KYUTECH-LSSE-19899034

Doctoral Dissertation

Multi-View Animal Behaviour Analysis with Deep Learning

Salvador BLANCO NEGRETE

September 2023

Department of Life Science and Systems Engineering
Graduate School of Life Science and Systems Engineering
Kyushu Institute of Technology

A Doctoral Dissertation
submitted to Graduate School of Life Science and Systems Engineering,
Kyushu Institute of Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Salvador BLANCO NEGRETE

Thesis Committee:

Professor Hiroaki WAGATSUMA	(Co-supervisor)
Professor Tomohiro SHIBATA	(Supervisor)
Professor Keiichi HORIO	(Co-supervisor)
Associate Professor Hideaki KAWANO	

Multi-View Animal Behaviour Analysis with Deep Learning*

Salvador BLANCO NEGRETE

Abstract

Understanding and analyzing animal behaviour is essential in various scientific disciplines, including neuroscience, psychology, ecology, genetics, and pharmacology. Automated detection and analysis of animal behaviours can significantly enhance research efficiency and accuracy in these areas. However, existing systems often rely on engineered features and are restricted to single-view analysis. This dissertation presents a novel approach to multi-view animal behaviour detection using deep learning that does not require pose estimation or other engineered features, works with small amounts of data, and is flexible. This dissertation introduces two primary contributions. The first involves adapting state-of-the-art human pose estimation systems for animal pose estimation. A multiple-monkey pose estimation system and a multi-view 3D pose estimation for marmosets are introduced. The objective is to demonstrate the challenges and limitations of using pose estimation and feature engineering in general for behaviour analysis. The "In the Wild" dataset, known as MacaquePose, was collected containing monkey images in various environments to train a deep neural network for 2D multi-monkey pose estimation. At the same time, the multi-view 3D marmoset dataset was collected in a laboratory setting. The second core contribution presents a novel multi-view behaviour detection system. The system captures behaviours using various perspectives, allowing a comprehensive understanding of the animal's movements. The system uses three neural networks; the first neural network (NN1) extracts Regions of Interest (ROIs) for each view, and NN2 is a classification network that creates a heat map that encodes confidence for the desired

*Doctoral Dissertation, Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, KYUTECH-LSSE-19899034, February 20, 2022.

behaviour across views and time within a time window. NN3 is trained to create a final prediction from the heat map. This approach allows the use of small amounts of data; it avoids using pose estimation or other engineered features, and the three networks can be trained separately or reused, making it easy to adapt to new behaviours or animals. The developed system was trained to detect rats' Wet Dog Shake (WDS) behaviour, making it the first to detect WDS behaviour. The WDS behaviour is relevant in studying various animal disease models, including acute seizures, morphine abstinence, and nicotine withdrawal. This behaviour has a short duration; it occurs spontaneously and infrequently, making it challenging to detect and analyze accurately. Still, using three views, the developed system can detect it with a precision of 0.91 and recall of 0.86. Notably, while this dissertation talks mainly about detecting WDS behaviour in rats, the developed multi-view deep learning system holds potential for broader applications in analyzing various animal behaviours. Multiple camera views enhance the system's ability to generalize across perspectives, add redundancy, and reduce occlusion, resulting in more accurate and robust behaviour detection. Therefore, it can be adapted and extended to detect and analyze other animal behaviours in diverse species. In conclusion, this dissertation presents a novel approach to multi-view animal behaviour detection using deep learning. The developed system opens new avenues for research in animal behaviour and welfare, and its potential applicability across diverse species makes it a valuable tool for studying complex animal behaviours.

Keywords:

Animal Behaviour, Nonhuman Primates pose estimation, Rat Kainate model, Wet dog shaking detection, Deep Learning

Contents

List of Figures

List of Tables

1 Introduction

The study of animal behaviour is fundamental to understanding the intricacies of animal cognition, communication, and social dynamics. It plays a vital role in various fields, including neuroscience, psychology, ecology, genetics, and pharmacology. Traditionally, animal behaviour analysis has relied on manual observation and annotation, which can be time-consuming, subjective, and prone to human error [1]. Advances in computer vision and machine learning, particularly deep learning, have provided researchers with new tools and methods for automating the analysis of animal behaviour, leading to more accurate and efficient research outcomes [2].

One challenge in animal behaviour analysis is accurately detecting and recognizing specific behaviours using a single camera perspective. Although multi-view systems like DeepLabCut's markerless pose estimation offer multiple perspectives, their scope is limited to estimating the pose of a single individual [3]. Developing advanced multi-view techniques that account for complex group dynamics and interactions between multiple animals is crucial for a comprehensive understanding of animal behaviour across diverse contexts and species [4]. This dissertation's main contribution is the development of a novel multi-view deep learning framework tailored for the analysis of animal behaviour, exemplified by its application to the detection of wet dog shake behaviour in rats. Additionally, this work also encompasses the adaptation of the OpenPose [5] algorithm for Multiple-Monkey pose estimation, the curation of an ample dataset featuring macaque monkey images, and the adaptation of a state-of-the-art human multi-view 3D pose estimation method for implementation with marmosets; these endeavours with non-human primates serve to highlight the challenges associated with the utilization pose and other engineered features in general, for behaviour detection.

1.1 Motivation of the Research

The motivation behind this research lies in the potential benefits of leveraging multi-view deep learning techniques for animal behaviour analysis. By utilizing multiple cameras to capture animal movements from different perspectives, a more complete and detailed representation of animal behaviour can be obtained, enabling researchers to gain a deeper understanding of complex behaviours and their underlying mechanisms [3]. This enhanced understanding can lead to new discoveries and insights into animal cognition, communication, and social dynamics, ultimately contributing to more informed decisions and actions in areas such as psychology, neuroscience, pharmacology, biology, conservation, animal welfare, human health and others.

Moreover, the application of deep learning techniques to animal behaviour analysis opens up new avenues for research in the development of innovative therapies and interventions for animal and human health [6]. The ability to automatically detect and analyze specific behaviours, such as the wet dog shaking behaviour in rats, can significantly impact the study of various animal disease models, including acute seizures, morphine abstinence, and nicotine withdrawal [7, 8, 9]. Early detection of behavioural abnormalities can inform targeted interventions and improve animal welfare outcomes. Another key motivation for this research is the need for efficient and accurate pose estimation in non-human primates. Presently, pose estimation techniques predominantly center around human subjects, leaving a gap in the analysis of non-human primate behaviour. By adapting state-of-the-art human pose estimation systems for non-human primates, curating a comprehensive dataset of macaque monkey images, and a multi-camera 3D dataset for marmosets, this research aims to provide resources to scholars working in the study of animal behaviour.

1.2 Thesis Overview

This dissertation seeks to answer the following research questions:

- How can advances in multi-view classification and multi-view human activity recognition be adapted to improve the accuracy and robustness of animal behaviour analysis?
- What specific challenges unique to animal behaviour analysis arise when integrating multi-view, and how can deep learning models be adapted to address these challenges?

This dissertation takes inspiration from the fields of multi-view classification and multi-view human activity recognition, incorporating and refining techniques to address the unique challenges inherent in animal behaviour analysis. Specifically, it investigates the unique challenges posed by multi-view integration in the animal behaviour analysis context, it explores optimal deep learning models and presents a novel multi-view behaviour detection system designed for animals. This system is tested on the Wet Dog Shake (WDS) behaviour in rats.

Chapter 1 provides the motivation for this research endeavour. Chapter 2 introduces animal behaviour analysis, multi-view approaches, and deep learning models in the context of animal behaviour analysis, reviewing the existing literature on multi-view classification, multi-view human activity recognition and its potential applicability to animal behaviour analysis. Chapter 3 outlines the multi-camera configurations. It covers the hardware setup of the Marmoset Multi-view Capture System and the Rat Multi-view Capture System and includes details of the animal preparation. Chapter 4 Details the methodology, covering data collection, processing, and deep-learning models used to address the challenges of animal behaviour analysis. Chapter 5 presents experimental results within the multi-view animal behaviour analysis context. Finally, Chapter 6 summarizes research findings and outlines future research directions.

2 Related Work

This chapter provides a comprehensive overview of the field of animal behaviour, with an emphasis on multi-view behaviour detection and pose estimation through computer vision and deep learning methodologies. By presenting the current scientific landscape, it highlights the challenges and limitations of traditional approaches. It pinpoints the gaps in the existing literature, mainly the limitations in single-view systems, that this dissertation seeks to fill. The overview of prior research permits taking inspiration from the advances in multi-view classification and multi-view human activity recognition; it also highlights the challenges associated with pose estimation for behaviour detection.

2.1 Animal Behaviour Analysis

The study of animal behaviour is crucial for understanding the intricacies of animal cognition, communication, and social dynamics. Animal behaviour research spans various fields, including neuroscience, psychology, ecology, genetics, and pharmacology [1]. Traditional methods for studying animal behaviour have relied on manual observation and annotation, but recent advances in computer vision and machine learning have provided more accurate and efficient alternatives. These technological advancements have made it possible to automate the analysis of animal behaviour and derive novel insights that were previously difficult to uncover using traditional techniques.

2.1.1 Manual Observation and Annotation

Manual observation and annotation have been the backbone of animal behaviour analysis for decades. Researchers observe animals either in the field or in controlled environments and manually record their behaviours [10]. While these

methods have provided valuable insights, they come with several limitations and challenges. Manual annotation can be time-consuming, subjective, and prone to human error [11]. Additionally, observer bias may influence the results, and the quality of annotations can vary significantly between individuals [12]. The limited scope of manual observations may also hinder the detection of rare or subtle behaviours, and the inability to scale to large datasets constrains the potential for comprehensive analysis [1].

These challenges have motivated the development of more accurate and automated techniques for animal behaviour analysis. With advancements in technology, researchers have started exploring new methods to address the limitations associated with manual observation and annotation. These innovative approaches include using computer vision and machine learning techniques, which provide more efficient and accurate ways to analyze animal behaviour. [13, 4, 3].

2.1.2 Computer Vision and Machine Learning Techniques

With the advent of computer vision and machine learning, researchers have begun to leverage these technologies for animal behaviour analysis. Computer vision techniques enable the extraction of visual information from images or videos, while machine learning algorithms can learn patterns in this data, allowing for the automatic detection and recognition of specific behaviours. This shift towards automated methods has significantly improved the accuracy and efficiency of animal behaviour analysis, enabling researchers to study larger datasets and uncover new insights [13].

Machine learning methods, including support vector machines (SVMs), principal component analysis (PCAs), and various other techniques, have been applied to animal behaviour analysis [14]. However, the recent rise of deep learning, a subfield of machine learning, has revolutionized the field. Deep learning algorithms, especially convolutional neural networks (CNNs), have achieved remarkable results in image recognition and classification tasks, making them well-suited for animal behaviour analysis [15, 16].

Deep learning-based approaches have been applied to various aspects of animal behaviour analysis, including pose estimation, action recognition, and tracking [17]. These methods offer the potential to overcome many limitations of man-

ual observation and annotation, allowing for more accurate and efficient analysis of animal behaviour. For example, pose estimation techniques can provide detailed information on an animal's posture and movement, which can be crucial for understanding specific behaviours [3].

Despite the promising results of deep learning in animal behaviour analysis, challenges still need to be addressed, such as generalization across different species and behaviours and the limitations of single-view systems [4]. Furthermore, the need for large annotated datasets to train deep learning models can be a barrier to entry for some researchers [18], and the interpretability of these models remains an open question [19].

To address some of these challenges, some researchers have started exploring using multi-view systems, which combine data from multiple cameras to provide a more comprehensive understanding of animal behaviour. By leveraging the power of multi-view deep learning techniques, it is possible to overcome some of the limitations of traditional single-view systems and push the boundaries of animal behaviour analysis even further [3, 4].

2.2 Single-view Pose Estimation

Single-view pose estimation refers to the process of estimating the position and orientation of an object or subject in a single image or video frame. In the context of animal behaviour analysis, pose estimation is critical for understanding the movements and actions of animals, as it allows researchers to track and analyze specific body parts or landmarks. While single-view pose estimation has been widely employed in various applications, it has several limitations when applied to animal behaviour analysis. These limitations include the inability to handle occlusions, variations in appearance, and changes in scale and perspective [20].

2.2.1 Human Pose Estimation

Human pose estimation is a well-established field within computer vision, focusing on determining the positions and orientations of human body parts in images or videos. Over the past decade, deep learning-based methods have become the state-of-the-art in human pose estimation, significantly improving the accuracy

and efficiency of the process [21]. These methods typically employ convolutional neural networks (CNNs) to learn and infer the spatial relationships between body parts. Popular approaches for human pose estimation include OpenPose [5], which employs a multi-stage CNN architecture and Part Affinity Fields (PAFs) to detect body part locations and estimate their associations, and DeeperCut [22], which leverages a deeper architecture and image-conditioned pairwise terms for improved pose estimation. HRNet [23] stands out as one of the best pose estimators, utilizing a high-resolution network architecture to maintain fine-grained spatial information throughout the network, resulting in highly accurate pose estimation across different scales.

2.2.2 Non-human Primate Pose Estimation

Despite the progress made in human pose estimation, the development of pose estimation methods for non-human primates has been relatively limited. Most existing methods are adaptations of human pose estimation techniques [24]. Moreover, the availability of annotated datasets for non-human primates is scarce, making it challenging to train and validate pose estimation models for these animals. In this thesis, the MacaquePose dataset was developed to address the lack of available annotated datasets for non-human primates. There is still a need for further research and development in this area, including expanding the variety of non-human primate species studied, creating larger and more diverse datasets, and exploring new model architectures and training strategies. Additionally, current pose estimation methods for non-human primates face limitations in handling extreme occlusions and large variations in scale, which should be addressed in future research.

2.3 Multi-view Behaviour Analysis

Multi-view behaviour analysis aims to capture and analyze the actions and movements of subjects from multiple perspectives simultaneously. This approach addresses the limitations of single-view analysis, providing a more comprehensive understanding of the subject's behaviour and improving the accuracy of pose estimation and action recognition.

2.3.1 Multi-view Image Classification

Multi-view image classification involves the integration of information from different views or modalities to improve the classification of images. This technique is widely used in various applications, including object recognition, scene understanding, and biomedical image analysis. Multi-view classification methods typically rely on feature extraction and fusion techniques, which combine information from multiple views to create a more robust and accurate representation of the subject [25]. Moving from multi-view image classification to more complex scenarios, the analysis of human actions in multi-view settings has gained considerable attention. In the following subsection, the focus is on how multi-view systems have been applied to human action recognition, with a particular emphasis on 3D pose estimation.

