RESEARCH METHODS

Chimpanzee face recognition from videos in the wild using deep learning

Daniel Schofield[1]*†, Arsha Nagrani[2]*†, Andrew Zisserman[2], Misato Hayashi[3], Tetsuro Matsuzawa[3], Dora Biro[4], 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.

[1]Primate Models for Behavioural Evolution Lab, Institute of Cognitive and Evolutionary Anthropology, University of Oxford, Oxford, UK. [2]Visual Geometry Group, Department of Engineering Science, University of Oxford, Oxford, UK. [3]Primate Research Institute, Kyoto University, Inuyama, Japan. [4]Department of Zoology, University of Oxford, Oxford, UK. [5]Gorongosa National Park, Sofala, Mozambique. [6]Interdisciplinary Center for Archaeology and Evolution of Human Behaviour (ICArEHB), Universidade do Algarve, Faro, Portugal. [7]Centre 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 multiclass classification, classification error is only very rarely uniformly distributed for all classes. To understand where the recognition

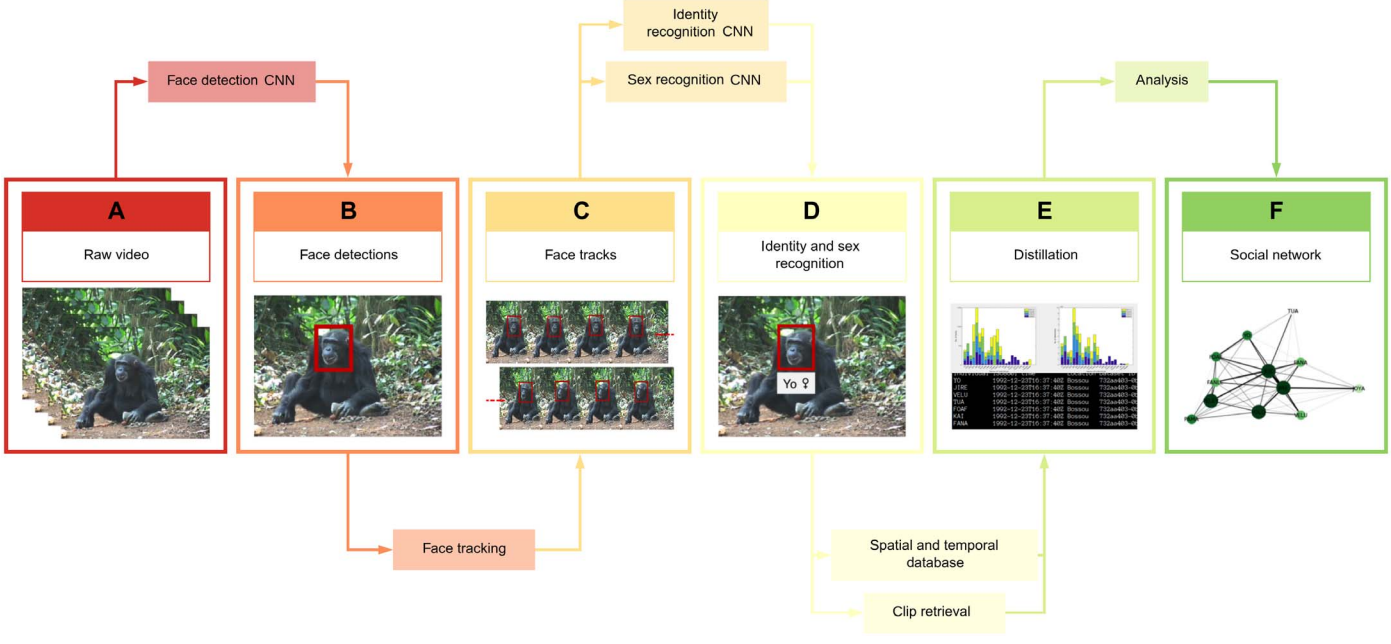

Fig. 1. Fully unified pipeline for wild chimpanzee face tracking and recognition from raw video footage. The pipeline consists of the following stages: (**A**) Frames are extracted from raw video. (**B**) Detection of faces is performed using a deep CNN single-shot detector (SSD) model. (**C**) Face 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)

