Automated Recognition of Grooming Behavior in Wild Chimpanzees

Yana van de Sande (yana.vandesande@ru.nl)
Radboud University Nijmegen

Lara Southern (lasouthern@uos.de)
Institute of Cognitive Science, Comparative BioCognition, Osnabrück University, Artilleriestrasse 34, 49076, Osnabrück, Germany

Wim Pouw (wim.pouw@ru.nl)
Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen

Abstract

Video recording is a widely used tool for studying animal behavior, especially in fields such as primatology. Primatologists rely on video data to analyze and research topics such as social grooming to uncover subtle mechanisms behind complex social behavior and structures. Insights into these social behaviors may provide us with a better understanding of our closest living relatives, but also new theories and insights into our own behavior. However, analyzing this type of data using manual annotation is currently a time-consuming task. Here we present an end-to-end pipeline to track chimpanzee (*Pan troglodytes)* poses using DeepLabCut (DLC) which then serves as input to a support vector machine. This classifier was trained to detect role transitions within grooming interactions. We replicate a recent study showing that DLC has usability value for chimpanzee data collected in natural environments. Our combined method of tracking and classification is remarkably successful in detecting the presence of grooming, indicating the directionality and a change in turn with an accuracy above 86% on unseen videos. We can identify particular pose features used in the classification of grooming, which will contribute to the exploration of turn-taking dynamics on a scale that would otherwise be difficult to attain with traditional methods. Finally, our pipeline can in principle be applied to recognize a variety of other socially interactive behaviors that are largely recognizable by (joint) postural states.

Keywords: automatic behavior recognition, pose estimation, DeepLabCut, social interaction, grooming, chimpanzee, in the wild, naturalistic environment

Introduction

One of the main challenges in the analysis of primate behavior in the wild is the annotation process. Data is often recorded continuously in natural environments, from various angles and conditions. These conditions result in large datasets consisting of video material with high levels of variation in quality. The current manual processing and the analysis of this data takes a lot of time and effort, in part because the annotation process requires both standardization and validation. Additionally, as a researcher needs to search for relevant behaviors (e.g. grooming) while often also needing to select for specific individuals (e.g. those of a certain social rank, sex and/or age). This process of scanning across large video data sets is extremely time intensive.

Grooming behavior in primatesSocial grooming is one of the most important social interactions for establishing and fortifying social standing (Dunbar, 1988; Mielke et al., 2018; Seyfarth & Cheney, 2012). *Social grooming*, or *grooming* from here forward, is "a tactile and highly visual behavior in which one individual closely scrutinizes and meticulously picks through the hair of another, removing debris and insects with their hands and/or mouth" (Goodall, 1986). While researchers recognize grooming's significance extends beyond basic hygiene, the specific ways these interactions are structured remain a topic of investigation. Additionally, social grooming has been proposed as a context to study turn-taking dynamics in communication due its high levels of negotiation, and as such a fundamental building steppingstone for the evolution of language and higher (social) cognitive abilities (Fedurek et al., 2015; Grueter et al., 2012; Levison & Torreira, 2015; Pika, 2014). Therefore, primatologists are not only calling for unified methods, but for more long-term data collection and ways to analyze more data (Pika et al., 2018; Radhakrishna & Jamieson, 2018; Strier, 2003). Here novel technologies offer potential solutions, promising a more efficient and replicable approach while maintaining transparency.

Related work

With recent advances in deep learning, interdisciplinary methodologies have opened new ways to process and analyze data. The progress in this domain has been focused on two tasks that are often combined. Firstly, there have been developments in animal pose tracking. Secondly, there has been significant development in behavior classification.

With regards to animal pose tracking, Wiltshire et al. (2023) utilized DeepLabCut (DLC) (Lauer et al., 2022; Mathis et al., 2018) for pose estimation in wild chimpanzees to validate its use for in-the-wild data. DLC was applied in the context of multi-animal videos and two models were trained on data that represented raw data as much as possible. Wiltshire et al. (2023) report the tracking performance of their models to be substantially better than inter-human coding variation. While developments in primate pose tracking are accelerating (Desai et al., 2023; Hu et al., 2023; Wu et al., 2019), Wiltshire et al. (2023) were able to stress-test for the first time that DLC is a promising tool for pose-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

tracking of chimpanzees in the wild. We used their results and reflections as a reference point for aspects of this study.