2.3.2 Multi-view Human Action Recognition

Multi-view human action recognition has received significant attention due to its potential applications in surveillance, sports analysis, and human-computer interaction [26]. By leveraging multiple camera perspectives, this approach can better handle occlusions and variations in appearance, scale, and viewpoint. In multi-view human action recognition (MVHAR), two primary approaches have been explored. The first approach involves training end-to-end neural networks, as demonstrated by previous works such as [27, 28, 29]. However, the effectiveness of this approach heavily relies on the availability of large MVHAR datasets, which are commonly used in human action recognition research [30, 31, 32, 33, 34]. Unfortunately, equivalent datasets for animal behavior analysis are not readily accessible, and producing them would incur substantial costs and efforts.

The second approach in MVHAR involves feature extraction, with skeleton features being widely adopted in existing studies [35, 36]. However, utilizing skeleton features requires the use of pose estimators, which demand extensive amounts of labeled data for accurate training [22, 5]. The challenge lies in accurately labeling animal data, as it requires a deep understanding of the target animal's anatomy to produce correct and meaningful labels [18].

In summary, both approaches present unique challenges when applied to an-

imal behavior analysis. While end-to-end neural networks require large animal datasets that are currently scarce, using skeleton features necessitates precise pose estimators and specialized labeling expertise. As part of this research, we have sought to address these challenges by developing a multi-view behavior classification system that overcomes the limitations of both approaches, enabling efficient and accurate behavior analysis in animals without extensive reliance on large datasets or complex feature engineering.

Multi-view 3D Pose Estimation

Multi-view 3D pose estimation focuses on estimating the 3D positions and orientations of body parts or landmarks from multiple camera views. By fusing information from different perspectives, it aims to provide a more accurate and complete representation of the subject's movements. Several approaches have been proposed for multi-view 3D pose estimation in human action recognition. For instance, the EpipolarPose framework [37] leverages the epipolar geometry constraints between different camera views to improve 3D pose estimation accuracy.

Another notable method is VoxelPose [20], which employs a bottom-up approach to predict 3D human poses in a voxelized space. VoxelPose uses a two-stage process that first detects 2D human body confidence maps and then projects these confidence maps from all views into a 3D space to estimate individual human poses in 3D. By leveraging volumetric representations and fusing information from multiple views, VoxelPose can achieve improved performance in 3D human pose estimation tasks.

Despite the advancements in multi-view 3D pose estimation for human action recognition, there is still limited research on applying these techniques to animal behaviour analysis. The following section will discuss the current state of multi-view systems in animal behaviour analysis and the potential for further research in this area.

2.3.3 Multi-view Systems in Animal Behaviour Analysis

Despite the advancements in multi-view human action recognition, the application of multi-view systems in animal behaviour analysis remains relatively limited [3]. Challenges include the need for accurate pose estimation in non-human primates and the detection of specific behaviours, such as the wet dog shaking behaviour in rats. Some recent studies have begun to explore the potential of multi-view systems for animal behaviour analysis, demonstrating the benefits of this approach in terms of improved accuracy, robustness, and ability to handle occlusions [4, 38]. However, further research is needed to develop and validate multi-view systems for various animal species and behaviours, ultimately enhancing our understanding of complex animal behaviours and their underlying mechanisms.

2.4 Wet Dog Shaking Behaviour

Wet dog shaking (WDS) is a distinctive type of behaviour exhibited by a diverse range of animals, such as rodents, African lions, and giant pandas, characterized by rapid whole-body shaking motions. These movements involve the rapid oscillation of the animal's body, generating forces that propel water droplets away from its fur or skin, similar to a wet dog's actions to remove water from its fur [39]. This behaviour has been observed across different species and environments, suggesting an adaptive and functional significance. The shaking motions may serve various purposes depending on the species and context. Studying WDS in depth could reveal valuable insights into the complex interactions between animals and their environments.

2.4.1 Wet Dog Shaking Behaviour in Rats

In the context of rats, WDS behaviour holds particular significance for studying animal disease models, including acute seizures, morphine abstinence, and nicotine withdrawal, as it provides valuable insights into their underlying mechanisms [7, 8, 9]. Triggered by factors such as drug administration, cold water exposure, or tactile stimulation around the ears, this behaviour manifests as an abrupt,

convulsive shudder of the head, neck, and trunk, which mirrors the movements seen in dogs. The neuropharmacology of WDS is still uncertain, but evidence points to the possible involvement of central serotonin (5-HT) pathways [40].

Similar to how other behaviours, such as self-grooming, have been employed as a tool for understanding neurological and psychiatric disorders in neuroscience[41, 42], WDS also holds the potential for advancing our knowledge of various animal disease models. Naturally occurring WDS is rare and nearly absent when observed over short periods [43]. However, it becomes more prevalent under specific experimental conditions, such as the Rat Kainate (KA) model, as described below.

Rat Kainate Model

The Rat Kainate (KA) model is an animal model of temporal lobe epilepsy induced by the administration of kainic acid, a potent neurotoxin. This model is characterized by an overexpression of WDS behaviour, making it an ideal subject for the study of WDS detection methods. Soon after the administration of KA, rats experience an unusually high amount of WDS for approximately one hour until class IV and/or class V seizures appear [44, 45]. In the three experiments used for this study, the animals experienced an average of 139 WDS events during the hour of high WDS activity after KA administration. The average duration was 0.33s with a standard deviation of 0.11s.

Figure 2 presents a time-lapse of WDS behaviour recorded in this study (t=0s previous to the start WDS behaviour, t=0.1s-0.4s WDS behaviour, t=0.5s ending of WDS behaviour). This behaviour is evocative of the movement seen in dogs.

The Rat KA model is the foundation for developing and testing a multi-view system to detect WDS automatically in this dissertation. This novel approach could provide valuable insights by analysing WDS behaviour and its implications in various animal disease models.

2.4.2 Traditional Detection Methods

Traditional detection methods for WDS behaviour in rats predominantly depend on manual observation and annotation by researchers [46, 47]. While these meth-

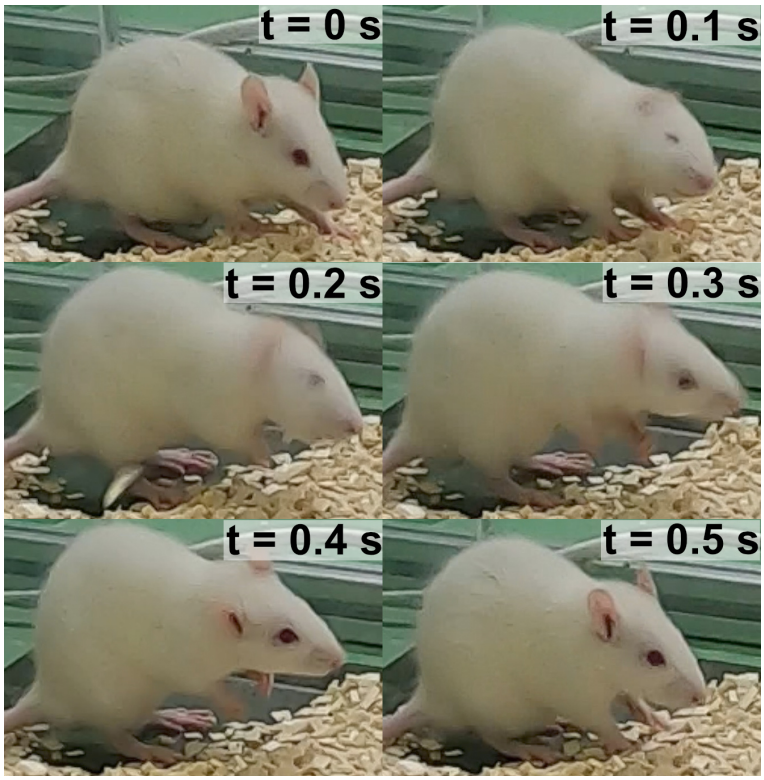

Figure 2.1: A time-lapse of WDS behaviour recorded in this study (t=0s previous to the start WDS behaviour, t=0.1s-0.4s WDS behaviour, t=0.5s ending of WDS behaviour).

ods can yield effective results, they come with several drawbacks, including the time-consuming nature of the process, the subjectivity introduced by human judgment, and the potential for errors due to human oversight. Moreover, continuously monitoring WDS behaviour can prove challenging, particularly in long-term studies where researchers may be required to observe and annotate numerous instances over extended periods.

Observer fatigue, resulting from the repetitive and labour-intensive nature of manual observation, can compromise the accuracy of the annotations. Notably, for WDS, the common technique of fast-forwarding video to analyze behaviours is not feasible due to the short duration of WDS episodes, which in the experiments performed for this work had an average duration of 0.33 seconds. Fast-forwarding video is a method often employed to review and identify behavioural events of

interest more quickly by speeding up the playback of the recorded video [48]. However, since WDS episodes are so brief, this fast-forwarding technique would likely result in researchers missing the crucial WDS events, making it unsuitable for detecting such short-duration behaviour.

Additionally, variability between different or even the same observer at different times may lead to inconsistencies in the data collected. This variation can hinder the reliability of the conclusions drawn from the analysis of WDS behaviour and limit the reproducibility of the study results. Consequently, there is a need for more efficient, objective, and reliable methods of detecting and analyzing WDS behaviour in rats to overcome these challenges and enhance the quality of research in this area.

2.4.3 WDS Deep Learning-based Detection Techniques

Deep learning techniques have shown promise in various behavioural detection tasks, demonstrating their potential for automating and enhancing the analysis of complex behaviours [13]. However, to date, there is no established system for detecting WDS behaviour in rats using deep learning methods. This dissertation aims to address this gap by developing a novel deep learning-based detection system for WDS behaviour in rats, leveraging multi-view camera systems to improve accuracy and robustness. The development of such a system would contribute to the advancement of animal behaviour analysis and significantly impact the study of animal disease models.

3 Multi-Camera Setups and Animal Preparation

This chapter describes the multi-camera setups and systems used for marmoset 3D pose estimation and wet dog shaking detection in rats. It also includes information on the preparation of the rats for the recordings.

3.1 Marmoset Multi-view Capture System

The Marmoset Multi-view Capture System was designed to record the behaviour of marmosets in a controlled environment using multiple synchronized cameras. The system consists of eight high-resolution cameras strategically placed around a 1.5 x 1.5-meter marmoset enclosure to ensure comprehensive coverage of the entire area Figure 3.1. Each camera in the system is equipped with a high-quality lens, offering a wide field of view to maximize the coverage of the marmoset's activities within the enclosure. The cameras can capture images at a resolution of 2048 x 1536 px and a frame rate of 24 frames per second, providing detailed and smooth video recordings of the animals' movements and interactions. A dedicated hardware-based synchronization system was employed to ensure precise synchronization of the cameras. This system uses a central control unit to distribute accurate timing signals to each camera to ensure synchronization and consistency across the various viewpoints. This synchronization is critical for the precise reconstruction of 3D poses and the analysis of marmoset behaviour from the multi-view recordings. The marmoset enclosure is designed to provide a comfortable and stimulating environment for the animals while facilitating the capture of various complex naturalistic behaviours in high-quality video data. The enclosure includes features such as two pole structures for climbing and feeding

stations, encouraging natural marmoset behaviours and social interactions. The enclosure acrylic walls are constructed using non-reflective materials to minimize glare and reflections that might interfere with the video quality. The captured video data from the eight cameras are stored on a high-capacity, multi-channel recording system, which allows for simultaneous recording and playback of video data from all cameras. This feature enables researchers to review the captured footage in real-time, allowing for quick adjustments to the experimental setup if necessary.

Figure 3.1: A collage of the eight synchronized camera views capturing the marmosets within their 1.5 x 1.5-meter enclosure. The images have been selectively cropped to emphasize the marmosets and showcase the diverse perspectives provided by the multi-view capture system.

3.2 Rat Multi-view Capture System

The experiments involving rats were conducted in accordance with the Guide for Care and Use of Laboratory Animals at the Graduate School of Life Science and Systems Engineering of the Kyushu Institute of Technology (Sei#2021-003).

3.2.1 Animal Preparation

The experiments were performed with three rats aged 4–5 weeks and weighing between 104.0 and 152.5 grams (Japan SLC inc). There were eight to ten days of adaptation before the experiment. Light (12 h light - 12 h dark), humidity (50±5), and temperature (23±1) factors were regulated. Administering 0.05% kainic acid (KA) (5mg/kg) after anesthesia (3.5 isoflurane) intraperitoneally using the repeated low-dose protocol (Hellier et al., 1998). The injections were administered every hour, three times in total. The animals were recorded for one hour immediately after the third KA injection.

3.2.2 Hardware Setup

The live mouse tracker 50 × 50 cm^2 plastic cage [13] was utilized for the experiments. To create a machine learning dataset, three sessions were recorded with up to four cameras: a camcorder, a GoPro HERO8 Black, and two GoPro HERO7 Silver cameras. A camera was set on each side of the cage. The camcorder was mounted using a tripod, as this is a common setup for manual labelling [42]. The GoPro cameras were attached to the plastic enclosure using a suction cup mount and positioned at the center top of the panel as illustrated in Figure 3.2. The angle was adjusted to capture the entire enclosure. The angle and position of the cameras changed slightly between experiments, as the cameras were mounted and dismounted between different sessions. All cameras were set to a resolution of 1080p at 30 fps. The GoPro HERO8 features three digital lenses; the wide digital lens was used. Videos were synchronized using Apple Final Cut Pro Multicam editing workflow.

The study was conducted using the cloud service Google Colab with a Tesla V100-SXM2-16GB GPU graphics card and an Intel (R) Xeon (R) CPU @ 2.20GHz hardware configuration.

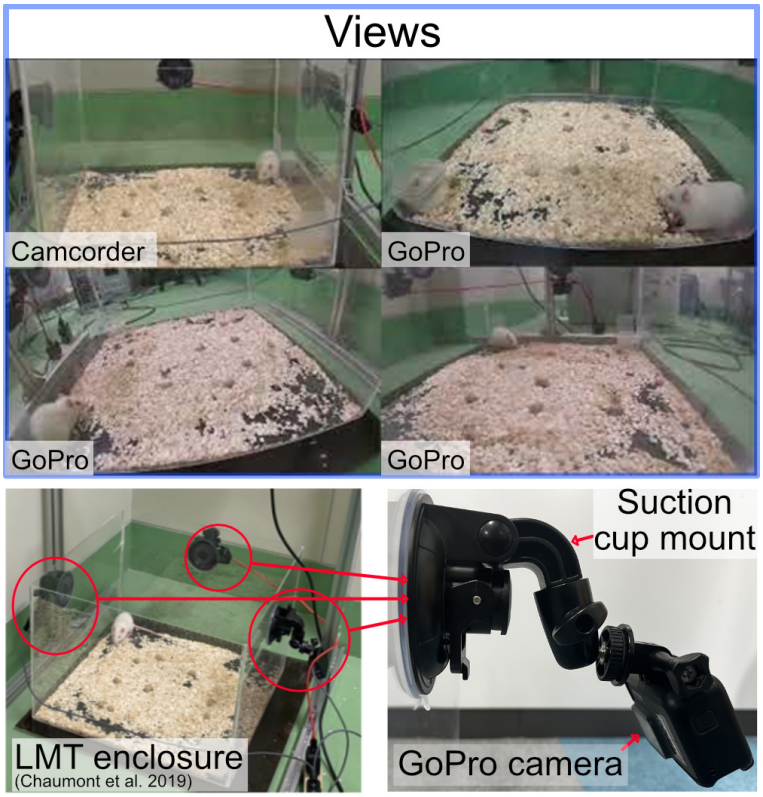

Figure 3.2: Top: the rat being depicted from different perspectives. Left bottom: the open-source LMT Acrylic enclosure. Bottom right: GoPro Camera attached by a suction cup mount.

