Automated Rodent Behavior Recognition:
Machine Learning Strategies, Behavioral Challenges and Practical Solutions

Elsbeth A. van Dam

DONDERS SERIES

RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

Automated Rodent Behavior Recognition:

Machine Learning Strategies, Behavioral Challenges and Practical Solutions

Elisabeth Aafje van Dam

Radboud Dissertations Series:
ISSN: 2950-2772 (Online); 2950-2780 (Print)

Published by Radboud University Press, Nijmegen

ISBN: 978-94-6515-086-4
DOI: 10.54195/9789465150864
Free download at: https://doi.org/10.54195/9789465150864

Cover image: GPT-4 with DALL·E 3
Cover design: Proefschrift AIO/ Guntra Laivacuma
Printing: DPN Rikken/Pumbo

Financial support for the printing of this thesis was kindly supported by Radboud University

RADBOUD
UNIVERSITY
PRESS

Automated Rodent Behavior Recognition:

Machine Learning Strategies, Behavioral Challenges and Practical Solutions

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op
maandag 16 juni 2025
om 10:30 uur precies

door

Elisabeth Aafje van Dam

geboren op 15 juni 1969
te Doetinchem

Promotoren:

Prof. dr. M.A.J. van Gerven
Prof. dr. L.P.J.J. Noldus

Manuscriptcommissie:

Prof. dr. R.J.A. van Wezel
Prof. dr. J.R. Homberg
Dr. J.C. van Gemert (Technische Universiteit Delft)

Contents

Chapter 1

General introduction

1.1 Measuring rodent behavior

Animal behavior studies have been instrumental for our understanding of biological and psychological phenomena for a long time (Aristotle, 384-322 BC; (Darwin, 1872; Tinbergen, 1951). Given their genetic similarities with humans, rats and mice serve as integral models for human diseases, allowing researchers worldwide to probe potential drug therapies for psychiatric and neurological disorders. The behavior of transgenic rodents provides valuable insights into the genetic underpinnings of brain disorders, and the function of specific proteins and genes.

Measuring animal behavior objectively is crucial for reliable and reproducible assessments, for instance when evaluating the effectiveness or safety of medicines. Long-term observation is important to monitor the well-being of laboratory animals and obtain as much data as possible, thereby reducing the total number of animals needed in the experiments. However, manual recording of behavior by human observers is time-consuming (and therefore costly), tedious and error-prone, making it unsuitable as input for systems that perform continuous, unsupervised operation. Therefore, there has been an ongoing effort in the scientific community to develop techniques for automated measurement of behavior.

Since the 1990s, we have seen automated behavioral observation tools evolve (Baran et al., 2022) from electro-mechanical devices such as infrared photobeam sensors that record the amount of activity of a laboratory rat in a test arena, to early video analysis systems that monitor the location and direction of the movement (Spruijt et al., 1992; Noldus et al., 2001). After that, behavior measurement from video evolved with multiple body-point tracking and with quantification of behavior classes derived from the animal's pose (Rousseau et al., 2000). At the same time, the need was expressed to detect more high-level behaviors like 'rearing' and 'grooming', and for assessing behavior in more natural conditions and for longer durations, for better behavior recognition performance and for tools to analyse behavioral patterns(Spruijt and De Visser, 2006). This led to the development of high-throughput behavior analysis software in 2012 that is described in **Chapter 2** of this thesis. The method is intended for long duration recordings of a rodent in a home cage and is able to recognize ten different categories of behavior from a continuous stream of video input (Van Dam et al., 2013).

Such automated systems provide a quick, consistent annotation with an accuracy comparable to human annotators, while being immune to the bias, drift, and limitations inherent in human observers. These advancements have become more significant with the increase of large recording datasets and computational hardware capacity, which has propelled progress in automated behavior measurement across animal species in general. In essence, automated assessment of rodent behavior from camera recordings has become

indispensable in behavioral research, enhancing the speed of rodent behavior analysis and helping to ensure more consistent behavior annotations.

However, available systems are still restricted in the sense that they either have a limited interpretation of the input signals (measuring for instance low level features such as the amount of activity or displacement, or simple activities such as walking), or in the sense that they are tailored to a specific research setup, or required extensive development efforts and large amounts of training data. As a consequence, in frequent situations where no off-the-shelf measurement systems are available for the specific research setup or where the behaviors that need to be observed are more complex, researchers are left with the laborious and error-prone method of manually scoring their data. These problems impede progress in areas where the interpretation of behavior is a core objective.

This thesis covers seventeen years of research, at first instance to find a method of automatically annotating the rodent behaviors that are most relevant for behavioral neuroscience research, and following that, to improve recognition to be more accurate, more robust, more generic and more flexible. It starts in 2007, before neural networks were rediscovered and before the amount of artificial intelligence (AI) hardware and tools exploded to what is available now, in 2024. The field of computer vision and pattern recognition has expanded to the field of AI, with enormous progress in image and text processing to serve diverse, specific tasks, to an extent that many experts in the field believe that artificial general intelligence (AGI), with generic cognitive capabilities and perhaps even understanding, is within reach (Morris et al., 2023; Roser, 2023). What progress has this brought so far to the field of rodent behavior recognition? What are the limitations that are shared between the different methods, and can we identify their causes, the reasons why the limitations occur? And is it possible to combine the strengths of humans and AI to be both flexible and efficient while remaining accurate?

1.2 Automated behavior recognition basics

Automated behavior recognition strives to infer behaviors or specific types of activity from a stream of sensor data. There are many ways to infer behaviors from input data. In general, the process consists of multiple consecutive steps (see Figure 1.1): preprocessing the input data to ensure the right input quality and format, deriving features from the input data to extract relevant information, preprocessing the features to reduce the dimensionality, classification of the features by mapping them onto one of the behavior categories and finally, applying post-processing to smoothen the result or correct inconsistencies. Each of the stages can be either simple, or can be an advanced system by itself such as tracking the animal in a video to derive location and trajectory features or a deep learning classification network for the detection of the behaviors.

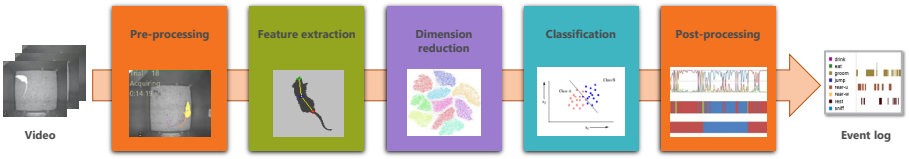


Figure 1.1: General diagram showing the stages of the automated behavior recognition process, from the raw video input to the resulting event log that contains the behavioral events.

With machine learning, tuning and optimization of feature extraction and classification can be automated using algorithms that learn and improve from experience. In this thesis, several types are applied. If an algorithm can compare the actual outcomes to ideal outcomes that have been annotated in advance (ground truth), it is considered supervised learning. If the algorithm only uses the data itself, for instance to group or segment parts of the data based on its distribution or pattern, it is called unsupervised learning. Optimization of algorithms can also be done using with only a small proportion of the ground truth annotation (semi-supervision). There is a body of work dedicated to estimating the desired output from limited amounts of data. One of those is active learning that seeks to find the most informative parts of the data that it needs annotation for in order to optimize its learning and improve its predictive performance. Finally, there is self-supervised learning where the algorithm uses part of the data to reconstruct or predict the remaining part of the input data. It is mostly used to learn a lower-dimensional representation of the input data, preserving only the relevant information needed for reconstruction (auto-encoding) or, in case of timeseries, to predict the future data (auto-regression).

Another important distinction between behavior classification methods is the way that time is handled. Behaviors are typically temporal and usually cannot be inferred from still images or body postures alone. The temporal aspect can be taken into account by either incorporating surrounding context data in the features, in classification, or both.

1.3 Developments in automated behavior recognition technology

1.3.1 Progress in video processing and activity detection

Around 2015, the field of automated behavior recognition, or activity detection as it was called in the human domain, was shifting from traditional computer vision and from machine learning to deep learning. Traditional methods such as the Automated Behavior Recognition (ABR) system described in **Chapter 2** require handcrafted feature detection, which makes use of domain-specific knowledge to extract the relevant information from the data. The quality of the detection relies mainly on the quality of the features, which is mostly influenced by the quality of the tracking. Traditional tracking from video is based on computer vision techniques, especially on background detection and active shape modeling. Given carefully designed and normalized features, it is possible to learn to estimate activities from a relatively small number of around 50 examples, for instance with statistical approaches like Linear or Quadratic Discriminant Analysis (LDA and QDA), mostly preceded by dimension reduction like Principle Component Analysis (PCA), or projective methods like Support Vector Machines (SVM) or decision trees like Random Forests (RF).

With deep learning, feature extraction is automated. Especially in image processing, convolutional neural networks (CNNs) have made the detection of objects much more flexible and efficient. For temporal data, a combination of convolutional (2D and 3D-CNNs) and recurrent neural networks (RNNs), including Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) became popular, detecting activities in short video clips or in wearable sensors data streams. In 2018, Transformer models were first introduced in text processing with the BERT model (Devlin et al., 2018), with the benefit of better handling of long-range dependencies. These successes were the prelude of a new generation of Generative Pretrained Transformer (GPT) models such as OpenAI's natural language processing chatbot ChatGPT (OpenAI, 2022a). For images, the text-to-image model Dall-E was released (OpenAI, 2022b)). Very recently in 2024, Sora was announced

for text-to-video generation ((Brooks et al., 2024)). The focus of video generation is still largely on 3D-consistency during camera movements with very limited capabilities for variation in activities, although simulations of walking, swimming and eating individuals look impressive. Automated generic modeling and generation of natural, diverse behavior in videos is still a far away goal.

1.3.2 Rodent behavior recognition

In the domain of rodent behavior recognition, several traditional systems are described in the literature (for instance VSAmbr (Jhuang et al., 2010), ABR (Van Dam et al., 2013), JAABA (Kabra et al., 2013), idTracker (Pérez-Escudero et al., 2014), Moseq (Wiltschko et al., 2015)). The ABR system described in **Chapter 2** of this thesis has been part of the commercial EthoVision XT video tracking system that Noldus Information Technology offers for rodent behavior research (`www.noldus.com/ethovision`). The first version of the ABR software was applicable for rat behavior and was released in 2012 as part of EthoVision XT 10. In 2014, the software was extended with a module for the recognition of mouse behavior and was additionally adapted to handle videos from multiple frame rates as well as live recordings with real-time, simultaneous inference for up to eight cages. The rat and the mouse behavior recognition modules are currently used in more than 300 academic and industrial laboratories around the world. Although a strong effort was made to make the modules robust and although they perform well in varying circumstances, they are not sufficiently flexible to handle all of the use cases of rodent behavior researchers.

Since 2016, deep learning approaches have been published for rodent behavior recognition (Eyjolfsdottir et al., 2016). Deep learning has been especially beneficial in detection and tracking of rodents, with for instance with open source tools such as DeepLabCut (Mathis et al., 2018) and SLEAP (Pereira et al., 2022). SimBA (Goodwin et al., 2024) and BehaveNet (Batty et al., 2019) were the first open source tools to add behavior analysis. A recent survey describes the features and pros and cons of multiple available open source tools (Isik and Unal, 2023). As is clear from the growing number of publications, the field of rodent behavior recognition is flourishing and will be for the coming years, as new AI methodologies and techniques are being explored.

1.4 Outline of the thesis

This thesis describes how rodent behavior can be annotated automatically, and explores how this can be improved to be more robust, more generic and more flexible. The recognized behaviors are restricted to the behaviors performed by a singly-housed rodent (rat or mouse) and recorded by a camera that is placed above the animal cage with constant background and infrared lighting. Examples of behavior are for instance ‘walk’, ‘groom’ or ‘scratch’. **Chapter 2** describes the classical approach, namely supervised classification of generic, hand-crafted features derived from the videos after tracking. It presents the Automated Behavior Recognition (ABR) system for the recognition of various specific rat behaviors that are the most annotated behavior by hand: ‘drink’, ‘eat’, ‘groom’, ‘jump’, ‘rear unsupported’, ‘rear wall’, ‘rest’, ‘sniff’, ‘twitch’ and ‘walk’. In this study, ABR is validated on an unseen dataset by comparison with manual behavioral scoring by an expert. The effects of drug treatment on certain behavioral categories were measured and compared for both analysis methods. **Chapter 3** explores if more generic classification of rodent behavior is possible by inferring rat behaviors directly from the video frames in an end-to-end manner, using deep learning. The performance is evaluated within and across experimental setups. It shows that using a 3D-convolutional network in conjunction with data augmentation strategies improves within-setup dataset performance over

the traditional ABR system. However, it also shows that improvements do not transfer to videos in different experimental setups. Finally, possible causes and cures are discussed. **Chapter 4** elaborates on the main reasons why rodent behavior recognition does not reach 100% accuracy in general and performs poorly on certain specific behaviors. Three aspects of behavior dynamics are distinguished that are difficult to automate. These aspects are isolated in an artificial dataset and with the artificial data, the results are reproduced using state-of-the-art behavior recognition models. These newer models use self-supervised learning to first generate a lower-dimensional representation of the data before classification.

The last research chapter, **Chapter 5**, elaborates on the practical solutions and tools that help behavior researchers annotate new behaviors for which no supervised classifier was previously trained, allowing for more flexible classification by tailoring the annotation to the needs of a particular research experiment with specific behaviors that the researcher wants to investigate, in new experimental setups. The first part of the chapter, **Chapter 5a** proves the robustness and generic applicability of the ABR features by using them to classify 'scratch' behavior of mice in two different datasets, from high-speed video recordings. In the second part, **Chapter 5b**, the possibility is investigated to combine manual annotation with AI assistance into a hybrid solution, in such a way that the manual annotation process is more efficient than a fully manual annotation and that the end result is more precise than a fully automated annotation result. The benefit of active learning is presented on the behaviors 'stretched attend' and 'unsupported rearing'.

The thesis concludes with a general discussion in **Chapter 6**, in which the contributions of the research chapters are highlighted and in which the shortcomings and future solutions are explained. It furthermore gives an outlook on new ways of behavioral analysis that take advantage of fully unsupervised detection of behavioral effects and allow behavioral researchers to explore their data and the behavioral effects between experiment groups interactively. The discussion chapter ends with a reflection on the ethical implications of rodent behavior recognition research.

Chapter 2

An automated system for the recognition of various specific rat behaviors

This chapter has been published as E.A. van Dam, J.E. van der Harst, C.J.F ter Braak, R.A.J. Tegelenbosch, B.M. Spruijt and L.P.J.J. Noldus (2013). An automated system for the recognition of various specific rat behaviors. *Journal of Neuroscience Methods*, 218(2), 214–224. https://doi.org/10.1016/j.jneumeth.2013.05.012

2.1 Introduction

Rats and mice are widely used as models for human diseases, and their behavior is studied in laboratories around the world to find new drugs for psychiatric and neurologic disorders. Furthermore, the behavioral phenotype of transgenic rodents is used as a read-out in the search for the genetic basis of brain disorders and to reveal the underlying functional role of proteins and genes. Difficulties in the reproducibility and reliability of behavioral data have been known for a number of years (Crabbe et al., 1999; Würbel, 2002; Wahlsten et al., 2003). One of the primary sources of difficulty is the limited sustained attention of human observers, especially under dimmed light conditions, resulting in predominantly short-lasting behavioral observations. Golani and colleagues showed that a very precise ethogram and consistent time and space conditions are crucial to describe animal behavior accurately (Drai et al., 2001; Fonio et al., 2009; Benjamini et al., 2011).

For the tracking and analysis of rodent location, body contour and mobility, computer vision systems exist that observe animals in real time from an overhead infrared-sensitive video camera. A summary of home-cage testing systems based on computer vision and other sensor techniques was provided by Spruijt and De Visser (2006).

For the analysis of more specific body postures and behavioral patterns, however, researchers still rely on human observation. However, manual annotation is labour intensive, error-prone and subject to bias as a consequence of individual interpretation. In contrast, automated annotation is repeatable, objective and consistent, and it saves time and effort.

Research in behavior recognition from video mainly focuses on human activities. During the past decade, many methods have been proposed to recognise activities such as 'walking', 'waving' or 'punching' (Aggarwal and Ryoo, 2011). Rodents do not have rigid limbs that make behaviors look different and, hence, easier to distinguish; the behaviors that interest biologists and neurologists can be very subtle. There is not a clear difference in animal posture or movement intensity between 'eating' and 'grooming snout' or between 'drinking' and 'sniffing the drink nipple'. Moreover, because rodents are nocturnal animals, their behavior is preferably studied under dimmed or infrared light. This means

that the automated system cannot use color information, which is an important cue in human tracking. Conversely, there are many difficulties in human activity recognition that are not present in animal lab recordings; cameras and backgrounds are static and stable, and occlusions can be avoided.