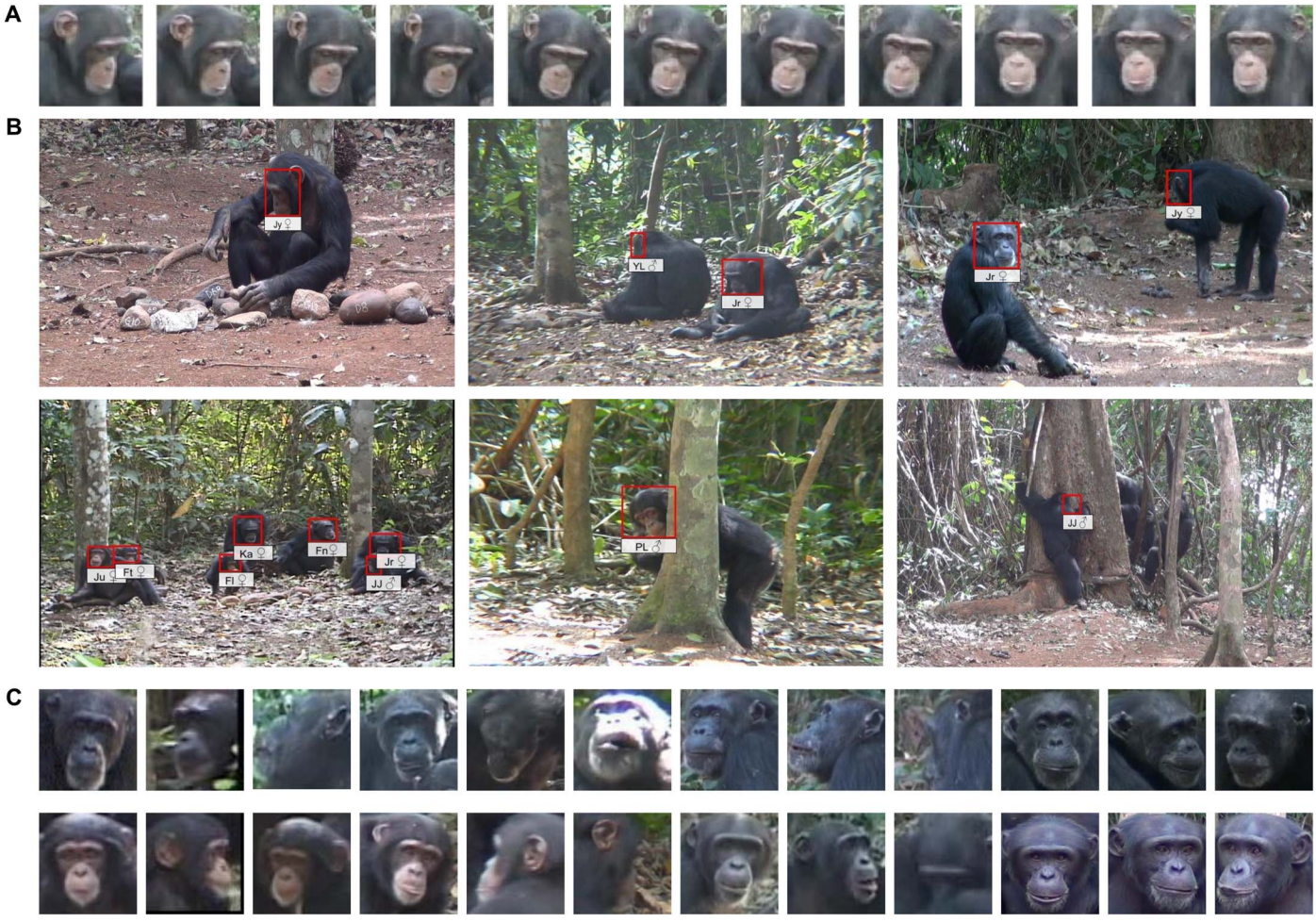

Fig. 2. Face recognition results demonstrating the CNN model's robustness to variations in pose, lighting, scale, and age over time. (**A**) Example of a correctly labeled face track. The first two faces (nonfrontal) were initially labeled incorrectly by the model but were corrected automatically by 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

