ORGANISMAL BIOLOGY

Automated audiovisual behavior recognition in wild primates

Max Bain[1]*, Arsha Nagrani[1]†, Daniel Schofield[2], Sophie Berdugo[2,3], Joana Bessa[4], Jake Owen[4], Kimberley J. Hockings[5], Tetsuro Matsuzawa[6], Misato Hayashi‡§, Dora Biro[4,7], Susana Carvalho[2,8,9,10], Andrew Zisserman[1]

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 biologgers, 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

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

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

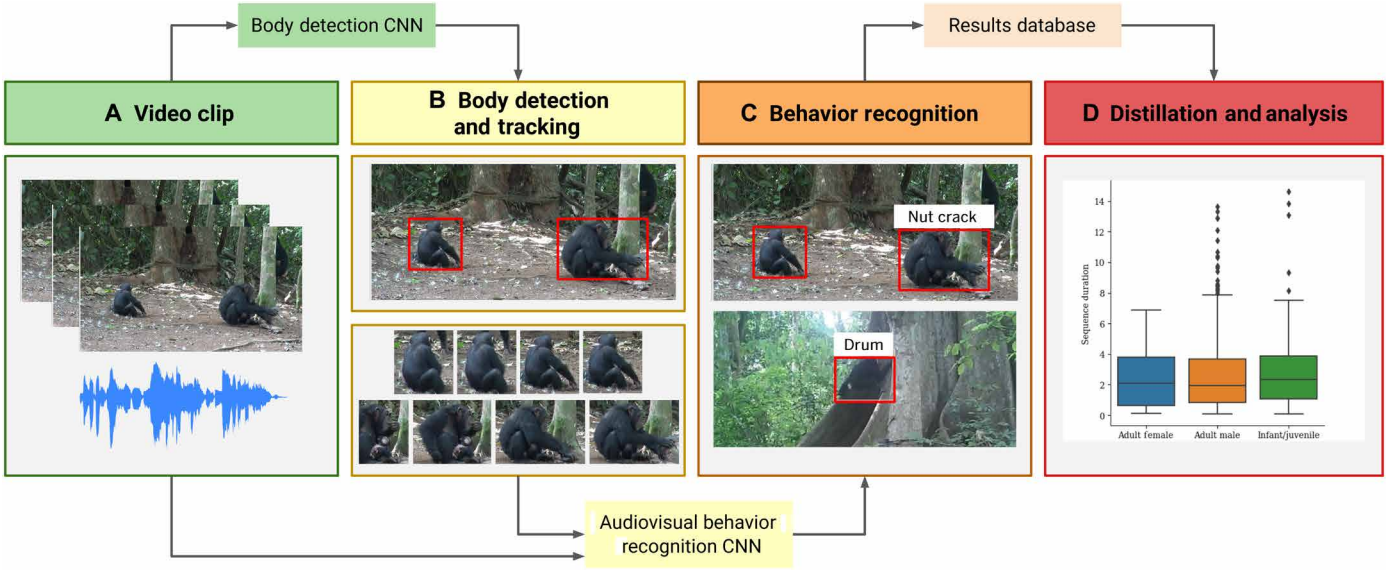

Fig. 1. Fully unified pipeline for wild chimpanzee behavior recognition and analysis from raw video footage. The pipeline consists of the following stages: (**A**) Frames and audio are extracted from raw video. (**B**) Body detection is performed over the video frames using a deep CNN single-shot detector (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).

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**

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 ± SD = 9.21 ± 9.49%; and females, mean ± SD = 7.97 ± 9.19%), while juveniles required longer nut cracking sequences per nut consumed than adult males and females (males, mean ± SD =16.8 ± 6.46 s; females, mean ± SD =15.7 ± 10.41 s; and juveniles/infants, mean ± 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 ± SD = 2.21 ± 1.80 s) than immature individuals (mean ± SD = 2.72 ± 2.39 s) and adult females (mean ± 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 ± SD = 1.74 ± 1.13 s; and females, mean ± SD = 2.17 ± 1.09 s) than agonistic display (males, mean ± SD = 2.31 ± 2.61 s; and females, mean ± SD = 2.78 ± 1.78 s) and travel drumming bouts (males, mean ± SD = 2.27 ± 1.35 s; and females, mean ± SD = 2.89 ± 1.97 s).