4 Methodology

4.1 Multiple-Monkey 2D Pose Estimation

4.1.1 Dataset and Annotations

The MacaquePose monkey dataset, utilized for multiple-monkey 2D pose estimation, comprises 13,083 RGB images featuring 16,393 monkeys that were used to train the model [18]. These images, either sourced from the internet through Google Open Images, captured in zoos, or taken at the Primate Research Institute of Kyoto University, depict monkeys engaging in various activities and social interactions, including feeding, grooming, sleeping, fighting, climbing, and swimming. The dataset presents numerous challenges for pose estimation, such as diverse backgrounds, occlusions, noise, and other artefacts. Of all the images, 93.75% were allocated for training and the remaining 6.25% for validation. Table 1 provides further details on the dataset's distribution.

Each monkey in the dataset is annotated with 17 labels corresponding to different body features, as illustrated in Figure 4.1. If a body feature is present, its (x,y) coordinates are noted, and an additional bit indicates whether the body feature is visible or occluded.

4.1.2 Network Architecture

HyperPose, an open-source implementation, is employed as part of the TensorLayer project [49]. TensorLayer is integrated with Google's TensorFlow Framework for machine learning and deep learning. In this study, ResNet18 is used as the network's backbone [50]. Although employing a deeper backbone like ResNet50 might yield better results, ResNet18 allows the network to remain lightweight. This reduces the training and execution time and computational

Figure 4.1: Top: Examples of pictures and labels in the present dataset.

requirements, making it more accessible. Following the same rationale, the network's head is a lightweight implementation of OpenPose [5]. The network takes body feature-related labels and masks for each monkey as input. As an output, the network generates score maps and estimates preliminary part affinity fields. After a refinement stage, the monkeys' skeletons are generated. Once the model is trained, it is exported to the Open Neural Network Exchange (ONNX) format, allowing it to run in the Nvidia C++ TensorRT high-performance deep learning inference framework.

4.1.3 Evaluation and Training

The Average Precision (AP) metric is used to evaluate the model. During training, standard augmentation methods such as random rotation, shifts, and flips were employed. A graph of the training loss is also provided. The network processes images with a maximum dimension of 640 pixels, either in height or width. The model is trained on an Nvidia GeForce GTX TITAN X graphics card for up to 100,000 iterations with a batch size of 8. The training takes approximately 24 hours. The same computer setup is used for evaluation and inference.

4.2 Marmoset 3D Pose Estimation

4.2.1 Dataset and Annotations

The dataset for Marmoset 3D Pose Estimation consisted of 41 short clips, from which random frames were extracted, resulting in a total of 3,000 annotated marmoset bodies. Out of these, 2,700 were used for training, while the remaining 300 were reserved for validation. Both 2D and 3D annotations were provided for each marmoset body in the dataset.

4.2.2 Network Architecture

The network architecture for Marmoset 3D Pose Estimation consisted of two main components: HigherHRNet [51] for 2D pose estimation and VoxelPose [20] for 3D pose estimation as illustrated in Figure 4.2.

HigherHRNet:

HigherHRNet is a state-of-the-art deep learning architecture designed for high-resolution 2D human pose estimation. It adopts a multi-resolution approach to predict human keypoints and generates heatmaps, which represent the likelihood of keypoint locations. The primary advantage of HigherHRNet lies in its ability to maintain high-resolution representations throughout the network, enabling more accurate and precise keypoint localization.

In the context of the Marmoset 3D Pose Estimation, HigherHRNet was adapted to estimate 2D poses for marmosets. It was pre-trained using the MacaquePose dataset, allowing for faster convergence and improved marmoset pose estimation task performance. The output of HigherHRNet consisted of 2D heatmaps representing the likelihood of each marmoset body part's location.

4.2.3 VoxelPose:

VoxelPose is a 3D pose estimation method that leverages the 2D heatmaps generated by the HigherHRNet to estimate the corresponding 3D poses. It works by constructing a 3D heatmap, also known as a voxel grid, from the 2D heatmaps.

The voxel grid is created by lifting the 2D heatmaps into 3D space, followed by the process of matching and associating 2D keypoints across different camera views.

Then a 3D Convolutional Neural Network processes the voxel grid to generate the final 3D pose predictions. The network learns to identify and match keypoints across different views, accounting for occlusions and ambiguities that may arise in the 3D reconstruction process.

By combining HigherHRNet's 2D pose estimation capabilities with VoxelPose's 3D pose estimation approach, the network architecture enables accurate and efficient marmoset 3D pose estimation. This methodology allows the use of the pre-trained MacaquePose dataset, reducing the training data required for the Marmoset 3D Pose Estimation task and improving the system's overall performance.

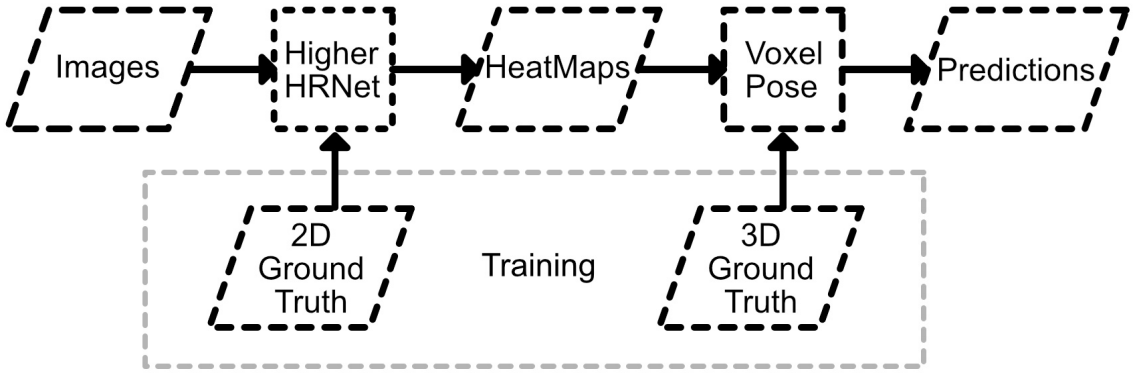

Figure 4.2: 3D Pose Estimation Process. The diagram illustrates the steps involved in the 3D pose estimation process, starting with the 2D pose estimation using HigherHRNet, followed by the generation of 2D heatmaps, and finally, the estimation of 3D poses using VoxelPose.

4.3 Wet Dog Shaking Detection in Rats

In this section, the methodology employed to detect WDS behaviour in rats using the data collected from the multi-view system is detailed.

4.3.1 Data Collection and Annotation

The dataset for machine learning was created using three one-hour recordings captured from 3-4 cameras (views) at a frame rate of 30 fps. These recordings were made following the third KA injection. Two of these recordings were used for training, and one was used for validation. The target behaviour, WDS, is a rapid oscillation of the body that occurs naturally as a spontaneous behaviour [39]. As WDS is a rare behaviour, only KA-treated rats were considered for this study. In the three experiments, the rats experienced 149, 220, and 49 WDS events with a mean duration of 0.33s (approximately 10 frames) and a standard deviation of 0.11s.

Each frame was annotated either as WDS or NWDS (Not Wet Dog Shaking). To create an object localization dataset, image features were first extracted from a sub-sample of frames and then principal component analysis was applied to reduce the dimensions. After clustering the frames using K-means, the images were annotated with the help of cloud services Roboflow and Labelbox. This resulted in a dataset containing 1.5k images for training.

For the WDS classification dataset, the Object Detection Network was used to obtain the Region of Interest (ROI) predictions and the images were cropped accordingly. All frames containing WDS behaviour and an equal number of NWDS behaviour images were selected, yielding a balanced dataset with 25,544 frames.

4.3.2 Neural Networks

The framework consists of three neural networks as illustrated in Figure 4.4: NN1 for object localization, NN2 for image classification, and NN3 for predicting the final score from a feature map that encodes multiple views and time, as illustrated in Figure 4.5. TensorFlow on the Google Colab platform was used for implementation.

Object Localization Network

For object localization, the Single Shot MultiBox Detector (SSD) approach [52] with MobileNetV2 as the backbone [53] was employed. SSD maintains high accuracy while being fast. The SSD consists of a single feed-forward convolutional

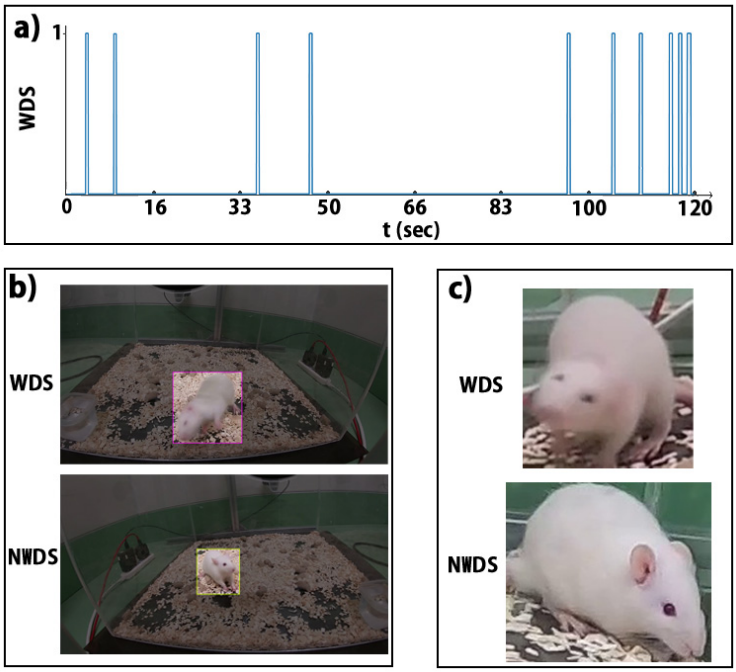

Figure 4.3: Types of Annotations for Wet Dog Shaking Behaviour in Rats. (a) Timestamp-based labelling of rat behaviour. (b) Object detection annotation, including the region of interest and behaviour class label. (c) Cropped image showcasing the rat and its corresponding behaviour class label.

network with three stages. The TensorFlow 2 Detection Model Zoo API was used for implementation.

Image Classification Network

A Convolutional Neural Network (CNN) was trained to classify images into two categories: WDS and NWDS. Image classification can be described as assigning a label from k categories to an image x, represented by the function 4.1:

$$f : x \mapsto \{1, \ldots, k\} \tag{4.1}$$

In this case, $k = 2$, corresponding to WDS and NWDS categories. MobileNetV2 [53] was used as the backbone of the classifier, and the previously

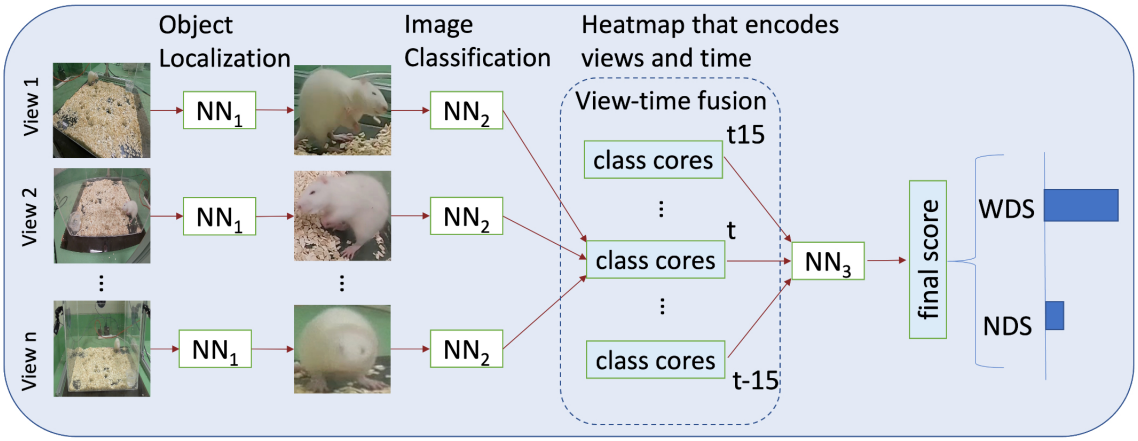

Figure 4.4: Overview of the Wet Dog Shaking Detection Method. The process involves three neural networks: NN1 for object localization, identifying regions of interest within the input images; NN2 for image classification, categorizing the cropped regions based on the presence of WDS behaviour; and NN3 for predicting the final score by analyzing a feature map encoding time and multi-view information, ultimately determining the likelihood of WDS behaviour occurrence.

described WDS classification dataset was used for training.

Multi-view Integration and Time Series Analysis

To account for multiple views and temporal information, the image classification equation was modified as 4.2 and a score fusion multi-view image classification algorithm was employed. A separate network, NN3, was trained to predict the final score from the output of NN2, which generates predictions for each view from a specific time window. NN2 analyses all frames from all views within the time window and creates a vector map of class scores for each camera angle. The time window is user adjusted; in the case of this study, the time window of t-15 to t+15 was selected. This choice aligns with the average duration of the WDS behaviour, 0.33 seconds average, as shown in Figure 4.5. This vector map is then used as input for NN3 to generate the final prediction as illustrated in Figure 4.5.

$$f : X_V \to \{1, \ldots, k\} \mid X_V = \{x_1, x_2, \ldots, x_{n_V}\} \tag{4.2}$$

$$f : X_V^t \to \{1, \ldots, k\} \mid {X_V}^t = \begin{pmatrix} x_1^{t-n} & \cdots & x_1^t & \cdots & x_1^{t+n} \\ x_2^{t-n} & \cdots & x_2^t & \cdots & x_2^{t+n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n_v}^{t-n} & \cdots & x_{n_v}^t & \cdots & x_{n_v}^{t+n} \end{pmatrix} \tag{4.3}$$

In Equation 4.2, f is a function that maps a set of images X_V from n_V views to one of k categories. In Equation 4.3, the function f is extended to incorporate time information, denoted by the superscript t, and multiple views, denoted by the subscript n_V. The matrix represents the images from different views and timepoints, with t being the current time and n indicating the number of time steps before and after the current time.