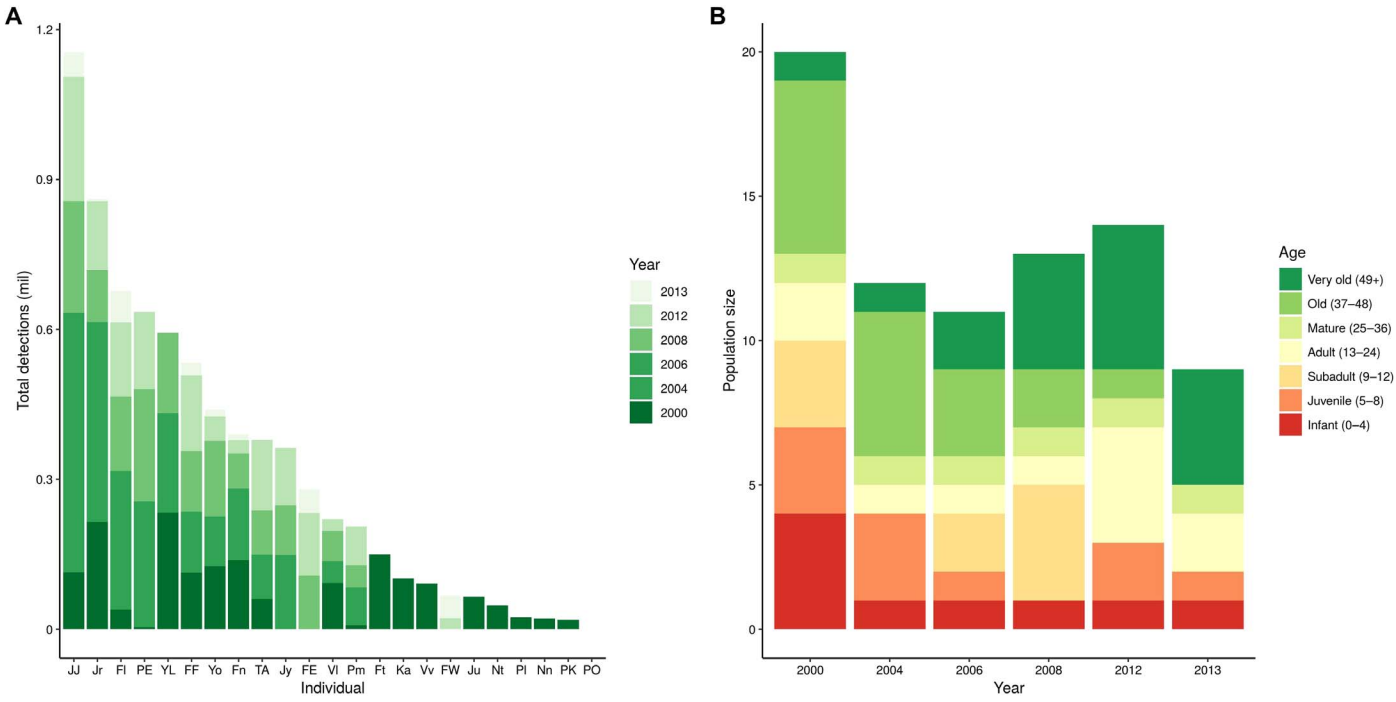

Fig. 3. Face detection and recognition results. (**A**) Histograms of detection numbers for individuals in the training and test years of the dataset (2000, 2004, 2006, 2008, 2012, and 2013). (**B**) Output of model for number of individuals detected in each year and proportion of individuals in different 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 ± 24.33%), while novice annotators (n = 3) performed significantly worse (20.67 ± 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

