AUTOMATED CLASSIFICATION OF SOCIAL BEHAVIORS OF PRIMATES IN NATURAL ENVIRONMENTS

AN EXPLORATION OF DISTINCT DEEP LEARNING ARCHITECTURES AND MODALITIES

LAURA J. HAGEDORN

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE

DEPARTMENT OF
COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

STUDENT NUMBER

2024795

COMMITTEE

dr. Görkem Saygili
dr. Sharon Ong

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 19, 2024

WORD COUNT

8756

ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisor, Dr. Görkem Saygili, and all the employees at the Biomedical Primate Research Center and Utrecht University for their support and guidance.
And, as always, in loving memory of my father and mentor, Georg Hagedorn, who once introduced me to the field of computer science, and would have loved to see me graduate from it.

AUTOMATED CLASSIFICATION OF SOCIAL BEHAVIORS OF PRIMATES IN NATURAL ENVIRONMENTS

AN EXPLORATION OF DISTINCT DEEP LEARNING ARCHITECTURES AND MODALITIES

LAURA J. HAGEDORN

Abstract

This research explores three deep learning models and modalities for automatic pose estimation and social action recognition of long-tailed macaques (*Macaca fascicularis*) at the Biomedical Primate Research Center (BPRC) in Rijswijk, The Netherlands. RGB data was collected from a dual calibrated camera system located in a semi-natural enclosure over one month, and a pose estimation algorithm using YOLOv8 was trained to generate 2D and 3D skeleton data. The findings revealed that a Spatial-Temporal Graph Convolutional Network (ST-GCN) based on 2D skeleton data exhibited the highest performance with an accuracy of 75%. The study demonstrates the feasibility of detecting naturally occurring, "between pairs" level actions in semi-natural enclosures between two primates, without necessitating specialized equipment or expensive cameras.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

The video data used in this research was provided by the Biomedical Primate Research Center (BPRC) in Rijswijk. The videos were acquired through two cameras located in a semi-natural enclosure at the BPRC. The cameras were installed in February and constantly recording in February and March without further interference. All images and figures are created by the author, except for figure 8 which displays an illustration of the triangulation process and was taken from Ekberg, Daemi, and Mattsson (2017), which is clearly cited in the thesis. Several software packages have been used for this research. Pose estimation was achieved by utilizing the deep learning framework for object detection proposed by You Only Look Once version 8 (YOLOv8) (Jocher, Chaurasia, & Qiu, 2023). Camera calibration and triangulation relies on OpenCV (Bradski, 2000). For action recognition, the OpenMMLab's (M. Contributors, 2020) action recognition framework MMACTION2 (Contributors, 2020) was heavily modified and adapted. The reused/adapted code fragments are clearly indicated. The code is entirely written in Python 3.12 and available at `https://github.com/laurahgdrn/NHP-AR`. In terms of writing, the author used assistance with the language of the paper. A generative language model, OpenAI (2024), was used to improve the author's original content, for paraphrasing, spell checking and grammar. No other typesetting tools or services were used.

2 INTRODUCTION

Macaques are essential to the study of human psychology, (cognitive) neuroscience, disease research, and drug development because of the biological similarity and homology between macaques and humans (Gardner & Luciw, 2008; Hannibal, Bliss-Moreau, Vandeleest, McCowan, & Capitanio, 2017; Warren et al., 2020; Xue & Deng, 2023). Crucial components of these studies are behavioral analyses, which are costly and time-consuming when done manually. Besides this, behavioral measurements have historically been confined to single motor modalities, typically involving bodily constraints, which provide limited comprehension of natural behavior and induce strong ethical concerns about the animal's welfare. Similarly, other experiments that involved tracking or identifying animals required markers attached to body parts - ranging from superficial tags or collars to implanted chips - to track positions (Foster et al., 2014; Gilja, Chestek, Nuyujukian, Foster, & Shenoy, 2010; Vargas-Irwin et al., 2010) which induces ethical concerns as they can irritate and even harm the animals. To address this challenge, there is a growing interest in the application of automatic pose estimation and behavior classification, because of their potential to boost these analyses while also allowing for a more comprehensive understanding of natural behavior (Knaebe, Weiss, Zimmermann, & Hayden, 2022). By automating the process of monitoring macaques in their naturalistic settings, researchers can maintain high standards of animal welfare while conducting meaningful observational research (Bala et al., 2020; Hannibal et al., 2017; Voloh et al., 2023; Xue & Deng, 2023) without constant human presence. This is crucial as it minimizes interference and allows animals to behave uninhibitedly, providing a more authentic depiction of their (inter)actions. Moreover, it can also enable the identification of behavioral patterns indicative of underlying psychological states, such as anxiety or depression (Hayden, Park, & Zimmermann, 2022) as well as other injuries or illnesses as it has been shown that macaques mask signs of injury or illness in the presence of humans (Gaither et al., 2014).

Therefore, improved understanding of macaque behavior aids in the ethical treatment and care of these animals in research settings. By reducing the need for invasive methods and enhancing data accuracy, automated systems contribute to the refinement of experimental practices, aligning with ethical guidelines and public expectations for animal welfare (Knaebe et al., 2022). Furthermore, insights gained from macaque studies can inform broader societal discussions on animal cognition, behavior, and welfare (Xue & Deng, 2023).

2.1 *Social Behavior*

The social environment is an important predictor of health and mortality risk in social mammals (Simons, Michopoulos, Wilson, Barreiro, & Tung, 2022). The costs and benefits of group living are not distributed evenly among group members and can significantly impact an individual's overall fitness (Schülke et al., 2022). Disparities arise based on factors such as relative spatial position within the group, dominance rank, and the frequency and diversity of friendly physical contact with other members (Aguilar-Melo, Calme, Pinacho-Guendulain, Smith-Aguilar, & Ramos-Fernández, 2020; Schülke et al., 2022). The advantages individuals gain from spatial cohesion include heightened protection from predators (LaBarge, Allan, Berman, Margulis, & Hill, 2020; Sirot & Touzalin, 2009), improved access to social information about potential risks or resources (LaBarge et al., 2020) as well as cooperative defense mechanisms, improved vigilance, collaborative defense of resources, and more efficient foraging (Schülke et al., 2022). Finally, the social environment can predict molecular, physiological and life-history outcomes (Simons et al., 2022). In other words, social animals with more and stronger social relationships live longer, healthier lives (Simons et al., 2022; Testard, Tremblay, & Platt, 2021). There are several social interactions animals can engage in to strengthen their bond and social status, with the two predominant ones being grooming (Kaburu et al., 2019; Simons et al., 2022) and playing (Shimada & Sueur, 2018; Solanki, Lalremruati, Lalchhuanawma, et al., 2020a). Automatic action recognition of grooming and playing behaviors in macaques can revolutionize the understanding of their social structures and dynamics, and it also allows researchers to collect and analyze vast amounts of data with greater efficiency and accuracy than traditional methods (Knaebe et al., 2022; Pereira et al., 2019). To be precise, this approach allows for continuous monitoring and real-time analysis, providing a more comprehensive picture of social interactions within macaque groups. Automatic action recognition can help identify subtle patterns and changes in social behavior that may indicate health issues, stress, or shifts in group hierarchy. For instance, a decrease in grooming frequency might signal social isolation or health problems, while an increase in play behavior could imply a stable and cohesive social environment (Hayden et al., 2022). By detecting these changes early, interventions can be implemented to improve animal welfare and manage social dynamics accurately.

2.1.1 *Grooming*

As mentioned above, primates live in cohesive social groups, in which they maintain stable relationships through affiliative behaviors such as grooming (Jablonski, 2021). Grooming involves manipulating the body surface, which includes all forms of care and attention. In social primates, grooming is the primary currency of social affiliation (Simons et al., 2022; Solanki et al., 2020a). For instance, several different primate species, including macaques and larger apes, spent up to 20% of their day grooming (Solanki, Lalremruati, Lalchhuanawma, et al., 2020b). In addition to establishing and maintaining affiliative relationships, grooming is also believed to reduce tension and aggression between individuals (Simons et al., 2022; Solanki et al., 2020b), and promotes well-being by directly eliminating ectoparasites such as lice, fleas, and ticks. Individuals lacking grooming interactions exhibit heightened anxiety levels and reduced fertility compared to individuals with regular social interactions (Jablonski, 2021). From birth to adulthood, grooming plays a pivotal role in building and sustaining trust-based relationships, which are crucial for individual well-being and reproductive success. Grooming facilitates the formation, maintenance as well as reconciliation of social ties and therefore promotes emotional stability among individuals and encourages group cohesion (Jablonski, 2021). Moreover, grooming behavior is highly influenced by the social rank of an individual (Simons et al., 2022): higher-ranking individuals generally engage in less grooming and are more often groomed by subordinates (Seyfarth, 1977). Additionally, Solanki et al. (2020a) stated that grooming behavior varies across gender. They found that in long-tailed macaques, female-female pairs tend to focus their grooming on the face and frontal areas, whereas male-male pairs prefer grooming the back and tail. Also, grooming occurs more frequently as the animals age, often gradually replacing playing behavior.

2.1.2 *Playing*

The term 'play' is not strictly defined, but it usually has the following characteristics: the performance of the behavior is not functional, but it is spontaneous, voluntary, intentional and pleasurable (Kellman & Radwan, 2022). Similar to grooming, playing strengthens social bonds and group stability. In many social animal species, playing is thought to fulfill various social functions, including the acquisition of social knowledge and skills, as well as the reinforcement of social bonds and cognitive developments (Beltran Frances et al., 2020; Wright et al., 2018). To be more precise, social play intertwines cooperation, communication, and reciprocal actions among individuals. Additionally, play incorporates behavioral elements from various social situations, including conflict, mating, and hunting, blurring the distinction between play and non-play behaviors (Wright et al., 2018). The type and display of play vary by species, but generally, play activity increases during juvenile hood but declines at sexual maturity (Mayhew, Funkhouser, & Wright, 2020). According to Mayhew et al. (2020), play is also more often observed between young male juveniles, likely to test their increasing physical strength in competition with other individuals.

3 RELATED WORK

3.1 *Pose Estimation*

Pose estimation, a well-explored subfield within computer vision, is a process that aims at locating different body parts to obtain a representation. A majority of this research is focused on human pose estimation (HPE) with applications in human-computer interaction, motion analysis including action recognition and prediction, healthcare and augmented and virtual reality (Munea et al., 2020; C. Wang, Zhang, & Ge, 2021; Zheng et al., 2023).

Pose estimation from 2-dimensional input like images or videos has been explored extensively in the last decade, starting off with traditional methods that required hand-crafted feature extraction techniques for different body parts (Zheng et al., 2023), utilizing methods like pictorial structures and graphical models. These approaches, though effective in constrained environments, faced challenges with variability in poses and complex backgrounds (Toshev & Szegedy, 2014). With the increasing availability of computational resources, deep learning approaches have outperformed many of the traditional computer vision methods and are currently considered the state-of-the-art designs for pose estimation (Zheng et al., 2023). In general, pose estimation involves two main steps: 1) the localization of body joints/key points, and 2) creating a valid pose configuration from the detected key points (Munea et al., 2020). The localization of the key points is mainly achieved through supervised learning methods, where models are given a large set of images of, e.g., humans in different situations and positions, along with their manually annotated key points.

Human pose estimation can be divided into single-person or multi-person pose estimation, with the latter being much more challenging than the former. Multi-person pose estimation requires a deep understanding of the logical arrangement of key points, enabling the distinction of key points belonging to distinct individuals (Kocabas, Karagoz, & Akbas, 2018). While human pose estimation has been largely explored (He, Zhang, Ren, & Sun, 2016; Kocabas et al., 2018; Toshev & Szegedy, 2014), only a small group of researchers has investigated pose estimation for non-human primates (NHP). The detection and pose estimation of NHPs can be more challenging than of humans as frequent interactions cause occlusions and complicate the association of detected key points to the correct individuals, as well as having highly similar-looking animals that interact closely. The detection and pose estimation of macaques pose additional challenges due to their limbs having high degrees of freedom. One of the most famous open-source frameworks for animal pose estimation is DeepLabCut (Mathis et al., 2018) which applies transfer learning with deep neural networks and has been shown to enable accurate pose estimations of mice (Mathis et al., 2018), flies, cheetahs, horses, and fish (Nath et al., 2019). Most recently, Lauer et al. (2022)) extended the DeepLabCut framework by introducing multi-animal pose estimation. Bala et al. (2020) introduced OpenMonkeyStudio for 3D pose estimation of primates. It is a marker-less motion capture system for long-tailed macaques. The implementation incorporates pose estimation using a deep neural network and utilizes 62 cameras, generating multi-view image streams that significantly enhance annotated data through 3D multi-view geometry. However, the training data utilized to train the model was collected from a tiny cage with almost no distracting objects. The model's applicability in a natural and macaque-friendly environment is questionable.