In the literature, a few systems have been described that can automatically recognise animal behaviors that are more complex than locomotion and pose. For instance, Dankert et al. (2009) used action detection in the recognition of aggression and courtship behavior of insects. For rodents, Rousseau et al. (2000) were the first to show that the detection of specific behaviors was possible. They applied neural network techniques to recognise nine solitary rat behaviors from body shape and position, recorded from the side-view. The behavior of 63.7% of the frames was correctly recognised compared to human-annotated ground truth. In 2005, Dollár et al. (2005) recognised mouse behavior from the classification of sparse spatio-temporal features, reaching an accuracy of 72%. Steele et al. (2007) used alterations in home-cage behavior for detecting perturbations in neural circuit function based on pose estimation. In 2010, Jhuang et al. (2010) predicted mouse strain type with an accuracy of 90% by comparing the relative frequencies of eight automatically detected behaviors. The features that they used were generated based on a computational model of motion processing in the human brain, followed by classification using a Hidden Markov Model Support Vector Machine (SVMHMM). They achieved an overlap between the generated 8-class behavior annotation and human-annotated ground truth of 77.3%. This is a considerable result that is on par with human annotation, which had a measured agreement of only 71.6% according to the same article. Poor inter-observer agreement is a well-known problem reported by List et al. (2005), among others, who also addressed the difficulties of performance evaluation when the ground truth is ambiguous.

Recently, Burgos-Artizzu et al. (2012) created a system for the recognition of the social behavior of mice, from both top and side views; this system included the solitary behaviors 'clean', 'drink', 'eat', 'up' and 'walk'. Their approach was based on spatio-temporal and trajectory features and was extended with a temporal context model. They calculated the performance not as the percentage agreement over all video frames, but they instead took the average recognition rate per behavior to account for the imbalance in behavior frequencies. The average recognition rate over 13 behavior classes was 61%. They measured a human inter-observer agreement of 70%. The authors remarked that human disagreement was almost entirely associated with the labelling of 'other' behaviors, whereas the automatic approach made more mistakes discerning among the specified behaviors. Removing 'other' from their performance measurement resulted in a human recognition rate of 91% and an automated recognition rate of 66%.

All of these systems show that, in principle, it is possible to recognise rodent behavior from video footage. However, the current systems have limitations. Most importantly, for all of the systems, changes in experimental setup, such as cage layout and position or camera distance and resolution, require re-training the classification algorithms. Some behaviors are restricted to location either due to the small cage or by definition. For example, for Jhuang et al. (2010) eating behavior could only be detected close to the feeder. However, rodents often take pieces of food to eat elsewhere in the cage. It cannot be excluded that the classification relies on location for these behaviors as location is part of the features in these systems. The second limitation is that not everything can be observed from the side view. Although the side view provides a better perspective for some behavior bouts, other episodes where the animal is facing away from the camera are difficult to observe, and even the manually annotated ground truth has to be estimated from uncertain clues. Finally, there is a risk in training a behavioral system using a Hidden Markov Model, in which the state transition probabilities are learned from the training sequence. For drug-treated animals, the behavioral transition probabilities are likely to be

altered. These changed transitions are a result of the experiment, not part of the model, and researchers will want to analyse the altered transition data.

A common feature of all the studies mentioned above is that training and testing videos are recorded in exactly the same setup. With the system presented here, we take recognition a step further by generalising the applicability to robust detection in videos with a setup not seen before by the algorithm. The variations in setup concern the animal size, strain, camera distance, illumination, cage layout, and cage background.

The structure of the paper is as follows. In Section 2.2, we describe the technical aspects of the proposed Automated Behavior Recognition (ABR) system, followed by a description of the two-way validation. First, we perform a straightforward frame-by-frame comparison of ABR with frame-accurate manual annotation. We evaluate videos recorded in the same setup as the training videos as well as on videos recorded in a different setup. Second, we perform an experimental study to validate ABR on a large set without the need to supply frame-accurate manual annotation. For this, we compared drug treatment effects detected by ABR to those detected by human observation. Rats are treated with two types of psychopharmaca that are well-known for their effects on behavior: a stimulant drug (Amphetamine) and a sedative drug (Diazepam). Pharmacological validation is achieved by analysing the type and direction of the drug effects detected by both methods, as well as a comparison of the behavior frequencies and durations across 5-minute intervals. The results of the two validation methods are in Section 2.3. We present the study conclusions in Sections 2.4 and 2.5.

2.2 Materials and methods

2.2.1 Rat behavior recognition system

In this study, image processing, machine learning and pattern recognition techniques are combined to create a system for automated behavior recognition in rats. The ABR system can categorise video data into behaviors: 'drink', 'eat', 'groom', 'jump', 'rear-unsupported' (standing on hind legs), 'rear-wall' (standing on hing legs with front paws leaning against the wall), 'rest', 'sniff', 'twitch', and 'walk'. These are the categories that are currently annotated by hand in neurobehavioral research protocols and from which the cognitive, motivational and emotional state of the animal is indirectly inferred. The system can be deployed online using a video stream; there is no need to store the entire video or perform multiple iterations to obtain the annotation. The only information needed from the user is the boundaries of the cage region in the video image and the animal size. Information about the cage layout such as the location of the feeder, drink nipple and the wall is preferred but not necessary. ABR uses an overhead camera view as opposed to the side-view camera that is applied in other systems. Behavior recognition from an overhead camera is technically more challenging because body postures are less prominently related to specific behaviors such as 'rearing', but it offers significant advantages. With side-view observation, the visibility of specific behaviors depends on the orientation of the animal relative to the camera; multiple cameras are needed to capture all views, and researchers must ensure that the backgrounds of all views are blank and remain constant. A top-view camera allows multiple locomotion parameters to be recorded (e.g., velocity, distance travelled, and walking pattern), and it allows the drink bottle, feeder and operant devices to be mounted to the cage. This is more practical in the lab and can even be applied in rack-mounted home-cage systems. For an example of the classification performed by ABR, see the supplementary video 'ABR demo'. Supplementary material related to this article can be found, in the online version, at `http://dx.doi.org/10.1016/j.jneumeth.2013.05.012`.

Training set

The system was trained using a high-quality dataset recorded at Janssen Research and Development in Beerse, Belgium. The dataset consisted of 25.3 video hours of six Sprague-Dawley rats individually housed in a PhenoTyper® 4500 cage (`http://www.noldus.com/phenotyper`, Noldus Information Technology, Wageningen, Netherlands) (Figure 2.1) at 720×576 pixel resolution, 25 frames per second and with infrared lighting. Subsets of these recordings were annotated by a trained observer using The Observer XT 10.0 annotation software, leading to a dataset of 254,652 frames comprising 13 behavior classes and 37 behavioral elements. Too few instances of the behaviors 'dig' and 'gnaw' occurred to be effectively used in training, and these behaviors were labelled as 'other' behavior. To reduce noise and enable frame-by-frame evaluation, the annotations have been made frame-accurate. Manual annotation takes time and concentration; a trained expert can score approximately two hours in succession with this concentration. Frame-accurate annotation took one hour for every five minutes of video.

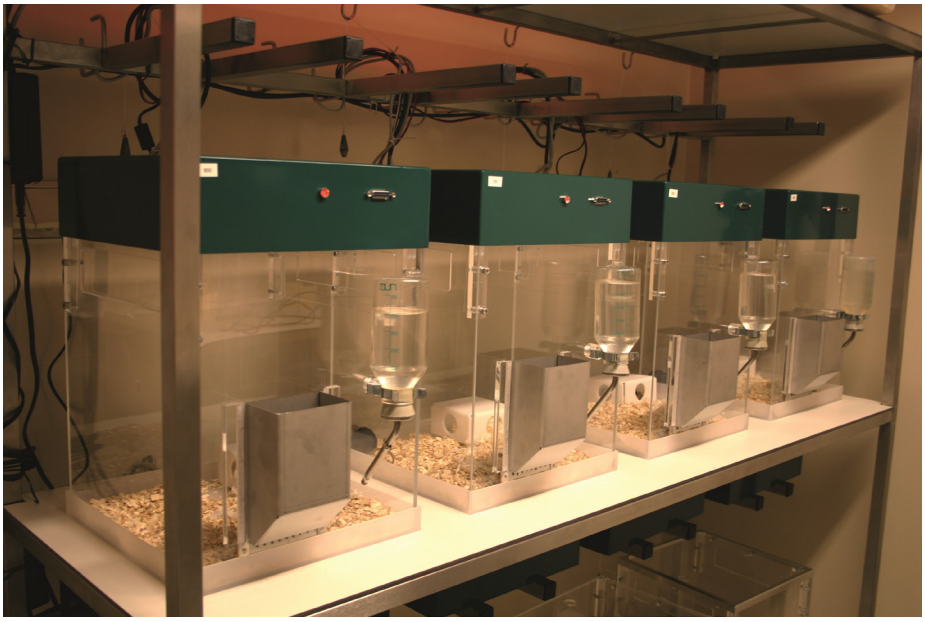


Figure 2.1: PhenoTyper 3000 cages

For the training dataset, all frames with unspecified or unusual behavior were removed; for instance, one case involved the animal grooming while lying on its back. Additionally, frames where features contained missing data were removed from the dataset. For every remaining behavior class, a random sub-selection of 3000 frames was made. For the distribution estimation of zone distances, bout duration per behavior and minimal gap sizes between behavior bouts of the same category, all frames of the training set were used.

Features for behavior recognition

From the video data, two feature groups were generated, with a total of 169 features per video frame: tracking features and motion features. Each of these features was carefully chosen to capture different aspects of the animal posture and movement.

Tracking

From the tracking analysis, the following features were generated:

- Movement of the animal's centre of gravity, nose point and tail base
- Animal body contour features (shape and area)
- Zone information: distance to feeder, drink nipple and wall

The features were obtained using EthoVision XT 8.0, a video tracking system developed by Noldus Information Technology (2011a,b). We used the detection parameters 'dynamic background subtraction' and 'model-based tracking'.

Motion features

From the motion analysis, the following features were generated:

- Motion statistics at multiple sliding temporal windows
- Motion intensity
- Motion periodicity

The drawback of including temporal information in the features is that near the temporal boundaries of the behavior bouts, the features become dependent on the previous or following behavior bouts. This is especially the case for short-lasting behaviors, such as 'twitch', where there is a risk of training on the context instead of on the characteristics of the behavior itself. The motion features were estimated by calculating the optical flow of adjacent frames at regions of interest on the animal body using the Lucas–Kanade algorithm (Lucas and Kanade, 1981). Next, the flow vector field on the animal's contour was compressed to a motion profile line of every frame. These profile lines were stitched together to form a normalised 2D motion map of the animal over time. On this motion map, sliding time window statistics (mean, variance, range of motion intensity) were calculated. Periodicity was retrieved by the use of log-Gabor filters in the temporal direction (Field, 1987).

Classification

The generated features are classified in multiple stages. After dimensionality reduction by means of Fisher linear discriminant analysis (LDA) (Fukunaga, 1990), a set of the most distinctive behavioral elements are classified for all behaviors of interest with a quadratic classifier based on normal densities (Duda et al., 2001). From these elements, the composite behavior is deduced. 'Rearing' behavior, for instance, consists of three behavioral elements: rising, standing on hind legs and coming down. 'Grooming' behavior can consist of many elements (grooming snout, grooming fur, and scratching). 'Eat' behavior can consist of any one of the elements 'eat-at-feeder', 'eat-from-floor', or 'eat-from-hand'. Because these behavioral elements all have different characteristic postures, they are recognised separately and combined later. After classification, the probabilities of behavioral elements are known for each frame. For the location-restricted behaviors 'drink' and 'rear-wall' and the behavioral element 'eat-at-feeder', these probabilities can be modified by multiplication with the distance-to-location probability given a particular behavior; these are derived from offline training distributions for these locations and behavior combinations. This step is optional, and it is the only time that zone information is used. Temporal smoothing is applied to the modified behavioral element probabilities. This is performed based on offline trained distributions of behavior durations and durations of gaps between

Table 2.1: Diversity in test datasets

Dataset	Duration	Resolution	Setup
Video 1	13.7 min	720×576	One video from the within-setup dataset
Video 2	10.5 min	360×320	Half resolution, different sawdust
Video 3	5.0 min	720×576	Different strain (Wistar)
Video 4	5.3 min	768×576	Visible light
Video 5	2.5 min	720×576	Visible light, no sawdust

consecutive behaviors of the same type. The retrieved likelihoods are combined to ten high-level behavior classes by taking the maximum probability of the behavioral elements, with a reject threshold set to 0.25. This threshold means that all frames with a maximum class probability below this value are labelled as 'other'. As a final step, the annotation is smoothed again.

2.2.2 Measuring frame-by-frame accuracy

Test videos with a different setup

Apart from the high-quality video dataset discussed in Section 2.2.1 we also annotated four videos with different resolution, animal strain, illumination and background. The recording conditions of these videos were used to test both the flexibility and robustness of the classification system. All videos were recorded with a static overhead camera at a frame rate of 25 frames per second and with a pixel resolution such that a walking rat is at least 46 pixels long. A constant level of illumination was provided by either visible or infrared light. The lighting maintained sufficient contrast between the animal and the background. Only one animal was present in each cage, and the animal was entirely visible during the whole session. The cage layout (feeder, drink nipple, and wall) was provided. The total time of the test videos was 37.0 minutes, i.e. 55,521 frames. Figure 2.2 and Table 2.1 show the diversity of these data.

Figure 2.2: Different setups used for testing. See Table 2.1 for description.

Test set

Initially, we used all videos for training and testing and performed a general leave-one-out cross-validation. However, tests showed that overall performance dropped when noisier datasets were added to the training set. Therefore, because the high-quality dataset was large and we were especially interested in the performance on unseen datasets of lower quality, we trained only on the high-quality dataset and used the other videos for testing. As a reference, we also evaluated the ABR performance on one video from the high-quality dataset. For this test, we retrained the ABR on the remaining high-quality training sets to ensure that the classifier had not seen this video before testing. In addition, there were no recordings of the same animal in both training and test videos to avoid individual animal bias. For all test sets, the frames with missing data were removed. Missing data occur when the animal is not (entirely) visible in the arena.

2.2.3 Validation in a pharmacological experiment

Experimental setup

To validate ABR on a larger scale, we conducted a study to compare the treatment effects detected by ABR to those found by human observers. Rats were treated with two types of psycho-pharmaca [1]: a stimulant drug (Amphetamine, dose of 2.5 mg/kg) and a sedative drug (Diazepam, dose of 1.0 mg/kg). Saline treatment at an equal volume was used as a control. Table 2.2 presents an overview of the experimental schedule. The experiments were performed in adherence to the legal requirements of Dutch legislation on laboratory animals (Wod/Dutch 'Experiments on Animals Act') and were approved by an Animal Ethics Committee ('Lely-DEC'). To minimise the required number of animals, animals served as their own control; both treatments were applied to the same animal with a washout period of four days. Furthermore, each drug treatment was preceded by a saline treatment 24 hours prior to provide a separate baseline per treatment, thereby excluding a time effect. It was intentionally decided to start with the Diazepam treatment for all animals and not to use a mixed design. This decision was based on previous studies where the animals seemed to be context-conditioned after Amphetamine treatment, and a subsequent saline injection resulted in anticipatory activity (Van der Harst, personal observation). Amphetamine-induced behavioral sensitisation has been reported extensively in the literature (Vanderschuren et al., 1999; Do Nascimento Alvarez et al., 2006) also in relation to conditioning (Drew and Glick, 1988, 1990). Therefore,it was decided to administer the stimulant drug at the end of the experiment. The behavior of the animals was recorded on video for one hour post-injection. The period of 10–35 minutes post-injection, containing the peak-effects of both treatments, was analysed using both ABR and manual annotation. The validation consisted of a comparison of (the direction of) treatment effects detected by both analysis methods on certain behavioral categories (Section 2.2.3) using 5-minute intervals as the statistical units (Section 2.2.3). The behavior durations in these 5-minute intervals are graphically summarised in a biplot, which highlights the variation in duration of behaviors within and between treatments and within and between observation methods.

Animals and housing

The test subjects were four male Sprague-Dawley rats that were six weeks old (Hsd:SD, Harlan, The Netherlands) and weighed approximately 150–200 gram on arrival. After arrival, the animals were housed socially with a reversed day/night cycle (9:00–21:00 red lights on, 21:00–9:00 white lights on) in cohorts of two animals in a Macrolon IV-S cage (Techniplast, Italy) with a heightened lid, a shelter/climbing-partition object and two water bottles. Water and standard lab chow (CRM-E, SpecialDiet Services, United Kingdom) was available ad libitum and was refreshed weekly. The animals were allowed to habituate to the reversed day/night cycle, housing and management for 12–14 days before the experiments. The animals were handled on a regular basis during this habituation phase to avoid any effects of handling on the rats' behavior during the experiment. After the general habituation phase, the animals were transferred individually to the automated home cage, PhenoTyper 4500, equipped with a feeder and drink bottle. On the third day of PhenoTyper-housing, the animals were habituated to the experimental procedures to avoid any effects of these procedures (e.g., entering the room, starting camera-recording, handling the animals and placing them back in the PhenoTyper) on the behavior of the animals during the actual test. After 4 days of habituation to the new home cage, the

[1] Doses are based on dose–response tests in a related study by Dunne et al. (2007) [dose volume: 2 ml/kg, s.c.], in combination with the results from a pilot study in which a dose of 2 mg/kg Diazepam caused sedation.

