzee face detector with two classes: background and chimpanzee face.

3) The detector was then run over all the frames of the video (extracted at 25 fps), giving us face detections in every frame.

Evaluation protocol for face detector. Evaluation was performed on a held-out test set using the standard protocol outlined by Everingham *et al.* (39). The precision/recall curve was computed from a method’s ranked output. Recall was defined as the proportion of all positive examples ranked above a given rank, while precision is the proportion of all examples above that rank which are from the positive class. For the purpose of our task, high recall was more important than high precision (i.e., false positives are less dangerous

than false negatives) to ensure no chimpanzee face detections were missed. Some false positives, such as the recognition of chimpanzee behinds as faces (e.g., fig. S1C), were automatically discarded by the tracking procedure (see the Face tracking section).

Programming implementation details. The detector was implemented using the machine learning library MatConvNet (40). The SSD detector was trained on two Titan X GPUs for 50 epochs (where 1 epoch consists of an entire pass through the training set) using a batch size of 32 and 2 sub-batches. Flip, zoom, path, and distort augmentation was used during preprocessing with a zoom factor of 4. The ratio of negatives to positives while training was 3, and the overlap threshold was 0.5. The detector was trained without batch normalization.

Face tracking

Our tracking pipeline consisted of the following steps, allowing us to obtain face tracks from raw video:

1) Shot detection: Each video was divided into shots, which are continuous segments of footage. A single face track does not exist over multiple shots.

2) Tracking: Face detections within each shot were then grouped into face tracks. The KLT tracker was implemented in MATLAB and optimized for speed and complexity constraints. This specific tracking model achieved temporally smooth tracking of animal faces.

3) Post processing: False positives are rarely tracked over multiple frames, so tracks shorter than 10 frames were discarded as false positives, and detections within a track were smoothed.

Face recognition

Our recognition pipeline consisted of the following steps:

1) In the training stage, face tracks were labeled with the identity of the chimpanzee. This involved selecting the label from a menu for each face track in the training video. We then created a simple web-based annotation interface enabling a single human annotator to label face tracks with their identity; this enabled millions of chimpanzee faces to be labeled with only a few hours of manual effort. The first, middle, and last frames in a track are displayed, and a single label was applied to each track. A screenshot is provided in fig. S2.

2) This annotation was then used to train a chimpanzee face recognizer. Recognition involved automatically assigning every face track to an identity and sex label. This stage was performed first at a frame level, and results were then aggregated to obtain a single label per track. This aggregation was implemented by averaging the pre-softmax logit predictions for every frame in a single track. For identity recognition, the goal was to classify the faces into 23 individual identity classes, as well as 1 additional class for false-positive tracks, which are “nonface” tracks mistakenly detected and tracked. This ensured an end-to-end system with no manual interference, i.e., false positives did not need to be discarded manually. Since the number of individuals in our dataset was small compared with human face datasets (typically in the thousands), we also added an additional class of “negatives,” which consisted of individuals outside of our dataset. Since it was not important to label these individuals, we obtained these faces from other sources such as films. The only purpose that these negatives served was to make the identity classifier more discriminative for the individuals of interest: 1.9 hours of videos of an estimated ~100 unrelated chimpanzees (including all chimpanzee subspecies) were used as negatives, from the DVDs *Chimpanzee* (Disney, 2012) and *The New Chimpanzees* (National Geographic, 1995), and 6759 faces were extracted using our detector and tracking model and labeled “face” or “nonface.”

For identity recognition, we trained the network by minimizing the softmax log loss (cross-entropy) over 25 classes

$$L = - \sum_{n=1}^N \left(y_c(x_n) - \log \sum_{j=1}^{C_t} e^{y_j(x_n)} \right)$$

where x_n is a single face input to the network, y_j is the pre-softmax activation for class j of size C_t , and c is the true class of x_n . Since the classes were heavily unbalanced (the number of training examples for each individual varies greatly; see Fig. 3A), the loss was weighted according to the frequency distribution of the classes. These weights for each class c are specified as

$$a_c = \frac{n_{\text{total}} - n_c}{n_c}$$

where n_{total} is the total number of training examples, and n_c is the number of training examples for class c . The weighted loss can then be expressed as

$$L_w = \sum_c a_c l(c)$$

where $l(c)$ is the component of the loss for class c .

For sex classification, the network was trained using a binary cross entropy loss over the two classes: male and female. By using the entire training set, the sex classification model was trained to achieve invariance to identity and age.

With the exception of (18), previous works on primate face recognition (13–16, 19) required face cropping and then pose normalization using manually annotated landmarks. We applied no such pose normalization. The input to the face subnetwork was an RGB image, automatically cropped from the source frame using the detector to include only the face region. We extended the bounding boxes provided by the detector by 20% to include more context and resized the image to 224×224 . Standard aggressive on-the-fly augmentation strategies were used during training (41), including random horizontal flipping, brightness, and saturation jittering, but we did not extract random crops from within the face region.

3) The recognizer is then applied to all face tracks, assigning a label (identity or nonface) to each of them.

We can run this on as many videos as required. At this stage, we have labeled face tracks for all the chimpanzees in the video. Note that steps 1) and 2) only need to be done once (the training stage). For new videos, only step 3) has to be run.

Extending to new individuals. Learning new IDs does not require complex surgery of the network or a full retraining from scratch. We demonstrated this by training on our identity disjoint train split of only 19 individuals (see Results). We observed a drop in identity recognition performance using only these individuals. We then added in the remaining four individuals by simply modifying the final layer to include hidden units for each of these four individuals and training on them. Doing this, we achieved a similar performance as we reported before (79.1% frame-level accuracy) but requiring less training time (only 10 epochs, compared with 50 epochs when training from scratch) and, hence, one-fifth the computational resources. The only change required to the architecture was the number of units in the final layer.

Programming implementation details. The networks for recognition were trained on three Titan X GPUs for 50 epochs using a batch size of 256. We trained both models end to end via stochastic gradient descent with momentum (0.9) weight decay (5×10^{-4}) and a logarithmically decaying learning rate (initialized to 10^{-2} and decaying to 10^{-8}). Both models were trained from scratch with no pretraining.

Social network analysis

We sampled networks at a high resolution using every frame to record co-occurrences between intervals. We started by defining co-occurrence $\alpha_{ij}(\delta t)$ of individuals i and j as the number of times they were recorded to co-occur in a given time interval δt . This was then normalized by the total number of times individual i is observed in this time period with another individual

$$N_i = \sum_j \alpha_{ij}(\delta t)$$

and the total number of times individual j is observed with another individual

$$N_j = \sum_i \alpha_{ij}(\delta t)$$

This gives us the final co-occurrence of individuals i and j in time interval δt to be

$$\alpha_{\text{norm}}(i,j)(\delta t) = \frac{\alpha_{ij}(\delta t)}{\alpha_{ij}(\delta t) + (N_i - \alpha_{ij}(\delta t)) + (N_j - \alpha_{ij}(\delta t))}$$

Since the camera pans and zooms, frames that contained only a single individual (due to zooming) were ignored. Networks were visualized, and network metrics (table S5) were extracted using the Python library NetworkX.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at <http://advances.sciencemag.org/cgi/content/full/5/9/eaaw0736/DC1>

Fig. S1. Face detector results.

Fig. S2. Screenshots of the web-based annotation interfaces.

Fig. S3. Screenshots from the web-based experiment testing human annotator performance at identifying individual chimpanzees in cropped images.

Fig. S4. Frame-level accuracy of model with variation in chimpanzee face resolution.

Table S1. Name, ID code, sex, age, and years present for every chimpanzee at Bossou within the dataset analyzed.

Table S2. Summary statistics of training and testing datasets for recognition model.

Table S3. Identity and sex recognition results for accuracy on all faces and frontal faces only in the test set.

Table S4. Confusion matrix for the 13 individuals in the test set.

Table S5. Metrics of Bossou social networks derived from co-occurrences of detected individuals in video frames.

Movie S1. Video demo of automated identity and sex recognition of wild chimpanzees at Bossou, achieved through our deep learning pipeline.

REFERENCES AND NOTES

1. A. Caravaggi, P. B. Banks, A. C. Burton, C. M. V. Finlay, P. M. Haswell, M. W. Hayward, M. J. Rowcliffe, M. D. Wood, A review of camera trapping for conservation behaviour research. *Remote Sens. Ecol. Conserv.* **3**, 109–122 (2017).
2. T. Nishida, K. Zamma, T. Matsusaka, A. Inaba, W. C. McGrew, *Chimpanzee Behavior in the Wild: An Audio-Visual Encyclopedia* (Springer Science & Business Media, 2010).
3. T. Clutton-Brock, B. C. Sheldon, Individuals and populations: The role of long-term, individual-based studies of animals in ecology and evolutionary biology. *Trends Ecol. Evol.* **25**, 562–573 (2010).

4. A. Swanson, M. Kosmala, C. Lintott, R. Simpson, A. Smith, C. Packer, Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Sci. Data* **2**, 150026 (2015).
5. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press, 2016), vol. 1.
6. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436–444 (2015).
7. L. Yang, P. Luo, C. Change Loy, X. Tang, A large-scale car dataset for fine-grained categorization and verification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2015), pp. 3973–3981.
8. O. M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in *Proceedings of the British Machine Vision Conference* (BMVC, 2015), vol. 1, p. 6.
9. Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman, Vggface2: A dataset for recognising faces across pose and age, in *Automatic Face & Gesture Recognition* (2018).
10. H. S. Kuhl, T. Burghardt, Animal biometrics: Quantifying and detecting phenotypic appearance. *Trends Ecol. Evol.* **28**, 432–441 (2013).
11. A. G. Villa, A. Salazar, F. Vargas, Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks. *Ecol. Informatics* **41**, 24–32 (2017).
12. M. S. Norouzzadeh, A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, J. Clune, Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E5716–E5725 (2018).
13. D. Crouse, Z. Richardson, S. Klum, A. Jain, A. L. Baden, S. R. Tecot, LemurFacID: A face recognition system to facilitate individual identification of lemurs. *BMC Zool.* **2**, 2 (2017).
14. C. L. Witham, Automated face recognition of rhesus macaques. *J. Neurosci. Methods* **300**, 157–165 (2018).
15. C. A. Brust, T. Burghardt, M. Groenenberg, C. Kading, H. S. Kuhl, M. L. Manguette, J. Denzler, Towards Automated Visual Monitoring of Individual Gorillas in the Wild, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2017), pp. 2820–2830.
16. D. Deb, S. Wiper, A. Russo, S. Gong, Y. Shi, C. Tymoszek, A. Jain, Face recognition: Primates in the wild. arXiv:1804.08790 [cs.CV] (24 April 2018).
17. A. Freytag, E. Rodner, M. Simon, A. Loos, H. S. Kuhl, J. Denzler, Chimpanzee faces in the wild: Log-euclidean CNNs for predicting identities and attributes of primates, in *German Conference on Pattern Recognition* (2016), pp. 51–63.
18. A. Loos, A. Ernst, An automated chimpanzee identification system using face detection and recognition. *EURASIP J. Imag. Vid. Process.* **2013**, 1–17 (2013).
19. A. Loos, T. A. M. Kalyanasundaram, Face recognition for great apes: Identification of primates in videos, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, 2015), pp. 1548–1552.
20. T. Matsuzawa, T. Humle, Y. Sugiyama, *The Chimpanzees of Bossou and Nimba* (Springer, 2011).
21. Y. Sugiyama, Population dynamics of wild chimpanzees at Bossou, Guinea, between 1976 and 1983. *Primates* **25**, 391–400 (1984).
22. T. Matsuzawa, Field experiments of Tool-Use, in *The Chimpanzees of Bossou and Nimba*, M. Matsuzawa, T. Humle, Y. Sugiyama, Eds. (Springer, 2011), pp. 157–164.
23. V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
24. T. Matsuzawa, T. Humle, K. Koops, D. Biro, M. Hayashi, C. Sousa, Y. Mizuno, A. Kato, G. Yamakoshi, G. Ohashi, Y. Sugiyama, M. Kourouma, Wild chimpanzees at Bossou-Nimba: Deaths through a flu-like epidemic in 2003 and the green-corridor project. *Primate Res.* **20**, 45–55 (2004).
25. C. Tomasi, T. Kanade, *Detection and Tracking of Point Features* (Tech. Rep. CMU-CS-91-132, Carnegie Mellon Univ., 1991).
26. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)* (IEEE, 2009), pp. 248–255.
27. J. Yosinski, J. Clune, Y. Bengio, H. Lipson, in *Advances in Neural Information Processing Systems* (2014), pp. 3320–3328.
28. D. R. Farine, H. Whitehead, Constructing, conducting and interpreting animal social network analysis. *J. Anim. Ecol.* **84**, 1144–1163 (2015).
29. J. Krause, S. Krause, R. Arlinghaus, I. Psorakis, S. Roberts, C. Rutz, Reality mining of animal social systems. *Trends Ecol. Evol.* **28**, 541–551 (2013).
30. D. J. Anderson, P. Perona, Toward a science of computational ethology. *Neuron* **84**, 18–31 (2014).
31. A. Pérez-Escudero, J. Vicente-Page, R. C. Hinz, S. Arganda, G. G. De Polavieja, idTracker: Tracking individuals in a group by automatic identification of unmarked animals. *Nat. Methods* **11**, 743–748 (2014).
32. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 770–778.
33. S. Schneider, G. W. Taylor, S. Linquist, S. C. Kremer, Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods Ecol. Evol.* **10**, 461–470 (2019).
34. Y. Sugiyama, Demographic parameters and life history of chimpanzees at Bossou, Guinea. *Am. J. Phys. Anthropol.* **124**, 154–165 (2004).
35. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in *European Conference on Computer Vision* (2016), pp. 21–37.
36. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in *International Conference on Learning Representations* (2015).
37. A. Dutta, A. Gupta, A. Zisserman, VGG Image Annotator (VIA) (2016); www.robots.ox.ac.uk/~vgg/software/via.
38. K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: Delving deep into convolutional nets, in *Proceedings of the British Machine Vision Conference* (BMVC, 2014).
39. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **88**, 303–338 (2010).
40. A. Vedaldi, K. Lenc, MatConvNet: Convolutional neural networks for MATLAB, in *Proceedings of the 23rd ACM International Conference on Multimedia* (ACM, 2015), pp. 689–692.
41. A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems* (NIPS, 2012), pp. 1097–1105.

Acknowledgments: We are grateful to Kyoto University's Primate Research Institute for leading the Bossou Archive Project and supporting the research presented here, and to IREB and DNRST of the Republic of Guinea. This study marks the 30th anniversary of the opening of Bossou's outdoor laboratory and is dedicated to the researchers who have collected data there since 1988. Video data of wild chimpanzees collected in Guinea was noninvasive and approved by the ethics committee at the Primate Research Institute, Kyoto University. All research conducted conforms to the ethics guidelines set by the Association for the Study of Animal Behaviour. **Funding:** This work is supported by the EPSRC program grant Seebibyte: Visual Search for the Era of Big Data (EP/M013774/1), and the Cooperative Research Program of Primate Research Institute, Kyoto University. A.N. is funded by a Google PhD fellowship in machine perception, speech technology, and computer vision. D.S. is funded by the Clarendon Fund, Boise Trust Fund, and Wolfson College, University of Oxford. S.C. is funded by the Leverhulme Trust (PLP-2016-114). T.M. is funded by MEXT-JSPS (#16H06283) and LGP-U04, as well as the Japan Society for the Promotion of Science (JSPS) Core-to-Core Program CCSN. **Author contributions:** D.S., A.N., and A.Z. conceived the research. D.S. and A.N. analyzed the data and prepared the manuscript. D.S. provided annotations for training, and A.N. developed the software. A.Z., D.B., and S.C. provided project oversight, supervised the research, and prepared the manuscript. T.M. and M.H. reviewed the manuscript and supervised the research conducted in Japan and Guinea. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 15 November 2018

Accepted 2 August 2019

Published 4 September 2019

10.1126/sciadv.aaw0736

Citation: D. Schofield, A. Nagrani, A. Zisserman, M. Hayashi, T. Matsuzawa, D. Biro, S. Carvalho, Chimpanzee face recognition from videos in the wild using deep learning. *Sci. Adv.* **5**, eaaw0736 (2019).

Chimpanzee face recognition from videos in the wild using deep learning

Daniel SchofieldArsha NagraniAndrew ZissermanMisato HayashiTetsuro MatsuzawaDora BiroSusana Carvalho

Sci. Adv., 5 (9), eaaw0736. • DOI: 10.1126/sciadv.aaw0736

View the article online

<https://www.science.org/doi/10.1126/sciadv.aaw0736>

Permissions

<https://www.science.org/help/reprints-and-permissions>

Use of think article is subject to the [Terms of service](#)

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS. Copyright © 2019 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

Supplementary Materials for

Chimpanzee face recognition from videos in the wild using deep learning

Daniel Schofield*, Arsha Nagrani*, Andrew Zisserman, Misato Hayashi, Tetsuro Matsuzawa, Dora Biro, Susana Carvalho

*Corresponding author. Email: daniel.schofield@anthro.ox.ac.uk (D.S.); arsha@robots.ox.ac.uk (A.N.)

Published 4 September 2019, *Sci. Adv.* **5**, eaaw0736 (2019)
DOI: 10.1126/sciadv.aaw0736

The PDF file includes:

Fig. S1. Face detector results.

Fig. S2. Screenshots of the web-based annotation interfaces.

Fig. S3. Screenshots from the web-based experiment testing human annotator performance at identifying individual chimpanzees in cropped images.

Fig. S4. Frame-level accuracy of model with variation in chimpanzee face resolution.

Table S1. Name, ID code, sex, age, and years present for every chimpanzee at Bossou within the dataset analyzed.

Table S2. Summary statistics of training and testing datasets for recognition model.

Table S3. Identity and sex recognition results for accuracy on all faces and frontal faces only in the test set.

Table S4. Confusion matrix for the 13 individuals in the test set.

Table S5. Metrics of Bossou social networks derived from co-occurrences of detected individuals in video frames.

Legend for movie S1

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/5/9/eaaw0736/DC1)

Movie S1 (.mp4 format). Video demo of automated identity and sex recognition of wild chimpanzees at Bossou, achieved through our deep learning pipeline.

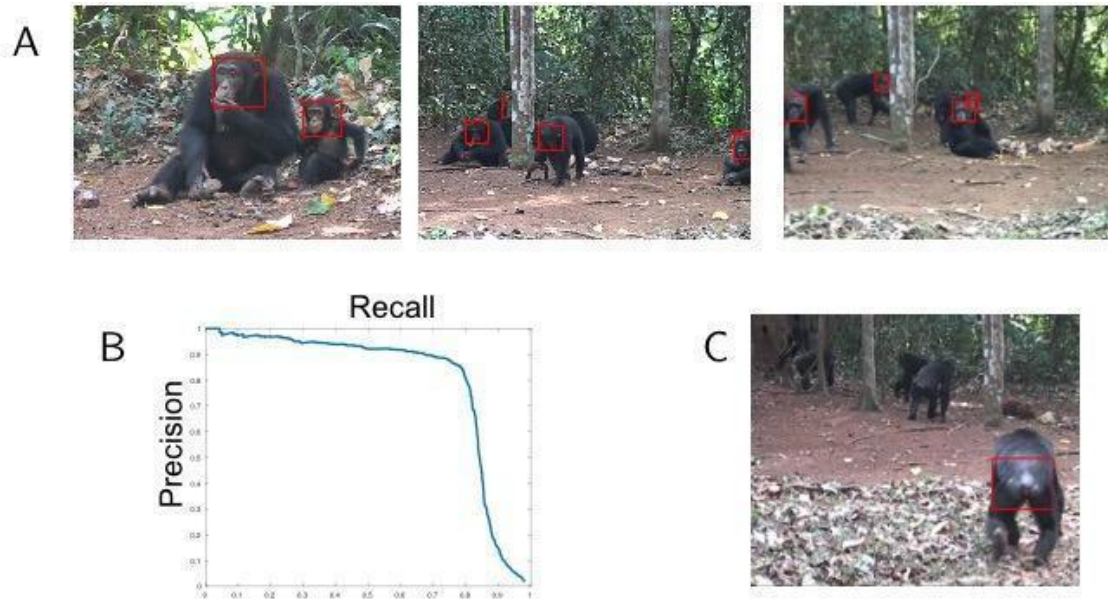


Fig. S1. Face detector results. A. Examples of chimpanzee face detections: the detector is trained to achieve invariance to scale, extreme poses and occlusions due to branches and other individuals. B. The precision-recall curve for the detector. The AP, summarising the shape of this curve, is 81%. C. An example of a false positive detection. Most false positive detections are of chimpanzee behinds. (Photo credit: Kyoto University, Primate Research Institute)

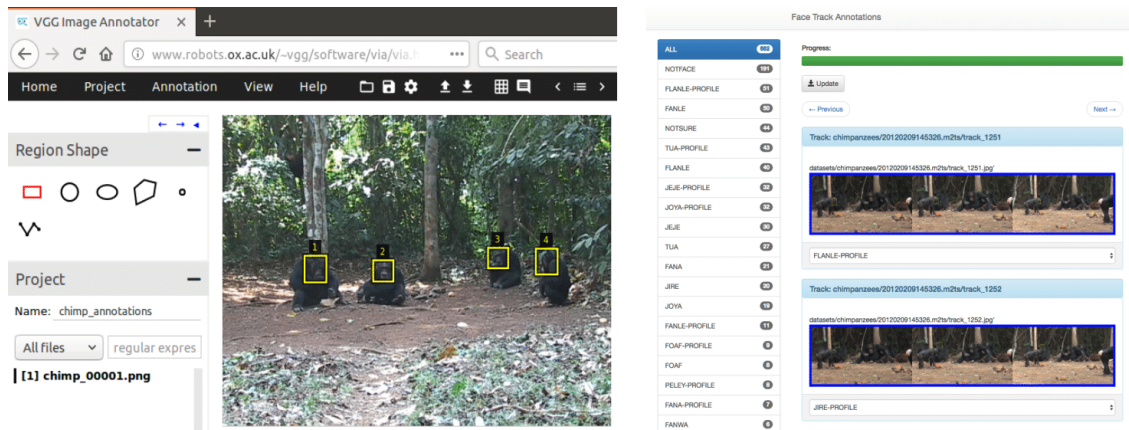


Fig. S2. Screenshots of the web-based annotation interfaces. (Left) Interface for the detector and (Right) interface for the face-tracks. The task for detection is to draw a tight bounding box around each chimpanzee face. The task for the face-tracks is to provide an identity label for each track, and to label the track as frontal or profile. (Photo credit: Kyoto University, Primate Research Institute)

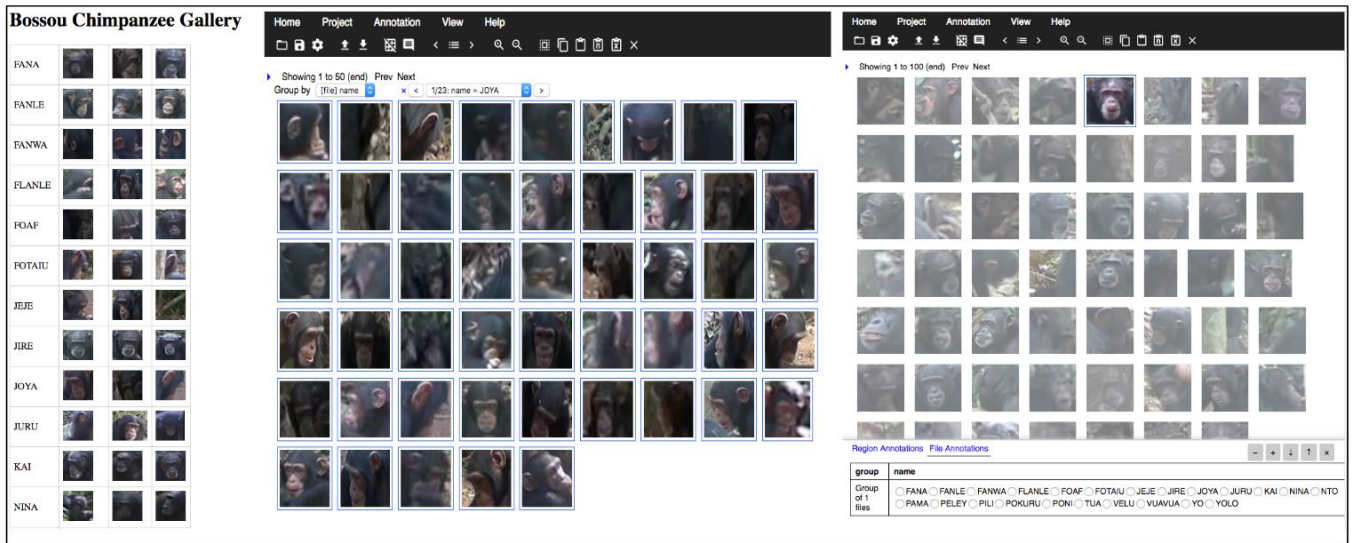
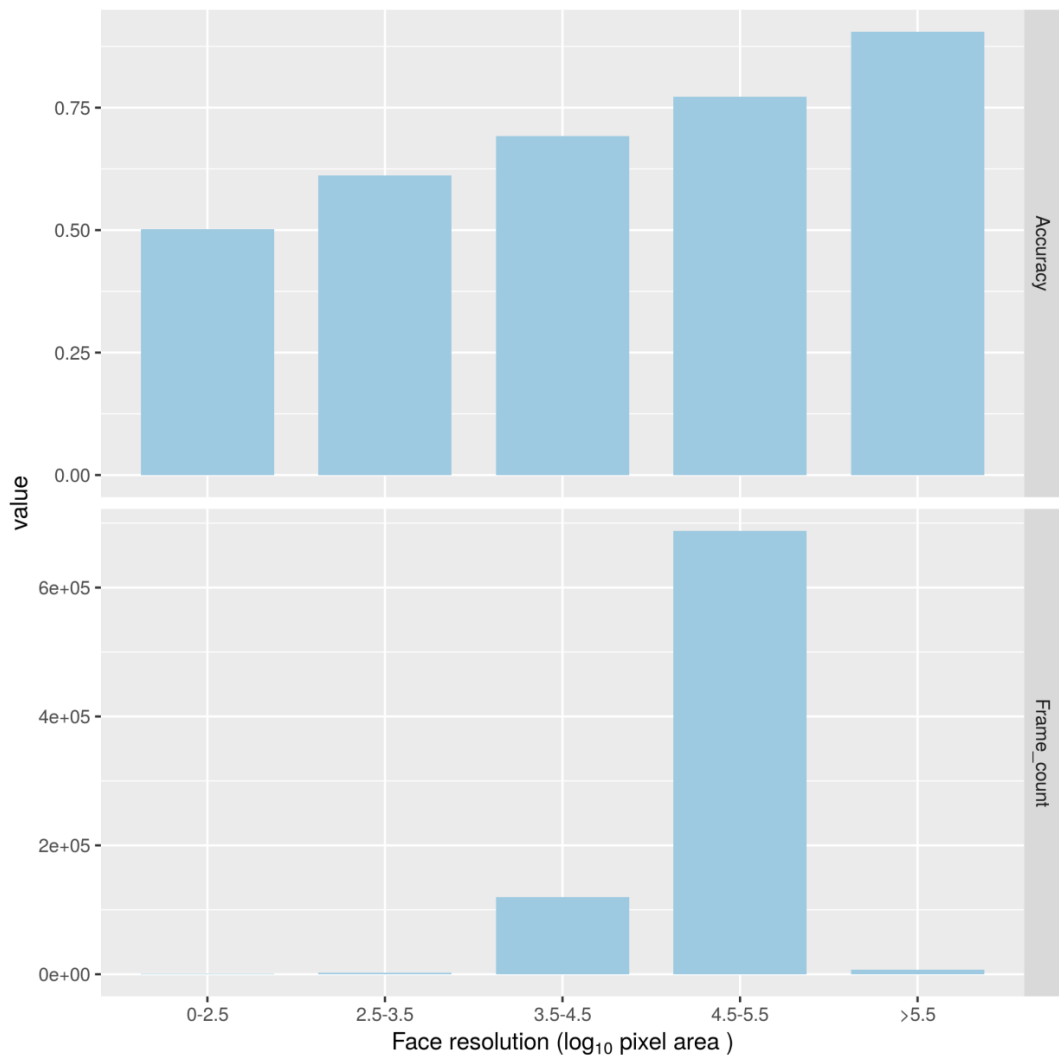


Fig. S3. Screenshots from the web-based experiment testing human annotator performance at identifying individual chimpanzees in cropped images. The experiment is performed in three separate web-browsers using the web-based VIA software interface: A table for cross-comparison of individuals (left), a gallery of 50 images per individual (center) and 100 images for identification by annotator (right). (Photo credit: Kyoto University, Primate Research Institute)



Frame Level accuracy	0.5015	0.6112	0.6915	0.7723	0.9048
No. of frames	325	2387	119776	687717	6715
Resolution (pixel area on a log scale)	0-2.5	2.5-3.5	3.5-4.5	4.5-5.5	>5.5

Fig. S4. Frame-level accuracy of model with variation in chimpanzee face resolution. Top graph shows that model accuracy improves monotonically with face resolution (pixel area). Bottom graph shows the variation in face resolution present in the dataset, with faces as small as 60 pixels. This variation is due to both camera resolution and size of faces on the screen as individuals move closer or further away and the camera pans and zooms.

Table S1. Name, ID code, sex, age, and years present for every chimpanzee at Bossou within the dataset analyzed. Ages are shown in years, and are estimated for those individuals that were already adults at the time long-term research at Bossou began in 1976 (indicated by *). Age ranges correspond to each individual's age during the year(s) indicated in the rightmost column.

Individual	ID code	Sex	Age range	Years present
TUA	TA	M	43-55*	2000-2012
Kai	Ka	F	50*	2000
Nina	Nn	F	46*	2000
Nto	Nt	F	7	2000
Fana	Fn	F	44-57*	2000-2013
Fanle	Fl	F	3-16	2000-2013
FANWA	FW	M	1-2	2012-2013
FLANLE	FE	M	1-6	2008-2013
Fotaiu	Ft	F	9	2000
FOAF	FF	M	20-33	2000-2013
Jire	Jr	F	42-54*	2000-2013
Joya	Jy	F	0-9	2004-2013
JEJE	JJ	M	3-16	2000-2013
Juru	Ju	F	7	2000
Velu	VI	F	41-54*	2000-2013
Vuavua	Vv	F	9	2000
Yo	Yo	F	41-54*	2000-2013
YOLO	YL	M	9-17	2000-2008
Pama	Pm	F	33-46*	2000-2012
PELEY	PE	M	2-15	2000-2012
PONI	PO	M	7	2000
Pili	PI	F	13	2000
POKURU	PK	M	4	2000

Table S2. Summary statistics of training and testing datasets for recognition model.

Columns show the number of individuals, hours, frames, face-detections and face-tracks that comprise the training and testing datasets in each year. We did not test on the year 2000 as most of the individuals in this year had too few (less than 30) tracks for training. We reserved data from the years 2006 (interpolated year) and 2013 (extrapolated year) solely for testing.

Year	Indvs	Train				Test			
		Hours	Frames	Face-detections	Face-tracks	Hours	Frames	Face-detections	Face-tracks
2000	20	8.9	802,872	2,141,685	3,800				
2004	12	10.4	935,381	1,717,809	2,616	0.7	61,480	81,148	120
2006	11					5.6	526,870	1,100,922	2,488
2008	13	9.3	835,207	2,086	4,426	1.7	161,657	425,041	1,029
2012	13	7.5	676,279	1,955,676	4432	0.9	82,221	230,882	1,118
2013	9					2.1	186,266	267,848	585
Total	23	36.1	3,249,739	7,901,740	15,274	11	1,018,494	2,105,841	5,340

Table S3. Identity and sex recognition results for accuracy on all faces and frontal faces only in the test set. Columns show accuracy calculated as the percentage of correctly labelled faces obtained from the model (compared to ‘ground truth’ labels provided by a human annotator). Frontal faces are defined as faces facing the camera where both eyes of the individual are visible. Accuracies were calculated for each year and an overall average is obtained by weighting the accuracy for each year by the number of detections. We report both the frame level accuracy (without track-level averaging) and the accuracy obtained using our final predictions (with track-level averaging).

	All faces				Frontal faces only			
	Identity		Sex		Identity		Sex	
Year	Frame level	Track level	Frame level	Track level	Frame level	Track level	Frame level	Track level
2004	85.51	92.67	95.37	99.82	87.28	93.95	96.57	100
2006	76.86	91.51	85.41	95.35	83.58	92.25	88.62	96.30
2008	81.83	94.71	90.08	97.33	83.98	95.36	92.27	99.24
2012	82.70	92.88	88.65	96.57	87.29	94.10	91.02	98.06
2013	75.63	90.35	89.99	99.08	77.09	91.37	90.31	99.83
Weighted Av.	79.12	92.47	87.34	96.16	84.30	95.07	90.12	97.36

Table S4. Confusion matrix for the 13 individuals in the test set. The colour gradient is from yellow to dark green for numbers of increasing value. The names along the top refer to the true class label and the names on the left refer to the prediction made by the model. Numbers along the diagonal represent per class accuracies, and numbers off the diagonal represent error proportions. The largest proportion of incorrect predictions is between Jy and FE, both infants in the dataset, followed by Jy and JJ (siblings) and YL and JJ (both light skinned males). The highest error proportions (≥ 0.10) are highlighted in red.

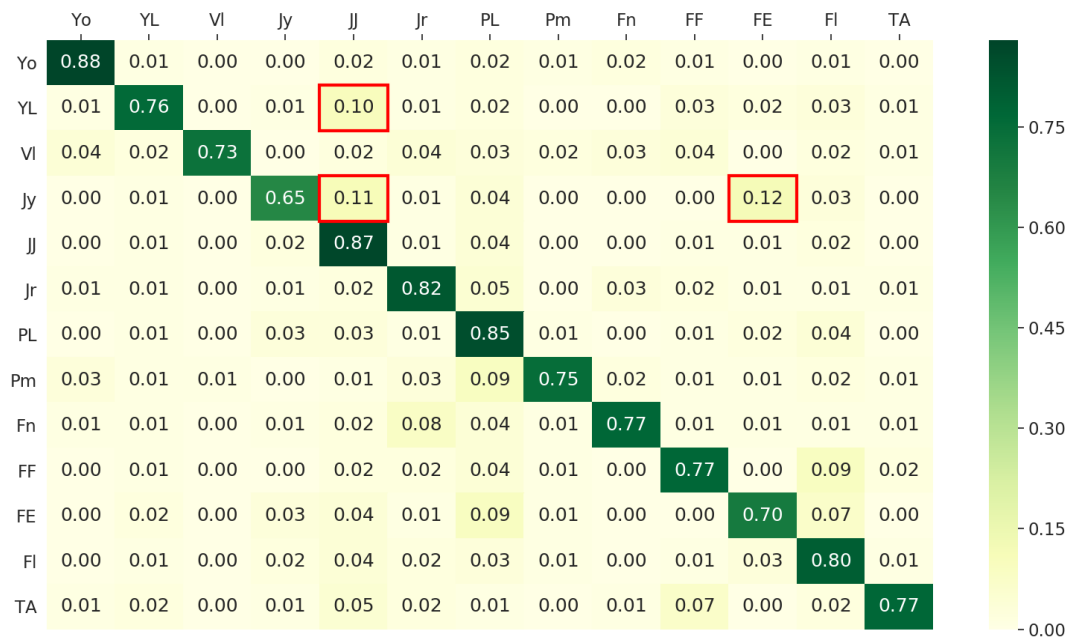


Table S5. Metrics of Bossou social networks derived from co-occurrences of detected individuals in video frames. Nodes correspond to individuals, edges to pairwise co-occurrences. Average degree is calculated as the average number of edges for each individual in the network.

Year	Hours of observation	Co-occurrences	Nodes	Edges	Av. degree
2000*	8.9	1363291	19	144	15.16
2004	10.4	585890	12	62	10.33
2008	9.3	1150418	13	78	12.00
2012	7.5	993042	13	66	10.15

* Individuals with fewer than 50 detections were removed from the network (n = 1)

Movie S1. Video demo of automated identity and sex recognition of wild chimpanzees at Bossou, achieved through our deep learning pipeline.

4

Automated face recognition using deep neural networks produces robust primate social networks and sociality measures

Published in special issue for *Methods in Ecology and Evolution* on ‘Beyond observations: applying contemporary methods across the fields to the study of animal sociality in the wild’

Automated face recognition using deep neural networks produces robust primate social networks and sociality measures

Daniel P. Schofield^{*1,2}, Gregory F. Albery^{3,4}, Josh A. Firth⁵, Alexander Mielke^{1,6}, Misato Hayashi^{7,8}, Tetsuro Matsuzawa^{7,9,10}, Dora Biro^{11,5}, and Susana Carvalho^{1,12,13}

¹Primate Models for Behavioural Evolution Lab, Institute of Human Sciences, University of Oxford, 64 Banbury Road, Oxford, OX2 6PN, UK

²Visual Geometry Group, Department of Engineering Science, University of Oxford, Oxford, UK

³Department of Biology, Georgetown University, Washington D.C., United States of America

⁴Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany

⁵Department of Biology, University of Oxford, Oxford, UK

⁶University of St Andrews, St Andrews, UK

⁷Chubu Gakuin University, Seki, Japan

⁸Japan Monkey Centre, Inuyama, Japan

⁹Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125, USA

¹⁰College of Life Sciences, Northwest University, Xi'an 710069, China

¹¹Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA

¹²Interdisciplinary Center for Archaeology and Evolution of Human Behaviour (ICArEHB), Universidade do Algarve, Faro, Portugal

¹³Gorongosa National Park, Sofala, Mozambique

March 2023

*Corresponding author, email: daniel.schofield@eng.ox.ac.uk

Abstract

1. Longitudinal video archives of behaviour are crucial for examining how sociality shifts over the lifespan in wild animals. New approaches adopting computer vision technology hold serious potential to capture interactions and associations between individuals in video at large scale; however, such approaches need a priori validation, as methods of sampling and defining edges for social networks can substantially impact results.

2. Here, we apply a deep learning face recognition model to generate association networks of wild chimpanzees using 17 years of a video archive from Bossou, Guinea. Using 7 million detections from 100 hours of video footage we examined how varying the size of fixed temporal windows (i.e aggregation rates) for defining edges impact individual-level gregariousness scores.

3. The highest and lowest aggregation rates produced divergent values, indicating that different rates of aggregation capture different association patterns. To avoid any potential

bias from false positives and negatives from automated detection, an intermediate aggregation rate should be used to reduce error across multiple variables. Individual-level network-derived traits were highly repeatable, indicating strong inter-individual variation in association patterns across years and highlighting the reliability of the method to capture consistent individual-level patterns of sociality over time. We found no reliable effects of age and sex on social behaviour and despite a significant drop in population size over the study period, individual estimates of gregariousness remained stable over time.

4. We believe that our automated framework will be of broad utility to ethology and conservation, enabling the investigation of animal social behaviour from video footage at large scale, low cost and high reproducibility. We explore the implications of our findings for understanding variation in sociality patterns in wild ape populations. Furthermore, we examine the trade-offs involved in using face recognition technology to generate social networks and sociality measures. Finally, we outline the steps for the broader deployment of this technology for analysis of large-scale datasets in ecology and evolution.

Key words: primate sociality, chimpanzee, social structure, social networks, face recognition, deep learning, computational methods

Data accessibility: Data and code used in this study are available in a Github repository at <https://github.com/dschof/DeepPrimateNetworksMEE> and archived in Zenodo at <https://zenodo.org/record/8063792> (Schofield et al. 2023).

1 Introduction

Social behaviour has important implications for survival and reproduction, and understanding sociality is therefore important for understanding animals' evolution and ecology (Abbot and Rubenstein 2017, Krause and Ruxton 2002). Using interaction patterns and spatial associations between group members, biologists can gain insight into the structure, function and evolutionary consequences of sociality for animals in the wild (Alexander 1974, Hinde 1976, Silk and Kappeler 2017). Social network analysis has proven to be a useful tool for quantifying individual-level patterns of association and group-level social structure in a wide array of species (Croft et al. 2008, Krause et al. 2015), including birds (Aplin et al. 2015a, St Clair et al. 2015), cetaceans (Allen et al. 2013) and primates (Sueur et al. 2011). Long-term field sites and repeated observations of individuals across many years are required to capture the demographic fluctuation of populations and the ecological and evolutionary processes driving species' adaptations and social behaviour (Clutton-Brock and Sheldon 2010). Large-scale image datasets, in particular remote camera trap databases and video archives (Burton et al. 2015, O'Connell et al. 2010) at high temporal resolution, have potential to examine behavioural change over time. Conventional methods employ human researchers to code behavioural information; however, this is becoming increasingly difficult to scale as video databases accumulate, limiting the scope of research and the statistical power of analysis. For typical coding tasks such as individual identification from camera trap footage, learning to reliably identify habituated individuals can take a matter of weeks for experienced coders, whilst training inexperienced coders or for more challenging tasks such as labelling unhabituated individuals or multiple behavioural variables, it can take months of research effort (McCarthy et al. 2019).

Recent developments in artificial intelligence (AI) and the field of computer vision, boosted by advances in deep learning (LeCun et al. 2015), have opened up the possibility for automating behavioural analysis from video (Vidal et al. 2021, Weinstein 2018). Individual detection, tracking and recognition using deep learning can be used to measure patterns of association and social networks at high frame rates (Schofield et al. 2019). Automated micro-sensing technologies such as animal-borne radio frequency identification (RFID) tags (Firth and Sheldon 2016, Gelardi et al. 2020), proximity loggers (St Clair et al. 2015) and GPS trackers (Strandburg-Peshkin et al. 2015) can track animal behaviour at high resolution

and accuracy, but require individual animals to be fitted with devices, which introduces logistical and ethical challenges that can be prohibitively invasive for some species such as chimpanzees (Dore et al. 2020). Use of deep learning for individual identification (Chen et al. 2020, Ferreira et al. 2020b) can efficiently generate sociality information for a wider array of species that are difficult to observe in situ or cannot be tagged or tracked using biologgers. Visual data from archives are valuable as they can reveal the nature and behavioural context of individual interactions. Video data can be collected through different methods such as systematic remotely operated camera trap grids or hand operated video cameras. Although use of video offers a valuable way to capture a detailed record of events, it also presents several visual challenges. For example, remote camera traps, which are fixed in situ, can miss animal interactions due to their limited fields of vision. Hand operated cameras can move more freely to capture social behaviors by panning or zooming, but this movement can cause motion blur or introduce observer bias by focusing on specific individuals or behaviors. Additionally, video data collected in the wild often has low resolution, poor lighting, and may be occluded by vegetation, making it a challenge to capture animal associations accurately and without bias.

Once video has been collected, behavioural information, such as the distribution of individuals in video across time and space, can be automatically extracted using individual recognition classifiers to produce a time series of individual detections and co-presence (Schofield et al. 2019). This data can be summarised to capture social relationships by generating ‘association’ based social networks, recording ‘co-occurrences’ if individuals appear together at a set location, within a certain threshold of distance, or within a time window, as an indication of the rate of interaction and the strength of the relationship - these relationships are represented as ‘edges’/edge weights between ‘nodes’ in a network (Farine and Whitehead 2015). Automating data collection of social behaviour through video-based analysis using AI enables high-throughput sampling compared to coding this information manually from in situ observation or retrospectively from video records, and thus provides greater statistical power which is required for robust estimates of these social indices (Franks et al. 2010, James et al. 2009). However, social interaction is highly dynamic, and choosing a set threshold for defining ‘grouping events’ of edges/edge weights between individuals in a network (otherwise referred to here as the aggregation rate i.e the size of the temporal window used for defining edges/co-occurrences and aggregating individual associations) to detect the

‘real’ biological signal for capturing meaningful associations is a considerable challenge (Ferreira et al. 2020a) ; furthermore, these thresholds are likely to be species, population specific, and may need to be adjusted depending on the study design or context. Thus far few studies have examined how the choice of sampling strategies, methods of data collection and network aggregation can impact biological results (Castles et al. 2014, Davis et al. 2018, Farine 2015, Gelardi et al. 2020, He et al. 2022). Choosing a fixed threshold for the size of temporal window used introduces trade-offs which need further investigation. For example, using high thresholds/short time windows (e.g. recording interactions every second) increases data throughput, but can aggregate too much noise (Farine 2018) and may introduce false negatives by failing to detect social interactions (Type II error). Conversely, using low thresholds/aggregation rates with larger time windows may increase false positives by aggregating individuals who were not interacting (Type I error), thereby producing networks which are lacking in variation and too low in resolution (Wang et al. 2014). The aim is to record associations at a high enough temporal resolution to capture meaningful variation in sociality without oversampling and introducing errors to the social networks.

In this study, we use a long-term video archive of wild chimpanzees (*Pan troglodytes verus*) from Bossou in Guinea to investigate the reliability of face recognition for producing social networks and sociality estimates at the individual level. To assess potential biases from use of human-operated cameras in our dataset (e.g individuals missed out of shot), we examine how varying the size of fixed temporal windows for sampling grouping events affects the patterns and consistency of social behavior estimates over time. Bossou offers a unique opportunity to examine social behaviour in the context of tool use, where chimpanzees have been continuously observed over 30 years at a clearing in the forest to crack nuts using stone tools (Matsuzawa 1994, Sugiyama and Koman 1979). Using footage recorded annually at the so-called ‘outdoor laboratory’ we examine three measures of gregariousness in chimpanzees: a) individual social network centrality (Strength) b) the average size of the party in which an individual is observed (Mean Group Size) and c) the proportion of time spent alone (Time Alone). The main aims of this paper are to 1) investigate how the aggregation rate/size of the temporal window used for capturing co-occurrences affects these estimates, 2) investigate how stable and repeatable these measures are at the individual level across years, and 3) model whether any demographic and life history variables (age and sex) are related to this variation. We provide future directions for the wider roll-out of the method and

139 suggestions for next steps of data collection and analysis.

140 **2 Methods**

141 **2.1 Study site and subjects**

142 Data were collected from the wild chimpanzee field research site of Bossou, southeastern Guinea, West
143 Africa. Long-term observations of a single community of wild chimpanzees (*Pan troglodytes verus*) have
144 been conducted here since 1976 (Matsuzawa et al. 2011, Sugiyama 1984). In 1988, researchers established
145 a quasi-experimental ‘outdoor laboratory’ in a natural forest clearing (Matsuzawa 1994) to observe nut
146 cracking behaviour, which peaks naturally in June-July and November-December (Yamakoshi 1998). In
147 this study, data were collected across 17 years (1996-2017) from film recording at the outdoor laboratory
148 during each dry season (Dec-Feb) whereby provisioning of nuts and stones by researchers allowed effective
149 observation of nut-cracking behaviour. Approximately 20-30 hours of footage were collected annually
150 using tripod-mounted human operated cameras, sometimes from multiple angles. The outdoor lab was
151 opened daily during daylight hours throughout the field season (typically across 1-2 months), and video
152 was only recorded if individuals attended, until they left the forest clearing. Video data collection was
153 noninvasive and approved by the ethics committee at Kyoto University’s Primate Research Institute.
154 All field research in Guinea conformed to ethical guidelines outlined by the Association for the Study of
155 Animal Behaviour.