3.2 *Action Recognition*

Action recognition is a challenging task in the field of computer vision. In general, action recognition refers to automatically detecting human behaviors and gestures. Action recognition can be roughly divided into four groups: atomic level, between human and object, between pairs, and within groups (Morshed, Sultana, Alam, & Lee, 2023), with the atomic level being the easiest and within groups the most challenging act to decipher. Similar to automatic pose estimation, action recognition has received a growing attention in the last decade (Morshed et al., 2023). This trend can be attributed to its diverse applications across various domains such as human-computer interaction (HCI) (Gammulle et al., 2023), criminal settings and surveillance systems (Peng, Shi, Varanka, & Zhao, 2021; Sujith, 2014), virtual and augmented reality (Ma et al., 2021), as well as health care (Bibbò & Vellasco, 2023), and sign language (Thangali, Nash, Sclaroff,

& Neidle, 2011; Varadaraju, 2013). In clinical settings, action recognition systems can aid in stroke rehabilitation and assessing parkinson's severity (Bibbò & Vellasco, 2023; Morshed et al., 2023).

Action recognition can be performed on RGB (Red, Green, Blue) images and videos, 2D/3D skeleton data, and other modalities (Sun et al., 2022). It is different from other computer vision tasks, such as object detection or pose estimation, because the additional temporal dimension provides crucial information. Therefore, complex models which are able to simultaneously track the spatial and temporal domain had to be developed. Recurrent Neural Networks (RNNs) are specifically designed to model sequential and temporal data which is why they have a wide range of applications across various fields such as Natural Language Processing (NLP), Time-Series Analysis and Forecasting, and Action Recognition (W. Li, Wen, Chang, Nam Lim, & Lyu, 2017; Tyagi & Abraham, 2022). However, RNNs fail at recognizing long-term dependencies, due to the vanishing and gradients (Noh, 2021) which is why RNN-based architectures like Long Short Term Memory (LSTM) models (Donahue et al., 2015; Sun et al., 2022; Yue-Hei Ng et al., 2015) or Gated Recurrent Units (GRU) (Dwibedi, Sermanet, & Tompson, 2018; Kim, Lee, & Lee, 2018) have been widely explored for HAR and have achieved high performances.

Other common models for Human Action Recognition (HAR) from RGB data are 2D Convolutional Neural Networks (2D-CNNs) (Simonyan & Zisserman, 2014) which usually consist of a two-stream framework comprising a spatial network, that takes the single frames as input, and a temporal network, that receives multi-frame-based optical flows. Consequently, the spatial stream learns appearance features while the temporal stream captures motion features. Another CNN-based approach is the so-called SlowFast Network, which has been introduced by Feichtenhofer, Fan, Malik, and He (2019)) and showed promising results. Instead of operating on two separate streams in the spatial and temporal domain, a SlowFast network processes one temporal stream, but sampled at different frame rates. In more detail, a typical SlowFast network consists of a slow pathway, which performs convolutional operations on frames with a large temporal stride, and a fast pathway, which performs convolutional operations at a high frame rate (Feichtenhofer et al., 2019). The SlowFast network is computationally *lightweight* compared to similar methods (Feichtenhofer et al., 2019) and it does not require optical flow, enabling end-to-end learning from the raw data, while still achieving state-of-the-art performance for action classification and detection in videos (C.-F. R. Chen et al., 2021; Feichtenhofer et al., 2019).

Collecting RGB data is typically straightforward and provides detailed visual information about the scene being captured, but recognizing actions from RGB data presents its own challenges due to background variations, differing viewpoints, scale discrepancies, and lighting conditions. Additionally, RGB videos are often large, resulting in substantial computational expenses when attempting to model the spatio-temporal context.

Because of this, some works have focused on minimizing data dimensionality by extracting the most relevant information from videos, such as skeleton sequences that represent the trajectories of body joints and capturing essential movements (C. Wang & Yan, 2023). Not only does skeleton data require less computational resources, it is also a more robust and compact representation of movements as it mitigates issues related to viewpoint variations, occlusions, background clutter, and lighting conditions (C. Wang & Yan, 2023). Human skeleton data can be acquired using motion sensors or by applying pose estimation algorithms (L. Wang et al., 2018; Q. Wang, Zhang, & Asghar, 2022). Because skeleton data resides in a non-Euclidean space, it poses challenges for traditional deep learning methods to fully exploit their potential (Monti et al., 2017; Peng et al., 2021). Fortunately, the emergence of Geometric Deep Learning has introduced solutions such as the Graph Convolutional Network (GCN), specifically designed to tackle action recognition tasks using skeleton data and currently provide one of the most commonly used frameworks for skeleton-based action recognition (Monti et al., 2017; Peng et al., 2021). GCN-based methods are designed to perform convolutional operations on graph data (L. Wang et al., 2018). In the context of skeleton-based HAR, the most prominent work has been devoted towards Spatial-Temporal Graph Convolutional Networks (ST-GCN) to capture motion and temporal dependencies on serialized skeleton data (Cai, Jiang, Han, Jia, & Lu, 2021; Huang et al., 2020; Peng et al., 2021; Yan, Xiong, & Lin, 2018). Conventional ST-GCNs consist of a series of individual ST-GCN blocks that apply spatial and temporal graph convolutions alternately over the skeleton graph (Yan et al., 2018).

Although 2D skeleton data mitigates some of the challenges that arise from using RGB data for action recognition, it is still sensitive to occlusions and lack of depth perception. 3D skeleton data offers significant advantages compared to 2D data. The depth information provides a more accurate spatial understanding of the (human) body, which can enhance pose estimation and action recognition accuracy (Peng et al., 2021). Additionally, 3D skeletons are less affected by changes in viewpoint, maintaining consistency across different camera angles. They also preserve spatial relationships more effectively, improving the discrimination between various actions and poses. Furthermore, 3D data better handles occlusions by estimating obscured body parts, resulting in more complete skeleton representations and higher action recognition performance (Peng et al., 2021). There are two common approaches to generate 3D skeleton data: (1) using cameras with depth sensors, such as Microsoft's Kinect series (Bilesan, Komizunai, Tsujita, & Konno, 2021; Shahroudy, Liu, Ng, & Wang, 2016), or extracting poses from multiple views and then performing triangulation (Iskakov, Burkov, Lempitsky, & Malkov, 2019; Qiu, Wang, Wang, Wang, & Zeng, 2019; Tome, Toso, Agapito, & Russell, 2018). Similar to action recognition from 2D skeleton data, ST-GCNs and other GCN-based models are the most dominant methods for 3D skeleton based action recognition (Peng et al., 2021).

As mentioned in section 2, Bala et al. (2020) introduced OpenMonkeyStudio, a system designed for automated markerless 3D pose estimation of macaques, which utilizes a multi-camera setup with 62 cameras surrounding an 8m^3 enclosure. They employed a deep learning-based architecture tailored for macaque anatomy, integrating CNNs for key point detection and a graph-based optimization for 3D pose reconstruction. Their system achieved an average detection accuracy of approximately 95% for key points when validated against manually annotated ground truth data.

C. Li et al. (2023) developed a 2D skeleton-based deep learning model for pose estimation of cynomolgus monkeys. Bone recognition was achieved utilizing a high-resolution network (HRNet) and a MaskTrack R-CNN to track the monkeys' positions. The positional information for each monkey was then extracted and fed into a HRNet to generate a heatmap, achieving a detection accuracy of 98.8%. For action recognition, C. Li et al. (2023) proposed a two-stream model based on temporal shift and split attention (TSSA) with a ResNet-50 backbone and self-attention mechanisms added to each layer. Evaluating performance with top-1-accuracy, the model achieved 98.99% in detecting semantic actions such as climbing, jumping, and moving down.

Social action recognition between pairs or groups of primates has not yet been explored. Close interactions, such as grooming and playing, result in increased proximity between individuals, complicating the recognition of their actions. Furthermore, the greater degrees of freedom in macaques' extremities pose additional challenges, as well as the scarcity of action recognition data sets Bala et al. (2020).

3.3 *Current study*

This study expands the existing literature in several ways. First, the recorded video footage comes from two stabilized cameras in a semi-natural environment where the primates live in groups of up to 15 individuals. The recorded area includes platforms at different heights, swings, ropes, food, and other toys. The animals are allowed to enter or leave the recorded area at any time, therefore the amount of animals in a frame ranges between 0 and 15. Although each of these factors poses a unique challenge to computer vision algorithms, they also allow for the detection of natural behavior while ensuring the animal's well-being. Many studies presented earlier trained their pose estimation models in highly restricted and unnatural laboratory settings (Bala et al., 2020; C. Li et al., 2023; Martini, Bognár, Vogels, & Giese, 2024; Nakamura et al., 2016). Second, prior studies have largely focused on detecting single (C. Li et al., 2023; Mathis et al., 2018; Nakamura et al., 2016) or at maximum two animals (Bala et al., 2020; Martini et al., 2024). This study aims to track and detect multiple animals simultaneously by focusing on their social interactions. Again, this approach ensures the animal's well-being and the observation of natural actions and will provide relevant insights into social behavior, revolutionizing our understanding of social structures and dynamics. It allows for the identification of subtle patterns and changes in social behavior that may be indicative of health issues, stress, or shifts in group hierarchy (Hayden et al., 2022), further ensuring the well-being of macaques and other social animals in captivity.

As mentioned in the previous section, data sets for NHP action recognition are non-existent (or not openly accessible). This study will introduce the first action recognition data set for classifying grooming and playing behaviors of macaques, consisting of three modalities: RGB data, 2D skeleton data and 3D skeleton data. RGB data will be collected from two cameras, a pose estimation algorithm will be trained to generate 2D skeletons and triangulation will be applied to create 3D skeleton data.

Finally, this research aims to answer the following research questions:

RQ1 *To what extent can deep neural networks be employed to identify grooming and playing behaviors among long-tailed macaques?*

- SubQ1 *How does their performance vary across different datasets?*
- SubQ2 *How does the duration (number of frames) of the input video affect the models' performances?*

RQ2 *How can deep learning and computer vision methods be leveraged to increase observational research of macaques while ensuring animals' well-being?*

4 METHODS

The data for this research was provided by the Biomedical Primate Research Center (BPRC).

4.1 *Biomedical Primate Research Center (BPRC)*

The Biomedical Primate Research Centre (BPRC) in Rijswijk, The Netherlands, is a scientific research institute that conducts biomedical research on serious diseases and its macaque research has led to groundbreaking discoveries in areas such as infectious diseases (e.g., HIV, hepatitis, or corona), chronic illnesses (e.g., Alzheimer's or Parkinson's, multiple sclerosis or arthritis, or post-traumatic stress disorder (PTSD)), and immunology (BPRC, 2024). In addition, a significant area of research conducted at the BPRC is dedicated to ethology. Ethology refers to the study of animal behavior (e.g., eating, mating, sleeping, and collaborating). Macaques at the BPRC live in so-called pseudo-natural groups, which mimic the social dynamics of groups living in the wild. Research of the ethology group at the BPRC largely focuses on the origin and evolution of the social behavior of primates and, consequently, humans (BPRC, 2024).

While the pseudo-natural habitats contribute to the welfare of the primates, they pose challenges for direct observation. For instance, in specific experiments, particularly those focused on social interactions and group dynamics of macaques, researchers from the ethology group rely on hours of video material recorded using hand cameras. It has been estimated that a detailed analysis of a video takes roughly three times its duration (Ardoin & Sueur, 2023). Automatic pose detection can significantly enhance the quality and quantity of observational research at BPRC without disrupting the primates' pseudo-natural environment.

4.2 *Hardware and Software Specifications*

Several software packages have been used for this research. Pose estimation was achieved by utilizing the deep learning framework for object detection proposed by You Only Look Once version 8 (YOLOv8) (Jocher et al., 2023). Camera calibration and triangulation relies on OpenCV (Bradski, 2000). For action recognition, the OpenMMLab's (M. Contributors, 2020) action recognition framework MMACTION2 (Contributors, 2020) was heavily modified and adapted. The deep learning models were trained on a Windows machine with an NVIDIA GeForce RTX 4090, CUDA 11.8 (NVIDIA, Vingelmann, & Fitzek, 2020) and PyTorch 2.3.0 (Paszke et al., 2017). The code is entirely written in Python 3.12 and available at `https://github.com/laurahgdrn/NHP-AR`.

4.3 *Data Collection*

The data provided for this project consists of video footage from two cameras that cover a small part of one of the cages at the BPRC. The two cameras have a 1/2.8" Progressive Scan CMOS image sensor with a 2.8 to 12 mm focal length and a 2560 × 1440 resolution. The observed area covers approximately $96m^3$ (8m length x 4m width x 3m height). The cameras were constantly recording in February and March.