Table 2.2: Schematic overview of the experimental setup – see also description in Section 2.2.3

Arrival animals	Day 0
Habituation, incl handling and fixation[a]	Day 0–15
Start experiment – housing in PhenoTyper[b]	Day 16
Habituation test procedure [2×]	Day 19
Baseline 1 (saline-1)	Day 20
Diazepam	Day 22
Re-housing (socially) in Macrolon[c]	Day 22
Habituation test procedure [2×]	Day 26
Baseline 2 (saline-2)	Day 27
Amphetamine	Day 29
Re-housing (socially) in Macrolon	Day 29

[a] During habituation to the reversed day-night cycle, housing and management, the animals are also habituated to procedures such as handling (being picked-up and held by a human) and fixation (restraining procedure to be able to inject the animals), to prevent stress of these procedures during the experiment.
[b] PhenoTyper 4500 cage.
[c] Macrolon IV-S cage.

experiment started with the first baseline recording (after saline injection) (Table 2.2). Weighing of the animals and other activities, such as cage-cleaning, were always performed after testing to prevent any influence of these procedures on the behavior of the animals during monitoring.

Annotation and behavioral categories

The video files were analysed by an annotation expert using The Observer XT 10.0 annotation software and by the ABR system. This process resulted in two datasets containing 25 min of continuous behavioral observation for each of the four treatments (Saline-Baseline-1, Diazepam, Saline-Baseline-2, Amphetamine; Table 2.2). It is important to note that in relation to the validation method (i.e., comparison of the direction of the effects as detected by human observations and by automated observation), the applied ethograms for the observation methods were not identical. The human annotator used an ethogram of 26 elements that was previously applied in behavioral research (Van der Harst et al., 2003a,b). The human annotator was very well trained in this ethogram, which supported reliable scoring. The ABR ethogram consisted of 11 elements (including an 'other' category). To facilitate a comparison of the annotation methods, some of the behaviors were grouped into categories. The two ethograms with the behavior definitions are provided in Table 2.3 and Table 2.4. The grouping of behavioral categories is presented in Table 2.5. The 'sniff' category of ABR was mapped onto more categories than the 'sniffing' categories used by human annotators because ABR used a broader definition of 'sniff', where sniffing was not restricted to nose sniffing while holding the head and body still. Some of the behavior that ABR categorised as 'sniff' was defined as 'mobile exploration' by the human observer. However, 'mobile exploration' was grouped with 'walk' by the ABR because these behaviors showed greater overlap. Another difference in definition was the human 'root/dig' category. When the rat used its entire body or moved around sawdust with either its nose or paws, ABR defined this as 'other' behavior. However, when it moved its nose around on the floor, ABR scored this behavior as 'sniff'. The same was true for the human 'gnaw/nibble' category, which additionally overlapped with ABR 'eat'. There was no ambiguity between the two ethograms for the categories 'drink', 'groom', 'rear' and 'rest'.

Table 2.3: Ethogram of the observed behavioral elements, derived from Van der Harst et al. (2003a)

Human	Description
Drink	Licking at the spout of the water bottle
Eat	Gnawing/eating food pellets (either from feeder or from pieces that are held by the fore-paws)
Groom	Washing the muzzle with fore-paws (including licking fore-paws) or grooming the fur or hind-paws by means of licking or chewing
Scratch	Scratching muzzle, head or body with one the hind-paws
Hop/jump	Hopping (moving forward with small hops) or jumping (big forward or upward jump(s))
Jerk	Sudden convulsive movement of the head or body
Shake	Shaking the head or entire body
Rear supported	Exploring while standing in an upright posture, leaning with front paws against the cage-wall or other object (not present in the current setup)
Rear unsupported	Exploring while standing in an upright posture, unsupported
Resting (lie and sit)	Lying or sitting without obvious exploration, including sleeping (eyes closed)
Attention	Alert posture (sitting or lying with slightly lifted head, apparently listening and/or looking around)
Sniffing air	Sniffing in the air
Sniffing other	Sniffing at sawdust, walls or other objects
Root/dig	Rooting with the muzzle or digging with the front paws in the sawdust
Gnaw/nibble	Gnawing or nibbling on sawdust, droppings or at the walls or floor of the cage
Walk/forward move	Moving forward in a clear direction (more than three steps) without obvious exploration
Mobile exploration	Exploring the surroundings (sniffing, attention) while moving forward or around
Other	Behavior other than defined in this ethogram, or when it is not visible what behavior the animal displays
Stretch	Stretching body, often in combination with stretching fore- and/or hind-paws
Stretched attend	Stretched posture with head/nose pointed forward, often in combination with one lifted fore-paw
Undefined transition	Short transition (point-event) during the display of a specific behavior, without interrupting this behavior[a]
Yawn	Yawning
Scan	Slow sideways swaying of the head and anterior part of the body (typical behavior for albino rats)
Freeze	Stiffening of the entire body, including immobility of the whiskers and auricles
Circling/chase tail	Circling around own axis or chasing tail

[a] Ambulation while rearing, change of body-position during grooming (falling backward when going from grooming muzzle to licking genitals).

Table 2.4: Ethogram of the automated behavior recognition system ABR, used for single housed rats in home cage or open field.

Behavior	Event type	Description
Drink	State	Drinking from the drink nipple
Eat	State	Eat at feeder or from floor or eating while holding food in front paws
Groom	State	Grooming snout, head, fur or genitals. Includes scratch and licking of paws during a grooming session
Jump	State	Fast displacement, taking off with both hind legs at the same time
Rear unsupported	State	Standing on hind legs unsupported. Rearing events include the rise and descend
Rear wall	State	Standing on hind legs with front paws leaning against the wall. Rearing events include the rise and descend
Rest	State	Resting without hardly moving, either sit or lying down. Includes sleeping. No interest in environment
Sniff	State	Slight movements of the head in order to gather information about the environment, possibly with slight, discontinuous displacement. The category includes: sniff air, wall, floor and other objects
Twitch	Point	Sudden and short movement of the body or head. Includes body shake, head shake
Walk	State	The rat moves to another place. Hind legs must move as well
Other	State	Any behavior other than described above

Table 2.5: Merged ethogram. The behavioral categories were merged for analysis. Categories of the human ethogram 'root/dig' and 'gnaw/nibble' were ambiguous for the ABR ethogram but have been merged into the 'explore' category.

Human	ABR	Merged
Drink	Drink	Drink
Eat	Eat	Eat
Groom, scratch	Groom	Groom
Hop/jump	Jump	Jump
Jerk, shake	Twitch	Twitch
Rear-supported	Rear wall	Rear
Rear-unsupported	Rear unsupported	
Resting (lie and sit)	Rest	Rest
Attention, sniffing-air, sniffing-other, root/dig, gnaw/nibble	Sniff	Explore
Walk/forward-move, mobile-exploration	Walk	Walk
Other, stretch, stretched-attend, undefined-transition, yawn, scan, freeze, circling/chase-tail, head-exploration	Other (including dig and gnaw)	Other

Statistical analysis

The frequencies and durations of the nine behaviors of the combined ethogram were recorded in five non-overlapping intervals of five minutes for each rat, treatment and observation method, as explained in the previous sections. This analysis was achieved by processing and integrating the data of both observation methods in The Observer XT 10.0 software. The duration of five minutes per analysis interval was chosen to maximise the level of detail and amount of information and to minimise statistical pseudo-replication and distortion due to a lack of synchrony between the logs. This process yielded five records per rat and treatment; there were four treatments and 20 pairs of records for each of the four rats, each pair consisting of an ABR record and a human-scored record. Prior to statistical analysis, durations were logarithmically transformed (natural logarithm) after the addition of the minimum non-zero value (because ln(0) does not exist). This transformation made distributions more symmetric and fit with models for percentage change on the original scale. The duration of 'twitch' was discarded because the human 'twitch' was scored as a point event, so it had no duration. For both observation methods separately, the effect of Diazepam and Amphetamine was defined as the difference of the means between the transformed data of the drug treatment and the corresponding control treatment. Data were tested for significance using a two-tailed paired t-test and a Wilcoxon signed-rank test using R (R Development Core Team, 2010). The mean on a ln-scale corresponds approximately to the median on the original scale. We also compared the annotations of ABR and human directly by Spearman rank correlation. For both tests, the level of significance was set at 0.05. Frequencies were analysed by a method designed for count data, namely, a log-linear regression assuming a quasi-Poisson distribution (Faraway, 2006; McCulloch et al., 2008). By including rat and interval combinations in the model, this regression is similar to a paired t-test for count data. The treatment effect in this model was tested with a deviance-based F-test (Faraway, 2006; McCulloch et al., 2008).

The variation in the duration of the eight state behaviors (nine minus 'twitch') in all 160 records is graphically summarised by a log-ratio principal component analysis (Aitchison and Greenacre, 2002), as implemented in Canoco 5 (Ter Braak and Šmilauer, 2012). This analysis focuses on the (logarithm of) ratios of durations and thereby avoids the problem of pseudo-correlations between behaviors. These arise because the total duration is equal to the observation interval, so if one behavior lasts longer, the others must be shorter. The result of the analysis is presented in a biplot (Greenacre, 2012) with points for records and arrows for behaviors that point in the direction of maximum increase across the diagram. In the biplot, each ABR record is connected with a line to the corresponding human observed record. Each of the 80 five-minute intervals is thus represented by a line segment. The mean positions of the records of the treatment observation method combinations are also shown and similarly connected. The resulting biplot highlights the variation in duration of behaviors within and between treatments, within and between observation methods and within and between the 80 individual five-minute intervals.

2.3 Results

2.3.1 Frame-by-frame accuracy

Interpretation of the performance figures is not trivial when comparing precision and recall results between behavior classes or between videos. The amount of 'other' data influences the result, especially for the precision calculation. In other research on action recognition, this problem is often avoided by creating a balanced sub-selection of target class data and using it to generate random training and test subsets. In this way, the behaviors in

the training and test sets are equally balanced. In our tests, we chose the most stringent validation by testing on unseen animals and video setups. This makes it impossible to use a test set that has the same behavior histogram as the training set because it is not possible to select videos where different animals behave the same.

Accuracy on videos with same setup as the training dataset

Figure 2.3 displays the correspondence of the ground truth and the generated annotation over time for reference test video 1. The overlap between manual and automatic annotation measured on the subset of target behaviors was 71%. As is clear from the confusion matrix of video 1 in Figure 2.4, the recall was rather good for most behaviors, but the recall was 0 for the 'rest' behavior. This can be explained by the fact that only one 'rest' event occurred in this video. This event was significantly shorter than normal 'rest' events, and because it was in between two 'eat' events, the 'rest' annotation was lost during the temporal smoothing. The main mistakes on this test set involved the 'drink behavior being misclassified as the 'sniff' behavior. The same misclassification was also observed for behavior that was manually labelled as 'other' ('dig', 'gnaw' and unspecified, often explorative behavior, in total 22% of the data). The 'sniff' versus 'other' confusion plays a major role in the low agreement observed for the entire video (62%). The computation time needed by ABR to label video 1 was 23 fps, i.e., near real-time.

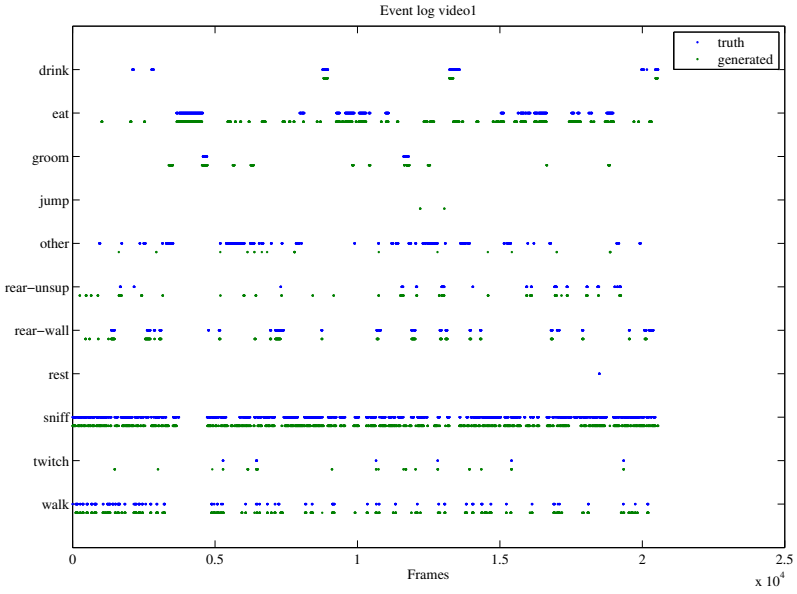


Figure 2.3: Ground truth and generated annotation over time for test Video 1. 71% of the target frames (all frames except 'other') is correctly labeled by ABR. The main confusions are 'sniff' as 'eat', 'sniff' as 'walk', 'eat' as 'sniff', 'sniff' as 'rear-unsupported', accounting for 26%, 20%, 9% and 7% of the mistakes, respectively.

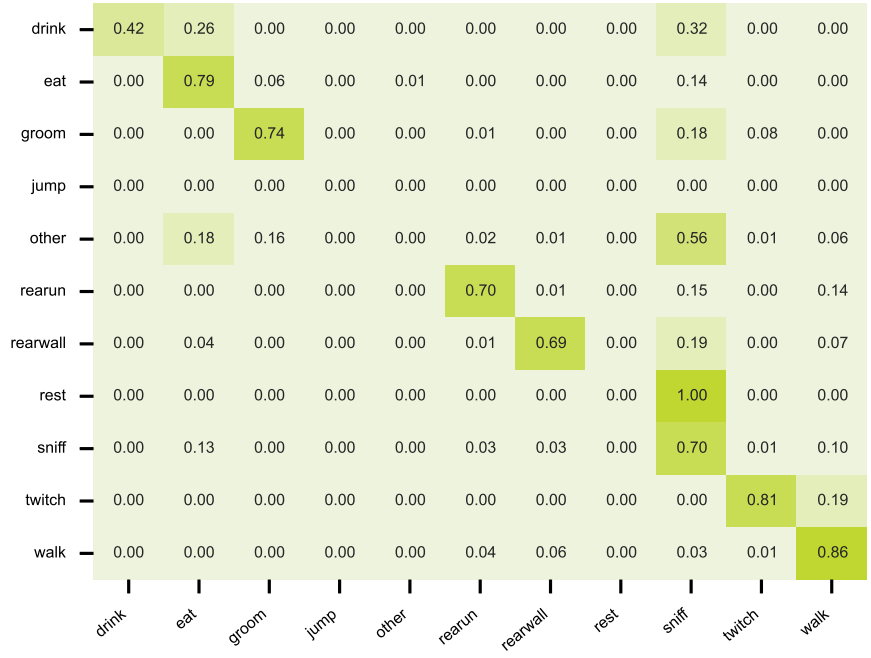


Figure 2.4: Confusion matrix of ABR results on video 1. Rows denote ground truth, the columns denote the decisions. So the first row indicates that 42% of the 'drink' behavior was scored as 'drink', 26% was misinterpreted as 'eat' and 32% was considered as 'sniff'

Accuracy on videos with a different setup

In Table 2.6, the results obtained with the other test videos are presented. It shows two percentages of overlap between manually labelled ground truth and ABR-generated annotation; the first percentage is calculated only on the subset of target frames and the second on the entire video, including 'other' behavior. Overall agreement is always lower because the 'other' class is too diverse to train on and because at the same time the 'other' class resembles the target behaviors, thus making it difficult to distinguish. The 'other' category is often scored in case of a transition between two behaviors. However, it is important to note that the results of test videos 2 to 5 are comparable to the results of reference video 1.

Table 2.6: Percentage overlap between ground truth and ABR annotation on the test videos. Since this is a 10-class problem, the chance agreement is 10% for the agreement on target class subset, and 9% for the agreement on all frames.

Test video	Overlap on target classes	Overlap on entire video, including 'other'
Video 1	71%	62%
Video 2	65%	59%
Video 3	80%	72%
Video 4	67%	60%
Video 5	65%	65%

Table 2.7: ABR recall per behavior compared to benchmark in Jhuang et al. (2010) and to Burgos-Artizzu et al. (2012). These systems do not classify behaviors 'jump', 'sniff' and 'twitch'. Class 'mmove' stands for 'micromovement' behavior that is defined by 'small movements of the animal's head or limbs'. For comparison, behaviors 'rear-unsupported' and 'rear-wall' are grouped.