156 **2.2 Video pre-processing**

157 To minimise computation for image processing and data storage, while ensuring consistent sampling
158 across years, an average of 6 hours of video content (5.99 ± 1.46 hrs) distributed across 15.29 ± 24.14
159 videos was randomly sampled per field season for 17 years spanning 1996 to 2017, totalling 101.82 hours
160 of footage from manually digitised tapes (see Table **S1**). To avoid data re-sampling, maintain consistency
161 and improve performance of our recognition model, we used one medium-distance camera angle per
162 year to capture clear footage of individual faces, avoiding the low-resolution from wider camera angles

where face detection is challenging. Videos were processed using a deep learning face recognition pipeline as described in (Schofield et al. 2019) using Pytorch (Paszke et al. 2019) and a ‘Titan X’ Graphical Processing Unit (GPU). For each video in the dataset, a CSV file produced a spatiotemporal database with coordinates of individual detections in each video frame. Multiple CSV files were collated into a large dataframe comprising ~ 7 million frame-level detections using the package ‘Pandas’ (McKinney 2011) in Python version 3.7.3 (Oliphant 2007).

The face recognition method uses supervised deep neural networks and makes a prediction for each face detection from 23 fixed identity classes (and an additional class for mistaken “nonface” false positive detections). Some individuals were rarely sighted and thus not included in training data (2002 $n = 1$, 1996-1999 $n = 1$). For these ‘unknown’ chimpanzees, the model is forced to falsely predict an identity from one of the fixed classes; to reduce the number of false positives, individuals known not to be present in that year (those yet to be born, or having died or dispersed) were removed from the dataframe using field notes/records. Continuous video sequences often contain shot-changes - jumps in time where the camera was switched on and off. To prevent an impact on the social network analysis (individuals being incorrectly grouped/associated across different time periods), a Python shot-detector (Castellano 2021) was used with its default sensitivity value to mark these points. Given the variable frame rate of videos, frames were converted into seconds for aggregation to reduce the amount of observations for analysis and provide consistent units of time for sampling networks.

2.3 Social network adjacency matrices

To quantify social networks of the Bossou group we adopted the ‘Gambit of the group’ methodology, which calculates association scores based on the observation of all individuals within a group at a single point in time, rather than using behavioural interactions such as grooming or physical contact (Franks et al. 2010). Videos are split into a set of fixed time windows/observation points (or ‘grouping events’) where individuals detected in the same time window are recorded as co-occurring in that group. This produces networks represented as weighted adjacency matrices, where the association scores for each pair of individuals (dyads) are weighted using the simple ratio index (Hoppitt and Farine 2018) (see section below) to reflect the proportion of times they co-occurred at the outdoor laboratory (Schofield et al.

2019). These matrices form the basis of social networks, where ‘nodes’ represent individual chimpanzees, and ‘edges’ the strength of association between individuals. These association-based networks thus infer social connections without measuring them directly, but when sampled at high rates produce robust association indices for easily observed groups (Farine and Whitehead 2015, Silk et al. 2015).

2.4 Correlations across aggregation rates

In the context of video data collection at the outdoor laboratory, individuals were recorded at approximately a 15-20m distance. The general protocol for the camera shots was to capture all individuals within the video frame, covering most of the 7x20m space of the outdoor lab. However, occasionally the camera would pan or zoom to capture specific bouts of activity, such as a special case of tool use or social interaction, which may only include the 1-5m surrounding the focal individuals (Fig. 1). Wider shots are more likely to capture all individuals at the forest clearing, whereas zoom shots capture individuals in close proximity. The size of the temporal window used for recording social associations thus has key implications for capturing social associations. Sometimes the detector can fail to detect faces when they are obscured from view (for example, by vegetation or another individual) or the individual moves out of shot (e.g Fig. 1, 3-4). Larger time windows (e.g. spanning multiple seconds, minutes or even hours) may be less sensitive to these detection errors, as only a single detection is needed for individuals to co-occur in the time window - this means individuals can co-occur, despite having moved temporarily out of shot (the observational area). Larger time windows may group together individuals who never interacted - for example, if they enter and leave the forest clearing at different points in time. Furthermore, a single misidentification in an image can incorrectly aggregate individuals across this larger time range (for example, a 10 minute interval).

To examine how temporal window size impacts gregariousness scores and social networks, we constructed weighted adjacency matrices for a broad range of temporal windows/fixed aggregation rates (1-3600 seconds at 10 second intervals), resulting in 360 methodological replicates. Social networks were generated and analysed in R v. 4.1.0 (R Core Team 2021). For each aggregation rate, we generated a ‘group-by-individual’ matrix detailing the co-occurrences between all chimpanzees, which was weighted using the simple ratio index (SRI): $\text{Detections}_{A,B} / (\text{Detections}_A + \text{Detections}_B - \text{Detections}_{A,B})$ (Cairns and Schwager 1987).

This represents the proportion of grouping events/time windows in which focal individuals were detected together. One social network was produced for each field season, which was resampled using each aggregation rate, producing 6120 networks for comparison for 17 years of data. To examine how varying the size of the temporal window affected the sociality scores we used Pearson's correlation coefficient to compare social network values (strength) calculated from weighted networks generated using face recognition ('AI') at different aggregation rates (1-3600 seconds) against the most extreme aggregation rate (every second). We also mapped how the mean values for Time Alone, Mean Group Size and Network Density changed as a function of aggregation rate across this time interval range.

2.5 Statistical analysis

2.5.1 Sociality measures

We generated common individual-level network and sociality variables used to study animal social networks (Farine and Whitehead 2015, Whitehead 2008) and gregariousness across the life span in primates (Machanda and Rosati 2020). Individual metrics were used to examine correlations between different sampling methods, and as response variables in Generalised Linear Models (GLMMs). We calculated measures including: (1) 'Strength' - the sum of an individual's social associations or the proportion of time spent associating with others; (2) 'Mean Group Size' - the average number of co-detections at any one time across all time windows as a measure of party size and (3) 'Time Alone' - the proportion of 'grouping events'/time windows where an individual was detected alone as a gauge of solitariness. We calculated a 'global' network measure to quantify a feature of the overall structure of the network ('Density') - the proportion of edges (social connections) to the total potential edges in the weighted network - to examine how aggregation rates affected the overall cohesiveness and connectivity of the network.

2.5.2 Models

For every year in the dataset, each individual has one value for each of our outcome variables/gregariousness scores (conditional on being detected by our recognition model at least once in that year). Consistent

242 experimental and filming protocols at a single location and during the same time of year reduced the
243 potential impact of nuisance/extraneous variables on our sociality scores (e.g such as from variations
244 caused by environmental factors, locality and seasonality). For two variables (Strength and Mean Group
245 Size) we fitted linear mixed models (LMM) with Gaussian error structure and identity link function using
246 the package “lme4” (Bates et al. 2007) in the R v. 4.1.0 statistical software package (R Core Team 2021).
247 To model the proportion data for ‘Time Alone’, we fitted a Generalized Linear Mixed Model (GLMM)
248 with a beta regression error structure, logit link function and Restricted Maximum Likelihood (REML)
249 using the package “glmmTMB” (Brooks et al. 2017). In order to fit the beta regression model using
250 glmmTMB, the zero-inflated proportion scores (individuals who were not observed alone) were adjusted
251 by 0.001. We examined whether demographic variables (Age and Sex) influenced our three individual
252 sociality metrics (sampled at an intermediate 120 second aggregation rate). All model sets retained the
253 same structure for fixed and random effects: Individual Identity and Year were included as categorical
254 random effects; Sex, Age in years, and Population Size were included as fixed effects. Given that age
255 has been shown to follow a quadratic curve in primate sociality and cognition (Lacreuse et al. 2020),
256 we included it as a quadratic term for all models. We scaled all of our continuous predictor variables
257 to have a mean of 0 and a standard deviation of 1 to encourage model fitting and aid interpretability
258 (Schielzeth 2010). No further transformation was applied to the quadratic age term (which represents
259 the square of the z-transformed Age variable). To examine the effect that our test variables (Age and
260 Sex) had on individual sociality scores, we used likelihood ratio tests (Harrison et al. 2018, Nakagawa
261 and Schielzeth 2010) comparing our full model containing the fixed predictors of interest, and a reduced
262 regression/‘null’ model, which only contained our control predictor variables (Year of Observation and
263 Individual as random effects, and Population Size as a control fixed effect). This approach differs from
264 the use of null models that rely on randomisations or permutations to address bias in network studies
265 (e.g Farine and Carter 2022). For each variable, a further likelihood test was performed comparing the
266 full model with a null model with individual identity as a random effect removed to examine the impact
267 it had on our outcome variables (Nakagawa and Schielzeth 2010) (See ‘Repeatability’ section below).
268 For each full-null model comparison, we used the ‘anova’ and ‘compare’ functions in the “performance”
269 package in R (Lüdtke et al. 2021) to assess model fit and obtain p-values. For all models we checked for

collinearity, homogeneity of variance and the normality of residuals using the ‘check_model’ function in the “performance package”, which did not reveal any significant abnormalities. ‘Time Alone’ was modelled using beta regression which showed slight violation of normality of residuals due to outliers in the dataset, indicating that our model under-predicted some large values at the tail end of the distribution (Fig. **S6**).

2.5.3 Repeatability

Repeatability measures the consistency of inter-individual differences over time. Traits that show a high level of repeatability have high between-individual variance and low within-individual variance, which indicates that individuals have differing and stable tendencies or strategies for interacting/associating with group members over time (Tkaczynski et al. 2020). To identify the individual-level repeatability of these metrics, we assessed the proportion of variance in our gregariousness variables that were attributable to the individual identity random effect in our models (Nakagawa and Schielzeth 2010). We compared full models containing all fixed and random effects, and null models without the random intercept and slope for individual identity, to test whether individual identity explained variance in our outcome variables and improved model fit (Tkaczynski et al. 2020). We used the ‘r.squaredGLMM’ function from the ‘MuMin’ package in R (Bartoń 2021) to calculate individual R^2 for a model containing individual identity as the only random effect (dropping Year) to examine the repeatability/strength of the effect that individual identity had.

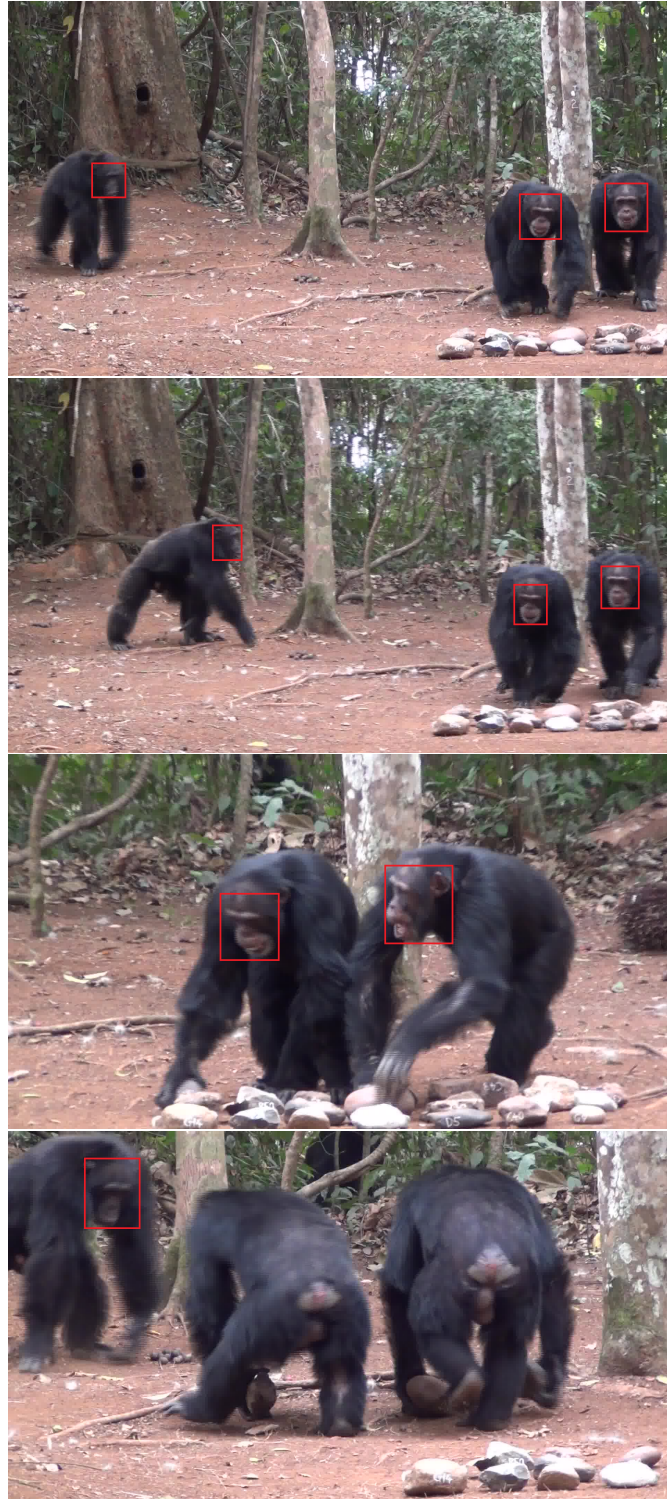
3 Results

3.1 Aggregation rates

Figure **2** summarises the changes in correlations and mean values of social network and sociality variables across rates aggregation. For Strength (Fig. **2A**), we found a steep drop-off in correlation values when comparing networks from the highest aggregation rate (1 second) to other aggregation rates, indicating that the highest and lowest aggregation rates produced divergent social network scores. This indicates that extreme aggregation rates may be capturing different association patterns, and may also be introducing errors affecting the network scores. The network density ranged from 0.94 - 0.99 (Fig. **2C**) which indicates

295 that as the size of the aggregation time window increases, a greater proportion of individuals are observed
296 together, reducing variation in association rates and network scores. This high range of connectivity in
297 the network is also likely due to the small group size and a feature specific to the Bossou population (see
298 discussion in Supplementary section 4.5). Sociality variables ('Mean Group Size' and 'Time Alone') also
299 appear to be sensitive to the size of the temporal window, with the highest aggregation rate producing
300 more extreme values before stabilising at around the 600 second aggregation rate. This suggests that
301 short temporal windows (below a minute) are susceptible to bias when the camera zooms or captures a
302 single individual in shot, therefore overestimating the time individuals spend alone and underestimating
303 the party size. Using a low aggregation rate with larger time windows has the inverse effect, reducing
304 the number of observations, overestimating party size and underestimating individual solitariness as
305 more individuals are grouped together. Collectively across measures, visual assessment of these curves
306 indicates that an intermediate value of 50-120 seconds avoids extreme values and the potential noise
307 generated from oversampling, whilst still capturing enough variation in the networks to make inferences
308 about individual-level sociality patterns, however this optimal window may shift depending on the time
309 span of the study and amount of data collected.

Figure 1. Sequence from a 15-second clip of footage filmed at the outdoor laboratory in 2012. The example detections depict three adult males and highlight shot variations resulting from the moving camera, as well as potential biases caused by missed detections when individuals turn their heads or move out of shot. Images 1-4 (top to bottom) show: (1-2) a medium shot of three individuals entering; (3) a zoom shot of two males (with one peripheral male out of shot); and (4) two male turning their backs and being missed by our detector.



310 **3.2 Modelling results**

311 Effect sizes and measures of model fit from likelihood ratio tests for each model set are presented in Table
312 **1** and **2**. We found no significant differences between the effect of age and sex on sociality scores for
313 Strength ($X^2 = 5.18$, $df = 2$, $P = 0.075$) and Mean Group Size ($X^2 = 2.58$, $df = 2$, $P = 0.27$). We found
314 weak evidence for a quadratic age effect on Time Alone (AIC difference = 2.47)(Fig. **S8C**). Repeatability
315 scores for each sociality variable are presented in Table **3**. For all models, individual identity improved
316 model fit: Strength ($X^2 = 116.14$, $df = 1$, $P < 0.001$), Mean Group Size ($X^2 = 11.70$, $df = 1$, $P <$
317 0.001) and Time Alone ($X^2 = 20.11$, $df = 1$, $P < 0.001$). This indicates that Bossou chimpanzees showed
318 inter-individual variation and high repeatability (consistency) in gregariousness over time. For all of our
319 models, individual identity had a significant impact on model fit; however, the size of the effect varied -
320 the variation explained by the random effect of individual identity was highest for Strength (0.35) followed
321 by Time Alone (0.13) and Mean Group Size (0.03). Inter-individual variation in Strength, Mean Group
322 Size and Time Alone are visualised in Figure **3**.

Figure 2. Networks aggregated using different fixed time windows/aggregation rates (1-3600 seconds). A) Strength values from each network generated at different aggregation rates, correlated using Pearson's to the smallest temporal window (1 second). B) Mean value for group size and Time Alone (scaled to be between 0 and 1) for each aggregation rate. C) Network density for each network and aggregation rate. Dashed lines indicate range for intermediate aggregation (50-120 seconds) which avoids extreme values

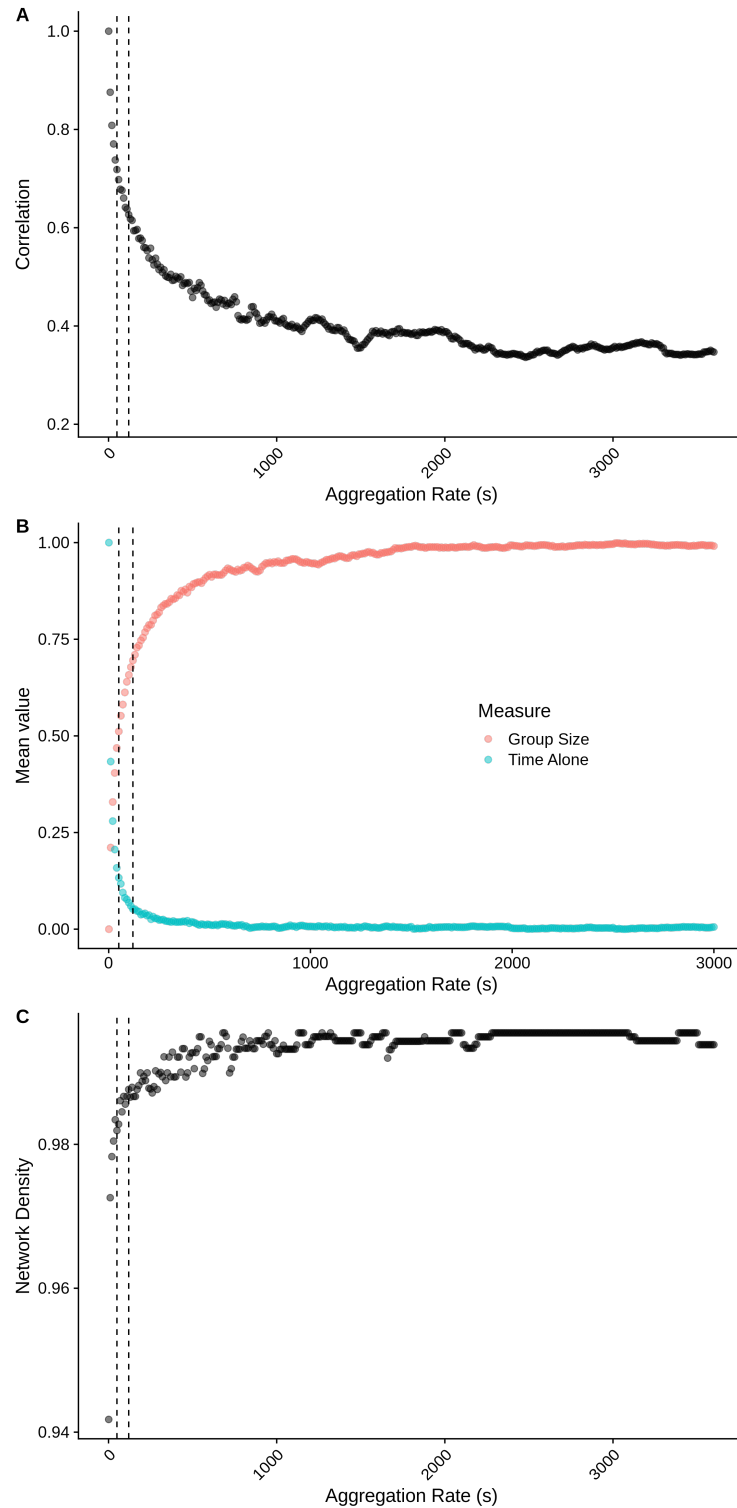


Figure 3. Individual random effect coefficients (x-axis indicates mean and variation) representing inter-individual variation in chimpanzee sociality measures for Strength, Mean Group Size and Time Alone at the 120 second aggregation rate. Scores above or below zero represent higher or lower than average expression in outcome variables for each individual. Error bars indicate 95% confidence intervals for the effects.

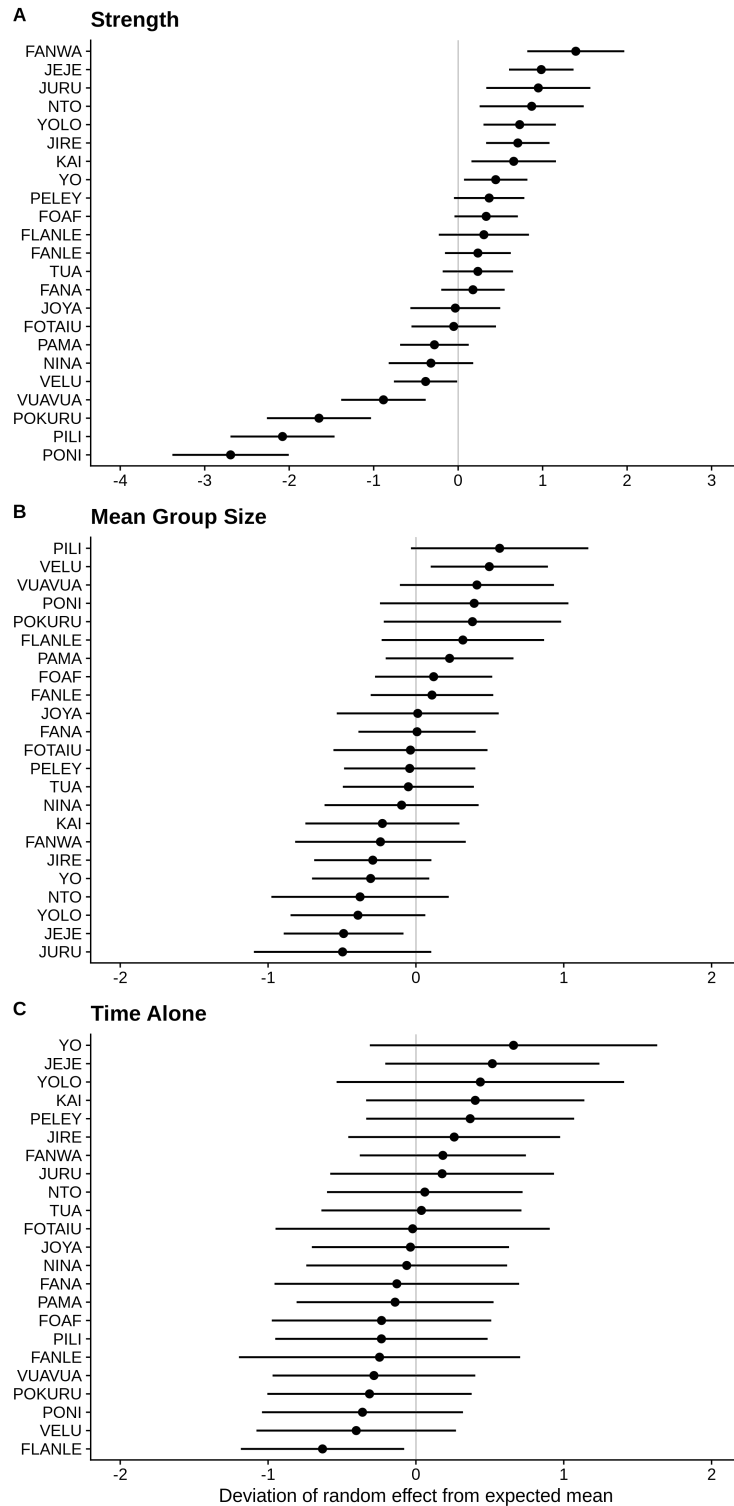


Table 1. Model results: Overview of model parameters for linear mixed models fitted using the ‘lmer’ and ‘glmmTMB’ packages in R.

<i>Outcome variable</i>	<i>Term</i>	<i>Estimate</i>	<i>SE</i>
Strength	Intercept	3.50	0.35
	Sex	-0.93	0.47
	Age ²	-0.11	0.08
	Population size	0.79	0.17
Mean Group Size	Intercept	3.50	0.35
	Sex	-0.93	0.47
	Age ²	-0.11	0.08
	Population size	0.79	0.17
Time Alone	Intercept	-2.91	0.18
	Sex	-0.24	0.24
	Age ²	-0.20	0.08
	Population size	-0.29	0.08

Table 2. Likelihood ratio test comparing model fit between full model containing all fixed and random effects, and a null model with fixed predictor variables of interest (age and sex) removed for each of our outcome variables: ‘Strength’, ‘Mean Group Size’ and ‘Time Alone’. For each likelihood ratio test we report the AIC, BIC and R^2 for the full and null model, and the difference between them (null-full) when the predictors of interest (age and sex) are removed from the full model. We also present X^2 and P-value for significance, and AIC for glmmTMB beta regression model (a difference of ~ 2 or less in AIC indicates no difference between models).

<i>Outcome variable</i>	<i>Model comparison</i>	<i>AIC</i>	<i>BIC</i>	<i>Conditional R^2</i>	<i>Marginal R^2</i>	<i>X^2</i>	<i>df</i>	<i>P value</i>
Strength	Full model	537.61	561.62	0.88	0.33	-	-	-
	Null model	538.79	555.94	-	-	-	-	-
	Full-Null	1.18	-5.68	-	-	5.18	2	0.075
Mean Group Size	Full model	657.48	681.49	0.78	0.52	-	-	-
	Null model	656.06	673.21	-	-	-	-	-
	Full-Null	-1.42	-8.28	-	-	2.58	2	0.27
Time Alone	Full model	-1027.52	-1003.52	0.27	0.08	-	-	-
	Null model	-1030.00	-1012.85	-	-	-	-	-
	Full-Null	-2.47	-9.33	-	-	-	-	-

Table 3. Individual chimpanzee repeatability in sociality variables. Variance in outcome variables attributed to individual identity as a random effect in linear models (R^2), spanning 17 years of video footage, and aggregated at 120 second time windows. We report the AIC/BIC values for the full/null model with the random effect of identity removed, and the X^2 and the P-value significance of this difference. We report ‘ID model’, which is the full model with ID as the only random effect (dropping year), to show the strength of the effect of individual ID, as a measure of individual-level repeatability of each trait. Individual R^2 is conditional R^2 (variance of fixed and random effects combined) - R^2 (the variance of only fixed effects).

<i>Outcome variable</i>	<i>Model comparison</i>	<i>AIC</i>	<i>BIC</i>	<i>Conditional R^2</i>	<i>Marginal R^2</i>	<i>X^2</i>	<i>df</i>	<i>P value</i>
Strength	Full model	537.61	561.62	0.88	0.33	-	-	-
	Null model	651.75	672.32	-	-	-	-	-
	Full-Null	114.14	110.71	-	-	116.14	1	<0.001
	ID model	-	-	0.71	0.35	0.35	-	-
Mean Group Size	Full model	657.48	681.49	-	-	-	-	-
	Null model	667.19	687.76	-	-	-	-	-
	Full-Null	9.70	6.28	-	-	11.70	1	<0.001
	ID model	-	-	0.58	0.56	0.03	-	-
Time Alone	Full model	1027.52	1003.52	-	-	-	-	-
	Null model	1009.41	988.84	-	-	-	-	-
	Full-Null	18.11	14.68	-	-	20.11	1	<0.001
	ID model	-	-	0.23	0.10	0.13	-	-

4 Discussion

In this study we tested the reliability of automated face recognition for measuring association-based social networks and individual-level gregariousness scores, using a long-term video archive dataset of wild chimpanzees. We examined how the size of the temporal window used for capturing co-occurrences affects the robustness of estimates of chimpanzee association scores over time, and the impact of potential errors and biases generated from different rates of aggregation on these measures. Overall, our results reveal that the size of temporal window used for aggregation of social networks should be a key consideration when measuring gregariousness using computer vision-based identification of individuals in video footage. Although networks produced across different aggregation rates were generally well correlated, extreme aggregation rates produced divergent scores, and our sociality variables Mean Group Size and Time Alone showed a steep drop off in values (Fig. 2B), indicating that care should be taken to evaluate how variables change as a function of aggregation rate. We found that selecting an intermediate aggregation rate avoids potential noise and extreme values generated from under- and over-aggregating. This approach improves the researcher's ability maximize the intensity of data collection while reducing errors, to capture meaningful variation in networks and gregariousness scores. However, the behavioural context of the study species and the choice of measures should inform the sampling strategy and size of temporal window used for capturing social behaviour.

4.1 Robustness of method

Our dataset was sampled annually within a 1-2 month window of data collection spanning 17 years. This is within the range of observation time for other studies generating animal social networks from intensive data collection over short periods (Davis et al. 2018, Feczko et al. 2015, Gelardi et al. 2020). The correlation of networks produced using larger temporal windows (lower aggregation rates) appeared to break down at around 1000s (Fig 2A), indicating that networks and measures generated at this aggregation rate were unreliable, which may be driven in part by too few data points and the sparse networks it produced. Examining the performance of the models, individual repeatability, and the distributions of sociality variables across all aggregation rates could help to investigate the source of

diverging values and potential bias in the networks further. For other automated tracking technologies, such as GPS, there is often a trade-off between rate of sampling and the longevity of battery life (He et al. 2022, McCann et al. 2021); sparse sampling can be surprisingly low for producing robust association networks if collected over many months (Davis et al. 2018). Thus, consideration should be given to the span of data collection in video databases - for our small seasonal snapshots, more hours of observation within each year would be required to generate robust measures using larger temporal windows. Data collected consistently across the field season through continuous monitoring (such as remote camera traps) may generate robust and stable network estimates using larger windows of aggregation. In the context of our study, using high aggregation rates may more closely approximate proximity networks (measuring how often individuals are within several metres of one another), whereas sparser aggregation rates may approximate party membership at a coarser level. In our study we have also not examined how sparse networks from using larger temporal windows produce false positives, due to the potential impact of a single misidentification for aggregating individuals within a large time window, and in future additional thresholds can be implemented to reduce these errors (e.g requiring a certain amount of time for individuals to spend together to co-occur within a time window).

Our results indicate that our recognition model produced individual-level estimates of sociality that were replicable over time at our intermediate rate of aggregation, highlighting the robustness of the method to produce reliable sociality estimates. Our finding that gregariousness was variable between individuals, and consistent within individuals across years, is consistent with the wider pattern shown by many other species, namely that behavioural patterns such as social network positions remain stable across the lifespan (Aplin et al. 2015b, Blaszczyk 2017, Finger et al. 2017, Jacoby et al. 2014, Wuerz and Krüger 2015). This also supports previous findings on repeatability of sociality in other chimpanzee groups which indicate that gregariousness becomes fixed during development through some form of canalisation, and remains consistent throughout their lifetime (Altschul et al. 2018, Tkaczynski et al. 2020).

4.2 Extending to other species and contexts

This system was developed and tested for a single group of chimpanzees, and thresholds are likely to vary depending on the ecological context and species (Webster and Rutz 2020). However the methodology for

model training are not species-specific (Schofield et al. 2019), and it is relatively simple to implement as it only requires a time series of individual detections, which means it can be flexibly applied for different forms of video data collection, species and ecological contexts. Our approach using Gambit of the Group for co-occurrences was able to generate enough variation and consistency in social measures, despite the relatively low image quality of archival data, moving cameras, small population size and specific context of a single forest clearing at Bossou. For video datasets collected over short time-frames (e.g over several hours, days or weeks), in constrained contexts where many individuals can be observed clearly in the same shot or do not have freedom to vary their social groups (e.g in an enclosure in captivity), then additional classifiers which track proximity or behavioural interactions may be required (see future directions below). The approach outlined here is particularly useful for species that exhibit fission-fusion dynamics and for data collection where grouping events are recorded over many months (Franks et al. 2010, Silk et al. 2015) and thus can be applied for either hand-held video focal follow data or long-term camera trap projects. For any approach, consideration of the size and quality of the shot should be considered, and studies using automated recognition classifiers should maximise the field of vision and image quality to capture all visible individuals whilst maintaining a close enough shot to detect individual faces reliably. Use of a full body recognition classifier should be a consideration due to its potentially greater recall than face recognition, depending on the size of the species, how recognizable individuals are by their bodies and the ecological context (e.g how often individuals are occluded by vegetation) (Bain et al. 2019, Ferreira et al. 2020b).

4.3 Benefits of using AI for animal sociality

A significant advantage of this system is that once the neural networks have been trained, archive footage can be used to optimise strategies for data collection of social behaviour from longitudinal datasets. Manual data collection of social networks undertaken retrospectively via video coding or through behavioural observation in situ is time-consuming and can take many times the length of the video depending on the task (Anderson and Perona 2014). Our method for generating social indices is entirely automatic, so once models are trained no video coding is required, which reduces potential for human observer bias and improves scalability; previous work found that without contextual information

researchers struggle to identify individuals from video footage, being outperformed by the recognition model in both speed and accuracy for this fine level task (Schofield et al. 2019). This is particularly relevant for long-term field sites where researchers with different levels of experience and interest can introduce variation in how data is collected from year to year. A key problem in the field of animal social networks is generating sufficient data for a robust analysis (Farine and Whitehead 2015, Franks et al. 2010). This automated approach is scalable and can increase the quantity of observations, increasing the statistical power and reliability estimation of the ‘real’ networks, without the need for re-watching and coding video.

4.4 Future directions

Despite these significant advantages, there are limitations of using AI. Deep neural networks take significant time and resources in terms of storage, annotation and processing of video data. Currently, investment is required by engineers to build, maintain and run the pipelines; these require technical knowledge, reducing the availability and practicality of the approach in the field and leading to bottlenecks for data preparation, annotation, storage and training between engineers and biologists. In order for this technology to be more widely adopted, open datasets and end-to-end pipelines need to become more transparent, accessible and efficient to use through open-access software projects (Mathis and Mathis 2020, Nath et al. 2019). The original development for our system took over a year to implement, however the bulk of this time was spent on acquisition of the data, expert annotation and development of the tools to train models from scratch. However it should be possible to train and fine-tune models in much less time as more open frameworks and models are released and computational resources become more affordable. Supervised methods are useful for retrospective analysis of complete archives when individuals are clearly visible and identifiable, as is the case with our dataset. However, applying these methods becomes significantly more challenging for large unhabituated populations with unknown individuals. Annotating data for training requires significant effort, and incorporating new individuals into the model may require modifying the final layer of the neural network and retraining (Schofield et al. 2019). New methods are emerging to improve the efficiency and flexibility of computer vision pipelines. Active learning models can automatically locate images of interest for human annotators, and lightweight on-device/real-time recognition systems can

improve the ease with which data is stored and annotated (Norouzzadeh et al. 2021). New methods for re-identification of individuals are emerging (Miele et al. 2021), such as metric learning, which can quantify similarity between images in a dataset of individuals, or one-shot meta-learning, training networks to learn categories without human annotation, without needing to retrain from scratch, or when classes are data poor (Hospedales et al. 2021).

The flexibility of such systems will enable development of more sophisticated analysis - for example, layering individual identification with behavioural recognition (Bain et al. 2021). Such new approaches will enable capture of a richer array of animal behaviour, to examine how individual-sociality, behaviour and networks interact in real-time - for example, to trace how new behaviours spread through a population via social learning, using dynamic network-based diffusion analysis (Hasenjager et al. 2021, Hobaiter et al. 2014). Despite over 60 years of field research, few studies have leveraged a longitudinal approach to examine how sociality patterns change over time and are impacted by demographics, such as reductions in population size, the loss of key individuals to death or dispersal, lack of emigration, reductions in fertility and ageing population structures (Sheldon et al. 2022, Shizuka and Johnson 2019). For large-camera trap databases and continuous monitoring projects, systematic comparison of multiple communities using automated classifiers for video data has potential to examine behavioural variation of many animal groups across large temporal and geographical ranges.

4.5 Conclusion

Capturing social associations using automated face recognition and social networks is a powerful tool for unlocking video archives and revealing the social structure of wild animal populations; however, investigation of the adequate thresholds for defining edges is required to provide robust insights into individual-level patterns of sociality. Overall, we find that automated face recognition using deep neural networks provides a robust method for producing chimpanzee social networks and sociality measures, and brings several key advantages including reduction in time investment in data collection and human observer bias, high through-put sampling for statistical power, and optimisation of network aggregation strategy for modelling biological processes. The method produces repeatable social networks, which show that inter-individual differences in sociality patterns remain stable over time in a group of wild

chimpanzees. Using this automatic framework, we find that although there was individual-level variation in social behaviour, age and sex do not appear to be strong drivers of individual gregariousness and social structure for the Bossou community in this data. This work provides a framework for evaluating the consistency and reliability of using AI to measure social behaviour from longitudinal archive datasets; it moves towards the widespread adoption of numerous deep learning algorithms which are able to generalise behavioural data collection across multiple species and habitats. Ultimately this methodology builds the platform to model the flexibility and variation in social behaviour and social networks in different ecological contexts and populations, over larger temporal and geographic ranges and across the lifespan in wild animals.

Acknowledgements: We thank Primate Research Institute, Kyoto university, the Direction Nationale de la Recherche Scientifique et Technique (DNRST) and Institute de Recherche Environnementale de Bossou (IREB) for supporting research, and all the researchers and field assistants who assisted with data collection at Bossou since 1988. We thank guest editors Thibaud Gruber and Erica van de Waal for inviting us to contribute this study to the special issue, and Damien Farine, one anonymous reviewer, Andrea Migliano and Alex Alvergne for their helpful feedback which greatly improved the manuscript. We also thank Arran J. Davis, João Coelho and Jean Debarros for their advice on statistics, and the Visual Geometry Group (VGG), Department of Engineering, University of Oxford, for their support, development and resources for the project (Andrew Zisserman, Arsha Nagrani, Max Bain, Abhishek Dhutta, Ernesto Coto Rios). Funding: Clarendon Fund and Boise Trust Fund, Wolfson College, University of Oxford (D.S.); Kyoto University Primate Research Institute for Cooperative Research Program (M.H. and D.S.); National Geographic Society (S.C.); St Hugh's College, University of Oxford (S.C.); Templeton World Charity Foundation grant no. TWCF0316 (to D.B.); MEXT-JSPS (no. 16H06283), LGP-U04, the Japan Society for the Promotion of Science (T.M.).

Author contributions: Daniel Schofield conceived of the study, curated the dataset, carried out formal analysis, visualization and wrote the original manuscript. Gregory Albery, Josh Firth, Alexander Mielke contributed with code and advice for the methodology. Susana Carvalho and Dora Biro supervised

and revised the work. Susana Carvalho, Dora Biro, Misato Hayashi and Tetsuro Matsuzawa provided resources, collected data, and supervised research in Japan and in Guinea. All authors reviewed the manuscript and provided feedback.

Conflict of Interest:

The authors have no conflicts of interest to declare.

Statement on inclusion: Our study was conducted using archival footage, thus, no fieldwork was undertaken for this research. Our authorship team has a broad geographical distribution, which represents a diversity of perspectives. We benefit from a long-standing international collaboration with local Guinean counterparts, including the IREB and DNRST, to support research at Bossou. However, we acknowledge that our study did not include scientists from Guinea, the country from where the data was originally collected. This limitation is due, in part, to the lack of local researchers actively engaged in this area of research at the study site. We recognise the significance of supporting and including local researchers and field assistants in scientific research, and we advocate for and encourage concerted efforts towards this objective in future research.

References

- Abbot, P. and Rubenstein, D., editors (2017). *Comparative social evolution*. Cambridge University Press Cambridge, UK.
- Alexander, R. D. (1974). The evolution of social behavior. *Annu. Rev. Ecol. Syst.*, 5(1):325–383.
- Allen, J., Weinrich, M., Hoppitt, W., and Rendell, L. (2013). Network-based diffusion analysis reveals cultural transmission of lobtail feeding in humpback whales. *Science*, 340(6131):485–488.
- Altschul, D. M., Hopkins, W. D., Herrelko, E. S., Inoue-Murayama, M., Matsuzawa, T., King, J. E., Ross, S. R., and Weiss, A. (2018). Personality links with lifespan in chimpanzees. *Elife*, 7:e33781.
- Anderson, D. J. and Perona, P. (2014). Toward a science of computational ethology. *Neuron*, 84(1):18–31.

- 509 Aplin, L. M., Farine, D. R., Morand-Ferron, J., Cockburn, A., Thornton, A., and Sheldon, B. C. (2015a).
510 Experimentally induced innovations lead to persistent culture via conformity in wild birds. *Nature*,
511 518(7540):538–541.
- 512 Aplin, L. M., Firth, J. A., Farine, D. R., Voelkl, B., Crates, R. A., Culina, A., Garroway, C. J., Hinde,
513 C. A., Kidd, L. R., Psorakis, I., Milligan, N. D., Radersma, R., Verhelst, B. L., and Sheldon, B. C.
514 (2015b). Consistent individual differences in the social phenotypes of wild great tits, *parus major*.
515 *Anim. Behav.*, 108:117–127.
- 516 Bain, M., Nagrani, A., Schofield, D., Berdugo, S., Bessa, J., Owen, J., Hockings, K. J., Matsuzawa,
517 T., Hayashi, M., Biro, D., Carvalho, S., and Zisserman, A. (2021). Automated audiovisual behavior
518 recognition in wild primates. *Sci Adv*, 7(46):eabi4883.
- 519 Bain, M., Nagrani, A., Schofield, D., and Zisserman, A. (2019). Count, crop and recognise: Fine-Grained
520 recognition in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*
521 *Workshops*, pages 0–0.
- 522 Bartoń, K. (2021). MuMIn: Multi-Model inference: <https://cran.r-project.org/package=mumin>.
- 523 Bates, D., Sarkar, D., Bates, M. D., and Matrix, L. (2007). The lme4 package. *R package version*, 2(1):74.
- 524 Blaszczyk, M. B. (2017). Consistency in social network position over changing environments in a
525 seasonally breeding primate. *Behav. Ecol. Sociobiol.*, 72(1):11.
- 526 Brent, L. J. N., Ruiz-Lambides, A., and Platt, M. L. (2017). Family network size and survival across the
527 lifespan of female macaques. *Proc. Biol. Sci.*, 284(1854).
- 528 Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug,
529 H. J., Machler, M., and Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages
530 for zero-inflated generalized linear mixed modeling. *R J.*, 9(2):378–400.
- 531 Burton, A. C., Neilson, E., Moreira, D., Ladle, A., Steenweg, R., Fisher, J. T., Bayne, E., and Boutin,
532 S. (2015). Wildlife camera trapping: a review and recommendations for linking surveys to ecological
533 processes. *Journal of Applied Ecology*, 52(3):675–685.

- 534 Cairns, S. J. and Schwager, S. J. (1987). A comparison of association indices. *Anim. Behav.*,
535 35(5):1454–1469.
- 536 Castellano, B. (2021). Pyscenedetect: <https://github.com/breakthrough/pyscenedetect>.
- 537 Castles, M., Heinsohn, R., Marshall, H. H., Lee, A. E. G., Cowlshaw, G., and Carter, A. J. (2014). Social
538 networks created with different techniques are not comparable. *Anim. Behav.*, 96:59–67.
- 539 Chen, P., Swarup, P., Matkowski, W. M., Kong, A. W. K., Han, S., Zhang, Z., and Rong, H. (2020). A
540 study on giant panda recognition based on images of a large proportion of captive pandas. *Ecol. Evol.*,
541 10(7):3561–3573.
- 542 Clutton-Brock, T. and Sheldon, B. C. (2010). Individuals and populations: the role of long-term,
543 individual-based studies of animals in ecology and evolutionary biology. *Trends Ecol. Evol.*,
544 25(10):562–573.
- 545 Croft, D. P., James, R., and Krause, J. (2008). *Exploring Animal Social Networks*. Princeton University
546 Press.
- 547 Davis, G. H., Crofoot, M. C., and Farine, D. R. (2018). Estimating the robustness and uncertainty of
548 animal social networks using different observational methods. *Anim. Behav.*, 141:29–44.
- 549 Dore, K. M., Hansen, M. F., Klegarth, A. R., Fichtel, C., Koch, F., Springer, A., Kappeler, P., Parga,
550 J. A., Humle, T., Colin, C., Raballand, E., Huang, Z.-P., Qi, X.-G., Di Fiore, A., Link, A., Stevenson,
551 P. R., Stark, D. J., Tan, N., Gallagher, C. A., Anderson, C. J., Campbell, C. J., Kenyon, M., Pebsworth,
552 P., Sprague, D., Jones-Engel, L., and Fuentes, A. (2020). Review of GPS collar deployments and
553 performance on nonhuman primates. *Primates*, 61(3):373–387.
- 554 Farine, D. R. (2015). Proximity as a proxy for interactions: issues of scale in social network analysis.
555 *Anim. Behav.*, 104(104):e1–e5.
- 556 Farine, D. R. (2018). When to choose dynamic vs. static social network analysis. *Journal of animal*
557 *ecology*, 87(1):128–138.