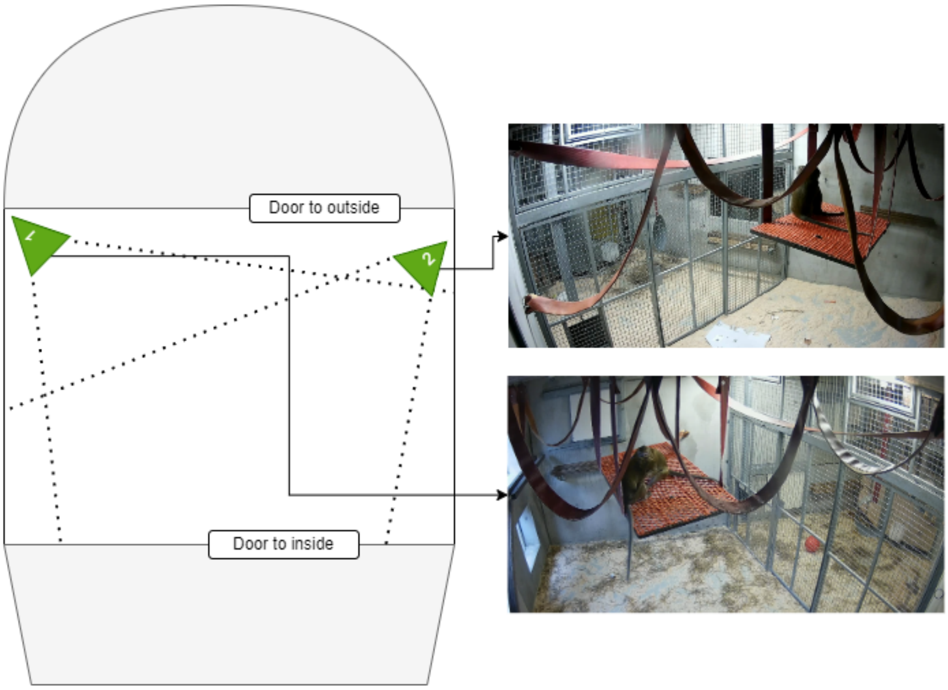


Figure 1: A sketch of the cage, including the camera positions and their views.

4.4 *Data Preparation and Preprocessing*

Several days of video data were available for this study. To minimize manual workload, an adaptive background subtraction algorithm was applied to extract movement. Background subtraction is a widely employed technique in different computer vision applications (Garcia-Garcia, Bouwmans, & Silva, 2020). It aims at classifying an image into foreground and background (Figure 2). In this case, background subtraction has been applied to extract movement, which was achieved by measuring the difference in pixels between a frame and its predecessor. If the pixel difference is below a certain threshold (30 pixels) over a period of 5 consecutive frames, it means that no relevant movement is displayed. Parts without movement were then removed from the input videos, and the segments were manually classified into grooming and playing behavior. This resulted in 261 video segments, from which 159 display grooming behavior and 102 playing behavior. With an average number of 134 frames per clip ($min. = 101, max. = 167$), the total number of frames is estimated to be around 28,710. The videos were then split into training, test, and validations with the following ratios: 0.7, 0.15, and 0.15, respectively.

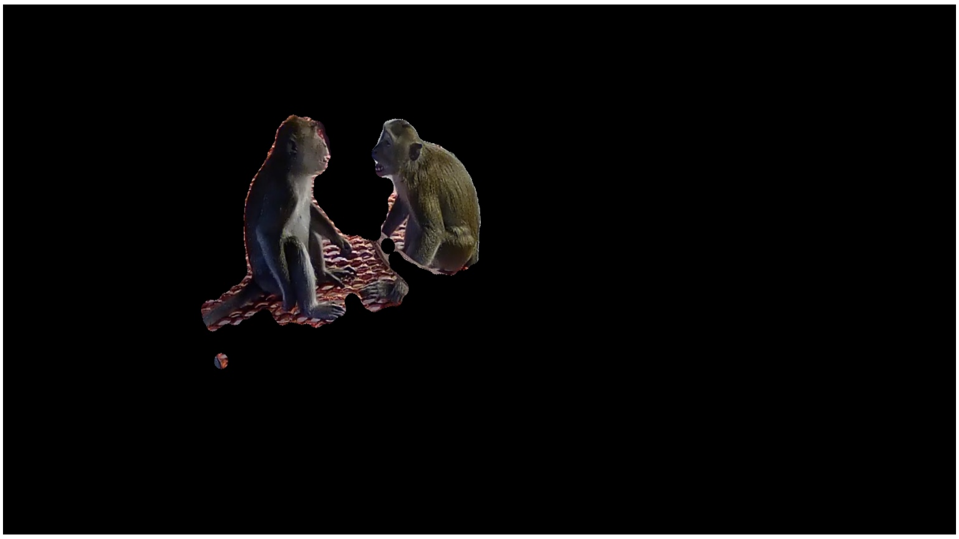


Figure 2: The result of the background subtraction algorithm.

4.4.1 *RGB Data*

For the RGB data condition, no additional processing steps were necessary.

4.4.2 *2D Skeleton Data*

For 2D skeleton action recognition, a pretrained YOLOv8 pose estimation model (subsection 4.8) was employed to automatically predict the poses for each macaque in each frame.

4.4.3 *3D Skeleton Data*

To construct 3D poses, the pretrained YOLO pose estimation model is applied in such a way that it simultaneously iterates through the synchronized frames from both camera view points and predicts the key points. Each key point for each detected macaque is then triangulated using the Direct Linear Transforms (DLT) function. Because the triangulation process

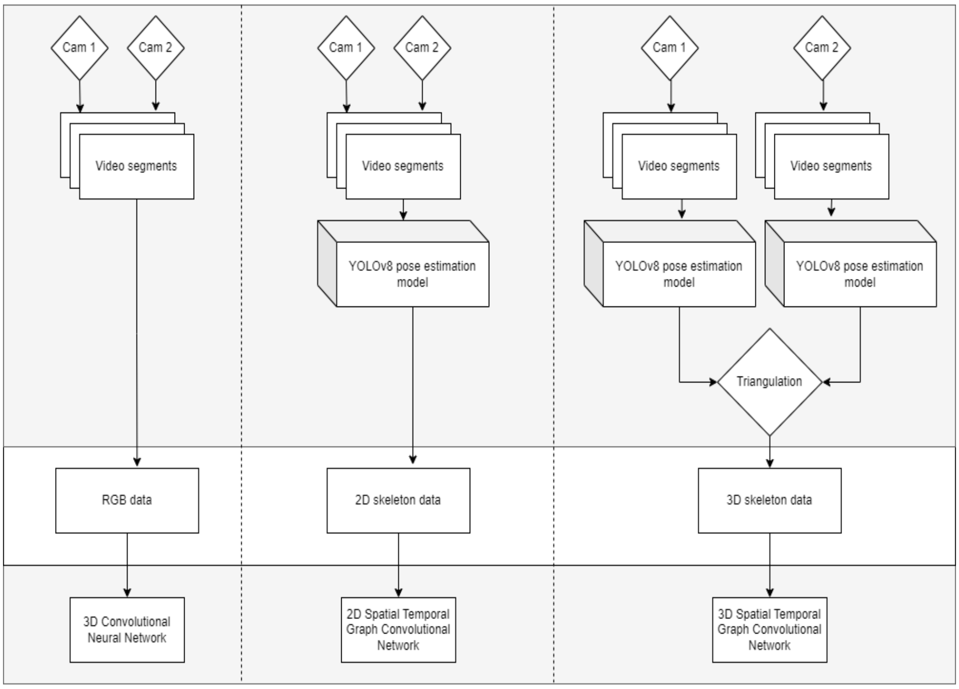

Cam 1
Cam 2
Video segments
YOLOv8 pose estimation model
Triangulation
RGB data
2D skeleton data
3D skeleton data
3D Convolutional Neural Network
2D Spatial Temporal Graph Convolutional Network
3D Spatial Temporal Graph Convolutional Network

Figure 3: The preprocessing steps for the three modalities, which were then used to train distinct deep learning models.

(described in subsection 4.7) requires two viewpoints to triangulate the key points of one animal, the data set for 3D skeleton data is half the size of the 2D skeleton data set.

4.5 *Data Augmentation*

Data augmentation is a common technique during the training phase of deep learning models, and generally describes the process of artificially generating more data samples (Taylor & Nitschke, 2018). Data augmentation does not only improve the performance of deep learning models, but also their robustness and generalizability. In the present research, data augmentation was applied to the raw videos with respect to class imbalance (Table 1). Data augmentation with respect to class imbalance simply refers to augmenting more samples of the underrepresented class (in this case, playing). Similar to Yun, Oh, Heo, Han, and Kim (2020), the videos have been rotated (vertically and horizontally) and cropped. To ensure that no relevant information is lost while cropping, the YOLO model was applied to detect macaques in each frame and crop a fixed size $(500x500pixels)$ around the center of the union of their detected bounding boxes.

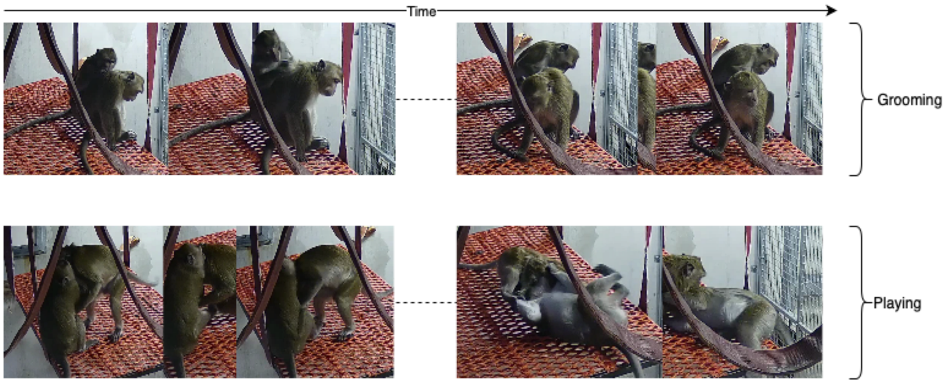


Figure 4: An example of the raw video data for each action class.

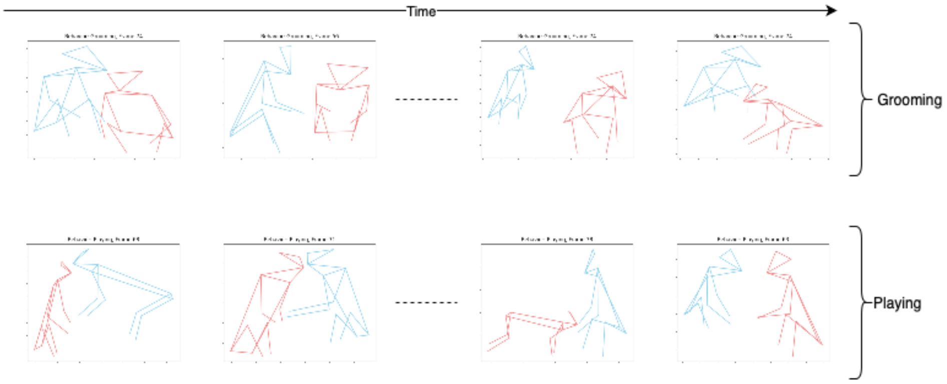


Figure 5: An example of the serialized 2D skeleton data for each action class.

4.6 *Camera Calibration*

The goal of camera calibration techniques is to obtain the extrinsic and intrinsic camera parameters. To compute the intrinsic and extrinsic parameters, it is necessary to find specific points of which the relative positions are known. A common method to achieve this is utilizing a checkerboard pattern (J. Chen et al., 2020; Placht et al., 2014). Initially, images of a checkerboard are captured from various viewpoints and OpenCV is used to automatically detect checkerboard patterns in these images (Bradski, 2000). Given the known dimensions of the checkerboard, the corresponding 3D points (x, y, x) can be mapped to the 2D image points (x, y), allowing for the calculation of distortion coefficients. Distortion coefficients account for focal length (f_x, f_y) and optical centers (c_x, c_y) of one camera. They can be used to create a camera matrix:

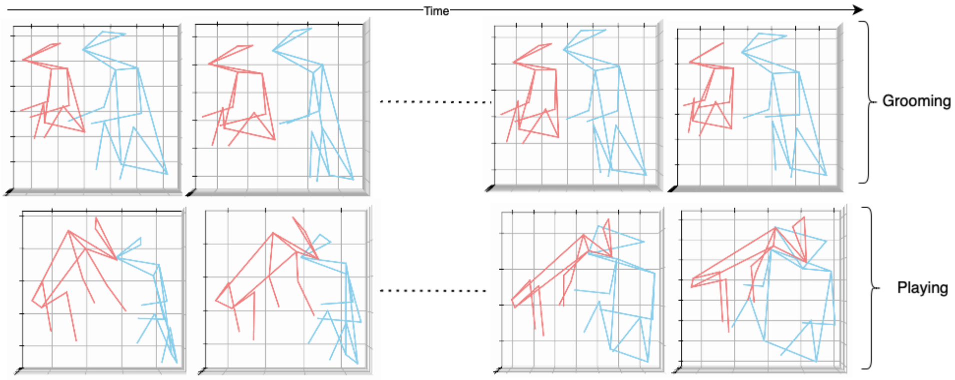


Figure 6: An example of the serialized 3D skeleton data for each action class.