The NN3 model architecture as illustrated in Figure 6 begins with an input layer, followed by a Conv2D layer with 30 filters and a kernel size of (3, 5). The output is flattened and passed through a Dense layer with 20 units and ReLU activation. Batch Normalization and Dropout layers are employed to improve generalization. Finally, a Dense output layer with 2 units with softmax activation provides classification probabilities for the input data. To further refine and eliminate gaps in the results obtained from the network, we utilize one median and one minimum filter.

Illustrated in Figure 4.6, the NN3 model architecture initiates with an input layer, progressing to a Conv2D layer featuring 30 filters and a kernel size of (3, 5). Subsequently, the output is flattened and directed through a Dense layer comprising 20 units activated by. To enhance generalization and prevent overfitting, Batch Normalization and Dropout layers are incorporated. The final stage involves a Dense output layer, encompassing 2 units with softmax activation, which yields classification probabilities for the input data, enabling precise determination of the target behavior. To further refine the results and rectify gaps, we incorporate both a median filter and a minimum filter. By employing these post-processing filters, the model delivers more coherent and refined behavior classification outcomes, enhancing its overall robustness and accuracy.

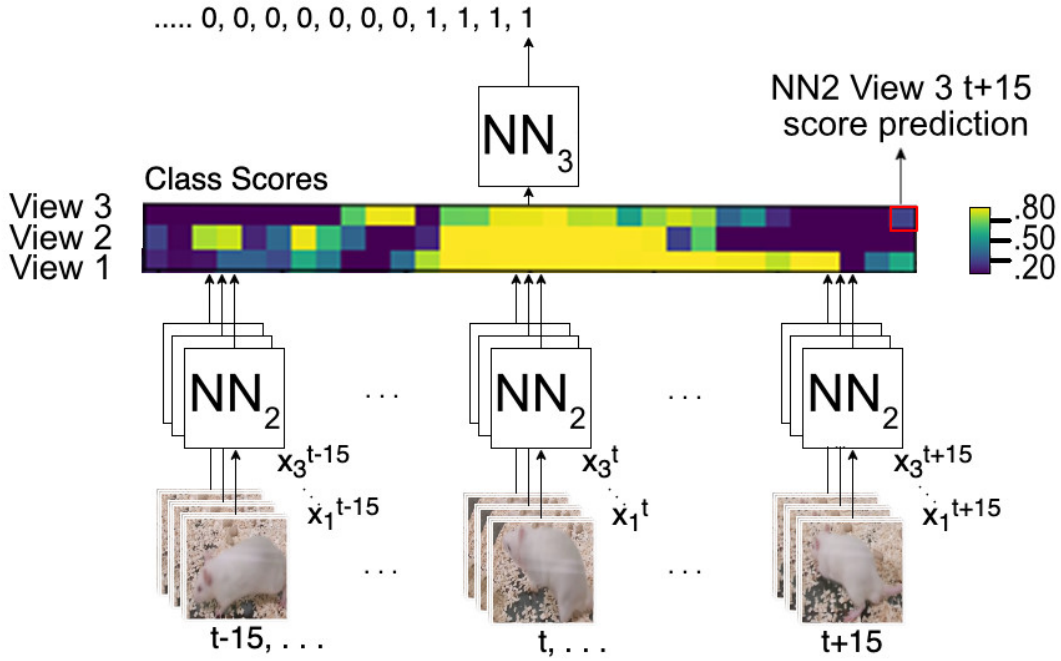

Figure 4.5: Multi-view integration and time series analysis framework. NN2 analyzes all frames from $t-15$ to $t+15$ from all views to create a vector map of class scores. NN3 generates the final prediction from the vector map.

Figure 4.6: Diagram of the NN3 model architecture, which consists of a Conv2D layer, a flatten layer, a Dense layer and ReLU activation, along with Batch

5 Results and Discussion

In this chapter, the results obtained from the experiments conducted for multiple-monkey 2D pose estimation, 3D pose estimation, and wet dog shaking detection in rats are presented. The implications of these results and their potential applications in animal behaviour analysis are discussed.

5.1 Multiple-Monkey 2D Pose Estimation

The model was trained up to 100,000 iterations, taking around 24 hours. Figure 5.1 shows that the loss significantly decreased during the first 10,000 iterations. Afterwards, the loss continues to drop at a slower but steady rate. The graph also compares the training performance of the monkey model to a network trained with the human MSCOCO 2017 dataset [54]. The monkey model's performance is comparable to the human one, and in fact, it performs slightly better. It is remarkable to see the monkey model achieving close results to the human model since the monkey dataset is approximately ten times smaller. The COCO training set contains over 100,000 person-labelled instances. The better results on the monkey model could be explained by the MSCOCO dataset being more challenging. The MSCOCO dataset contains numerous objects and animals in addition to humans, while the monkey dataset primarily consists of monkeys.

5.1.1 Evaluation

To evaluate the performance of the trained model, AP was used as a metric. The monkey-trained model was compared against the same network trained on the MSCOCO2017 human dataset. The results reported on the original OpenPose were also included as a reference [5]. It is important to note that a fair comparison is not possible due to the differences in the Original OpenPose implementation

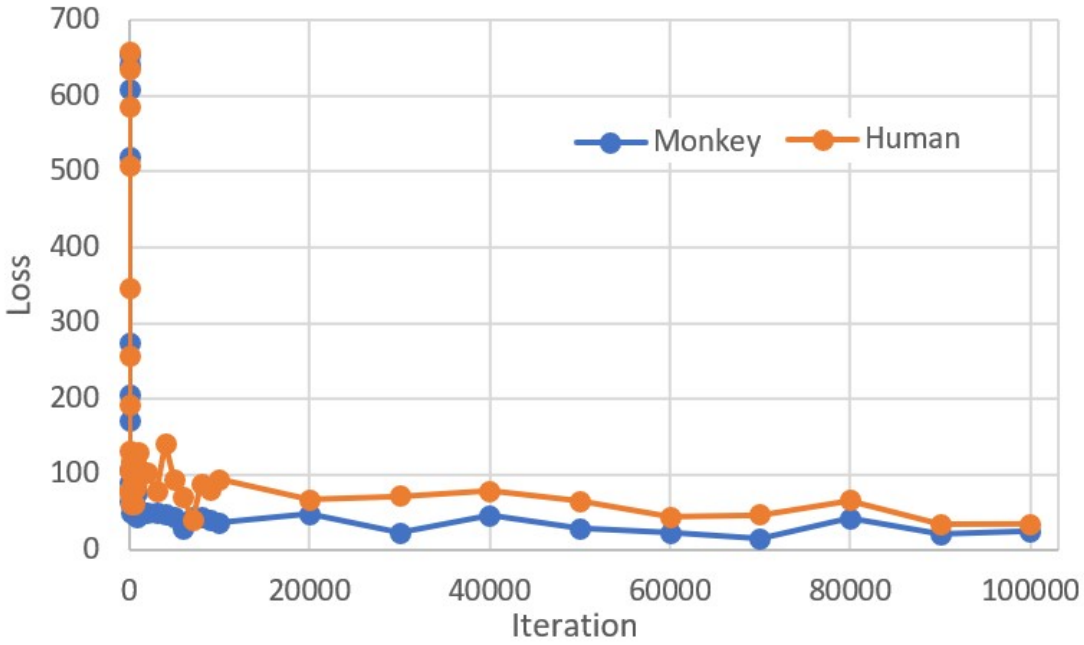

Figure 5.1: Training Loss Comparison for Monkey and Human Models. The plot illustrates the training loss over time for the Monkey model (blue curve) and the Human model (orange curve

Network	Dataset	Ap50	Ap75	ApM	ApL
Lightweight Openpose + ResNet18	**MSCOCO 2017**	**76.2**	**38.0**	**35.5**	**43.2**
Lightweight Openpose + ResNet18	MSCOCO 2017	58.7	10.4	18.2	28.1
Original OpenPose	MSCOCO 2016	83.4	66.4	55.1	68.1

Table 5.1: Performance comparison of the Monkey model and the Human pose trained model using AP50, AP75, APM, and APL metrics. The Monkey model was evaluated on 29 randomly selected images, while the Human pose trained model was evaluated on 11 randomly selected images.

and backbone. The original OpenPose backbone is based on the VGG-19 [55], while the monkey-trained model is based on ResNet18 [50] and uses a lightweight implementation of OpenPose. Additionally, the original OpenPose is evaluated on the MSCOCO2016 dataset.

Table 5.1 shows the results of Ap50, Ap75, ApM, and ApL. Twenty-nine randomly selected images from the evaluation set were chosen to evaluate the Monkey model. Similarly, to evaluate the human pose-trained model, eleven randomly selected images from the evaluation set were selected. Compared to the network trained on humans, the monkey network achieves better performance in all four APs; this can be attributed to the more challenging human dataset. The original OpenPose has a better performance compared to the monkey model.

5.1.2 Visual Assessment

Figures 5.2 5.3 5.4 show images from the evaluation set. These images were not seen during training. Figure 5.2, in addition to the final result, also shows the score maps and PAF. The output score maps confirm that the network can infer monkey body features and their relations in unseen images represented by the generated PAF.

Failure cases

Some of the most common failure cases are illustrated in Figure 5.3: (a) Although a monkey is detected, some of its body parts are not; in most cases, this could be

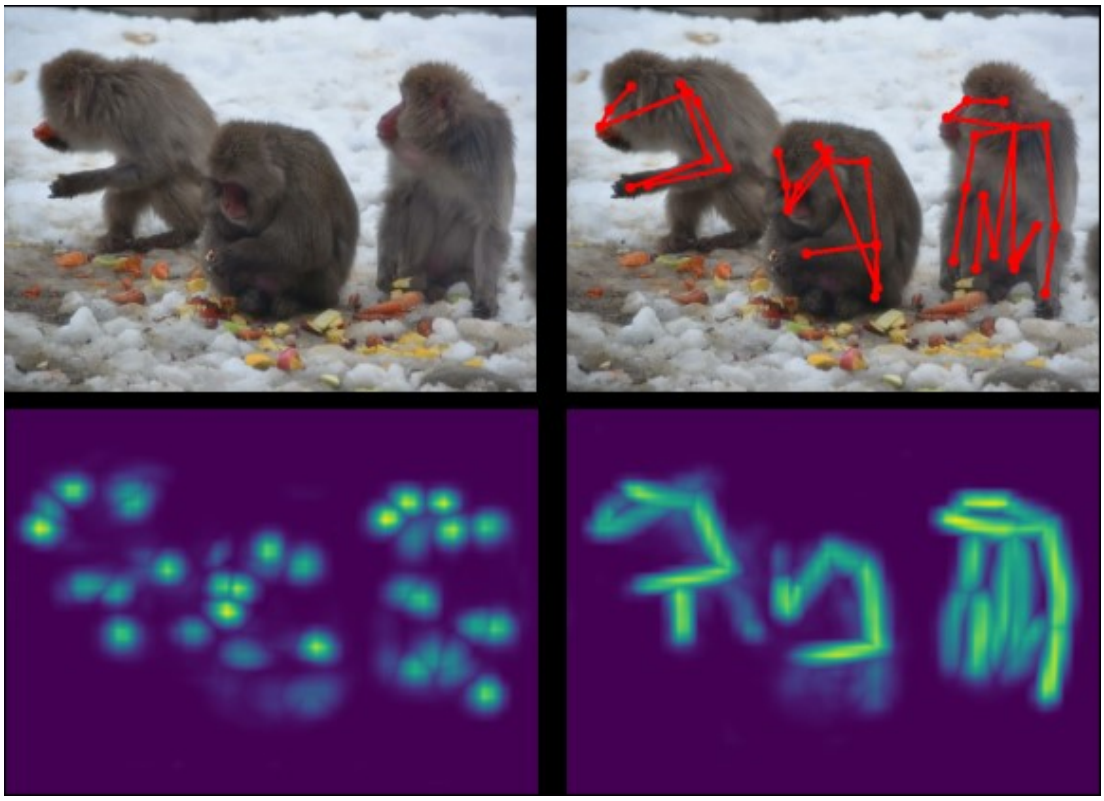

Figure 5.2: Visualization of the pose estimation process for a monkey. The input image (top left) is processed to generate the score map prediction (bottom left) and the PAF prediction (bottom right). The final output (top right) shows the detected key points and skeletal structure overlayed on the input image.

attributed to a low confidence score on the missing body features. (b) A monkey is not detected; this tends to happen more when facial body features are not visible. In some of these cases, monkeys are difficult to spot, even for a human, due to blending with the environment or small scale. (c) In a few cases, some body parts of a nonexistent monkey are detected in the background. (d) False body features of an existing monkey are detected in the background; in most instances, the falsely detected body feature is close to the subject, but in a few cases, it could appear as a random detection. (e) A skeleton prediction takes body parts from different monkeys, or in some cases, two generated monkey skeletons share a single body feature. These errors tend to occur when monkeys are closely interacting.

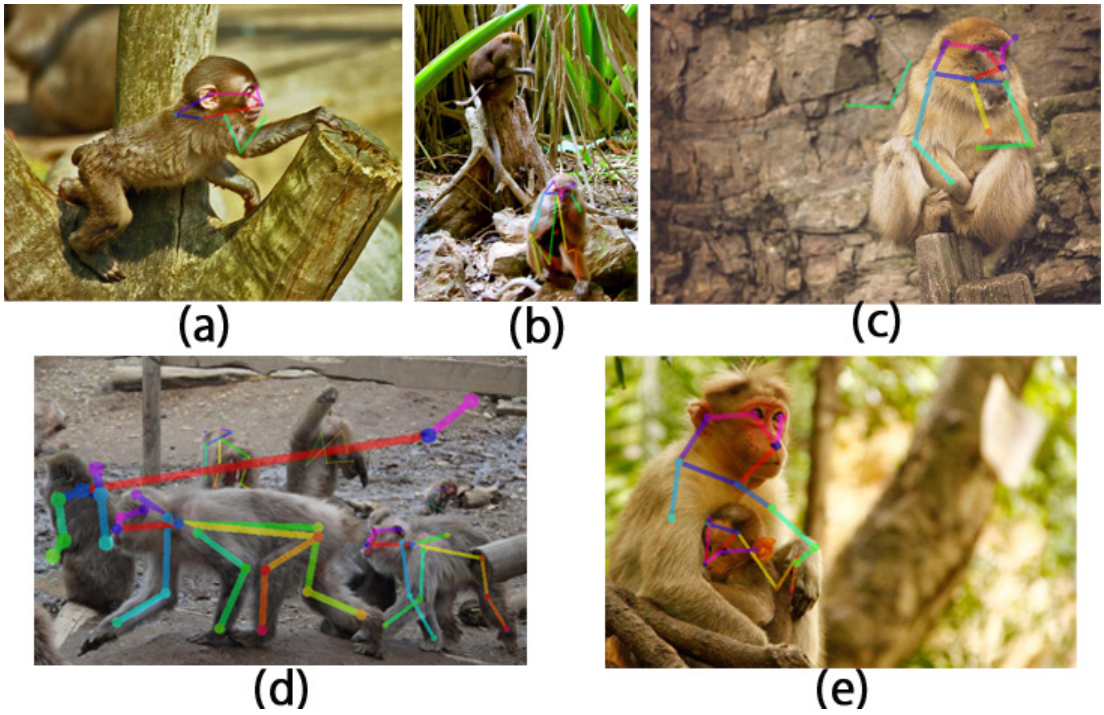

Figure 5.3: Sample failure cases in pose estimation: (a) Missing body parts; (b) Undetected monkey; (c) False detection in the background; (d) Random detections; (e) Shared body part detection among multiple individuals and incorrect association of body parts from different individuals to the same skeleton.