- 558 Farine, D. R. and Carter, G. G. (2022). Permutation tests for hypothesis testing with animal social
559 network data: Problems and potential solutions. *Methods in Ecology and Evolution*, 13(1):144–156.
- 560 Farine, D. R. and Whitehead, H. (2015). Constructing, conducting and interpreting animal social network
561 analysis. *J. Anim. Ecol.*, 84(5):1144–1163.
- 562 Feczko, E., Mitchell, T. A. J., Walum, H., Brooks, J. M., Heitz, T. R., Young, L. J., and Parr, L. A. (2015).
563 Establishing the reliability of rhesus macaque social network assessment from video observations. *Anim.*
564 *Behav.*, 107:115–123.
- 565 Ferreira, A. C., Covas, R., Silva, L. R., Esteves, S. C., Duarte, I. F., Fortuna, R., Theron, F., Doutrelant,
566 C., and Farine, D. R. (2020a). How to make methodological decisions when inferring social networks.
567 *Ecol. Evol.*, 10(17):9132–9143.
- 568 Ferreira, A. C., Silva, L. R., Renna, F., Brandl, H. B., Renoult, J. P., Farine, D. R., Covas, R., and
569 Doutrelant, C. (2020b). Deep learning-based methods for individual recognition in small birds. *Methods*
570 *Ecol. Evol.*, 11(9):1072–1085.
- 571 Finger, J. S., Guttridge, T. L., Wilson, A. D. M., Gruber, S. H., and Krause, J. (2017). Are some sharks
572 more social than others? short- and long-term consistencies in the social behavior of juvenile lemon
573 sharks. *Behav. Ecol. Sociobiol.*, 72(1):17.
- 574 Firth, J. A. and Sheldon, B. C. (2016). Social carry-over effects underpin trans-seasonally linked structure
575 in a wild bird population. *Ecol. Lett.*, 19(11):1324–1332.
- 576 Franks, D. W., Ruxton, G. D., and James, R. (2010). Sampling animal association networks with the
577 gambit of the group. *Behav. Ecol. Sociobiol.*, 64(3):493–503.
- 578 Gelardi, V., Godard, J., Paleressompouille, D., Claidiere, N., and Barrat, A. (2020). Measuring social
579 networks in primates: wearable sensors versus direct observations. *Proc. Math. Phys. Eng. Sci.*,
580 476(2236):20190737.
- 581 Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E. D.,

- 582 Robinson, B. S., Hodgson, D. J., and Inger, R. (2018). A brief introduction to mixed effects modelling
583 and multi-model inference in ecology. *PeerJ*, 6:e4794.
- 584 Hasenjager, M. J., Leadbeater, E., and Hoppitt, W. (2021). Detecting and quantifying social transmission
585 using network-based diffusion analysis. *Journal of Animal Ecology*, 90(1):8–26.
- 586 He, P., Klarevas-Irby, J. A., Papageorgiou, D., Christensen, C., Strauss, E. D., and Farine, D. R. (2022).
587 A guide to sampling design for gps-based studies of animal societies. *Methods in Ecology and Evolution*.
- 588 Hinde, R. A. (1976). Interactions, relationships and social structure. *Man*, 11(1):1–17.
- 589 Hobaiter, C., Poisot, T., Zuberbühler, K., Hoppitt, W., and Gruber, T. (2014). Social Network Analysis
590 Shows Direct Evidence for Social Transmission of Tool Use in Wild Chimpanzees. *PLoS Biol.*,
591 12(9):e1001960.
- 592 Hoppitt, W. J. and Farine, D. R. (2018). Association indices for quantifying social relationships: how to
593 deal with missing observations of individuals or groups. *Animal Behaviour*, 136:227–238.
- 594 Hospedales, T. M., Antoniou, A., Micaelli, P., and Storkey, A. J. (2021). Meta-Learning in neural
595 networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP.
- 596 Humle, T. (2011a). The 2003 epidemic of a flu-like respiratory disease at Bossou. In *The chimpanzees of*
597 *Bossou and Nimba*, pages 325–333. Springer.
- 598 Humle, T. (2011b). Location and ecology. In *The chimpanzees of Bossou and Nimba*, pages 13–21.
599 Springer.
- 600 Jacoby, D. M. P., Fear, L. N., Sims, D. W., and Croft, D. P. (2014). Shark personalities? repeatability of
601 social network traits in a widely distributed predatory fish. *Behav. Ecol. Sociobiol.*, 68(12):1995–2003.
- 602 James, R., Croft, D. P., and Krause, J. (2009). Potential banana skins in animal social network analysis.
603 *Behav. Ecol. Sociobiol.*, 63(7):989–997.
- 604 Krause, J., James, R., Franks, D. W., and Croft, D. P. (2015). *Animal Social Networks*. Oxford University
605 Press.

- 606 Krause, J. and Ruxton, G. D. (2002). *Living in Groups*. Oxford University Press.
- 607 Lacreuse, A., Raz, N., Schmidtke, D., Hopkins, W. D., and Herndon, J. G. (2020). Age-related decline
608 in executive function as a hallmark of cognitive ageing in primates: an overview of cognitive and
609 neurobiological studies. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 375(1811):20190618.
- 610 Langergraber, K., Mitani, J., and Vigilant, L. (2009). Kinship and social bonds in female chimpanzees
611 (pan troglodytes). *Am. J. Primatol.*, 71(10):840–851.
- 612 LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.
- 613 Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P., and Makowski, D. (2021). Performance: An R
614 package for assessment, comparison and testing of statistical models. *J. Open Source Softw.*, 6(60):3139.
- 615 Machanda, Z. P. and Rosati, A. G. (2020). Shifting sociality during primate ageing. *Philos. Trans. R.*
616 *Soc. Lond. B Biol. Sci.*, 375(1811):20190620.
- 617 Mathis, M. W. and Mathis, A. (2020). Deep learning tools for the measurement of animal behavior in
618 neuroscience. *Current opinion in neurobiology*, 60:1–11.
- 619 Matsuzawa, T. (1994). Field experiments on use of stone tools by chimpanzees in the wild. In de Waal F.
620 B. M., Wrangham R. W., McGrew W. C. , Heltne P. G., editor, *Chimpanzee cultures*, pages 351–370.
621 Cambridge, MA: Harvard University Press.
- 622 Matsuzawa, T., Humle, T., and Sugiyama, Y. (2011). *The chimpanzees of Bossou and Nimba*. Springer
623 Science & Business Media.
- 624 McCann, R., Bracken, A. M., Christensen, C., Fürtbauer, I., and King, A. J. (2021). The relationship
625 between GPS sampling interval and estimated daily travel distances in chacma baboons (*papio ursinus*).
626 *Int. J. Primatol.*, 42(4):589–599.
- 627 McCarthy, M. S., Després-Einspenner, M.-L., Farine, D. R., Samuni, L., Angedakin, S., Arandjelovic, M.,
628 Boesch, C., Dieguez, P., Haverkamp, K., Knight, A., Langergraber, K. E., Wittig, R. M., and Kühl,
629 H. S. (2019). Camera traps provide a robust alternative to direct observations for constructing social
630 networks of wild chimpanzees. *Anim. Behav.*, 157:227–238.

- 631 McKinney, W. (2011). pandas: a foundational python library for data analysis and statistics. *Python for*
632 *high performance and scientific computing*, 14(9):1–9.
- 633 Miele, V., Dussert, G., Spataro, B., Chamaillé-Jammes, S., Allainé, D., and Bonenfant, C. (2021).
634 Revisiting animal photo-identification using deep metric learning and network analysis. *Methods in*
635 *Ecology and Evolution*, 12(5):863–873.
- 636 Mitani, J. C. (2009). Male chimpanzees form enduring and equitable social bonds. *Anim. Behav.*,
637 77(3):633–640.
- 638 Mitani, J. C., Merriwether, D. A., and Zhang, C. (2000). Male affiliation, cooperation and kinship in
639 wild chimpanzees. *Anim. Behav.*, 59(4):885–893.
- 640 Mitani, J. C. and Watts, D. P. (2001). Why do chimpanzees hunt and share meat? *Anim. Behav.*,
641 61(5):915–924.
- 642 Nakagawa, S. and Schielzeth, H. (2010). Repeatability for gaussian and non-gaussian data: a practical
643 guide for biologists. *Biol. Rev. Camb. Philos. Soc.*, 85(4):935–956.
- 644 Nath, T., Mathis, A., Chen, A. C., Patel, A., Bethge, M., and Mathis, M. W. (2019). Using DeepLabCut
645 for 3D markerless pose estimation across species and behaviors. *Nat. Protoc.*, 14(7):2152–2176.
- 646 Norouzzadeh, M. S., Morris, D., Beery, S., Joshi, N., Jovic, N., and Clune, J. (2021). A deep active
647 learning system for species identification and counting in camera trap images. *Methods Ecol. Evol.*,
648 12(1):150–161.
- 649 O’Connell, A. F., Nichols, J. D., and Ullas Karanth, K. (2010). *Camera Traps in Animal Ecology:*
650 *Methods and Analyses*. Springer Science & Business Media.
- 651 Oliphant, T. E. (2007). Python for scientific computing. *Computing in Science Engineering*, 9(3):10–20.
- 652 Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N.,
653 Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S.,
654 Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). PyTorch: An imperative style, High-Performance
655 deep learning library. In *Advances in Neural Information Processing Systems*, volume 32.

- 656 R Core Team (2021). R: A language and environment for statistical computing. *R Foundation for*
657 *Statistical Computing, Vienna, Austria.*
- 658 Rosati, A. G., Hagberg, L., Enigk, D. K., Otali, E., Emery Thompson, M., Muller, M. N., Wrangham,
659 R. W., and Machanda, Z. P. (2020). Social selectivity in aging wild chimpanzees. *Science*,
660 370(6515):473–476.
- 661 Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods*
662 *Ecol. Evol.*, 1(2):103–113.
- 663 Schofield, D., Nagrani, A., Zisserman, A., Hayashi, M., Matsuzawa, T., Biro, D., and Carvalho, S. (2019).
664 Chimpanzee face recognition from videos in the wild using deep learning. *Sci Adv*, 5(9):eaaw0736.
- 665 Schofield, D. P., Albery, G. F., Firth, J. A., Mielke, A., Hayashi, M., Matsuzawa,
666 T., Biro, D., and Carvalho, S. (2023). dschof/deepprimatenetworksmee. zenodo.
667 <https://doi.org/10.5281/zenodo.8063792>.
- 668 Sheldon, B. C., Kruuk, L. E., and Alberts, S. C. (2022). The expanding value of long-term studies of
669 individuals in the wild. *Nature Ecology & Evolution*, pages 1–3.
- 670 Shizuka, D. and Johnson, A. E. (2019). How demographic processes shape animal social networks. *Behav.*
671 *Ecol.*, 31(1):1–11.
- 672 Silk, J. B. and Kappeler, P. M. (2017). Sociality in primates. In D. R. Rubenstein, P. A., editor,
673 *Comparative Social Evolution*, pages 253–283. Cambridge University Press Cambridge, UK.
- 674 Silk, M. J., Jackson, A. L., Croft, D. P., Colhoun, K., and Bearhop, S. (2015). The consequences of
675 unidentifiable individuals for the analysis of an animal social network. *Anim. Behav.*, 104:1–11.
- 676 St Clair, J. J. H., Burns, Z. T., Bettaney, E. M., Morrissey, M. B., Otis, B., Ryder, T. B., Fleischer,
677 R. C., James, R., and Rutz, C. (2015). Experimental resource pulses influence social-network dynamics
678 and the potential for information flow in tool-using crows. *Nat. Commun.*, 6:7197.
- 679 Strandburg-Peshkin, A., Farine, D. R., Couzin, I. D., and Crofoot, M. C. (2015). Shared decision-making
680 drives collective movement in wild baboons. *Science*, 348(6241):1358–1361.

- 681 Sueur, C., Jacobs, A., Amblard, F., Petit, O., and King, A. J. (2011). How can social network analysis
682 improve the study of primate behavior? *American Journal of Primatology*, 73(8):703–719.
- 683 Sugiyama, Y. (1984). Population dynamics of wild chimpanzees at Bossou, Guinea, between 1976 and
684 1983. *Primates*, 25(4):391–400.
- 685 Sugiyama, Y. (1988). Grooming interactions among adult chimpanzees at Bossou, Guinea, with special
686 reference to social structure. *International Journal of Primatology*, 9:393–407.
- 687 Sugiyama, Y. (2004). Demographic parameters and life history of chimpanzees at Bossou, Guinea.
688 *American Journal of Physical Anthropology: The Official Publication of the American Association*
689 *of Physical Anthropologists*, 124(2):154–165.
- 690 Sugiyama, Y. and Fujita, S. (2011). The demography and reproductive parameters of Bossou chimpanzees.
691 In Matsuzawa, T., Humle, T., and Sugiyama, Y., editors, *The Chimpanzees of Bossou and Nimba*, pages
692 23–34. Springer Japan, Tokyo.
- 693 Sugiyama, Y. and Koman, J. (1979). Tool-using and -making behavior in wild chimpanzees at Bossou,
694 Guinea. *Primates*, 20(4):513–524.
- 695 Tkaczynski, P. J., Mielke, A., Samuni, L., Preis, A., Wittig, R. M., and Crockford, C. (2020). Long-term
696 repeatability in social behaviour suggests stable social phenotypes in wild chimpanzees. *R Soc Open*
697 *Sci*, 7(8):200454.
- 698 Vallejo Seco, G., Ato García, M., Fernández García, M. P., and Livacic Rojas, P. E. (2013). Multilevel
699 bootstrap analysis with assumptions violated. *Psicothema*, 25(4):520–528.
- 700 Vidal, M., Wolf, N., Rosenberg, B., Harris, B. P., and Mathis, A. (2021). Perspectives on individual animal
701 identification from biology and computer vision. *Integrative and Comparative Biology*, 61(3):900–916.
- 702 Wang, W., Neuman, E. J., and Newman, D. A. (2014). Statistical power of the social network
703 autocorrelation model. *Soc. Networks*, 38:88–99.
- 704 Webster, M. M. and Rutz, C. (2020). How STRANGE are your study animals? *Nature*,
705 582(7812):337–340.

- 706 Weinstein, B. G. (2018). A computer vision for animal ecology. *J. Anim. Ecol.*, 87(3):533–545.
- 707 Whitehead, H. (2008). *Analyzing Animal Societies: Quantitative Methods for Vertebrate Social Analysis*.
708 University of Chicago Press.
- 709 Wuerz, Y. and Krüger, O. (2015). Personality over ontogeny in zebra finches: long-term repeatable traits
710 but unstable behavioural syndromes. *Front. Zool.*, 12 Suppl 1:S9.
- 711 Yamakoshi, G. (1998). Dietary responses to fruit scarcity of wild chimpanzees at Bossou, Guinea: possible
712 implications for ecological importance of tool use. *American Journal of Physical Anthropology: The*
713 *Official Publication of the American Association of Physical Anthropologists*, 106(3):283–295.

714 Supplementary

Table S1. Information on videos for each year in the dataset. ‘Total observation time’ refers to the hours of video data sampled for each year and ‘mean duration’ refers to the the average length of the video clips.

Year	Videos	Total Observation Time (hrs)	Total Observation Time (mins)	Mean Video Duration (mins)	Individuals Detected	Detections
1996	5	9.6	576.05	115.21	17	917267
1998	5	5.22	313.04	62.61	19	237708
1998-1999	5	5.11	306.43	61.29	19	339924
2000	5	5.21	312.35	62.47	19	421945
2002	5	6.34	380.3	76.06	16	340489
2002-2003	6	5.9	353.81	58.97	15	340743
2003-2004	5	5.77	346.13	69.23	15	349485
2004-2005	5	6.87	411.98	82.4	12	576859
2005-2006	5	6.64	398.15	79.63	12	711635
2007	6	8.18	491.06	81.84	11	260948
2008-2009	5	6.57	393.96	78.79	13	670218
2010	5	6.86	411.31	82.26	12	406441
2012	36	5.19	311.25	8.65	13	315879
2013	99	5.17	310.24	3.13	9	378089
2015	6	3.04	182.29	30.38	9	203513
2016	39	5.47	328.44	8.42	9	249140
2017	18	4.7	282.18	15.68	8	358271

Figure S1. Kernel density plot for distribution of individual-level Strength values

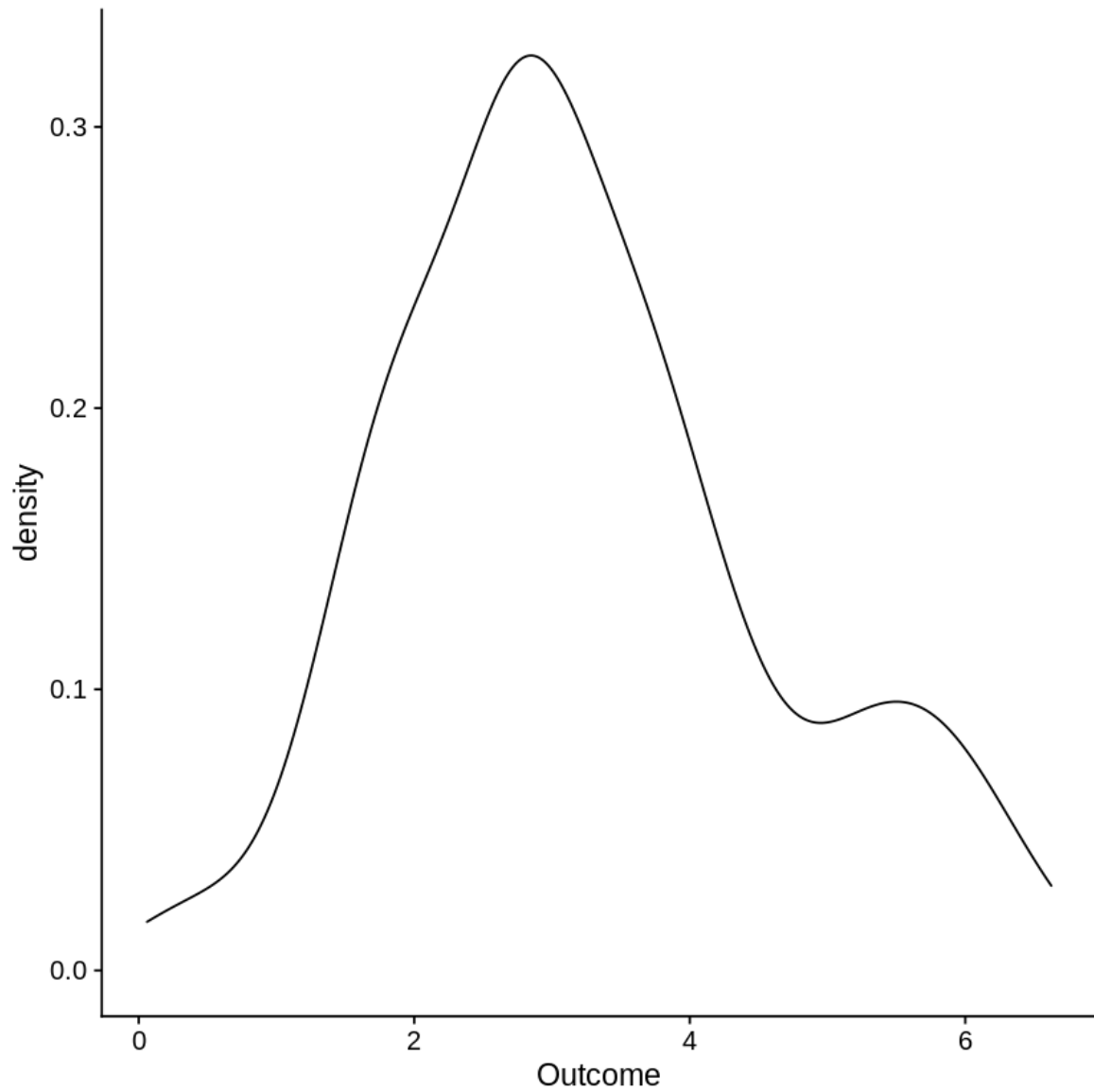


Figure S2. Output from normality checks using the ‘check_model’ function in the ‘performance package’ in R (Lüdtke et al. 2021) for the individual Strength variable.

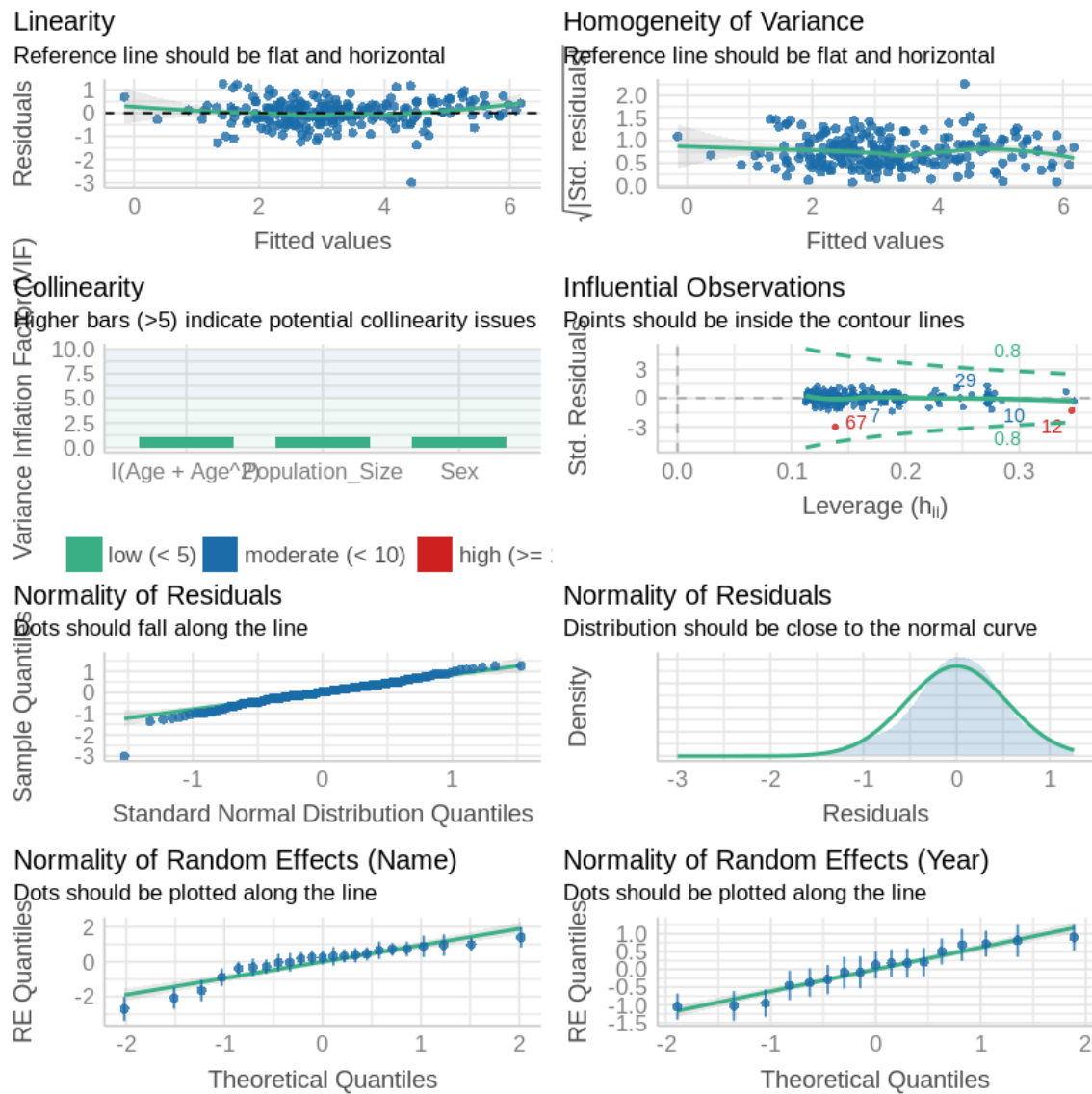


Figure S3. Kernel density plot for distribution of individual-level Mean Group Size values.

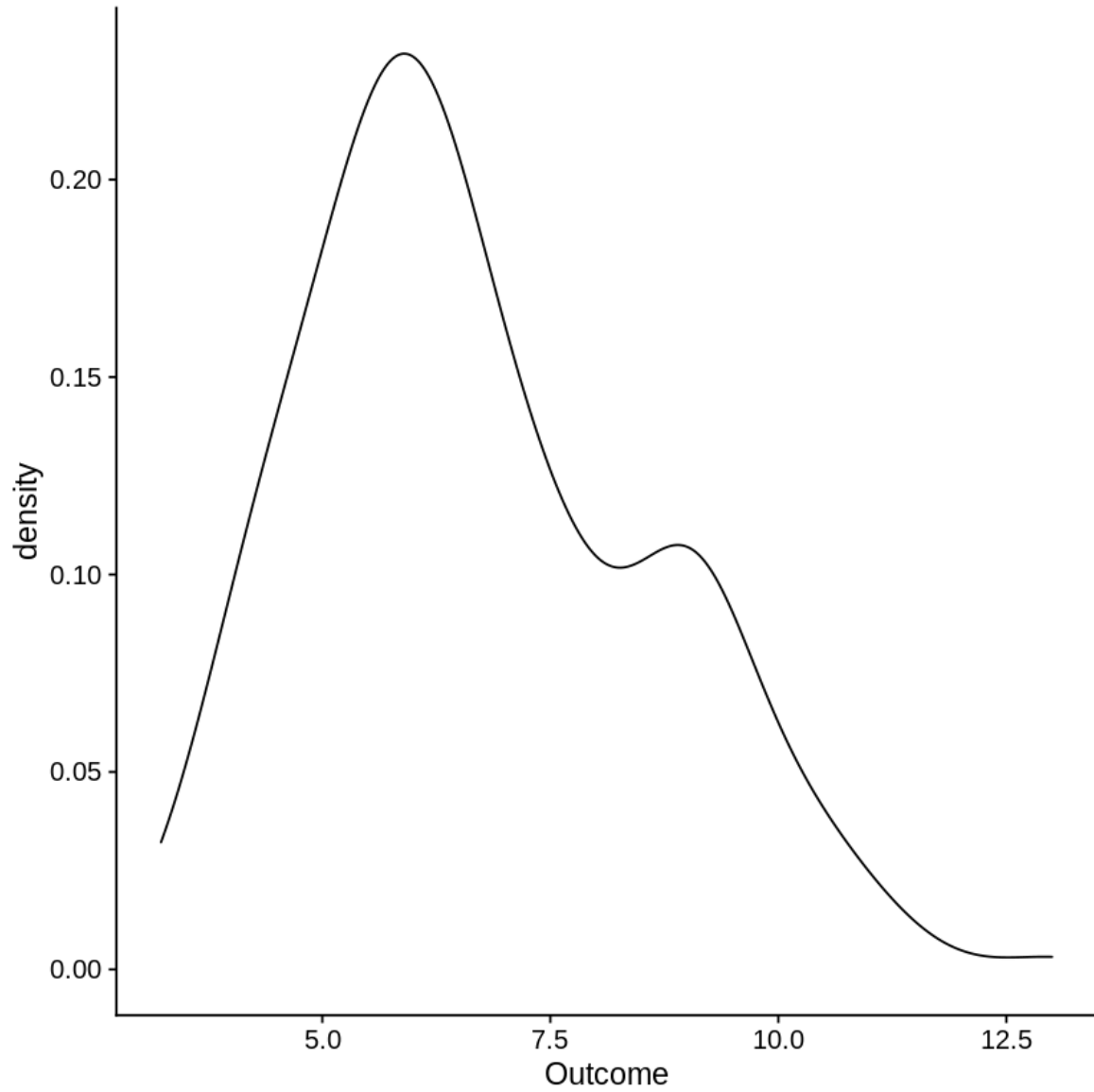


Figure S4. Output from normality checks using the ‘check_model’ function in the ‘performance’ package in R (Lüdtke et al. 2021) for the Mean Group Size variable.

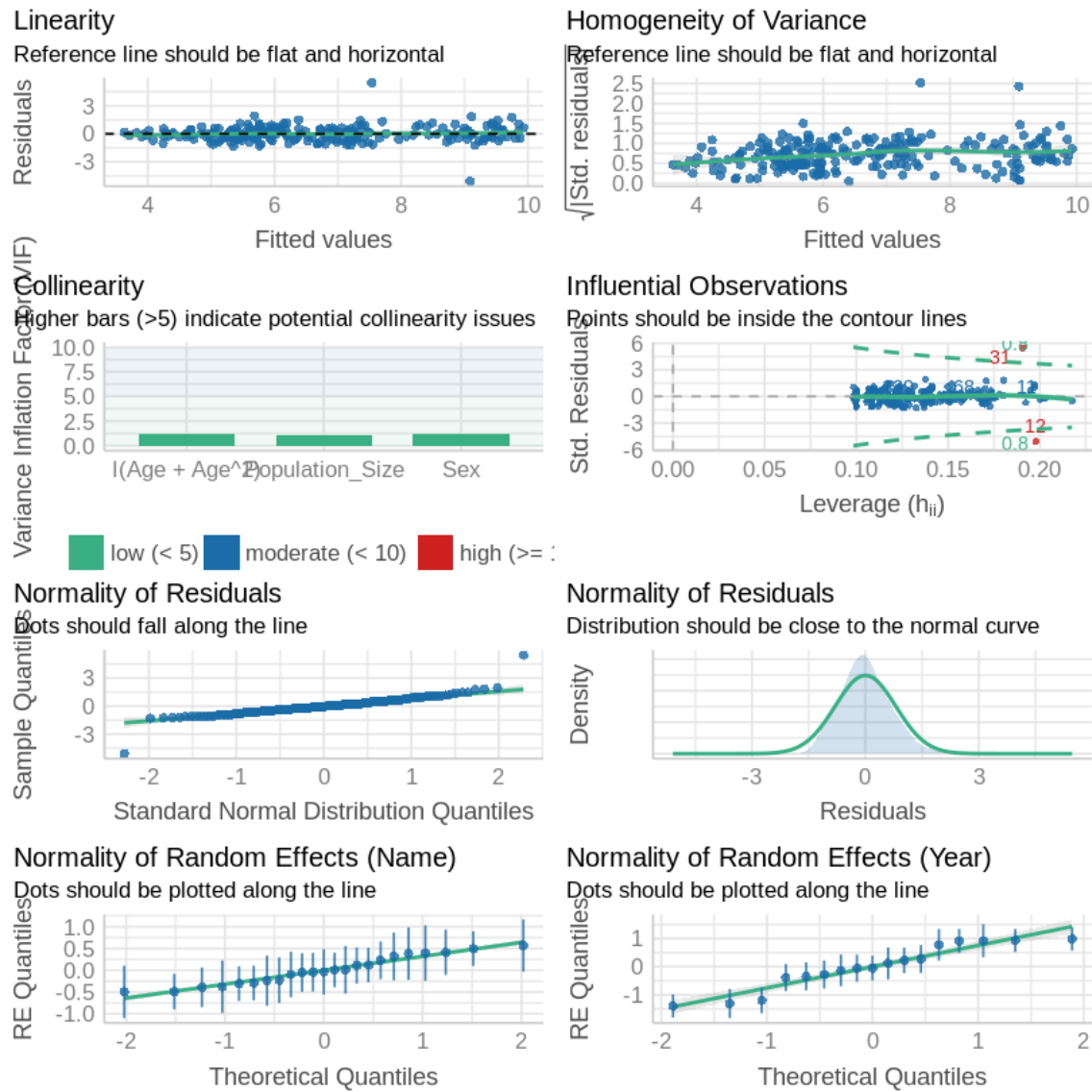


Figure S5. Kernel density plot for distribution of individual-level ‘Time Alone’ values.

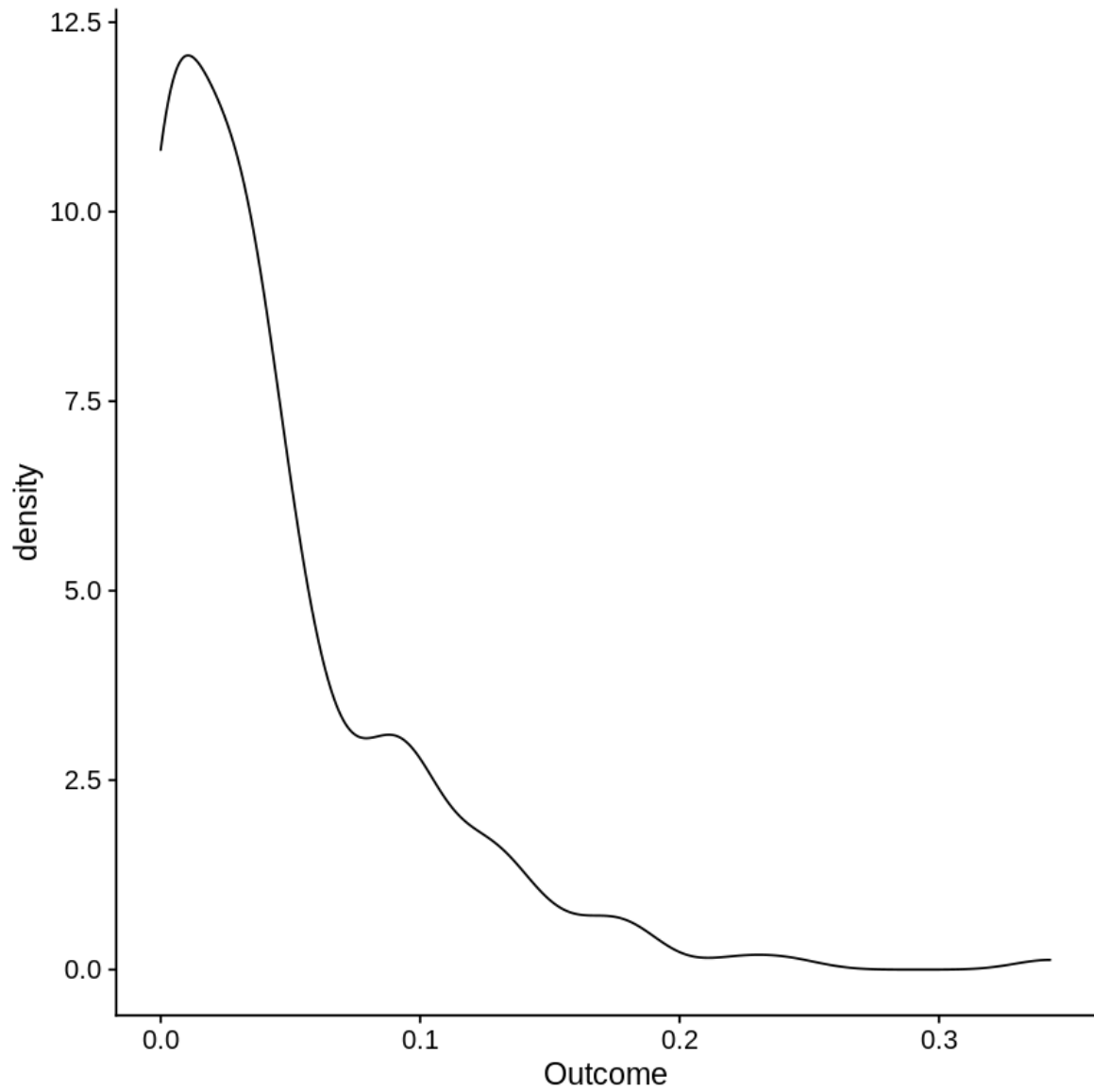
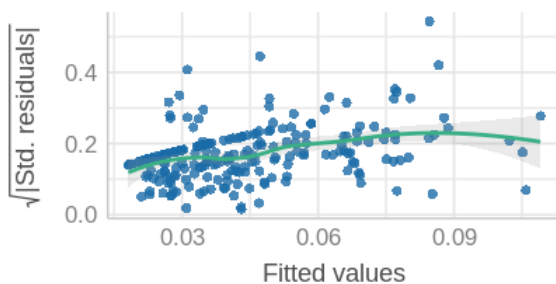


Figure S6. Output from normality checks using the ‘check_model’ function in the ‘performance’ package in R (Lüdtke et al. 2021) for the individual-level Time Alone variable. Normality checks of Time Alone reveal that the assumption of normality has been violated with a slightly longer tail than would be expected from a normally distributed variable and beta distribution, indicating our model is systematically underestimating the upper tail end of our distribution (see distribution of predicted values below). However this may be due to the nature of our dataset which has some outliers/inflated proportion scores, above that which would be expected for a chimpanzee (i.e above 10% of their time spent alone). Furthermore it has been demonstrated that violation of normality assumptions does not significantly affect regression estimates for mixed models (Vallejo Seco et al. 2013). Although the reference line for homogeneity of variance is not perfectly horizontal, overall there is an equal spread around this line indicating that the amount of heteroskedascity is unproblematic.

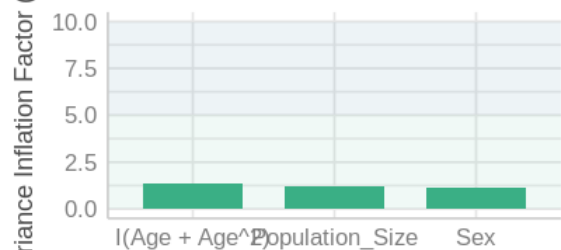
Homogeneity of Variance

Reference line should be flat and horizontal



Collinearity

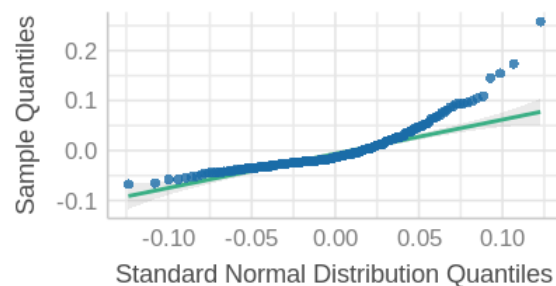
Higher bars (>5) indicate potential collinearity issues



low (< 5) moderate (< 10) high (>= 10)

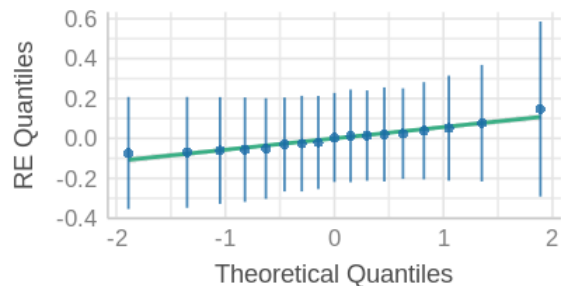
Normality of Residuals

Dots should fall along the line



Normality of Random Effects (Year)

Dots should be plotted along the line



Normality of Random Effects (Name)

Dots should be plotted along the line

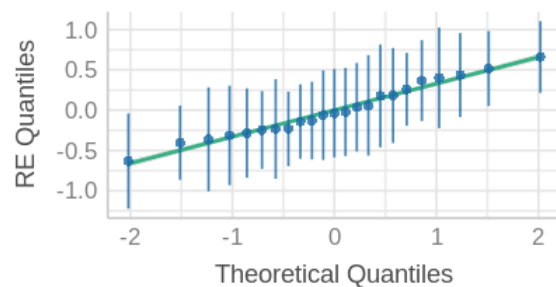


Figure S7. Bossou Age distribution (scaled to be within 0 and 1) by Sex. Across years there is a small proportion of middle aged individuals, with a large core of elderly females.

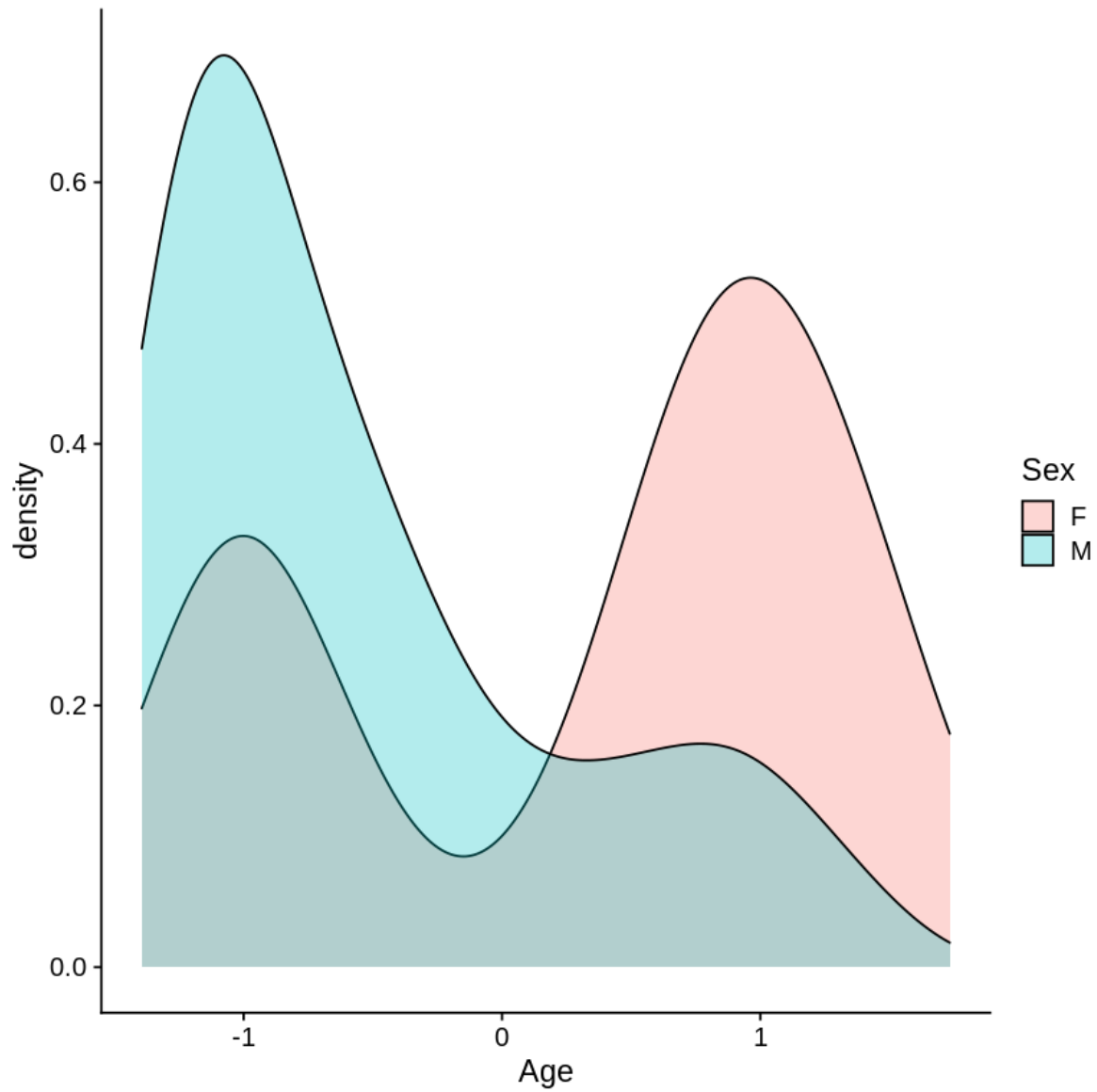


Figure S8. Model outputs from data generated using intermediate 120 second sampling/aggregation rate, visualising the effect of Age on each outcome variable (left column). Corresponding forest plots (right column) showing the effect of each fixed effect (SexM - effect of Males compared to Females), Age (fitted as a quadratic term) and Population Size (as a controlled fixed effect). Point estimates denote the effect size between each variable on outcome variable, error bars the 95% confidence intervals, and asterisks statistical significance (with estimated confidence intervals not overlapping with zero). Model parameters are detailed in Table 1 and 2. Note that visualised model fit for Age for Strength and Mean Group Size (left) are non-significant. To assess whether variables of interest (Age and Sex) significantly impact our model's predictive ability, check significance of likelihood ratio tests which are detailed in Table 2.

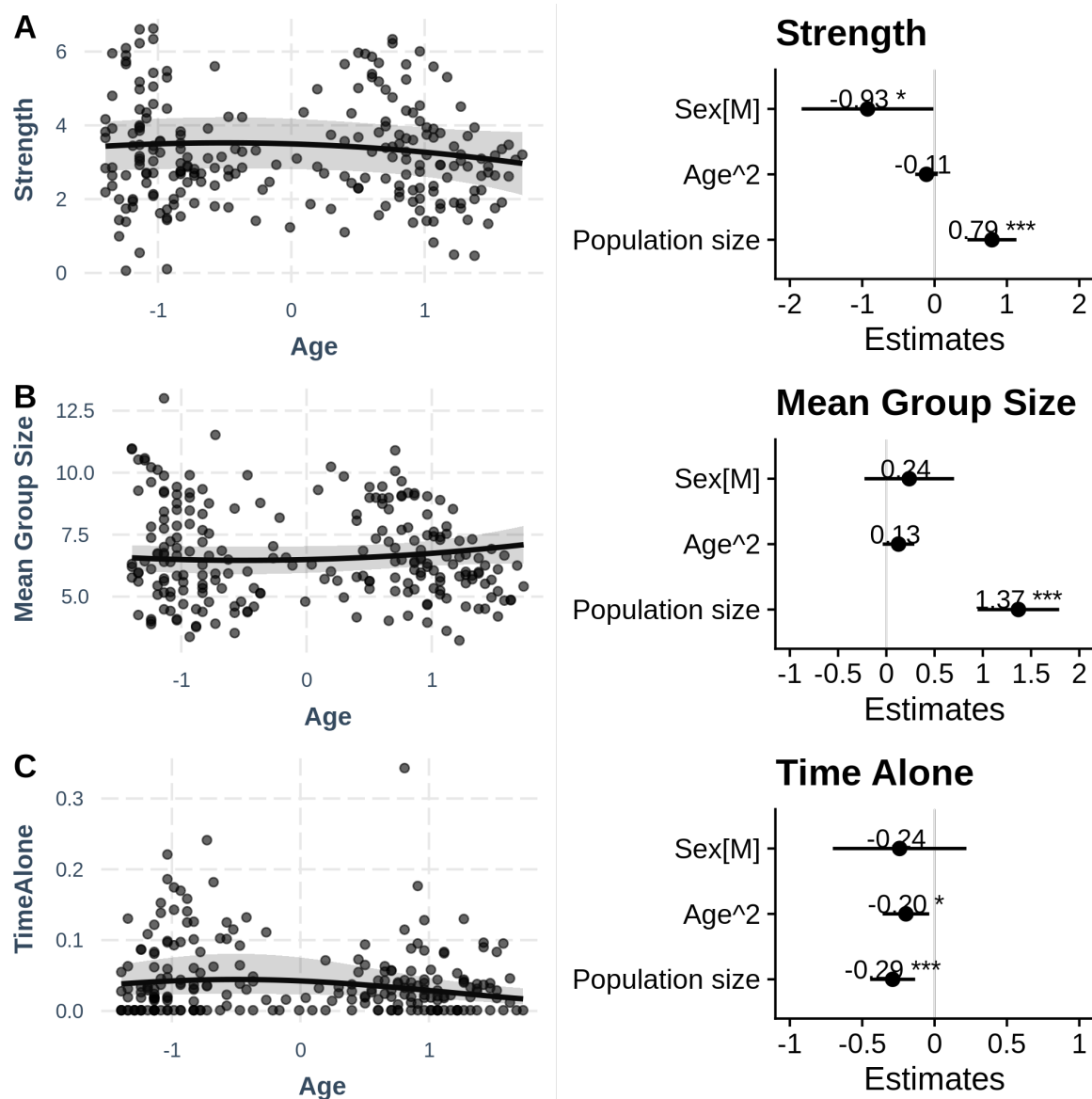


Figure S9. A zoom shot of two old females at the outdoor lab in 2012 eating palm fruit (left: Velu, right: Yo) (photo credit: Tetsuro Matsuawa).