Class	*Before video augmentation*		*After video augmentation*	
	Video Segments	**Frames**	**Video Segments**	**Frames**
Grooming	159	17,490	531	45,450
Playing	102	11,200	511	44,100
Total	**261**	**28,690**	**1044**	**114,840**

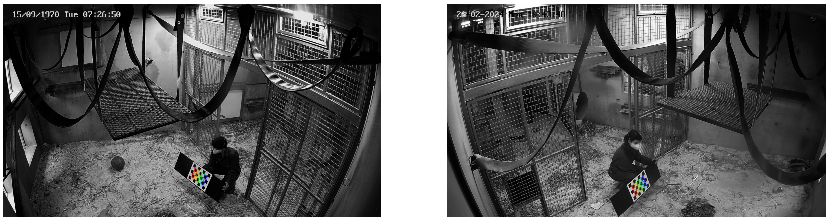
Table 1: The amount of the extracted video segments and frames for each class before and after video augmentation with respect to class imbalance.

$$cameramatrix = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 0 \end{pmatrix}$$

Extrinsic parameters corresponds to rotation and translation vectors which translate the coordinates of a 3D point to a coordinate system. They are described by rotation (R) and translation (t) vectors. The projection matrix P of a single camera is computed by multiplying the camera matrix (K) and the combined rotation and translation matrices:

$$P = K[R, t]$$

The projection matrix can then be used to project a 3D point (x, y, z) onto a 2D image plane. It transforms the 3D point in the camera coordinate system to homogeneous image coordinates by the scaling factor λ:



(a) Camera view 1

(b) Camera view 2

Figure 7: An example of two synchronized frames used for calibration after applying gamma correction, binary threshold, and erosion.

$$\begin{bmatrix} \lambda x \\ \lambda y \\ \lambda \end{bmatrix} = P \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

After calibration, the Root Mean Squared Error (RMSE) was 3.191 pixels. In general, a RMSE below 1 is desirable (Remondino & Fraser, 2006), but the slightly increased value of 3 pixels in this study is likely to be attributable to the camera set up at the BPRC. Because the cameras are located at high ceilings inside the cage and resolution is not optimal, several image enhancements techniques had to be applied to ensure proper detection of the checkerboard corners (Figure 7), which include gamma correction, binary threshold, and erosion. Besides this, camera calibration is generally applied for stereo vision, where two cameras are positioned side by side. However, in this case, the cameras were positioned opposite to each other, requiring additional adjustments through hard-coding to ensure accurate calibration of their perpendicular orientations.

4.7 *Triangulation*

Triangulation is a fundamental concept in computer vision and 3D reconstruction (J. Chen et al., 2020; Qiu et al., 2019; W. Wu, Xu, Liang, Mei, & Peng, 2020) and originates from the field of projective geometry. With the projection matrices (P) obtained for each camera, the triangulation process involves finding the 3D coordinates of the feature point by intersecting the back-projection rays from each camera's image plane (Brill, 1987). This intersection yields the estimated 3D coordinates of the point in the world coordinate system, as shown in Figure 8. A common method to achieve this is the Direct Linear Transformation (DLT) function, which is a mathematical method to solve a linear system using Singular Value Decomposition (SVD) (Brill, 1987).

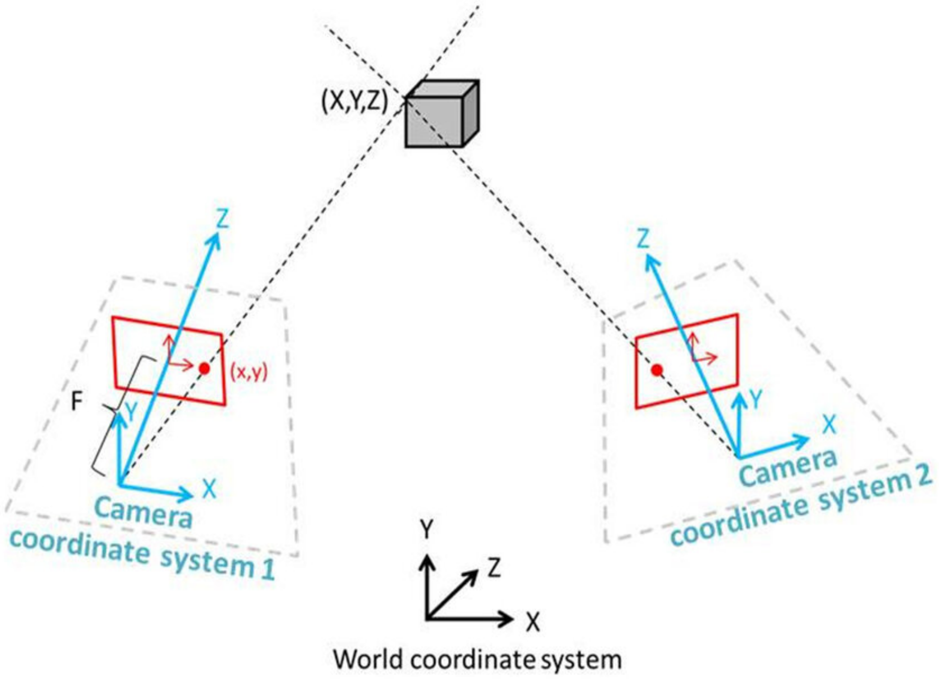


Figure 8: The triangulation process for a dual calibrated camera system. From (Ekberg et al., 2017).

4.8 *Pose Estimation of Macaques using YOLOv8*

As mentioned in section 2, action recognition data sets of macaques do not exist and need to be created from scratch. To create 2D and 3D skeleton data from videos, a You Only Look Once (YOLO) model was trained to perform pose estimation on macaques.

YOLO (You Only Look Once) is a cutting-edge real-time object detection system (Jocher et al., 2023) widely deployed across diverse research domains including agriculture (Tian et al., 2019; D. Wu, Lv, Jiang, & Song, 2020), medicine (Al-Masni et al., 2018; Nie, Sommella, O'Nils, Liguori, & Lundgren, 2019), autonomous vehicles, and security systems (Bhambani, Jain, & Sultanpure, 2020; Kumar, Narasimha Swamy, Kumar, Purohit, & Raju, 2021). The model processes input images through a pretrained convolutional neural network serving as the backbone, comprising 24 layers followed by two fully connected layers. This network predicts bounding box coordinates and their associated probabilities. Only the bounding box with the highest Intersection over Union (IoU) score is retained. IoU measures the overlap between predicted and ground truth bounding boxes by evaluating the ratio of their intersection to their union, ensuring accurate object localization and detection.

4.8.1 *Pretraining*

The model has been pretrained on the MacaquePose data set (Labuguen et al., 2021) for 200 epochs. The MacaquePose data set consists of more than 12,000 images of macaques in various positions and locations. Each macaque is annotated with 17 key points.

4.8.2 *Hyperparameter Optimization*

To ensure optimal performance of the YOLO pose estimation model, hyperparameter optimization was performed (Figure 9). Hyperparameters refer to those parameters that cannot be directly estimated from data learning and must be set before training, in contrast to the model parameters that can be initialized and updated throughout the learning process (e.g., the weights of neurons in neural networks) (Yang & Shami, 2020). In addition to better performance, hyperparameter optimization also increases the reproducibility of deep learning models. For YOLO, Hyperparameters include, among many others, the learning rate, the weight decay, and the momentum (Jocher et al., 2023). An overview of the best hyperparameters can be found in the appendix (section 11).

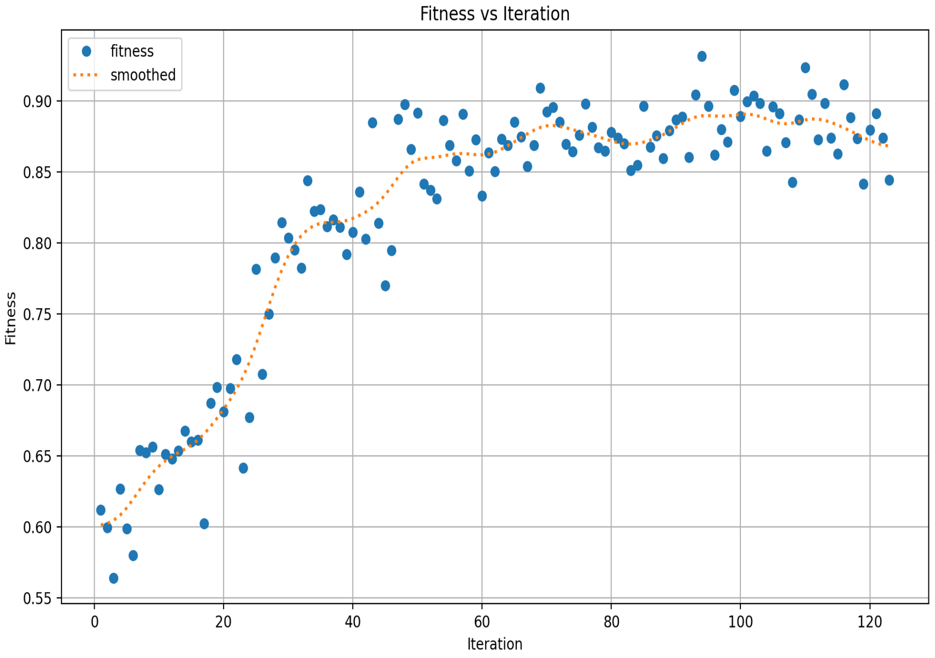


Figure 9: The fitness scores during hyperparameter optimization over 120 iterations.

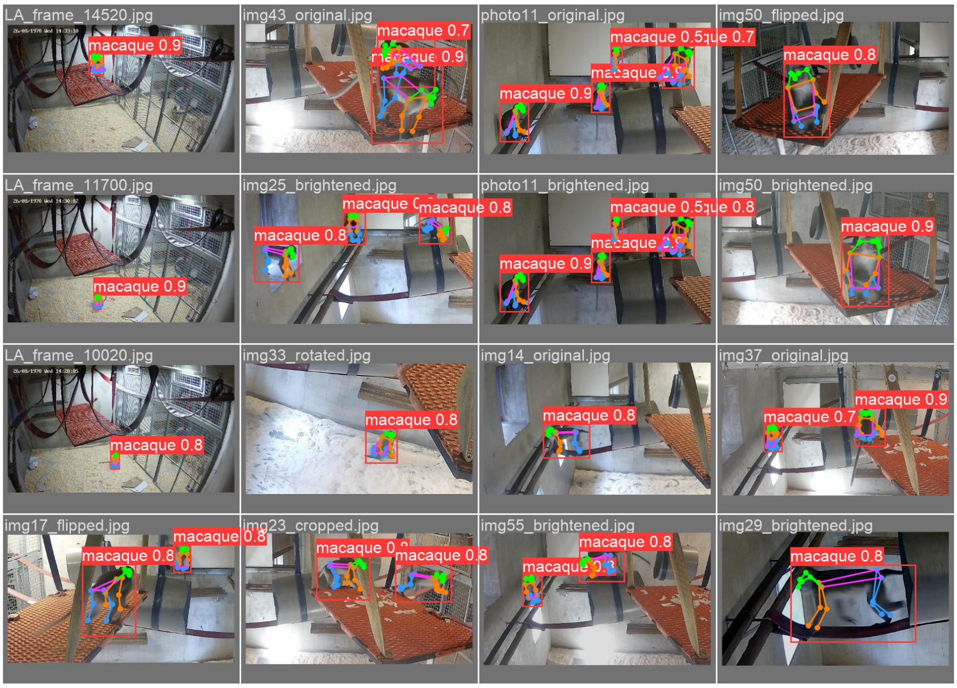


Figure 10: The performance of the model (*model 3*, Table 2) on the BPRC validation set after training and fine-tuning. A demonstration video is available here: `https://youtu.be/n0voSPY1U0I`.

4.8.3 *Intermediate Evaluation of the Pose Estimation Model*

After the hyperparameter tuning was completed, the model was evaluated on different data sets. A key evaluation metric for object detection models is mean Average Precision (mAP). It combines the precision (P) and recall (R) scores for multiple predictions. Recall refers to the models' capability to make predictions for a class compared to the total labels that the class has. Precision is the ratio of correct predictions to all predictions made by the model. Combining these two metrics results in a curve, showing the trade-off as we change the classification threshold: while mAP50 considers the precision at a 50% intersection over union (IoU) threshold, mAP50-95 calculates the average precision over a range of IoU thresholds, from 50% to 95%. A higher mAP suggests better performance. Table 2 shows that for the first model, which was trained and evaluated on the MacaquePose data set, 98% of the detected bounding boxes are correctly localized and 94% of all instances of macaques present in the images are successfully identified, reflecting a high accuracy in the spatial localization and identification of macaques within the images. The mean average precision of 97% calculated at an IoU threshold of 0.50, and 90% at an IoU threshold between 0.50 and 0.95 indicates a high accuracy at correctly detecting the bounding

	Box				Pose			
Model	P	R	mAP50	mAP50-95	P	R	mAP50	mAP50-95
1	0.98	0.94	0.97	0.90	0.95	0.89	0.90	0.72
2	0.77	0.77	0.75	0.58	0.35	0.33	0.14	0.02
3	0.94	0.88	0.92	0.73	0.72	0.64	0.61	0.59

Table 2: The performance of the pose estimation model after being trained on the MacaquePose data set and validated on the MacaquePose data set (*model 1*), trained on the MacaquePose data set and validated on the BPRC data (*model 2*), and fine-tuned on the BPRC data and evaluated on the BPRC data (*model 3*).

boxes in the images. Looking at the pose metrics, we can see that the pose precision is at 0.957 which indicates that 95% of the predicted poses are accurately localized with a recall of 89%. While the mean average precision at an IoU threshold of 0.50 is at 90.2%, it drops to roughly 70% at an IoU threshold between 0.50 and 0.95. The models' performance on the MacaquePose data set is satisfactory, but performance decreased to 35% when predicting key points in the videos from the BPRC (see model 2 in Table 2). The drop in performance is likely due to the high heterogeneity of the two data structures: While the MacaquePose data set consists of high quality images where the macaques are large and centered, the videos contain strong movement and often the monkeys seem relatively small which makes them more difficult to detect. Because of this, the model was fine-tuned on manually annotated frames which significantly improved its performance on the BPRC data: Now, 95% of the detected bounding boxes are correctly localized and 84% of all instances of macaques present in the images are successfully identified. The mean average precision calculated at an IoU threshold of 0.50 is 92%, and 75% at an IoU threshold between 0.50 and 0.95. This indicates a moderate accuracy at correctly detecting the bounding boxes in the images. Looking at the pose metrics, we can see that 72% of the predicted poses are accurately localized, with a recall of 65%. While the mean average precision at an IoU threshold of 0.50 is at 61%, it drops to roughly 59% at an IoU threshold between 0.50 and 0.95.