Success cases

Figure 5.4 shows a sample of successful detections on unseen images from the dev set. The sampled images display various activities that include eating, playing, jumping, crawling, and standing. The backgrounds are rich and varied, ranging from natural sceneries with plants and trees to a city with a sea view. Figure 6 Success detection sample: challenging and varied environments, multiple individuals, a variety of activities, occlusion, and social interactions.

Figure 5.4: Successful pose detection examples in challenging scenarios. The figure showcases a variety of environments, multiple individuals, diverse activities, occlusions, and social interactions, demonstrating the robustness of the pose estimation model.

5.1.3 Real-time Performance

The trained model was exported to the Open Neural Network Exchange (ONNX) format and ran in the Nvidia C++ TensorRT high-performance deep learning

inference framework. The model in the ONNX format has a size of 31 MB. In the testing process, it was possible to run the model in real-time using the webcam stream. The Hyperpose open-source project reports being able to run at 60 fps using the default OpenPose with the ResNet18 as a backbone and a resolution of 432 x 368 images. As the trained model uses a lightweight version of OpenPose, it should yield superior performance.

5.2 Marmoset 3D Pose Estimation

In this section, the outcomes of the marmoset 3D posture estimation model are reported. A single clip with two monkeys and a more difficult set of four clips with several monkeys and close interactions were used for the evaluation.

5.2.1 Evaluation

For the single clip with two monkeys, the model achieved the performance metrics presented in Table 5.2. The AP@25 and AP@50 scores were 0.2897 and 0.3800, respectively. These results indicate a reasonable level of accuracy in detecting the 3D poses of the monkeys, with over 50% of the estimated poses having an error of less than 500mm.

Scenario	**AP@25**	**AP@50**
One clip with two monkeys	0.2897	0.3800
Four clips with close interactions	0.04	0.13

Table 5.2: Evaluation results for the Marmoset 3D Pose Estimation model in different scenarios.

In the more challenging scenario with four clips containing multiple monkeys and close interactions, the model's performance was lower, with AP@25: 0.04 and AP@50: 0.13. This decrease in performance can be attributed to the increased complexity of the scenes, where close interactions and occlusions between the monkeys make the estimation task more difficult.

5.2.2 Visual Inspection

Upon visual inspection of the results, the 3D pose estimation model demonstrated the ability to estimate marmosets' poses in various challenging scenarios. However, some limitations were observed, such as difficulty handling extreme occlusions and large variations in scale.

Failure Cases

Despite the model's success in various situations, there were instances where the model failed to estimate the 3D poses of the marmosets accurately. Common failure cases include extreme occlusions, where one marmoset is almost entirely hidden by another, or when the marmosets are very close to each other, making it difficult to distinguish between their body parts. As illustrated in the supporting Figure 5.5, the 3D pose prediction shows a successful pose prediction for an isolated marmoset. However, for a group of three interacting marmosets, the model failed to estimate their poses accurately. It is also important to note that the three marmosets belonged to different development stages and, thus, sizes. In this case, body parts were incorrectly assigned to other individuals within the group, resulting in an inaccurate combined pose. This highlights the challenges faced by the model in estimating poses for closely interacting or occluded marmosets.

Success Cases

In contrast to the failure cases, the model successfully estimated the 3D poses of marmosets in a variety of situations. The successful cases often involved marmosets that were clearly visible and not occluded, allowing the model to predict their poses accurately. Furthermore, when the marmosets were positioned at a reasonable distance from each other, the model could effectively distinguish between their body parts and produce accurate pose estimations.

As shown in Figure 5.6, the model was capable of generating accurate 3D pose predictions for two separate marmosets in both 3D space and when mapped to the image. These successful cases highlight the model's effectiveness in handling non-human primate pose estimation when individuals are not closely interacting

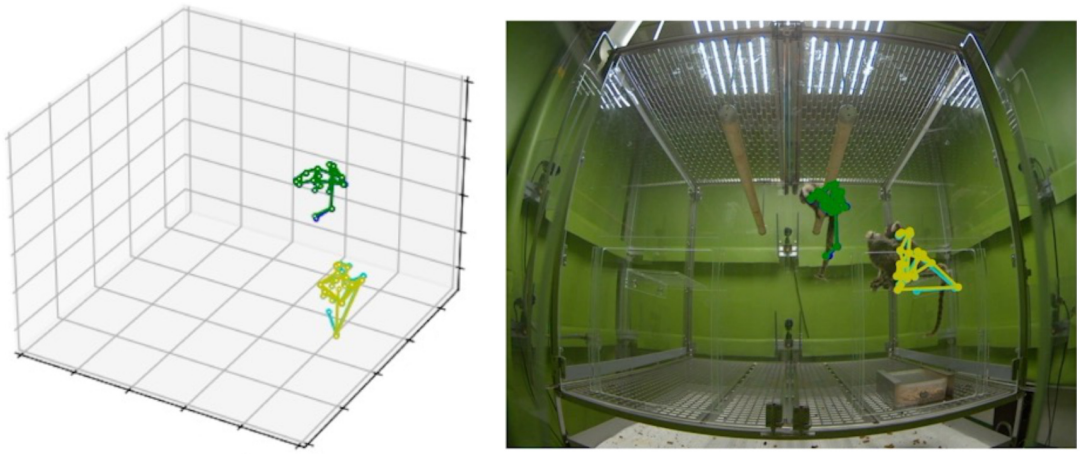

Figure 5.5: Illustration of 3D pose prediction for marmosets. On the left, the 3D pose prediction in a 3D space, and on the right, the 3D prediction is projected to the image. The figure shows an isolated marmoset with a successful pose prediction and a group of three interacting marmosets with a failed pose estimation, resulting in body parts being assigned to different individuals within the group, creating an inaccurate combined pose.

or experiencing significant occlusion.

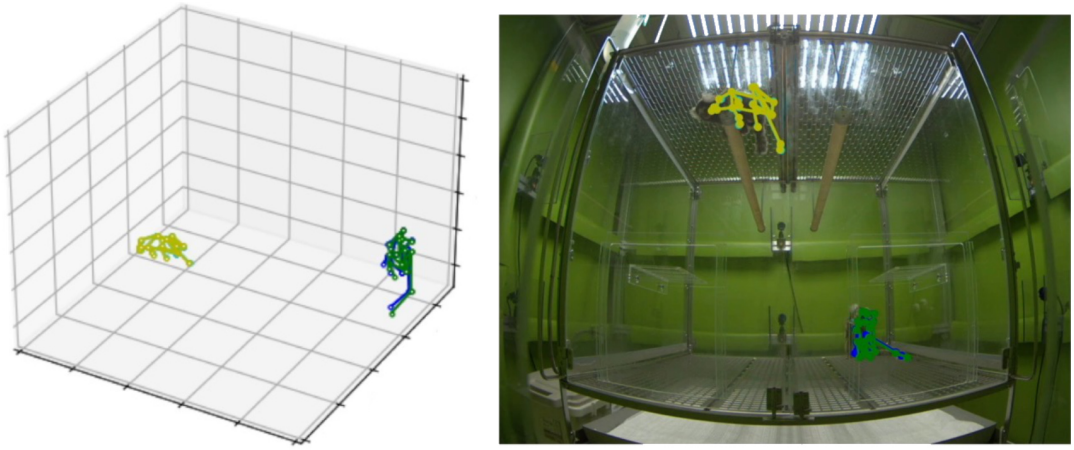

Figure 5.6: Illustration of successful 3D pose prediction for marmosets. On the left, the 3D pose prediction in a 3D space, and on the right, the 3D prediction mapped to the image. The figure shows two separate marmosets with accurate pose estimations, demonstrating the effectiveness of the model when individuals are not closely interacting.

5.2.3 Discussion

The Marmoset 3D Pose Estimation model demonstrates promising performance in simpler scenes, such as the single clip with two monkeys. However, when dealing with more challenging scenarios involving multiple monkeys and close interactions, the model's performance decreases. The observed limitations, such as difficulty handling extreme occlusions and large variations in scale, suggest that further improvements to the model are necessary for more robust performance in complex situations.

One major limitation is the low amount of available training data compared to human pose estimation datasets. The scarcity of annotated marmoset data may limit the model's ability to generalize to more complex scenes and handle various challenging cases. Collecting and annotating more marmoset pose data could potentially improve the model's performance by providing a more diverse and representative training set, allowing the model to learn more robust features

and better handle complex situations.

To better adapt the model to marmosets, some changes have already been implemented, such as modifying the size of feature volumes and cuboid proposals and training the HRNet to predict heatmaps specifically for marmosets. However, these adjustments may not be sufficient to fully account for the unique behaviours and characteristics of marmosets. Marmosets exhibit more dynamic movements, such as jumping and rapidly changing their shape, compared to humans, who typically stand and walk. Further adaptations to the model to accommodate these specific behaviours could lead to improved performance in marmoset pose estimation.

In addition to expanding the training dataset and refining the model's adaptations, other strategies could potentially improve the model's performance in handling challenging scenarios. For example, incorporating additional context information from the environment, utilizing a more robust method for handling occlusions, or employing a multi-stage approach for pose estimation that refines initial predictions. Further research and development are needed to address these limitations and enhance the model's applicability in real-world scenarios involving marmoset behaviour studies.

In conclusion, the Marmoset 3D Pose Estimation model shows potential for application in marmoset behaviour studies. However, its limitations, particularly the low amount of available data and the need for further model adaptations to marmoset-specific behaviours must be addressed to achieve better performance and robustness. Increasing the amount of annotated marmoset data, making further model adjustments, and exploring additional modelling techniques could lead to significant improvements in the model's performance, making it more suitable for real-world scenarios.

5.3 Wet Dog Shaking Detection in Rats

In this section, the results of the efforts to develop a multi-view animal behaviour classification system for detecting Wet Dog Shaking (WDS) behaviour in rats are presented. The evaluation of the system includes the assessment of object detection and image classification performance, as well as the effectiveness of

multi-view integration in improving the detection of WDS events.

5.3.1 Object Detection Results

The object detection network was trained on a dataset containing 1572 images, with 1500 images used for training and 72 for validation. The network achieved an average Intersection over Union (IoU) of 0.98 on the validation dataset, indicating successful localization of the animal in the image. However, the network's classification score was insufficient for practical application, achieving 0.79 precision and 0.52 recall. This led to the decision to use this network only for predicting the Region of Interest (ROI).

5.3.2 Image Classification Results

The image classification network (NN2) was trained on a dataset consisting of 25,549 images, using 24,920 of those images for training purposes and the remaining 629 for validation. The network achieved a 92% accuracy on the validation set. However, the 8% error rate, although seemingly low, is translated into numerous false positives when predicting the WDS behaviour.

Figure 5.7 illustrates the issue more clearly, showing the raw per-frame predictions from NN2 across three different views (blue) alongside the ground truth (red). The figure highlights the presence of many false positives generated by the network, despite the overall high accuracy. The result seems even worse when taking into account the fact that WDS behaviour occurs just 0.38% of the time in an already overexpressive-WDS disease model during the experiments, making false positives more noticeable and detrimental.

This observation demonstrates that relying solely on image classification using NN2 is not sufficient for accurately detecting the onset and termination of the WDS behaviour. To improve the detection of this behaviour, additional strategies, such as incorporating multiple views and temporal information with the help of NN3, should be employed to reduce false positives and enhance the overall performance of the system.

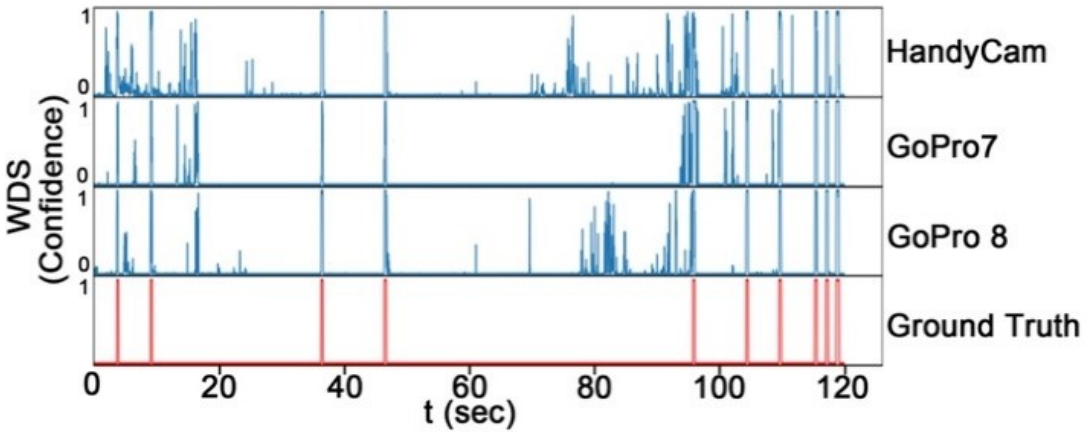

Figure 5.7: Figure displaying the raw per-frame predictions generated by the image classification network (NN2) across three distinct views, depicted in blue, in contrast with the ground truth marked in red.

5.3.3 Multi-view Performance Evaluation

A 1-hour experiment featuring 49 WDS events was conducted, examining the impact of using different numbers of views, ranging from one to three cameras, and assessing the multi-view system's effectiveness. The significant benefits of incorporating additional views into the approach were demonstrated by evaluating the system's performance using recall and precision metrics, as well as considering time analysis. As illustrated in Table 5.3, precision remained consistently high across all camera configurations, highlighting the system's capacity to identify WDS events while minimizing false positives accurately. Additionally, recall improved as the number of views increased, suggesting an enhancement in the system's ability to detect WDS events without missing true instances. This boost in recall emphasizes the importance of integrating multiple cameras into the methodology and considering time analysis to capture WDS behaviour better, ultimately leading to a more accurate and dependable detection process.