Test setup	Our system		Jhuang	Human	Commercial system	Burgos-Artizzu
	Equal to training setup Video 1	Unseen Video 2–5	Same as training setup	Same as training setup	Unknown	Same as training setup
Drink	0.42	0.93	0.72	0.78	0.63	0.49
Eat	0.79	0.76	0.75	0.87	0.73	0.53
Groom	0.74	0.58	0.70	0.57	0.30	0.47
Jump	–	0.48	–	–	–	–
Mmove	–	–	0.83	0.64	0.64	–
Rear	0.70	0.68	0.70	0.78	0.35	0.62
Rest	0.00	0.16	0.94	0.95	0.96	–
Sniff	0.70	0.76	–	–	–	–
Twitch	0.81	0.75	–	–	–	–
Walk	0.86	0.60	0.55	0.68	0.69	–

Benchmark to other systems

In Table 2.7, the recall results of each behavior class are compared to the results of the side-view systems described in Jhuang et al. (2010) and the top view system of Burgos-Artizzu et al. (2012). Although one has to be cautious when comparing the results of datasets with different setups, species, ethograms and annotators, the table shows that performance is similar across systems.

2.3.2 Results of the pharmacological experiment

Frequencies of behavior

Figure 2.5 displays the mean behavior frequencies measured over the 5-min intervals for the treated animals and their control (baseline measurement after saline injection) for both treatments and both annotation methods. The frequencies measured by ABR are consistently higher than the frequencies measured by human observation. However, log-linear regression shows a clear correspondence in the significance and direction of the treatment effects ($p < 0.05$). For the Amphetamine treatment, the two methods disagree only on the behaviors 'eat' and 'twitch'. For the Diazepam treatment, the methods agree on all behaviors. Both methods found no significant differences between the two control groups, except for the behavior 'jump' according to ABR (Figure 2.7).

Durations of behavior

For both methods separately, a paired t-test and a two-tailed Wilcoxon signed-rank test were performed on the log durations per five-minute intervals, comparing behavior post-treatment with the corresponding baseline measurement (i.e., Diazepam versus saline-1, Amphetamine versus saline-2). Both tests indicated that the same behaviors were significant ($p < 0.05$, except for the 'drink' behavior in the Diazepam versus saline-1 comparison, which was not significant using the Wilcoxon test but was significant in the t-test for ABR. Table 2.8 shows an overview of the observed effects, based on the t-test. For the effects of Amphetamine, both methods detected a significant decrease in the durations of 'drink', 'groom' and 'rest' and a significant increase in the durations of 'explore', 'rear' and 'walk'

Figure 2.5: Comparison of the means of behavior frequencies over the 5-min intervals for the treated animals and their control (baseline measurement after saline injection), for both treatments and both annotation methods. For the Diazepam treatment, the methods (manual scoring and ABR) agree on the significance and the direction of the treatment effect on all behaviors frequencies ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; # trend toward significance $p < 0.1$). For the effect of the Amphetamine treatment the methods disagree only on behaviors 'eat' and 'twitch'. See Section 2.3.2.

($p \leq 0.05$ for all cases and both methods). For 'eat', there was no agreement, due to one video containing a long sequence (25 min) during which the rat sat at the feeder performing stereotypic behavior that consisted of short head movements. Both methods scored mostly 'sniff' during this period (Human 93% and ABR 72%), but ABR also scored 5% 'eat', 11%'rear' and 6% 'groom'. For the Diazepam treatment, both methods showed significant decreases in the 'drink', 'eat', 'groom' and 'rear' behaviors and a significant increase in 'rest' ($p \leq 0.05$ for all cases and both methods). For 'jump', there was agreement on the direction of the effect, but not on the significance. ABR also indicated significant decreases in 'explore' and 'walk' behaviors after Diazepam treatment, whereas the human annotation data did not confirm such an effect. The different conclusions for the effect of Diazepam on 'explore' can be explained by the grouped ethogram; in the human annotation, 'root/dig' and 'gnaw/nibble' are mapped onto 'explore', and this is in fact what the Diazepam-treated animals displayed. Human annotation did find a significant decrease in the other elements of the 'explore' category, 'sniffing' and 'attention' behavior. As for 'walk', a close look at the human annotation data revealed a rather long false positive 'walking' event (6.79 s) that lasted almost 1/3 of the entire 'walk' duration in one video. We also compared the saline-1 and saline-2 treatments to test for the stability of the behaviors in the control treatments. Neither the paired t-test nor the Wilcoxon test showed any significant differences. The minimum observed p-value in these 8×2 tests was 0.16.

We also compared the annotations of ABR and human directly. The Spearman rank

Table 2.8: Treatment effects comparing the treatment with its own baseline (i.e. Diazepam vs saline-1, Amphetamine versus saline-2) expressed as mean difference on ln-scale (with standard error of difference between means between parentheses) on the behavioral categories as found by ABR and human annotation. The p-values (NS: $p > 0.05$) result from a two-tailed paired t-test on the ln-durations per 5-min intervals. 'twitch' is excluded from this analysis since it is scored as a point event with no duration. The methods agree about significance in 12 of the 16 cases (marked cells). There were no opposite effects. For an explanation of the disagreements see Section 2.3.2.

	Amphetamine		Diazepam	
	cline2-3	cline4-5		
	ABR	Human	ABR	Human
Drink	\1.46** (0.44)	\1.74** (0.57)	\1.05* (0.46)	\1.62** (0.56)
Eat	\1.00NS (0.60)	\1.96** (0.58)	\1.10* (0.46)	\1.75** (0.52)
Explore	1.29** (0.41)	2.02** (0.42)	\1.10* (0.41)	\0.48NS (0.44)
Groom	\2.20** (0.58)	\2.21** (0.39)	\2.99** (0.59)	\1.85** (0.53)
Jump	0.20NS (0.26)	\0.15NS (0.27)	\0.71** (0.24)	\0.45NS (0.23)
Rear	3.56** (0.49)	2.40** (0.71)	\1.95** (0.49)	\2.45** (0.49)
Rest	\3.76** (0.72)	\3.09** (0.80)	2.70** (0.60)	2.75** (0.73)
Walk	1.49* (0.60)	1.75* (0.62)	\1.21* (0.53)	\0.85NS (0.79)

* $p < 0.05$
** $p < 0.01$

Table 2.9: Comparison of ABR and human annotation over all intervals. Spearman rank correlation was significant for all behavior log durations and frequencies. For instance for 'groom' and 'rear' duration correlation was 0.69 and 0.86 respectively (for n = 80 intervals (5 min), $p < 0.001$).

	Duration	Frequency
Drink	0.64	0.60
Groom	0.69	0.55
Jump	0.63	0.62
Twitch	n.a.	0.57
Rear	0.86	0.78
Rest	0.92	0.62
Explore	0.85	0.75
Walk	0.89	0.89

correlation was significant for all log durations and frequencies of behaviors. For instance, for 'groom' and 'rear', the correlation was 0.69 and 0.86, respectively (for n=80 intervals (5 min), $p < 0.001$; Table 2.9). Figure 2.6 shows the behaviors (arrows) and records (symbols) projected onto a plane formed by the two major axes. The axes were found by the log-ratio PCA on the matrix of durations of eight behavioral categories in 160 records (80 5-min intervals scored by both ABR and a human). Each ABR record in the plot is conected by a line with its corresponding human manual observation record, and the mean positions of records of each treatment and observation method combination are shown (big symbols). Most of the variation in behavior duration is due to the treatment (Amphetamine or Diazepam), as indicated by the large distance between the positions of the treatment means positions using either observation method. The biplot also confirms that the mean difference between the saline treatments (intermediate between Amphetamine and Diazepam) is small. More importantly, in the context of this paper, the mean difference between the observation methods is small, as indicated by the small distance between the mean positions for each treatment. Not only is the mean difference small, but also the lines connecting the corresponding ABR and human records are also short compared to the overall differences between records. For some intervals of the saline treatment, the difference between observation methods is relatively large, but these are still small compared to the large overall differences within the saline treatments. In contrast, records from intervals of Amphetamine treatment are similar, both between intervals and between observation methods, and the same holds true for Diazepam.

2.4 Discussion

The experimental study with drug treatment demonstrated that ABR detects similar effects on behavior that are found by human observers for both Amphetamine and Diazepam. For Amphetamine, both methods found significant decreases in the 'drink', 'groom' and 'rest' behaviors and significant increases in the 'explore', 'rear' and 'walk' behaviors. For Diazepam, both methods showed significant decreases in 'drink', 'eat', 'groom' and 'rear' and a significant increase in 'rest' behavior. ABR also found significant decreases in the 'jump', 'explore' and 'walk' behaviors.

The differences between the drug effects detected by each method are partially explained by interpretation differences. One could say that using different ethograms inevitably introduces difficulties. However, some of the behavior classes overlap and interpretative differences between different observers are inevitable. The use of different ethograms reveals, rather than introduces, the problem. For the more clearly defined behaviors such as 'groom', 'rear' and 'rest', the agreement between the methods is unquestionable. The overlapping behavior classes also explain the difference in observed frequencies. In these cases, ABR switches between behaviors, whereas humans make interpretation decisions for longer periods. This switching occurs, for instance, between the behaviors 'eat-at-feeder', 'sniff' and 'rear-wall' and between the categories 'eat-floor' and 'sniff'. Other differences in the reported drug effects are caused by inherent errors in both methods. It is important to note that the types of errors are different between the methods. The errors in ABR are always systematic; for instance, ABR may consistently misinterpret a certain behavior. For example, lifting the head and sniffing air with a pure vertical head movement but without lifting the front paws is mistaken for unsupported rearing. However, the systematic mistakes are always related to the behavior, not to the recording duration, as is likely to be the case with human annotation (e.g., due to a loss of concentration). Another mistake is that the automatic annotation of 'rest' is too conservative: very short walking bouts of only one or two small steps are ignored by the automatic annotation but are scored by the human annotator. Because the animal moved very slowly

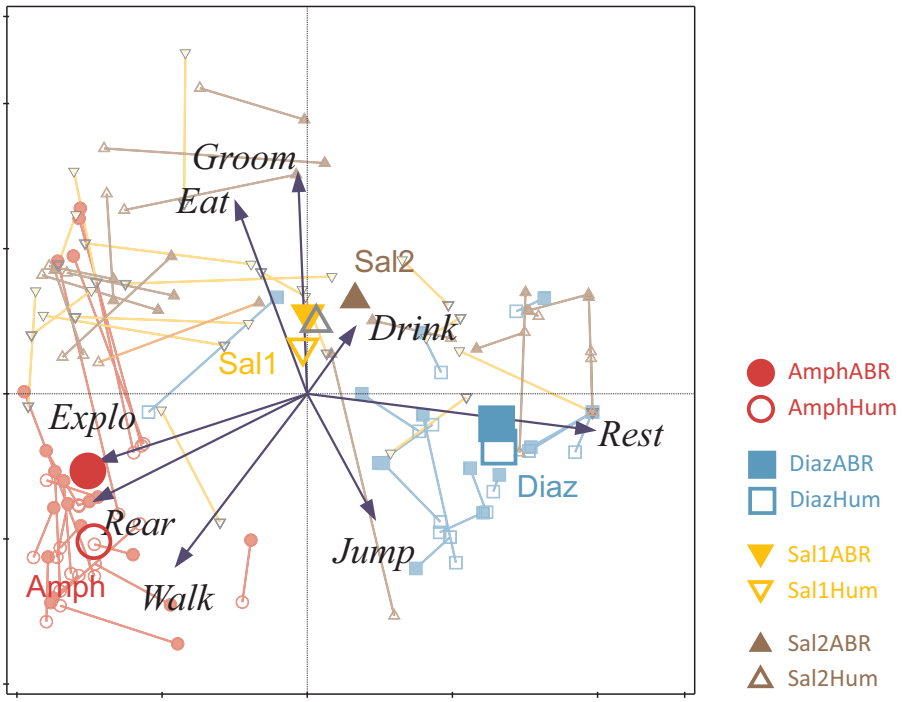


Figure 2.6: Logratio principal component biplot of durations of eight behaviors (arrows from the plot origin pointing in the direction of maximum change in duration) and 160 rat records (point symbols indicating treatment and observation method) with lines connecting each ABR observation record (closed symbol) with its corresponding human observation record (open symbol) and larger symbols indicating the mean positions of the treatment–observation method combinations. Treatments are Amphetamine (circles), Diazepam (squares), first saline treatment (down triangle) and second saline treatment (up triangle). The diagram displays 70% of the variance in the log-ratios of behavioral durations and highlights the small difference between observation methods compared to the large differences between and even within treatments. The behaviors 'explore', 'rear' and 'walk' last longest with Amphetamine (at the left hand side), 'rest' last longest with Diazepam (at the right hand side) and 'groom' and 'eat' last longest with the saline treatments (in the center of the diagram).

in this case, these movements are prominent to the human annotator, whereas the same event would be less prominent if it had occurred in the behavior sequence of a very active animal. One explanation is that human observers are very sensitive to behavior context and adjust their interpretation to the overall behavior of the animal. In some cases,this adjustment may be desirable, for instance, when an animal performs the same behavior differently than 'normal' due to treatment or genetic variance. When the deviation is large, the ABR considers it to be different from normal behavior. To account for this, either a translation needs to be made from the normal ethogram to the deviant one or the system needs to be trained to recognise the deviant behaviors. The same influence of context emerges in the annotation of behavior that is not listed in the ethogram, such as the stereotypical head movements at the feeder by a rat treated with Amphetamine. ABR

does not recognise this unseen behavior as 'other' because it is not notably different with respect to shape, motion intensity or direction; thus, ABR switches among behaviors that are most alike. The human annotator could recognise it as a new type of behavior and choose the best-fitting category for its assignment. These effects of the scoring context on the annotation itself remain an interesting subject for future studies. Further investigation will note whether ABR can reveal behavioral effects that are more subtle than the well-known effects of Amphetamine and Diazepam. A major advantage of automated behavior measurement is that it offers the opportunity to measure behavior over very long periods of time, thereby increasing the amount of data and widening the window of opportunity for the detection of effects.

2.5 Conclusions

The ABR system is able to recognise the most relevant rat behaviors in a fully automated manner. Due to the inevitable interpretation differences between human observers, 100% agreement between human and automated annotation is not feasible. However, the ABR-human correspondence is similar to human inter-observer agreement, i.e., generally acceptable to a level of 70–85% (Bateson and Martin, 2007). Unlike humans, however, the system operates with consistent and sustained attention, thereby allowing 24-hour observation of animals without human subjectivity. Automated annotation is repeatable, objective and consistent, and it saves time and effort. To our knowledge, ABR is the only system that can recognise behaviors across different setups. It therefore outperforms other known systems by offering equal reliability without the need for on-site training that requires labour-intensive hand-labelled data. ABR has the additional benefit of a more practical top-view camera position. ABR functionality will be integrated into the EthoVision software package. In the future, ABR will extend recognition to mouse behavior and social behavior and will make ABR suitable for videos of other frame rates and live recordings.

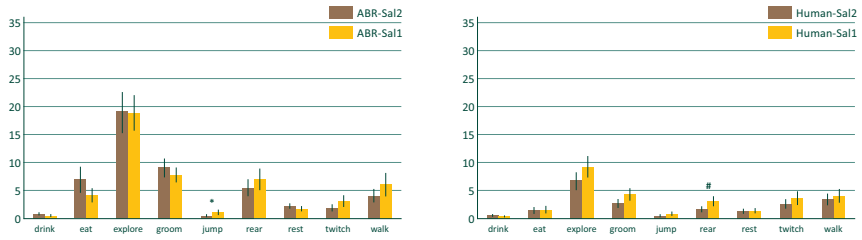


Figure 2.7: Comparison of the means of behavior frequencies over the 5-min intervals for the control groups (baseline measurement after saline injection), for both annotation methods. Both methods found no significant differences between the two control groups, except for the behaviour 'jump' according to ABR. ($^{*}p < 0.05$)

Chapter 3

Deep learning improves automated rodent behavior recognition within a specific experimental setup

This chapter has been published as E.A. van Dam, L.P.J.J. Noldus and M.A.J. van Gerven (2020). Deep learning improves automated rodent behavior recognition within a specific experimental setup. *Journal of Neuroscience Methods*, 332, 108536. `https://doi.org/10.1016/j.jneumeth.2019.108536`

3.1 Introduction

Observation and analysis of rodent behavior are widely used in studies in neuroscience and pharmacology. Laboratory rats and mice are valuable animal models for psychiatric and neurological disorders to study the behavioral effects of genetic variation, pharmacological treatment, optogenetic stimulation, and other interventions. However, manual annotation of animal behavior by human observers is labor-intensive, error-prone and subjective. Several automated systems are available that have been reported to perform on par with human annotators. They offer the advantage of quick and consistent annotation and are insensitive to bias, drift and the limited sustained attention of human observers. Yet most of them can only recognize behaviors as performed in the training material, recorded in the exact same setting as the training environment. This works fine with in standardized test cages, but in reality a lot of variation exists between rodent test environments used in different laboratories. The performance of the behaviors might also vary with treatment. There is still no adequate solution that works out of the box in the diverse real-life scenarios faced by behavioral researchers.