The final pose estimation model, *model 3*, was then applied to generate a series of 2D skeletons for each clip. These 2D skeletons were triangulated (subsection 4.7) to create 3D poses.

4.9 *Deep learning Models for Action Recognition*

Several deep learning models were trained to perform action recognition of macaques based on three modalities: video segments, 2D skeleton data, and 3D skeleton data. For action recognition from video segments and skeleton data, the MMACTION2 (Contributors, 2020) framework is adapted. MMAction2 is an open source toolkit based on PyTorch that is widely used in research for diverse experiments concerning human action recognition (Y. Chen, 2024; Chiura & van der Haar, 2023; Voropaev, Magomedov, & Alfimtsev, 2024). MMACTION2 is powered by OpenMM-lab (M. Contributors, 2020), which is an open-source computer vision algorithm system. The whole framework is strictly focused on human action recognition, which is why several modifications and adaptations to the source code were necessary to perform Non-Human Primate Action Recognition (NHP-AR). All models, independent of the modality of the data, aim to capture spatial-temporal context. The duration of an input video affects action recognition by providing more temporal context, which can improve accuracy, but also introduces potential noise and increases computational load (Gowda, Rohrbach, & Sevilla-Lara, 2021). In this study, the duration is especially relevant as macaques often switch their behavior almost instantly, and grooming behavior can smoothly transition into playing behavior or the other way around. To ensure a fair comparison across models, common parameters unrelated to specific architectural differences were standardized. For example, the optimization strategy for each model was a stochastic gradient descent (SGD) with a fixed learning rate of 0.1, momentum of 0.9, and weight decay of 0.0005. Each model was trained for 20 epochs.

4.9.1 *SlowFast for Action Recognition from Raw Video Segments*

The SlowFast architecture has been chosen for this NHP-AR task because of several reasons. SlowFast networks are *lightweight* while still achieving state-of-the-art performance. It effectively captures varied temporal dynamics by processing video frames at multiple temporal resolutions, making it suitable for distinguishing between fast movements (e.g., playing) and slower actions (e.g., grooming). The framework consists of a slow pathway, operating at low frame rate, designed to capture spatial semantics, and a fast pathway, operating at high frame rate, capturing motion at fine temporal resolution (Feichtenhofer et al., 2019). The slow pathway is characterized by deeper networks with larger receptive fields, while the fast pathway utilizes shallower networks for faster processing of temporal information. This study's model utilizes a backbone consisting of a Residual Neural Network with 50 layers, including residual blocks, which enable the model to learn hierarchical representations of input data. The temporal stride was set to 16 for the slow pathway, and 2 for the fast pathway. Lateral connections, connecting the slow and the fast pathway, are inserted every four residual blocks.

4.9.2 *2D Spatial-Temporal Graph Convolutional Network (2D-ST-GCN) for Action Recognition from 2-dimensional skeleton data*

ST-GCNs are designed to analyze the spatial configuration of the joints as well as their temporal dynamics (Yan et al., 2018). The construction of the spatial-temporal graph on the skeleton sequences occurs in two stages. Firstly, joints within a single frame are linked with edges based on the connectivity inherent in the primate body structure. Subsequently, each joint is connected to its corresponding joint in the consecutive frame. The ST-GCN model comprises 9 layers of spatial-temporal graph convolution units with a temporal kernel size of 9. Within each ST-GCN unit, a Residual Neural Network (RNN) mechanism is applied to enhance feature representation. For temporal convolution layers, the strides of the 4th and 7th layers are set to 2, acting as pooling layers. The classification head of the model is a Graph Convolutional Network (GCN) head.

4.9.3 *3D Spatial-Temporal Graph Convolutional Network (3D-ST-GCN) for Action Recognition from 3-dimensional skeleton data*

The architecture of the 3D-ST-GCN is identical to the 2D-ST-GCN described above. Merely the shape of each key point changes from (x, y) to (x, y, z).

5 EVALUATION

The most dominant evaluation metric for action recognition is top-1-accuracy (Contributors, 2020; C. Li et al., 2023), which simply refers to the ratio of the amount of correctly predicted instances to the total number of instances. "Correct" in this sense means that the prediction with the highest probability refers to the ground truth. While this is a reliable metric in other studies that contain up to 400 different action classes, it might not be sufficient for this specific research, where behavior is only classified in two categories. Therefore, precision and recall (section 4.8.3) will be utilized to evaluate the performance of the action recognition models[1].

Model	Input	Memory	Accuracy	Precision	Recall
SlowFast	RGB	8116	0.61	0.52	0.45
2D-ST-GCN	2D skel.	1186	0.75	0.79	0.81
3D-ST-GCN	3D skel.	1186	0.71	0.39	0.53

Table 3: Comparing the three action recognition models based on memory usage per epoch, accuracy, precision, and recall.

6 RESULTS

The comparative analysis of three different deep learning models for action recognition on social interactions between macaques reveals significant variations in performance across various metrics. The SlowFast model, which uses RGB input, demonstrates the highest memory usage per epoch (8116 MB) but yields a moderate accuracy of 0.46, with a precision of 0.52 and recall of 0.45. Accuracy in this context refers to the proportion of correctly identified actions (grooming or playing) out of all predictions made by the model. Precision indicates how many of the predicted grooming or playing actions were accurate. Recall, on the other hand, measures the proportion of true positive predictions among all actual positive instances, reflecting the model's ability to identify all instances of grooming and playing actions present in the dataset. The 2D-ST-GCN model exhibits superior performance with an accuracy of 0.75, precision of 0.79, and recall of 0.81, while requiring substantially less memory (1186 MB). The 3D-ST-GCN model, which uses 3D skeletal data, shows a notable drop in precision (0.29) despite having the same memory usage as the 2D-ST-GCN, and its accuracy (0.53) and recall (0.53) are also lower.

- SlowFast: https://youtu.be/YwU50N9mQTE
- 2D-ST-GCN: https://youtu.be/qcv0es_Itcs
- 3D-ST-GCN: https://youtu.be/xrrjLvEGywI

7 CROSS-DATASET PERFORMANCE ANALYSIS

To demonstrate generalizability, the best performing action recognition model, the 2D-ST-GCN, was tested on new videos from a different group. The structure of the enclosure is similar to the training data, but the individuals differ in appearance, size, and age. The 2D-ST-GCN has been selected

[1] mAP is a metric specific to object detection tasks, which is why it will not be adapted.

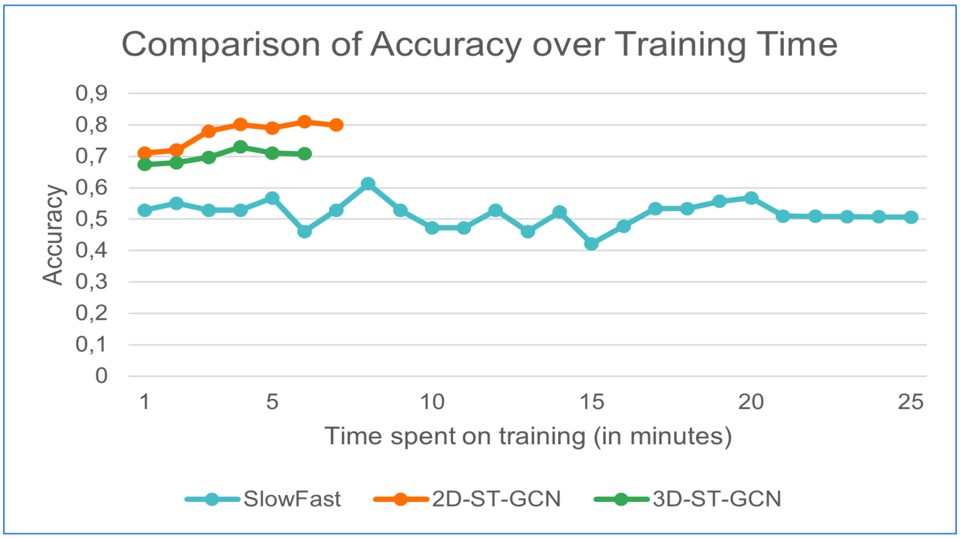


Figure 11: The top-1-accuracy of the three models over training time in minutes.

not exclusively because of its high performance. As mentioned earlier, skeleton data is more robust towards changes in lighting, appearance, and viewpoints. Therefore, it can be assumed that the 2D-ST-GCN performs better on new, unseen data. The 3D-ST-GCN will not be tested on the new data simply because the others enclosures do not accommodate several cameras, eliminating the possibility of 3D pose estimation.

Similar to the training data acquisition, adaptive background subtraction techniques were applied to extract movement, and the YOLO model (subsection 4.8) was utilized to filter segments in which there are at least two macaques present. Without additional manual work, this procedure resulted in 50 video segments. The videos were then manually classified (grooming, playing, or nothing) and compared to the models' prediction for each segment. The "nothing" class refers to "no significant activity" and is assigned to segments where the prediction probability for both grooming and playing behaviors is below 0.8.

As mentioned in subsection 3.2, many parameters can affect the performance of action recognition. One major factor is the number of frames sampled from the input video. In this study, the duration is especially relevant as macaques often switch their behavior almost instantly, and grooming behavior can smoothly transition into playing behavior or the other way around. If frame sampling is too high, those subtle transitions may go undetected. Conversely, if the frame number is too low, important details may be missed, making it difficult to accurately distinguish between different behaviors. Therefore, selecting an optimal amount of frames is essential for capturing the nuanced and rapid behavior changes

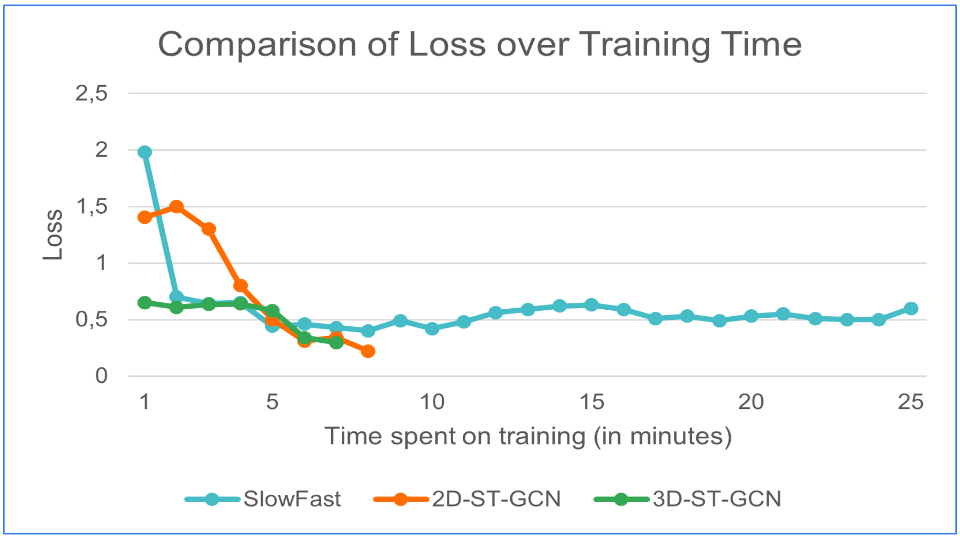


Figure 12: The loss of the three models over training time in minutes.

Sampled frames	Accuracy	Precision	Recall
50	0.47	0.61	0.47
100	0.66	0.65	0.64
200	0.59	0.68	0.59

Table 4: The effect of the number of sampled frames on the performance of the 2D-ST-GCN.

of macaques. To explore the optimal frame rate, three sampling rates have been tested: 50, 100, and 200 (Table 4).

8 DISCUSSION

This research focused on recognizing social interactions, grooming and playing, between macaques using different deep learning architectures. Grooming has a crucial function in establishing relationships and ensuring emotional stability within groups, and also aids in physical well-being by combating ectoparasites (Simons et al., 2022; Solanki et al., 2020b). Similarly, playing in social animals serves to strengthen group cohesion and facilitate social learning (Beltran Frances et al., 2020). Distinguishing those close interactions of macaques from raw video data is not a trivial task. Both behaviors require close spatial proximity, which introduces challenges for computer vision models. The temporal domain is crucial to distinguish the two actions from each other, as single frames (or skeletons) from both actions can be identical.