A direct comparison with previous methods is not feasible, as no existing systems could detect WDS behaviour. Moreover, modifying existing behaviour detection systems, such as the Live Mouse Tracker [13], to incorporate WDS detection is challenging. However, in the absence of previous automated WDS

Views	Precision	Recall
Three	0.91	0.86
Two	0.97	0.65
One	0.90	0.57

Table 5.3: Comparison of WDS recall and precision metrics with three different camera configurations in the 1-hour validation recording.

detection methods, we can use manual labelling as the benchmark for comparison. Manual labelling is considered the prior method for such analysis. Although time-consuming and labour-intensive, manual labelling provides a reliable reference for evaluating our system's performance. Precision and recall metrics are particularly useful in this context. Recall allows us to measure the system's ability to identify all instances of WDS behaviour compared to the human-annotated labels (ground truth).

To further assess the effectiveness of our system with various camera setups, Figure 5.8 has been incorporated, showcasing the ROC curves for three distinct scenarios: single-view, two-views, and three-views. These ROC curves effectively demonstrate the balance between sensitivity and specificity for each configuration. Notably, the outcomes reveal that the inclusion of multiple views enhances the overall performance of our system. However, it is important to note that the ROC curves are generated based on raw per-frame data, while the results presented in Table 2 are computed on a per WDS event basis. The integration of multiple views proves beneficial in refining the system's capabilities.

5.3.4 Visual Inspection

Figure 5.9 presents a visual comparison of the final predictions of the multi-view system for WDS behaviour in a two-minute video sample extracted from the 1-hour validation experiment session. The two-minute sample is analyzed with different numbers of views (from one up to three views) in blue, alongside the ground truth in red. Out of the 10 WDS events represented in the ground truth, the system detected 6 WDS events with one view, 8 with two views, and all 10 with three views. In all cases, no false negatives were observed. Upon

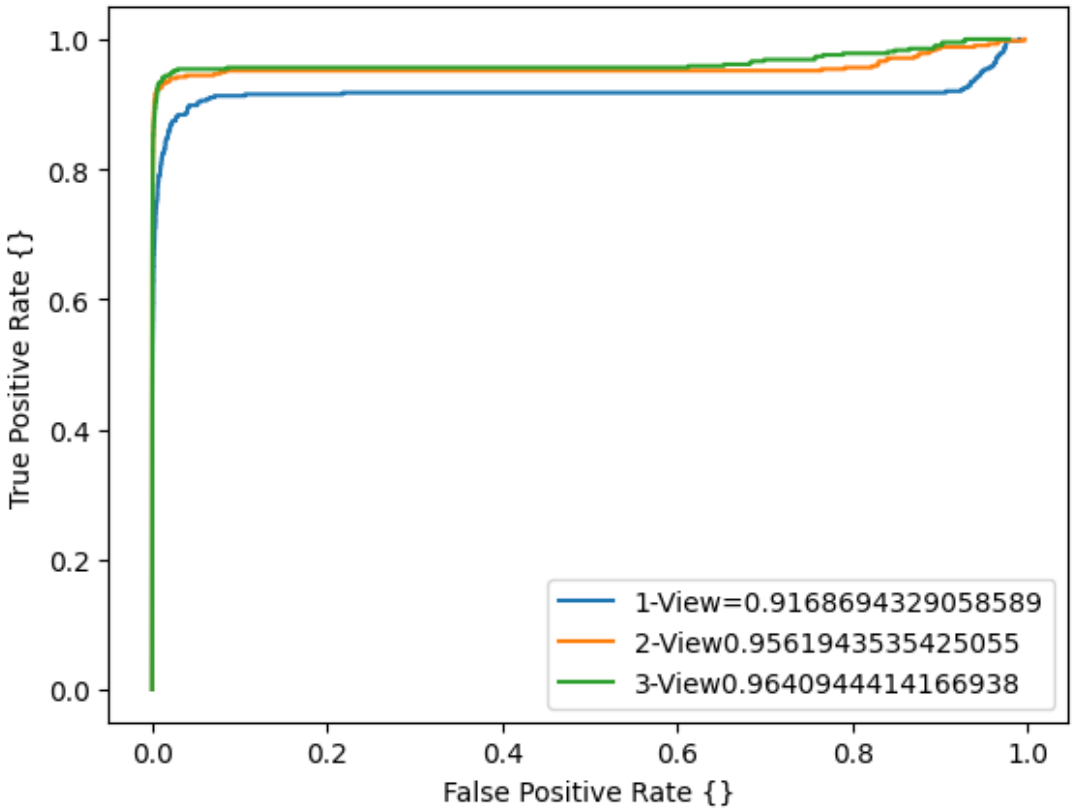

Figure 5.8: Receiver operating characteristic (ROC) curves for one, two, and three view configurations

visually inspecting the predictions in Figure 5.9, it becomes evident that the number of WDS events recalled increases with the addition of more cameras while maintaining zero false positives. This improvement in recall can be attributed to the fact that multiple views offer better coverage of the subject's behaviour, reducing the likelihood of missing true instances due to occlusion or orientation. Moreover, the increased information from multiple views also aids the system in discerning true WDS events and other similar actions, thereby reducing the number of false positives. These observations are consistent with the results in Table 5.3, as the recall values improve with the inclusion of more views, confirming the benefits of using multiple views for more accurate and reliable WDS behaviour detection.

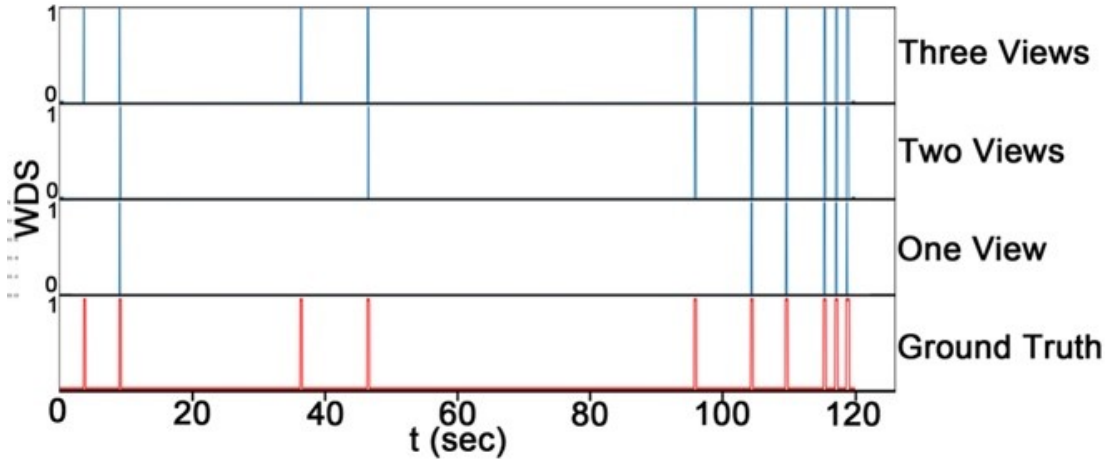

Figure 5.9: Figure displaying the raw per-frame predictions generated by the image classification network (NN2) across three distinct views, depicted in blue, in contrast with the ground truth marked in red.

Failure cases

In this section, analyze the instances where the multi-view behavior classification system failed. Throughout the 1-hour validation experiment, we observed a total of four false positives, which can be seen in Figure 5.10. Among these cases, two instances depicted the rat engaging in rearing behavior, while the other two instances showed walking behavior.

Figure 5.10: The false positives (four in total) during the 1 h validation experiment where the rat exhibits rearing behavior (two cases) and walking behavior (two cases).

The false positives in rearing behavior can be attributed to the occasional occurrence of forelimb elevation during Wet Dog Shaking (WDS) behavior, leading to a resemblance with rearing, as demonstrated in Figure 2.1. To ensure the model can distinguish between rearing and WDS, we conducted an analysis of the first 10 minutes of the validation experiment. During this period, we identified seven instances of rearing behavior, and remarkably, none of them were falsely detected as WDS. This observation suggests that the model generally exhibits the capability to distinguish between rearing behavior and WDS accurately. As for the false positives involving the rat's walking, we noted that the rat frequently moved throughout the experiment, yet only two instances were mistakenly classified as WDS behavior. In Figure 5.7, we showcase the raw per-frame predictions generated by the image classification network (NN2) across three distinct views, with the ground truth marked in red. Interestingly, between 15 and 17 seconds, all views indicated some probability of WDS, despite the rat actually walking, as illustrated in Figure 5.11. However, the final prediction by NN3 successfully filtered out the noise. While the raw classification predicted a high probability of WDS, reaching around 80 percent, true WDS events exhibited probabilities close to 100 percent, demonstrating the system's remarkable capability to distinguish walking from WDS in most cases.

In the case of false negatives, no discernible pattern was found to explain these misclassifications. Therefore, it is imperative to conduct further investigation and potentially make adjustments to the model to address these instances and significantly enhance the overall performance of the behavior classification system.

5.3.5 Discussion

Inspired by the human decision-making process that frequently depends on visual information from a variety of angles, a multi-view animal behaviour classification system for detecting Wet Dog Shaking (WDS) behaviour in rats was developed, which showed promising results. This system addresses unique challenges in animal behaviour classification, such as the lack of datasets and the variety of animals across different species used for experiments.

The evaluation of the system included assessments of object detection, image classification performance, and the effectiveness of multi-view integration in im-

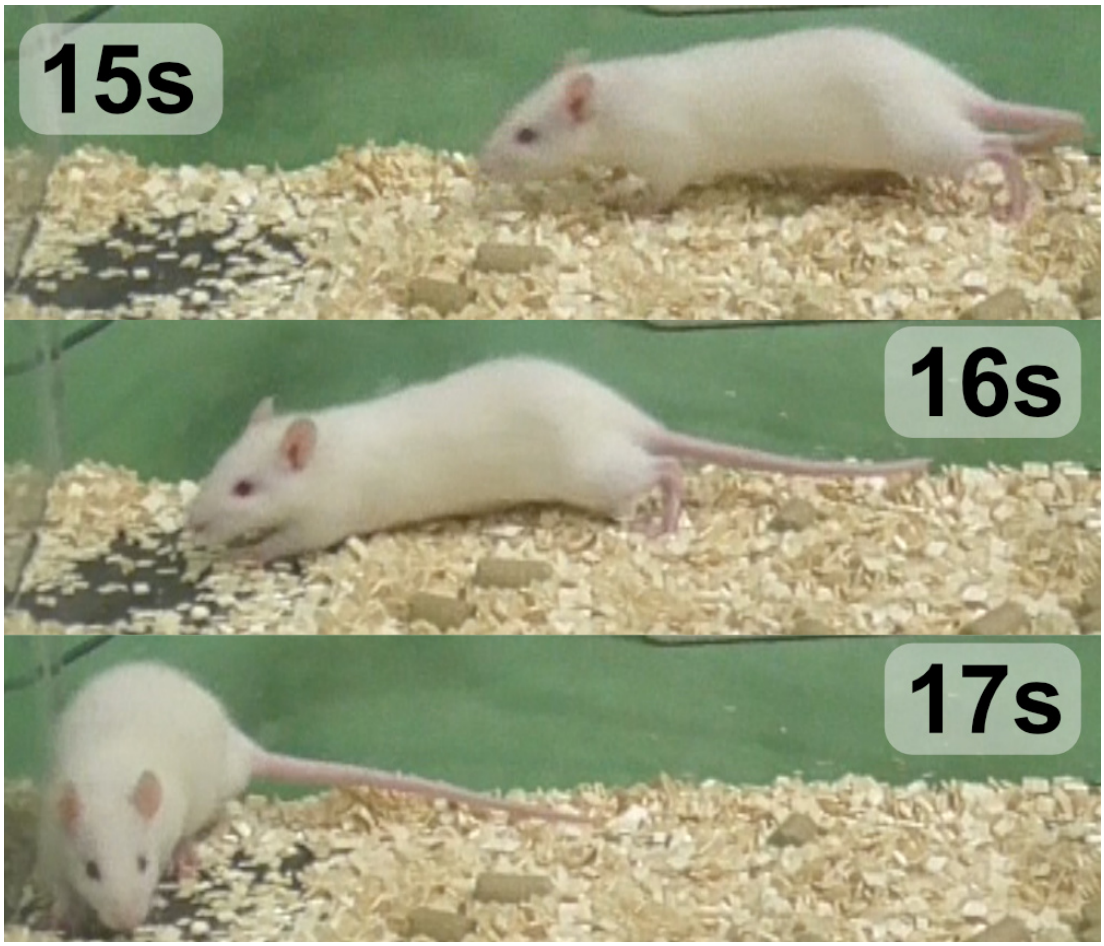

Figure 5.11: Time-lapse of the rat at the 15-17s mark, displaying the rat's walking behaviour, NN2 displays high probability of WDS behaviour

proving WDS event detection. The object detection network achieved successful animal localization in the image, with an average Intersection over Union (IoU) of 0.98 on the validation dataset. However, the network's classification score needed to be revised for practical application, leading to the decision to use it only for predicting the Region of Interest (ROI). The image classification network (NN2) achieved a 92% accuracy on the validation set. Nevertheless, the 8% error rate resulted in a significant number of false positives when predicting WDS behaviour, demonstrating that relying exclusively on image classification with NN2 was insufficient for accurately detecting WDS events.

Incorporating multiple views and temporal information with the help of NN3 proved to be an effective strategy to reduce false positives and enhance the system's overall performance. The system's recall and precision improved by adding more cameras, achieving 0.91 precision and 0.86 recall in the WDS behaviour classification task. This improvement can be attributed to the better coverage of the subject's behaviour and the increased information available from multiple views, which aids in discerning true WDS events and other similar actions, thereby reducing the number of false positives.

The findings confirm the benefits of using multiple views for more accurate and reliable WDS behaviour detection. The developed multi-view system shows potential for further application in the field of animal behaviour analysis, including other rat disease models where WDS is present and adaptation for other subjects, classification tasks, and animals. The system is designed to be easily adapted and used with different animals, behaviours, and other classification tasks while training with little data. It can also be easily adapted to be used with different amounts of cameras and/or for new environments by only fine-tuning the third network (NN3).

In the context of other rat disease models, the method should work where WDS is present, either as a spontaneous behaviour or in various disease models. This is because no studies suggest significant WDS locomotion differences between disease models. Additionally, the system could be adapted for use with other subjects and classification tasks, including other animals and behaviours.