Developments have also been made regarding behavioral classification in primatology. Sakib and Burghardt (2019) utilized convolutional neural network (CNN) architectures to demonstrate the viability of automated behavior detection and recognition. Their system succeeded in detecting and recognizing nine core ape behaviors (camera interaction, climbing down, climbing up, hanging, running, sitting, sitting on back, standing, and walking) in challenging camera trap footage with a top-3 average score of 94.07% after cross-validation.

Bain et al. (2021) used multimodal video-audio recordings to recognize nut cracking and buttress drumming. The efficacy of using deep neural network models for a biological application was shown using a CNN approach: multimodal classification of nut cracking resulted in an average precision of 0.77 and buttress drumming 0.86. A much higher precision was obtained in comparison to unimodal classification (audio only: nut cracking = 0.30, buttress drumming = 0.81, visual only: nut cracking = 0.76, and buttress drumming = 0.64).

These studies, however, focus on activities primarily conducted in solitude. In addition, the performance of these classifiers is highly dependent on the behavior itself; showing better results in highly distinguishable behaviors such as standing vs sitting relative to more similar behaviors such as walking vs running. New tools that address interactive behaviors are emerging (Hu et al., 2023), but it is unsure how robust behavior classification can be in in-the-wild data. Furthermore, ideally, computational approaches allow for some way to inform the researcher what low-level features contained in animal poses contribute to the classification of some behavior. Deep learning approaches in classification often complexify the explainability of what features contribute to classifications.

Current study

In this study, we investigate DLC-driven pose tracking combined with machine classification on social grooming data in wild central chimpanzees (*Pan troglodytes* as an alternative to manual annotation. Building on Wiltshire et al. (2023)'s insights, we examine the robustness of DLC pose estimation. We then take this analysis one step further, by investigating how pose estimates can be used to automatically detect the highly social and intricate behavior that is grooming, without compromising interpretability and explainability. We do so by using support vector machines.[1]

[1.] For a more elaborate explanation about the methodology used in this paper, please refer https://github.com/yanavdsande/chimps_grooming

Method

Operationalization

We operationalized grooming using the following definition: we must see one hand of the actor (groomer) touching a body part of the receiver (groomee) and the groomer's gaze must be directed towards the body of the groomee.

A grooming session was defined as a period of time during which two individuals were involved in a grooming interaction, without any change of behavior, and ended when both individuals stopped grooming for more than 30 seconds (following: Fedurek & Dunbar 2009; Newton-Fisher & Lee, 2011; Kaburu & Newton-Fisher, 2013).

A grooming bout was defined as a period of continuous grooming in a given direction within a grooming session. From observations, one can distinguish *one-directional grooming*, and *mutual grooming*. One-directional grooming is defined as a grooming bout where the groomed individual does not reciprocate grooming within the same session. In mutual grooming situations, both individuals groom each other at the same time.

Data

Data Collection The data used in the current study includes *N*=10 chimpanzees from the Rekambo community living in the rainforest of Loango National Park, Gabon, Africa. Their territory is part of the Loango National Park, which includes various habitat types. Over 301 days a total of 2107 hours of focal behavioral data were collected for a study by Southern et al. (in press). Individuals were followed for as long as possible from first encounter until nesting and focal individual adult males were under observation for a mean ± SD duration of 210.1 ± 74.5 h. All behavioral data were entered and recorded through an ethogram created using Cybertracker software (Cybertracker version 3.51).

Video Selection The videos were recorded in 4K quality using a SONY AX53 in a maximum of 5-minute increments resulting in videos with 25.4 fps. At the beginning of our data exploration phase, we had access to a total of 164 videos with a total length of 5 hours 42 minutes.

We assessed the quality of the video footage based on the following factors: visibility of the chimpanzees (operationalized by amount of natural occlusion e.g. occlusion through surroundings such as leaves, trees, or environmental factors, or occlusion through body parts), the contrast of the footage, the stability of the video after processing through a stabilizer. If apes were not clearly distinguishable in their environment, the videos were excluded. Our final dataset consisted of 75 videos (duration: 3 hours 4 minutes 3 seconds).