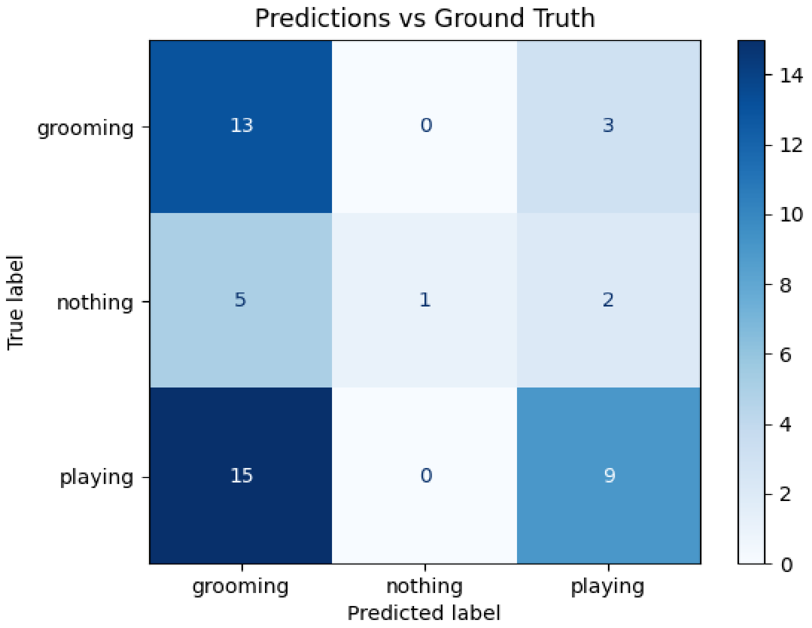


Figure 13: Confusion matrix of the ST-GCN model with a sampling rate of 50 on the new test set. The model correctly identified 13 instances of grooming but misclassified 3 as playing and none as 'nothing'. For playing, it only classified 9 instances correctly. It also misclassified playing as grooming 15 times.

This research proposed a YOLOv8 pose estimation model to automatically extract 2D poses of macaques from video data. After hyperparameter optimization and fine-tuning, the best performing model achieved an accuracy of about 70% at predicting key points, which is significanly lower than the markerless system proposed by Bala et al. (2020) which achieved an average detection accuracy of around 95% for key body parts when compared to manually annotated ground truth data. This is likely due to the different setups in both studies. In the study by Bala et al. (2020), the researchers utilized a highly professional setup, involving 62 cameras that encircle an open 8m^3 enclosure specifically designed for optimal pose estimation. This controlled environment minimizes occlusions and ensures comprehensive coverage from multiple angles, facilitating more accurate key point detection. In contrast, our study utilized only two cameras to cover a larger area of 96m^3. These cameras were placed in an existing enclosure where the macaques naturally reside. This more naturalistic and less controlled environment presents numerous challenges, such as increased occlusions, varying lighting conditions, and limited camera angles, which can significantly impact the accuracy of pose estimation. C. Li et al. (2023) achieved a top-1-accuracy of 98% at detecting semantic actions

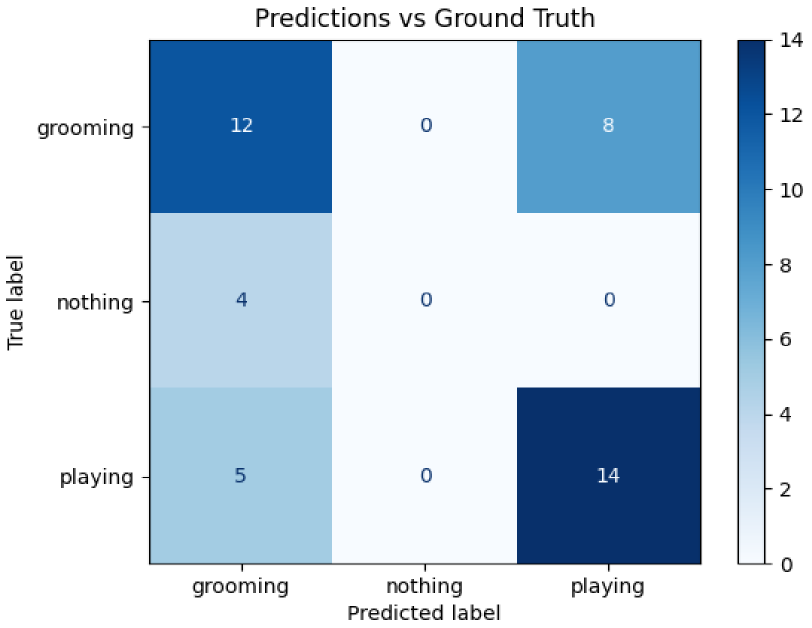


Figure 14: Confusion matrix of the ST-GCN model with a sampling rate of 100 on the new test set. With an increased frame rate sampling, the model correctly identified 12 instances of playing as playing but misclassified 8 as grooming and none as 'nothing'. It accurately identified 14 occurrences of playing but misclassified 5 as grooming.

such as climbing, walking, and sitting. The best model proposed in this research achieved an accuracy of 75% at classifying social interactions. This discrepancy is likely due to the fact that social interactions are more challenging to detect due to additional uncertainty and occlusions.

In the following, the research questions presented in section 2 will be revisited.

RQ1: To what extent can deep neural networks be employed to identify grooming and playing behaviors among long-tailed macaques?

This study explored three different deep learning architectures and modalities. The best performing model was a Spatial-Temporal Graph Convolutional Network, which was trained on 2D skeleton data and achieved a satisfying accuracy of 75%. This suggests that the 2D-ST-GCN model is both more efficient and effective for this task compared to the other proposed models. It achieves a high precision (79%) and recall (81%), which means it not only makes fewer false positive predictions compared to the other models, but also successfully identifies a high proportion of

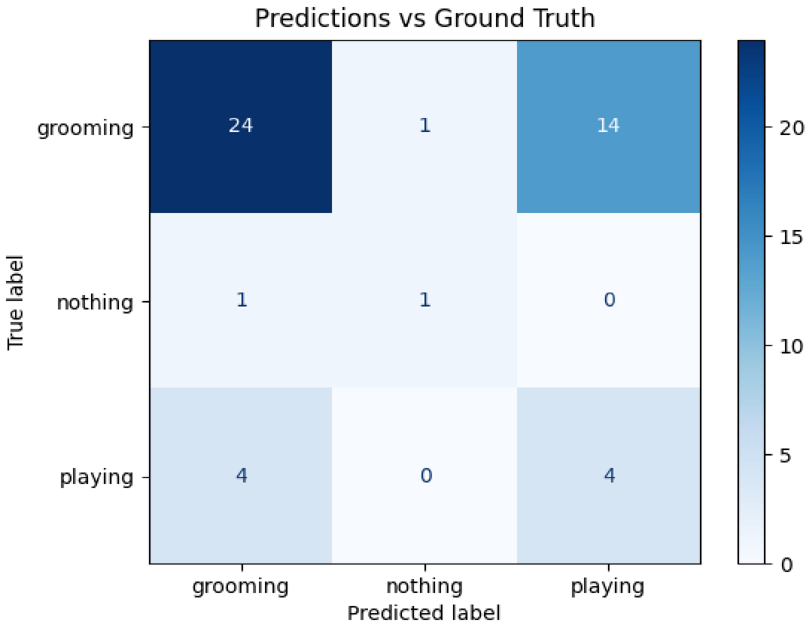


Figure 15: Confusion matrix of the ST-GCN model with a sampling rate of 200 on the new test set. The model correctly identified 24 instances of grooming as grooming, but misclassified 14 as playing and 1 as 'nothing'. It correctly classified 1 instance of grooming but misclassified 1 as playing. It only classified 4 instances of playing behavior correctly, and misclassified 4 as grooming. The strong performance for grooming and the weaker performance for playing indicate a possible imbalance or overlap in the feature representation of these classes.

the actual grooming and playing actions. Similar to the model architectures, each modality has its own advantage and disadvantage concerning usability, scalability, and computational resources. RGB data is easy to acquire and usually does not require complex preprocessing, making it a reasonable choice for researchers in distinct fields without advanced programming skills. However, video data requires large storage capacities and significant computational power for processing and analysis. Additionally, video data analysis sensitive to lighting conditions, occlusions, and camera angles. On the other hand, 2D skeletal data can reduce the amount of information to process and focus on the movement patterns, but the acquisition process is more complex, which makes it less accessible. It also relies on pose estimation algorithms, which can introduce errors. However, 2D skeletons are still affected by occlusions and different view points or camera angles. 3D skeleton data reduces the impact of occlusions and viewpoint changes, while also providing accurate spatial information about the subject's movements. However, the generation of 3D data requires even more sophisticated capture methods, such as multiple camera setups or depth sensors, which can be costly and technically demanding to implement. One major advantage of 2D and 3D skeleton data is the reduced data dimensionality, which is also reflected by the amount of training time that was necessary to train the deep learning models: The RGB-based model took 25 minutes to reach an accuracy of 60%, while the skeleton-based AR models both achieved an accuracy of more than 70% within 5 minutes of training.

SubQ1: How does their performance vary across different datasets?

When the 2D ST-GCN model was applied to video data from a different enclosure, its performance accuracy dropped to 66%, demonstrating limited robustness and generalizability across new groups and viewpoints. This performance decline is likely due to the heterogeneity between the two groups. Although 2D skeleton data is generally robust against changes in appearance (C. Wang & Yan, 2023), the macaques in the two groups differ in age and gender, which may influence their grooming and playing style. While the training group mainly consists of older males, the test group is larger and more diverse in terms of age and gender. Solanki et al. (2020a) has shown that male and female macaques have distinct grooming preferences, and Mayhew et al. (2020) stated that young individuals, especially males, are usually more playful. These variations in social behaviors likely contributed to the model's performance drop, as it may have overfitted to the grooming and playing style of adult males, thus failing to capture the broader range of behaviors present in a more diverse group.

SubQ2: How does the duration (number of frames) of the input video affect the models' performances?

As mentioned in subsection 3.2, the duration of an input video plays a crucial role in action recognition, as it provides additional temporal context that can enhance the accuracy of identifying actions. However, longer videos also introduce the risk of incorporating extraneous information, or noise, which can complicate the recognition process and increase the computational burden (Gowda et al., 2021). In the context of this study, the video duration is particularly significant due to the behavioral patterns of macaques. These animals frequently exhibit rapid shifts in behavior, such as transitioning almost instantaneously from grooming to playing and vice versa. As a result, accurately capturing and recognizing these behaviors requires a balance between a video length that is sufficient to understand the context and transitions, yet concise enough to minimize noise and maintain computational efficiency. To explore this effect, three sample sizes have been explored on the performance on the 2D ST-GCN: 50 frames, 100 frames, and 200 frames per sample. It has been shown that the model's best performance is achieved at 100 frames per sample. At this frame rate, the model effectively balanced the trade-off between temporal context and noise, capturing enough detail to accurately identify transitions between grooming and playing behaviors without being overwhelmed by unnecessary information.

How can deep learning and computer vision methods be leveraged to increase observational research of macaques while ensuring animals' well-being?

Automatic pose estimation and action recognition have several advantages compared to traditional methods. By minimizing constant human presence and interference, this approach allows animals to act uninhibitedly, which provides a more authentic depiction of their actions and interactions. Moreover, automatic action recognition facilitates the identification of subtle patterns and changes in social behavior that may signify underlying health issues, stress, or shifts in group dynamics. Detecting such changes early allows for suitable interventions, enhancing animal welfare. In similar experiments, single primates are held in highly unnatural and laboratory settings Bala et al. (2020,?); C. Li et al. (2023,?); Martini et al. (2024); Mathis et al. (2018); Nakamura et al. (2016,?). This study proves that automated methods can be effectively applied in natural settings, thereby improving the ecological validity of behavioral observations and ensuring the well-being of the animals.

9 LIMITATIONS

Notably, all limitations concerning data collection are attributable to animal welfare. The macaques present in the videos could move freely through the inside and outside compartments of their pseudo-natural enclosure. Due to this, data collection was scarce and declined as the year progressed because of improving weather conditions. The data set is not only small compared to other action recognition data set, the samples are also challenging in the sense of different lighting conditions and large groups of primates engaging in close activities.

The cameras' position and quality introduced another limitation. The cameras were positioned at perpendicular orientation at the ceiling of one of the inner compartments. Their angles towards the ground differed, and the area covered by both cameras was not optimal, which further hampered the process of data collection. Next to low resolution, the cameras did not provide time stamps for efficient frame synchronization. Eventually, frame synchronization was achieved by strict hard-coding, not accounting for frequent individual breaks of recording, which again affected the generation of 3D poses from both video streams.

The current model was designed and trained specifically to detect interactions involving only two macaques. Consequently, interactions that involve three or more animals are not detected by the model, leading to the loss of valuable information about more complex social dynamics. For instance, in a scenario where four macaques performing grouped grooming behavior, the model would only detect two of the macaques participating in the interaction. Although the model might still accurately identify the action as grooming, it fails to capture the full extent of the social interaction. This limitation prevents a comprehensive analysis of social behaviors that involve multiple participants, such as multi-animal play or complex grooming chains. Understanding these multi-animal interactions is crucial for a more complete insight into the social structure and dynamics of macaque groups. Therefore, enhancing the model to detect and analyze interactions involving multiple animals simultaneously would be an important step towards achieving a more thorough understanding of macaque social behavior.