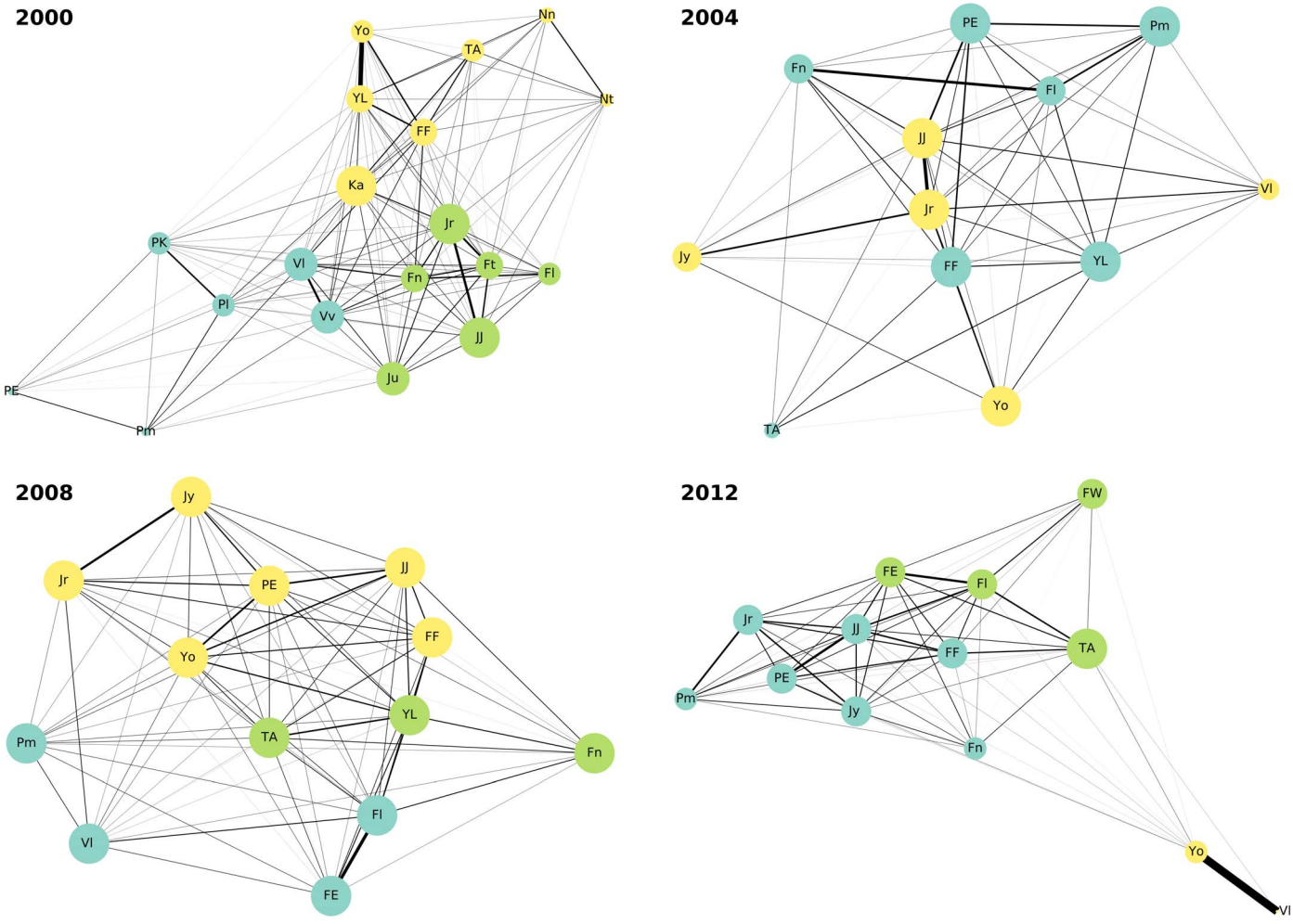

Fig. 4. Social networks of the Bossou community generated from co-occurrence matrices constructed using detections of the face recognition model. Each node represents an individual chimpanzee. Node size corresponds to the individual's degree centrality—the total number of "edges" (connections) they have (the higher the degree centrality, the larger the node). Node colors correspond to subclusters of the community as identified independently in each year using the Louvain community detection algorithm (*23*). Individuals whose ID codes begin with the same letter belong to the same matriline; IDs in capital letters correspond to males, while IDs with only the first letter capitalized correspond to females (see table S1). Within these clusters, as predicted, mothers and young infants have the strongest co-occurrences, and kin cluster into the same subgroups.

cues, e.g., behavioral context, full body posture and movement, handedness, and proximity to other individuals, while those coding video footage have the possibility to replay scenes.

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 with smaller datasets (*27*). Hence, the features learnt by our deep model will likely also be useful for other primate-related tasks, even if the datasets are smaller.

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

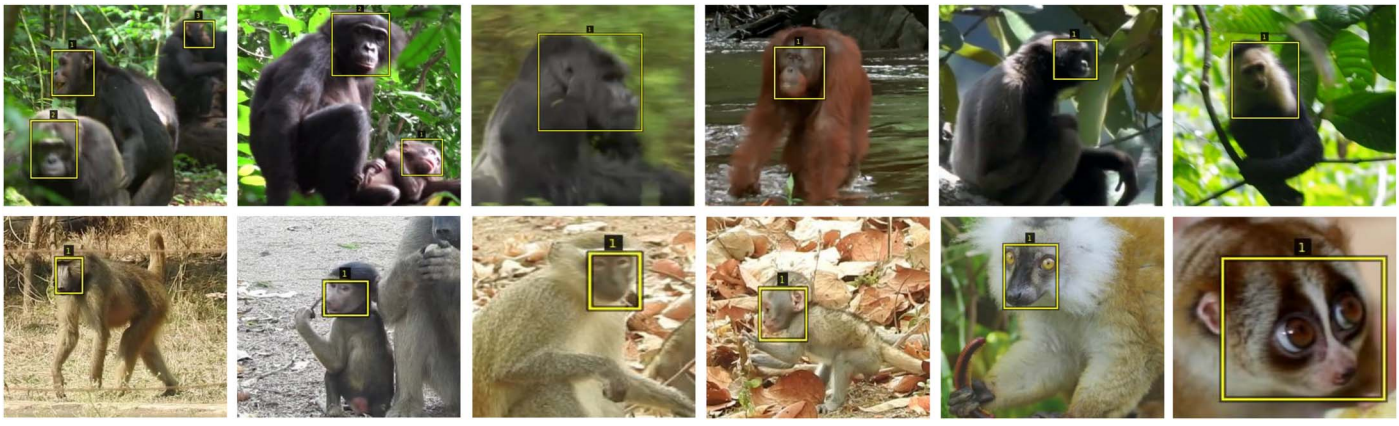

Fig. 5. Preliminary results from the face detector model tested on other primate species. Top row: *P. troglodytes schweinfurthii*, *Pan paniscus*, *Gorilla beringei*, *Pongo pygmaeus*, *Hylobates muelleri*, and *Cebus imitator*. **Bottom row:** *Papio ursinus* (x2), *Chlorocebus pygerythrus* (x2), *Eulemur macaco*, and *Nycticebus coucang*. Image sources: Chimpanzee: www.youtube.com/watch?v=c2u3NKXbGeo; Bonobo: www.youtube.com/watch?v=JF8v_HWvfLc&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 spatiotemporal 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

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

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

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

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 $\alpha_{i,j}(\delta t)$ 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

$$N_i = \sum_j \alpha_{i,j}(\delta t)$$

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

$$N_j = \sum_i \alpha_{i,j}(\delta t)$$

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

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

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. Kühl, 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, LemurFaceID: 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. Kühl, 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. Kühl, 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).

Science Advances

Chimpanzee face recognition from videos in the wild using deep learning

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

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