During the last 20 years, there have been several publications on automated rodent behavior recognition from video. One of the first to publish on this topic were Rousseau et al. (2000), who trained a neural network on pose features and reached an overall agreement of 63% on nine rat behaviors (49% average recall). Note that overall agreement is calculated over frames only, whereas average recall is the average of the proportion of correct frames per class. Subsequently, Dollár et al. (2005) used the bag-of-words approach for activity recognition with 72% average recall on five mouse behaviors. This was followed by the work of Jhuang et al. (2010), who applied a Support Vector Machine and Hidden Markov Model (SVM-HMM) on a combination of biologically-inspired video filter output and location features. They report 76% average recall for eight mouse behaviors. They also compared human to human scoring, which resulted in 72% overall agreement and 76% average recall. Van Dam et al. (2013) presented the EthoVision XT RBR system for automated rat behavior recognition that applies Fisher dimension reduction followed by a

quadratic discriminant on highly normalized, handcrafted features derived from tracking and optical flow. They addressed the importance of cross-setup validation in order to assess out-of-sample generalization, and reached 72% overall agreement and 63% average recall on ten classes for both within-setup and across-setup evaluation.

Meanwhile activity recognition research for human activities progressed tremendously, particularly with the advent of deep learning (Simonyan and Zisserman, 2014; Tran et al., 2015; Wang et al., 2016; Carreira and Zisserman, 2017; Huang et al., 2017; Tran et al., 2018). Deep neural networks allow an abstraction from input data to output categories by learning increasingly higher-level representations of the input. By feeding labeled input examples to the network, the network can compare its own output with the desired output and can amplify features that are important for discrimination and ignore irrelevant information. Deep networks vary in their architecture: the number and size of layers and the way information can flow through. Ideally, the network learns the mapping from input data to output class without any preprocessing, in a so-called end-to-end manner.

A few deep learning models have been applied to rodent behavior: Kramida et al. (2016) applied a Long short-term memory (LSTM) model to the features of the Very Deep Convolutional Networks architecture (VGG) and report 7% overall failure on a highly imbalanced test set with four mouse behavior classes. More recently Le and Murari (2019) applied a combination of three-dimensional Convolutional Neural Network (3D-CNN) and LSTM on the dataset from Jhuang et al. (2010) and report comparable results as Jhuang et al. with only end-to-end input. Finally, Jiang et al. (2019) propose a hybrid deep learning architecture with a combination of unsupervised and supervised layers followed by an HMM. They outperform Jhuang et al. on their mouse behavior dataset with overall agreement of 81.5% vs 78.3% and an average recall of 79% vs 76%. They also show that, after retraining, their method is applicable to another task with different classes in a slightly different setup.

As stated above, in order to be useful in behavioral research, an automated system must be able to recognize behaviors independent of treatment and laboratory setup. Good recognition performance on a dataset recorded in one setup is an important step. However, retraining supervised systems on a new setup requires a lot of data and brings back the manual annotation task for a significant number of video segments. Three approaches are possible to get closer to the goal. One direction is to standardize laboratory setups (Arroyo-Araujo et al., 2019). Second is to aim for quick adaptation of a classifier towards a new setup with minimal annotation effort, i.e. learn from limited data. The third is to strive for generic recognition with robust and adaptive methods.

Deep learning might provide the key to achieving these goals. Development time is reduced since laborious handcrafting of features is not needed anymore. Without dedicated features we might also be less restricted in the application, and less preprocessing avoids noise being introduced by it. Furthermore, trained networks can be partially reused so we don't need to train from scratch in a different but comparable scenario. The downside of neural networks is the amount and variety of data needed to train them.

The goal of this study is to compare our earlier handcrafted rodent behavior classification system to end-to-end classification by an advanced deep learning network for action recognition, to evaluate the flexibility of the recognition on unseen setups, and to learn how it can be improved. In Section 3.2 we explain network, metrics, datasets, sampling and augmentation. In Section 3.3 we present classification results of within-setup and across-setup recognition, which we discuss in Section 3.4. We conclude in Section 3.5.

3.2 Materials and methods

We address two questions. First, what is the performance of an advanced action recognition deep learning network on a rodent behavior dataset? We experiment on a dataset of short rat behavior clips and apply two different input schemes, namely a) end-to-end input without preprocessing, versus b) region-based input from tracking information, i.e. regions of interest around the animal + optical flow to capture the motion. We train with and without data augmentation. The second question we address is aimed to investigate applicability in real-life scenarios: what is the performance of this network on continuous videos and across setups? We evaluate videos of rat behavior using the best performing input scheme and compare within-setup and across-setup classification performance.

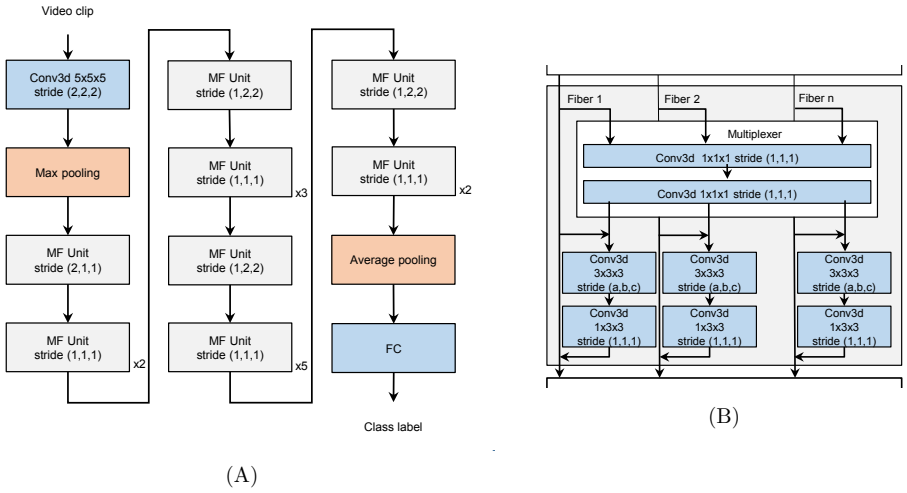


Figure 3.1: Architecture of Multi-Fiber network. (a) The overall architecture. (b) The internal structure of each Multi-Fiber Unit. The diagrams were reconstructed from Chen et al. (2018).

3.2.1 Network

As network architecture we used the Multi-Fiber network (MF-Net) described in Chen et al. (2018). Figure 3.1 shows the diagram of the network. The choice of this network was based on its good performance on the currently most important benchmark datasets for activity recognition, e.g. UCF-101 (Soomro et al., 2012), HMDB-51 (Kuehne et al., 2011) and Kinetics (Carreira and Zisserman, 2017), and its efficiency compared to other well performing networks, namely it needs 9× less calculations than I3D (Carreira and Zisserman, 2017) and 30× less calculations than R(2+1)D (Tran et al., 2018) to get to the same results. The crux of the MF-Net architecture is that it uses an ensemble of lightweight networks (fibers) to replace a complex neural network, reducing the computational cost while improving recognition performance. Multiplexer modules are used to facilitate information flow between fibers. We used the available code [1] with modified sampling, augmentation and performance metric. Furthermore we adjusted the number of layers and kernels to deal with our specific input layout considering resolution and channels.

[1] `https://github.com/cypw/PyTorch-MFNet`

The network consists of one three-dimensional convolutional (conv3d) layer followed by four multi-fiber convolution (MFconv) blocks. Each MFconv block consists of multiple MF units, and each MF unit consists of four (five for the first block unit) conv3d layers. All conv3d layers are followed by batch normalization and a rectified linear unit (ReLU). The final layers of the network are an average pooling layer and a fully connected layer. Since the middle layer of every MF unit uses a (3,3,3) kernel there is temporal convolution in 17 conv3d layers and additionally in the last average pooling layer during aggregation of the final eight frames.

We did not initialize the network with a pretrained model since our input channel layout differs from the colored 3-channel human activity datasets that available pretrained networks are trained on.

3.2.2 Metrics

In large scale activity recognition nowadays the most popular performance metric is top-1 or top-k accuracy, where the top-1 accuracy denotes the overall agreement across frames, i.e. the proportion of the input where the model's prediction was right and top-k denotes the proportion of the input where the target class was in the model's top k most likely predictions. However, these measures are misleading for imbalanced datasets with equally important classes as is the case with sampling from continuous videos. Suppose the dominant class covers 80% of the samples and the network classifies all samples as belonging to this class. Then the overall agreement of this obviously bad classifier would be 80%. More informative measures in this situation are precision and recall per class, precision of a class being the proportion of found frames that is correct and recall being the proportion of correct frames found. In this study, we use average recall as an aggregated measure, calculated by taking the average of the behavior class recalls. Although this does not take into account the precision, all ill-labeled samples contribute negatively to the average recall since we take all classes into account. In comparison to the averaged F1-score, false positives of rare classes have more negative impact than those of frequent classes, which is preferable. For comparison with related work, we also report overall agreement per video for the cross-setup evaluation, calculated as the proportion of correct frames per video.

Note that behavior itself is not discrete and behavior changes take time. Therefore, it is good to keep in mind that 100% accuracy is not feasible because of inherent ambiguity at behavior bout boundaries.

In all experiments and evaluations, frames not belonging to one of the nine classes are left out of the evaluation. The goal of this study is to compare handcrafted feature classification to end-to-end classification, within and across setups. Although the question how to detect 'other' behavior is important for applicability it was left outside the scope of this study.

3.2.3 Within-setup experiments on clips

Dataset

For the within-setup experiments we used the high quality dataset described by Van Dam et al. (2013). It consists of 25.3 video hours of six Sprague-Dawley rats individually housed in a PhenoTyper 4500 cage (`http://www.noldus.com/phenotyper`, Noldus Information Technology, Wageningen, Netherlands) at 720×576 pixel resolution, 25 frames per second and with infrared lighting, hence gray-scale. Subsets of these recordings (∼2.7 hour in 14 subvideos) were annotated by a trained observer using The Observer XT 10.0 annotation software (`http://www.noldus.com/observer`), and manually checked and aligned afterwards to ensure frame accurate and consistent labeling. Checking and alignment took one hour per five-minute video for 14 classes (including subatomic classes). In this study we

focused on the nine most frequent state behavior classes 'drink', 'eat', 'groom', 'jump', 'rest', 'rear unsupported', 'rear wall', 'sniff' and 'walk'. The tenth behavior, 'twitch', is a point behavior without annotated duration and was left out of the comparison.

The data is presented to the network in two different ways. End-to-end input consists of the gray-scale videos resized to square 224×224 resolution. The square crop was made from the center of the arena, after the resize. The end-to-end model trained on the clips subset is referred to as E2e-c. As an alternative to the end-to-end input we removed the tracking task and provided the model with a 88×88 moving region-of-interest around the animal. Frame motion information was added with the optical flow (x and y) in the second and third channel. The tracking, flow calculation and cropping was done with EthoVision XT 14.0 (`http://www.noldus.com/ethovision`), which was modified for this purpose. The second type of input format is referred to as Roi+flow.

Network details

Because the input resolution differs for the end-to-end (224×224×1×32) and Roi+flow inputs (88×88×3×32) the network layouts are slightly different. The main difference is that the max pooling layer was omitted in the Roi+flow layout, because the Roi+flow resolution needs less spatial reduction. For both networks the total size is 7.7M parameters. See Table 3.1 and Table 3.2 for more details.

Table 3.1: Details of the MF-Net architecture in end-to-end experiments.

layers	#MF units	#channels	#frames	width	height	kernel	stride	padding
input		1	32	224	224			
conv3d		16	16	112	112	5,5,5	2,2,2	3,3,3
maxpool		16	16	56	56	1,3,3	1,2,2	1,1,1
MFconv1	3	96	8	56	56			
MFconv2	4	192	8	56	56			
MFconv3	6	384	8	14	14			
MFconv4	3	768	8	7	7			
avg pool			1	1	1	8,7,7	1,1,1	
FC			400					

Table 3.2: Details of the MF-Net architecture in Roi+flow experiments.

layers	#MF units	#channels	#frames	w	h	kernel	stride	padding
input		3	32	88	88			
conv3d		32	32	44	44	3,3,3	1,2,2	1,1,1
MFconv1	3	96	16	44	44			
MFconv2	4	192	8	22	22			
MFconv3	6	384	8	11	11			
MFconv4	3	768	8	6	6			
avg pool			1	1	1	8,6,6	1,1,1	
FC			400					

Sampling

The within-setup experiments are applied on a set of behavior clips sampled from the within-setup dataset. We perform 4-fold cross-validation over different random train/test splits (80/20 per class). In each fold there are 2314 training clips and 398 test clips. Each clip contains 32 consecutive frames. The clip label is the behavior in the middle of the clip i.e. the annotation of the 17^{th} frame. Clips were randomly picked with the constraint that

there is no behavior transition in the middle of the clip, between frame 14 and 19. In the training set, the clips have a maximum overlap of 29 frames and there were never more than four clips selected per behavior bout, with a maximum of 400 clips per behavior. For testing, the maximum overlap is 25 frames and there are never more than two clips selected per behavior bout, with a maximum of 50 clips per behavior. For the exact number of test clips per behavior see Table 3.4. Clips from the same behavior bout are always together in the same split, so either in the training or in the test set.

Data augmentation

To prevent overfitting, the data is augmented by applying a random combination of the following known filters: resized crop, horizontal and vertical flip, inverse, rotation (90/180/270 degrees), luminance variation (brightness, contrast and gamma), additional Gaussian noise, additional salt & pepper noise, image blur. Additionally we applied two new filters: video cutout and dynamic illumination change. Video cutout is the 3D version of 2D-cutout introduced by DeVries and Taylor (2017). It implies adding occlusions to the clip by replacing randomly located cuboids with the mean clip value. Dynamic illumination change is created by adding a random 3D Gaussian to the clip, which has the effect of gradually turning on or dimming a spotlight on a random time and location in the clip. For Roi+flow the flow was calculated after random rotation and inverse of the video frame, and modified implementations where made for the flipping filters to flip the optical flow vectors. Resized crop was omitted and luminance variation was only applied to the gray-scale channels. Dynamic illumination change was also omitted for Roi+flow since it would affect the optical flow calculation.

After augmentation, the clips were normalized to have a mean of 0 and standard deviation of 1. Normalization was done per channel to avoid mixing image and optical flow information.

3.2.4 Continuous and cross-setup experiments

Dataset

In these experiments we used the cross-setup validation dataset from (Van Dam et al., 2013) illustrated by Figure 3.2 and Table 3.3. It contains one video from the within-setup dataset and four videos recorded with different resolution, animal strain, illumination, background and feeder and spout positions. Frame rate and camera viewpoint were not changed, and all recordings were made with constant lighting and good contrast between animal and background. Table 3.5 presents the performance of the conventional RBR system on these videos.

Figure 3.2: Stills of the five videos with different setups used for cross evaluation.

Sampling

In order to estimate robustness in real-life scenarios we next evaluate the performance across experimental setups and on continuous videos. Unlike the within-setup experiments

Table 3.3: Description of the five videos with different setups used for cross evaluation.

Dataset	Duration	Resolution	Setup
Video 1	13.7 min	720×576	One video from the within-setup dataset
Video 2	10.5 min	360×320	Half resolution, different sawdust
Video 3	5.0 min	720×576	Different strain (Wistar)
Video 4	5.3 min	768×576	Visible light
Video 5	2.5 min	720×576	Visible light, no sawdust

that were conducted on a balanced subset of clips and ignored clips around behavior bout transitions, the model is now deployed on sliding-window clips (32 frames wide, step size 1 frame). These clips contain more ambiguous data than the subset of clips used in our within-experiments and the set is not balanced anymore.

In the cross-setup experiments we consider only the end-to-end input scheme. We applied the E2e-c model that was trained on the entire balanced clips dataset (2712 clips) to the sliding-window clips of the test videos. Alternatively, we retrained the model on all sliding-window clips from the within-setup dataset (32 frames wide, step size 4). This model is referred to as E2e-s. The sliding-window clips set is much bigger (52560 clips) and not balanced anymore. To account for the imbalance during training we used weighted random sampling. This way during every epoch the less frequent behaviors are presented to the network more often. Since random augmentation is applied, the network sees different versions of the clips. For evaluation of within Video 1, the models were retrained without the clips of Video 1.