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

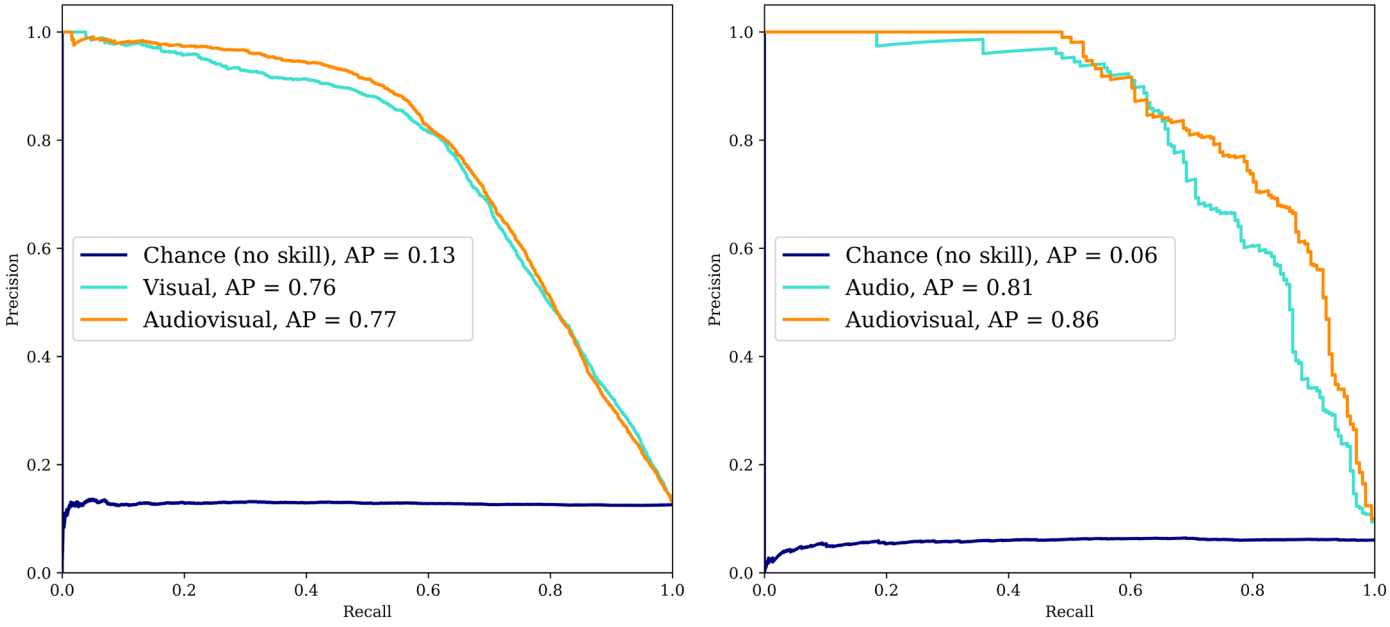

Fig. 3. Performance of the audio, visual, and audiovisual models for individual-level behavior classification. The curves for nut-cracking (**left**) and buttress-drumming (**right**) demonstrate that audiovisual outperforms single-modality methods. Instances where the behavior is either visually or 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