The 2D and 3D skeleton data utilized in this research relied on a pose estimation algorithm. Although the pose model was pretrained and fine-tuned, its key point predictions were not flawless, which introduced noise to the data. This noise is intensified when triangulated (Bartol, Bojanić, Petković, & Pribanić, 2022). Other studies have explored more advanced triangulation techniques for 3D (human) pose estimation, such as learnable triangulation through the usage of heat maps. Heat maps have a large advantage against key points, because they are represented as a probability instead of a single point. When overlapping the heat maps from several viewpoints for each key point, machine learning models can learn to efficiently estimate the 3D location of a point based on the heat maps. This process is more robust towards noise and the resulting 3D poses are more accurate (Bartol et al., 2022). In addition to this, the triangulation approach proposed in this research requires two 2D points to compute one point in 3D space. If a primate is only detected from one camera view, its 3D pose cannot be retrieved, which further reduced the data set size of the 3D skeletons. This is likely to be the reason why the 3D-ST-GCN did not perform better than the 2D-ST-GCN, in contrast to the establishment that 3-dimensional skeleton data is generally more accurate and robust than 2D skeleton data (Peng et al., 2021). The SlowFast model underperformed compared to the other two models, while also requiring more memory and more time for training. The mediocre performance could be attributable to the low amount of training epochs. To guarantee a fair comparison, each model was trained for 20 epochs. However, in the original paper that introduced the SlowFast architecture (Feichtenhofer et al., 2019), the amount of training epochs was set to 256. The same amount was used by Xiao, Lee, Grauman, Malik, and Feichtenhofer (2020). The initialized amount of epochs for skeleton-based action recognition models is usually much lower (J. Chen et al., 2020; Yan et al., 2018).

10 CONCLUSION

This study conducted a comprehensive exploration of various deep learning techniques tailored for the intricate tasks of social action recognition in primates. The video footage for this study was collected from two calibrated cameras in a semi-natural enclosure at the Biomedical Primate Research Center (BPRC). The macaques were able to move freely through the inside and outside compartments, and leave or enter the recorded area at any given time. By doing this, not only the observation of natural behavior is guaranteed but also high standards of well-being. Three distinct modalities (RGB, 2D skeleton, and 3D skeleton data) and deep learning architectures (SlowFast, 2D-ST-GCN, and 3D-ST-GCN) were explored to detect two social behaviors: grooming and playing. For skeleton data acquisition, a YOLOv8 pose estimation model was trained from scratch and used to perform pose estimation on the video sequences. The 2D skeleton data points were then triangulated to create 3D skeletons. The highest performance was achieved by the 2D-ST-GCN, with an accuracy of 75%. This study shows that it is possible to detect naturally occurring, "between pairs" level actions in semi-natural cages between two primates, without the need of special equipment or expensive cameras. In addition to the presented action recognition models, even the preprocessing procedures presented in this research enhance observational research. Instead of having to scan through hours of videos, researchers at the BPRC or other research institutes can make use of the background subtraction algorithm together with the pretrained YOLO model to almost instantly filter out sequences where macaques are present.

11 FUTURE WORK

Grooming and playing serve several social purposes between pairs, with frequent occurrences indicating healthy and stable groups. The deep learning methods proposed in this study enable automatic monitoring of these social interactions. The same approach can be extended to a broader range of behaviors. For example, automatically detecting behaviors that negatively influence the social bonds, such as aggression, provide valuable insights into social processes like reconciliation. Additionally, a combination of the proposed action recognition models together with identification of individuals would enable the automatic generation of social networks. As mentioned in the introduction, play incorporates behavioral elements from various social situations, blurring the distinction between play and non-play behaviors (Wright et al., 2018). Therefore, it would be interesting to see whether a model could learn to distinguish these subtle differences between play and non-play. Furthermore, the model proposed in this research is currently capable of performing inference only on offline data. Implementing real-time inference and action recognition can reduce computational resources by immediately identifying and selecting relevant sequences for storage.

Finally, the ST-GCN did not perform well across different groups and viewpoints. Future research should focus on training the models on larger and more diverse data sets to avoid overfitting.

REFERENCES

Aguilar-Melo, A. R., Calme, S., Pinacho-Guendulain, B., Smith-Aguilar, S. E., & Ramos-Fernández, G. (2020). Ecological and social determinants of association and proximity patterns in the fission-fusion society of spider monkeys (ateles geoffroyi). *American Journal of Primatology*, *82*(1), e23077.

Al-Masni, M. A., Al-Antari, M. A., Park, J.-M., Gi, G., Kim, T.-Y., Rivera, P., . . . Kim, T.-S. (2018). Simultaneous detection and classification of breast masses in digital mammograms via a deep learning yolo-based cad system. *Computer methods and programs in biomedicine*, *157*, 85–94.

Ardoin, T., & Sueur, C. (2023). Automatic identification of stone-handling behaviour in japanese macaques using labgym artificial intelligence. *arXiv preprint arXiv:2310.07812*.

Bala, P. C., Eisenreich, B. R., Yoo, S. B. M., Hayden, B. Y., Park, H. S., & Zimmermann, J. (2020). Automated markerless pose estimation in freely moving macaques with openmonkeystudio. *Nature communications*, *11*(1), 4560.

Bartol, K., Bojanić, D., Petković, T., & Pribanić, T. (2022). Generalizable human pose triangulation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11028–11037).

Beltran Frances, V., Castellano-Navarro, A., Illa Maulany, R., Ngakan, P. O., MacIntosh, A. J., Llorente, M., & Amici, F. (2020). Play behavior in immature moor macaques (macaca maura) and japanese macaques (macaca fuscata). *American Journal of Primatology*, *82*(10), e23192.

Bhambani, K., Jain, T., & Sultanpure, K. A. (2020). Real-time face mask and social distancing violation detection system using yolo. In *2020 ieee bangalore humanitarian technology conference (b-htc)* (pp. 1–6).

Bibbò, L., & Vellasco, M. M. (2023). *Human activity recognition (har) in healthcare* (Vol. 13) (No. 24). MDPI.

Bilesan, A., Komizunai, S., Tsujita, T., & Konno, A. (2021). Improved 3d human motion capture using kinect skeleton and depth sensor. *Journal of robotics and mechatronics*, *33*(6), 1408–1422.

BPRC. (2024). *Research achievements.* https://www.bprc.nl/index.php/nl/onderzoeksgebieden#missievanbprc. (Accessed: 10.05.2024)

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Brill, M. H. (1987). Triangulating from optical and sar images using direct linear transformations. *Photogrammetric engineering and remote sensing*, *53*(8), 1097–1102.

Cai, J., Jiang, N., Han, X., Jia, K., & Lu, J. (2021). Jolo-gcn: mining joint-centered light-weight information for skeleton-based action recognition. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 2735–2744).

Chen, C.-F. R., Panda, R., Ramakrishnan, K., Feris, R., Cohn, J., Oliva, A., & Fan, Q. (2021, June). Deep analysis of cnn-based spatio-temporal representations for action recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 6165-6175).

Chen, J., Wu, D., Song, P., Deng, F., He, Y., & Pang, S. (2020). Multi-view triangulation: Systematic comparison and an improved method. *Ieee Access*, *8*, 21017–21027.

Chen, Y. (2024). Action detection in badminton courts using ava dataset and mmaction2 architecture with slow fast model. *Highlights in Science, Engineering and Technology*, *85*, 783–789.

Chiura, T. B., & van der Haar, D. (2023). Offensive play recognition of basketball video footage using actionformer. In *International conference on human-computer interaction* (pp. 447–454).

Contributors. (2020). *Openmmlab's next generation video understanding toolbox and benchmark.* https://github.com/open-mmlab/mmaction2.

Contributors, M. (2020). *Openmmlab computer vision foundation.* https://github.com/open-mmlab/mmc.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2625–2634).

Dwibedi, D., Sermanet, P., & Tompson, J. (2018). Temporal reasoning in videos using convolutional gated recurrent units. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops* (pp. 1111–1116).

Ekberg, P., Daemi, B., & Mattsson, L. (2017, 02). 3d precision measurements of meter-sized surfaces using low cost illumination and camera techniques. *Measurement Science and Technology*, *28*. doi: 10.1088/1361-6501/aa5ae6

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019, October). Slowfast networks for video recognition. In *Proceedings of the ieee/cvf international conference on computer vision (iccv).*

Foster, J. D., Nuyujukian, P., Freifeld, O., Gao, H., Walker, R., Ryu, S. I., ... Shenoy, K. V. (2014). A freely-moving monkey treadmill model. *Journal of neural engineering*, *11*(4), 046020.

Gaither, A. M., Baker, K. C., Gilbert, M. H., Blanchard, J. L., Liu, D. X., Luchins, K. R., & Bohm, R. P. (2014). Videotaped behavior as a predictor of clinical outcome in rhesus macaques (macaca mulatta). *Comparative Medicine*, *64*(3), 193–199.

Gammulle, H., Ahmedt-Aristizabal, D., Denman, S., Tychsen-Smith, L., Petersson, L., & Fookes, C. (2023). Continuous human action recognition for human-machine interaction: a review. *ACM Computing Surveys*, *55*(13s), 1–38.

Garcia-Garcia, B., Bouwmans, T., & Silva, A. J. R. (2020). Background subtraction in real applications: Challenges, current models and future directions. *Computer Science Review*, *35*, 100204.

Gardner, M. B., & Luciw, P. A. (2008). Macaque models of human infectious disease. *ILAR journal*, *49*(2), 220–255.

Gilja, V., Chestek, C. A., Nuyujukian, P., Foster, J., & Shenoy, K. V. (2010). Autonomous head-mounted electrophysiology systems for freely behaving primates. *Current opinion in neurobiology*, *20*(5), 676–686.

Gowda, S. N., Rohrbach, M., & Sevilla-Lara, L. (2021). Smart frame selection for action recognition. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 35, pp. 1451–1459).

Hannibal, D. L., Bliss-Moreau, E., Vandeleest, J., McCowan, B., & Capitanio, J. (2017). Laboratory rhesus macaque social housing and social changes: Implications for research. *American Journal of Primatology*, *79*(1), e22528.

Hayden, B. Y., Park, H. S., & Zimmermann, J. (2022). Automated pose estimation in primates. *American journal of primatology*, *84*(10), e23348.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer* *vision and pattern recognition* (pp. 770–778).

Huang, Z., Shen, X., Tian, X., Li, H., Huang, J., & Hua, X.-S. (2020). Spatio-temporal inception graph convolutional networks for skeleton-based action recognition. In *Proceedings of the 28th acm international conference on multimedia* (pp. 2122–2130).

Iskakov, K., Burkov, E., Lempitsky, V., & Malkov, Y. (2019). Learnable triangulation of human pose. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 7718–7727).

Jablonski, N. G. (2021). Social and affective touch in primates and its role in the evolution of social cohesion. *Neuroscience*, *464*, 117–125.

Jocher, G., Chaurasia, A., & Qiu, J. (2023). *Ultralytics yolov8.* Retrieved from https://github.com/ultralytics/ultralytics

Kaburu, S. S., Marty, P. R., Beisner, B., Balasubramaniam, K. N., Bliss-Moreau, E., Kaur, K., ... McCowan, B. (2019). Rates of human–macaque interactions affect grooming behavior among urban-dwelling rhesus macaques (macaca mulatta). *American Journal of Physical Anthropology*, *168*(1), 92–103.

Kellman, J., & Radwan, K. (2022). Towards an expanded neuroscientific understanding of social play. *Neuroscience & Biobehavioral Reviews*, *132*, 884–891.

Kim, P.-S., Lee, D.-G., & Lee, S.-W. (2018). Discriminative context learning with gated recurrent unit for group activity recognition. *Pattern Recognition*, *76*, 149–161.

Knaebe, B., Weiss, C. C., Zimmermann, J., & Hayden, B. Y. (2022). The promise of behavioral tracking systems for advancing primate animal welfare. *Animals*, *12*(13), 1648.

Kocabas, M., Karagoz, S., & Akbas, E. (2018). Multiposenet: Fast multi-person pose estimation using pose residual network. In *Proceedings of the european conference on computer vision (eccv)* (pp. 417–433).

Kumar, P., Narasimha Swamy, S., Kumar, P., Purohit, G., & Raju, K. S. (2021). Real-time, yolo-based intelligent surveillance and monitoring system using jetson tx2. In *Data analytics and management: Proceedings of icdam* (pp. 461–471).

LaBarge, L. R., Allan, A. T., Berman, C. M., Margulis, S. W., & Hill, R. A. (2020). Reactive and pre-emptive spatial cohesion in a social primate. *Animal behaviour*, *163*, 115–126.

Labuguen, R., Matsumoto, J., Negrete, S. B., Nishimaru, H., Nishijo, H., Takada, M., ... Shibata, T. (2021). Macaquepose: a novel "in the wild" macaque monkey pose dataset for markerless motion capture. *Frontiers in behavioral neuroscience*, *14*, 581154.