3.3 Results

All experiments were conducted on a Dell Precision T5810 with 32GB memory and a NVIDIA Titan X (Pascal) GPU with 12 GB, running Ubuntu 18.04, with Python 3.7 using the PyTorch framework (0.4.1). MF-Net with end-to-end input ran the forward call at ∼230 frames/sec. The annotation speed of RBR (including video IO, tracking and feature extraction by EthoVision) is ∼124 frames/sec on a CPU (Dell Precision T3620 with 8GB, Intel Xeon E3-1240 v6 @3.7 GHz), which is almost five times faster than real-time.

3.3.1 Within-setup evaluation on clips

Figure 3.3 presents violin plots showing the classification results on the balanced clips dataset with and without data augmentation for the two different input schemes, for all folds. The end-to-end input scheme with data augmentation yields the best result of 75% average recall. The results per behavior are listed in Table 3.4, for both average fold and best fold. The effect of increasing data augmentation is shown in Figure 3.4. The confusion matrices in Figure 3.5 show that accuracy is high for almost all classes, the biggest confusion coming from 'jump'/'walk' and from 'sniff'/'eat'. From the loss curves in Figure 3.6 we observe that the network overtrains without data augmentation and that the network can learn longer for the more difficult end-to-end task. Experiments with smaller sized networks (less layers) did not improve Roi+flow test performance.

3.3.2 Across-setup evaluation on continuous videos

First, we evaluated the E2e performance on continuous videos. We compare two models: E2e-c trained on the cleaner and balanced clip dataset, and E2e-s trained on the much bigger but noisier dataset of sliding-window clips as it contains also clips with behavior bout transitions in the middle. We test both models on the sliding-window clips of Video

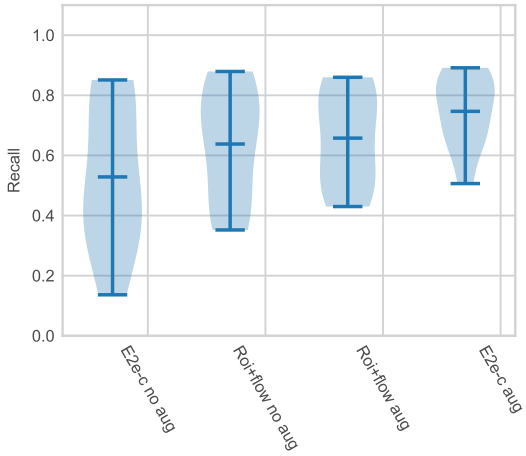


Figure 3.3: Average recall per class of the end-to-end and Roi+flow models after 4-fold within-setup evaluation on the balanced clips dataset, with and without augmentation.

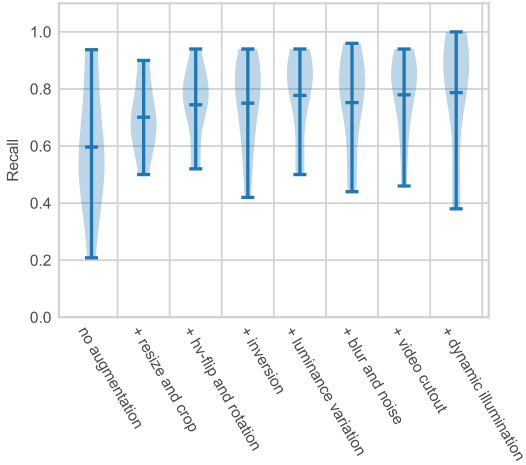


Figure 3.4: The effect of increasing data augmentation on the average class recall using the end-to-end model.

1. In Table 3.4 we see that having good performance on an unseen set of clips is not enough to guarantee performance on continuous videos. Instead, the E2e-s model performs better on all behaviors except 'rest' (that only has 16 frames). Figure 3.7 shows the event log for within-setup Video 1.

Next, we evaluated the E2e-s model on our set of videos in varying setups. Table 3.5 presents the overall agreement per video, Table 3.6 shows the recall per behavior. Compared to handcrafted-feature classification, E2e-s outperforms RBR on the within-setup evaluation, but not on the cross-validation task. This holds for all four cross-setup videos and for all classes except 'groom' and 'jump'. Performance was decreased especially for Video 3, which is mostly due to the large amount of false negatives for 'sniff' and false

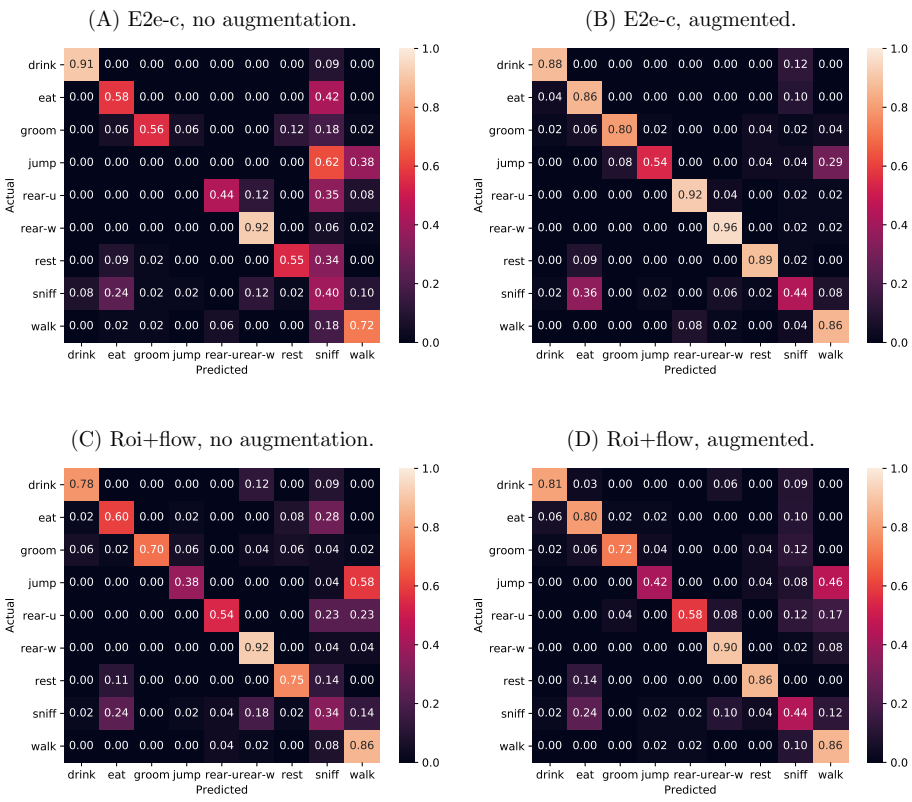
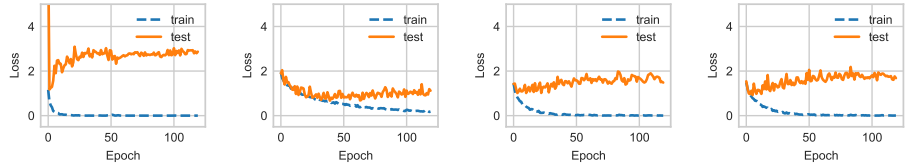


Figure 3.5: Confusion matrices with results of the end-to-end and the Roi+flow model on the within-setup test clips, with and without augmentation.



(A) E2e-c, no augmentation. (B) E2e-c, augmented. (C) Roi+flow, no augmentation. (D) Roi+flow, augmented.

Figure 3.6: Train and test losses while training the end-to-end and Roi+flow model with and without augmentation on the clipped, within-setup dataset. Horizontal axis is training iteration, vertical axis is loss. Once the training loss is zero, the network cannot learn anymore from the training set. Only b) E2e-c-augmented does not overtrain and learns best.

positives on 'rest' (see the event log in Figure 3.8). From the results per class the misclassifications of 'drink', 'eat' and 'rear wall' behaviors stand out compared to RBR.

Table 3.4: Recall per behavior of the end-to-end model, tested on clips test sets and on continuous video, all within-setup. For the model applied to the clips test set both the 4-fold and best fold results are presented.

	Clips test set			Video 1		
	clips	4-fold E2e-c	best fold E2e-c	frames	E2e-c	E2e-s
Drink	(32)	0.82	0.88	(689)	0.00	**0.39**
Eat	(50)	0.82	0.86	(3186)	0.44	**0.53**
Groom	(50)	0.80	0.80	(331)	0.68	**0.86**
Jump	(24)	0.67	0.54	(0)	-	-
Rear unsupported	(48)	0.67	0.92	(559)	0.56	**0.76**
Rear wall	(50)	0.86	0.96	(2908)	0.86	**0.94**
Rest	(44)	0.67	0.89	(16)	**0.79**	0.00
Sniff	(50)	0.51	0.44	(2453)	0.41	**0.80**
Walk	(50)	0.89	0.86	(10290)	0.68	**0.89**
Average recall	(398)	0.75	0.79	(18872)	0.55	**0.65**
Overall agreement	(398)	0.76	0.80	(18872)	0.48	**0.77**

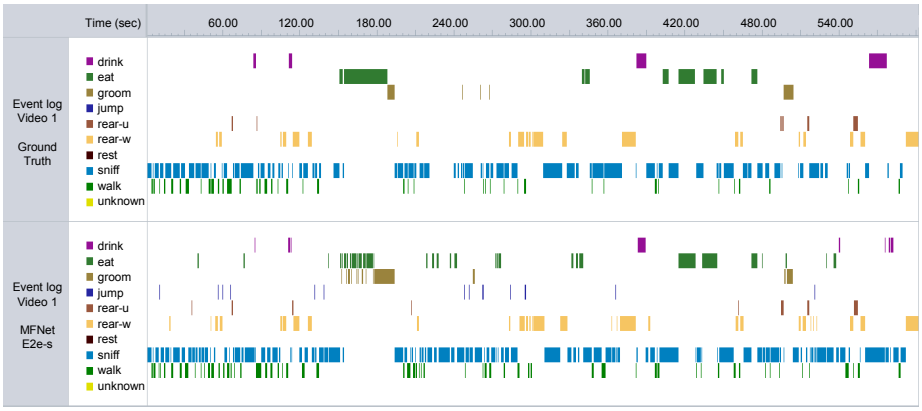


Figure 3.7: Event logs for manually labeled ground truth (above) and automatic end-to-end annotation (below) on Video 1 (within-setup evaluation).

Table 3.5: Overall agreement of the end-to-end model with augmentation after evaluation on unseen continuous videos, within and across setups, compared to the results of hand-crafted RBR classification. Note that datasets are not balanced.

	RBR	E2e-s
Within		
Video 1	0.71	**0.77**
Across		
Video 2	**0.65**	0.50
Video 3	**0.80**	0.27
Video 4	**0.67**	0.51
Video 5	**0.65**	0.59

3.4 Discussion

First, we interpret the within-setup mistakes of E2e-c. Looking at the confusion matrix in Figure 3.5 we see that most confusion comes from 'jump'/'walk', 'sniff'/'eat' and to

Table 3.6: Recall per behavior of the handcrafted RBR classification and the end-to-end model with augmentation, evaluated on unseen continuous videos within and across setups.

	Within Video 1 frames	RBR	E2e-s	Across Video 2-5 frames	RBR	E2e-s
Drink	(689)	**0.42**	0.39	(1181)	**0.93**	0.38
Eat	(3186)	**0.79**	0.53	(6691)	**0.76**	0.10
Groom	(331)	0.74	**0.86**	(1900)	0.58	**0.73**
Jump	(0)	-	-	(73)	0.48	**0.99**
Rear unsupported	(559)	0.70	**0.76**	(871)	**0.74**	0.63
Rear wall	(2908)	0.69	**0.94**	(2248)	**0.63**	0.05
Rest	(16)	0.00	0.00	(1003)	0.16	**0.27**
Sniff	(2453)	0.70	**0.80**	(15352)	**0.67**	0.65
Walk	(10290)	0.86	**0.89**	(2127)	**0.60**	0.55
Average recall	(18872)	0.62	**0.65**	(31446)	**0.62**	0.45

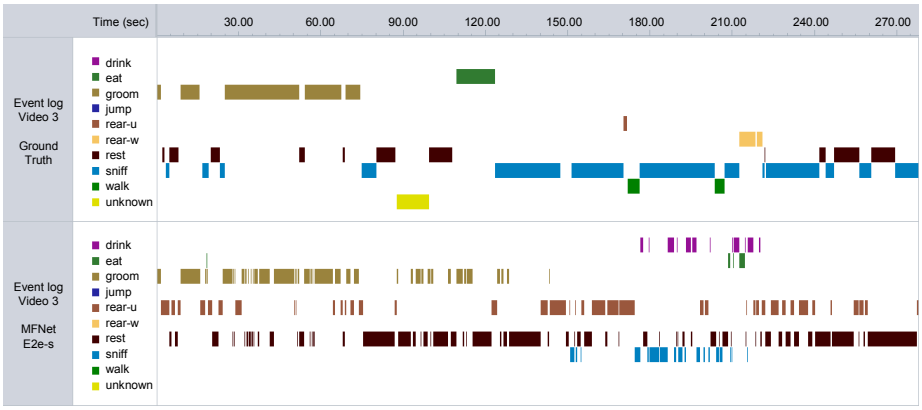


Figure 3.8: Event logs for manually labeled ground truth (above) and automatic end-to-end annotation (below) on Video 3 (across-setup evaluation).

a lesser extent from 'eat'/'sniff' and 'sniff'/'walk'. These are understandable mistakes since these are gradually overlapping behaviors that can be performed more or less at the same time and hence easily subject to interpretation differences. In these cases, automated annotation is probably even more consistent than human annotation that is more sensitive to context.

Second, we interpret the mistakes made by E2e-s on continuous video, within setup. The event logs for Video 1 (Figure 3.7) show very good correlation between human and E2e-s annotation. It stands out that there are more behavior switches in the E2e-s annotation. This suggests that although E2e-s classification contains many temporal filters, it could still benefit from post processing, either by explicitly averaging the soft-label output over time or by adding a recurrent layer after the FC layer. Many of the related work methods use recurrence in their classification. This is helpful to smoothen the output and helps the algorithm to suppress detection of unlikely behavior sequences. However, this makes these systems less applicable to annotate behavior of drug-treated animals. In these cases, the behavioral transition probabilities might be altered, and instead of being part of the model, these changed transitions are a result of the experiment.

Thirdly, let us examine the poor results of E2e-s on the cross-setup tasks. Looking at event logs for Video 3 in Figure 3.8 it is notable that many 'sniff' frames are mistakenly

detected as 'rest' or 'rear unsupported'. Also, many 'eat' behaviors are mistaken for 'rest'. Looking at the video reveals that the animal in Video 3 is very cautious and pauses a lot during its movements. Although this behavior was labeled by the human as 'sniff' it is a type of sniff that was not in the training data where the animals are more at ease. RBR does not suffer from these mistakes, possibly because the decision making is more integrated over time and 'rest' will be only detected when the animal does not move for a longer period.

The E2e model especially fails to recognize environment-dependent behaviors in the cross-setup task. Behaviors 'drink', 'eat' and 'rear wall' score below 40% while the recognition of these three behaviors is over 80% in the within-setup evaluation. Handcrafted RBR has an advantage since the location of the drinking spout, feeder zone and walls are provided by the user. However, the network should be able to 'see' the feeder and the edges of the floor in all setup videos, and deduce that the drinking spout is always on the side of the arena. Also during augmentation all clips are rotated and resized hence the model should be robust to changes in the exact position of walls, feeder and spout.

Future work will be to experiment with adding a recurrent layer to the network, adding augmentation that varies backgrounds and adding explicit visible environment cues to the input video, such as a floor map. Alternatively, we can optimize networks for specific setups. Still the most challenging problem will be to address the unseen behavior variation that caused wrong automated annotations of Video 3. A first step can be to detect abnormal behavior sequences and let the user tell the network how to interpret the sequence. This requires learning from fewer data examples.

3.5 Conclusion

In this study, we addressed the problem of automated rodent behavior recognition and compared the accuracy performance of an advanced deep learning approach (MF-Net) to conventional handcrafted classification (RBR). For within-setup performance on a clipped dataset we showed that MF-Net with end-to-end input outperforms both handcrafted RBR and MF-Net with Roi+flow input, provided sufficient data augmentation. For cross-setup performance on continuous video, we showed that MF-Net with end-to-end input could not outperform RBR. We argue that the end-to-end model has difficulty recognizing environment cues and is not robust to differences in behavior sequences observed, which is a problem for animals behaving different than normal, for instance due to treatment. We conclude that deep learning networks give us good performance on fixed setups with known behavior, but that more research is necessary to reach adaptive and flexible human-like performance that is independent of the setup and behavior performance.