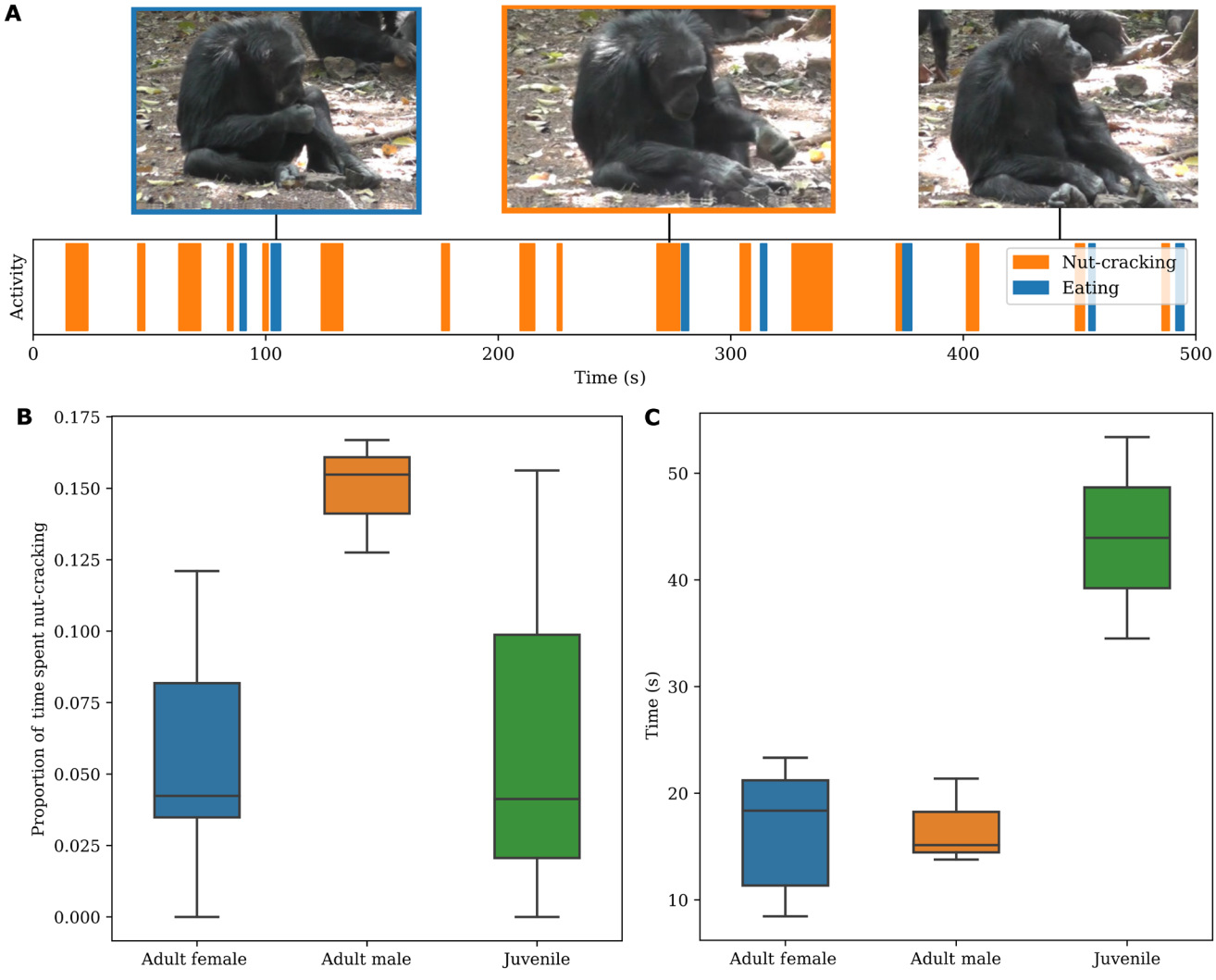

Fig. 4. Nut cracking analysis. (**A**) An example activity sequence following a single individual over the course of a video. The blank white spaces are any activities that are not nut cracking or eating. Note that eating typically follows nut cracking events. (**B**) Proportion of time spent nut cracking 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 biologgers. 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.