In conclusion, the developed multi-view animal behaviour classification system demonstrates the advantages of using multiple views for accurate and reliable

behaviour detection. The system's adaptability opens up possibilities for further application in the field of animal behaviour analysis.

6 Conclusion

In this dissertation, two multi-view deep learning frameworks for animal behaviour analysis are introduced, first targeting pose estimation, adopting VoxelPose to non-human primates and second a novel feature-independent approach for behaviour detection, exemplified by its use in detecting wet dog shaking behaviour in rats. By harnessing sophisticated deep learning techniques and multi-view data from various angles, these frameworks showcase their potential to enhance behaviour analysis accuracy and robustness beyond the scope of conventional single-view systems. The primary contributions of this work can be summarized as follows:

1. The first Multiple-Monkey pose estimation system, called "Multiple Monkey Pose Estimation Using OpenPose," OpenPose was trained for non-human primates, such as macaque monkeys.

2. An extensive dataset of macaque monkey images in the wild, known as "MacaquePose," has been assembled, serving as a valuable resource for researchers in the field. Its natural environments and wide range of data have already proven useful in numerous studies, enabling researchers to gain a deeper understanding of macaque monkey behaviour in natural environments.

3. To further enhance the analysis of non-human primate behaviour, 3D pose images of monkeys and marmosets were collected in a laboratory setting. Utilizing these images, deep neural networks were trained specifically for 3D pose estimation in non-human primates. The resulting system is capable of performing multi-view pose estimation with a high degree of accuracy, providing a more complete and detailed representation of the animals'

movements and behaviours. This advancement in 3D pose estimation techniques allows researchers to study non-human primates more effectively and comprehensively.

4. A novel multi-view approach for the detection of wet dog shaking (WDS) behaviour in rats was developed, marking the first system capable of detecting this behaviour. WDS is a short-duration behaviour relevant to the study of various animal disease models, including acute seizures, morphine abstinence, and nicotine withdrawal. By better detecting and analyzing WDS, this approach could significantly impact the study of these disease models. A multi-view system was designed and implemented to simultaneously capture WDS behaviour from multiple perspectives, providing a more comprehensive understanding of the rats' movements. Classifiers for each view were trained to detect WDS behaviour, and an additional network was developed to fuse the classification results and perform time analysis.

5. The first wet-dog shake dataset was compiled, containing over three hours of video footage from multiple views. This extensive dataset offers a valuable resource for researchers seeking to study this particular animal behaviour. It introduces a novel non-human multi-view dataset for activity recognition with a practical application, supporting the development of advanced recognition techniques and fostering further research.

In conclusion, the innovative multi-view deep learning framework presented in this dissertation has the potential to improve animal behaviour analysis. By capturing a complete representation of animal movements from multiple perspectives, the approach will enable researchers to gain a more comprehensive understanding of complex behaviours, leading to new discoveries and insights into animal cognition, communication, and social dynamics. Furthermore, the application of deep learning techniques to animal behaviour analysis opens up new avenues for research in animal welfare, conservation, and the development of innovative therapies and interventions for human and animal health. As the methods and techniques introduced in this dissertation continue to be refined and expanded upon, we hope they will have a transformative impact on the broader scientific

community and society as a whole, ultimately contributing to a deeper understanding of the intricate world of animal behaviour.

6.1 Limitations

This thesis introduces two successful frameworks that employ deep learning for multi-view animal behaviour analysis. The first framework has adapted VoxelPose for pose estimation in non-human primates, specifically marmosets. The second framework presents a novel feature-independent approach for behaviour detection, as exemplified by its application in detecting wet dog shaking behaviour in rats. However, it is important to acknowledge the presence of limitations. This section discusses some of these limitations, with a specific focus on the proposed multi-view detection system applied to WDS behaviour in rats, as outlined below:

- Limited species and behaviours: The current framework for pose estimation focuses on non-human primates.
- Reliance on annotated data: The pose estimation framework depends on annotated datasets to train deep learning models. Gathering and annotating extensive datasets can be laborious and difficult, particularly for uncommon behaviours or hard-to-access species. While the multi-view behaviour detection system presented here mitigates this by not relying on feature engineering, investigating unsupervised or semi-supervised learning methods could be even better.
- Challenging environments: The performance of the multi-view behaviour detection framework for WDS could be impacted by environmental factors like varying lighting conditions, occlusions, and complex backgrounds. Notably, the dataset was gathered within a single environment, specifically the same laboratory setting. However, the cameras were mounted and dismounted, and the LMT acrylic cage was moved between experiments, introducing some degree of variability into the dataset. Consequently, future endeavours could introduce new environments.

- Computational complexity: The deep learning models used in this framework may require significant computational resources, which can be a limiting factor, especially for real-time applications. In this work, an attempt was made to mitigate this limitation by utilizing computational cloud platforms, such as Google Colab. Moreover, expanding cameras, enhancing frame rate, or incrementing resolution may entail higher hardware expenses and necessitate additional processing capabilities..

These limitations highlight areas for future research and improvement, contributing to the development of more robust and versatile tools for animal behaviour analysis.

6.2 Future Work

This thesis introduces two successful frameworks that employ deep learning for multi-view animal behaviour analysis. The first framework, VoxelPose, has been adapted for pose estimation in non-human primates, specifically marmosets. The second framework presents a novel feature-independent approach for behaviour detection, exemplified by its use in detecting wet dog shaking behaviour in rats. However, several open questions for further research within this field persist. This section proposes potential solutions for future investigation, with particular emphasis on the proposed multi-view detection system applied to WDS behaviour in rats, as outlined below:

How can the multi-view behaviour detection proposed framework be trained to include more species and additional behaviours?

Future work could be trained in other animal species and behaviours. This could be achieved by expanding the existing datasets to include more annotated images and videos of different animals performing various behaviours. .

Can the multi-view behaviour detection framework be adapted to handle real-time analysis ?

Future efforts could focus on simplifying deep learning models or improving existing ones to address the challenge of real-time analysis. These changes might involve methods like reducing the neural network size and exploring ways to speed up hardware. Additionally, using ONNX could help access hardware enhancements or run the system on powerful local hardware, as the proposed system currently uses Google Colab. ONNX, or Open Neural Network Exchange, is an open-source framework that facilitates interoperability between different deep learning frameworks. It allows seamless sharing of trained models between various platforms and tools, enabling efficient deployment and optimization.

How can the proposed multi-view behaviour detection framework be adapted to work in more challenging environments with limited visibility or occlusions?

The current framework could be improved to handle challenging environments by incorporating additional sensors or data modalities, such as infrared cameras, depth sensors, to supplement the multi-view camera data. By fusing this addi-

tional information with the existing multi-view data, the deep learning models could be better equipped to handle occlusions and low visibility scenarios.

How can the insights gained from the multi-view behaviour analysis be integrated with other behavioural or physiological data to provide a more holistic understanding of an animal's state? Future research could focus on developing methods to integrate the multi-view behavioural analysis results with other relevant data sources, such as physiological measurements, hormone level measurements, or environmental factors. This integration could be achieved by developing models that consider these additional data sources as input features or by fusing the results of separate models in a hierarchical or ensemble-based approach. By incorporating these additional dimensions, the framework could provide a more holistic understanding of an animal's state, ultimately leading to more accurate and meaningful conclusions about animal behaviour and welfare.

How can the proposed multi-view behaviour detection framework be adjusted to function with multiple subjects? Currently, NN1, utilized for object localization, can identify one or more subjects, rendering it suitable for multi-subject detection. However, addressing the issue of assigning identity across frames is crucial, given that the network operates on a per-frame basis. This should not be challenging, as it is a well-explored problem with numerous solutions accessible.

How could the accuracy of the system be improved?

To enhance the system's accuracy, several strategies could be explored. Incorporating more cameras into the setup could provide a broader perspective of the behaviour and add redundancy, which is particularly beneficial for capturing complex actions. Increasing the dataset size by recording more instances of the behaviour or introducing recordings in novel environments could contribute to better generalization. Similarly, increasing the frame rate might capture finer details of the behaviour. However, each of these strategies comes with its own set of challenges and complications that need to be carefully considered. For instance, incorporating more cameras could lead to increased hardware costs and require additional processing resources. Collecting a substantial dataset could be time-consuming and necessitate significant annotation efforts. Similarly, a higher

frame rate would result in a larger volume of data for processing and storage. In this study, we carefully balanced factors such as the number of cameras, frame rate, and dataset size, taking into account the technical complexities and resource demands associated with each potential enhancement.

The future work proposed in this section seeks to address key open questions in the field of multi-view animal behaviour analysis using deep learning. The tools introduced in this dissertation hold the potential to advance the field of animal behaviour analysis significantly. By harnessing the advancements in deep learning techniques and leveraging multi-view data, these frameworks offer the means to achieve a more accurate and comprehensive understanding of animal behaviours. The utilization of these tools is anticipated to yield findings that not only enhance our insights into animal behaviour but also propel broader scientific progress. The study of animal behaviour often provides valuable insights that can parallel and shed light on human behaviour, given the shared underlying behavioural patterns and neural mechanisms. As scientific knowledge continues to evolve, these tools could lay the groundwork for even more sophisticated approaches capable of deciphering the intricacies of human behaviour. This would empower researchers to unravel the intricacies of human actions, interactions, psychological states, and other unknown connections. The horizon holds the promise of a deeper grasp of both animal and human behaviour, ultimately contributing to a more comprehensive understanding of the complexities of life itself.

Acknowledgements

First, I am immensely grateful to Professor Tomohiro Shibata, who has generously shared his wisdom, mentorship, and encouragement during my academic journey. His unwavering support has enabled and inspired me to push through challenges and thrive in my research. I would also like to express my sincere appreciation to Professor Kiyohisa Natsume, Professor Jumpei Matsumoto, and Professor Takaaki Kaneko for providing the necessary tools and materials. I am grateful for their invaluable input and guidance on this project. Additionally, I extend my thanks to the remaining thesis committee members: Professor Hiroaki Wagatsuma, Professor Keiichi Horio and Professor Hideaki Kawano. Their astute insights and constructive feedback were instrumental in bringing this thesis to fruition.

I would like to express my sincere gratitude to my friends and colleagues in the Shibata laboratory. The moments spent working alongside my fellow lab members and the enjoyable times we shared will always be cherished. I offer special thanks to my Kyutech seniors; I would like to acknowledge Doctor Arai Hirofumi as an exceptional example of hard work and commitment to helping others. Watching him constantly strive for self-improvement has been a tremendous inspiration to me. Additionally, I want to highlight Rollyn Labuguen as a true embodiment of kindness. Their continuous guidance has been invaluable, and this research would not have been possible without their unwavering support. I would also like to thank the lab secretaries Ohta San, Iwashita San, and Tanaka San, as well as the entire Kyutech staff for their unwavering assistance and support. I extend my gratitude to ITQ Professor Jose Joaquin Castellanos Galindo, who encouraged me to pursue my dreams of graduate studies and helped as a mentor. And, of course, I am deeply grateful to my parents and siblings in Mexico for their constant encouragement throughout my academic journey and life.

Lastly, I am deeply grateful to the Mexican Government's National Council of Science and Technology (CONACYT) for their assistance with my education. If not for the CONACYT Scholarship, my aspiration to study in Japan would have remained unrealized. I sincerely appreciate this incredible opportunity.

Publication List

Journal Papers

[1] Salvador Blanco Negrete, Hirofumi Arai, Kiyohisa Natsume, and Tomohiro Shibata. "Multi-view image-based behavior classification of Wet-Dog Shake in Kainate Rat Model". In: *Frontiers in Behavioral Neuroscience* (2023).

[2] Rollyn Labuguen, Jumpei Matsumoto, Salvador Blanco Negrete, Hiroshi Nishimaru, Hisao Nishijo, Masahiko Takada, Yasuhiro Go, Ken-ichi Inoue, and Tomohiro Shibata. "MacaquePose: A Novel "In the Wild" Macaque Monkey Pose Dataset for Markerless Motion Capture". In: *Frontiers in Behavioral Neuroscience* 14 (2021). ISSN: 1662-5153. DOI: 10.3389/fnbeh.2020.581154. URL: https://www.frontiersin.org/articles/10.3389/fnbeh.2020.581154.

Conference Papers

[1] Hirofumi Arai, Salvador Blanco Negrete, Rollyn Tiu Labuguen, Tomohiro Shibata, and Kiyohisa Natsume. "Behavior and Emotional Modification at Latent Period of Rat Epilepsy Model". In: *2021 60th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*. Tokyo, Japan, 2021, pp. 546–551. URL: https://ieeexplore.ieee.org/abstract/document/9555283.

[2] Rollyn Labuguen, Salvador Blanco Negrete, Tonan Kogami, Wally Enrico Ingco, and Tomohiro Shibata. "Performance Evaluation of Markerless 3D Skeleton Pose Estimates with Pop Dance Motion Sequence". In: *2020 Joint 9th International Conference on Informatics, Electronics and Vision (ICIEV) and 2020 4th International Conference on Imaging, Vision and*

Pattern Recognition (icIVPR). 2020. DOI: 10.1109/ICIEVicIVPR48672.2020.9306581.

[3] Rollyn Labuguen, Dean Bardeloza, Salvador Blanco Negrete, Jumpei Matsumoto, Ken-Ichi Inoue, and Tomohiro Shibata. "Primate Markerless Pose Estimation and Movement Analysis Using DeepLabCut". In: *Conference: 2019 Joint 8th International Conference on Informatics, Electronics and Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision and Pattern Recognition (icIVPR)*. 2019, pp. 297–300. DOI: 10.1109/ICIEV.2019.8858533.

[4] Rollyn Labuguen, Vishal Gaurav, Salvador Blanco Negrete, Jumpei Matsumoto, Kenichi Inoue, and Tomohiro Shibata. "Monkey Features Location Identification Using Convolutional Neural Networks". In: *The 28th Annual Conference of the Japanese Neural Network Society*. Tokyo, Japan, 2018. DOI: 10.1101/377895. URL: https://www.biorxiv.org/content/early/2018/07/28/377895.

[5] Salvador Blanco Negrete, Ravi Joshi, Rollyn Labuguen, Jumpei Matsumoto, and Tomohiro Shibata. "Mouse Anatomical Cardinal Planes and Axes Towards Augmentation for Behavior Analysis". In: *2018 Joint 7th International Conference on Informatics, Electronics and Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision and Pattern Recognition (icIVPR)*. 2018, pp. 195–199. DOI: 10.1109/ICIEV.2018.8641000.

Workshop Papers

[1] Salvador Blanco Negrete, Ravi Joshi, Rollyn Labuguen, Jumpei Matsumoto, and Tomohiro Shibata. "Multiple Monkey Pose Estimation Using OpenPose". In: *25th International Conference on Pattern Recognition (ICPR 2020)*. 2021. DOI: https://doi.org/10.1101/2021.01.28.428726. URL: https://homepages.inf.ed.ac.uk/rbf/VAIB20PAPERS/vaibsb20.pdf.