Frame selection We used the K-means algorithm implemented in DLC to extract frames from the video for the analysis. We extracted 20 frames from each video (corresponding to the best members of 20 clusters). The number of clusters was independent of the length of a video.

13 frames, in which the pose could not be clearly determined due to camera movement, were manually excluded, bringing our data set to 1487 frames. We found the distributions of the scene features for the selected frames to be similar to the distributions of the scene features of the videos.

For each picked frame, we added a quality control measure named occlusion to control for the visibility of body parts. The occlusion ratio was based on four criteria: occlusion through the environment, occlusion due to a body, occlusion caused by both environment and body, and no occlusion. Each of these categories received a score based on how many body parts could be tracked based on annotations.

Landmark selection The dataset after frame selection consisted of images with representative 2D poses of the apes exhibiting grooming behavior. Kinematic information was extracted through 16 body parts (markers; see Figure 1) and their relative two-dimensional positions in the frame (in pixel coordinates). The marker selection was based on previous research (Desai et al., 2023) and further discussed with the primatologists of the comparative biology group at Osnabrück University.

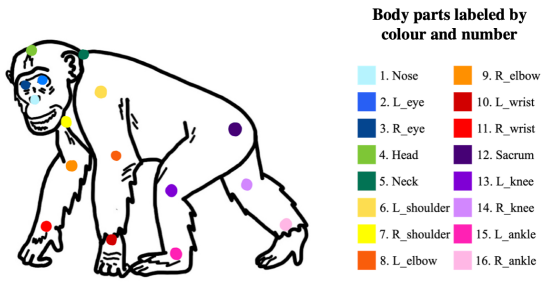

Figure 1. A schematic depiction of all landmarks annotated on the body of a chimpanzee. Each of these landmarks is marked, whenever visible, on each individual within the frame.

Manual annotation of the frames All frames were hand-annotated by the researcher using Napari (Chiu and Clack, 2022). Occluded yet inferable body parts were also annotated, given that if humans can infer the location of a certain body part in a still image, there is contextual information available which can be used by the network to infer the body part as well. We used the names of the chimpanzees to identify the individuals. Females and infants were marked as spectators.

Automatic annotation of frames We trained a DLC model with RESNET101 architecture on the manually annotated frames using 95% of the data for the train set and 5% for the validation set. Training was completed by using Google Colab. The GPUs used varied over time, based on their availability, but the training was mainly conducted on a T4 with 15gb RAM. Graphic feedback was provided by Tensorboard and plotted from the log information output during training. Mathis et al. (2018) and Lauer et al. (2022) recommend terminating training when the loss plateaus; we found that the loss plateaus at +/- 20.000 iterations for all models. This is less than the advised number of iterations that Lauer et al. recommend for multi-animal training (between 50k and 100k). However, due to our limited resources and the danger of overfitting, we decided to investigate the model performance after training for 20.000 iterations.

We evaluated our models by comparing our results with Wiltshire et al. (2023) and through inter coder reliability between two human annotators (see Figure 2). Due to uneven rankings caused by the missing head marker in Wiltshire et al.'s data, we were only able to calculate the Spearman's rank correlation coefficient between our model and the intercoder reliability, which was significant ($r = 0.92$, $p < 0.05$). For illustrative purposes, we included Wiltshire et al.' s rankings in Figure 2.

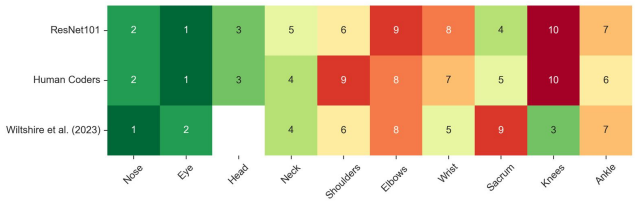

Figure 2. The rank of the body part relative to the other body parts based on the mean RMSE is printed in the cell. Labelability is printed on a scale from 1(green) to 10 (red), where 1 is best labelable and 10 worst labelable. Labelability of the current models (Wiltshire et al. vs. RESNET101 trained by the authors) indicate how well they correspond to a human annotator (ground truth), and for human coders how well they correspond between each other. The labelability shows that we get comparable specificities to body parts in our currently trained model relative to Wilshire's model.