715 **Social and ecological context at Bossou**

716 Bossou represents a unique case study due to its local ecology and demographics (Humle 2011b). The
717 population is geographically isolated, and there is high resource availability (Yamakoshi 1998). With an
718 absence of inter-community competition and low rates of population turnover, the group has remained
719 stable over time, with a core membership of 7 females and no more than 4 adult males at one time; no
720 female immigrants have joined the community since research started in 1976 (Sugiyama 2004, Sugiyama
721 and Fujita 2011). Contrary to other communities which are characterised by volatile male-male interactions
722 (Langergraber et al. 2009, Mitani 2009, Mitani et al. 2000, Mitani and Watts 2001), Bossou females
723 play a central role in maintaining peace, and grooming is more common than in other communities
724 (Sugiyama 1988). Typically, chimpanzees show a high degree of fission-fusion, whereby groups break
725 off into subgroups or ‘parties’. The high density and repeatability of individual social network positions
726 found in our study supports the idea that the Bossou community may have a higher level of gregariousness
727 than other populations (Sugiyama 1984). At Bossou, a large proportion of the group eats, travels and
728 rests together, and there is less pressure to split into separate parties due to the small and declining
729 population size, especially in later years after a disease epidemic in 2003 (Humle 2011a).

730 **Sociality over the lifespan**

731 In contrast to research which shows sex differences and declines in social network centrality and gregariousness
732 in primates (Brent et al. 2017, Machanda and Rosati 2020), we found no impact of age and sex on
733 individuals’ social network positions and mean group size over time. Research at Kibale National Park
734 found that solitariness increased with age in adult males (Rosati et al. 2020). Using our approach we
735 detected a weak quadratic effect of age on solitariness - in the social context of nut cracking at Bossou,
736 solitariness appeared to show a trend towards increasing over infancy, peaking in middle age, and declining
737 as older chimpanzees spent less time alone at the outdoor laboratory (Fig. **S8C**) - however, given this
738 detected effect was weak, further data is needed to explore whether this result is consistent with a wider
739 pattern specific to the Bossou community. Individual differences could be related to cognitive decline
740 and physical frailty associated with the ageing demographic at Bossou (Fig. **S9, S7**) or due to other

741 aspects such as social dominance which allows individuals to access nuts and preferred tools ahead of the
742 group. The use of our automated framework allows refinement of such fine-scale association patterns and
743 future work should examine how markers of cognitive decline and physical ageing can also be drivers of
744 variation in social networks and individual-level gregariousness across the lifespan.

5

Automated audiovisual behavior recognition in wild primates

Published in *Science Advances*

ORGANISMAL BIOLOGY

Automated audiovisual behavior recognition in wild primates

Max Bain^{1*}, Arsha Nagrani^{1†}, Daniel Schofield², Sophie Berdugo^{2,3}, Joana Bessa⁴, Jake Owen⁴, Kimberley J. Hockings⁵, Tetsuro Matsuzawa⁶, Misato Hayashi[§], Dora Biro^{4,7}, Susana Carvalho^{2,8,9,10}, Andrew Zisserman¹

Large video datasets of wild animal behavior are crucial to produce longitudinal research and accelerate conservation efforts; however, large-scale behavior analyses continue to be severely constrained by time and resources. We present a deep convolutional neural network approach and fully automated pipeline to detect and track two audiovisually distinctive actions in wild chimpanzees: buttress drumming and nut cracking. Using camera trap and direct video recordings, we train action recognition models using audio and visual signatures of both behaviors, attaining high average precision (buttress drumming: 0.87 and nut cracking: 0.85), and demonstrate the potential for behavioral analysis using the automatically parsed video. Our approach produces the first automated audiovisual action recognition of wild primate behavior, setting a milestone for exploiting large datasets in ethology and conservation.

INTRODUCTION

The field of ethology seeks to understand animal behavior from both mechanistic and functional perspectives and to identify the various genetic, developmental, ecological, and social drivers of behavioral variation in the wild (1). It is increasingly becoming a data-rich science: Technological advances in data collection, including biologists, camera traps, and audio recorders, now allow us to capture animal behavior in an unprecedented level of detail (2). In particular, large data archives including both audio and visual information have immense potential to measure individual- and population-level variation as well as ontogenetic and cultural changes in behavior that may span large temporal and spatial scales. However, this potential often goes untapped: The training and human effort required to process large volumes of video data continue to limit the scale and depth at which behavior can be analyzed. Automating the measurement of behavior can transform ethological research, open up large-scale video archives for detailed interrogation, and be a powerful tool to monitor and protect threatened species in the wild. With rapid advances in deep learning, the novel field of computational ethology is quickly emerging at the intersection of computer science, engineering, and biology, using computer vision algorithms to process large volumes of data (3).

The aim of this paper is to automate animal behavior recognition in wild footage. Deep learning-based behavior recognition has thus far been shown in constrained laboratory settings (4, 5) or using still images (6) and has yet to be effectively demonstrated on unconstrained video footage recorded in the wild. Measuring animal behavior from wild footage presents substantial challenges—often, behaviors are hard to detect, obscured by motion blur, occlusion, vegetation, poor resolution, or lighting. If successful, then the tools would enable exploration of a multitude of research questions in ethology and conservation. Increasingly, research is revealing fine-scale variation between individuals and populations of wild animals (7); however, capturing this variation is often laborious and not feasible on the large scale through manual annotation. Automated approaches allow us to examine in more detail the variation, through cross comparison of animal groups in a wide variety of contexts. Detailed time series data of individual behavior enables integration of time depth perspectives into field research to more comprehensively reconstruct how behavior develops across the life span (ontogenetically) as well as examine how other processes such as social transmission, demography, and ecology interact to drive behavior change over time (8). These detailed behavioral data are also a crucial component of conservation research: They enable us to investigate how anthropogenic pressures such as climate change and habitat fragmentation disrupt animal behavior (9) (migratory patterns, foraging, reproduction, etc.) and to develop novel behavioral metrics to monitor the risks to and viability of threatened populations (10, 11). Here, we demonstrate the potential of such an approach by developing a system for the automated classification of two distinct wild chimpanzee behaviors with idiosyncratic audiovisual features: nut cracking and buttress drumming. We also analyze pilot data of sex and age differences in percussive behaviors (nut cracking and drumming) from longitudinal archive and camera trap datasets. Chimpanzees are an ideal species for testing behavioral recognition; owing to their large fission-fusion societies, complex sociality, and behavioral flexibility, they exhibit exceptionally rich behavioral repertoires (12). Our target behaviors, nut cracking and buttress drumming, differ in their function—extractive tool use versus long-distance communication, respectively—but both involve

¹Visual Geometry Group, Department of Engineering Science, University of Oxford, Oxford, UK. ²Primate Models for Behavioural Evolution Lab, Institute of Human Sciences, School of Anthropology and Museum Ethnography, University of Oxford, Oxford, UK. ³Social Body Lab, Institute of Human Sciences, School of Anthropology and Museum Ethnography, University of Oxford, Oxford, UK. ⁴Department of Zoology, University of Oxford, Oxford, UK. ⁵Centre for Ecology and Conservation, College of Life and Environmental Sciences, University of Exeter, Exeter, UK. ⁶Division of the Humanities and Social Sciences, California Institute of Technology, 1200 E. California Blvd., MC 228-77, Pasadena, CA 91125, USA. ⁷Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA. ⁸Gorongosa National Park, Sofala, Mozambique. ⁹Centre for Functional Ecology, Department of Life Sciences, Coimbra University, Coimbra, Portugal. ¹⁰Interdisciplinary Centre for Archaeology and Evolution of Human Behaviour, Algarve University, Faro, Portugal. *Corresponding author. Email: maxbain@robots.ox.ac.uk

†Now at Google AI Research.

§Present address: Chubu Gakuin University, 30-1 Naka Oida-cho, Kakamigahara, Gifu 504-0837, Japan.

§Present address: Japan Monkey Centre, Kanrin-26, Inuyama, Aichi 484-0081, Japan.

Copyright © 2021
The Authors, some
rights reserved;
exclusive licensee
American Association
for the Advancement
of Science. No claim to
original U.S. Government
Works. Distributed
under a Creative
Commons Attribution
NonCommercial
License 4.0 (CC BY-NC).

Downloaded from https://www.science.org on December 18, 2021

percussive actions that produce distinctive sounds, i.e., the pounding of a hammer stone against a nut balanced on an anvil stone and the pounding of hands or feet against large buttress roots. Whereas nut cracking is limited to some West African and Cameroon chimpanzees (*Pan troglodytes verus* and *Pan troglodytes ellioti*), buttress drumming is a universal behavior across all chimpanzee communities (12).

In relation to previous works using deep learning, individual reidentification has been a critical first step toward full automation (13, 14), but this alone cannot capture the full complexity of behaviors that animals perform in the wild across space and time. Existing methods have used deep learning for markerless pose estimation to track the movement of animal body parts (15), but pose estimation models perform poorly at recognizing actions using posture and limb movements alone (16). Other approaches have used single-image analysis to identify basic activities of wild animals using tagged information from camera traps, but these fail to capture the dynamic sequences of behavior required for detailed analysis (17). Recent advances in human action recognition in the field of computer vision have used three-dimensional (3D) convolutional neural networks (CNNs) (18), which incorporate spatiotemporal information across video frames (19), but thus far have only been applied to animal species to produce broad behavioral classification limited to the visual domain (20).

Given that both behaviors have strong audio and visual signatures, we recognize actions using both audio and visual streams. Our automatic framework consists of two stages: (i) body detection and tracking of individuals through the video (localization in space and time) and (ii) audiovisual action recognition (Fig. 1 and movie S1). Audio allows us to determine temporal segments where the nut cracking and buttress drumming occur (“scene level”) but does not pinpoint the individual responsible. By visually detecting and tracking all chimpanzees that appear in the video, frame by frame, we are able to determine the spatial position of each individual present.

The next stage of our framework uses both the scene level audio and the visual content of each track to specify which individual is performing the behavior (“individual level”). Both stages in our pipeline use a deep CNN model (see Materials and Methods).

The audio stream can also be used to provide a preview mechanism to filter out behavioral sequences for human annotators to label (21), substantially reducing the time required to collect annotations. This is achieved using an audio-only action recognition model (which operates at the scene level) and can identify “proposals” or short video sequences where the action is likely to occur. A human annotator then only verifies whether the sequence contains the action or not. This allows us to efficiently create a labeled action recognition dataset that can be used to train the second stage of our automatic pipeline.

Our method is able to identify where fine-grained movements such as striking and drumming are occurring in time and space automatically. It consists of a deep CNN model, which predicts actions using audio only, visual only, and both audio and visual modalities together. We demonstrate the use of our pipeline on two different data sources: For nut cracking, we use part of a longitudinal video archive recorded by human-operated camcorders at an “outdoor laboratory” in Bossou, Guinea (14, 22); while for buttress drumming, data were collected between 2017 and 2019 by 25 motion triggered cameras in Cantanhez National Park, Guinea-Bissau (23). Last, we also demonstrate possible next steps in behavioral analysis enabled by the automatically parsed video. This approach represents the first automated audiovisual action recognition of species in the wild.

RESULTS

For nut cracking, we apply our pipeline to 40.2 hours of video containing 2448 nut cracking sequences (see Materials and Methods for definition of a sequence), resulting in a total of 24,700 individual body tracks (linked detections through video frames of the same

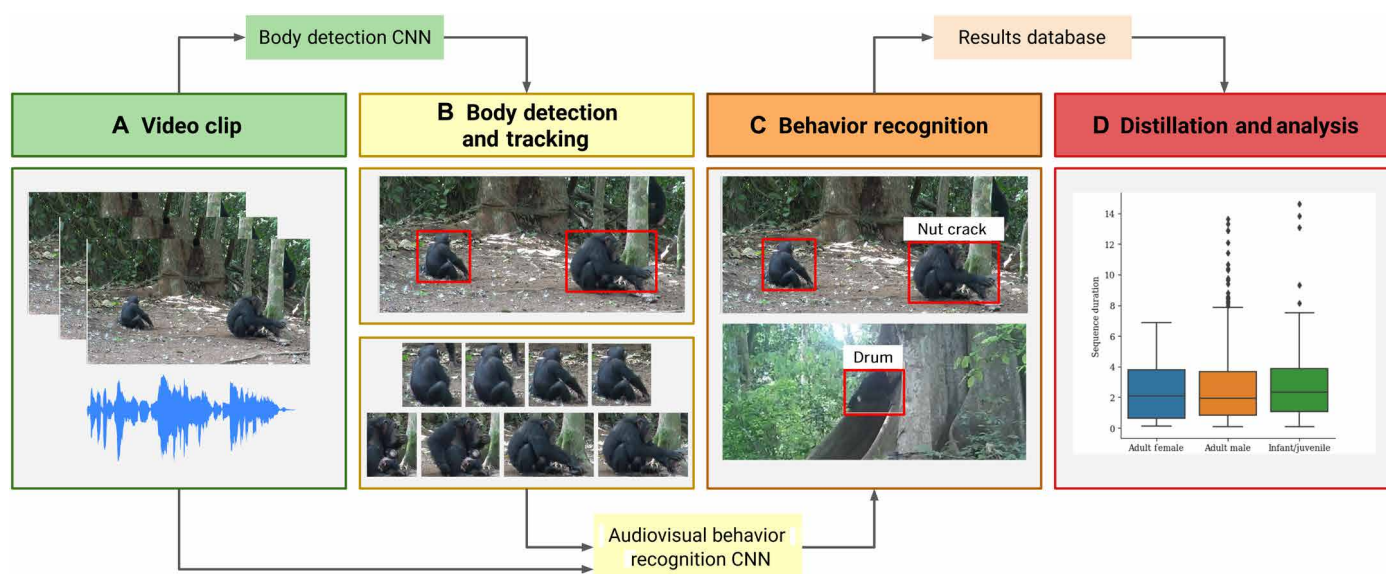


Fig. 1. Fully unified pipeline for wild chimpanzee behavior recognition and analysis from raw video footage. The pipeline consists of the following stages: (A) Frames and audio are extracted from raw video. (B) Body detection is performed over the video frames using a deep CNN single-shot detector (SSD) model, and the detections are tracked using a Siamese tracker. (C) The body tracks are classified (e.g., is this individual cracking nuts?) using the audio data and spatiotemporal visual information for the track by a deep CNN audiovisual behavior model. The system only requires the raw video as input and produces labeled body tracks and metadata as temporal and spatial information. This automated system can be used to perform large-scale analysis (D) of behavior. Photo credit: Kyoto University, Primate Research Institute.

individual; Fig. 1C and fig. S1B). The training set for our model consists of data taken from three years (2004, 2008, and 2012), while we test the performance of our model entirely on data from a different year (2013) to demonstrate generalizability over time. Our audio-only nut cracking recognition CNN model obtains high average precision (85%; Table 1) at the scene level. Results at a scene level only detect time periods where nut cracking is being performed in the video, but they do not isolate the nut cracker, given that multiple individuals may be nut-cracking in the video at the same time (Fig. 2). We also predict results at an individual level, identifying whether a particular individual is nut-cracking or not. Our chimpanzee body detector achieved an average precision of 92% (fig. S1), and our nut cracking recognition model performed well on different poses and lighting conditions typical of videos recorded in the wild (Fig. 2 and movie S1), achieving an overall average precision of 77% at an individual level (Table 1 and Fig. 3).

For buttress drumming, 10.8 hours of camera trap footage are analyzed, resulting in a total of 1251 drumming sequences. We trained our model on data from two chimpanzee communities (Cabante and Caiquene-Cadique) and evaluated our model on manually labeled held-out test data from a third community (Lautchandé). Data from an additional community (Cambeque) are included in the analysis. Our drumming recognition CNN model achieved 87% average precision at a scene level (using audio only) and 86% average precision at an individual level (Table 1 and Fig. 3).

To demonstrate the potential applications of this framework, we used the output of our automatic pipeline to further characterize nut cracking and buttress drumming behaviors. For nut cracking, we trained a visual classifier to identify eating events: This model followed the protocols of the visual-only drumming and nut cracking classifiers and sought to identify instances when food was passed from hand to mouth (an indication of successful nut cracking). Given that an individual typically eats in conjunction with nut cracking, our audio prescreening narrowed down the search space, allowing us to efficiently label 896 body tracks of individuals consuming nuts. This enabled us to analyze, as a function of age/sex class, the average time spent nut cracking per eating event (a proxy for the number of nuts successfully cracked and consumed) (Fig. 4). For buttress drumming, we automatically detect the first and last beats of each drumming bout to precisely measure drumming bout length as a function of age/sex class, allowing us to map the distribution of drumming events throughout the day (Fig. 5; details of the

automatic beat detection method are found in the “Analysis” section in Materials and Methods).

For Bossou chimpanzees, nut cracking bouts were predominantly performed by adult males ($n = 4665$ bouts) followed by adult females ($n = 5485$ bouts) and juveniles ($n = 2134$ bouts), while infants ($n = 1$) were not observed nut-cracking. The mean time spent nut-cracking and the proportion of time spent nut-cracking differed between age/sex groups. Adult males spent a greater proportion of their time nut-cracking than adult females (males, mean $\pm$ SD = $9.21 \pm 9.49\%$; and females, mean $\pm$ SD = $7.97 \pm 9.19\%$), while juveniles required longer nut cracking sequences per nut consumed than adult males and females (males, mean $\pm$ SD = 16.8 ± 6.46 s; females, mean $\pm$ SD = 15.7 ± 10.41 s; and juveniles/infants, mean $\pm$ SD = 43.4 ± 39.0 s) (Fig. 4C and table S3), confirming previous reports on the ontogeny of nut cracking (24). This suggests that adult males consumed the greatest number of nuts.

For buttress drumming, we analyzed 992 drumming bouts; the majority of bouts were performed by adult males ($n = 845$), confirming previous observations that this is a predominantly male activity (25), and occurred throughout the day, following a bimodal distribution with peaks in the morning and in the afternoon (Fig. 5B). When analyzing bout duration, adult males had, on average, shorter bouts (mean $\pm$ SD = 2.21 ± 1.80 s) than immature individuals (mean $\pm$ SD = 2.72 ± 2.39 s) and adult females (mean $\pm$ SD = 2.75 ± 1.79 s). There is a marked variation within each age/sex group, especially in adult males (min = 0.21 s and max = 19.48 s). In addition, drumming context (travel, feeding, and agonistic display) was analyzed for both adult males and adult females. In both groups drumming, during “travel” was the most common (509 bouts for males and 34 bouts for females), followed by “agonistic display” (225 bouts for males and 25 bouts for females), and lastly “feeding” (111 bouts for males and nine bouts for females). The proportion of drumming events for different contexts was approximately equal for both adult males and females (fig. S3). All drumming performed by immature individuals was done in a “play” context. Drumming bout duration varied between contexts in both adult males and adult females as well as between sexes. Feeding drumming bouts were, on average, shorter (males, mean $\pm$ SD = 1.74 ± 1.13 s; and females, mean $\pm$ SD = 2.17 ± 1.09 s) than agonistic display (males, mean $\pm$ SD = 2.31 ± 2.61 s; and females, mean $\pm$ SD = 2.78 ± 1.78 s) and travel drumming bouts (males, mean $\pm$ SD = 2.27 ± 1.35 s; and females, mean $\pm$ SD = 2.89 ± 1.97 s).

Table 1. Recognition results for both nut cracking and buttress drumming. We provide a baseline (random), which shows the chance performance of a random classifier. Bold indicates the highest performing method for the task.

Task	Method	Average precision	
		Nut cracking	Buttress drumming
I. Scene level	Random	0.09	0.11
	Audio	0.85	0.87
II. Individual level	Random	0.12	0.13
	Audio	0.30	0.81
	Visual	0.76	0.64
	Audiovisual	0.77	0.86

DISCUSSION

Overall, our model demonstrates the efficacy of using deep neural network architectures for a biological application: the automated recognition of percussive behaviors in a wild primate. Unlike older, rule-based automation methods, our method is entirely based on deep learning and is data-driven. It also improves on previous single-frame methods by being video-based: It uses 3D convolutions in time (18) to reason about temporal information, which is important for action detections, and exploits the multimodality of video to use the audio and visual streams jointly to classify behaviors.

Often, it is challenging to curate datasets large enough to train action recognition models without sifting through a significant amount of footage (20 and 4.1% of footage yielded our behaviors of interest for the nut cracking and buttress drumming data, respectively). A key aspect of our approach is the use of audio as a prescreening



Fig. 2. Behavior recognition results demonstrate the CNN model's robustness to variations in pose, lighting, scale, and speed of action. Example of correctly labeled body tracks from unseen and unheard videos (nut cracking and drumming for the top two and bottom two rows, respectively). Middle two rows: Multiple individuals nut-cracking and buttress-drumming showing variations in lighting, pose, background, and number of chimpanzees. Photo credit: Kyoto University, Primate Research Institute; The Cantanhez Chimpanzee Project.

mechanism, which substantially cuts down the large search space of video for annotation.

Furthermore, we do not constrain the video data in any way, as is done commonly for deep learning methods applied to primate recognition and analysis by aligning individual detections or selecting

for age, resolution, or lighting (26). Instead, we are able to perform the task “in the wild” and ensure an end-to-end pipeline that will work on raw video with minimum preprocessing (Fig. 2). We also demonstrate that our method is applicable to both long-term targeted field video recordings (including in a field experimental setting) and to

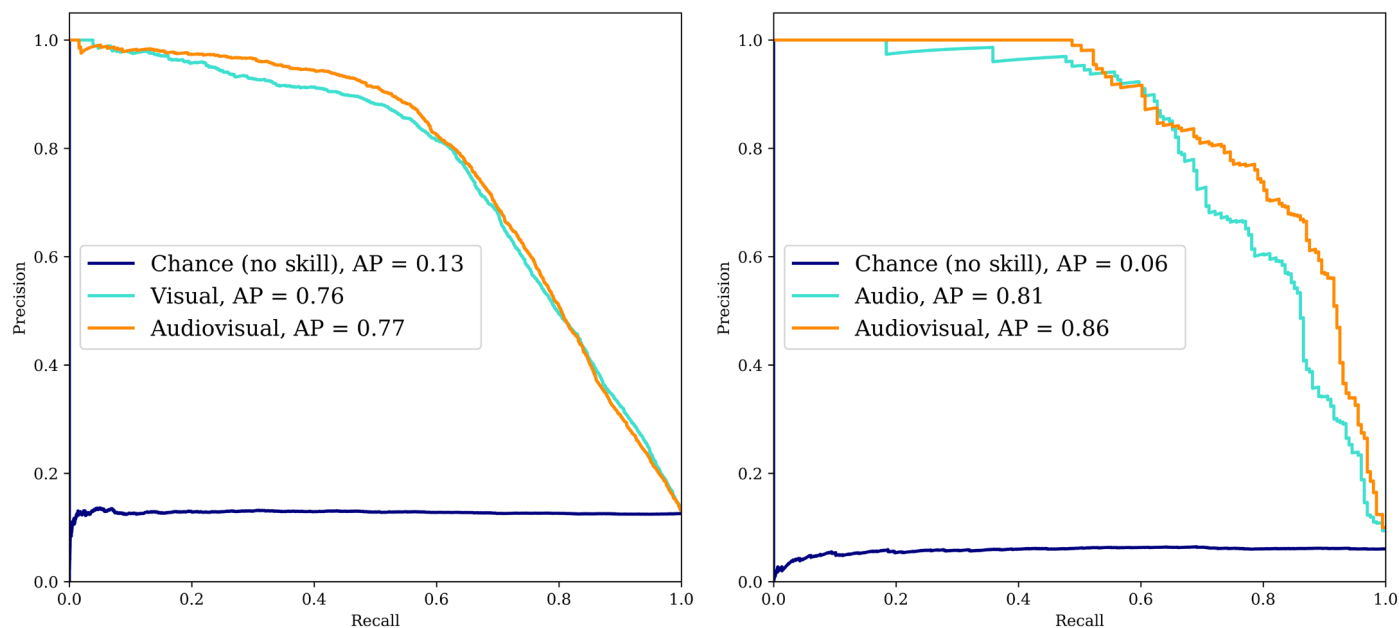


Fig. 3. Performance of the audio, visual, and audiovisual models for individual-level behavior classification. The curves for nut-cracking (left) and buttress-drumming (right) demonstrate that audiovisual outperforms single-modality methods. Instances where the behavior is either visually or audibly occluded can be compensated by using the other modality (AP: Average Precision).

remote monitoring camera trap datasets, demonstrating its usefulness across different data collection protocols.

The pipeline can be applied to data where only audio or only visual information is available (e.g., camera trap recordings where the behavior occurs off-screen, video recordings with noisy or corrupted audio, or microphone-only recordings). The benefits of our audio method are that it is not affected by visual distractors such as lighting, pose, size, and occlusion and is also computationally cheaper to run. Certain actions (such as buttress drumming) are also more discriminative in the audio space than the visual space (Table 1) and hence require less training data. Audio also allows greater coverage, by detecting actions beyond the field of view of the camera. Our visual-only method, on the other hand, provides the added benefit of allowing localization at an individual level, predicting which individuals are performing particular actions. This is a key advantage for potential future applications in, for example, the monitoring of individual behavior and welfare, both in the wild and in captive settings. Our audiovisual model combines the benefits of both modalities. For drumming, we demonstrate that our model works well even on camera trap data from locations unseen by our model during training (which therefore might contain different tree species) and communities of chimpanzees.

Our model's performance demonstrates the effectiveness of using multimodal deep learning for behavioral recognition of individual animals in longitudinal video archives and camera trap datasets in the wild. Using a novel combination of data collection methods (automated classifiers and manual annotations) and video datasets (archival footage and camera traps), we validate our approach by reproducing known findings on the ontogeny of nut cracking from the existing literature (24) and go further to gain preliminary insights into drumming behavior in unhabituated communities as well as revealing potential sex and age differences in different contexts (previously neglected in published work). Ultimately, the integration of

computer vision and ethology using automated behavior recognition can aid behavioral research and conservation, moving beyond inferences of social structure and demographics that can be inferred using individual identification [e.g., (14)] to capturing the full complexity and dynamics of social interactions and behaviors. Typically, the time and resources required for manual data collection of multiple behaviors (either through in situ observation or retrospective video coding) prohibits analysis of large scale datasets. Adopting automated behavioral recognition is scalable, increasing the speed, quantity, and detail of data that can be collected and analyzed. Once classifiers have been trained, such work can move beyond broad classification of general behavioral states (eating, resting, etc.) to include fine-grained analysis at multiple layers/dimensions of behavior (27)—for example, using pose estimation to quantify postural kinematics or detect the number and order of elements in a behavioral sequence (e.g., nut cracking strikes) or investigating temporal co-occurrences between the behavior of individuals in the same group. We also envisage that our method could have a large impact in conservation science. Anthropogenic pressures are increasingly affecting animal behavior, with habitat fragmentation and population loss posing an imminent threat to “cultural species” through the erosion of behavioral diversity (28). Automating the measurement of behavioral diversity and activity budgets could be crucial for developing more sophisticated metrics to monitor the health and stability of wild populations (10).

There are some limitations to our study, notably that the audio preview step is limited to actions that contain a distinctive sound (such as percussion). Nonetheless, none of the pipeline steps are specific to primate behavior, and the method can be readily applied to other animal species and behaviors. Furthermore, behaviors that are audio distinctive exist in multiple domains, and we envisage possible applications for our pipeline in, for example, marine and terrestrial animal communication (vocalizations), movement (wing

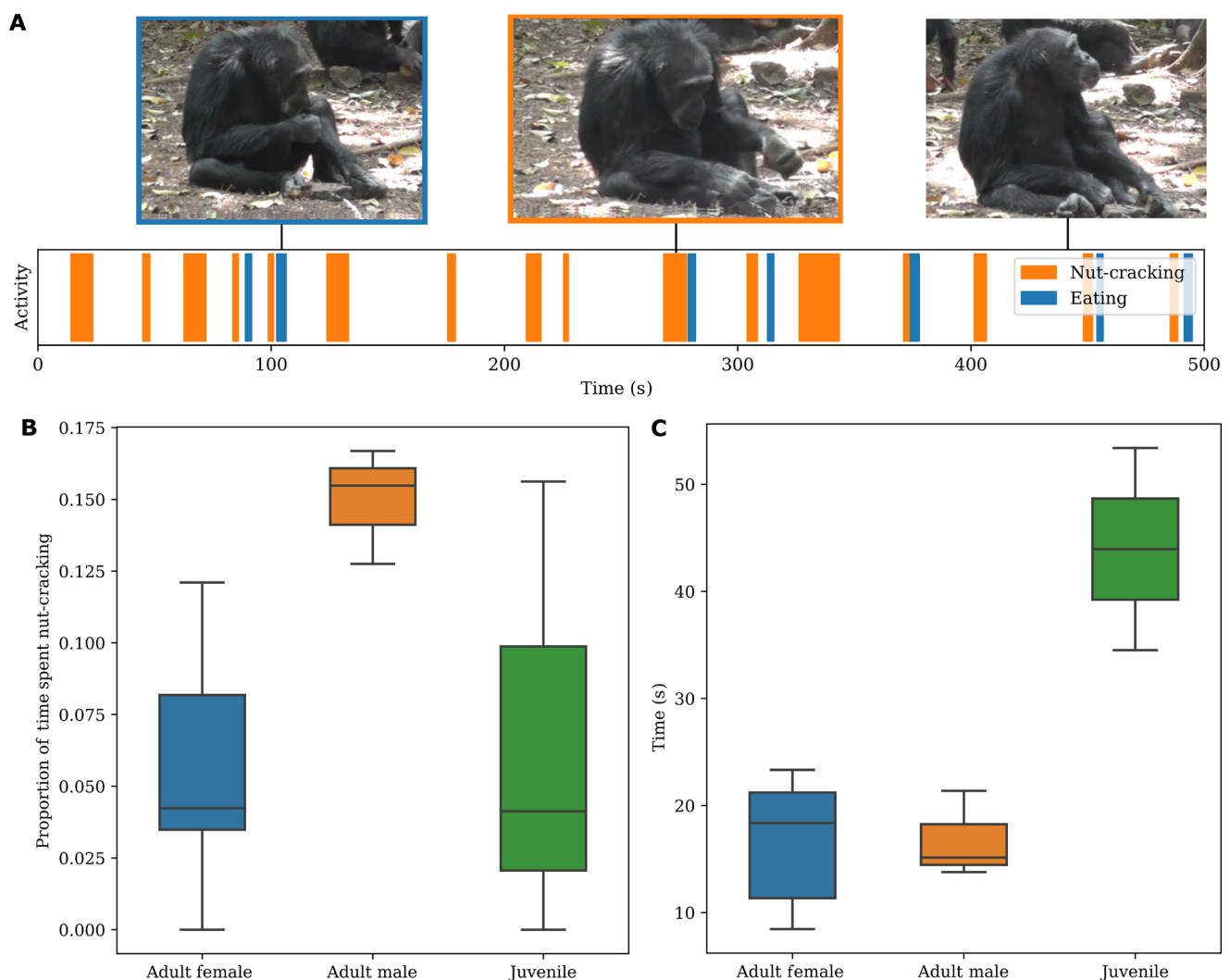


Fig. 4. Nut cracking analysis. (A) An example activity sequence following a single individual over the course of a video. The blank white spaces are any activities that are not nut cracking or eating. Note that eating typically follows nut cracking events. (B) Proportion of time spent nut cracking as a fraction of total time visible. (C) Average time spent nut cracking per eating event. Computed by dividing the cumulative time spent nut-cracking over the total number of eating events as a function of age and sex. Photo credit: Kyoto University, Primate Research Institute.

flapping and stepping), self-maintenance (scratching), aggression (hitting, slapping, and screaming), and foraging (tearing, smashing, and chewing). These analyses could be performed on data not only from remote sensors but also from animal-borne audio-only biologists. Another limitation concerns the fact that, for individual-level recognition, our method is heavily reliant on the performance of the body detector: Individuals that are not detected or tracked cannot have their behavior classified. For example, the detector often fails to detect infants on their mother's backs, although for our present analyses, this poses no problem, because young chimpanzees do not nut-crack or buttress-drum while being carried. For behaviors that specifically require a visual classifier (such as successful nut cracking being identified through the hand-to-mouth motion of eating), visual occlusion or motion blur poses challenges. However, we note that the body detector has far fewer missed individual detections than other methods that are reliant on face detection (14, 22). Future

directions to improve our pipeline include adopting active learning, which minimizes annotator effort by automatically selecting informative samples from a pool of unannotated data for a human to annotate to retrain the network (29). In addition, self-supervised learning enables label-free pretraining, initializing the model in such a way that reduces the annotation requirement for training (30).

Our pipeline provides a critical first step in large-volume automated behavioral coding and represents a breakthrough in measuring behavior. It will permit detailed intraindividual, interindividual, and cross-site comparisons, automated collection of activity budgets, and longitudinal studies of behavior at individual and population levels, enabling detailed investigation into ontogeny, cultural evolution, and the persistence/decline of behavioral variation over time and how these relate to environmental change. It has transformative potential to science, setting a milestone for exploiting large datasets in ethology and conservation.

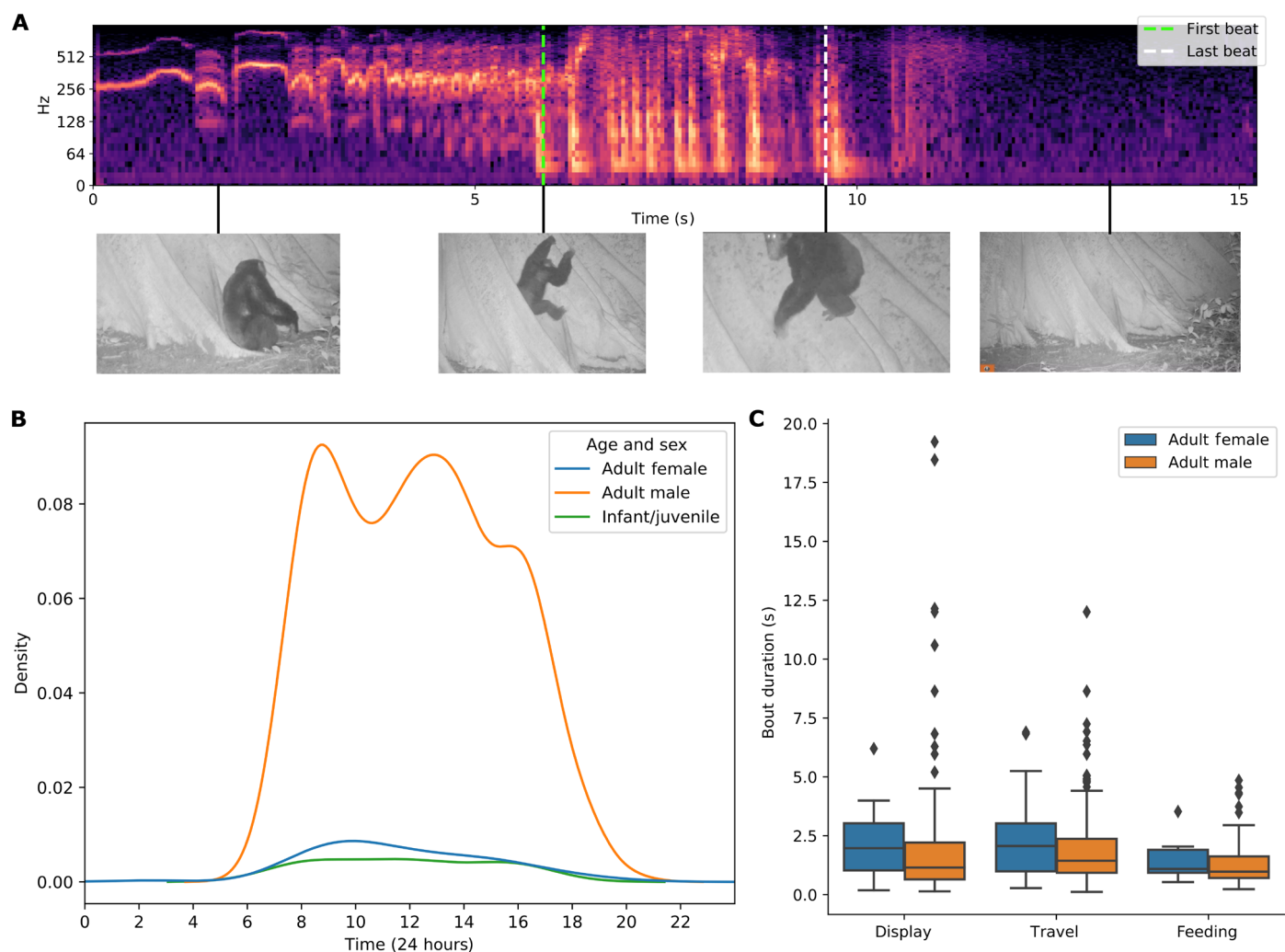


Fig. 5. Buttress drumming analysis. (A) Spectrogram showing a detected drumming bout delineated by the first and last beats, with video frames visualized. (B) Kernel density estimation plot showing the diel distribution of buttress drumming bouts, based on hh:mm:ss data captured by camera traps. (C) Duration (in seconds) of buttress drumming bouts by context (agonistic display, travel, and feeding) and by age/sex (adult females and adult males). Photo credit: The Cantanhez Chimpanzee Project.

MATERIALS AND METHODS

Video archive

Description of actions

Nut cracking has been described as the most complex tool-use behavior in wild chimpanzees, with the nut cracker typically combining three objects (31, 32). It involves placing a hard-shelled nut on an anvil and then using a hammer to pound the nut until the edible kernel is exposed—sometimes one or two wedges are used to stabilize the anvil. We defined nut cracking “sequences” as beginning when the hammer is raised before the initial strike of a nut and ending when the hammer makes contact with the nut or anvil for the final time before the nut is consumed or abandoned or the camera is moved away. Sequences often included multiple strikes for a single nut.

Buttress drumming is a universal and frequent behavior across all chimpanzee communities, but there is much left to understand about its functions and potential cross-community variation. Drumming occurs when a chimpanzee slaps or stamps rhythmically on the buttress of a tree, often accompanied by a distinct vocalization called a pant hoot. Multiple functions of drumming have been proposed,

including long-distance communication (25, 33) and intimidation accompanying agonistic displays (34). Distinctive individual drumming patterns and pant hoot vocalizations are thought to act as signals that coordinate group movement and distribution when traveling, as well as containing information about the individual’s identity (35). These distinct drumming patterns have been described for both males and females (36), but male chimpanzees appear to drum more frequently when traveling (35). We defined buttress drumming sequences as beginning when the first beat was detected and ending with the last beat; any behavior, such as pant hoots, occurring immediately before the first beat or immediately after the last beat were not included. Beats were detected visually, when at least one hand or foot was in contact with the buttress, and auditorily, when the distinct beat sound was heard.

Structure of the data

Nut cracking at Bossou, Guinea. Data used were collected in the Bossou forest, southeastern Guinea, West Africa, a long-term chimpanzee field site established by Kyoto University in 1976 (14). Bossou is home to an outdoor laboratory: A natural forest clearing (7 m by 20 m)

located in the core of the Bossou chimpanzees' home range (07°39'N and 008°30'W) where raw materials for tool use—stones and nuts—are provisioned, and the same group has been recorded since 1988 (14, 22). The use of standardized video recording over many field seasons has led to the accumulation of more than 30 years of video data, providing unique opportunities to analyze chimpanzee behavior over multiple generations. In total, we analyzed 43.1 hours of video footage.

Buttress drumming in Cantanhez National Park, Guinea-Bissau. Data used were collected by camera traps ($n = 25$) deployed in the home ranges of four different communities (Caiquene-Cadique, Lautchandé, Cambeque, and Cabante) in Cantanhez National Park, Southern Guinea-Bissau, West Africa (11°14'17.2"N and 15°02'16.9"W) between February 2017 and December 2018. Chimpanzees in Cantanhez National Park inhabit an agroforest landscape and are not habituated to researchers. The camera traps were set up in areas that chimpanzees frequented and pointed to trees with large buttress roots with clear signs of wear from chimpanzee buttress drumming. Some cameras were moved during the study period to account for seasonal changes in chimpanzee ranging patterns and when a new area of interest was located. Cameras were motion sensitive and were set to record 1-min video clips when triggered. Approximately 41,000 video clips were collected over the study period, of which 4745 contained footage of chimpanzees, spanning a total of 47.2 hours of video footage.

Dataset splits: Training, testing, and analysis. We divide the dataset into different sections. Part of the data is manually annotated by human annotators, which provides data for training and testing our automated framework (described in the "Methods" section). The remaining data are unlabeled by humans. Our framework is applied to these unlabeled data automatically (this stage is referred to as inference) for analysis (described in the "Analysis" section). Dataset statistics are provided in table S1.

Methods

Our pipeline for the detection of audio-discriminative percussive behaviors consists of the following two stages: (i) chimpanzee detection and tracking and (ii) audiovisual action recognition (Fig. 1).

To efficiently collect annotations for the second stage (ii) audiovisual action recognition, we also use an additional "audio preview stage" (described below in the "Audio action recognition" section) only when collecting training data. This optional audio preview stage uses audio only to determine temporal segments where the behaviors (nut cracking and drumming) occur at a scene level. This markedly reduces the total search space of the video, allowing for efficient annotations, used to train a model on both scene-level audio and the visual content of each track to determine which individual is carrying out the behavior.

With this trained model, our pipeline can then be applied directly to previously unseen videos without any human input. At this point, we do not require the audio preview stage and only use (i) detection and tracking and (ii) audiovisual action recognition.

All stages in the method are implemented using deep CNNs. For audio previewing, we train a CNN on the spectrogram image of the audio. For detection, we use a single-shot detector (SSD) object category detector (37) to detect individuals. The detections for an individual are then grouped across frames (time) using a pretrained tracker. The final audiovisual action recognition stage involves a spatiotemporal CNN for the visual features and a spectrogram CNN for the audio. The training data were obtained by using the

VGG Image Annotator (VIA) annotation tool (38). We provide a detailed description for each stage of the pipeline in the following sections and then describe how the analysis is carried out given the detected behaviors.

Audio action recognition

With the audio data alone, our framework is able to classify actions at the scene level. The nut cracking and buttress drumming audio classifier achieved 85 and 87% average precision, respectively, on unseen test data (Table 1).

Network architecture. For the audio model, we use a 2D CNN (ResNet-18), pretrained on VGGSound (39). The output is passed through two linear layers and then a final predictive layer with two neurons and a softmax activation function, resulting in a binary classifier for each target action.

Inputs. We use short-term magnitude spectrograms as input to a ResNet-18 model. All audio is first converted to single-channel, 16-bit streams at a 16-kHz sampling rate for consistency. Spectrograms are then generated in a sliding window fashion using a hamming window with a width of 32 ms and a hop of 10 ms, with a 512-point fast Fourier transform. This gives spectrograms a size of 257×201 for 3 s of audio. The resulting spectrogram is integrated into 64 mel-spaced frequency bins with a minimum frequency of 125 Hz and a maximum frequency of 7.5 kHz, and the magnitude of each bin is log-transformed. This gives log mel spectrogram patches of 64×201 bins, used as input to the CNN.

Augmentations. Temporal jittering of 0.5 s is used as well as augmentation to positive samples by randomly adding background audio samples (audio that does not contain nut cracking and buttress drumming).

Training. Binary cross-entropy is used as the training objective, along with an Adam optimizer with a learning rate of 5×10^{-3} .

Audio preview for manual annotation. Videos in the wild (including from camera traps) contain a lot of dead footage, where the actions of interest may be captured rarely. Manually searching through all this footage is a labor-intensive task. Hence, we use an inexpensive and computationally efficient prescreening method to automatically sift through many hours of footage, proposing short videos that contain the action and discarding the rest. This is done using the audio alone, because our actions of interest are all percussive and make a distinct sound.

The audio model is applied using a sliding window of size 3 s, with a stride of 0.5 s over the raw video footage. This produces a probability score $P(\text{action})$ of the action of interest being present within each temporal window. We then use the most confident 7% of windows (using the probability score as the confidence) for discrete video labeling, resulting in 2418 discrete, 3-s long video proposals to be annotated. The more expensive body detection and tracking is performed only on these "audio proposals." The body tracks are visualized on the proposals, allowing the annotators to label each actor in the proposal with a binary label denoting whether or not they are performing the action. Given that the drumming video footage is already segmented into short clips and annotated, the audio preview step was not required for the buttress drumming data at training time, so it was only used for nut cracking here.

At inference time, the audio preview can be used as a filtering step first before the full framework, providing computation savings. Because audio is much cheaper computationally than the full framework (detection, tracking, and audiovisual classification), this can be useful in resource-constrained environments such as running

the framework on the camera traps themselves. Because this work was not constrained in terms of compute, we did not use the audio preview step at inference. For buttress drumming, the trade-off is minimal; a computation saving of 64% still captures 97% of drumming events. For nut cracking, the trade-off is greater; a computation saving of 64% captures 70% of nut cracking events. As there are many off-screen nut cracking events, the sound of nut cracking is not definitively on-screen.

Visual detection and tracking

A prerequisite for our method of automated detection of primate behavior is the detection and tracking of the target animal, producing spatiotemporal tracks following individuals through time. Deep learning has proved to be highly successful at object detection and tracking, and previous works describe the protocol and results of this applied to footage of wild animals (13, 14, 40, 41).

In more detail, we follow the same protocol as in (13), which involves fine-tuning an SSD object detector (37) on bounding box annotations of chimpanzee bodies. Because the two datasets contain very different sources of footage, including camera traps for drumming and direct longitudinal recordings for nut cracking (the former containing night vision, varied lighting, and out of focus blur; with the latter having higher quality video but consisting of close-ups as well as medium shots), we separately fine-tune the two object detectors, one for each dataset.

For the nut cracking dataset, we fine-tune on 16,000 bounding box annotations across 5513 video frames. For the buttress drumming dataset, we fine-tune on 2200 bounding box annotations across 2137 video frames. All video frames were sampled every 10 s.