Chapter 4

Disentangling rodent behaviors to improve automated behavior recognition

This chapter has been published as E.A. van Dam, L.P.J.J. Noldus and M.A.J. van Gerven (2023). Disentangling rodent behaviors to improve automated behavior recognition. *Frontiers in Neuroscience*, volume 17:1198209. https://doi.org/10.3389/fnins.2023.1198209

4.1 Introduction

Automated observation and analysis of behavior is important to facilitate progress in many fields of science, especially in behavioral studies on neurological and psychiatric disorders or drug discovery, where rodents (mice and rats) are still the most commonly used model animals in preclinical research. With increasingly large image datasets and computational hardware capacity, we have seen a tremendous progress in pose estimation for many different animal species (Mathis et al., 2018; Lauer et al., 2022). In behavior recognition, the progress has not been that evident. Available systems recognize behaviors with a reliability of around 70-75% (Van Dam et al., 2020), or are trained and tested on footage from the same recording session, for a limited set of specific behaviors. However, in order to be useful in behavioral research, automated systems that can recognize behavioral activities must be able to recognize them independent of animal genetic background, drug treatment or laboratory setup. To match human-level performance in annotating behavior, we need to improve accuracy, robustness and genericity of automated systems. Accuracy means good precision and recall per behavior, robustness means consistent accuracy across experimental setups, and genericity means that the same method is applied to all behaviors. Three approaches are at hand. First is to standardize laboratory setups, i.e. the test environment in which the animals are observed (Grieco et al., 2021). This limits the variance but leaves the animal- and treatment-related variation. Second is to aim for quick adaptation of the recognition system towards a new setup with minimal annotation effort, i.e. fine-tuning or retraining. This requires new ground truth data and brings back the manual annotation task for a significant number of video segments. Moreover and more importantly, researchers who need to compare animal behavior between treatment groups need one measurement system instead of separately trained observation models. The third approach is to explicitly strive for generic recognition with robust methods, which is in principle possible as humans can do so.

In this paper, we investigate where we stand with respect to the goal of generic recognition, and what is needed when we raise the bar for future automated behavior recognition,

that is, (1) to recognize ethologically relevant behaviors, (2) recognize behaviors robustly across experimental setups, and (3) recognize new behaviors with limited data and fine-tuning effort.

Robustness across experimental setups requires that the system can handle variation in three aspects, namely appearance, behavior execution, and behavioral sequence. For the behavior class performed, the appearance of the animal is irrelevant, i.e. whether the animal is white or black, long or small, thick or slim, long or short-haired. The same applies to the appearance of the environment, such as the walls, floor, feeder, drink spout or enrichment objects. While their presence may enable or limit certain behaviors, their color and texture should not affect recognition. Behavior recognition should also be invariant to how behaviors are executed, i.e. differences in event duration, pace and subbehavioral pattern. In addition to the usual event variations, behavior execution varies by physical or emotional state, and by individual animal, depending on strain, gender, age, history and medication. Furthermore, execution varies due to different layout of the environment, such as the size of the cage or the height of the drink spout. The third aspect for which automated recognition systems need to be robust is the sequence of the behaviors performed, as the treatment of animals affects the frequencies of specific behaviors. Behavior recognition systems that use history or recurrence such as hidden Markov models (HMMs), recurrent neural networks (RNNs) or 3D convolutional neural networks (3D-CNNs) train on temporal context and hence on behavioral context, and will have difficulty to recognize the behavior events when applied in a different context.

There are multiple ways to increase robustness of behavior classification systems. The best way is to train on larger and more diverse datasets. This is costly and it is not always possible to cover all experimental diversity beforehand. Alternatively, we can factor out variance up front by normalizing the input. By using tracked body points we can focus on the poses and dynamics, and solve most of the appearance bias (Graving et al., 2019). Furthermore, there are training 'tricks' to improve a model's internal robustness, such as dropout and variational encoding of latent variables (Goodfellow et al., 2016). We can also add variance by augmentation of the input, altering the input in ways that leave the behavior intact. Most data augmentation methods used are augmentations of appearance, such as size, scale or pixel intensity (Krizhevsky et al., 2017).

Behavior execution differences and behavior sequence differences are differences in dynamics. We believe that focus on variation in dynamics can improve behavior modeling substantially. If we can normalize and augment the behavior execution and behavior sequence, classification will be more robust. Stretching and folding the time-series to alter the speed and intensity of the movement is one way, but we can also vary the sequence of the behavior events as well as the subbehavioral pattern. To vary the sequence of the behaviors we need to detect the events and how they follow each other. To vary the subbehavioral patterns per behavior, we need to understand the type and characteristics of the possible subbehaviors and how they are combined. We give examples of composite rodent behaviors in Section 4.3.1. We further expect that breaking down composite behaviors into subbehaviors will also highlight subtle yet essential constituents and thereby will increase the detection accuracy of behaviors that are otherwise too difficult to separate from behaviors that are alike and more frequent.

The main idea of this paper is that acknowledging the hierarchical and composite structure of behavior can bring automated behavior recognition to the next level and a step closer to human-like annotation performance. If we could leave out the appearance variation and measurement errors and if we had endless amount of training data, to what extent are state-of-the-art networks able to model behavior dynamics?

We illustrate and explain three types of composite behaviors in Section 4.3.1. These compositions are present in the rat dataset described in Section 4.3.2. Next, we describe an

artificial dataset that contains these compositions in an abstracted form and can be used to study the limits of automation models without input noise or lack of data (Section 4.3.2). Finally, we present behavior recognition results on both the rat and the artificial data in Section 4.4 and draw conclusions in Section 4.5

4.2 Related work

4.2.1 Supervised behavior recognition

An effective recipe for training a recognition system is to record a dataset, annotate it and use supervised learning to train a classifier to recognize the behaviors. The classifier iteratively finds the best optimization path to get as close to the ground truth as it can, using all the cues it can find. Hence, the quality and robustness of the resulting classifier is always dependent on the representational value of the data trained on. In order to be robust to using cues that are only coincidentally or concurrently related to the behavior classes, data augmentation is applied to the input: typically, image transformations like flipping, scaling and rotation. Deep learning models are very good at finding informative cues, but this also means they are sensitive to using cues that only apply within the training dataset. In almost all studies that describe behavior recognition systems, the test set is recorded in the same setup, with animals from the same strain and treatment as those in the training set. Previous work shows that although deep models can reach better performance than conventional methods, the performance is less transferable to different experiment settings (Van Dam et al., 2020). Supervised methods that have been applied are conventional methods such as bag-of-words (Dollár et al., 2005), Bayesian classification (Van Dam et al., 2013) or tree-based classifiers used in MARS (Segalin et al., 2021) and SimBA (Goodwin et al., 2024). Perez and Toler-Franklin (2023) provide an overview of CNN-based approaches, such as 2D, Two Stream networks and 3D-CNNs, often combined with recurrent head to model the temporal dependencies. In recent years, major advances in deep learning classification are made using Transformer architectures that are designed to pick up the most relevant context without constraints on how far away that context is. Sun et al. (2023) report that multiple Transformer-derived networks applied to trajectory data improve the classification of social rodent behavior.

4.2.2 Data-driven approaches

During the past 10 years, data-driven approaches have been presented that learn the constituent modules of behavior from the data itself. MoSeq from the Datta Lab introduced behavior syllables or motifs as behavior components (Datta, 2019) and uses autoregression filters for classification (Wiltschko et al., 2020; Costacurta et al., 2022). TREBA (Sun et al., 2021), and VAME (Luxem et al., 2022) use self-supervised learning with recurrence on sliding temporal windows to create latent representations that are used as input in supervised downstream tasks. These methods are capable of accurately predicting phenotypes and behaviors from videos withheld from the training dataset. Self-supervision is very useful when the amount of training data is small compared to the network complexity, and in discovering new significant behavior motifs or patterns. For image classification tasks, Newell (2022) showed that, with self-supervised pretraining, the top accuracy plateau is reached faster and with less data. Nonetheless, as in supervised training, accuracy increase stops around 75-80% (Sun et al., 2023). What most models have in common is the assumption that behavior consists of a sequence of behavior states and that the subject switches from one state to the next. The underlying assumption is that states can be inferred either statistically by learning the underlying state-switching process from the observed samples (HMMs), or by sliding window classification.

4.2.3 Hierarchical approaches

Other research recognizes that behavior can be looked upon at different levels and different scales, and that detection can be improved when models are trained at multiple hierarchical levels simultaneously. Gupta and Gomez-Marin (2019) show that worm behavior is organized hierarchically and derive a context-free grammar to model this. Casarrubea et al. (2018) apply T-pattern analysis to study the deep structure of behavior in different experimental contexts. Kim et al. (2019) introduce a variational approach to learn hierarchical representation of time-series on navigation tasks. Finally, Luxem et al. (2022) detect behavioral motifs in an unsupervised manner and let human experts assign labels to communities of these motifs obtained from motif traversal analysis. Recent work that most closely resembles our representation of hierarchical structure in rodent behavior is that of Weinreb et al. (2024). This work builds upon MoSeq and extends the auto-regressive model (AR-HMM) by Switching linear dynamical systems (SLDS). They distinguish three hierarchical levels, namely behavior syllables, pose dynamics and keypoint coordinates. Their main purpose however is to denoise the input that contains erroneous keypoint jitter introduced by failing tracking.

4.3 Materials and methods

4.3.1 Behavior

In the following we provide a description of the constituents that make up behavior, give different examples of composite behavior and describe other factors that make automated behavior recognition non-trivial. We derived these constituents and compositions after visual inspection of the failures of rat behavior classification that we report in Section 4.4.1, as well as from the results on various other datasets reported over the years by users of the keypoint-based behavior recognition module RBR from Van Dam et al. (2013) that is part of the EthoVision XT video tracking system.

Behavior constituents

Figure 4.1, panel (A) shows a representation of behavior seen as switching states. The samples are the observed poses, extended with derived features at the consecutive timestamps. It is implied that all observations are related to a single behavior state, and that state switches are abrupt. This is the way behavioral data is labeled that is used as ground truth for training recognition systems and that the system gets to see either one-by-one or in a sliding window with fixed length. However, when we as humans annotate behavior, we evaluate the samples differently and distinguish more than switching states. Subjective experience suggests that we predict future motion, and only take a closer look when we see deviation of what we expect, regardless of the subject or the behaviors at hand. This interpretation of the human brain as a prediction machine is supported by research in cognitive neuroscience (Keller and Mrsic-Flogel, 2018; Heilbron et al., 2022). We seem to build a belief about the goal pose and intentional state of the subject, based on the observed poses over time. When what we see no longer resembles our belief, we take a closer look, in order to revise our belief. That is, we go through the following stages of observation and inference: The subject displays behavior A - the subject no longer displays behavior A - the subject is in transition to another behavior - the subject is in transition towards either behavior B, C or D - the subject is doing behavior B. We evaluate the consecutive poses until we see that the subject arrives at a new key pose and infer the behavior from that. Sometimes we have to wait for a sequence of key poses before a decision can be made. In a transition between behaviors, the intermediate poses are merely pose changes to get

from one key pose to another. They are necessary because subjects can only move around in space and time in a continuous manner. Yet, they do not define the behaviors, but are defined entirely by the previous and the next key pose. The constituents that form behavior are therefore not only states that determine the samples. Apart from states, we can also distinguish transitions, key poses with no duration and sequential combinations of these.

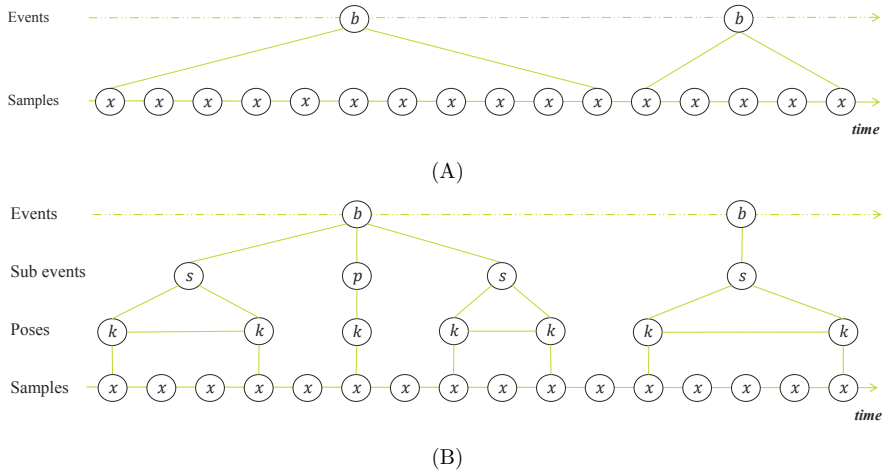


Figure 4.1: Different representations of behavior **(A)** Behavior seen as switching states with successive behaviors b at event level and observations x at sample level (one for every time-step). **(B)** Behavior seen as hierarchically structured constituents and transitions, with two intermediate levels, namely a sub-event level and pose level. The sub-event level contains state events s and point events p. The pose level contains key poses k. Key poses are body postures that are held by the animal during one sub-event. The intermediate samples between successive sub events are transition samples between different key poses.

With this is mind we propose a new representation of behavior, shown in Figure 4.1, panel (B). It shows a representation of behavioral components and how they can be combined, which resembles what we see when we annotate behavior. While we are labeling the events, we perceive behavior as a sequence of state events, point events and transitions. State events are defined by key poses with a certain variation and duration, whereas point events are defined by key poses with zero or minimal duration. Note that we use point event slightly different than is common among ethologists, who use point event to indicate that the duration is not relevant for analysis. Here we want to emphasize that the behavior is characterized by a momentary key pose. Transitions are the transitional movements between different key poses (also known as movement epenthesis). Behaviors are combinations of these constituents. If we can build automated models that can detect these constituents, we can improve the recognition.

Finally a note on what should not be modelled, namely the dependencies between the behaviors at the top-level. We need to make sure that the recognition of a behavior is not dependent on the occurrence of specific previous behavior events. The behavior transition matrix is an output of an experimental test and this information should not be used during training to optimize the recognition, for if the sequence changes because of treatment effects, the detection will be hampered. In practice this means that we must have a sufficient amount of diverse training data, either by collection or augmentation.

Three examples of composite rodent behavior

We illustrate the composition of behavior into a sequence of transitions, state events and point events with three examples of rodent behavior in Figure 4.2.

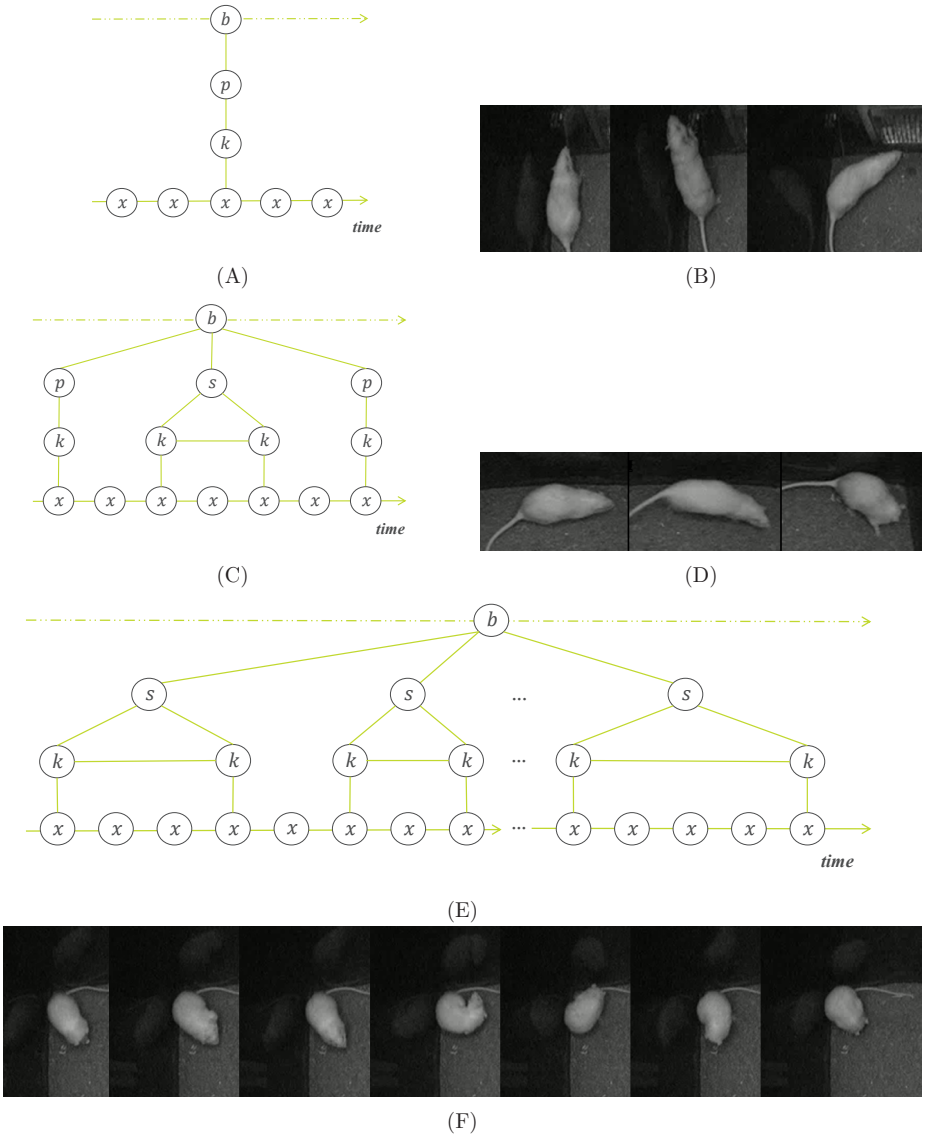


Figure 4.2: Three examples of structured behavior, each with a schematic representation and a selection of frames from a single event. Panels **(A)** and **(B)** show a point event (rearing), **(C)** and **(D)** show an ordered composition (jumping) that consists of three sub events, namely a point event (take-off), a state event (stretched pose) and another point event (landing). Panels **(E)** and **(F)** illustrate an unordered sequence of state events (grooming).