Creating hand-crafted features for grooming detection Features were crafted to express properties of the data that we hypothesized predict grooming behavior. We first used the markers to create 2D vectors that describe the bones of the body parts. For example: the left wrist coordinates and left elbow coordinates constructed the lower left arm vector. An additional eleven vectors were created (see Figure 3, blue vectors): both lower arms (wrist to elbow), both upper arms (elbow to shoulder), shoulders (shoulder to shoulder), neck (head to neck), spine (neck to sacrum), both upper legs (sacrum to knee), and both lower legs (knee to ankle). After constructing skeleton vectors, we constructed features that can be grouped in three categories: within features, between features, and flexion/extension features. Within features include the skeleton vectors, body markers and their self-embeddings, e.g. the position of each body part relative to every other body part expressed as the normalized Euclidean distance between the two (see Equation 1 and Figure 3, red vectors), or the distance between wrist and shoulder (a proxy for limb extension/flexion, see Figure 3, yellow vector).

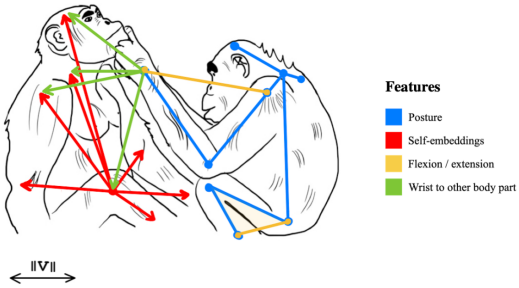

Figure 3. A Sketch of Different Constructed Features During a Grooming Interaction. The different types of features that capture the kinematic data in a grooming interaction.

Equation 1: Calculation of normalized Euclidean distance between two bodyparts

$$\Delta bodypart_A, bodypart_B = \frac{\sqrt{(x_{bodypart_A} - x_{bodypart_b})^2 + (y_{bodypart_A} - y_{bodypart_b})^2}}{Norm}$$

$$Norm = \sqrt{(X_{sacrum} - X_{neck})^2 + (Y_{sacrum} - Y_{neck})^2}$$

Between features include the distance of the wrist of one ape participating in a grooming bout and the body parts of the other ape (see Figure 3, green vectors). This distance was calculated using the pseudocode in Code Block 1. All distances were normalized by the Euclidean distance between the sacrum and the neck, such that distances are comparable between apes of different sizes.

Code block 1: Pseudo code minimum distance point E to line segment AB

```
Input: point A (x,y), point B (x,y), Point E (x,y)

        1. calculate vector AB , BE, and AE
        2. calculate the dot product of AB and BE
        (AB_BE)
        3. calculate the dot product AB and AE
        (AB_AE)

 4. calculate the minimum distance:

# if AB_BE > 0:
        # output: euclidean distance between B and E
# elif AB_AE < 0:
        # output: euclidean distance between A and E
# else:
        # output: |EF| = |(AB X AE)/|AB||
(perpendicular distance)
```

Constructing binary labels for the data All frames in our data set are paired with a variety of labels that describe the grooming situation. We labeled a total of 464 pictures as Right_grooms_Left (40%), 354 pictures as Left_grooms_Right (30%), 72 images as Mutual_Grooming (6%) and 271 No_Grooming (23%).

Support Vector Machine (SVM)

Training Two support vector machines (SVM) were trained on the manually annotated frames and their grooming labels. One SVM used a linear kernel and one used a non-linear third-order polynomial kernel. We used an 80/20 split for train and test set. Missing data points in the features were approximated using a mean strategy imputer. We weighed errors to compensate for class imbalance. Because we corrected for class imbalance, chance level performance is 25%.

Results For all classes we find good classification performance (see Table 1a and Table 1b). Interestingly, the most difficult class to classify is mutual grooming, which is classified as right grooming left in 14.3% of the cases or as left grooming right in 21.4% of the cases, but the grooming behavior is always recognized. Similar results were found in the linear kernel case (see Table 1a and Table 1b).