Tracking. The object tracker used to link the resulting detections through time is a pretrained Siamese network. Pairs of detections in consecutive frames with a Jaccard overlap greater than 0.5 are given as input to the network. Detection pairs with a similarity score greater than 0.5 are deemed to be from the same track.

Evaluation for the detectors. Evaluation is performed on a held-out test set using the standard protocol outlined in (42). The precision-recall curve is computed from a method's ranked output. Recall is defined as the proportion of all positive examples above a given rank, while precision is the proportion of all examples above that rank, which are from the positive class. For the purpose of our task, high recall is more important than high precision (i.e., false positives are less dangerous than false negatives) to ensure that no chimpanzees are missed. The Bossou and Cantanhez detectors achieved average precision scores of 0.92 and 0.91 on their respective test sets. Precision-recall curves for both detectors are shown in fig. S1A.

Programming implementation details. The detector was implemented using the machine learning library PyTorch and trained on two Titan X Graphical Processing Units (GPUs) for 20 epochs (where 1 epoch consists of an entire pass through the training set) using a batch size of 32 and two subbatches. Flip, zoom, path, and distort augmentation was used during preprocessing with a zoom factor of 4. The ratio of negatives to positives while training was 3, and the overlap threshold was 0.5. The detector was trained without batch normalization. The tracker was also implemented in PyTorch.

Audio-visual action recognition

Network architecture. For the visual stream, we use a 3D ResNet-18, with 3D convolutions (30). The output is passed through two linear layers and then a final predictive layer, with two neurons and a softmax activation function. For the audiovisual fusion model, 512 di-

mensional embeddings from the ResNet backbone in each stream are concatenated and then passed to the final predictive layer, with two neurons and a softmax activation function.

Inputs. For the audio stream, the preprocessing is identical to the "Audio action recognition" stage. Video frames are sampled at 25 frames per second, and all detections are resized to 128×128 —we feed in 40 frames over 2.5 s, with three red-green-blue channels each, sampled randomly during training and uniformly during inference. This gives final inputs of size $40 \times 128 \times 128 \times 3$.

Augmentations. Standard augmentation techniques are applied to the visual inputs: color jittering, random cropping, and horizontal flipping. For the audio, we repeat the augmentations in the "Audio action recognition" section.

Training. All models are trained with a binary cross-entropy loss. In this stage, we use the annotations obtained from the "Audio action recognition" stage of the pipeline to train the model.

Evaluation. Evaluation for the action recognition models is performed on a held-out test set, the statistics of which are supplied in Table 1. The audiovisual fusion model performed the best at the individual level for both nut cracking and buttress drumming (77 and 86%, respectively), demonstrating its robustness across domains and actions and demonstrating its efficacy over audio or vision alone.

The models are evaluated on their precision recall at either the scene level or individual level. For the scene level, we evaluate the audio-only model with a stride of 0.5 s and a forgiveness collar of 0.5 s. For the individual level, we evaluate the audio, visual, and audiovisual models with a stride of 0.5 s per track and a forgiveness collar of 0.5 s.

Implementation details. The networks for action recognition were trained on four Titan X GPUs for 20 epochs using a batch size of 16. We trained both models end to end via stochastic gradient descent with momentum (0.9) weight decay (5×10^{-4}) and a logarithmically decaying learning rate (initialized to 10^{-2} and decaying to 10^{-8}). The visual stream is initialized with weights from (30), and the audio model is initialized with weights pretrained on VGG-Sound (39).

Action-specific implementation details

Nut cracking analysis: Success detection. To further analyze nut cracking behaviors, we additionally measure another action: passing food from hand to mouth, which is an indication of successful nut cracking. Here, the shell has been successfully cracked and the individual passes the kernel to their mouth using their hand; henceforth, this action is referred to as "eating." Because this behavior has a strong visual signature, we train a visual classifier to determine this. This model follows the protocol of the visual-only drumming and nut cracking classifiers. The training labels for eating events were gathered from the audio preview proposals, totaling 896 tracklets of individuals eating. While the audio preview searches for nut cracking, eating is often found shortly after successful nut cracking events, so the short audio proposals often contain this action as well. Furthermore, individuals often nut-crack together, resulting in multiple individuals in a video proposal. Training the model on data from 2004 and 2008 results in 89% accuracy in classifying eating on unseen tracks from 2012.

Buttress drumming duration analysis. We investigate the duration of drumming bouts by determining the start and the end beat in a drumming bout using audio-based beat detection. Beat detection is performed in an automated fashion by using low-pass filtering

and onset detection to the audio signal of the drumming bout. The audio sequence is first low pass-filtered using a Butterworth filter with a cutoff frequency of 800 Hz. Onset detection is then performed on the filtered audio waveform. We use the onset detection method provided by the Librosa Python toolbox. The hyperparameters were chosen to achieve the best beat counting accuracy on 30 drumming bouts hand-labeled with the number of beats. During evaluation, we apply a forgiveness collar of 0.25 s on either side of the drumming event boundaries to be more lenient toward imprecise boundary annotation. From the beat detections, we define the duration of a drumming bout to be the interval between the first and last beats. This beat detection method predicts drumming duration with a mean and median error of 0.205 and 0.131 s, respectively.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at <https://science.org/doi/10.1126/sciadv.abi4883>

[View/request a protocol for this paper from Bio-protocol.](#)

REFERENCES AND NOTES

1. N. Tinbergen, On aims and methods of ethology. *Z. Tierpsychol.* **20**, 410–433 (1963).
2. J. Krause, S. Krause, R. Arlinghaus, I. Psorakis, S. Roberts, C. Rutz, Reality mining of animal social systems. *Trends Ecol. Evol.* **28**, 541–551 (2013).
3. D. J. Anderson, P. Perona, Toward a science of computational ethology. *Neuron* **84**, 18–31 (2014).
4. O. Sturman, L. von Ziegler, C. Schläppli, F. Akyol, M. Privitera, D. Słominski, C. Grimm, L. Thieren, V. Zerbi, B. Grewe, J. Bohacek, Deep learning-based behavioral analysis reaches human accuracy and is capable of outperforming commercial solutions. *Neuropsychopharmacology* **45**, 1942–1952 (2020).
5. E. A. van Dam, L. P. J. J. Noldus, M. A. J. van Gerven, Deep learning improves automated rodent behavior recognition within a specific experimental setup. *J. Neurosci. Methods* **332**, 108536 (2020).
6. P. Swarup, P. Chen, R. Hou, P. Que, P. Liu, A. W. K. Kong, Giant panda behaviour recognition using images. *Glob. Ecol. Conserv.* **26**, e01510 (2021).
7. S. P. Kaufhold, E. J. C. van Leeuwen, Why intergroup variation matters for understanding behaviour. *Biol. Lett.* **15**, 20190695 (2019).
8. M. Cantor, A. A. Maldonado-Chaparro, K. B. Beck, H. B. Brandl, G. G. Carter, P. He, F. Hillemann, J. A. Klarevas-Irby, M. Ogino, D. Papageorgiou, L. Prox, D. R. Farine, The importance of individual-to-society feedbacks in animal ecology and evolution. *J. Anim. Ecol.* **90**, 27–44 (2021).
9. D. M. Dominoni, W. Halfwerk, E. Baird, R. T. Buxton, E. Fernández-Juricic, K. M. Fristrup, M. F. McKenna, D. J. Mennitt, E. K. Perkin, B. M. Seymoure, D. C. Stoner, J. B. Tennesen, C. A. Toth, L. P. Tyrrell, A. Wilson, C. D. Francis, N. H. Carter, J. R. Barber, Why conservation biology can benefit from sensory ecology. *Nat. Ecol. Evol.* **4**, 502–511 (2020).
10. F. Christiansen, M. H. Rasmussen, D. Lusseau, Inferring activity budgets in wild animals to estimate the consequences of disturbances. *Behav. Ecol.* **24**, 1415–1425 (2013).
11. A. Caravaggi, P. B. Banks, A. C. Burton, C. M. V. Finlay, P. M. Haswell, M. W. Hayward, M. J. Rowcliffe, M. D. Wood, A review of camera trapping for conservation behaviour research. *Remote Sens. Ecol. Conserv.* **3**, 109–122 (2017).
12. A. Whiten, J. Goodall, W. C. McGrew, T. Nishida, V. S. Reynolds, Y. Sugiyama, C. E. G. Tutin, R. W. Wrangham, C. Boesch, Charting cultural variation in chimpanzees. *Behaviour* **138**, 1481–1516 (2001).
13. M. Bain, A. Nagrani, D. Schofield, A. Zisserman, in *Workshop on Computer Vision for Wildlife Conservation, ICCV (IEEE, 2019)*.
14. D. Schofield, A. Nagrani, A. Zisserman, M. Hayashi, T. Matsuzawa, D. Biro, S. Carvalho, Chimpanzee face recognition from videos in the wild using deep learning. *Sci. Adv.* **5**, eaaw0736 (2019).
15. A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, M. Bethge, DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).
16. D. Shao, Y. Zhao, B. Dai, D. Lin, FineGym: A hierarchical video dataset for fine-grained action understanding, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (IEEE, 2020)*, pp. 2616–2625.
17. M. S. Norouzzadeh, A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, J. Clune, Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E5716–E5725 (2018).
18. S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 221–231 (2013).
19. J. Carreira, A. Zisserman, Quo Vadis, action recognition? A new model and the kinetics dataset, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (IEEE, 2017)*, pp. 6299–6308.
20. F. Sakib, T. Burghardt, Visual recognition of great ape behaviours in the wild. [arXiv:2011.10759 \[cs.CV\]](https://arxiv.org/abs/2011.10759) (21 November 2020); [http://arxiv.org/abs/2011.10759](https://arxiv.org/abs/2011.10759).
21. R. Gao, T.-H. Oh, K. Grauman, L. Torresani, Listen to Look: Action recognition by previewing audio, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (IEEE, 2020)*, pp. 10457–10467.
22. T. Matsuzawa, Field experiments on use of stone tools by chimpanzees in the wild, in *Chimpanzee Cultures*, R. W. Wrangham, W. C. McGrew, F. B. M. de Wall, P. G. Heltne, Eds. (Harvard Univ. Press, 1994), pp. 351–370.
23. J. Bessa, K. Hockings, D. Biro, First evidence of chimpanzee extractive tool use in Cantanhez, Guinea-Bissau: Cross-community variation in honey dipping. *Front. Ecol.* **9**, 625303 (2021).
24. D. Biro, N. Inoue-Nakamura, R. Tonooka, G. Yamakoshi, C. Sousa, T. Matsuzawa, Cultural innovation and transmission of tool use in wild chimpanzees: Evidence from field experiments. *Anim. Cogn.* **6**, 213–223 (2003).
25. A. C. Arcadi, D. Robert, C. Boesch, Buttress drumming by wild chimpanzees: Temporal patterning, phrase integration into loud calls, and preliminary evidence for individual distinctiveness. *Primates* **39**, 505–518 (1998).
26. D. Deb, S. Wiper, A. Russo, S. Gong, Y. Shi, C. Tymoszek, A. Jain, Face recognition: Primates in the wild, [arXiv:1804.08790 \[cs.CV\]](https://arxiv.org/abs/1804.08790) (24 April 2018); [http://arxiv.org/abs/1804.08790](https://arxiv.org/abs/1804.08790).
27. K. Huang, Y. Han, K. Chen, H. Pan, G. Zhao, W. Yi, X. Li, S. Liu, P. Wei, L. Wang, A hierarchical 3D-motion learning framework for animal spontaneous behavior mapping. *Nat. Commun.* **12**, 2784 (2021).
28. H. S. Kühl, C. Boesch, L. Kulik, F. Haas, M. Arandjelovic, P. Dieguez, G. Bocksberger, M. B. McElreath, A. Agbor, S. Angedakin, E. A. Ayimisin, E. Bailey, D. Barubiyo, M. Bessone, G. Brazzola, R. Chancellor, H. Cohen, C. Coupland, E. Danquah, T. Deschner, D. Dowd, A. Dunn, V. E. Egbe, H. Eshuis, A. Goedmakers, A.-C. Granjon, J. Head, D. Hedwig, V. Hermans, I. Imong, K. J. Jeffery, S. Jones, J. Junker, P. Kadam, M. Kambere, M. Kambi, I. Kienast, D. Kujiakwinja, K. E. Langergraber, J. Lapuente, B. Larson, K. Lee, V. Leinert, M. Llana, G. Maretti, M. Marrocoli, R. Martin, T. J. Mbi, A. C. Meier, B. Morgan, D. Morgan, F. Mulindahabi, M. Murai, E. Neil, P. Niyigaba, L. J. Ormsby, R. Orume, L. Pacheco, A. Piel, J. Preece, S. Regnaut, A. Rundus, C. Sanz, J. van Schijndel, V. Sommer, F. Stewart, N. Tagg, E. Vendras, V. Vergnes, A. Welsh, E. G. Wessling, J. Willie, R. M. Wittig, Y. G. Yuh, K. Yurkiw, K. Zuberbühler, A. K. Kalan, Human impact erodes chimpanzee behavioral diversity. *Science* **363**, 1453–1455 (2019).
29. M. S. Norouzzadeh, D. Morris, S. Beery, N. Joshi, N. Jovic, J. Clune, A deep active learning system for species identification and counting in camera trap images. *Methods Ecol. Evol.* **12**, 150–161 (2021).
30. T. Han, W. Xie, A. Zisserman, Video representation learning by deep predictive coding, in *Proceedings of the IEEE International Conference on Computer Vision Workshops (IEEE, 2019)*, pp. 1483–1492.
31. T. Matsuzawa, T. Humle, Y. Sugiyama, *The Chimpanzees of Bossou and Nimba* (Springer Science & Business Media, 2011).
32. S. Carvalho, E. Cunha, C. Sousa, T. Matsuzawa, Chaînes opératoires and resource-exploitation strategies in chimpanzee (Pan troglodytes) nut cracking. *J. Hum. Evol.* **55**, 148–163 (2008).
33. C. Boesch, Symbolic communication in wild chimpanzees? *Hum. Evol.* **6**, 81–89 (1991).
34. T. Nishida, *Chimpanzees of the Lakeshore: Natural History and Culture at Mahale* (Cambridge Univ. Press, 2011).
35. M. Babiszewska, A. M. Schel, C. Wilke, K. E. Slocumbe, Social, contextual, and individual factors affecting the occurrence and acoustic structure of drumming bouts in wild chimpanzees (Pan troglodytes). *Am. J. Phys. Anthropol.* **156**, 125–134 (2015).
36. V. Reynolds, *The Chimpanzees of the Budongo Forest: Ecology, Behaviour and Conservation* (OUP Oxford, 2005).
37. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, *European Conference on Computer Vision* (Springer, 2016), pp. 21–37.
38. A. Dutta, A. Zisserman, in *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)* (Association for Computing Machinery, 2019), pp. 2276–2279.
39. H. Chen, W. Xie, A. Vedaldi, A. Zisserman, VggSound: A large-scale audio-visual dataset, in *Proceedings of ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2020), pp. 721–725.
40. J. Yu, H. Su, J. Liu, Z. Yang, Z. Zhang, Y. Zhu, L. Yang, B. Jiao, A Strong Baseline for Tiger Re-ID and its Bag of Tricks, in *Proceedings of the IEEE International Conference on Computer Vision Workshops (IEEE, 2019)*, pp. 302–309.
41. P. Chen, P. Swarup, W. M. Matkowski, A. W. K. Kong, S. Han, Z. Zhang, H. Rong, A study on giant panda recognition based on images of a large proportion of captive pandas. *Ecol. Evol.* **10**, 3561–3573 (2020).
42. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **88**, 303–338 (2010).

Acknowledgments: We are grateful to Kyoto University's Primate Research Institute for leading the Bossou Archive Project and supporting the research presented here and to the IREB and DNRSIT of Guinea. This study is dedicated to all the researchers and field assistants who have collected data in Bossou since 1988. We thank the Instituto da Biodiversidade e das Áreas Protegidas (IBAP) for their permission to conduct research in Guinea-Bissau and for logistical support, research assistants and local guides for assisting with data collection, and local leaders for granting us permission to conduct research. We thank M. Ramon for collecting camera trap data in Cabante, Guinea-Bissau. **Funding:** This study was supported by EPSRC Programme Grants Seebibyte EP/M013774/1 and Visual AI EP/T028572/1; Google PhD Fellowship (to A.N.); Clarendon Fund (to D.S. and S.B.); Boise Trust Fund (to D.S., S.B., and J.B.); Wolfson College, University of Oxford (to D.S.); Keble College Sloane-Robinson Clarendon Scholarship, University of Oxford (to S.B.); Fundação para a Ciência e a Tecnologia, Portugal SFRH/BD/108185/2015 (to J.B.); Templeton World Charity Foundation grant no. TWCF0316 (to D.B.); National Geographic Society (to S.C.); St Hugh's College, University of Oxford (to S.C.); Kyoto University Primate Research Institute for Cooperative Research Program (to M.H. and D.S.);

MEXT-JSPS (no. 16H06283), LGP-U04, the Japan Society for the Promotion of Science (to T.M.); and Darwin Initiative funding grant number 26-018 (to K.J.H.). **Author contributions:** Conceptualization: D.S. Methodology: M.B., A.N., and A.Z. Data curation: M.B., D.S., J.B., S.B., and J.O. Data collection: D.B., S.C., T.M., M.H., K.J.H., and J.B. Software, formal analysis, and visualization: M.B. Supervision: A.Z., D.B., and S.C. Writing (original draft): M.B., A.N., D.S., and J.B. Writing (review and editing): A.Z., D.B., S.C., and K.J.H. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials and at the Dryad data repository (<https://datadryad.org/stash/share/UUfSTzSL9eTbAo-78pdaXPdaIJmdJzSuqhXcb48vHM>).

Submitted 12 March 2021

Accepted 23 September 2021

Published 12 November 2021

10.1126/sciadv.abi4883

Automated audiovisual behavior recognition in wild primates

Max BainArsha NagraniDaniel SchofieldSophie BerdugoJoana BessaJake OwenKimberley J. HockingsTetsuro MatsuzawaMisato HayashiDora BiroSusana CarvalhoAndrew Zisserman

Sci. Adv., 7 (46), eabi4883. • DOI: 10.1126/sciadv.abi4883

View the article online

<https://www.science.org/doi/10.1126/sciadv.abi4883>

Permissions

<https://www.science.org/help/reprints-and-permissions>

Use of this article is subject to the [Terms of service](#)

Science Advances (ISSN) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS. Copyright © 2021 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

Supplementary Materials for

Automated audiovisual behavior recognition in wild primates

Max Bain*, Arsha Nagrani, Daniel Schofield, Sophie Berdugo, Joana Bessa,
Jake Owen, Kimberley J. Hockings, Tetsuro Matsuzawa, Misato Hayashi, Dora Biro,
Susana Carvalho, Andrew Zisserman

*Corresponding author. Email: maxbain@robots.ox.ac.uk

Published 12 November 2021, *Sci. Adv.* **7**, eabi4883 (2021)
DOI: [10.1126/sciadv.abi4883](https://doi.org/10.1126/sciadv.abi4883)

The PDF file includes:

Supplementary Text
Figs. S1 to S3
Table S1
Legend for Movie S1

Other Supplementary Material for this manuscript includes the following:

Movie S1

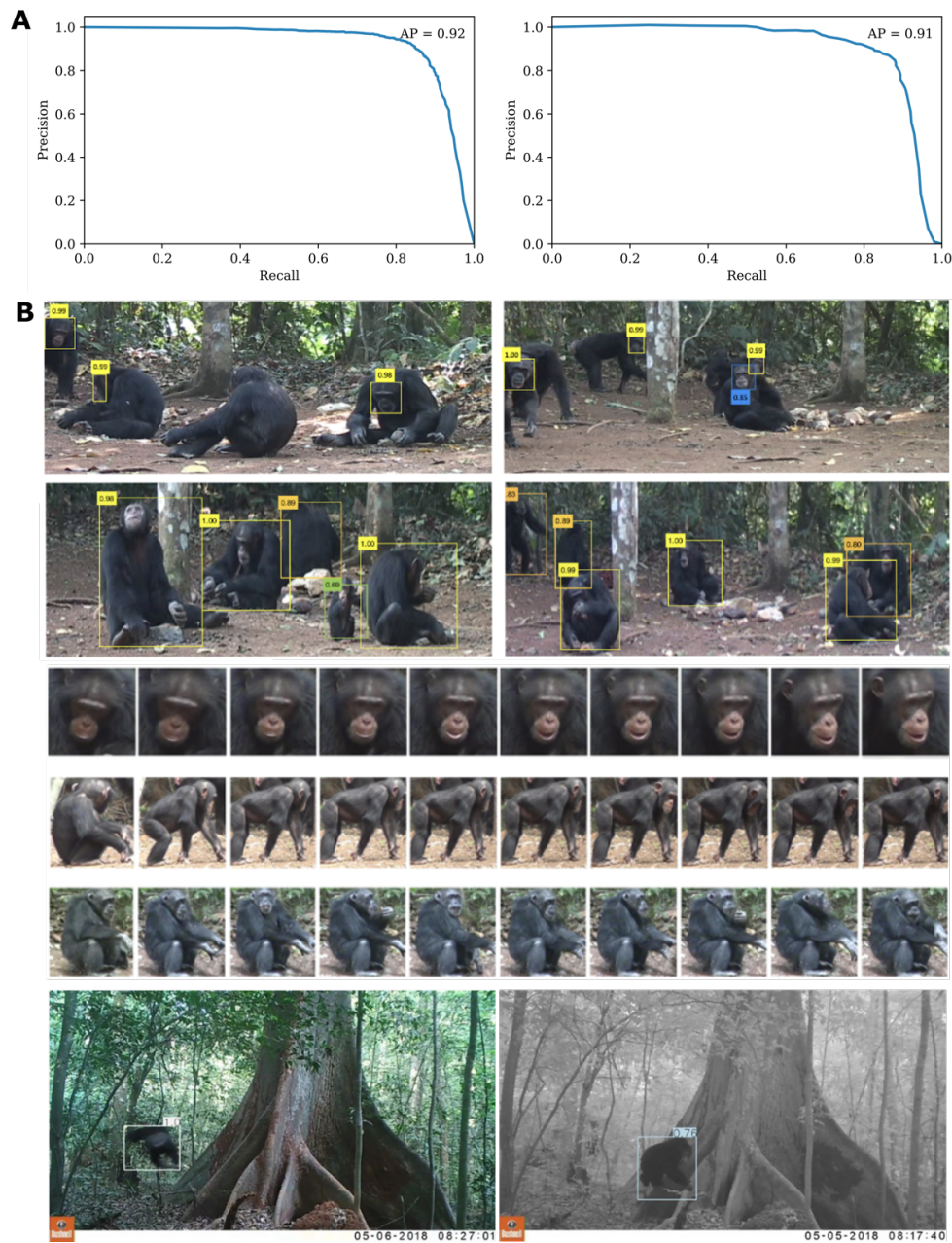


Fig. S1. Detection and Tracking Results: (A) Precision Recall Curves for the individual body detectors for Bossou (left) and Cantanhez (right), with their Average Precision (AP) scores labelled. (B) Qualitative examples of detections (detector confidences shown above each detection) and body tracks on both the Bossou and Cantanhez video footage. Photo credit: Kyoto University, Primate Research Institute; The Cantanhez Chimpanzee Project

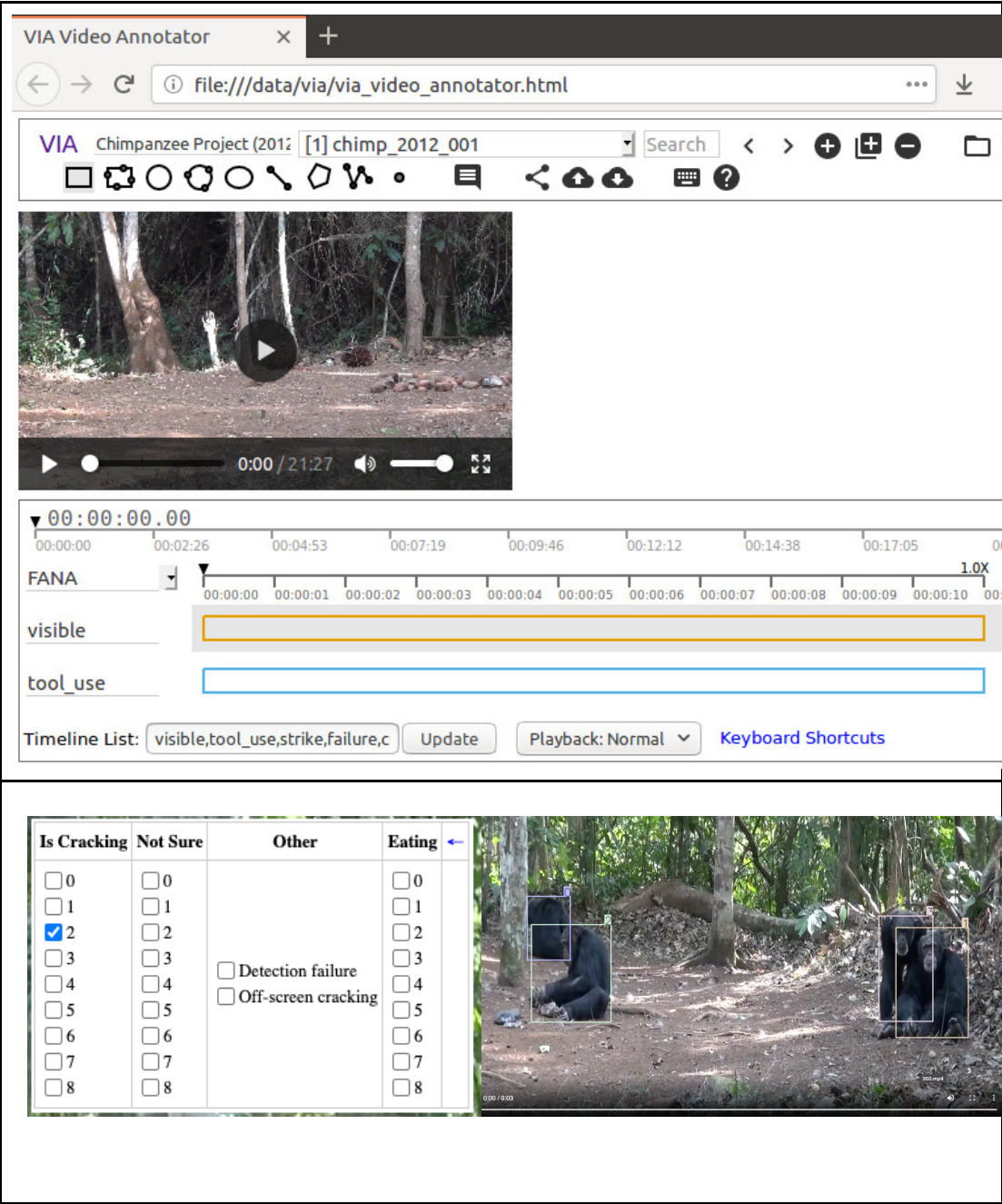


Fig. S2. Screenshots of the annotation tool, the VGG Image Annotator (VIA). VIA is a browser based, open-source annotation software which we modified for this purpose, facilitating temporal annotation of video.

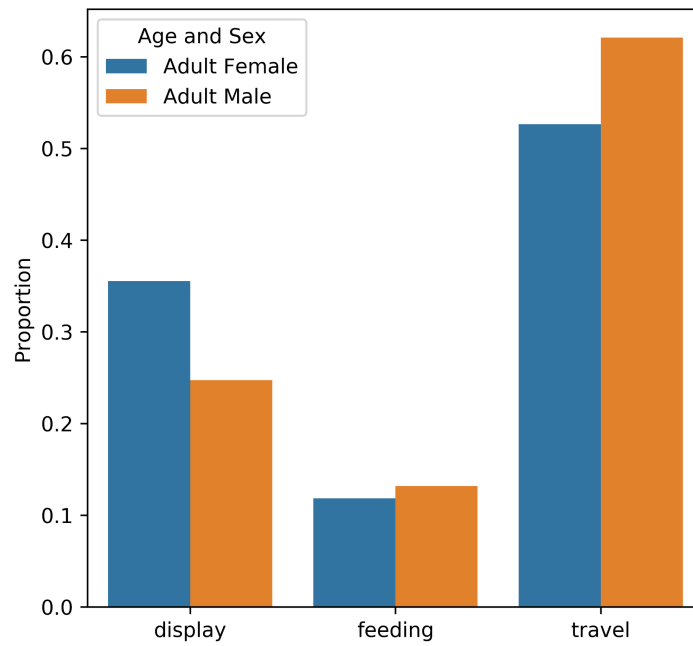


Fig. S3. Proportion of buttress-drumming bouts by context for Adult Males and Females. Context here consists of agnostic display, feeding and travel. Proportions for each context are approximately equal for both males and females.

Table S1. Training and Testing Data Statistics for (A) Nut-Cracking and (B) Buttress-Drumming for the Bossou and Cantahez video footage respectively.

(A)	Split	Sampled Years	Duration of video (hours)	# Nut-cracking Sequences	# Tracklets	# Tracklets containing <i>nut-cracking</i>
	Training	2012	10.5	864	11,400	508
	Training (Audio proposals)	2004, 2008, 2012	1.9	1,455	9,564	1,786
	Testing	2013	2.6	417	3,699	154
	Analysis	2012, 2013	13.1	1,281	15,099	662
(B)	Split	Sampled Communities	Duration of video (hours)	# Drumming Bouts	# Tracklets	# Tracklets containing <i>drumming</i>
	Training	Cabante, Caiquene-Cadique	32.8	778	2,526	1,249
	Testing	Lautchande	1.7	166	521	191
	Analysis	Cabante, Caiquene-Cadique, Lautchande, Cambeque	10.8	1,251	-	-

Movie. S1. Demonstration of automatic behaviour classification pipeline to example videos. Videos are first processed by the body detection tracking stage, followed by audiovisual behaviour classification.

6

Discussion

Contents

6.1	Summary	127
6.2	Key contributions	129
6.2.1	Data preservation and access	129
6.2.2	Deep learning	130
6.2.3	Social behaviour across the lifespan	131
6.2.4	Capturing behaviour	133
6.3	Future directions	134
6.3.1	The future of AI for ethology	134
6.3.2	Biological applications	137
6.4	Conclusion	143

6.1 Summary

The social lives of animals and the interplay of behaviour and ecology has long been of interest in ethology (Hinde 1976a; Tinbergen 1963). Measuring behaviour of our closest living primate relatives is central to understanding our own evolution (Muller 2017). Longitudinal video archives hold serious potential to measure social behaviour in wild animals over large temporal and spatial scales. However, so far, detailed examinations of longitudinal processes, such as cultural transmission and evolution in complex social systems, have been constrained by the lack of methods

for processing and analysing video data. Breakthroughs in computer vision enable a new methodological framework which can capture individual social interactions and behaviour change across the lifespan in animals.

In this thesis, I developed a new paradigm for measuring sociality and analysing behaviour of wild animals dynamically in video. Motivated by lack of time-depth for measuring animal behaviour, and inspired by the latest developments using cloud technologies in film and television video workflows, I pioneered a novel computational framework which leverages video data and computer vision to retrospectively analyse video archives. The scope of this research is broad and challenging, utilising the latest advances in computer science and machine learning, to measure a complex behavioural system in the form of chimpanzee tool use and social networks. The utilisation of longitudinal data, automated frameworks and fine-scale behavioural analysis offer considerable scope for capturing behavioural processes which require time-depth in other species, as well as shedding light on possible factors influencing aspects of early hominin evolution, such as technological and cultural evolution.

This work makes several key theoretical and methodological contributions, with its operational framework and technological advancements for measuring chimpanzee sociality, as well as the managing and processing of video in the wildlife sciences. In Chapter 1, I introduced the theoretical basis and motivation for the thesis, and outlined a novel framework for measuring animal behaviour using computer vision. I identified wild chimpanzee technology as an ideal candidate for applying this automated framework, and introduced the long-term field site of Bossou, Guinea. I then outlined a framework for using computer vision to utilise time-depth in video archives of wild animals for network analysis. In Chapter 2, I presented the Bossou archive dataset, and introduced the methodology and rationale for processing and managing large-scale video content archives. In Chapter 3 I showed that face recognition and deep neural networks can be applied to challenging data in the wild, and the output of this system can be used to generate social networks automatically. In the following Chapter 4, I evaluated the reliability of the method, using it to measure social structure and individual level variation for several sociality measures.

Chapter 5 extended this framework further, layering individual identification with full-body tracking and behaviour recognition.

6.2 Key contributions

Chimpanzees are an ideal candidate to measure behaviour over longitudinal time-scales in the wild, given their complex sociality and rich diversity of cultures. I synthesize data from the classic field site of Bossou, with its long-term video record of field data collection and systematic quasi-experimental research. High quality and detailed observations of nut-cracking behaviour, one of the most complex forms of tool use in wild chimpanzees, make Bossou an important and timely project to tackle new frameworks for time-depth in animal behaviour.

6.2.1 Data preservation and access

Curation of the Bossou video records represents a significant achievement of this thesis. Bossou's pioneering field experiments and collection of high quality image data across multiple years is representative of the emerging pattern of fieldwork in the wildlife sciences. The organisation and/or digitisation of these resources is often a bottleneck which limits access and the potential to leverage long-term data collection. Although other field sites, such as Gombe and Mahale, have systematically collected basic ecological and social data across years, how to efficiently and cost effectively consolidate and extract information from video data remains an unresolved challenge in the field. In Chapter 2, in response to this problem, I introduced a strategy for curating and managing large scale video datasets collected in the wild, laying the groundwork for further analysis of the Bossou archive, and consolidation of other critical archives of wild animals. The population of Bossou is now dwindling at seven individuals; this work ensures the protection of priceless data that can never be collected again, and the historical legacy of one of the few long-term field sites in Africa, opening up exciting new avenues of research for future generations of students and researchers.

Despite reliance on visual data in field research, its effective management is still limited. Innovative strategies for maintaining, accessing and analysing these resources are essential to advance research in ecology and evolution. In Chapter 1, I outline a novel interdisciplinary framework at the intersection of computer vision and ethology, which utilises several key components, to transform how behaviour can be modelled in the wild: long-term video archives, automated individual identification, behaviour recognition, and social network analysis. The set of protocols for managing, processing and analysing video data introduced using this computational framework will ultimately improve reproducibility in the field, open up new avenues of research and enable behaviour to be measured dynamically and at scale for animal populations in the wild.

6.2.2 Deep learning

The Bossou video archive presents a unique opportunity to capture how social behaviour, group dynamics and tool use evolve over generations (see section 6.3.2); however, developing the framework for measuring behaviour change at scale and across different populations requires the adoption of new technology and methods. In this thesis, a significant advance has been the creative application of computer vision to analyse behaviour from video footage. In Chapter 3 I make a key advance in the field, by developing face recognition for individual identification using deep neural networks, which performs effectively on challenging archival video in a wild setting. Repeated measurements over the lifetime of an individual are critical to record the development and fitness impacts and behavioural change over time (Clutton-Brock and Sheldon 2010a); individual detection, tracking and recognition is the first step towards this goal. Previous methods attempting to solve this challenge were limited to recognition on static images, trained from small datasets; these models inevitably failed when applied to challenging datasets such as archival footage or field data. Our approach, outlined in Chapter 3 introduced novel annotation frameworks which streamlined collaboration between engineers and primatologists. The use of tracking and detection to automatically produce tracks for labelling

increased the size of the dataset, and the ease and speed of annotation. By sampling the same individuals in a range of challenging poses and contexts, we demonstrated that deep learning provides a robust method to reliably recognise individuals in different life stages and over generations.

6.2.3 Social behaviour across the lifespan

Developing robust recognition systems for wild settings is a challenging, however the application of such automated systems to extract information for behavioural analysis adds an additional layer of complexity. Social relationships are core to animal societies (Rubenstein and Abbot 2017); for species with long-term, structured and stable social groups, there is strong selection for behavioural strategies to maintain relationships, as this can have key impacts on survival and reproduction (Thompson 2013). Quantifying social structure is thus critical to examine how social behaviour can relate to aspects of ecology, demography and life history at the individual and population-level (Shizuka and Johnson 2019). Social network analysis provides a useful framework for measuring the transmission of behaviour over time, and thus quantifying a network and social relationships remains a core first step in modelling cultural evolutionary processes more effectively. In Chapter 4, I leverage deep learning using neural networks to generate social networks, and validate the system for generating robust social networks and sociality metrics commonly used in the field. Using this system for the first time, I highlight the feasibility of automatically measuring primate associations dynamically across space and time in video. Typically, automated remote-sensing trackers or loggers can map the spatial movement and social associations between individuals, either based on co-occurrence at a specific location or proximity. I further delve into the practical application of an alternative method, using face recognition for producing social networks and sociality scores from video footage. Automated methods are now common in field research, however few papers evaluate the robustness of the system for collecting data using different sampling strategies or in different contexts (Davis et al. 2018). In Chapter 3 I demonstrated the high performance of neural networks

for individual recognition; however, further work was required to test whether the system can reliably produce robust networks, whether there is significant bias, and how sampling can best be used to optimise data collection. I show that oversampling using the methodology does indeed need to be considered before any behavioural analysis can take place. Although the system is robust at providing repeatable social network positions for individuals, sampling at extreme time intervals (too high or too low) can produce spurious results. Manually conducted data collection suffers from the issue of single based sampling approaches; a core advantage of using this automated framework is that a range of sampling values can be easily generated and compared in order to document the effect on biological results, facilitating a more robust use and evaluation of the approach in different contexts.

Having established the validity and robustness of the method, I explore how such a system can be used to examine aspects of individual social behaviour. Age and sex are key factors in shaping a multitude of facets in primate society, from dominance hierarchies, to reproduction and social structure (Machanda and Rosati 2020). In human society, the protracted lifespan has benefits for cooperative offspring care (Kramer 2010). Aspects of life history and ontogeny shape critical windows of learning for the adoption of cultural traits, such as nut-cracking (Matsuzawa, Biro, et al. 2008b). Social structure and the distribution of age and sex may impact a group's long-term viability, by disrupting group movement and sociality patterns (Albery, Clutton-Brock, et al. 2022). In isolated ageing communities like Bossou, eventually this could reduce learning events or proficient models for the generational transmission of behaviour (Rosati, Arre, et al. 2016). Longitudinal timescales are powerful in enabling close observation of individual behaviour changes across the lifespan. In primate societies, ageing is key in shaping how individuals modulate their relationships (Alberts 2019). Research has shown that chimpanzees display age effects in other communities, with individuals becoming more isolated in their old age (Rosati, Hagberg, et al. 2020); however, it is unclear how this affects social dynamics and the transmission of culture over generations and in different populations. Using information on the presence of individuals across time and space, crucial information

regarding relationships between individuals across the lifespan can be examined; at Bossou, this revealed the social stability of the small, highly interconnected community, despite loss of individuals. In the context of our co-occurrence dataset from the outdoor lab, age and sex did not significantly impact an individual's gregariousness; however, this automated framework has great potential to examine the landscape of sociality in different communities and contexts. Contrary to other findings which indicate social behaviour declines with age in certain species, our results suggest that small interconnected communities may buffer against the impact of age and individuals' social behaviour. Comparison of how specific ecological and community contexts can affect sociality patterns is important in highlighting inter-group variation in social behaviour over the life span, and further comparisons should be made before making species-level generalisations. This work is essential to lay the foundations for further exploring how sociality shifts across the lifespan for individuals and communities in the wild. Ultimately information about social structure, life history and demography will be crucial to incorporate into models of time-depth and cultural change (Shizuka and Johnson 2020).

6.2.4 Capturing behaviour

Individual identification and quantification of social structure and individual relationships, are key for the foundations of long-term community change at Bossou in response to ecological and anthropogenic pressures. Although deep learning is effective at generating social networks and sociality measures, animal societies are multi-layered; measuring the full complexity of animal movement and interactions over time in wild species requires methods for measurement of behaviour, for example to examine how tool use is socially transmitted through a population (See section 6.3.2). Nut-cracking represents one of the most cognitively complex and social behaviours in chimpanzees, passed down over generations, and thus is a useful candidate for applying deep learning to automatically capture individual behaviour over longitudinal time-frames. In Chapter 5, I outline a further advance, combining individual recognition with behaviour recognition, to provide a more complete time

series of individual movement and behaviour in video. This work demonstrates the generalisability of this automated framework by extending training data to other communities and behaviours (e.g. Cantanhez National Park, buttress drumming). Our system works effectively on both old archival footage, as well as remote camera traps from multiple communities, to show that the methodology is feasible for a range of challenging ecological settings and data types. Preceding work has focused on the visual domain, despite animal behaviour often being complex and multi-modal. In the field, researchers rely on other cues to measure or recognise behaviour, such as vocalisations, or audible movement. Utilising a multi-modal approach to detect behaviours such as nut-cracking is efficient both in terms of pre-screening for annotation processes (i.e a low cost way of locating behaviours of interest in archives), and in terms of the robustness of models for performing on footage that is challenging and variable in quality, where either video or audio are not clear.

6.3 Future directions

6.3.1 The future of AI for ethology

Computational frameworks for scientific collaboration

The Bossou archive is already supporting several projects; in future, moving towards a unified system which leverages use of cloud resources, remote collaborative editing and authenticated user access would be a blueprint for how research teams store, share and validate important datasets. This work is essential to streamline the workflows between model training, and behavioural/biological annotation. Such advances will help support the development of new deep learning tools, and their flexible deployment on video in the wildlife sciences. More robust infrastructure for managing large-scale video archives and complex annotation projects is needed to ensure these tools become accessible and reproducible. This also applies to primate labs where real-time analysis of live camera feeds could direct research efforts and improve animal welfare and researcher safety. For conservation applications, live ‘edge’ processing of video on small single board computers like the Raspberry Pi could trigger alerts to be sent across low-bandwidth remote networks, reducing the

costs of collecting processing data and thereby helping to continuously monitor ecosystems and biodiversity.

In this thesis I have shown how deep learning can transform how ethologists conduct behavioural analysis of species in the wild; however, such approaches are not without their challenges. Deep neural networks are data heavy, and to achieve the benchmark results many images had to be annotated. Coordinating training and testing with these annotation frameworks is intensive, after which processing and model evaluation is still required before any tools can be used. Developing the systems is an initial challenge, but then applying them effectively for behavioural analysis is not straightforward; extracting data from deep learning pipelines can be technical, and using automated frameworks may reduce human error but also introduce machine error, which requires extensive validation.

The work outlined in this thesis is also a story of scientific collaboration between disparate fields. Development of the systems requires considerable training and expertise in computational fields, and collaboration with the Department of Engineering was essential to achieve the objectives within the timeline of a thesis. Fields in machine learning and engineering sciences are extremely fast paced and expanding rapidly. Papers are often published as open source conference proceedings, and direct application of technology represents only a small part of the field - systems are often designed to beat benchmarks using constrained datasets to test theoretical computer-vision-specific problems, rather than tackle noisy, real-world applications. Due to the time-consuming and labour-intensive nature of field-based disciplines such as primatology and other biological fields, scientific process is slower, and data collection, conceptualisation and publication in scientific journals often extend the research timelines. Academic groups are naturally protective of this hard-won data. Consequently, collaboration can be challenging, given the differing emphasis and value placed on technical development and application, as well as the high level of accuracy and reliability required of any systems in real world research settings (Kerner 2020). This thesis represents

one of the first attempts to bridge these disparate fields and to extend computer vision applications to analysis of behavioural data in the field.

Towards open data and AI in the wildlife sciences

With those caveats in mind, several approaches are making the field more transparent, accessible, and powerful. As for any scientific collaboration, establishing good relationships, with clear outcomes and expectations is essential. Moving towards interdisciplinary research with researchers working closely with engineers should aim to produce systems and papers which credit both the development and field-specific application on both sides of the collaboration. Ultimately, technical advances are moving deep learning towards more flexible and deployable systems. Moving towards open datasets in the field of primatology can be made possible by means of remotely accessed and controlled video databases in the cloud, which protect both the use and accessibility of such valuable assets. Establishing reproducible workflows through open source software projects, which can be used ‘out the box’ by researchers with limited engineering experience has been done successfully for limited applications such as pose estimation in neuroscience (Mathis et al. 2018). Increasing the variety of animal behaviour training data will create more transferable models and bolster the generalisability of such systems to cope with the inherent visual challenges of footage collected in the wild.

New tools

In addition to more training data, reliance on the supervised approach to deep learning can make way for increasingly flexible automated systems. Chapters 3,4, 5, and Appendix B show the effectiveness of applying visual AI to automate behaviour analysis. So far, most approaches have been based on supervised learning and significant advances have been made in the field using this approach. Nevertheless, this still relies on human annotation which remains intensive for projects requiring large and complex ethograms with many behavioural variables. Future approaches might use self-supervised learning, which leverages existing labels in one domain

to make predictions for annotations in another (for example, using audio of nut-cracking to predict annotations for visual models). Self-supervised learning also has large potential to detect undiscovered behavioural patterns in a population, as researchers are constrained by their own set of knowledge and the limitations of the human visual system (Anderson and Perona 2014).

6.3.2 Biological applications

New directions in ecology and evolution

In Chapter 1 of this thesis, I outlined the core components needed for measuring individual behaviour automatically; however, more work is required to increase the complexity of analysis. Despite the key contribution this research has made for the study of chimpanzee social behaviour from video archives, there are many possible directions for future research. The general computational framework outlined here has a broad range of applications across numerous fields in the cognitive and wildlife sciences, both in the wild and in the laboratory; for example, visual data from remote sensing cameras, and spatial movement from animal tracking devices could be used in tandem to measure the behavioural context of collective behaviour (Hughey et al. 2018), detect species interactions (Høye et al. 2021), or track the gaze of freely-moving individuals in cognitive experiments (Kano et al. 2022).

Measuring sociality

Now that the base framework for individual identification, behavioural recognition and social network generation has been trained, tested and validated, there are many potential avenues of research for measuring sociality. Chapter 4 exhibits how this system can automatically and reliably quantify chimpanzee social relationships and generate robust social networks. Thus far, this system has relied on the ‘gambit of the group’ approach (Franks et al. 2010), which assumes individuals detected together in the same videos are interacting. Chapter 5 presents the final step in unifying an automated framework for measuring behaviour in the wild, creating possibilities for empirical and methodological work which combines social

network analysis with the fine-scale analysis of behaviour. More sophisticated methods can move beyond fixed time-window sampling; behavioural classifiers can be trained to capture a multilayered and complex array of social behaviours; for example, generating social networks via grooming, physical contact, distance or even gaze. At Bossou, applying interaction-based rather than association-based networks may reveal more subtle differences in chimpanzees as they age in the wild - for example, how selective individuals are in terms of whom they spend time with, and how these behavioural interactions change (e.g. such as levels of aggression) (Rosati, Hagberg, et al. 2020).