Lauer, J., Zhou, M., Ye, S., Menegas, W., Schneider, S., Nath, T., ... others (2022). Multi-animal pose estimation, identification and tracking with deeplabcut. *Nature Methods*, *19*(4), 496–504.

Li, C., Xiao, Z., Li, Y., Chen, Z., Ji, X., Liu, Y., ... others (2023). Deep learning-based activity recognition and fine motor identification using 2d skeletons of cynomolgus monkeys. *Zoological Research*, *44*(5), 967.

Li, W., Wen, L., Chang, M.-C., Nam Lim, S., & Lyu, S. (2017). Adaptive rnn tree for large-scale human action recognition. In *Proceedings of the ieee international conference on computer vision* (pp. 1444–1452).

Ma, Z., et al. (2021). Human action recognition in smart cultural tourism based on fusion techniques of virtual reality and som neural network. *Computational Intelligence and Neuroscience*, *2021*.

Martini, L. M., Bognár, A., Vogels, R., & Giese, M. A. (2024). Macaction: Realistic 3d macaque body animation based on multi-camera markerless motion capture. *bioRxiv*, 2024–01.

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.

Mayhew, J. A., Funkhouser, J. A., & Wright, K. R. (2020). Considering social play in primates: A case study in juvenile tibetan macaques. *The behavioral ecology of the Tibetan macaque*, 93.

Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., & Bronstein, M. M. (2017). Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 5115–5124).

Morshed, M. G., Sultana, T., Alam, A., & Lee, Y.-K. (2023). Human action recognition: A taxonomy-based survey, updates, and opportunities. *Sensors*, *23*(4), 2182.

Munea, T. L., Jembre, Y. Z., Weldegebriel, H. T., Chen, L., Huang, C., & Yang, C. (2020). The progress of human pose estimation: A survey and taxonomy of models applied in 2d human pose estimation. *IEEE Access*, *8*, 133330–133348.

Nakamura, T., Matsumoto, J., Nishimaru, H., Bretas, R. V., Takamura, Y., Hori, E., ... Nishijo, H. (2016). A markerless 3d computerized motion capture system incorporating a skeleton model for monkeys. *PloS one*, *11*(11), e0166154.

Nath, T., Mathis, A., Chen, A. C., Patel, A., Bethge, M., & Mathis, M. W. (2019). Using deeplabcut for 3d markerless pose estimation across species and behaviors. *Nature protocols*, *14*(7), 2152–2176.

Nie, Y., Sommella, P., O'Nils, M., Liguori, C., & Lundgren, J. (2019). Automatic detection of melanoma with yolo deep convolutional neural networks. In *2019 e-health and bioengineering conference (ehb)* (pp. 1–4).

Noh, S.-H. (2021). Analysis of gradient vanishing of rnns and performance comparison. *Information*, *12*(11), 442.

NVIDIA, Vingelmann, P., & Fitzek, F. H. (2020). *Cuda, release: 10.2.89.* Retrieved from https://developer.nvidia.com/cuda-toolkit

OpenAI. (2024). *Chatgpt: Gpt-4.* https://www.openai.com/chatgpt. (Accessed: 2024-06-08)

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., . . . Lerer, A. (2017). Automatic differentiation in pytorch.

Peng, W., Shi, J., Varanka, T., & Zhao, G. (2021). Rethinking the st-gcns for 3d skeleton-based human action recognition. *Neurocomputing*, *454*, 45–53.

Pereira, T. D., Aldarondo, D. E., Willmore, L., Kislin, M., Wang, S. S.-H., Murthy, M., & Shaevitz, J. W. (2019). Fast animal pose estimation using deep neural networks. *Nature methods*, *16*(1), 117–125.

Placht, S., Fürsattel, P., Mengue, E. A., Hofmann, H., Schaller, C., Balda, M., & Angelopoulou, E. (2014). Rochade: Robust checkerboard advanced detection for camera calibration. In *Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part iv 13* (pp. 766–779).

Qiu, H., Wang, C., Wang, J., Wang, N., & Zeng, W. (2019). Cross view fusion for 3d human pose estimation. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 4342–4351).

Remondino, F., & Fraser, C. (2006). Digital camera calibration methods: considerations and comparisons. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *36*(5), 266–272.

Schülke, O., Anzà, S., Crockford, C., De Moor, D., Deschner, T., Fichtel, C., . . . others (2022). Quantifying within-group variation in sociality—covariation among metrics and patterns across primate groups and species. *Behavioral Ecology and Sociobiology*, *76*(4), 50.

Seyfarth, R. M. (1977). A model of social grooming among adult female monkeys. *Journal of theoretical Biology*, *65*(4), 671–698.

Shahroudy, A., Liu, J., Ng, T.-T., & Wang, G. (2016). Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1010–1019).

Shimada, M., & Sueur, C. (2018). Social play among juvenile wild japanese macaques (macaca fuscata) strengthens their social bonds. *American Journal of Primatology*, *80*(1), e22728.

Simons, N. D., Michopoulos, V., Wilson, M., Barreiro, L. B., & Tung, J. (2022). Agonism and grooming behaviour explain social status effects on physiology and gene regulation in rhesus macaques. *Philosophical Transactions of the Royal Society B*, *377*(1845), 20210132.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, *27*.

Sirot, E., & Touzalin, F. (2009). Coordination and synchronization of vigilance in groups of prey: the role of collective detection and predators' preference for stragglers. *The American Naturalist*, *173*(1), 47–59.

Solanki, G., Lalremruati, P., Lalchhuanawma, K., et al. (2020a). Grooming pattern in captive macaques: A comparative study. *Environment Conservation Journal*, *21*(3), 127–135.

Solanki, G., Lalremruati, P., Lalchhuanawma, K., et al. (2020b). Grooming pattern in captive macaques: A comparative study. *Environment Conservation Journal*, *21*(3), 127–135.

Sujith, B. (2014). Crime detection and avoidance in atm: a new framework. *International Journal of Computer Science and Information Technologies*, *5*(5), 6068–6071.

Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., & Liu, J. (2022). Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence*, *45*(3), 3200–3225.

Taylor, L., & Nitschke, G. (2018). Improving deep learning with generic data augmentation. In *2018 ieee symposium series on computational intelligence (ssci)* (pp. 1542–1547).

Testard, C., Tremblay, S., & Platt, M. (2021). From the field to the lab and back: neuroethology of primate social behavior. *Current opinion in neurobiology*, *68*, 76–83.

Thangali, A., Nash, J. P., Sclaroff, S., & Neidle, C. (2011). Exploiting phonological constraints for handshape inference in asl video. In *Cvpr 2011* (pp. 521–528).

Tian, Y., Yang, G., Wang, Z., Wang, H., Li, E., & Liang, Z. (2019). Apple detection during different growth stages in orchards using the improved yolo-v3 model. *Computers and electronics in agriculture*, *157*, 417–426.

Tome, D., Toso, M., Agapito, L., & Russell, C. (2018). Rethinking pose in 3d: Multi-stage refinement and recovery for markerless motion capture. In *2018 international conference on 3d vision (3dv)* (pp. 474–483).

Toshev, A., & Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1653–1660).

Tyagi, A. K., & Abraham, A. (2022). Recurrent neural networks: Concepts and applications.

Varadaraju, A. T. (2013). *Exploiting phonological constraints for handshape* *recognition in sign language video* (Unpublished doctoral dissertation). Boston University.

Vargas-Irwin, C. E., Shakhnarovich, G., Yadollahpour, P., Mislow, J. M., Black, M. J., & Donoghue, J. P. (2010). Decoding complete reach and grasp actions from local primary motor cortex populations. *Journal of neuroscience*, *30*(29), 9659–9669.

Voloh, B., Eisenreich, B. R., Maisson, D. J., Ebitz, R. B., Park, H. S., Hayden, B. Y., & Zimmermann, J. (2023). Hierarchical organization of rhesus macaque behavior. *Oxford open neuroscience*, *2*, kvad006.

Voropaev, V., Magomedov, V., & Alfimtsev, A. (2024). Automatic generation of neurocoomics based on computer vision system data. In *2024 international conference on artificial intelligence, computer, data sciences and applications (acdsa)* (pp. 1–6).

Wang, C., & Yan, J. (2023). A comprehensive survey of rgb-based and skeleton-based human action recognition. *IEEE Access*.

Wang, C., Zhang, F., & Ge, S. S. (2021). A comprehensive survey on 2d multi-person pose estimation methods. *Engineering Applications of Artificial Intelligence*, *102*, 104260.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2018). Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, *41*(11), 2740–2755.

Wang, Q., Zhang, K., & Asghar, M. A. (2022). Skeleton-based st-gcn for human action recognition with extended skeleton graph and partitioning strategy. *IEEE Access*, *10*, 41403–41410.

Warren, W. C., Harris, R. A., Haukness, M., Fiddes, I. T., Murali, S. C., Fernandes, J., ... others (2020). Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science*, *370*(6523), eabc6617.

Wright, K. R., Mayhew, J. A., Sheeran, L. K., Funkhouser, J. A., Wagner, R. S., Sun, L.-X., & Li, J.-H. (2018). Playing it cool: Characterizing social play, bout termination, and candidate play signals of juvenile and infant tibetan macaques (macaca thibetana). *Zoological research*, *39*(4), 272.

Wu, D., Lv, S., Jiang, M., & Song, H. (2020). Using channel pruning-based yolo v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments. *Computers and Electronics in Agriculture*, *178*, 105742.

Wu, W., Xu, M., Liang, Q., Mei, L., & Peng, Y. (2020). Multi-camera 3d ball tracking framework for sports video. *IET Image Processing*, *14*(15), 3751–3761.

Xiao, F., Lee, Y. J., Grauman, K., Malik, J., & Feichtenhofer, C. (2020).

Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*.

Xue, P., & Deng, S. (2023). Smoothness-based consistency learning for macaque pose estimation. *Signal, Image and Video Processing*, *17*(8), 4327–4335.

Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 32).

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316.

Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4694–4702).

Yun, S., Oh, S. J., Heo, B., Han, D., & Kim, J. (2020). Videomix: Rethinking data augmentation for video classification. *arXiv preprint arXiv:2012.03457*.

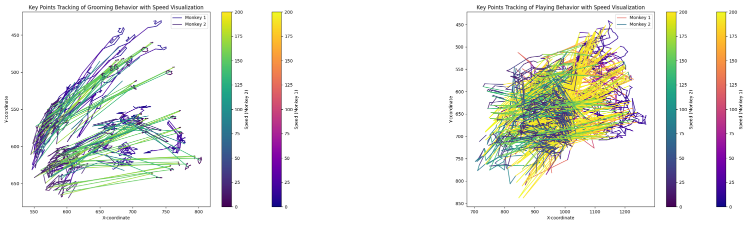
Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., . . . Shah, M. (2023). Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, *56*(1), 1–37.

APPENDIX A

12 BEST HYPERPARAMETERS FOR POSE ESTIMATION MODEL

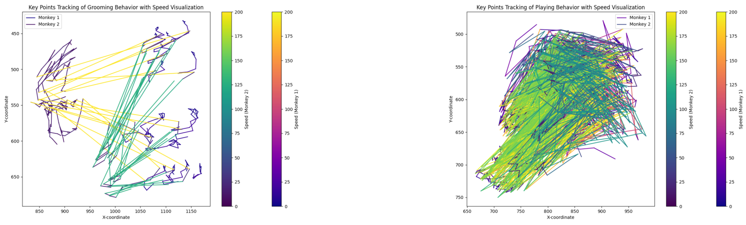
13 SPEED ANALYSIS

One might think that the mere speed at which the pixels or key points change their location from one frame to another can give an indication on the action label because grooming usually consists of small, calm movements of the wrists, while playing is generally composed of large changes over all key points (Figure 16). However, speed alone is not always a reliable indicator, as shown in Figure 17. Therefore, accurate action recognition requires the model to learn more complex patterns from the data.



(a) Grooming (b) Playing

Figure 16: An example of tracked key points over time where the speed of the key points is an adequate indicator of the action class.



(a) Grooming (b) Playing

Figure 17: An example of tracked key points over time where the speed of the key points is not an adequate indicator of the action class.

Name	Description	Value
weight_decay	Weight decay (L2 regularization)	0.00035
warmup_epochs	Number of epochs for warmup phase	2.43996
warmup_momentum	Initial momentum during warmup	0.39855
box	Box regression loss weight	6.11354
cls	Classification loss weight	0.66992
dfl	Distribution focal loss weight	1.90914
hsv_h	Hue augmentation factor	0.00832
hsv_s	Saturation augmentation factor	0.56134
hsv_v	Value augmentation factor	0.38388
degrees	Rotation augmentation degrees	0.0
translate	Translation augmentation factor	0.10243
scale	Scaling augmentation factor	0.13946
shear	Shear augmentation factor	0.0
perspective	Perspective augmentation factor	0.0
flipud	Vertical flip probability	0.0
fliplr	Horizontal flip probability	0.41099
mosaic	Mosaic augmentation factor	1.0
mixup	Mixup augmentation factor	0.0
copy_paste	Copy-paste augmentation factor	0.0

Table 5: Hyperparameter tuning results