Table 1a: Confusion Matrix multiclass SVM polynomial kernel and linear kernel

True class		Predicted class				
		L_Grooms_R	Mutual	No_Grooming	R_Grooms_L	total
L_Grooms_R	Polynomial	63 (86.3%)	2 (2.7%)	6 (8.2%)	2 (2.7%)	73
	Linear	61 (83.6 %)	2 (2.7%)	4 (5.5 %)	6 (8.2 %)	
Mutual	Polynomial	3 (21.4%)	9 (64.2%)	0 (0.0%)	2 (14.3%)	14
	Linear	4 (28.6 %)	8 (57.1%)	1 (7.1 %)	1 (7.1 %)	
No_Grooming	Polynomial	4 (7.8%)	1 (2.0%)	43 (84.3%)	3 (5.9 %)	51
	Linear	6 (11.8 %)	1 (2.0%)	32 (62.7 %)	12 (23.5 %)	
R_Grooms_L	Polynomial	2 (2.1 %)	1 (1.1%)	16 (17.0%)	75 (79.8 %)	94
	Linear	3 (3.2 %)	0 (0.0 %)	7 (7.5 %)	84 (89.4 %)	
Total		72	13	65	82	232

Note. Accuracy polynomial kernel: 81.9%, accuracy linear kernel: 79.7%

Table 1b: Statistics Multiclass SVM polynomial kernel and linear kernel

		L_Grooms_R	Mutual	No_Grooming	R_Grooms_L	Macro Avg.	Weighted avg.
Precision	Polynomial	0.875	0.69	0.662	0.915	0.786	0.833
	Linear	0.824	0.727	0.727	0.816	0.774	0.79
Recall	Polynomial	0.863	0.64	0.843	0.798	0.787	0.819
	Linear	0.836	0.571	0.627	0.893	0.732	0.797
F1-score	Polynomial	0.869	0.66	0.741	0.852	0.782	0.822
	Linear	0.830	0.640	0.674	0.853	0.749	0.793

Behavior classification - Automatic labeling using DLC We presented the trained model with a randomly selected video that was not included in the train dataset, the test dataset of the DLC model, nor the train/test set of the SVM. In this video, two males are grooming each other. All detections of body parts created by DLC were stored and filtered on a likelihood of > 0.8.

Automated Behavior classification

Performance evaluation To validate our model, a comparison was made to a hand labeled video of the

grooming directionality in the video. Manual labeling was done by the time duration of the grooming bout expressed in seconds. We multiplied the amount of seconds with the frame rate to project the directionality label on frame level. The accuracy and precision were computed to assess the performance on unseen, video data.

Video creation We used PILLOW to draw the predicted class on each frame and used FFMPEG to create a video with the created labels (Figure 3). Additionally, the moment where the class changes is now flagged as a change in grooming context. A csv file with timestamps is created including all labels, the flag for a change in grooming context (also known as a bout), and more specifically where a turn change occurs (a change in grooming directionality that involves a switch of groomer / groomee identity). Turn was operationalized as a change in grooming directionality lasting longer than a second (25 frames) and the video to keep the human in the loop.

Results

After comparing the class labels predicted by our SVM with the ground truth class labels, we find an accuracy of 0.863, meaning that we were able to correctly predict the grooming directionally in 86% of the video (Table 2a/b). All endings of the grooming bouts are correctly recognized and no false positives or false negatives for turn changes were found. Although sometimes a different label was predicted, it never lasted a consecutive number of frames to count as a turn change (Figure 4).

Table 2a: Confusion Matrix polynomial Multiclass SVM

True class	Predicted class				
	L_Grooms_R	Mutual	No_Grooming	R_Grooms_L	total
L_Grooms_R	0	0	0	0	0
Mutual	0	148	0	14	162
No_Grooming	0	0	0	0	0
R_Grooms_L	0	25	0	98	123
Total	0	173	0	112	285

Note. Accuracy: 86.3%

Table 2b: Statistics polynomial Multiclass SVM

	L_Grooms_R	Mutual	No_Grooming	R_Grooms_L	Macro Avg.	Weighted avg.
Precision	-	0.855	-	0.875	0.865	0.864
Recall	-	0.914	-	0.797	0.856	0.863
F1-score	-	0.692	-	0.589	0.641	0.648

Our results show that, in 86% of the frames, grooming behavior was correctly recognized, and it was able to predict the occurrence of a turn. Looking closer at the data, we see that the closer to a turn the more wrongly predicted frames are present (Figure 4). The saturation of wrongly predicted frames shows a shift in turn-relevant kinematic information.