Social transmission

Such work expanding the different ways of quantifying social structure opens up new directions for quantifying social transmission of behaviour through networks. Detecting whether behaviours are learned socially, or individually, is a challenge for the field where variables cannot be controlled experimentally (Whiten, Caldwell, et al. 2016). Coding for networks and time series of behaviours manually is extremely time-consuming, and thus far few studies have mapped the spread of behaviours over long time-frames through network based diffusion analysis. Once trained, our method can be efficiently used to automatically generate social networks, and along with training behavioural classifiers (e.g. eating of a novel food resource) can then examine whether the uptake of this behaviour follows the pathways of social structure within the community (Hoppitt 2017).

Detecting novel behaviours

Another key area in animal behaviour research is identification of spontaneously acquired or modified behaviours never previously observed, otherwise defined as ‘innovations’ (Nishida, Matsusaka, et al. 2009; O’Brien and Shennan 2009). Capturing innovation or modifications of behaviour dynamically remains one of the biggest challenges, for example, to measure the open-endedness of cultural evolutionary change - detecting variation in behaviour and breaking down a fluid

behavioural sequence into arbitrary discrete units is both challenging and time-consuming for human researchers. In future, deep learning classifiers can be used to examine behaviour in increasing detail to search for patterns and subtle variation. This is not always feasible in video footage but has been done to great effect using manual annotation; for example, for action pattern analysis breaks up traits into discrete units to reveal the structure and complexity of a behavioural trait such as tool use (Tan et al. 2015). Developing classifiers which are unsupervised (clustering novel potential behaviours independent of pre-defined behaviours by human annotators) will enable modelling processes beyond a single behavioural variable in isolation - e.g nut-cracking or eating a single species of nut - to measure the drivers of the size and diversity of full behavioural repertoires over time.

Efficiency and complexity

A core aspect to ethological analysis is the quantification of the evolutionary fitness of behavioural traits. Measuring efficiency, from the kinematics of movement to the performance and energetic pay-off of tool use, is important to many aspects of ecology and evolution, and is seen as a driving force of adaptation and behavioural change, with complexity an emergent property of selective forces (Gruber, Chimento, et al. 2022). In Chapter 5, we show how efficiency of nut-cracking can be measured as a function of the proportion of time spent nut-cracking to eating. More fine-scale methods of identification can be used; for example, ‘beat’ detection to count frequency, and timing for strikes to calculate individual efficiency metrics. Quantification of efficiency need not be using action recognition classifiers - pose estimation can be used to detect body parts and movement. Two dimensional pose estimation can capture landmarks, but now more sophisticated methods can infer three dimensional structure of objects. Such approaches are bolstered by the ability to attach Inertial Measurement Units (IMUs), which can be used to reconstruct human 3D pose orientation and acceleration in outdoor environments (von Marcard et al. 2017). New approaches could use motion sensors to test this in the lab, and apply it to naturalistic video contexts. These approaches provide promising avenues

for quantifying the efficiency and complexity of motion from video; this can help show the selective advantage of different traits that may be driving their behavioural adoption. A sophisticated development could use gaze direction to capture how direct observation and learning events correlate with movement patterns and the adoption and dissemination of behaviours. Gaze direction in the wild has been conducted in computer vision and is easily translatable to animal behaviour now that individual tracking has been implemented (Marin-Jimenez et al. 2019). Rather than capturing the diffusion of a behaviour at a single location (Hobaiter, Poisot, et al. 2014), learning events can be more dynamically modelled by capturing gaze direction/peering behaviours of different variants across time and space.

Cultural evolution

The combination of these different elements and future developments can be used to model cultural evolution. Capturing how behaviour is progressively modified and transmitted over many generations in the wild remains a considerable challenge in the field of cultural primatology and animal behaviour; most research is conducted on short-term projects, and thus there is a lack of longitudinal datasets for fine-scale analysis. Therefore, significant barriers for studying culture evolution in non-humans are due to bottlenecks in data availability and accessibility, and video archives offer a unique opportunity to test these hypotheses on cumulative culture in non-humans.

The theoretical and methodological framework, and tools developed in this thesis provide the platform to tackle this problem. At Bossou, using social network analysis to quantify the structure of the network, and behaviour recognition to identify nut-cracking events, there is future potential to combine network analysis with behaviour recognition to map any changes to nut-cracking in the Bossou population (such as the adoption of a novel nut-species, or tool or cracking technique). The nutritional and energetic benefits of nut-cracking for individual reproduction and survival seem clear, but selection need not be solely genetic/biological, and cumulative change could also be related to cultural selection on the inherent structure of the behaviour itself to improve its learnability for acquisition and transmission (Gruber,

Chimento, et al. 2022). There has been little work examining the individual-level variation in action patterns, and social transmission and changes in complexity or efficiency of behaviors over generations has yet to be demonstrated in wild populations over longitudinal time-scales. As has been detected in other species (Tan et al. 2015; Schofield, McGrew, et al. 2018), subtle inter-individual variation in action patterns and efficiency may be detectable in nut-cracking at Bossou and may be subject to change over time, such as the use of varying grips, movement of strikes, selection of nuts and choice of tool-use combinations.

The increasingly intensive provisioning of nuts and stones at the outdoor laboratory may have increased activity and exposure to nut-cracking learning opportunities, which could have influenced the culture of nut-cracking, its behavioural variation and efficiency over several generations. Such potential improvements in the behaviour from increased exposure need not be unidirectional, and may also decline in later years with the changing population dynamics of Bossou, as the population age, structure and size changes due to disease epidemics, death and the dispersal of individuals (Humle 2011c). Nut-cracking may constitute a behavioural stasis, or time-frames longer than decades may be needed to detect any behavioural change. If performance of tool use in wild chimpanzees is transmissible over generations, and increases or decreases in complexity, it may challenge the view that culture and hominin technology is static and evolves in a unidirectional or gradualistic manner. Furthermore, it may indicate that elements of CC may have deeper evolutionary roots in the animal kingdom.

Expanding to other populations and species

The research presented in this thesis has been conducted for a single group of chimpanzees; individual recognition models in this thesis were trained on 21 individuals spanning 1996 to 2017. A further ten individuals are yet to be added, and data from the earliest years (1992) needs to be incorporated in order to maximise the generational time and sample size. Expanding the number of individuals from Bossou and other communities will make the systems more robust to varying

ecological settings and data types, ultimately making them more transferable. Incorporating new individuals will improve model generalisability, and has large potential to extend the application to show both cross-community variation across multiple populations, as well as within-community behavioural change over time. We show the potential transferable applicability of these systems by testing our detector model on multiple species, and in Chapter 5 we make a start on this expansion by comparing camera trap data from multiple communities in West Africa. The framework outlined in this thesis has already had a significant impact on the field; collaborations with other sites and teams of researchers (e.g Gombe and the Jane Goodall Institute) are underway to provide face recognition for other chimpanzee populations (*Pan troglodytes schweinfurthii*). An ever-expanding set of populations and groups of primates, alongside greater sampling and time depth, offers exciting new avenues to explore the landscape of sociality and behavioural variation in different ecological contexts and populations.

Conservation

By expanding to other species, application of our models will have a substantial impact on both ethological and conservation science. Increasing scale and diversity of datasets will boost the flexibility and wider uptake of automated frameworks for wildlife monitoring. The work presented here has been supervised and retrospective, but increasingly these systems can be designed to work in situ, to collect data over the long term. Low cost devices which can be used and integrated with remote data backup, will ensure increased coverage of the behaviour of wild populations. Individual identification and tracking is essential for monitoring demography. Behaviour is key to conservation - social networks and culture can inform us about the health and viability of populations (Brakes et al. 2021). For example, automated social network analysis can help to identify keystone individuals for vaccination to prevent disease transmission, and more sophisticated behavioural metrics using behavioural classifiers can help quantify the flexibility and adaptability of populations. Chimpanzees and other primate species are increasingly under threat

(Estrada et al. 2017). The work presented in this thesis preserves vital history of chimpanzees in the wild for future generations of researchers to analyse, as well as new scalable methods which can be deployed to measure behaviour across large temporal and geographic ranges, to monitor and protect primates in the wild.

6.4 Conclusion

The ultimate aim of this thesis evolved from a focus on models for social behaviour and cultural evolution in non-humans, and, of necessity, developed a radical new paradigm in response to the challenges of measuring time depth of animal behaviour in the wild. This framework is extendable to handling video data and leveraging computer vision algorithms for many research fields and could transform the collaboration of multi-disciplinary teams. In Chapter 1, I outlined the value of longitudinal data for measuring the ecology and evolution of behaviour over time, but ethnographic records often lack enough detail to measure fine-scale change. Adopting tools from computer vision will transform the field of animal behaviour and enable the measurement of behavioural variation more dynamically and at scale using video archives. The computational framework presented in this thesis bridges the disparate fields of computer science and primatology. It enables investigation of multiple decades worth of data on chimpanzee tool-use over several generations, and develops a strategy which can be used to examine behaviour from a time-depth perspective across multiple species and populations. The quantification of long-term changes in tool use in a group of wild chimpanzees could have broad implications for modelling our evolutionary origins of culture and technology.

Appendices

A

Cumulative culture in non-humans:
overlooked findings from Japanese
monkeys?

Published in *Primates*



Cumulative culture in nonhumans: overlooked findings from Japanese monkeys?

Daniel P. Schofield¹ · William C. McGrew² · Akiko Takahashi³ · Satoshi Hirata⁴

Received: 17 August 2017 / Accepted: 2 December 2017 / Published online: 27 December 2017
© The Author(s) 2017. This article is an open access publication

Abstract

Cumulative culture, generally known as the increasing complexity or efficiency of cultural behaviors additively transmitted over successive generations, has been emphasized as a hallmark of human evolution. Recently, reviews of candidates for cumulative culture in nonhuman species have claimed that only humans have cumulative culture. Here, we aim to scrutinize this claim, using current criteria for cumulative culture to re-evaluate overlooked qualitative but longitudinal data from a nonhuman primate, the Japanese monkey (*Macaca fuscata*). We review over 60 years of Japanese ethnography of Koshima monkeys, which indicate that food-washing behaviors (e.g., of sweet potato tubers and wheat grains) seem to have increased in complexity and efficiency over time. Our reassessment of the Koshima ethnography is preliminary and nonquantitative, but it raises the possibility that cumulative culture, at least in a simple form, occurs spontaneously and adaptively in other primates and nonhumans in nature.

Keywords Cumulative culture · Japanese macaque · Food processing · Traditions · Ethnography

Introduction

Nineteenth-century anthropologists such as Tylor (1871) and Morgan (1877) championed ideas of a linear, evolutionary progression of human society through cultural stages from ‘savagery’ to ‘civilization.’ Although this thinking has faded away, some disciplines, such as anthropology, archaeology, and psychology, maintain the view that human society occupies a solitary pinnacle in the animal kingdom (Whiten and van Schaik 2007; Mesoudi 2011, 2016). This thinking occurs in often-cited cases of superlative human achievements: space exploration, modern medical technology, invention of calculus, etc. These cases represent the culmination of the achievements of many persons over

multiple generations, each making gradual modifications to the innovative advances that preceded them. The evolutionary process driving this remarkable enhancement is said to be ‘cumulative culture’ (CC), conceived generally as the ever-increasing, additive complexity or efficiency of cultural performance over time. Authors from various academic disciplines assert that CC is what separates *Homo sapiens* from all other living species cognitively and behaviorally (e.g., anthropology, Hill 2009; archaeology, Haidle et al. 2015; psychology, Tennie et al. 2009; Tomasello 2009; primatology, Perry 2009; philosophy, Sterelny 2009; even neuroscience, Somel et al. 2013; but cf. ethology, Whitehead and Rendell 2015).

CC has been characterized as a ‘ratchet,’ yielding progressive innovation and improvement over generations (Tomasello et al. 1993). The process can be seen as repeated inventiveness that leads to incrementally better adaptation; that is, more efficient, secure, convenient, etc. survival and reproduction. The conceptual consensus is that high-fidelity information transmission and cognitively complex social learning in humans facilitates the expression of behavior and the products of behavior, prevents ‘slippage,’ and allows the modification of cultural traits to ratchet up, advancing beyond what any individual could achieve alone (Tennie et al. 2009). Advocates of this position (Galef 2009; Hill

✉ Daniel P. Schofield
daniel.schofield@anthro.ox.ac.uk

¹ Institute of Cognitive and Evolutionary Anthropology, University of Oxford, Oxford OX2 6PE, UK

² School of Psychology and Neuroscience, University of St. Andrews, St. Andrews KY15 9JH, UK

³ National Museum of Emerging Science and Innovation, 2-3-6 Aomi, Koto-ku, Tokyo, Japan

⁴ Wildlife Research Center, Kyoto University, Kyoto 606-3201, Japan

2009; Tomasello 2009) maintain that there is no evidence, either experimental or observational, that any other species possesses CC.

Recent evidence of innovation, transmission, propagation, diffusion, and intergroup variation in behavior in a wide range of species in nature has bolstered arguments for animal culture, stressing continuity between behavioral mechanisms of humans and nonhumans (McGrew 2004; Whitehead and Rendell 2015). Proposed evidence for CC in nonhumans in nature includes tool use in chimpanzees (*Pan troglodytes*; Sanz and Morgan 2009) and New Caledonian crows (*Corvus moneduloides*; Hunt and Gray 2003), social games in white-faced capuchin monkeys (*Cebus capucinus*; Perry 2011), and stone handling (SH) by Japanese macaques (Leca et al. 2007a, 2012). Recent reviews of this evidence have dismissed these claims as flawed and inconclusive, leaving the sceptics unconvinced (e.g., Dean et al. 2014).

All of the above-cited examples (except for SH, see below) rely almost entirely on indirect tests of CC; that is, they are based on inferential, retrospective reconstruction and lack chronological (real-time) evidence of cumulative change. They rely on cross-sectional rather than longitudinal data. For example, termite-fishing tools used by wild chimpanzees at Goulougo are more efficient if they have frayed tips than the unfrayed tips used elsewhere. This suggests that Goulougo chimpanzees have advanced the design of this extractive technology by modifying and so improving their tools' tips (Sanz and Morgan 2009). But there are not yet enough long-term ethological data, nor a corresponding archaeological record, to test this idea at their study site.

Here, we revisit the well-known early studies of the Japanese monkey (*Macaca fuscata*) at Koshima, as potential candidates for cumulative culture in nonhumans.

For over six decades, researchers have observed the monkeys on this offshore islet in southern Japan; it is the longest continuing study of any nonhuman primate species (Kawamura 1959; Matsuzawa 2015). From 1948 to 2016, 627 individuals have been recorded in total, spanning many generations (Takahashi et al., pers. comm.). Imanishi and colleagues from Kyoto University pioneered three new methods in primatology: long-term behavioral monitoring, provisioning to habituate subjects for closer observation, and individual identification. Serendipitously, Imo, a juvenile female, invented 'sweet potato washing' in 1953, and wheat washing in 1956, and the spread of these behaviors throughout the group is a textbook example of spontaneous nonhuman culture, as followed from inception (e.g., Boyd and Silk 2009). This unparalleled collection of longitudinal data on these behaviors over more than 60 years allows a unique opportunity to investigate changes over many generations (Matsuzawa 2015). The potential for CC at Koshima has been asserted before, but only in brief and general terms (Avital and Jablonka 2000; Jablonka et al. 2014).

We aim to re-evaluate the claim that no CC occurs in nonhumans. We scrutinize proposed criteria for assessing CC, based on a recent review (Dean et al. 2014). Tackling CC requires an operational definition, so we devise a framework for revisiting and re-assessing the data from *Macaca fuscata* at Koshima.

Criteria for cumulative culture

To examine the extent or absence of CC across species, Dean et al. (2014) sought to assess the behavioral and cognitive repertoires of species, mostly primates, in the wild. Their assessment entailed a two-step process. The first step was to establish a trait as cultural using the 'method of exclusion' (Whiten et al. 2001). This we take as given for the Koshima macaques, as one of us (McGrew 1998, 2009) has argued extensively.

Dean et al.'s second step required the trait in question to be *cumulative*: that there is direct evidence that the trait has changed over time in a directional or progressive manner, resulting in an enhanced level of *complexity*. Following Tennie et al. (2009), they stated that to be deemed cumulative, a behavioral trait must go beyond what a single individual could have invented alone (Dean et al. 2014, p 5). According to these criteria, no nonhuman species reviewed by them passed the test.

Operationalizing cumulative culture

In order to evaluate the claim that CC is a phenomenon that is evolutionarily unique to humans (Tennie et al. 2009, p 2405), CC needs to be defined more explicitly and precisely, thus allowing systematic, quantitative, and explicit comparisons across species. CC has been defined often in recent years, but most definitions are imprecise and make no attempt to be operational (empirically testable), such as: "...innovations are progressively incorporated into a population's stock of skills and knowledge, generating ever-more-sophisticated repertoires" (Shipton and Nielsen 2015, p 332). We more pragmatically define CC as a modification (change in the sequence or form of behavioral elements) of a cultural trait (i.e., acquired via social learning) that enhances its complexity, efficiency, security, or convenience. We acknowledge that efficiency, being a broad variable, may be too general, thus we specify two aspects of efficiency with obvious adaptive value: security and convenience. Each of these four enhancements requires more discussion. Finally, as used above, modification should be distinguished from 'step-wise traditions,' as proposed by Tennie et al. (2009).

Efficiency and complexity

Dean et al. (2014) distinguished between efficiency and complexity, saying that either can be used systematically to ascertain CC, though few have tackled exactly how to measure these features. Here, we define efficiency as ‘net benefits per unit time of performance of a behavioral pattern.’ Thus, ‘net benefits’ represents the composite, proximate pay-off of energy, time, and risk accrued by an organism. For example, efficiency can be quantified as the amount of food consumed or number of feeding events per unit time, as has been done in studies of chimpanzees in captivity (Yamamoto et al. 2013; Davis et al. 2016) or in nature (McGrew and Marchant 1999; Sanz and Morgan 2009). On a wider, proximal front, efficiency can refer to greater security, convenience, comfort, even pleasure (e.g., Stewart et al. 2007), but no one has attempted to quantify this (so far as we know).

For complexity, behaviors can be understood as hierarchically organized programs of action, or ‘cultural recipes’ in which a series of dependent actions and subgoals are performed to achieve an overall end goal (Charbonneau 2015). The action chain of a hierarchically organized sequence can be broken down (parsed) into separate units, in which each unit is “a single action that results in an observable change to an item” (Byrne and Byrne 2001, p 503). Although fluid and continuous behaviors may be split or lumped arbitrarily differently in such parsing, the hierarchical and sometimes recursive unfolding of these actions can be quantified as a measure of the cognitive complexity of a task (Byrne and Byrne 2001; Byrne 2007). Such systematically structured chains appear to be transmissible between individuals (Claidière et al. 2014).

Modification

Dean et al. (2014) distinguished accumulation from modification. They defined the former as “...addition of knowledge or behavior patterns to the behavioral repertoire of an individual or population,” (p 4) such as the addition of a new food item to a diet. For example, if a chimpanzee that already eats oranges then adds lemons (Takahata et al. 1986), this is just an accumulative augmentation of another citrus fruit to the diet. Such dietary enlargement does not constitute CC, as it does not entail the modification of a trait so that complexity or efficiency is increased. However, Dean et al. did not specify exactly what modification is. For example, if a chimpanzee cracks a nut using a wooden hammer when it previously only used stones, is this accumulation or modification? The new raw material, wood, might have superior qualities, such as being in more plentiful supply, making it more efficient to obtain. That is, the change of hammer type could represent CC, or it might be just the substitution of another raw material to produce an alternative

percussor. Here we define modification as a change to the hierarchical sequence of cultural acts (i.e., behavior acquired by social learning), by addition, deletion, or substitution, which increases the effectiveness of completing a task or attaining a goal. By this behavioral standard, a change from stone to wooden hammers would be only accumulative, not CC (see Luncz et al. 2015 for reporting of stone–wood hammer choice and change).

Japanese macaques

We now present key ethnographic data from a nonhuman primate species that is a potential case of CC. Japanese macaques inhabit a wide latitudinal range, from the subtropics to snowy mountains, making them useful for both inter- and intragroup comparisons of behavior. Japanese monkeys are behaviorally flexible and innovative, displaying stone handling (Nahallage et al. 2016), aquatic thermoregulation (Zhang et al. 2007), exploitation of marine resources (Leca et al. 2007b), specific forms of social interaction (Nakagawa et al. 2015), and food washing (Hirata et al. 2001). Of these, only for SH has the case been made for CC (Leca et al. 2012). Decades of data show cultural change, with the repertoire of SH elements increasing in number and diversity, but this seems to reflect accumulation rather than modification resulting in clear, cumulative progression. No evidence has been presented that SH is adaptive (functional), which suggests no increase in efficiency, etc. All of the SH variants are independent, simple behavioral patterns that lack complexity, so we exclude SH as exemplifying CC. Although the Koshima macaques are well documented and often cited in cultural primatology (de Waal 2001), their basic ethnography has usually been inexplicably ignored; here we seek to clarify their status.

Koshima is an islet of 32 ha that is 300 m off the Kyushu mainland, in Miyazaki Prefecture, Japan (Watanabe 2001) (Fig. 1). The island has two main ecotypes: a hilly area (reaching 113 m asl) covered in thick evergreen forest, and a sandy beach and shoreline on the west side of the islet, where the macaques have been provisioned since 1952. A freshwater stream runs from the forest through the beach to the sea. The macaques have always lived in the forested areas, but after provisioning started they began to emerge intermittently to forage on the shore, where they still spend much of their time (Watanabe 1994) (Fig. 2).

Sweet potato washing

Food washing represents hierarchically organized sequences of behavior, and has been observed elsewhere in macaques (e.g., *Macaca fascicularis*; Tan et al. 2016). At Koshima, researchers began provisioning to tempt macaques onto the beach for clearer observation conditions. The macaques



Fig. 1 View from shoreline of the beach and forest on Koshima island (Photo by Akiko Takahashi)



Fig. 2 Japanese macaques on the beach at Koshima (Photo by Akiko Takahashi)

were given mainly two staples: unwashed sweet potatoes were dumped and unhusked wheat grains were scattered on the sand. (Wild monkeys are well known to be crop-raiders that “grub out” sweet potatoes and pilfer rice from farmers’ fields; Kawamura 1972.) Initially, the macaques used their hands or body hair to brush away sand from the potatoes (Watanabe 1994). However, in 1953, a 1.5-year-old juvenile female named Imo started ‘sweet potato washing’ (SPW) (‘dip and brush’ in Table 1), and the behavior quickly spread to other individuals in the group (Kawai 1965). Four phases

of transmission followed. SPW initially spread by horizontal transmission among some of Imo’s immature peers. Then came transmission from young to old vertically upward from child to mother, and obliquely from younger to older siblings, followed by horizontal transmission to other adults. As SPW became more established, it spread from parous females to their offspring via downward vertical transmission (Hirata et al. 2001; Huffman and Hirata 2003) (Fig. 3).

Soon, SPW diversified to new variations (detailed in Table 1). It moved from fresh to salt water, from flinging to hand-held, then eventually to individual pools dug in the sand. Some changes, such as scavenging from others, seem less surprising, but there are no reports of such scavenging before provisioning began. So, we cannot know the relevant details of its emergence, but researchers at the time considered it notable enough to record as novel. These changes comprised seven progressive, cumulative steps, but not necessarily linear ones. Complexification need not be sequential, as diverse behavioral patterns may cross-fertilize one another in a kind of synergistic process (yet unstudied). For example, although scavenging and pirating by individuals may not be cumulative steps, these behaviors may have driven modifications for labor saving and increased protection of food (Kawai et al. 1992). Although SPW initially was hypothesized to enhance the palatability of the potatoes, recent findings show that SPW correlates with lowered

Table 1 Sweet potato processing chronology after provisioning at Koshima began in 1952. Kawai et al. (1992), Watanabe (1994), and Hirata et al. (2001)

Stages of apparent cultural change	Year first observed	Description of acts	Cumulative improvement
1. Brush	1952	Brush sand brush off with hand or fur	Cleaner foodstuff reduces wear on teeth from sand. Hygienic treatment may reduce risk of parasites
2. Dip and brush	1953	Dip potato in stream with one hand, and brush sand off with other	Washing more effective at removing sand, grit and soil than dry ‘brush’ variant
3. Immerse and roll	1955	Potato immersed and rolled underwater in stream	More vigorous treatment more effective at removing sand, etc. than variants 1 and 2
4. Rinse saltwater	1957	Wash potato in sea water	Wave action removes more sand, grit and soil. Flavour of potato enhanced (gustation)
5. Dip and gnaw	1958	Dip potato in sea water between repeated bites	Flavour of potato further enhanced bite-by-bite
6. Scavenge	1983	Gather pieces dropped/discarded by others	Reduced labor as less time and energy spent washing. Less risk of food being pirated?
7. Plunder	1983	Attack/threaten rinser and rob of cleaned potatoes	Reduced labor as less time and energy spent washing. Bigger portions of potatoes than 6
8. Private pool	1983	Dig own separate, more secluded pool for rinsing potato	Solitary eating decreases risk of scavengers/plunderers. Less stress means less hurried eating

**Fig. 3** Sweet potato washing on the beach at Koshima (Photo by Akiko Takahashi)

geo-helminthic infection (Sarabian and Macintosh 2015). Emergence of new hygienic variants (e.g., from simply dipping and brushing to immersing and rolling) may represent (unintentional but potentially useful) cumulative progress in reducing the risk of acquisition of harmful parasites.

Wheat washing

‘Wheat washing’ (WW, also called ‘sluicing’ or ‘placer mining,’ Hirata et al. 2001) entails wheat grains being scattered/dropped/rinsed in water. Initially when wheat grains were scattered on the beach, the monkeys painstakingly picked up the individual grains one-by-one with thumb and forefinger opposition. The first behavioral variant emerged in 1956, when Imo picked up a mixture of wheat grains and sand from the beach, carried this mixture to the water’s edge,

**Fig. 4** Japanese macaque washes wheat on Koshima beach (Photo by Akiko Takahashi)

and flung it into the water. The sand sunk to the bottom and separated from the wheat, which floated and was scooped off the surface. This invention of WW followed a similar transmission process to SPW, with initial spread among infants through play relations, their siblings, and mothers, and—after they matured and reproduced—vertically down to their offspring (Hirata et al. 2001) (Fig. 4).

After Imo’s innovation, more enhancements were added to the repertoire of WW (Hirata et al. 2001; Kawai et al. 1992; Watanabe 1994) (Table 2 gives details of the chronology). In 1959, specialized ‘snatchers’ reduced their own labor costs by threatening and plundering the washed wheat of other individuals. Other labor-saving or security-increasing innovations included scavenging lost grains floating downstream from other wheat-washers

Table 2 Wheat washing chronology after provisioning at Koshima began in 1952. Kawai et al. (1992), Watanabe (1994), and Hirata et al. (2001)

Stages of apparent cultural change	Year first observed	Description of act	Cumulative improvement
1. Throw	1956	Carry and drop sand/wheat mixture into water, separating wheat from sand. Skim floating wheat off surface	Grains separated from grit, so more easily consumed. Cleaner foodstuff saves wear on teeth and may reduce risk of parasites
2. Attack and plunder	1959	'Muggers' threaten and attack wheat washer and take grains	Labor-saving, as less energy expended collecting and washing
3. Scavenge	1962	Collect wheat grains floating downstream or in sea	Labor-saving, as less energy expended collecting and washing
4. Dribble	1970	Grasp sand and wheat mixture in hands, then repeatedly drop small amounts into water to prevent snatching by others	More secure processing against attacker/plunderers. More controlled to reduce loss of grain in water current or waves
5. Sweep	1971	Sweep wheat grains next to water's edge by hand into water	Labor-saving: less energy expended collecting and washing
6. Screen	1974	Grasp mixture in hands, then shake in water. Sand removed with grains remaining in clenched fists	More controlled than throwing. Increased security against attacker/plunderers
7. Mobile screen	1974	Walk into water while grasping mixture in hands while screening	Increased security against scavengers/plunderers. Labor-saving, and more controlled than the other variants—loss of wheat grains prevented
8. Private pool	1983	Dig small depression in sand, then sweep wheat grains into resulting pool and skim off surface	

('collecting'). With the accrued benefits of increased energy pay-offs from these subsistence strategies, population numbers swelled. In 1972, provisioning was reduced, leading to declines in population size. However, when the provisioning was reduced, the diversity of WW increased (Kawai et al. 1992; Watanabe 1994), and its techniques increased in complexity and efficiency. Having initially thrown the grains into the water, the macaques began to use a more secure hand-held technique, 'dribbling,' followed by 'sweeping,' 'screening,' and 'mobile sweeping.' Finally, individuals began to dig private pools in the sand for more sequestered, focused rinsing. These simple but effective measures protected against pirating snatchers and also reduced inadvertent loss of grains. Despite the end of regular provisioning in 1973, young monkeys still engage in WW (authors' observations), as obtaining wheat 2–3 times a week is enough to sustain WW (Takahashi et al., pers. comm.).

Evaluating the Koshima findings

Some critics (e.g., Laland and Hoppitt 2003) have discounted the Koshima findings because they emerged from provisioning. Opportunities offered by artificial feeding apparently triggered Imo's innovation, but humans made no further contribution to the monkeys' cultural evolutionary change. De Waal (2001, p 207–209) seems to have refuted this criticism, based on first-hand reports from Koshima. Researchers did not devise an experimental study; rather, they monitored the serendipitous appearance of behavioral novelty. The monkeys currently are provisioned with no more than 3 kg of wheat overall, 3 times a week; this supplement forms only a small proportion of their overall diet, most of which is natural (Takahashi et al., pers. comm.). The monkeys range freely in nature, so provisioning is merely an affordance provided by humans that the monkeys have exploited innovatively after encountering a newly available but intermittent food resource.

Other criticisms are that the behaviors shown by the Koshima monkeys are simple patterns common to all macaques, and so are not beyond what each macaque could learn individually (Visalberghi and Frigaszy 1990; Tomasello 1999), and that behavioral innovations spread too slowly to generate cumulative change (Galef 1992, but cf. de Waal 2001, p 209–210). The initial behaviors of hand-rubbing potatoes and picking up individual wheat grains may be basic to all macaques (although we have seen no evidence presented for this claim). However, the later Koshima behaviors, entailing more efficient and complex action, seem not to have been reported for any other species of *Macaca*. Instead, the emergent, later behaviors were unexpected and seem to show innovation and enterprise. The speed of diffusion or lack of it needs to be compared with the spontaneous spread

of other habits in other natural primate populations before assessing whether or not it was slow. Comparative data on diffusion rates of behaviors in Japanese macaques indicate that the behavioral type rather than purely the social context in learning best explains varying rates of diffusion (Huffman and Hirata 2003).

Finally, critics claim that SPW and WW behaviors were learned either individually by trial and error or by local/stimulus enhancement, in which one individual's behavior at a locality 'enhances' the stimulus, increasing the probability that a similar discovery will be made by another individual. Essentially, the claim is that each monkey 'reinvented the wheel' (Tomasello 1999), and the behaviors therefore might not be culturally transmitted (dependent on one's criteria for transmission). Thus, according to this viewpoint, the behavior of the Koshima macaques changed sporadically, even randomly, perhaps through a 'drift-like' process (Koerper and Stickel 1980), rather than by a progressive 'ratchet-like' process. At best, the critics say, macaque innovations such as SH are corruptions of existing behaviors that are "inaccurately transmitted between individuals without any further addition of complexity" (Dean et al. 2014, p 8). However, the published ethnographic data from the monkeys show patterns of spread affected by age, sex, and kinship, but not random appearance. Different matrilineal lines showed preferences for specific variants of SPW and WW (Hirata et al. 2001). Not surprisingly, the snatching of others' grains in WW was performed more by monkeys of dominant lineages (Kawai et al. 1992). Regardless of the social learning mechanisms involved (see below), enough change has occurred to suggest potentially cumulative increases in complexity and efficiency of the washing behaviors, as derived from 60 years of data from Koshima.

Discussion

The descriptive post hoc data presented here can be only suggestive, not conclusive. Qualitative reports need to be succeeded by quantitative testing of hypotheses. This is clear in principle but uncertain in practice, being dependent on current (not historic) conditions at Koshima. But, in the meantime, the qualitative reports reflect topical issues relating to CC, as outlined below.

High-fidelity transmission

The predominant social learning mechanisms of Japanese macaques remain unclear, and little research has been done on captive populations (Hirata et al. 2001). Nahallage et al.

(2016) suggest that SH is transmitted through stimulus enhancement and response facilitation. However, even if food-washing behaviors spread through simple forms of social learning such as these, they seem to have enabled the occurrence of CC. Further, recent evidence from studies of other nonhuman species indicates that culture occurs with simpler transmission mechanisms than previously thought (Logan et al. 2016). This suggests that high-fidelity learning, which is so often stated to be essential to CC (Lewis and Laland 2012; Tomasello 2016) may not be necessary (Sasaki and Biro 2017). Recent experimental evidence on humans suggests that high-fidelity social learning is useful but not necessary to generate CC (Caldwell 2015; Zwirner and Thornton 2015; but cf. Wasielewski 2014). Observational data from the Aka hunter-gatherers of the Congo basin suggests that most skills are learned by passive observation, not necessarily by direct teaching or imitation (Hewlett et al. 2011).

Individual and social learning

Although the analysis of Dean et al. (2014) was acute, their criteria for innovation seem overly restrictive. In particular, their criterion that a trait must be beyond individual innovative capacity is problematic, as hypotheses that are framed in the negative, i.e., 'X cannot do something' cannot be verified empirically, as it is logically impossible to prove the absence of something. It also seems nonsensical, as each actual innovation expands the imaginable limits of what any individual or species can or cannot do, ad infinitum. Nonhumans repeatedly surprise us by their inventiveness, making it impossible to say a priori what could or could not be achieved (Kummer and Goodall 1985; Nishida et al. 2009). But what of space shuttles, or mobile phone technology? The argument is that these cases would be impossible to invent *de novo*, and nonhumans have no such innovations. Although these cases from modern industrial society are impressive, such arguments commit the 'Space Shuttle Fallacy' (McGrew 2004). Most individual humans have not done these things, and using this criterion would exclude populations of *Homo sapiens* (e.g., traditional hunter-gatherer societies). Citing a hypothetical "zone of latent solutions" in nature, demarcated by the upper boundary of a species' cognitive skills (Tennie et al. 2009), does not help matters unless its validity is empirically testable in situ. This is a formidable challenge, and we await operational criteria (rather than proposed features) that will be applicable in an ecologically valid context.

Culture evolves and devolves

Contrary to the ratchet effect, proposed by Tennie et al. (2009), culture does not seem to evolve in the simple,

unidirectional progression that the analogy implies, and in theory there is no reason why CC should always be heading for more complexity or efficiency. Culture change does not equal CC; environments or demographics change and culture responds (Kolodny et al. 2015). For example, over millennia during the Holocene, Tasmanian hunter-gatherers lost valuable skills and technologies, apparently due to the population bottleneck from rising ocean levels at the end of the last glacial epoch (Henrich 2004). Critics claim that non-human traditions are few, predictable, and transient, but just as in human culture—under suitable conditions—nonhuman cultural traits in nature apparently increase in prevalence, complexity, and efficiency, as evidenced by the Koshima monkeys. Their behavior was not linearly progressive, as the ratchet implies, but often flexible, interchangeable, and intermittent (Avital and Jablonka 2000; Jablonka et al. 2014). Those ‘stages’ did not unfold one after the other; variants multiplied over time. Cultural variants seem to develop and fluctuate based on various factors, such as the inextricable interactions of environment, demography, and perhaps even gene frequencies. The Koshima monkeys show that with minimal exposure to wheat and sweet potatoes, cultural traits emerge and persist, apparently with improvement. However, as we have stressed above, hypothetical assertions require empirical testing to be conclusive.

Time depth

If nonhumans generate CC, why is there so little evidence of it? The Koshima data suggest that long-term data may be required to detect CC. Most studies and comparative data from the wild are from relatively brief snapshots (months or years rather than decades), which may be why nonhuman CC is so elusive. Time per se is not the issue, but rather the long generation times of large-brained, K-selected mammals, which may mean that decades are needed to compile the data. Many generations have followed on at Koshima after Imo’s innovations, but tracing the precise lineages for each behavioral pattern across multiple generations remains to be done. We know of a few such efforts to use chronological, archival data retrospectively to reconstruct cultural change (e.g., diffusion of ant fishing in Gombe chimpanzees, O’Malley et al. 2012), but we know of no such efforts to infer past CC.

Whether or not CC is unique to humans also depends greatly on how the phenomenon is defined. As with traditional arguments used against culture in nonhumans, using social learning processes (e.g., teaching and imitation) both to define and as evidence for CC is logically flawed, and restricts CC to humans; for this position to be valid, high fidelity mechanisms must be shown to be necessary for CC, not simply more effective (Gruber 2016). Definitions vary, but so long as they are precise, explicit, and operational, the

question of CC in nonhumans can be addressed in a comparative framework. Although cultural evolution is more rapid than genetic evolution, a challenge for comparative studies of CC is that macroevolutionary change in human evolution seems to operate at several orders of magnitude faster than in other taxa. Despite this challenge, the macaque evidence seems to show that 60+ years of study could be enough to trace progressive change. Imo’s initial innovations enhanced the performance of basic subsistence activities and triggered the transmission and modification of other variants, which seem to have increased in complexity, security, convenience, and efficiency.

Conclusion

So, why have the Koshima data, which have long resided in the public domain, been ignored in current discussions of CC? Perhaps present-day commentators have not carefully read the original, older ethnography (Kawamura 1959; Kawai 1965) or detailed synthetic accounts of its progression (Itani and Nishimura 1973), even if they sometimes cite these primary or secondary sources. Or, they cite later publications (Hirata et al. 2001; Kawai et al. 1992; Watanabe 1994) without going back to the original reports. Language is not a barrier to access: we have used only English-language sources, but we did not find any additional information available in Japanese-language publications or data archives.

Our review of Koshima ethnography indicates that the food-washing behaviors of the monkeys may have accumulated in complexity and efficiency, consistent with definitions of CC. This suggests that the evolutionary roots of CC are deep in the primate clade. Our reprise of the overlooked Koshima ethnographic record is not conclusive, as the precise changes in complexity and efficiency of the different washing behaviors remain to be tested. An example of how these ideas might be tested would be to survey which variants of the washing behaviors still exist today, and if so, determine which of the enhancements are more frequent when there are more monkeys on the beach or in close proximity. However, these findings at least appear to call into question the overwhelming current received wisdom that nonhuman animals cannot build upon behavioral improvements made by previous group members, and challenges the idea that only humans have history and cultural evolution that other animals lack (Mesoudi 2011). We believe that these data may dispel the idea that human uniqueness is the best null hypothesis and should encourage open-ended future research on nonhuman CC, at least for investigators seeking its evolutionary roots in ourselves. We hope to see human ethological studies of CC in operation in the real-world, spontaneous behavior of *Homo sapiens*.

Acknowledgements We thank: Rachel Kendal and Agumi Inaba for helpful comments and additional information; Corpus Christi College (Cambridge) for funding DS's visit to Koshima; Prof. Tetsuro Matsuzawa for funding WCM's visit to Koshima.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Avital E, Jablonka E (2000) Animal traditions: behavioural inheritance in evolution. Cambridge University Press, Cambridge
- Boyd R, Silk JB (2009) How humans evolved, 5th edn. W.W. Norton, New York
- Byrne RW (2007) Culture in great apes: using intricate complexity in feeding skills to trace the evolutionary origin of human technical prowess. *Philos Trans R Soc Lond B Biol Sci* 362:577–585. <http://doi.org/10.1098/rstb.2006.1996>
- Byrne RW, Byrne JME (2001) Estimating the complexity of animal behaviour: how mountain gorillas eat thistles. *Behaviour* 138:525–557
- Caldwell CA (2015) Experimental studies of cumulative culture in modern humans: what are the requirements of the ratchet? In: Mesoudi A, Aoki K (eds) Learning strategies and cultural evolution during the Palaeolithic. Springer, Tokyo, pp 145–154
- Charbonneau M (2015) All innovations are equal, but some more than others: (re)integrating modification processes to the origins of cumulative culture. *Biol Theory* 10:322–335. <http://doi.org/10.1007/s13752-015-0227-x>
- Claidière N, Smith K, Kirby S, Fagot J (2014) Cultural evolution of systematically structured behaviour in a non-human primate. *Proc Biol Sci* 281:20141541. <http://doi.org/10.1098/rspb.2014.1541>
- Davis SJ, Vale GL, Schapiro SJ et al (2016) Foundations of cumulative culture in apes: improved foraging efficiency through relinquishing and combining witnessed behaviours in chimpanzees (*Pan troglodytes*). *Sci Rep* 6:35953
- de Waal FBM (2001) The ape and the sushi master: cultural reflections by a primatologist. Basic Books, New York
- Dean LG, Vale GL, Laland KN et al (2014) Human cumulative culture: a comparative perspective. *Biol Rev Camb Philos Soc* 89:284–301. <http://doi.org/10.1111/brv.12053>
- Galef BG (1992) The question of animal culture. *Hum Nat* 3:157–178
- Galef BG (2009) Culture in animals. In: Laland KN, Galef BG (eds) The question of animal culture. Harvard University Press, Cambridge, pp 222–246
- Gruber T (2016) Great apes do not learn novel tool use easily: conservatism, functional fixedness, or cultural influence? *Int J Primatol* 37:296–316. <http://doi.org/10.1007/s10764-016-9902-4>
- Haidle MN, Bolus M, Collard M et al (2015) The nature of culture: an eight-grade model for the evolution and expansion of cultural capacities in hominins and other animals. *J Anthropol Sci* 93:43–70. <http://doi.org/10.4436/jass.93011>
- Henrich J (2004) Demography and cultural evolution: how adaptive cultural processes can produce maladaptive losses—the Tasmanian case. *Am Antiq* 69:197–214. <http://doi.org/10.2307/4128416>
- Hewlett BS, Fouts HN, Boyette AH, Hewlett BL (2011) Social learning among Congo Basin hunter-gatherers. *Philos Trans R Soc Lond B Biol Sci* 366:1168–1178
- Hill K (2009) Animal “culture”. In: Laland KN, Galef BG (eds) The question of animal culture. Harvard University Press, Cambridge, pp 269–287
- Hirata S, Watanabe K, Kawai M (2001) “Sweet potato washing” revisited. In: Matsuzawa T (ed) Primate origins of human cognition and behavior. Springer, Berlin, pp 487–508
- Huffman MA, Hirata S (2003) Biological and ecological foundations of primate behavioral traditions. In: Frigaszy DM, Perry S (eds) The biology of tradition: models and evidence. Cambridge University Press, Cambridge, pp 267–296
- Hunt GR, Gray RD (2003) Diversification and cumulative evolution in New Caledonian crow tool manufacture. *Proc Biol Sci* 270:867–874. <http://doi.org/10.1098/rspb.2002.2302>
- Itani J, Nishimura A (1973) The study of infrahuman culture in Japan: a review. In: Menzel EW (ed) Precultural primate behavior. Karger, Basel, pp 25–50
- Jablonka E, Lamb MJ, Zeligowski A (2014) Evolution in four dimensions: genetic, epigenetic, behavioral, and symbolic variation in the history of life. MIT Press, Cambridge
- Kawai M (1965) Newly acquired pre-cultural behaviour of the natural troop of Japanese monkeys on Koshima inlet. *Primates* 6:1–30. <http://doi.org/10.1007/BF01794457>
- Kawai M, Watanabe K, Mori A (1992) Pre-cultural behaviors observed in free-ranging Japanese monkeys on Koshima islet over the past 25 years. *Primate Rep* 32:143–153
- Kawamura S (1959) The process of sub-culture propagation among Japanese macaques. *Primates* 2:43–60. <http://doi.org/10.1007/BF01666110>
- Kawamura S (1972) The pre-culture in the Japanese macaques. In: Centre national de la recherche scientifique (ed) Modeles animaux du comportement humain: Paris, 8-11 decembre 1970. Editions du Centre national de la recherche scientifique, Paris, pp 155–179
- Koerper HC, Stickel EG (1980) Cultural drift: a primary process of culture change. *J Anthropol Res* 36:463–469
- Kolodny O, Creanza N, Feldman MW (2015) Evolution in leaps: the punctuated accumulation and loss of cultural innovations. *Proc Natl Acad Sci USA* 2015:201520492. <http://doi.org/10.1073/pnas.1520492112>
- Kummer H, Goodall J (1985) Conditions of innovative behaviour in primates. *Philos Trans R Soc Lond B Biol Sci* 308:203–214
- Laland KN, Hoppitt W (2003) Do animals have culture? *Evol Anthropol* 12:150–159. <http://doi.org/10.1002/evan.10111>
- Leca J-B, Gunst N, Huffman MA (2007a) Japanese macaque cultures: inter- and intra-troop behavioural variability of stone handling patterns across 10 troops. *Behaviour* 144:251–281
- Leca J, Gunst N, Watanabe K, Huffman MA (2007b) A new case of fish-eating in Japanese macaques: implications for social constraints on the diffusion of feeding innovation. *Am J Primatol* 69:821–828
- Leca J, Gunst N, Watanabe K, Huffman MA (2012) Thirty years of stone handling tradition in Arashiyama macaques: implications for cumulative culture and tool use in non-human primates. In: Leca J-B, Gunst N, Huffman MA (eds) The monkeys of Stormy Mountain: 60 years of primatological research on the Japanese macaques of Arashiyama. Cambridge University Press, Cambridge, pp 223–257
- Lewis HM, Laland KN (2012) Transmission fidelity is the key to the build-up of cumulative culture. *Philos Trans R Soc B Biol Sci* 367:2171–2180. <http://doi.org/10.1098/rstb.2012.0119>
- Logan CJ, Breen AJ, Taylor AH, Gray RD, Hoppitt WJ (2016) How New Caledonian crows solve novel foraging problems and what it means for cumulative culture. *Learn Behav* 44:18–28. <http://doi.org/10.3758/s13420-015-0194-x>
- Luncz LV, Wittig RM, Boesch C (2015) Primate archaeology reveals cultural transmission in wild chimpanzees (*Pan troglodytes verus*).