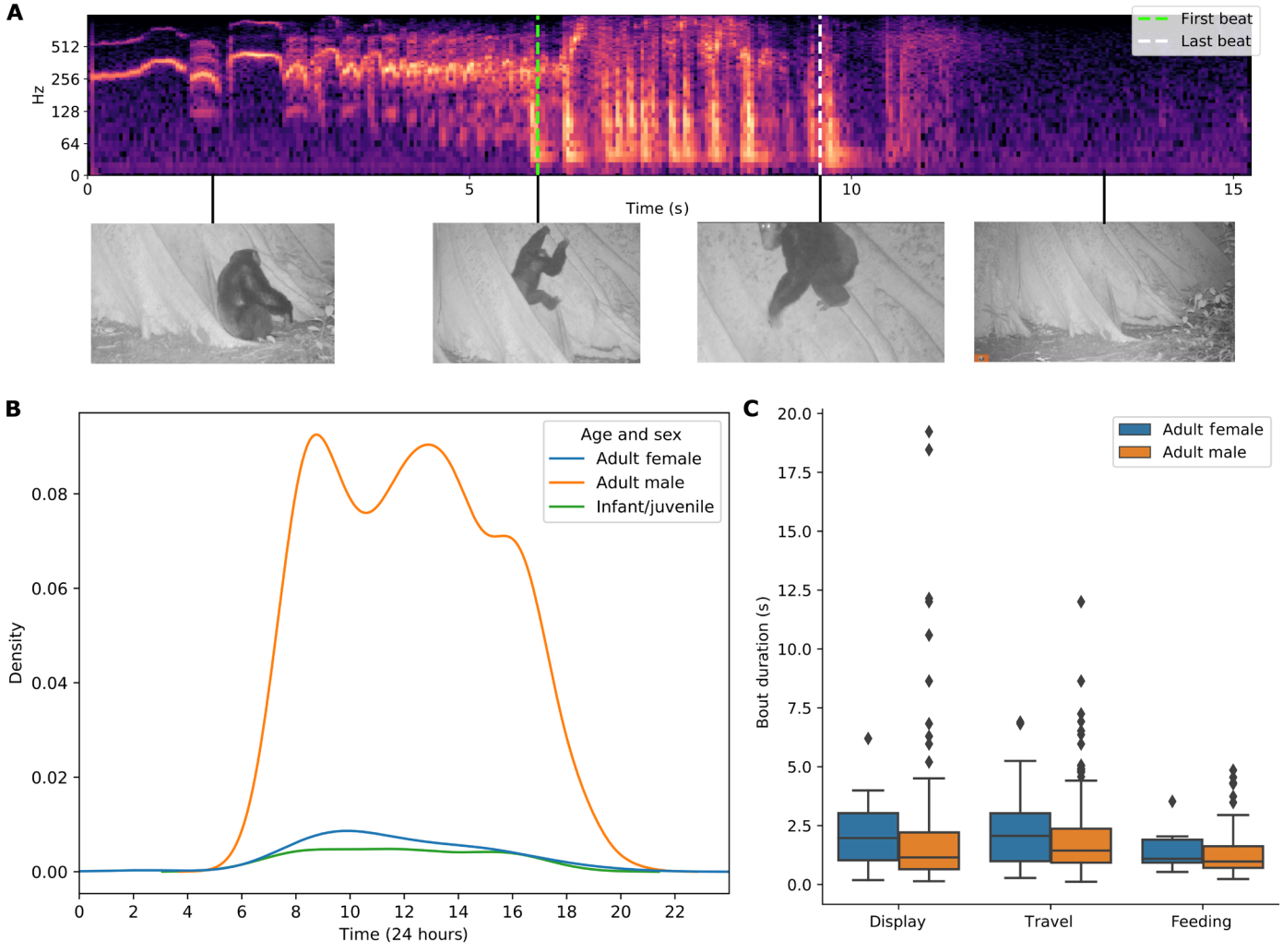

Fig. 5. Buttress drumming analysis. (**A**) Spectrogram showing a detected drumming bout delineated by the first and last beats, with video frames visualized. (**B**) Kernel density estimation plot showing the diel distribution of buttress drumming bouts, based on hh:mm:ss data captured by camera traps. (**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(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 dimensional 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 × 128 × 128 × 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 bouts 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. Schläppi, F. Akyol, M. Privitera, D. Slominski, 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. Tennessen, 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] (21 November 2020); http://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] (24 April 2018); http://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. Kujirakwinja, K. E. Langergraber, J. Lapuente, B. Larson, K. Lee, V. Leinert, M. Llana, G. Maretti, S. 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. Jojic, 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 dese 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. Slocombe, 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-78pdaXPdaIUJmdJzSuqhXcb48vHM).

Submitted 12 March 2021
Accepted 23 September 2021
Published 12 November 2021
10.1126/sciadv.abi4883

ScienceAdvances

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

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.