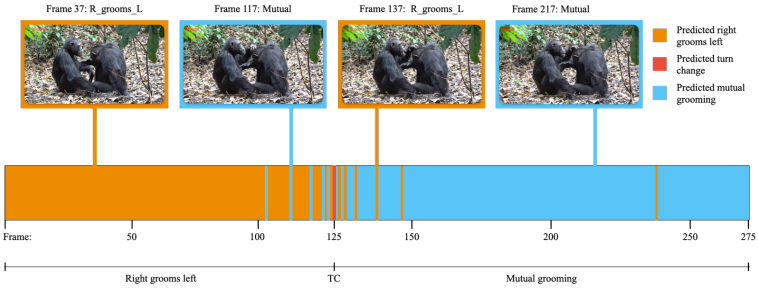

Figure 4. A representation of machine labeled frames over time. This is a schematic depiction of the labeled video. Predicted grooming directionality is depicted in the colored bar. Below there is a timeline of the ground truth label, based on the labeling by a primatologist. The time series is expressed in consecutive frames.

Discussion

We investigated the feasibility of using pose estimation through DLC to automatically label grooming interactions between chimpanzees in a natural setting. Using self- and other-pose embeddings as an input for an SVM classifier, our approach was able to recognize behaviors with a high granularity (80%), classifying not only grooming presence but also directionality (who is grooming who). Consequently, our techniques can be extended to various *directional* social behaviors in chimpanzees (and other species) that are characterized by distinct pose structures, such as greeting interactions, mother-infant interactions, food sharing, and displays of aggression. Additionally, given that we used raw recordings to extract the kinematic data, our approach is not limited to the anatomical features of chimpanzees and can be extended to other species. For example, human social touch could be automatically detected using our approach, or our approach could be leveraged for manual gesture detection (e.g., Ripperda et al., 2020). Furthermore, our framework requires limited pre-processing of video attributes, making it useful across different data collection protocols. In sum, this pipeline opens the door for further research on a range of different social interactions and across various other species living in natural environments.

The success of our pipeline is potentially of huge significance to primatologists. As previously mentioned, it is often hard to consistently annotate datasets due to large amounts of raw video footage. By automatically flagging the presence of a certain behavior in video data, our pipeline provides a pre-screening mechanism. Thus, even if researchers prefer manual annotation, rather than machine classification, our method will still allow for a reduction of search time for relevant behaviors. In the future, one idea is to build in a confidence measure where classifications falling below the confidence threshold trigger a GUI for the researcher to label the datapoint manually. This manual labeling would present an opportunity to train the model on

challenging or underrepresented data points, enhancing performance through 'human in the loop' training.

One of the advantages of a linear kernel SVM is the possibility to explore the coefficient space. Informal explorations revealed that the features our classifier uses to make a distinction between different directionalities seem to correspond with some observations and theory on grooming signals. For example, when the ape is positioned squarely in front the camera, the SVM uses the distance between left wrist and right wrist of that ape to determine the groomer identity. The relevant hypothesis here being that the greater the distance from left wrist to right wrist, the more the body is opened and therefore the more body surface is presented to the groomer (Fedurek et al., 2015), implicating that said ape is the groomee. Unfortunately, due to the lack of 3D information in our data it was not possible to test this hypothesis. It did however show that analyzing the coefficients of the linear classifier is potentially informative, which provides some value over and above black box deep learning based classifiers. The subtleties in pose captured from these extractions may in turn further inform human researchers, providing key insights into motion analysis of body signaling in social behavior.

Limitations

There are several limitations to our study. By limiting ourselves to 16 markers, we had the benefit of avoiding increased network complexity and a more laborious manual annotation process. Additionally, using the landmarks by Desai et al. (2023), we created comparable data between these two studies. However, including body parts relevant to the behavior under investigation, such as the fingers, lips and lower back in case of grooming, could have potentially conveyed more information about communicative cues, making the results more accurate. In the future it would be interesting to combine existing datasets including more annotated body parts to assess performance differences of our pipeline.