- Philos Trans R Soc Lond B Biol Sci 370:20140348. <https://doi.org/10.1098/rstb.2014.0348>
- Matsuzawa T (2015) Sweet-potato washing revisited: 50th anniversary of the Primates article. *Primates* 56:285–287. <https://doi.org/10.1007/s10329-015-0492-0>
- McGrew WC (1998) Culture in nonhuman primates? *Ann Rev Anthropol* 27:301–328
- McGrew WC (2004) The cultured chimpanzee: reflections on cultural primatology. Cambridge University Press, Cambridge
- McGrew WC (2009) Ten dispatches from the chimpanzee culture wars, plus postscript (revisiting the battlefronts). In: Laland KN, Galef BG (eds) *The question of animal culture*. Harvard University Press, Cambridge, pp 41–69
- McGrew WC, Marchant LF (1999) Laterality of hand use pays off in foraging success for wild chimpanzees. *Primates* 40:509–513. <http://doi.org/10.1007/BF02557586>
- Mesoudi A (2011) *Cultural evolution: how Darwinian theory can explain human culture and synthesize the social sciences*. University of Chicago Press, Chicago
- Mesoudi A (2016) Cultural evolution: integrating psychology, evolution and culture. *Curr Opin Psychol* 7:17–22
- Morgan LH (1877) *Ancient society*. Henry Holt, New York
- Nahallage CAD, Leca J-B, Huffman MA (2016) Stone handling, an object play behaviour in macaques: welfare and neurological health implications of a bio-culturally driven tradition. *Behaviour* 153:845–869
- Nakagawa N, Matsubara M, Shimooka Y et al (2015) Embracing in a wild group of Yakushima macaques (*Macaca fuscata yakui*) as an example of social customs. *Curr Anthropol* 56:104–120
- Nishida T, Matsusaka T, McGrew WC (2009) Emergence, propagation or disappearance of novel behavioral patterns in the habituated chimpanzees of Mahale: a review. *Primates* 50:23–36. <https://doi.org/10.1007/s10329-008-0109-y>
- O'Malley RC, Wallauer W, Murray CM, Goodall J (2012) The appearance and spread of ant fishing among the Kasekela chimpanzees of Gombe: a possible case of intercommunity cultural transmission. *Curr Anthropol* 53:650–663
- Perry S (2009) Are nonhuman primates likely to exhibit cultural capacities like those of humans? In: Laland KN, Galef BG (eds) *The question of animal culture*. Harvard University Press, Cambridge, pp 247–268
- Perry S (2011) Social traditions and social learning in capuchin monkeys (*Cebus*). *Proc R Soc B Biol Sci* 366:988–996. <https://doi.org/10.1098/rstb.2010.0317>
- Sanz CM, Morgan DB (2009) Flexible and persistent tool-using strategies in honey-gathering by wild chimpanzees. *Int J Primatol* 30:411–427. <https://doi.org/10.1007/s10764-009-9350-5>
- Sarabian C, Macintosh AJJ (2015) Hygienic tendencies correlate with low geohelminth infection in free-ranging macaques. *Bio Lett* 11:20150757. <https://doi.org/10.1098/rsbl.2015.0757>
- Sasaki T, Biro D (2017) Cumulative culture can emerge from collective intelligence in animal groups. *Nat Commun* 8:15049. <https://doi.org/10.1038/ncomms15049>
- Shipton C, Nielsen M (2015) Before cumulative culture. *Hum Nat* 26:331–345
- Somel M, Liu X, Khaitovich P (2013) Human brain evolution: transcripts, metabolites and their regulators. *Nat Rev Neurosci* 14:112–127. <https://doi.org/10.1038/nrn3372>
- Sterelny K (2009) Peacekeeping in the culture wars. In: Laland KN, Galef BG (eds) *The question of animal culture*. Harvard University Press, Cambridge, pp 288–304
- Stewart FA, Pruett JD, Hansell MH (2007) Do chimpanzees build comfortable nests? *Am J Primatol* 69:930–939
- Takahata Y, Hiraiwa-Hasegawa M, Takasaki H, Nyundo R (1986) Newly acquired feeding habits among the chimpanzees of the Mahale Mountains National Park, Tanzania. *Hum Evol* 1:277–284
- Tan AWY, Luncz L, Haslam M et al (2016) Complex processing of prickly pear cactus (*Opuntia* sp.) by free-ranging long-tailed macaques: preliminary analysis for hierarchical organisation. *Primates* 57:141–147. <https://doi.org/10.1007/s10329-016-0525-3>
- Tennie C, Call J, Tomasello M (2009) Ratcheting up the ratchet: on the evolution of cumulative culture. *Philos Trans R Soc Lond B Biol Sci* 364:2405–2415. <https://doi.org/10.1098/rstb.2009.0052>
- Tomasello M (1999) The human adaptation for culture. *Ann Rev Anthropol* 28:509–529
- Tomasello M (2009) The question of chimpanzee culture, plus postscript. In: Laland KN, Galef BG (eds) *The question of animal culture*. Harvard University Press, Cambridge, pp 198–221
- Tomasello M (2016) The ontogeny of cultural learning. *Curr Opin Psychol* 8:1–4
- Tomasello M, Kruger AC, Ratner HH (1993) Cultural learning. *Behav Brain Sci* 16:495–511
- Tylor EB (1871) *Primitive culture*. Murray, London
- Visalberghi E, Frigaszy DM (1990) Food-washing behaviour in tufted capuchin monkeys, *Cebus apella*, and crab-eating macaques, *Macaca fascicularis*. *Anim Behav* 40:829–836
- Wasielowski H (2014) Imitation is necessary for cumulative cultural evolution in an unfamiliar, opaque task. *Hum Nat* 25:161–179. <https://doi.org/10.1007/s12110-014-9192-5>
- Watanabe K (1994) Precultural behavior of Japanese macaques: longitudinal studies of the Koshima troops. In: Gardner RA, Gardner BT, Chiarelli B, Plooij FC (eds) *The ethological roots of culture*. Kluwer, Dordrecht, pp 81–94
- Watanabe K (2001) A review of 50 years of research on the Japanese monkeys of Koshima: status and dominance. In: Matsuzawa T (ed) *Primate origins of human cognition and behaviour*. Springer, Tokyo, pp 405–417
- Whitehead H, Rendell L (2015) *The cultural lives of whales and dolphins*. The University of Chicago Press, Chicago
- Whiten A, van Schaik CP (2007) The evolution of animal “cultures” and social intelligence. *Philos Trans R Soc Lond B Biol Sci* 362:603–620. <https://doi.org/10.1098/rstb.2006.1998>
- Whiten A, Goodall J, McGrew WC, Nishida T, Reynolds V, Sugiyama Y, Tutin CEG, Wrangham RW, Boesch C (2001) Charting cultural variation in wild chimpanzees. *Behaviour* 138:1481–1516
- Yamamoto S, Humle T, Tanaka M (2013) Basis for cumulative cultural evolution in chimpanzees: social learning of a more efficient tool-use technique. *PLoS ONE* 8:e55768. <https://doi.org/10.1371/journal.pone.0055768>
- Zhang P, Watanabe K, Eishi T (2007) Habitual hot-spring bathing by a group of Japanese macaques (*Macaca fuscata*) in their natural habitat. *Am J Primatol* 69:1425–1430
- Zwirner E, Thornton A (2015) Cognitive requirements of cumulative culture: teaching is useful but not essential. *Sci Rep* 5:16781. <http://doi.org/10.1038/srep16781>

B

Count, Crop and Recognise: Fine-Grained Recognition in the Wild

Published as conference paper for ICCV Workshop on Computer Vision for
Wildlife Conservation

Count, Crop and Recognise: Fine-Grained Recognition in the Wild

Max Bain^{1†}, Arsha Nagrani¹, Daniel Schofield² and Andrew Zisserman¹

¹Visual Geometry Group, Department of Engineering Science, University of Oxford

²Institute of Cognitive & Evolutionary Anthropology, University of Oxford

Abstract

The goal of this paper is to label all the animal individuals present in every frame of a video. Unlike previous methods that have principally concentrated on labelling face tracks, we aim to label individuals even when their faces are not visible. We make the following contributions: (i) we introduce a ‘Count, Crop and Recognise’ (CCR) multi-stage recognition process for frame level labelling. The Count and Recognise stages involve specialised CNNs for the task, and we show that this simple staging gives a substantial boost in performance; (ii) we compare the recall using frame based labelling to both face and body track based labelling, and demonstrate the advantage of frame based with CCR for the specified goal; (iii) we introduce a new dataset for chimpanzee recognition in the wild; and (iv) we apply a high-granularity visualisation technique to further understand the learned CNN features for the recognition of chimpanzee individuals.

1. Introduction

Recognising animal individuals in video is a key step towards monitoring the movement, population, and complex social behaviours of endangered species. Traditional individual recognition pipelines rely extremely heavily on the detection and tracking of the face or body, both for humans [6, 11, 18, 27, 34, 42, 56, 61, 64] and for other species [13, 48, 52, 60]. This can be a daunting annotation task, especially for large video corpora of non-human species where custom detectors must be trained and expert knowledge is required to label individuals. Furthermore, often these detectors fail for animal footage in the wild due to the occlusion of individuals, varying lighting conditions and highly deformable bodies.

Our goal in this paper is to automatically label individuals in every frame of a video; but to go beyond face and body recognition, and explore identification using the entire frame. In doing so we analyse the important trade

off between precision and recall for face, body and full-frame methods for recognition of individuals in video. We target the recognition of chimpanzees in the wild. Consider the performance of models at the three levels of face, body and frame (Figure 1). Face recognition now achieves very high accuracy [44, 51, 55] for humans due to the availability of very large datasets for training face detection [31, 49, 63, 65] and recognition [2, 7, 24, 32, 59]. The result is that the *precision* of recognising individuals will be high, but the *recall* may well be low, since, as mentioned above, face recognition will fail for many frames where the face is not visible. Using a body level model occupies a middle ground between face and frame level: it offers the possibility of recognising the individual when the head is occluded, e.g. by distinguishing marks or shapes in the case of animals, or by hair or clothes in the case of humans (albeit it is worth noting that changes in clothing can reduce this advantage – animals obligingly are unclothed). However, body detectors do not as yet have the same performance as face detectors, as animal bodies in particular are highly deformable and can often overlap each other. This means that bodies may be missed in frames, especially if they are small. A frame level model offers the possibility of very high recall (since there are no explicit detectors that can fail, as there are for faces and bodies). In addition, such a method can implicitly use higher-level features for recognition, such as the co-occurrence and spatial relationships between animal individuals (eg. infants are often present in close proximity to the mother). However, the precision may be low because of the challenge of the large proportion of irrelevant information present in the frame (in the case of body and particularly face detection, irrelevant information is removed).

In this paper we show that the performance of frame level models can be considerably improved by automatically zooming in on the regions containing the individuals. This then enables the best of both worlds: cheap supervision at the frame level, obviating the necessity to train and employ face or body detectors, and high recall; but with the precision comparable to face and body detection. We

[†]Correspondence at maxbain@robots.ox.ac.uk

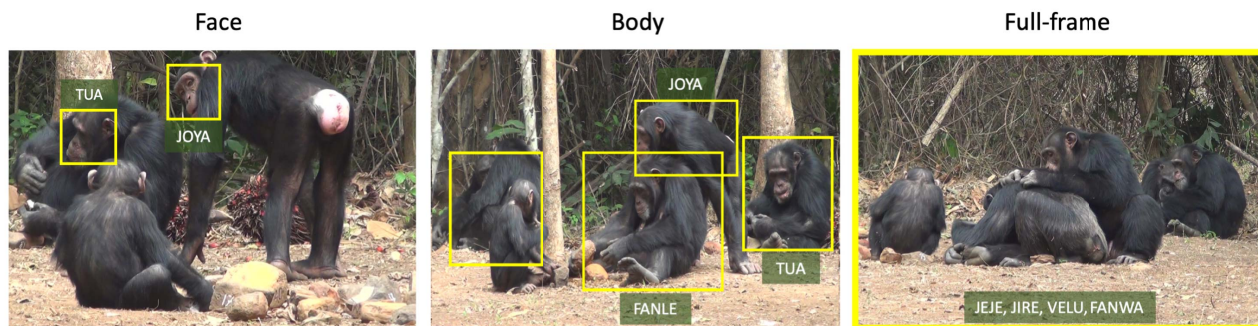


Figure 1. Levels of localisation that can be used to recognise individuals in raw footage. Left to right: (1) **Face**: high precision, but often individuals are not detected. (2) **Full Body**: while recall is higher, bodies can be incredibly difficult to detect due to their extremely deformable nature. (3) **Full Frame**: In this work we explore an architecture to recognise individuals using only frame level supervision.

make the following contributions: (i) we propose a multi-stage *Count, Crop and Recognise* (CCR) pipeline to recognise individuals from raw video with only frame level identity labels required for training. The first two *Count and Crop* stages proposes a rectangular region that tightly encloses all the individuals in the frame. The final *Recognise* stage then identifies the individuals in the frame using a multilabel classifier on the rectangular region *at full resolution* (Figure 2). (ii) We analyse the trade-offs between using our frame level model and other varying levels of localised supervision for fine-grained individual recognition (at a face, body and frame level) and their respective performances. (iii) We have annotated a large, ‘in the wild’ video dataset of chimpanzees with labels for multiple levels of supervision (face tracks, body tracks, frames) which is available at www.robots.ox.ac.uk/~vgg/research/ccr. Finally, (iv) we apply a high-granularity visualisation technique to further understand the learned CNN features for the recognition of chimpanzee individuals.

2. Related Work

Animal recognition in the wild: Video data has become indispensable in the study of wild animal species [8, 43]. However, animals are difficult objects to recognise, mainly due to their deformable bodies and frequent self occlusion [1, 4]. Further, variations in lighting, other individual flora, and motion blur create additional challenges. Taking inspiration from computer-vision based systems for humans, previous methods for species identification have focused on faces, for chimpanzees [13, 21], tigers [35, 37], lemurs [12] and even pigs [25]. Compared to bodies, faces are less deformable and have a fairly standard structure. However, unlike human faces or standard non-deformable object categories, there is a dearth of readily available detectors that can be used off the shelf to localize animals in a frame, requiring researchers to annotate datasets and train their own detectors. It is also often not clear which part of the animal is the most discriminative, e.g. for

elephants ears are commonly used [15], whereas for other mammals unique coat patterns such as stripes for zebras and tigers [37] and spots on Jaguars could be key for recognition [26]. Moving to a full-frame method obviates the need to identify a key discriminating region. Popular wildlife recognition datasets, such as iNaturalist [58], contain species level labels and in contrast to our dataset, typically contain a *single* instance of a class clearly visible in the foreground. While a valuable dataset does exist for the individual recognition of chimpanzees [21, 39], this dataset only contains cropped faces of individuals from zoo enclosures, less applicable to applications of conservation in the wild.

Human recognition in TV and film videos: The original paper in this area by Everingham *et al.* [18] introduced three ideas: (i) associating faces in a shot using tracking by detection, so that a face-track is the ‘unit’ to be labelled; (ii) the use of aligned transcripts with subtitles to provide supervisory information for character labels; and (iii) visual speaker detection to strengthen the supervision (if a person is speaking then their identity is known from the aligned transcript). Many others have adopted and extended these ideas. Cour *et al.* [11] cast the problem as one of ambiguous labelling. Subsequently, Multiple Instance Learning (MIL), was employed by [6, 27, 34, 61, 64]. Further improvements include: unsupervised and partially-supervised metric learning [9, 23]; the range of face viewpoints used (e.g. adding profile face tracks in addition to the original near-frontal face tracks) [19, 53]; and obtaining an episode wide consistent labelling [56] (by using a graph formulation and other visual cues). Recent work [42] has explored using *only* face and voice recognition, without the use of weak supervision from subtitles.

Frame level supervision: The task of labelling image regions given only frame level labels is that of weakly supervised segmentation: every image is known to have

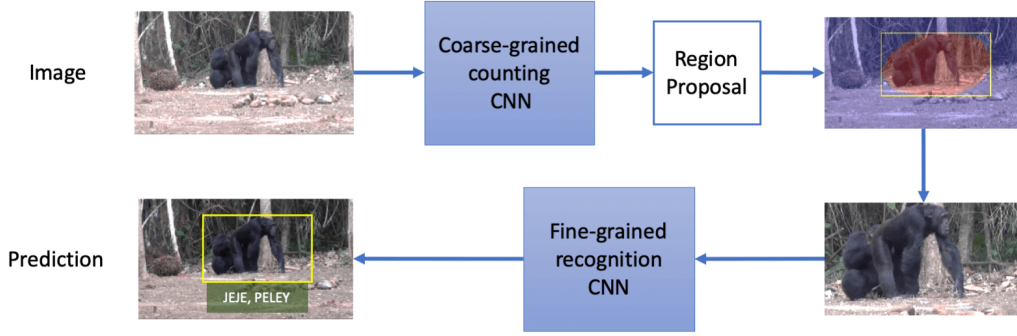


Figure 2. The *Count, Crop and Recognise* pipeline consists of three stages: (1) a coarse-grained counting network to count the number of individuals per frame, (2) a crop stage where the class activation maps from the counting network are used to localise regions of interest in the image, and (3) a fine-grained classifier trained on these cropped images.

(or not) – through the image (class) labels – one or several pixels matching the label. However, the positions of these pixels are unknown, and have to be inferred. Early deep learning works on this area include [33, 46, 47]. Our problem differs in that it is fine-grained – all the object classes are chimpanzees that must be distinguished, say, rather than the 20 PASCAL VOC classes of [33, 46, 47]. While there have been works on localising fine-grained objects with weak supervision [5, 22, 29], they deal only with the restricted case of one instance per image (i.e. an image containing a single bird of class *Horned Puffin*). As far as we know, we are the first to tackle the challenging task of classifying multiple fine-grained instances in a single frame with weak supervision.

3. Count, Crop and Recognise (CCR)

Our goal is, given a frame of a video, to predict all the individuals present in that frame. We would like to learn to do this task with only *frame-level* labels, i.e. no detections and hence no correspondences (who’s who). The major challenge with such a method, however is that frames contain a lot of irrelevant background noise (Figure 3), and the distinctions between different individuals is often very fine-grained and subtle (these fine details are hard to learn due to the limited input resolution of CNNs).

Hence we propose a multi-stage, frame level pipeline that automatically crops discriminative regions containing individuals and so eliminates as much background information as possible, while maintaining the high resolution of the original image. This is achieved by training a deep CNN with a coarse-grained counting objective (a much easier task than fine-grained recognition), before performing identity recognition. The method is loosely inspired by the weakly-supervised object detection method C-WSL [22], however, unlike this work, our method requires neither explicit object proposals nor an existing weakly supervised detection method. Since we do not require exact bounding boxes *per instance*, but simply a generic zoomed in region, we use

class guided activation maps to determine the region of focus. The multiple stages of our CCR method are described in more detail below. Precise implementation details can be found in Section 6.2, and a diagrammatic representation of the pipeline can be seen in Figure. 2.

Let $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$ be a single frame of the video and let $Y \in \{0, 1\}^k$ be a finite vector denoting which of the total k individuals are visible. $Y[i] = 1$ if the i -th individual is visible in $\mathbf{x}$, and $Y[i] = 0$ otherwise.

Count: We first train a parameterised function $c_\theta(\mathbf{x}')$, given a resized image input $\mathbf{x}' \in \mathbb{R}^{C \times H' \times W'}$ to count the number of individuals n within a frame. In general, we can cast this problem as either a multiclass problem or a regression problem. Since the number of individuals per frame in our datasets is small, we pose this counting task as one of multiclass classification, where the total number of individuals present can be categorised into one of the following classes $n \in \{0, 1, \dots, N\}$ where all counts of N or more are binned into a single bin, with N selected as a hyperparameter (in this work we use $N = 3$). The ‘Negatives’ class ($n = 0$) is very important for training. Labels for counting come for free with frame level annotation (total number of labels per frame, or $n = |Y|$). The loss to be minimised can then be framed as a cross-entropy loss on the target n values. In this work we instantiate $c(\mathbf{x}')$ as a deep convolutional neural network (CNN) with convolutional layers followed by fully connected layers. Generally $H', W' < H, W$ due to the discrepancy in resolution of raw images and pretrained CNNs.

Crop: Class Activation Maps (CAMs) [67] are generated from the counting model $c_\theta(\mathbf{x}')$ to localise the discriminative regions. For resized input image $\mathbf{x}'$, let $f_k(i, j)$ denote the activation of a unit k in the last convolutional layer, and w_k^n denote the weight corresponding to count n for unit k . The CAM, M_n , at each spatial location is given by:

$$M_n(i, j) = \sum_k w_k^n f_k(i, j) \quad (1)$$

describing the importance of visual patterns at different spatial locations for a given class, in this case a count. By upsampling the CAM to the same size of $\mathbf{x}$ (H, W) image regions most relevant to the particular category can be identified. The CAM is then normalised and segmented:

$$M_n^{norm}(i, j) = \frac{M_n(i, j) - \min_{i,j} M_n(i, j)}{\max_{i,j} M_n(i, j)} \quad (2)$$

$$M_n^{thresh}(i, j) = \begin{cases} 1, & \text{if } M_n^{norm}(i, j) > T \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $T \in [0, 1]$ is the chosen threshold value. The largest connected component in M_n^{thresh} is found using classical component labelling algorithms [20, 62], examples shown in Figure 3. The bounding box enclosing this component is used to crop the original input image $\mathbf{x}$ to get $\mathbf{x}_{crop}$, removing irrelevant portions of the image and permitting higher resolution of the cropping region.

Recognise: The cropped regions $\mathbf{x}_{crop}$ are used to train a fine grained recognition classifier $R_\phi(\mathbf{x}'_{crop})$ using the original frame-level labels Y . This second recognition classifier is also instantiated as a CNN, with different parameters ϕ , and trained for the task of multilabel classification, with one class for every individual in the dataset. We use a weighted Binary Cross-Entropy loss, where the weight w_i for each class i is: $w_i = f_{max}/f_i$, where f_{max} refers to the number of instances for the most populous class, and f_i is the number of instances for class i .

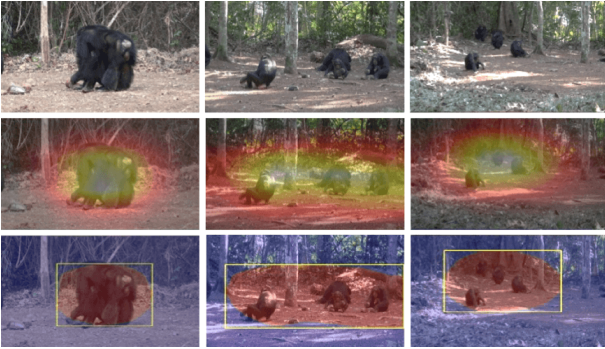


Figure 3. Region proposals for the Chimpanzee Bossou dataset. These are learnt via our counting CNN with no detection supervision at all. Top row: original frame; middle row: CAM for the count; bottom row: region proposal. Note in the second column, how the localisation works well even when the individuals are far apart from each other.

Why use counting to localise? Our method begs the following question: if a model must identify discriminative regions to be able to count individuals, surely it must also identify these regions to perform fine-grained recognition?

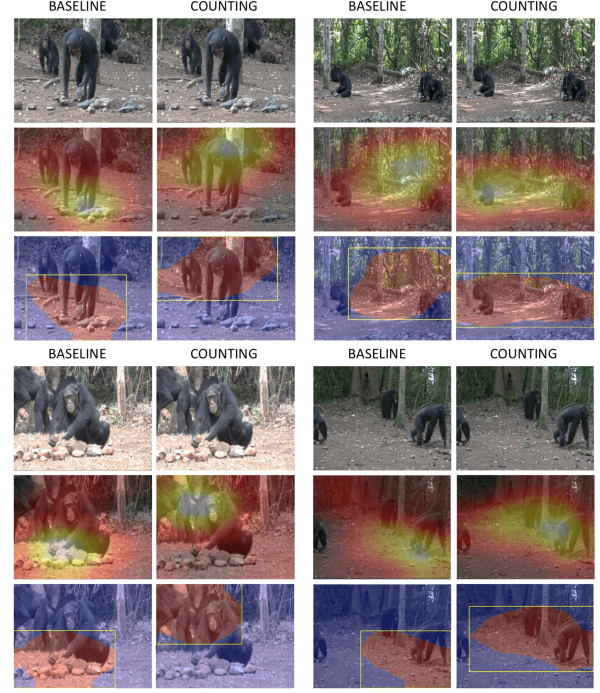


Figure 4. Region proposals for both the baseline (recognition) and counting method. Note how the baseline method mistakenly focuses on the background features, rocks and trees, to recognise individuals.

In this case we could just train the fine grained recognition network to obtain region proposals, crop regions and then retrain the recognition network in an iterative manner. However, counting objects is a much easier task than the fine-grained recognition of identities (a widely studied phenomenon in psychology, called subitizing [10] suggests that humans are able to count objects with a single glance if the total number of objects is small). We find that this leads to much better region proposals, as demonstrated in Figure 4 where we show proposals obtained from a counting model and from an identification model. By tackling an easier task first, our model is using a form of curriculum learning [3].

4. Face and Body Tracking and Recognition

In order to test recognition methods that explicitly use only face and body regions, we first create a chimpanzee face and body detection dataset, by annotating bounding boxes using the VIA annotation tool [16]. We then train a detector with these detection labels, and run the detector on every frame of the video. A tracker is then run to link up the detections to form *face-tracks* or *body-tracks*, which then become a single unit for both labelling and recognition. Examples are shown in Figure 5. Finally, we train a standard CNN multi-class classifier on the regions in the track using a cross-entropy loss on the identities in the dataset to train a recognition model.

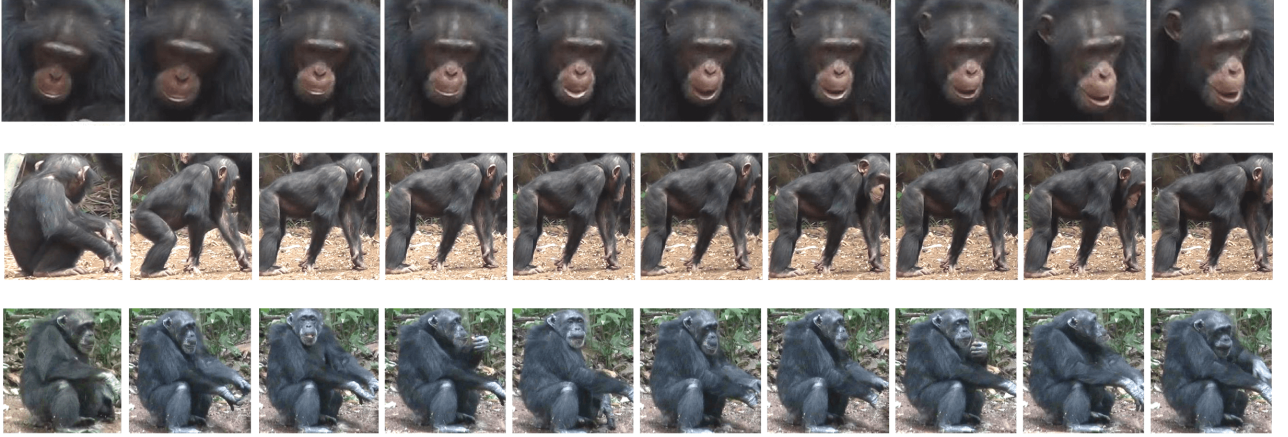


Figure 5. Chimpanzee tracks for face (top row) and body (bottom two rows).

5. Datasets

Chimpanzee Bossou Dataset: We use a large, un-edited video corpus of chimpanzee footage collected in the Bossou forest, southeastern Guinea, West Africa. Bossou is a chimpanzee field site established by Kyoto University in 1976 [30, 41, 50, 54]. Data collection at Bossou was done using multiple cameras to document chimpanzee behaviour at a natural forest clearing (7m x 20m) located in the core of the Bossou chimpanzees’ home range. The videos were recorded at different times of the day, and span a range of lighting conditions. Often there is heavy occlusion of individuals due to trees and other foliage. The individuals move around and interact freely with one another and hence faces in video have large variations in scale, motion blur and occlusion due to other individuals. Often faces appear as extreme profiles (in some cases only a single ear is visible). While the original Bossou dataset is a massive archive with over 50 hours of data from multiple years, in this paper we use roughly 10 hours of video footage from the years 2012 and 2013, of which we reserve 2 hours for testing. Chimpanzees are visible for the vast majority of this footage, therefore we also include sampled frames of just the forest background ($n = 0$) from other years to permit negative training for all methods.

Dataset annotation and statistics: We manually provide frame level annotations (i.e. name tags for the individuals present) for every frame in the videos using the VIA video annotation tool [17]. VIA is an open source project based solely on HTML, Javascript and CSS (no dependency on external libraries) and runs entirely in a web browser*. In addition, we compute face and body detections and tracks (as described in Section 4) and also label these tracks manually using the VIA tool. All identity labelling

*<http://www.robots.ox.ac.uk/~vgg/software/via/>

	hours	#frames	# individuals
train	8.30	830k	13
test	1.56	161k	10
total	9.86	992k	13

Table 1. Dataset Statistics for the Chimpanzee Bossou dataset. We annotate facetracks, bodytracks and identities at a frame level.

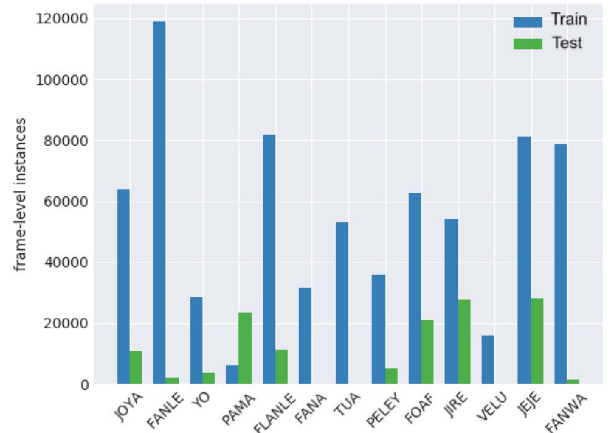


Figure 6. Instance frequency histograms for each individual in the Chimpanzee Bossou dataset.

was done by an expert anthropologist familiar with the identities in the archive. The statistics of the dataset are given in Table 1. The frame-level frequency histogram for each individual is shown in Figure 6, where an instance of an individual is defined as a frame for which the individual is visible.

6. Experiments

We first evaluate the performance of the face-track and body-track methods, in particular the proportion of frames that they can label (the frame recall), and their identity

recognition performance. This is then compared to the performance of the frame-level CCR method using average precision (AP) to analyse the trade-offs thoroughly. We also compare the CCR method to a simple baseline, where an identity recognition CNN is trained directly on the resized raw (not zoomed in) images $\mathbf{x}'$.

6.1. Evaluation Metrics

Detector Recall: The detector recall is the proportion of instances where faces (or bodies) are detected and tracked. This provides an upper bound on the number of individual instances that can be recognised from the video dataset using the face-track or body-track methods. We note that this is a function of two effects: (1) the visibility of the face or body in the image (faces could be turned away, be occluded etc); and (2) the performance of the detection and tracking method (i.e. is the face detected even if it is visible); though we do not distinguish these two effects here.

Identification Accuracy: This is the proportion of detections that are labelled correctly (each face-track or body-track can only be one of the possible identities).

System-level Average Precision (AP): For the face (and body) track methods, the precision and recall for each individual is computed as follows: all tracks are ranked by the score of the individual face classifier; if the track belongs to that individual, then all the frames that contain that track are counted as recalled; if the track does not belong to that individual, then the frames that contain that track are not recalled (but the precision takes these negative tracks into account), i.e. we only recall the frames containing a track if we correctly identify the individual in that track. For the frame level CCR method, the frames are ranked by the frame-level identity classifier, and the precision and recall computed for this ranked list. We then calculate both the micro and macro Average Precision score over all the individuals. Macro Average Precision (mAP) takes the mean of the AP values for every class, whereas Micro Average Precision (miAP) aggregates the contributions of all classes to compute its average metric. For our heavily class unbalanced datasets, the latter is a much better indicator of the overall performance. (histograms are provided in the supplementary material).

6.2. Implementation Details

CNN architecture and training: For a fair comparison, we use the following hyperparameters across *all* recognition models: a ResNet-18 [28] architecture pretrained on ImageNet [14] with input size $H', W' = 224$ i.e. for the counting CNN $c_\phi(\mathbf{x}')$, the fine-grained identity CNN $R_\phi(\mathbf{x}'_{crop})$, and the recognition CNNs used for both the body and the face models. This architecture achieves

a good trade-off between performance and number of parameters. In principle any deep CNN architecture could be used with our method. The models are trained and tested on every third frame from the video (to avoid the large amount of redundancy in consecutive frames). We use a batch size of 64; standard data augmentation (colour jittering, horizontal flipping etc.) but only random cropping on the negative ($n = 0$) samples. All models are trained end-to-end in PyTorch [45]. Models and code will be released.

Face and Body tracks: The face and body tracks were obtained by training a Single Shot MultiBox Detector (SSD) [38], on 8k and 16k bounding box annotations respectively. The annotations were gathered on frames sampled every 10 seconds from a subset of training footage as well as from videos from other years. The detectors are trained in PyTorch with 300×300 input resolution and the same data augmentation techniques as [38]. We use a batch size of 64 and train the detectors for 95k iterations with a learning rate of $1E-4$. We used the KLT [40, 57] and SPN [36] tracker to obtain face and body tracks respectively. During the recognition stage, predictions are averaged across a track.

Count, Crop and Recognise: The coarse-grained counting CNN is applied on the entire dataset and the CAM of the highest softmax prediction for each image recorded. The CAMs, just 7×7 int arrays, are saved cheaply as grey-scale images each of size 355 bytes. Alternatively, this can be performed online during training, albeit at a greater computational cost since the CAMs are recomputed every epoch. Before training the recognition stage, we upsample the CAMs to the size of its corresponding image and threshold with $T = 0.5$, perform full-resolution cropping and then resize back to 224×224 , the input size of the fine-grained identity CNN $R_\phi(\mathbf{x}_{crop})$. Fine-grained recognition is then performed on these cropped regions.

6.3. Results

Detector recall and identification accuracy: The performance is given in Table 2. It is clear that recall is a large limitation for both the face-track and body-track methods. The face detector recall is low (less than 40%), far lower than that of the body detector. This reflects the fact that the chimpanzee's faces are not visible in many frames, rather than failures of the face detector. Hence even a perfect face recognition system would miss many chimpanzee instances at the frame level. While the identification accuracy for chimpanzees, is slightly higher for faces than for bodies, the relatively high recall of the body-track method shows a clear advantage over faces.

	#instances	#tracks	recall (%)	test acc. (%)
face	1.02m	5k	39.9	71.3
body	1.64m	12k	64.0	70.5
frame	2.13m	-	100.0	-

Table 2. Face and body detector recall and identification test accuracy (acc.) results for the Bossou dataset. Recall is calculated as a percentage of the total number of instances annotated at a frame level, which we note as a theoretical upper bound of 100%.

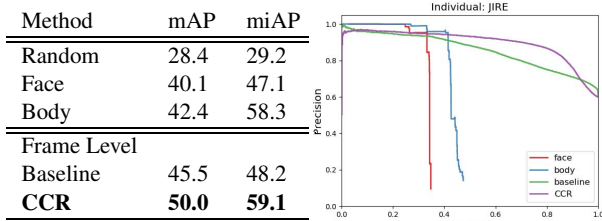


Figure 7. Left: Comparison of system level AP for all methods on the *test* set; Right: PR curves for a single individual from the Chimpanzee dataset.

System level AP: Results are given in Figure 7, left. We compare our CCR method to a simple baseline without the Count and Crop stages. CCR outperforms the baseline by a large-margin (more than 9% AP). The PR curve for the chimpanzee ‘JIRE’ (Figure 7, right), reiterates the results that face-track recall is the lowest, albeit with the highest precision. In contrast, the CCR method has far higher recall and with a similar level of precision. The overall AP values (Figure 7, left) show that the body-track AP is quite high, since it achieves a large boost in recall over the face-tracks with a very small drop in identification accuracy (less than 1%). We note that the CCR method, however, outperforms the body-track method as well. This is an impressive performance considering CCR requires only frame level supervision in training, and eschews the need to train a body detector.

7. Weakly Supervised Localisation of Individuals

Labelling individuals within a frame offers insight into social relationships by monitoring the frequency of co-occurrences and locations of the capturing cameras. However, unlike face and body detection, the frame level approach does not explicitly localise individuals within the frame, preventing analysis of the local proximity between individuals. To tackle this, we propose an extension to CCR which localises individuals without any extra supervisory data. This is shown in the examples of Figure 8.

Following a similar process to the ‘Crop’ stage in CCR, bounding boxes are generated for each labelled individual from CAMs extracted from the recognition model $R_\phi(\mathbf{x}'_{crop})$. The locations of the individuals are assumed

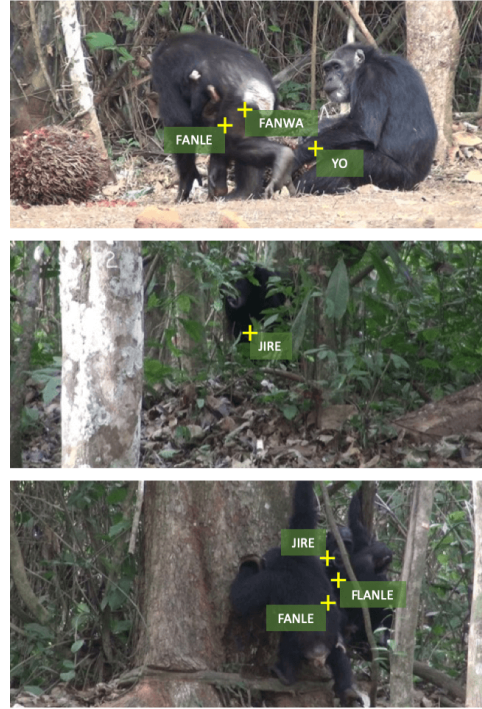


Figure 8. Weakly supervised localisation of individuals.

to be at the centroid of these bounding boxes, with qualitatively impressive results even when the individuals are grouped together.

8. Interpretability

In this penultimate section, we introduce a high-granularity visualisation tool to understand and interpret the predictions made by the face and body recognition models. These tease out the discriminative features learnt by the model for this task of fine-grained recognition of individuals. Understanding these features can provide new insights to human researchers.

A Class Activation Map (CAM) [67], introduced in Section 3, can be used to localise discriminative features but it does so at low resolution and thus cannot identify high-frequency features, such as edges and dots. An alternative visualisation method is Excitation Backprop (EBP) [66]. EBP achieves high-granularity visualisation via a top-down attention model, working its way down from the last layer of the CNN to the high resolution input layer. Activations are followed from layer to layer with a probabilistic Winner-Take-All process.

In Figure 9, we show the EBP visualisations from the face recognition model of example images of individuals in the Bossou dataset. When the ears are visible, the face model shows high activation on the ear region – similarly for the brow and mouth regions. Upon closer inspection of the original face images, the ears of each individual are in-

deed highly unique and distinguishable. The expert anthropologist, who manually labelled the dataset, noted that he doesn't pay particular attention to the ears when identifying the individuals. Perhaps our discovery of ear uniqueness in chimpanzees in this dataset, and possibly all chimpanzees, could improve expert's recognition of chimpanzee individuals.

The EBP visualisation for the body recognition model in Figure 9 reiterates the importance of the face and ears in distinguishing the individuals. Further, note Jejes hairless patch on his left leg in the top of Figure 9g and corresponding EBP activation, indicating that the body recognition model also uses distinguishing marks on the body. Similarly, Foaf's white spot above his upper lip (Figure 9e) is another region of high activation. The presence of the white spot was unbeknownst to the anthropologist who noted he would now use this information to identify Foaf in the future. These two examples show that a CNNs learned discriminative features for a specific individual can be visualised and interpreted by humans. Of course, these findings are not statistically relevant and quantitative analysis would be needed in order to determine the effectiveness of the use of recognition CNNs to train human experts.

9. Conclusion

We have proposed and implemented a simple pipeline for fine-grained recognition of individuals using only frame-level supervision. This has shown that a counting objective allows us to learn very good region proposals, and zooming into these discriminative regions gives substantial gains in recognition performance. Many datasets 'in the wild' have the property that resolution of individuals can vary greatly with scene depth, and with cameras panning and zooming in and out. Our frame-level method approaches the precision of face-track and body-track recognition methods, whilst now allowing a much higher recall. We hope that our newly created dataset will spur further work in high-recall frame-level methods for fine-grained individual recognition in video, and that our preliminary work on interpretability of CNNs for classifying individuals of species gives insight on identifying discriminative features.

Acknowledgments: This project has benefited enormously from discussions with Dora Biro and Susana Carvalho at Oxford. We are grateful to Kyoto University's Primate Research Institute for leading the Bossou Archive Project, and supporting the research presented here, and to IREB and DNRST of the Republic of Guinea. This work is supported by the EPSRC programme grant Seebibyte EP/M013774/1. A.N. is funded by a Google PhD fellowship; D.S. is funded by the Clarendon Fund, Boise Trust; Fund and Wolfson College, University of Oxford.

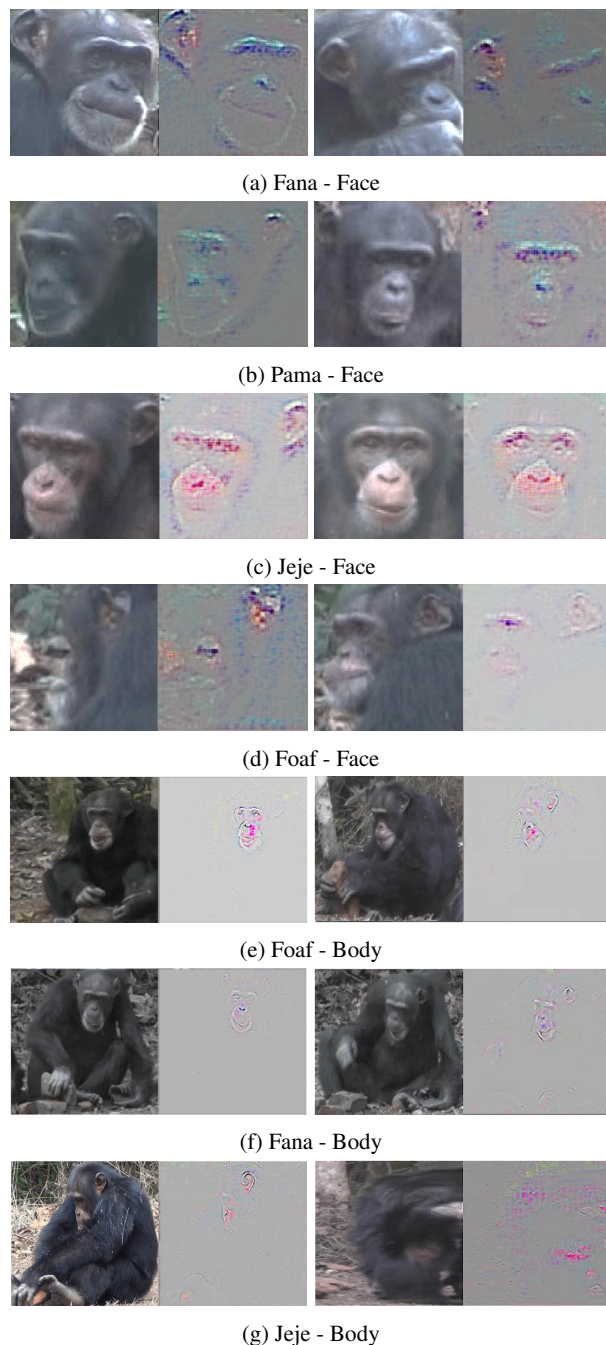


Figure 9. Excitation Backprop [66] visualisations (right) from the face and body recognition model for example images of individuals in the Chimpanzee Bossou dataset.