Poster Presentations

[1] Salvador Blanco Negrete, Dean Bardeloza, Rollyn Labuguen, Jumpei Matsumoto, Kenichi Inoue, and Tomohiro Shibata. "Development of Low-cost Markerless Phenotyping Systems for Mice and Monkey via DeepLabCut". In: *Jisedai Nou Purojyekuto Fuyu no Shinpojiumu (2018)*. 2018.

[2] Rollyn Labuguen, Salvador Blanco Negrete, Sanjay Dwivedi, and Tomohiro Shibata. "Development of Markerless Phenotyping Systems Towards Discovering Individuality". In: *International Conference on Machine Intelligence (2019)*. 2019.

Appendices

References

[1] Jeanne Altmann. "Observational study of behavior: sampling methods." In: *Behaviour* 49.3 (1974), pp. 227–67. ISSN: 0005-7959. DOI: 10.1163/156853974x00534. URL: https://brill.com/view/journals/beh/49/3-4/article-p227_3.xml%20http://www.ncbi.nlm.nih.gov/pubmed/4597405.

[2] David J. Anderson and Pietro Perona. "Toward a Science of Computational Ethology". In: *Neuron* 84.1 (2014), pp. 18–31. ISSN: 08966273. DOI: 10.1016/j.neuron.2014.09.005. URL: http://www.ncbi.nlm.nih.gov/pubmed/25277452%20https://linkinghub.elsevier.com/retrieve/pii/S0896627314007934.

[3] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. "DeepLabCut: markerless pose estimation of user-defined body parts with deep learning". In: *Nature Neuroscience* 21.9 (2018), pp. 1281–1289. ISSN: 15461726. DOI: 10.1038/s41593-018-0209-y. URL: http://dx.doi.org/10.1038/s41593-018-0209-y.

[4] Jumpei Matsumoto, Hiroshi Nishimaru, Taketoshi Ono, and Hisao Nishijo. "3D-Video-Based Computerized Behavioral Analysis for In Vivo Neuropharmacology and Neurophysiology in Rodents". In: 121 (2017), pp. 89–105. DOI: 10.1007/978-1-4939-6490-1. URL: http://link.springer.com/10.1007/978-1-4939-6490-1.

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. "OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. arXiv 2018". In: *arXiv preprint arXiv:1812.08008* (1812).

[6] Ahmet Arac, Pingping Zhao, Bruce H. Dobkin, S. Thomas Carmichael, and Peyman Golshani. "DeepBehavior: A Deep Learning Toolbox for Automated Analysis of Animal and Human Behavior Imaging Data." In: *Frontiers in systems neuroscience* 13.May (2019), p. 20. ISSN: 1662-5137. DOI: 10.3389/fnsys.2019.00020. URL: http://www.ncbi.nlm.nih.gov/pubmed/31133826%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6513883.

[7] Eddie Wei. "Assessment of precipitated abstinence in morphine- dependent rats". In: *Psychopharmacologia* 28.1 (1973), pp. 35–44. ISSN: 0033-3158. DOI: 10.1007/BF00413955. URL: http://link.springer.com/10.1007/BF00413955.

[8] K. Suemaru, H. Araki, Y. Kitamura, K. Yasuda, and Y. Gomita. "Cessation of chronic nicotine administration enhances wet-dog shake responses to 5-HT2 receptor stimulation in rats". In: *Psychopharmacology* 159.1 (2001), pp. 38–41. ISSN: 00333158. DOI: 10.1007/s002130100866.

[9] Doga Vuralli, Anne Sophie Wattiez, Andrew F. Russo, and Hayrunnisa Bolay. "Behavioral and cognitive animal models in headache research Cenk Ayata". In: *Journal of Headache and Pain* 20.1 (2019). ISSN: 11292377. DOI: 10.1186/s10194-019-0963-6.

[10] Philip N Lehner. *Handbook of ethological methods*. Cambridge University Press, 1998.

[11] Melissa Bateson and Paul Martin. *Measuring behaviour: an introductory guide*. Cambridge university press, 2021.

[12] Marian Stamp Dawkins. *Observing animal behaviour: design and analysis of quantitative data*. Oxford University Press, 2007.

[13] Fabrice de Chaumont, Elodie Ey, Nicolas Torquet, Thibault Lagache, Stéphane Dallongeville, Albane Imbert, Thierry Legou, Anne-Marie Le Sourd, Philippe Faure, Thomas Bourgeron, and Jean-Christophe Olivo-Marin. "Real-time analysis of the behaviour of groups of mice via a depth-sensing camera and machine learning". In: *Nature Biomedical Engineering* 3.11 (2019), pp. 930–942. ISSN: 2157-846X. DOI: 10.1038/s41551-019-0396-1. URL: https://www.nature.com/articles/s41551-019-0396-1.

[14] John Joseph Valletta, Colin Torney, Michael Kings, Alex Thornton, and Joah Madden. "Applications of machine learning in animal behaviour studies". In: *Animal Behaviour* 124 (2017), pp. 203–220.

[15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[16] Jana Wäldchen and Patrick Mäder. "Machine learning for image based species identification". In: *Methods in Ecology and Evolution* 9.11 (2018), pp. 2216–2225.

[17] Semih Günel, Helge Rhodin, Daniel Morales, João Campagnolo, Pavan Ramdya, and Pascal Fua. "DeepFly3D, a deep learning-based approach for 3D limb and appendage tracking in tethered, adult Drosophila". In: *Elife* 8 (2019), e48571.

[18] Rollyn Labuguen, Jumpei Matsumoto, Salvador Blanco Negrete, Hiroshi Nishimaru, Hisao Nishijo, Masahiko Takada, Yasuhiro Go, Ken-ichi Inoue, and Tomohiro Shibata. "MacaquePose: a novel "in the wild" macaque monkey pose dataset for markerless motion capture". In: *Frontiers in behavioral neuroscience* 14 (2021), p. 581154.

[19] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[20] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. "VoxelPose: Towards Multi-Camera 3D Human Pose Estimation in Wild Environment". In: *European Conference on Computer Vision (ECCV)*. 2020.

[21] Alexander Toshev and Christian Szegedy. "Deeppose: Human pose estimation via deep neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1653–1660.

[22] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. "Deepercut: A deeper, stronger, and faster multi-person pose estimation model". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*. Springer. 2016, pp. 34–50.

[23] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. "Deep High-Resolution Representation Learning for Human Pose Estimation". In: *CVPR*. 2019.

[24] Rollyn Labuguen, Vishal Gaurav, Salvador Negrete Blanco, Jumpei Matsumoto, Kenichi Inoue, and Tomohiro Shibata. "Monkey features location identification using convolutional neural networks". In: (2018).

[25] Marco Seeland and Patrick Mäder. "Multi-view classification with convolutional neural networks". In: *Plos one* 16.1 (2021), e0245230.

[26] Daniel Weinland, Remi Ronfard, and Edmond Boyer. "A survey of vision-based methods for action representation, segmentation and recognition". In: *Computer vision and image understanding* 115.2 (2011), pp. 224–241.

[27] Prasetia Utama Putra, Keisuke Shima, and Koji Shimatani. "Markerless Human Activity Recognition Method Based on Deep Neural Network Model Using Multiple Cameras". In: *2018 5th International Conference on Control, Decision and Information Technologies, CoDIT 2018* July (2018), pp. 13–18. DOI: 10.1109/CoDIT.2018.8394780.

[28] Prasetia Utama Putra, Keisuke Shima, and Koji Shimatani. "A deep neural network model for multi-view human activity recognition". In: *PLoS ONE* 17.1 January (2022), pp. 1–20. ISSN: 19326203. DOI: 10.1371/journal.pone.0262181. URL: http://dx.doi.org/10.1371/journal.pone.0262181.

[29] Shruti Vyas, Yogesh S. Rawat, and Mubarak Shah. "Multi-view Action Recognition Using Cross-View Video Prediction". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12372 LNCS (2020), pp. 427–444. ISSN: 16113349. DOI: 10.1007/978-3-030-58583-9_26.

[30] Daniel Weinland, Remi Ronfard, and Edmond Boyer. "Free viewpoint action recognition using motion history volumes". In: *Computer Vision and Image Understanding* 104.2-3 SPEC. ISS. (2006), pp. 249–257. ISSN: 10773142. DOI: 10.1016/j.cviu.2006.07.013.

[31] Nikolaos Gkalelis, Hansung Kim, Adrian Hilton, Nikos Nikolaidis, and Ioannis Pitas. "The i3DPost multi-view and 3D human action/interaction database". In: *CVMP 2009 - The 6th European Conference for Visual Media Production* November (2009), pp. 159–168. DOI: 10.1109/CVMP.2009.19.

[32] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song Chun Zhu. "Cross-view action modeling, learning, and recognition". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2014), pp. 2649–2656. ISSN: 10636919. DOI: 10.1109/CVPR.2014.339. arXiv: 1405.2941.

[33] Amir Shahroudy, Jun Liu, Tian Tsong Ng, and Gang Wang. "NTU RGB+D: A large scale dataset for 3D human activity analysis". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-Decem (2016), pp. 1010–1019. ISSN: 10636919. DOI: 10.1109/CVPR.2016.115. arXiv: 1604.02808.

[34] Zihui Guo, Yonghong Hou, Pichao Wang, Zhimin Gao, Mingliang Xu, and Wanqing Li. "FT-HID: a large-scale RGB-D dataset for first- and third-person human interaction analysis". In: *Neural Computing and Applications* 35.2 (Jan. 2023), pp. 2007–2024. ISSN: 0941-0643. DOI: 10.1007/s00521-022-07826-w. URL: https://link.springer.com/10.1007/s00521-022-07826-w.

[35] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. "View Adaptive Neural Networks for High Performance Skeleton-Based Human Action Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.8 (2019), pp. 1963–1978. ISSN: 19393539. DOI: 10.1109/TPAMI.2019.2896631. arXiv: 1804.07453.

[36] Cunling Bian, Wei Feng, Fanbo Meng, and Song Wang. "Global–local contrastive multiview representation learning for skeleton-based action recognition". In: *Computer Vision and Image Understanding* 229.November 2022 (Mar. 2023), p. 103655. ISSN: 10773142. DOI: 10.1016/j.cviu.2023.103655. URL: https://doi.org/10.1016/j.cviu.2023.103655%20https://linkinghub.elsevier.com/retrieve/pii/S1077314223000358.

[37] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. "Self-Supervised Learning of 3D Human Pose using Multi-view Geometry". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[38] Tanmay Nath, Alexander Mathis, An Chi Chen, Amir Patel, Matthias Bethge, and Mackenzie Weygandt Mathis. "Using DeepLabCut for 3D markerless pose estimation across species and behaviors". In: *Nature Protocols* 14.7 (July 2019), pp. 2152–2176. ISSN: 1754-2189. DOI: 10.1038/s41596-019-0176-0. URL: http://dx.doi.org/10.1038/s41596-019-0176-0%20https://www.nature.com/articles/s41596-019-0176-0.

[39] Andrew K Dickerson, Zachary G Mills, and David L Hu. "Wet mammals shake at tuned frequencies to dry". In: *Journal of the Royal Society Interface* 9.77 (2012), pp. 3208–3218.

[40] P Bedard and CJ Pycock. "'Wet-dog'shake behaviour in the rat: a possible quantitative model of central 5-hydroxytryptamine activity". In: *Neuropharmacology* 16.10 (1977), pp. 663–670.

[41] Allan V Kalueff, Adam Michael Stewart, Cai Song, Kent C Berridge, Ann M Graybiel, and John C Fentress. "Neurobiology of rodent self-grooming and its value for translational neuroscience". In: *Nature Reviews Neuroscience* 17.1 (2016), pp. 45–59.

[42] Hirofumi Arai, Masaya Shigemoto, and Kiyohisa Natsume. "Detection of the change in characteristics of self-grooming by the neural network in the latent period of the Rat Kainate Epilepsy model". In: *SICE Journal of Control, Measurement, and System Integration* 15.2 (2022), pp. 64–70.

[43] G Sperk, H Lassmann, H Baran, F Seitelberger, and O Hornykiewicz. "Kainic acid-induced seizures: dose-relationship of behavioural neurochemical and histopathological changes". In: *Brain research* 338.2 (1985), pp. 289–295.

[44] Maxime Lévesque and Massimo Avoli. "The kainic acid model of temporal lobe epilepsy". In: *Neuroscience & Biobehavioral Reviews* 37.10 (2013), pp. 2887–2899.

[45] Ludmyla Kandratavicius, Priscila Alves Balista, Cleiton Lopes-Aguiar, Rafael Naime Ruggiero, Eduardo Henrique Umeoka, Norberto Garcia-Cairasco, Lezio Soares Bueno-Junior, and Joao Pereira Leite. "Animal models of epilepsy: use and limitations". In: *Neuropsychiatric disease and treatment* (2014), pp. 1693–1705.

[46] Eddie Wei. "Assessment of precipitated abstinence in morphine-dependent rats". In: *Psychopharmacologia* 28 (1973), pp. 35–44.

[47] Yoshihisa Kitamura, Kazuhiko Shibata, Kozue Akiyama, Shizue Kimoto, Yoshika Fujitani, Kouhei Kitagawa, Hirotaka Kanzaki, Mamoru Ouchida, Kenji Shimizu, Hiromu Kawasaki, et al. "Increased DOI-Induced Wet-Dog Shakes in Adrenocorticotropic Hormone–Treated Rats Are Not Affected by Chronic Imipramine Treatment: Possible Involvement of Enhanced 5-HT2A–Receptor Expression in the Frontal Cortex". In: *Journal of pharmacological sciences* 106.1 (2008), pp. 100–106.

[48] Jennifer L Hellier, Peter R Patrylo, Paul S Buckmaster, and F Edward Dudek. "Recurrent spontaneous motor seizures after repeated low-dose systemic treatment with kainate: assessment of a rat model of temporal lobe epilepsy". In: *Epilepsy research* 31.1 (1998), pp. 73–84.

[49] Yixiao Guo, Jiawei Liu, Guo Li, Luo Mai, and Hao Dong. "Fast and Flexible Human Pose Estimation with HyperPose". In: *ACM Multimedia* (2021). URL: `https://github.com/tensorlayer/hyperpose`.

[50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[51] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. "HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation". In: *CVPR*. 2020.

[52] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. "Ssd: Single shot multibox detector". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer. 2016, pp. 21–37.

[53] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. "Mobilenetv2: Inverted residuals and linear bottlenecks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520.

[54] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. "Microsoft coco: Common objects in context". In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer. 2014, pp. 740–755.

[55] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).