Another limitation was that the pipeline was tested on a variety of images, but only tested on one example video. This particular video was not present in the training set or corresponding test set but was part of the data collection done by the same researcher, at the same location, with the same chimpanzees. Therefore, it is difficult to comment on the ecological validity of our approach at this juncture of the project; however, in theory, our method could be used for data collected using different data collection protocols. To test this statement, the next planned step is to test the performance of our pipeline on videos recorded at different field sites.

Notes on using machine learning in primatology

Increased generalizability of models leads to broader applicability across various types of behavioral data. This heightened applicability in turn contributes to the sustainability of research. A model only needs to be trained once and can be subsequently used as a pre-trained model; these pre-trained models are easily fine-tuned to answer other research questions. Using pre-trained models is computationally cheaper, more (environmentally) sustainable and requires less data than creating and training a model from scratch. It also contributes to the accessibility of machine learning techniques, since research facilities with fewer resources can also access pre-trained models. A similar strategy was employed in the DLC's model zoo (Kane et al., 2020). Taking the sustainability implications into account is of huge importance, especially in a field such as primatology where many species are classified as threatened.

The introduction of DLC has revolutionized the way we can conduct behavioral cross-species research. Our results show the continuing promise of machine learning techniques for the field of primatology. Yet there still lies a task for the community to improve the accessibility of these methods. In its current state adapting the software to different types of data requires programming and hardware knowledge. Our recommendations are to centralize documentation and make more extensive use of version control; recent advancements in the development of DLC show how collaboration and communication between fields and researchers increase the usability of the software.

Conclusion

This study contributed to the research of understanding social behavior in chimpanzees and how they coordinate their actions, which can be extended to human social interactions. Specifically, we make large scale data analysis possible on grooming events in the wild, in a non-vocal modality and in a social context where no current work has been published on the topic. Our findings allow for further exploration of how grooming events are coordinated, what social rules govern this behavior, and how interactive primitives can help form social networks. Additionally, our utilization of machine learning techniques underscores the potential of technology in studying wild primate behavior and automating the recognition of specific behaviors. Beyond the contribution to the field of primatology, this study highlights the growing pains of these advancements. Notably, the tracking and classification task applied to the current dataset presented significant technical challenges (e.g. footage recorded in natural surroundings, with multiple individual tracking) and the process therefore shed light on necessary improvements needed in software to fully harness computational methods in primatological research. We do however conclude this work by acknowledging these machine learning techniques hold promise for automating labor-intensive processing, opening up exciting new research avenues within this field.

References

Bain, M., Nagrani, A., Schofield, D., Berdugo, S., Bessa, J., Owen, J., ... & Zisserman, A. (2021). Automated audiovisual behavior recognition in wild primates. *Science advances*, *7*(46). 10.1126/sciadv.abi4883

Chiu, C. L., & Clack, N. (2022). napari: a Python Multi Dimensional Image Viewer Platform for the