References

- [1] H. M. Afkham, A. T. Targhi, J.-O. Eklundh, and A. Pronobis. Joint visual vocabulary for animal classification. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008. 2
- [2] A. Bansal, A. Nanduri, C. Castillo, R. Ranjan, and R. Chellappa. Umdfaces: An annotated face dataset for training deep networks. *arXiv preprint arXiv:1611.01484*, 2016. 1
- [3] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009. 4
- [4] T. L. Berg and D. A. Forsyth. Animals on the web. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, volume 2, pages 1463–1470. IEEE, 2006. 2
- [5] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016. 3
- [6] P. Bojanowski, F. Bach, , I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Finding actors and actions in movies. In *Proc. ICCV*, 2013. 1, 2
- [7] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *Proc. Int. Conf. Autom. Face and Gesture Recog.*, 2018. 1
- [8] A. Caravaggi, P. B. Banks, A. C. Burton, C. M. Finlay, P. M. Haswell, M. W. Hayward, M. J. Rowcliffe, and M. D. Wood. A review of camera trapping for conservation behaviour research. *Remote Sensing in Ecology and Conservation*, 3(3):109–122, 2017. 2
- [9] R. G. Cinbis, J. J. Verbeek, and C. Schmid. Unsupervised metric learning for face identification in TV video. In *Proc. ICCV*, pages 1559–1566, 2011. 2
- [10] D. H. Clements. Subitizing: What is it? why teach it? *Teaching children mathematics*, 5:400–405, 1999. 4
- [11] T. Cour, B. Sapp, and B. Taskar. Learning from ambiguously labeled images. In *Proc. CVPR*, 2009. 1, 2
- [12] D. Crouse, R. L. Jacobs, Z. Richardson, S. Klum, A. Jain, A. L. Baden, and S. R. Tecot. Lemurfaceid: a face recognition system to facilitate individual identification of lemurs. *Bmc Zoology*, 2(1):2, 2017. 2
- [13] D. Deb, S. Wiper, A. Russo, S. Gong, Y. Shi, C. Tyomozek, and A. Jain. Face recognition: Primates in the wild. *arXiv preprint arXiv:1804.08790*, 2018. 1, 2
- [14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 6
- [15] I. Douglas-Hamilton. *On the ecology and behaviour of the African elephant*. PhD thesis, University of Oxford, 1972. 2
- [16] A. Dutta, A. Gupta, and A. Zissermann. VGG image annotator (VIA). <http://www.robots.ox.ac.uk/~vgg/software/via/>, 2016. 4
- [17] A. Dutta and A. Zisserman. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM ’19, New York, NY, USA, 2019. ACM. 5
- [18] M. Everingham, J. Sivic, and A. Zisserman. “Hello! My name is... Buffy” – automatic naming of characters in TV video. In *Proc. BMVC*, 2006. 1, 2
- [19] M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automatic naming of characters in TV video. *Image and Vision Computing*, 27(5), 2009. 2
- [20] C. Fiorio and J. Gustedt. Two linear time union-find strategies for image processing. *Theoretical Computer Science*, 154(2):165 – 181, 1996. 4
- [21] A. Freytag, E. Rodner, M. Simon, A. Loos, H. S. Kühl, and J. Denzler. Chimpanzee faces in the wild: Log-euclidean cnns for predicting identities and attributes of primates. In *German Conference on Pattern Recognition*, pages 51–63. Springer, 2016. 2
- [22] M. Gao, A. Li, R. Yu, V. I. Morariu, and L. S. Davis. C-wsl: Count-guided weakly supervised localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 152–168, 2018. 3
- [23] M. Guillaumin, J. Verbeek, and C. Schmid. Multiple instance metric learning from automatically labeled bags of faces. In *Proc. ECCV*, pages 634–647, 2010. 2
- [24] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: Challenge of recognizing one million celebrities in the real world. *Electronic Imaging*, 2016(11):1–6, 2016. 1
- [25] M. F. Hansen, M. L. Smith, L. N. Smith, M. G. Salter, E. M. Baxter, M. Farish, and B. Grieve. Towards on-farm pig face recognition using convolutional neural networks. *Computers in Industry*, 98:145–152, 2018. 2
- [26] B. J. Harmsen, R. J. Foster, E. Sanchez, C. E. Gutierrez-González, S. C. Silver, L. E. Ostro, M. J. Kelly, E. Kay, and H. Quigley. Long term monitoring of jaguars in the cockscomb basin wildlife sanctuary, belize; implications for camera trap studies of carnivores. *PloS one*, 12(6):e0179505, 2017. 2
- [27] M. Haurilet, M. Tapaswi, Z. Al-Halah, and R. Stiefelhagen. Naming tv characters by watching and analyzing dialogs. In *WACV*, 2016. 1, 2
- [28] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6
- [29] T. Hu and H. Qi. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. *arXiv preprint*

- arXiv:1901.09891*, 2019. 3
- [30] T. Humle. Location and ecology. In *The chimpanzees of Bossou and Nimba*, pages 13–21. Springer, 2011. 5
- [31] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, UMass Amherst Technical Report, 2010. 1
- [32] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016. 1
- [33] A. Kolesnikov and C. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Proc. ECCV*, 2016. 3
- [34] M. Köstinger, P. Wohlhart, P. Roth, and H. Bischof. Learning to recognize faces from videos and weakly related information cues. In *avss*, 2011. 1, 2
- [35] H. S. Kühl and T. Burghardt. Animal biometrics: quantifying and detecting phenotypic appearance. *Trends in ecology & evolution*, 28(7):432–441, 2013. 2
- [36] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu. High performance visual tracking with siamese region proposal network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 6
- [37] S. Li, J. Li, W. Lin, and H. Tang. Amur tiger re-identification in the wild. *CoRR*, abs/1906.05586, 2019. 2
- [38] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. SSD: single shot multi-box detector. *CoRR*, abs/1512.02325, 2015. 6
- [39] A. Loos and A. Ernst. Detection and identification of chimpanzee faces in the wild. In *2012 IEEE International Symposium on Multimedia*, pages 116–119. IEEE, 2012. 2
- [40] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981. 6
- [41] T. Matsuzawa. Field experiments of tool-use. In *The Chimpanzees of Bossou and Nimba*, pages 157–164. Springer, 2011. 5
- [42] A. Nagrani and A. Zisserman. From benedict cumberbatch to sherlock holmes: Character identification in tv series without a script. In *Proc. BMVC*, 2017. 1, 2
- [43] T. Nishida, K. Zamma, T. Matsusaka, A. Inaba, and W. C. McGrew. *Chimpanzee behavior in the wild: an audio-visual encyclopedia*. Springer Science & Business Media, 2010. 2
- [44] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proc. BMVC*, 2015. 1
- [45] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 6
- [46] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 3
- [47] P. H. O. Pinheiro and R. Collobert. Weakly supervised semantic segmentation with convolutional networks. In *Proc. CVPR*, 2015. 3
- [48] H. Rakotonirina, P. M. Kappeler, and C. Fichtel. The role of facial pattern variation for species recognition in red-fronted lemurs (*eulemur rufifrons*). *BMC evolutionary biology*, 18(1):19, 2018. 1
- [49] D. Ramanan and X. Zhu. Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2886. Citeseer, 2012. 1
- [50] D. Schofield, A. Nagrani, M. Hayashi, T. Matsuzawa, D. Biro, and S. Carvalho. Chimpanzee face recognition from videos in the wild using deep learning. *Science Advances*, 5, 2019. 5
- [51] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 1
- [52] S. Sinha, M. Agarwal, M. Vatsa, R. Singh, and S. Anand. Exploring bias in primate face detection and recognition. In L. Leal-Taixé and S. Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 541–555, Cham, 2019. Springer International Publishing. 1
- [53] J. Sivic, M. Everingham, and A. Zisserman. “Who are you?” – learning person specific classifiers from video. In *Proc. CVPR*, 2009. 2
- [54] Y. Sugiyama. Population dynamics of wild chimpanzees at bossou, guinea, between 1976 and 1983. *Primates*, 25(4):391–400, 1984. 5
- [55] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015. 1
- [56] M. Tapaswi, M. Baeuml, and R. Stiefelhausen. “knock! knock! who is it?” probabilistic person identification in tv series. In *Proc. CVPR*, 2012. 1, 2
- [57] C. Tomasi and T. K. Detection. Tracking of point features. Technical report, Tech. Rep. CMU-CS-91-132, Carnegie Mellon University, 1991. 6
- [58] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 2
- [59] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, et al. Iarpa janus benchmark-b face

- dataset. In *CVPR Workshop on Biometrics*, 2017. 1
- [60] C. L. Witham. Automated face recognition of rhesus macaques. *Journal of neuroscience methods*, 300:157–165, 2018. 1
- [61] P. Wohlhart, M. Köstinger, P. M. Roth, and H. Bischof. Multiple instance boosting for face recognition in videos. In *DAGM-Symposium*, 2011. 1, 2
- [62] K. Wu, E. Otoo, and A. Shoshani. Optimizing connected component labeling algorithms. volume 5747, 04 2005. 4
- [63] J. Yan, X. Zhang, Z. Lei, and S. Z. Li. Face detection by structural models. *Image and Vision Computing*, 32(10):790–799, 2014. 1
- [64] J. Yang, R. Yan, and A. G. Hauptmann. Multiple instance learning for labeling faces in broadcasting news video. In *ACM Multimedia*, 2005. 1, 2
- [65] S. Yang, P. Luo, C.-C. Loy, and X. Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016. 1
- [66] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation backprop. *CoRR*, abs/1608.00507, 2016. 7, 8
- [67] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 3, 7

Works Cited

- Abdelrahman, Ahmed A. et al. (Mar. 7, 2022). “L2CS-Net: Fine-Grained Gaze Estimation in Unconstrained Environments”. In: *arXiv:2203.03339 [cs]*. version: 1. arXiv: 2203.03339. URL: <http://arxiv.org/abs/2203.03339> (visited on 03/31/2022).
- Alberts, Susan C. (2019). “Social influences on survival and reproduction: Insights from a long-term study of wild baboons”. In: *Journal of Animal Ecology* 88.1. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1365-2656.12887>, pp. 47–66. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/1365-2656.12887> (visited on 05/12/2022).
- Albery, Gregory F., Tim H. Clutton-Brock, et al. (Jan. 4, 2022). *Ageing red deer alter their spatial behaviour and become less social*. Section: New Results Type: article. bioRxiv, p. 2021.06.11.448092. URL: <https://www.biorxiv.org/content/10.1101/2021.06.11.448092v2> (visited on 05/12/2022).
- Albery, Gregory F., Lucinda Kirkpatrick, et al. (2021). “Unifying spatial and social network analysis in disease ecology”. In: *Journal of Animal Ecology* 90.1. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1365-2656.13356>, pp. 45–61. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/1365-2656.13356> (visited on 03/31/2022).
- Allen, Jenny A. (2019). “Community through Culture: From Insects to Whales”. In: *BioEssays* 41.11. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.201900060>, p. 1900060. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.201900060> (visited on 05/08/2022).
- Almeida-Warren, Katarina, Tetsuro Matsuzawa, and Susana Carvalho (Aug. 25, 2021). *Landscaping the behavioural ecology of primate stone tool use*. Section: New Results Type: article. bioRxiv, p. 2021.08.24.457369. URL: <https://www.biorxiv.org/content/10.1101/2021.08.24.457369v1> (visited on 05/16/2022).
- Alvergne, Alexandra, Markus Jokela, and Virpi Lummaa (June 29, 2010). “Personality and reproductive success in a high-fertility human population”. In: *Proceedings of the National Academy of Sciences* 107.26, pp. 11745–11750. URL: <https://pnas.org/doi/full/10.1073/pnas.1001752107> (visited on 05/15/2022).
- Anderson, David J. and Pietro Perona (Oct. 1, 2014). “Toward a Science of Computational Ethology”. In: *Neuron* 84.1, pp. 18–31. URL: <http://www.sciencedirect.com/science/article/pii/S0896627314007934> (visited on 02/12/2019).
- Bateson, Melissa and Paul Martin (2021). *Measuring behaviour: an introductory guide*. Cambridge University Press.

- Bateson, Patrick and Kevin N. Laland (Dec. 2013). “Tinbergen’s four questions: an appreciation and an update”. In: *Trends in Ecology & Evolution* 28.12, pp. 712–718. URL: <https://linkinghub.elsevier.com/retrieve/pii/S016953471300236X> (visited on 04/11/2022).
- Bengio, Yoshua, Aaron Courville, and Pascal Vincent (Aug. 2013). “Representation Learning: A Review and New Perspectives”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1798–1828.
- Biro, Dora, Noriko Inoue-Nakamura, et al. (Dec. 2003). “Cultural innovation and transmission of tool use in wild chimpanzees: evidence from field experiments”. en. In: *Anim. Cogn.* 6.4, pp. 213–223.
- Biro, Dora, Takao Sasaki, and Steven J. Portugal (July 2016). “Bringing a Time–Depth Perspective to Collective Animal Behaviour”. In: *Trends in Ecology & Evolution* 31.7, pp. 550–562. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0169534716300027> (visited on 07/14/2018).
- Boesch, Christophe et al. (Sept. 2020). “Chimpanzee ethnography reveals unexpected cultural diversity”. In: *Nature Human Behaviour* 4.9. Number: 9 Publisher: Nature Publishing Group, pp. 910–916. URL: <https://www.nature.com/articles/s41562-020-0890-1> (visited on 03/24/2022).
- Brakes, Philippa et al. (Apr. 28, 2021). “A deepening understanding of animal culture suggests lessons for conservation”. In: *Proceedings of the Royal Society B: Biological Sciences* 288.1949, p. 20202718. URL: <https://royalsocietypublishing.org/doi/full/10.1098/rspb.2020.2718> (visited on 03/23/2022).
- Brown, André E. X. and Benjamin de Bivort (July 2018). “Ethology as a physical science”. In: *Nature Physics* 14.7. Number: 7 Publisher: Nature Publishing Group, pp. 653–657. URL: <https://www.nature.com/articles/s41567-018-0093-0> (visited on 04/08/2022).
- Burkhardt, Richard W (1999). “Ethology, natural history, the life sciences, and the problem of place”. In: *Journal of the History of Biology* 32.3. Publisher: Springer, pp. 489–508.
- Byrne, Richard W and Lucy A Bates (2006). “Why are animals cognitive?” In: *Current Biology* 16.12. Publisher: Elsevier, R445–R448.
- Canteloup, Charlotte et al. (May 18, 2021). “Processing of novel food reveals payoff and rank-biased social learning in a wild primate”. In: *Scientific Reports* 11.1, p. 9550. URL: <https://www.nature.com/articles/s41598-021-88857-6> (visited on 03/18/2022).
- Cao, Qiong et al. (May 2018). “VGGFace2: A Dataset for Recognising Faces across Pose and Age”. In: *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*. 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), pp. 67–74.
- Carpenter, Clarence Ray (1935). “Behavior of red spider monkeys in Panama”. In: *Journal of Mammalogy* 16.3, pp. 171–180.
- Carreira, Joao and Andrew Zisserman (July 2017). “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017 IEEE Conference on Computer Vision

- and Pattern Recognition (CVPR). Honolulu, HI: IEEE, pp. 4724–4733. URL: <http://ieeexplore.ieee.org/document/8099985/> (visited on 06/15/2022).
- Carvalho, Susana, Dora Biro, Eugénia Cunha, et al. (Mar. 2012a). “Chimpanzee carrying behaviour and the origins of human bipedality”. In: *Current Biology* 22.6, R180–R181. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0960982212000826> (visited on 01/27/2019).
- (Mar. 2012b). “Chimpanzee carrying behaviour and the origins of human bipedality”. en. In: *Curr. Biol.* 22.6, R180–1.
- Carvalho, Susana, Dora Biro, William C McGrew, et al. (Oct. 2009a). “Tool-composite reuse in wild chimpanzees (Pan troglodytes): archaeologically invisible steps in the technological evolution of early hominins?” In: *Animal Cognition* 12 (S1), pp. 103–114. URL: <http://link.springer.com/10.1007/s10071-009-0271-7> (visited on 12/11/2019).
- (Oct. 2009b). “Tool-composite reuse in wild chimpanzees (Pan troglodytes): archaeologically invisible steps in the technological evolution of early hominins?” en. In: *Anim. Cogn.* 12 Suppl 1, S103–14.
- Carvalho, Susana, Eugénia Cunha, et al. (July 2008). “Chânes opératoires and resource-exploitation strategies in chimpanzee (Pan troglodytes) nut cracking”. en. In: *J. Hum. Evol.* 55.1, pp. 148–163.
- Carvalho, Susana, Tetsuro Matsuzawa, and William C McGrew (2013). “From pounding to knapping: how chimpanzees can help us to model hominin lithics”. In: *Tool use in animals: cognition and ecology*, pp. 225–241.
- Carvalho, Susana and William C McGrew (2012). “The origins of the Oldowan: Why chimpanzees (Pan troglodytes) still are good models for technological evolution in Africa”. In: *Stone Tools and Fossil Bones: Debates in the Archaeology of Human Origins*, pp. 201–221.
- Castellano, Brandon (2020). “Pyscenedetect”. In: *Last accessed*.
- Ching, Travers et al. (Apr. 30, 2018). “Opportunities and obstacles for deep learning in biology and medicine”. In: *Journal of The Royal Society Interface* 15.141. Publisher: Royal Society, p. 20170387. URL: <https://royalsocietypublishing.org/doi/full/10.1098/rsif.2017.0387> (visited on 03/31/2022).
- Christin, Sylvain, Éric Hervet, and Nicolas Lecomte (2019). “Applications for deep learning in ecology”. In: *Methods in Ecology and Evolution* 10.10. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13256>, pp. 1632–1644. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13256> (visited on 04/05/2022).
- Clutton-Brock, Tim (2012a). “Long-Term, Individual-Based Field Studies”. In: *Long-Term Field Studies of Primates*. Ed. by Peter M Kappeler and David P Watts. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 437–449.
- (2012b). “Long-term, individual-based field studies”. In: *Long-term field studies of primates*. Springer, pp. 437–449.
- Clutton-Brock, Tim and Ben C Sheldon (2010a). “Individuals and populations: the role of long-term, individual-based studies of animals in ecology and evolutionary biology”. In: *Trends Ecol. Evol.* 25.10, pp. 562–573.
- (2010b). “Individuals and populations: the role of long-term, individual-based studies of animals in ecology and evolutionary biology”. In: *Trends in ecology & evolution* 25.10, pp. 562–573.

- Couzin, Iain D. (Jan. 1, 2009). "Collective cognition in animal groups". In: *Trends in Cognitive Sciences* 13.1, pp. 36–43. URL: <https://www.sciencedirect.com/science/article/pii/S1364661308002520> (visited on 04/22/2022).
- Curio, Eberhard (June 1, 1996). "Conservation needs ethology". In: *Trends in Ecology & Evolution* 11.6, pp. 260–263. URL: <https://www.sciencedirect.com/science/article/pii/0169534796200461> (visited on 05/09/2022).
- Davis, Grace H., Margaret C. Crofoot, and Damien R. Farine (July 1, 2018). "Estimating the robustness and uncertainty of animal social networks using different observational methods". In: *Animal Behaviour* 141, pp. 29–44. URL: <https://www.sciencedirect.com/science/article/pii/S0003347218301210> (visited on 05/11/2022).
- Deng, Jia et al. (2009). "Imagenet: A large-scale hierarchical image database". In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, pp. 248–255.
- De Waal, Frans B M (Dec. 2003). "Silent invasion: Imanishi's primatology and cultural bias in science". en. In: *Anim. Cogn.* 6.4, pp. 293–299.
- Dickinson, Janis L, Benjamin Zuckerberg, and David N Bonter (2010). "Citizen science as an ecological research tool: challenges and benefits". In: *Annual review of ecology, evolution, and systematics* 41, pp. 149–172.
- Dutta, A, A Gupta, and A Zissermann (2016). *VGG Image Annotator (VIA)*. URL: <http://www.robots.ox.ac.uk/%20vgg/software/via/>.
- Dutton, Paul and Hazel Chapman (Jan. 2015). "New tools suggest local variation in tool use by a montane community of the rare Nigeria–Cameroon chimpanzee, *Pan troglodytes ellioti*, in Nigeria". In: *Primates* 56.1, pp. 89–100.
- Estrada, Alejandro et al. (2017). "Impending extinction crisis of the world's primates: Why primates matter". In: *Science Advances* 3.1, e1600946.
- Farine, Damien R and Hal Whitehead (Sept. 2015). "Constructing, conducting and interpreting animal social network analysis". In: *J. Anim. Ecol.* 84.5. Ed. by Sonia Altizer, pp. 1144–1163.
- Franks, Daniel W., Graeme D. Ruxton, and Richard James (Jan. 2010). "Sampling animal association networks with the gambit of the group". In: *Behavioral Ecology and Sociobiology* 64.3, pp. 493–503. URL: <http://link.springer.com/10.1007/s00265-009-0865-8> (visited on 07/27/2020).
- Friard, Olivier and Marco Gamba (2016). "BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations". In: *Methods in Ecology and Evolution* 7.11. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12584>, pp. 1325–1330. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12584> (visited on 04/05/2022).
- Gilby, Ian C (Apr. 2006). "Meat sharing among the Gombe chimpanzees: harassment and reciprocal exchange". In: *Anim. Behav.* 71.4, pp. 953–963.
- Goodall, Jane (1963). "Feeding behaviour of wild chimpanzees: a preliminary report". In: *Sym. Zool. Soc. Lond.* 10, pp. 39–48.
- (1965). "Chimpanzees of the gombe stream reserve". In: *Primate behavior: Field studies of monkeys and apes*.

- (1986a). “The chimpanzees of Gombe: Patterns of behavior”. In: *Cambridge Mass.*
- (1986b). “The chimpanzees of Gombe: Patterns of behavior”. In: *Cambridge Mass.*
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*.
<http://www.deeplearningbook.org>. MIT Press.
- Gruber, Thibaud, Michael Chimento, et al. (2022). “Efficiency fosters cumulative culture across species”. In: *Philosophical Transactions of the Royal Society B* 377.1843, p. 20200308.
- Gruber, Thibaud, Lydia V Luncz, et al. (June 2019). “Cultural change in animals: a flexible behavioural adaptation to human disturbance”. en. In: *Palgrave Communications* 5.1, pp. 1–9.
- Haddock, Steven Harold David and Casey W Dunn (2011). *Practical computing for biologists*. Issue: 57: 004 HAD. Sinauer Associates Sunderland, MA.
- Hinde, R. A. (1976a). “Interactions, Relationships and Social Structure”. In: *Man* 11.1. Publisher: [Wiley, Royal Anthropological Institute of Great Britain and Ireland], pp. 1–17. URL: <https://www.jstor.org/stable/2800384> (visited on 03/30/2022).
- Hinde, Robert A (1976b). “Interactions, relationships and social structure”. In: *Man*, pp. 1–17.
- Hobaiter, Catherine and Richard W Byrne (Sept. 2011). “The gestural repertoire of the wild chimpanzee”. en. In: *Anim. Cogn.* 14.5, pp. 745–767.
- Hobaiter, Catherine, Timothée Poisot, et al. (Sept. 30, 2014). “Social Network Analysis Shows Direct Evidence for Social Transmission of Tool Use in Wild Chimpanzees”. In: *PLoS Biology* 12.9. Ed. by Frans B. M. de Waal, e1001960. URL: <http://dx.plos.org/10.1371/journal.pbio.1001960> (visited on 07/20/2018).
- Hockings, Kimberley (Mar. 2007). “Human-chimpanzee coexistence at Bossou, the Republic of Guinea: a chimpanzee perspective”. en. PhD thesis.
- Hockings, Kimberley, Matthew R McLennan, et al. (Apr. 2015a). “Apes in the Anthropocene: flexibility and survival”. en. In: *Trends Ecol. Evol.* 30.4, pp. 215–222.
- Hockings, Kimberley J., Matthew R. McLennan, et al. (Apr. 1, 2015b). “Apes in the Anthropocene: flexibility and survival”. In: *Trends in Ecology & Evolution* 30.4, pp. 215–222. URL: <http://www.sciencedirect.com/science/article/pii/S0169534715000397> (visited on 12/09/2019).
- Hoppitt, Will (Dec. 5, 2017). “The conceptual foundations of network-based diffusion analysis: choosing networks and interpreting results”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1735, p. 20160418. URL: <https://royalsocietypublishing.org/doi/full/10.1098/rstb.2016.0418> (visited on 03/18/2022).
- Howard, Jeremy and Sylvain Gugger (2020). *Deep Learning for Coders with fastai and PyTorch*. O’Reilly Media.
- Høye, Toke T et al. (2021). “Deep learning and computer vision will transform entomology”. In: *Proceedings of the National Academy of Sciences* 118.2. Publisher: National Acad Sciences.
- Hughey, Lacey F. et al. (May 19, 2018). “Challenges and solutions for studying collective animal behaviour in the wild”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 373.1746. Publisher: Royal Society, p. 20170005. URL: <https://royalsocietypublishing.org/doi/full/10.1098/rstb.2017.0005> (visited on 05/20/2022).

- Humle, Tatyana (2011a). "Location and ecology". In: *The chimpanzees of Bossou and Nimba*. Springer Japan Tokyo, pp. 13–21.
- (2011b). "The 2003 epidemic of a flu-like respiratory disease at Bossou". In: *The chimpanzees of Bossou and Nimba*. Springer, pp. 325–333.
- (2011c). "The Tool Repertoire of Bossou Chimpanzees". In: *The Chimpanzees of Bossou and Nimba*. Ed. by Tetsuro Matsuzawa, Tatyana Humle, and Yukimaru Sugiyama. Tokyo: Springer Japan, pp. 61–71.
- Humle, Tatyana and Tetsuro Matsuzawa (June 2004). "Oil palm use by adjacent communities of chimpanzees at bossou and nimba mountains, west Africa". In: *Int. J. Primatol.* 25.3, pp. 551–581.
- Imanishi, Kinji (1941). "Seibutsu-no-sekai (The world of living things)". In: *Kobundo, Tokyo*.
- Imanishi, Kinji and Pamela J Asquith (2013). *A Japanese View of Nature: The World of Living Things by Kinji Imanishi*. Routledge.
- Inoue-Nakamura, Noriko and Tetsuro Matsuzawa (June 1997). "Development of stone tool use by wild chimpanzees (*Pan troglodytes*)". en. In: *J. Comp. Psychol.* 111.2, pp. 159–173.
- Jolly, Alison (1972). *The evolution of primate behaviour*. Macmillan.
- Jones, James Holland (2011). "Primates and the evolution of long, slow life histories". In: *Current Biology* 21.18. Publisher: Elsevier, R708–R717.
- Jumper, John et al. (Aug. 2021). "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873. Number: 7873 Publisher: Nature Publishing Group, pp. 583–589. URL: <https://www.nature.com/articles/s41586-021-03819-2> (visited on 03/31/2022).
- Kano, Fumihiro, Yuri Kawaguchi, and Yeow Hanling (Mar. 8, 2022). "Experimental evidence that uniformly white sclera enhances the visibility of eye-gaze direction in humans and chimpanzees". In: *eLife* 11, e74086. URL: <https://elifesciences.org/articles/74086> (visited on 05/20/2022).
- Kappeler, Peter, ed. (2010). *Animal Behaviour: Evolution and Mechanisms*. Berlin, Heidelberg: Springer Berlin Heidelberg. URL: <http://link.springer.com/10.1007/978-3-642-02624-9> (visited on 04/23/2022).
- Kappeler, Peter M. and David P. Watts, eds. (2012). *Long-Term Field Studies of Primates*. Berlin, Heidelberg: Springer Berlin Heidelberg. URL: <http://link.springer.com/10.1007/978-3-642-22514-7> (visited on 05/01/2022).
- Kerner, Hannah (2020). "Too many AI researchers think real-world problems are not relevant". In: *Opinion. MIT Technology Review*.
- Köhler, W (1925). "The mentality of apes, trans. E. Winter". In: Publisher: Vintage Books/Harcourt, Brace.(Original work published in 1925)[BB, RAG].
- Koops, Kathelijne, William C McGrew, and Tetsuro Matsuzawa (Jan. 2013). "Ecology of culture: do environmental factors influence foraging tool use in wild chimpanzees, *Pan troglodytes* versus?" In: *Anim. Behav.* 85.1, pp. 175–185.
- Koops, Kathelijne, Aly Gaspard Soumah, et al. (Jan. 2022). "Field experiments find no evidence that chimpanzee nut cracking can be independently innovated". en. In: *Nature Human Behaviour*, pp. 1–8. URL: <https://www.nature.com/articles/s41562-021-01272-9> (visited on 01/26/2022).

- Kortlandt, A (1967). "Experimentation with chimpanzees in the wild". In: *Neue Ergebnisse der Primatologie*, pp. 208–244.
- Kramer, Karen L. (2010). "Cooperative Breeding and its Significance to the Demographic Success of Humans". In: *Annual Review of Anthropology* 39.1. _eprint: <https://doi.org/10.1146/annurev.anthro.012809.105054>, pp. 417–436. URL: <https://doi.org/10.1146/annurev.anthro.012809.105054> (visited on 05/12/2022).
- Krause, Jens et al. (2013). "Reality mining of animal social systems". In: *Trends in ecology & evolution* 28.9, pp. 541–551.
- Kühl, Hjalmar S et al. (Mar. 2019). "Human impact erodes chimpanzee behavioral diversity". en. In: *Science* 363.6434, pp. 1453–1455.
- LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *nature* 521.7553, pp. 436–444.
- Lehmann, Julia, Bonaventura Majolo, and Richard McFarland (Jan. 1, 2016). "The effects of social network position on the survival of wild Barbary macaques, *Macaca sylvanus*". In: *Behavioral Ecology* 27.1, pp. 20–28. URL: <https://doi.org/10.1093/beheco/arv169> (visited on 03/31/2022).
- Lemon, Stephenie C. et al. (Dec. 2003). "Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression". In: *Annals of Behavioral Medicine* 26.3, pp. 172–181. URL: <https://academic.oup.com/abm/article/26/3/172-181/4631556> (visited on 04/05/2022).
- Lonsdorf, Elizabeth V, Lynn E Eberly, and Anne E Pusey (Apr. 2004). "Sex differences in learning in chimpanzees". en. In: *Nature* 428.6984, pp. 715–716.
- Luncz, Lydia V and Christophe Boesch (Jan. 2015). "The extent of cultural variation between adjacent chimpanzee (*Pan troglodytes verus*) communities; a microecological approach". en. In: *Am. J. Phys. Anthropol.* 156.1, pp. 67–75.
- Luncz, Lydia V, Roger Mundry, and Christophe Boesch (May 2012). "Evidence for cultural differences between neighboring chimpanzee communities". en. In: *Curr. Biol.* 22.10, pp. 922–926.
- Luncz, Lydia V, Giulia Sirianni, et al. (Mar. 2018). "Costly culture: differences in nut-cracking efficiency between wild chimpanzee groups". In: *Anim. Behav.* 137, pp. 63–73.
- Luncz, Lydia V, Magdalena S Svensson, et al. (Aug. 2017). "Technological Response of Wild Macaques (*Macaca fascicularis*) to Anthropogenic Change". en. In: *Int. J. Primatol.* 38.5, pp. 872–880.
- Machanda, Zarin P. and Alexandra G. Rosati (Nov. 9, 2020). "Shifting sociality during primate ageing". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 375.1811. Publisher: Royal Society, p. 20190620. URL: <https://royalsocietypublishing.org/doi/full/10.1098/rstb.2019.0620> (visited on 05/11/2022).
- Marchant, Linda F., William C. McGrew, and Irenäus Eibl-Eibesfeldt (1995). "Is Human Handedness Universal? Ethological Analyses from Three Traditional Cultures". In: *Ethology* 101.3. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1439-0310.1995.tb00362.x>, pp. 239–258. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1439-0310.1995.tb00362.x> (visited on 04/07/2022).

- Marin-Jimenez, Manuel J et al. (2019). "LAEO-Net: Revisiting People Looking at Each Other in Videos". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3477–3485.
- Martin, Paul and Patrick Bateson (2007). *Measuring Behaviour*.
- Mathis, Alexander et al. (Sept. 2018). "DeepLabCut: markerless pose estimation of user-defined body parts with deep learning". In: *Nature Neuroscience* 21.9. Number: 9 Publisher: Nature Publishing Group, pp. 1281–1289. URL: <https://www.nature.com/articles/s41593-018-0209-y> (visited on 06/10/2022).
- Matsuzawa, Tetsuro (May 1985). "Use of numbers by a chimpanzee". en. In: *Nature* 315.6014, pp. 57–59.
- (1994). "Field experiments on use of stone tools by chimpanzees in the wild". In: *Chimpanzee cultures*. Ed. by McGrew William C de Waal F. B. M. & Heltne P. G. Wrangham R W. Cambridge, MA: Harvard University Press, pp. 351–370.
- (Dec. 2003). "The Ai project: historical and ecological contexts". en. In: *Anim. Cogn.* 6.4, pp. 199–211.
- (2011a). "Field Experiments of Tool-Use". In: *The Chimpanzees of Bossou and Nimba*. Springer, pp. 157–164.
- (2011b). "Field Experiments of Tool-Use". In: *The Chimpanzees of Bossou and Nimba*. Springer, pp. 157–164.
- Matsuzawa, Tetsuro, Dora Biro, et al. (2008a). "Emergence of culture in wild chimpanzees: Education by master-apprenticeship". In: *Primate Origins of Human Cognition and Behavior*. Tokyo: Springer Japan, pp. 557–574.
- (2008b). "Emergence of culture in wild chimpanzees: education by master-apprenticeship". In: *Primate origins of human cognition and behavior*. Springer, pp. 557–574.
- Matsuzawa, Tetsuro, Tatyana Humle, and Yukimaru Sugiyama (2011). *The chimpanzees of Bossou and Nimba*. Springer Science & Business Media.
- Matsuzawa, Tetsuro and William C McGrew (July 2008). "Kinji Imanishi and 60 years of Japanese primatology". en. In: *Curr. Biol.* 18.14, R587–91.
- McCarthy, Maureen S et al. (2019). "Camera traps provide a robust alternative to direct observations for constructing social networks of wild chimpanzees". In: *Animal Behaviour* 157, pp. 227–238.
- McCulloch, Warren S and Walter Pitts (1943). "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4, pp. 115–133.
- McGrew, William C (Apr. 2010). "Evolution. Chimpanzee technology". en. In: *Science* 328.5978, pp. 579–580.
- (Jan. 2017). "Field studies of Pan troglodytes reviewed and comprehensively mapped, focussing on Japan's contribution to cultural primatology". en. In: *Primates* 58.1, pp. 237–258.
- McGrew, William C, RM Ham, et al. (1997). "Why don't chimpanzees in Gabon crack nuts?" In: *International Journal of Primatology* 18.3, pp. 353–374.
- McGrew, William C and CEG Tutin (1978). "Evidence for a social custom in wild chimpanzees?" In: *Man*, pp. 234–251.
- Mercader, Julio, Huw Barton, et al. (Feb. 2007). "4,300-year-old chimpanzee sites and the origins of percussive stone technology". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 104.9, pp. 3043–3048.

- Mercader, Julio, Melissa Panger, and Christophe Boesch (May 2002). "Excavation of a chimpanzee stone tool site in the African rainforest". en. In: *Science* 296.5572, pp. 1452–1455.
- Migliano, Andrea Bamberg and Lucio Vinicius (Jan. 31, 2022). "The origins of human cumulative culture: from the foraging niche to collective intelligence". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 377.1843. Publisher: Royal Society, p. 20200317. URL: <https://royalsocietypublishing.org/doi/full/10.1098/rstb.2020.0317> (visited on 03/29/2022).
- Mobbs, Dean et al. (July 21, 2021). "Promises and challenges of human computational ethology". In: *Neuron* 109.14, pp. 2224–2238. URL: <https://www.sciencedirect.com/science/article/pii/S0896627321003743> (visited on 03/30/2022).
- Muller, Martin N (2017). *Chimpanzees and human evolution*. Harvard University Press.
- Musgrave, Stephanie et al. (Jan. 2020). "Teaching varies with task complexity in wild chimpanzees". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 117.2, pp. 969–976.
- Nakamura, Michio et al. (Sept. 2015). *Mahale Chimpanzees: 50 Years of Research*. en. Cambridge University Press.
- Nishida, Toshisada (1968). "The social group of wild chimpanzees in the Mahali Mountains". In: *Primates* 9.3, pp. 167–224.
- Nishida, Toshisada, Takahisa Matsusaka, and William C. McGrew (Jan. 2009). "Emergence, propagation or disappearance of novel behavioral patterns in the habituated chimpanzees of Mahale: a review". In: *Primates* 50.1, pp. 23–36. URL: <http://link.springer.com/10.1007/s10329-008-0109-y> (visited on 05/13/2022).
- Nishida, Toshisada, Koichiro Zamma, et al. (2010). *Chimpanzee behavior in the wild: an audio-visual encyclopedia*. Springer Science & Business Media.
- Nissen, Henry W (1931). *A Field Study of the Chimpanzee: Observations of Chimpanzee Behavior and Environment in Western French Guinea. With a Foreword by Robert M. Yerkes*. Johns Hopkins Press.
- Northcutt, Curtis G., Anish Athalye, and Jonas Mueller (Nov. 7, 2021). "Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks". In: *arXiv:2103.14749 [cs, stat]*. arXiv: 2103.14749. URL: <http://arxiv.org/abs/2103.14749> (visited on 05/10/2022).
- O'Brien, Michael J. and Stephen J. Shennan, eds. (2009). *Innovation in Cultural Systems: Contributions from Evolutionary Anthropology*. The MIT Press. URL: <https://direct.mit.edu/books/book/3197/innovation-in-cultural-systemscontributions-from> (visited on 05/13/2022).
- Parkhi, Omkar M, Andrea Vedaldi, and Andrew Zisserman (2015). "Deep face recognition". In.
- Pereira, Talmo D., Joshua W. Shaevitz, and Mala Murthy (Dec. 2020). "Quantifying behavior to understand the brain". In: *Nature Neuroscience* 23.12. Number: 12. Publisher: Nature Publishing Group, pp. 1537–1549. URL: <https://www.nature.com/articles/s41593-020-00734-z> (visited on 05/07/2022).
- Plooij, Frans X et al. (May 2015). "An archive of longitudinal recordings of the vocalizations of adult Gombe chimpanzees". en. In: *Sci Data* 2, p. 150027.

- Ragusa, Francesco et al. (Oct. 12, 2020). “The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain”. In: *arXiv:2010.05654 [cs]*. version: 1. arXiv: 2010.05654. URL: <http://arxiv.org/abs/2010.05654> (visited on 03/31/2022).
- Rahwan, Iyad et al. (Apr. 2019). “Machine behaviour”. In: *Nature* 568.7753. Number: 7753 Publisher: Nature Publishing Group, pp. 477–486. URL: <https://www.nature.com/articles/s41586-019-1138-y> (visited on 04/08/2022).
- Rosati, Alexandra G., Alyssa M. Arre, et al. (May 11, 2016). “Rhesus monkeys show human-like changes in gaze following across the lifespan”. In: *Proceedings of the Royal Society B: Biological Sciences* 283.1830. Publisher: Royal Society, p. 20160376. URL: <https://royalsocietypublishing.org/doi/full/10.1098/rspb.2016.0376> (visited on 05/12/2022).
- Rosati, Alexandra G., Lindsey Hagberg, et al. (Oct. 23, 2020). “Social selectivity in aging wild chimpanzees”. In: *Science* 370.6515. Publisher: American Association for the Advancement of Science, pp. 473–476. URL: <https://www.science.org/doi/full/10.1126/science.aaz9129> (visited on 05/12/2022).
- Rowe, Noel and Marc Myers (2011). “Photography and video for field researchers”. en. In: *Field and Laboratory Methods in Primatology*. Ed. by Joanna M. Setchell and Deborah J. Curtis. 2nd ed. Cambridge: Cambridge University Press, pp. 305–318. URL: https://www.cambridge.org/core/product/identifier/CB09780511921643A025/type/book_part (visited on 01/26/2022).
- Rubenstein, Dustin R. and Patrick Abbot (2017). “The Evolution of Social Evolution”. In: *Comparative Social Evolution*. Ed. by Dustin R. Rubenstein and Patrick Abbot. Cambridge: Cambridge University Press, pp. 1–18. URL: https://www.cambridge.org/core/product/identifier/9781107338319%23CT-bp-1/type/book_part (visited on 05/04/2022).
- Sakura, Osamu and Tetsuro Matsuzawa (1991). “Flexibility of wild chimpanzee nut-cracking behavior using stone hammers and anvils: An experimental analysis”. In: *Ethology*.
- Samuel, Arthur L (1962). “Artificial intelligence: a frontier of automation”. In: *The Annals of the American Academy of Political and Social Science* 340.1, pp. 10–20.
- Sarabian, Cecile and Andrew J J MacIntosh (Nov. 2015). “Hygienic tendencies correlate with low geohelminth infection in free-ranging macaques”. en. In: *Biol. Lett.* 11.11.
- Schofield, Daniel, William C McGrew, et al. (Mar. 1, 2018). “Cumulative culture in nonhumans: overlooked findings from Japanese monkeys?” In: *Primates* 59.2, pp. 113–122. URL: <https://doi.org/10.1007/s10329-017-0642-7> (visited on 12/08/2019).
- Schofield, Daniel, Arsha Nagrani, et al. (2019). “Chimpanzee face recognition from videos in the wild using deep learning”. In: *Science advances* 5.9, eaaw0736.
- Setchell, Joanna M and Deborah J Curtis (Feb. 2011). *Field and Laboratory Methods in Primatology: A Practical Guide*. en. Cambridge University Press.
- Shizuka, Daizaburo and Allison E Johnson (June 8, 2019). “How demographic processes shape animal social networks”. In: *Behavioral Ecology*, arz083. URL: <https://academic.oup.com/beheco/advance-article/doi/10.1093/beheco/arz083/5512707> (visited on 12/10/2019).
- (Jan. 31, 2020). “The long view on demographic effects on social networks: a response to comments on Shizuka and Johnson”. In: *Behavioral Ecology* 31.1. Ed. by

- Leigh Simmons, pp. 19–20. URL:
<https://academic.oup.com/beheco/article/31/1/19/5618606> (visited on 09/07/2021).
- Strandburg-Peshkin, Ariana et al. (2015). “Shared decision-making drives collective movement in wild baboons”. In: *Science* 348.6241, pp. 1358–1361.
- Strum, Shirley C, Graham Stirling, and Steve Kalusi Mutunga (2015a). “The perfect storm: Land use change promotes *Opuntia stricta*’s invasion of pastoral rangelands in Kenya”. In: *Journal of Arid Environments* 118. Publisher: Elsevier, pp. 37–47.
- (July 2015b). “The perfect storm: Land use change promotes *Opuntia stricta*’s invasion of pastoral rangelands in Kenya”. In: *J. Arid Environ.* 118, pp. 37–47.
- Sugiyama, Yukimaru (1995a). “Drinking tools of wild chimpanzees at Bossou”. en. In: *Am. J. Primatol.* 37.3, pp. 263–269.
- (Apr. 1995b). “Tool-use for catching ants by chimpanzees at Bossou and Monts Nimba, West Africa”. en. In: *Primates* 36.2, pp. 193–205.
- Sugiyama, Yukimaru and Jeremy Koman (Oct. 1979). “Tool-using and -making behavior in wild chimpanzees at Bossou, Guinea”. en. In: *Primates* 20.4, pp. 513–524.
- Suzuki, Akira (Aug. 1965). “An ecological study of wild Japanese monkeys in snowy areas”. en. In: *Primates* 6.1, pp. 31–72.
- Swanson, Alexandra et al. (June 2015). “Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna”. en. In: *Sci Data* 2, p. 150026.
- Takeshita, Rafaela S C et al. (May 2018). “Beneficial effect of hot spring bathing on stress levels in Japanese macaques”. en. In: *Primates* 59.3, pp. 215–225.
- Tan, Amanda et al. (May 13, 2015). “There Is More than One Way to Crack an Oyster: Identifying Variation in Burmese Long-Tailed Macaque (*Macaca fascicularis aurea*) Stone-Tool Use”. In: *PLOS ONE* 10.5, e0124733. URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0124733> (visited on 12/08/2019).
- Thompson, Melissa Emery (2013). “Reproductive Ecology of Female Chimpanzees”. In: *American Journal of Primatology* 75.3. eprint:
<https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajp.22084>, pp. 222–237. URL:
<https://onlinelibrary.wiley.com/doi/abs/10.1002/ajp.22084> (visited on 05/16/2022).
- Tinbergen, Niko (1963). “On aims and methods of ethology”. In: *Zeitschrift für tierpsychologie* 20.4. Publisher: Wiley Online Library, pp. 410–433.
- Trask, Andrew W (2019). *Grokking deep learning*. Simon and Schuster.
- Tutin, CEG et al. (Jan. 1, 2001). “CHARTING CULTURAL VARIATION IN CHIMPANZEES”. In: *Behaviour* 138.11. Publisher: Brill, pp. 1481–1516. URL:
https://brill.com/view/journals/beh/138/11-12/article-p1481_11.xml (visited on 03/25/2022).
- Von Marcard, Timo et al. (Mar. 24, 2017). “Sparse Inertial Poser: Automatic 3D Human Pose Estimation from Sparse IMUs”. In: *arXiv:1703.08014 [cs]*. arXiv: 1703.08014. URL: <http://arxiv.org/abs/1703.08014> (visited on 03/25/2022).
- Voulodimos, Athanasios et al. (Feb. 1, 2018). “Deep Learning for Computer Vision: A Brief Review”. In: *Computational Intelligence and Neuroscience* 2018. Publisher: Hindawi, e7068349. URL:
<https://www.hindawi.com/journals/cin/2018/7068349/> (visited on 03/30/2022).

- Washburn, S L and Irvén DeVore (1961). “The Social Life of Baboons”. In: *SCIENTIFIC AMERICAN*, p. 11.
- Watanabe, Kunio, Akio Mori, and Masao Kawai (Jan. 1992). “Characteristic features of the reproduction of Koshima monkeys, *Macaca fuscata fuscata*: A summary of thirty-four years of observation”. en. In: *Primates* 33.1, pp. 1–32.
- Weinstein, Ben G (2018). “A computer vision for animal ecology”. In: *Journal of Animal Ecology* 87.3, pp. 533–545.
- Whiten, A, J Goodall, et al. (June 1999). “Cultures in chimpanzees”. en. In: *Nature* 399.6737, pp. 682–685.
- Whiten, Andrew (Nov. 2, 2019). “Cultural Evolution in Animals”. In: *Annual Review of Ecology, Evolution, and Systematics* 50.1, pp. 27–48. URL: <https://www.annualreviews.org/doi/10.1146/annurev-ecolsys-110218-025040> (visited on 12/09/2019).
- Whiten, Andrew, Francisco J. Ayala, et al. (July 25, 2017). “The extension of biology through culture”. In: *Proceedings of the National Academy of Sciences* 114.30, pp. 7775–7781. URL: <https://www.pnas.org/content/114/30/7775> (visited on 12/08/2019).
- Whiten, Andrew, Christine A Caldwell, and Alex Mesoudi (2016). “Cultural diffusion in humans and other animals”. In: *Current opinion in Psychology* 8. Publisher: Elsevier, pp. 15–21.
- Whiten, Andrew, Rachel A. Harrison, et al. (Jan. 31, 2022). “Collective knowledge and the dynamics of culture in chimpanzees”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 377.1843. Publisher: Royal Society, p. 20200321. URL: <https://royalsocietypublishing.org/doi/10.1098/rstb.2020.0321> (visited on 04/01/2022).
- Wild, Sonja et al. (2019). “Multi-network-based diffusion analysis reveals vertical cultural transmission of sponge tool use within dolphin matriline”. In: *Biology letters* 15.7, p. 20190227.
- Wittig, Roman M (2018). *40 years of research at the Tai Chimpanzee Project*.
- Yamakoshi, Gen (1998). “Dietary responses to fruit scarcity of wild chimpanzees at Bossou, Guinea: possible implications for ecological importance of tool use”. In: *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists* 106.3, pp. 283–295.
- (2011). “The “Prehistory” Before 1976: Looking Back on Three Decades of Research on Bossou Chimpanzees”. In: *The Chimpanzees of Bossou and Nimba*. Ed. by Tetsuro Matsuzawa, Tatyana Humle, and Yukimaru Sugiyama. Tokyo: Springer Japan, pp. 35–44.
- Yamakoshi, Gen and Yukimaru Sugiyama (Oct. 1995). “Pestle-pounding behavior of wild chimpanzees at Bossou, Guinea: A newly observed tool-using behavior”. en. In: *Primates* 36.4, pp. 489–500.
- Yerkes, Robert Mearns and Ada W Yerkes (1929). “The great apes.” In: Publisher: Yale University Press.
- Young, Tom et al. (Aug. 2018). “Recent Trends in Deep Learning Based Natural Language Processing [Review Article]”. In: *IEEE Computational Intelligence Magazine* 13.3. Conference Name: IEEE Computational Intelligence Magazine, pp. 55–75.
- Zuberbühler, Klaus and Roman M. Wittig (2011). “Field experiments with non-human primates: a tutorial”. In: *Field and Laboratory Methods in Primatology*. Ed. by

Joanna M. Setchell and Deborah J. Curtis. 2nd ed. Cambridge: Cambridge University Press, pp. 207–224. URL: https://www.cambridge.org/core/product/identifier/CB09780511921643A019/type/book_part (visited on 01/26/2022).