Research Community. *Microscopy and Microanalysis*, *28*(S1), 1576-1577.
Desai, N., Bala, P., Richardson, R., Raper, J., Zimmermann, J., & Hayden, B. (2023). OpenApePose, a database of annotated ape photographs for pose estimation. *Elife*, *12*, RP86873.
Dunbar, R. I. M. (1988*). Primate social systems*. Springer Science & Business Media.
Fedurek, P., Dunbar, R. I., & British Academy Centenary Research Project. (2009). What does mutual grooming tell us about why chimpanzees groom?. *Ethology, 115*(6), 566-575.
Fedurek, P., Slocombe, K. E., Hartel, J. A., & Zuberbühler, K. (2015). Chimpanzee lip-smacking facilitates cooperative behaviour. *Scientific reports, 5*(1), 13460.
Goodall, J. (1986). *The chimpanzees of Gombe: Patterns of behavior.* Belknap Press of Harvard Univ. Press.
Grueter, C. C., Matsuda, I., Zhang, P., & Zinner, D. (2012). Multilevel societies in primates and other mammals: introduction to the special issue. *International Journal of Primatology, 33*, 993-1001.
Hu, Y., Ferrario, C. R., Maitland, A. D., Ionides, R. B., Ghimire, A., Watson, B., ... & Ye, B. (2023). LabGym: Quantification of user-defined animal behaviors using learning-based holistic assessment. *Cell Reports Methods*, 3(3).
Kaburu, S. S., & Newton-Fisher, N. E. (2013). Social instability raises the stakes during social grooming among wild male chimpanzees. *Animal Behaviour, 86*(3), 519-527.
Kane, G. A., Lopes, G., Saunders, J. L., Mathis, A., & Mathis, W. M. (2020). Real-time, low-latency closed-loop feedback using markerless posture tracking. *eLife*, 9, e61909. https://doi.org/10.7554/eLife.61909
Lauer, J., Zhou, M., Ye, S., Menegas, W., Schneider, S., Nath, T., ... & Mathis, A. (2022). Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nature Methods*, *19*(4), 496-504.
Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. Frontiers in psychology, 6, 731.
Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience, 21*(9), 1281-1289. https://doi.org/10.1038/s41593-018-0209-y
Mielke A., Samuni L., Preis A., Gogarten J. F., Crockford C. & Wittig R. M. (2017). Bystanders intervene to impede grooming in Western chimpanzees and sooty mangabeys. *R. Soc*. open sci.4171296171296 http://doi.org/10.1098/rsos.171296
Newton-Fisher, N. E., & Lee, P. C. (2011). Grooming reciprocity in wild male chimpanzees. *Animal Behaviour, 81*(2), 439-446.
Pika, S. (2014). Chimpanzee grooming gestures and sounds: What might they tell us about how language evolved?. In *The social origins of language: early society, communication and polymodality* (pp. 129-140). Oxford University Press.
Pika, S., Wilkinson, R., Kendrick, K. H., & Vernes, S. C. (2018). Taking turns: bridging the gap between human and animal communication. *Proceedings of the Royal Society B*, *285*(1880), 20180598.
Radhakrishna, S., & Jamieson, D. (2018). Liberating primatology. *Journal of Biosciences*, *43*(1), 3-8
Ripperda, J., Drijvers, L., & Holler, J. (2020). Speeding up the detection of non-iconic and iconic gestures (SPUDNIG): A toolkit for the automatic detection of hand movements and gestures in video data. *Behavior research methods, 52*(4), 1783-1794.
Sakib, F., & Burghardt, T. (2020). Visual recognition of great ape behaviours in the wild. *arXiv preprint arXiv:2011.10759*. https://doi.org/10.48550/arXiv.2011.10759
Schofield, D., Nagrani, A., Zisserman, A., Hayashi, M., Matsuzawa, T., Biro, D., & Carvalho, S. (2019). Chimpanzee face recognition from videos in the wild using deep learning. *Science advances, 5*(9), eaaw0736
Seyfarth, R. M., & Cheney, D. L. (2012). The evolutionary origins of friendship. *Annual review of psychology 63*: 153-177. 10.1146/annurev-psych-120710-100337
Strier, K. B. (2003). Primatology comes of age: 2002 AAPA luncheon address. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, *122*(S37), 2-13. 10.1002/ajpa.10383
Wiltshire, C., Lewis-Cheetham, J., Komedová, V., Matsuzawa, T., Graham, K. E., & Hobaiter, C. (2023). DeepWild: Application of the pose estimation tool DeepLabCut for behaviour tracking in wild chimpanzees and bonobos. *Journal of Animal Ecology.*
Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., Girshick., R. (2019). Detectron2. *Retrieved from https://github.com/facebookresearch/detectron2*

Acknowledgements

We would like to thank the Agence Nationale des Parcs Nationaux (ANPN) and the Centre National de la Recherche Scientifique et Technique (CENAREST) of Gabon for their permission to conduct research in Loango National Park. Furthermore, we are thankful to all members of the Ozouga Chimpanzee Project (past and present) for their support and dedication to the Rekambo community. A special thanks goes to the directors of the Ozouga Chimpanzee Project, Tobias Deschner and Simone Pika for enabling the collection of this data. This research is part of a project funded by an EU-Consolidator grant (772000, TurnTaking) rewarded to Simone Pika from the European Research Council (ERC) under the Horizon 2020 research and innovation program. Wim Pouw is funded by a NOW VENI grant (VI.Veni 0.201G.047: PI Wim Pouw).