Figure 4.2, panels (A) and (B), show a typical rearing event, where the behavior

consists of a transition from the non-rearing key pose before the rearing, towards the short peeking pose in an upright body position, followed by the second transition towards the next non-rearing key pose. What happens often is the detection of a false-positive rearing event when the actual rearing pose does not occur but the animal is shortly retreating to change direction. However, the system detects the transitional movements, i.e. a forward movement or turn followed by a backward movement or turn. The point event in the middle that is defining the event as rearing is missing but the transition samples match most of the samples of the rearing events in the training set. Note that rearing events can also be state events, when the upright position is held for some time.

The next example in Figure 4.2, panels (C) and (D), is a jumping event that starts with the point event of the take-off, followed by a fast-forward movement and the landing as a second point event. These two point events, the take-off and the landing, are essential for the jumping behavior and distinguishes it from mere walking behavior. Yet the majority of the samples in the jumping event are in the fast-forward movement, so the behavior distributions of walking and jumping overlap considerably when all samples are weighted equally.

The third example in Figure 4.2 on panels (E) and (F) is a grooming event that is composed of multiple state events that are not strictly ordered although the common sequence is grooming snout, head, fur, genitals. The grooming-snout substate samples overlap considerably with substates of behaviors eating, sniffing and resting, but can nevertheless be identified as grooming because they are surrounded or followed by more outstanding grooming subevents. In this case it is the context of the surrounding substates that determine the decision when made by a human annotator.

Distribution characteristics of rodent behaviors

Apart from the challenging demands posed by the composite behaviors, additional characteristics of rodent behavior make automated recognition difficult. These are: high overlap between poses of different behavior classes, high variance between events of the same class, mixture of pose distributions for a subset of classes, unbalance of event frequency distributions hence little training data for rare but important classes, and finally, high variance in event duration, which makes it difficult to set global temporal scales for processing. We give examples of pose overlaps and present behavior event distributions in Figure 4.3.

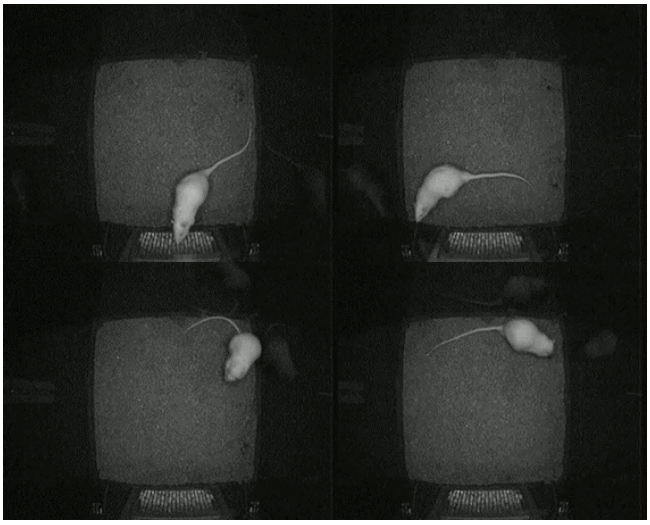
4.3.2 Data

To analyze the extent to which automated behavior recognition models are able to model rodent behavior in general and composite rodent behavior in particular, we experiment with two types of data: real rat behavior data and artificial abstracted behavior inspired by real rat behavior.

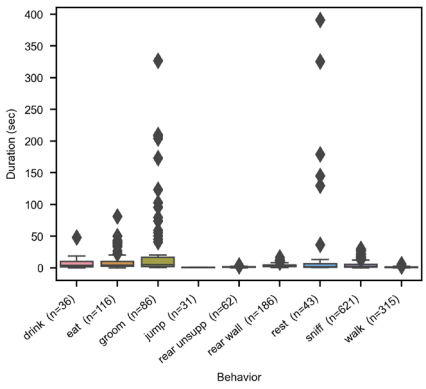
Rat behavior dataset

The rat behavior dataset was reused from previous work and is described in (Van Dam et al., 2013). It consists of 25.3 video hours of six Sprague-Dawley rats, each in a PhenoTyper 4500 cage [1] at 720×576 pixel resolution, 25 frames per second and with infrared lighting, hence gray-scale. Subsets of these recordings (∼2.7 hour in 14 subvideos) were annotated by a trained observer using The Observer XT 10.0 annotation software [2], and manually checked and aligned afterwards to ensure frame accurate and consistent labeling. In this study we focused on the nine most frequent behavior classes 'drink', 'eat', 'groom',

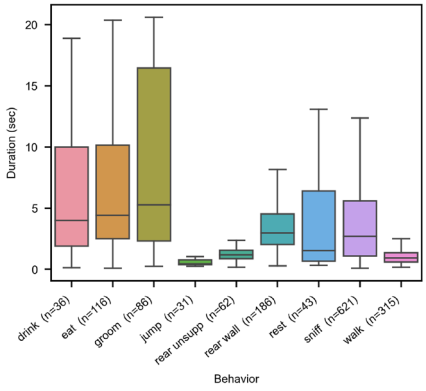
[1] http://www.noldus.com/phenotyper
[2] http://www.noldus.com/observer



(A)



(B)



(C)

Figure 4.3: Illustrations of the distribution characteristics of rat behavior that make automated recognition challenging. Panel **(A)** shows four examples of pose confusions. Clockwise, starting upper left, are the confusions (manual label/automated label) sniff/eat, sniff/drink, eat/groom. groom/eat. Panel **(B)** and panel **(C)** show the distributions of the behavior event durations on the entire rat dataset, with and without outliers, to illustrate the big differences across and within behavior classes.

'jump', 'rear unsupported', 'rear wall', 'rest', 'sniff' and 'walk'. To focus on the dynamics, we applied the same input preprocessing as was used in VAME by (Luxem et al., 2022), namely we tracked six body-points using DeepLabCut (Mathis et al., 2018), and aligned and normalized these.

Artificial time-series

In order to experiment with different types of behavior dynamics without suffering from incomplete or incorrect features or insufficient amount of data, we generated artificial time-

series of randomly sampled behavioral events, with predefined behavior components and substate dependencies inspired by the rodent behavior components. The sample features, or poses, are drawn from predefined distributions, with configurable variation across and inside events. Components are either point events or states with durations sampled from a distribution, and are concatenated by transition periods of two to eight samples. Per behavior event, we added fluctuations with configurable smoothness, amount and periodicity. As a last step, we added observation noise. The result is a configurable amount of time-series data that we can train the recognition models on, with configurable difficulty, depending on the number of behaviors, number of features, overlap in feature distributions, complexity of behavior structure, and amount of overlap between the constituents of different behaviors. With this procedure we generated two different datasets to experiment with: 1) artificial state behaviors and 2) artificial composite behaviors. The code to construct these datasets is publicly available [3]. In the code repository, we included the definitions for the artificial datasets used here, as well as an example with four features.

Artificial state behaviors The first artificial dataset contains only state behaviors, modeled after the varying distribution characteristics mentioned in Section 4.3.1. The feature distributions and an example time-series of state behaviors are plotted in Figure 4.4, panel (A) and (B). The following behaviors are included. First, behaviors with well separated pose (b01, b02), which should be easy to recognize and are added as sanity check. Second, behaviors with poses that are alike (b03, b04; confusion group 1). In real rat data there are behavior pairs have overlapping poses, for instance 'drink' and 'sniff'. Third, behaviors whose pose distributions are a mixture of poses (b05), for instance as 'groom' and 'eat'. Fourth, behaviors with uncommon event duration distributions, either long or short (b06, b07; confusion group 2). Examples in rat behavior are 'sleep' and 'twitch'. Fifth, periodic behaviors (b08, b09 overlapping with behavior b10; confusion group 3). Finally, we inserted pose transition samples between behavior events (b00).

Artificial composite behaviors The second artificial dataset contains two behaviors with well-separated pose (b01 and b02) and additionally the following composite behaviors. First, point behaviors, i.e. defined by key poses of zero or minimal duration, with transitions dependent on the key poses of surrounding events. Point behaviors are hard to detect because they may overlap with samples from state behaviors or with transition samples. An example in the rat behavior data are rearing events, where the surrounding frames are similar to sniffing poses. In the artificial dataset, the point behavior is b11, overlapping with b12. Second, ambiguous subbehaviors in unordered sequences: behaviors defined as an unordered sequence of subbehaviors that have their own distributions, and where some of these subbehavior distributions overlap with other behaviors (behavior1 = n x {A or B or X}, behavior2 = {P or Q or X}). In the rat behavior data this corresponds with the overlap between grooming-snout and eating events (b13, overlapping with b14: confusion group 4). Third, ambiguous subbehaviors in ordered sequences: behavior defined by a specific, fixed sequence of subbehaviors, where some of the subbehaviors also occur in the sequence of other behaviors (composite behavior A-X-B versus behavior P-X-Q). An example in the rat behavior data is jumping behavior that consists of take-off - stretched pose - landing. The stretched pose is also part of a walking sequence (b15, overlapping with b16: confusion group 5). Feature distributions and an example time-series of composite behavior are plotted in Figure 4.4, panel (C) and (D).

[3] `https://github.com/ElsbethvanDam/artificial_behavior_data`

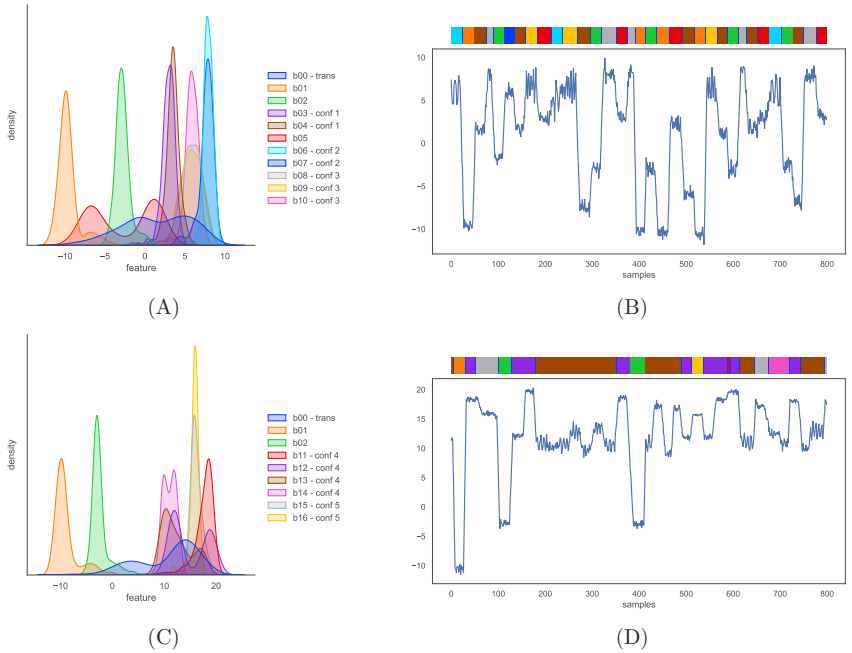


Figure 4.4: Artificial behavior feature distributions and 1-dimensional time-series example with the behavior event bar on top. **(A)** and **(B)** for artificial state behaviors, and **(C)** and **(D)** for a set including artificial composite behaviors, i.e. behaviors that can consist of a sequence of subbehaviors per event.

4.3.3 Classification models

We will now describe the two models we used to evaluate the current performance of automated rodent behavior recognition. The first model is a recurrent variational auto-encoder (RNN-VAE) that we applied to all the data. The second is a Transformer model for time-series that we applied to the artificial data.

A good approach is to train a recurrent variational auto-encoder (RNN-VAE) to get a behavior embedding for every short time window of length T (T=0.5 sec) and use this embedding as input for a small linear network that aims to find n behavioral motifs (n=30) from the data. The mapping to the motifs is then used to classify the final behaviors per sample in a supervised manner, using a linear classifier. We followed the network implementation of VAME (Luxem et al., 2022) with an encoder consisting of two bidirectional GRU layers (hidden dimension h=64) and a decoder of one GRU layer (h=32) plus a linear layer to map the input resolution of $T \times F$, where F denotes the feature dimensionality. The embedding size varies with the size of the features: For the rat data (14 features) we used embedding dimension d=30, and for the artificial data with only one feature we use d=6. The output of the encoder is the concatenation of the hidden RNN states. Before passing the output of the encoder to the decoder, a joint distribution is learned and sampled from during training, to ensure better robustness of the embedding. The n motifs are learned by including in the loss the clustering-based spectral regularization term (see Luxem et al. (2022) (supporting information), Ma et al. (2019)). In our experiments, we did not train the motifs and behavior classification separately, but instead added a supervised classification head. This means we allowed the network

to optimize embedding and motifs for both the decoding and the behavior classification task, by optimizing three losses: a self-supervised reconstruction loss, a clustering loss and a supervised classification loss. During training, the importance of the classification loss is gradually increased. Note that for supervised classification we could have omitted the motif cluster mapping. We kept it in because we want to investigate the model's ability to learn motifs for the difficult (rare, subtle, composite) behaviors.

As an alternative model, we replaced the RNN-VAE network with a Transformer network derived from LIMU-Bert (Xu et al., 2021), a Bert model for time-series, and applied it to the artificial datasets. The model has four encoder layers, each with four attention heads and a feed-forward layer with hidden size h=80. A linear decoder projects the encoded input back to the original input size $T \times F$. As in LIMU-Bert, to train the encoder, the input sequence of 20 samples is masked with a contiguous span of samples instead of individual samples to avoid trivial solutions (mask ratio=0.45), and only the spans are represented and predicted. After reconstruction, the entire original input sequence is encoded without masking and a slice of five samples is classified with a bidirectional GRU classification head (h=30). As before, the reconstruction loss and a supervised classification loss are trained simultaneously.

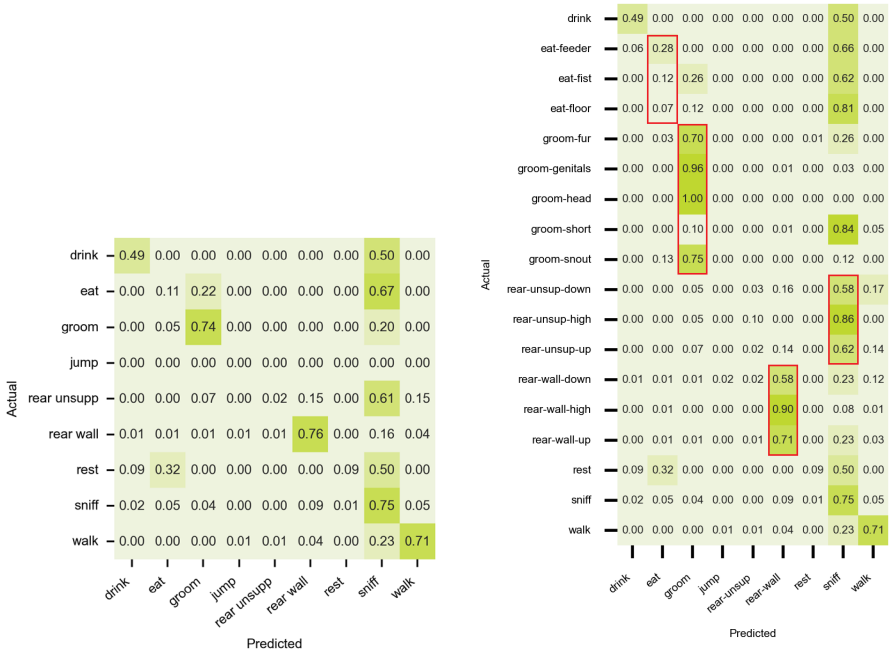
For all experiments, we performed a hyperparameter search with Optuna (Akiba et al., 2019) to ensure the best possible results. The tuned parameters are learning rate, number of hidden dimensions and the size of the embedding. For the Transformer network we also tuned the mask ratio and the window size of the slice that is sent to classification.

4.4 Results

4.4.1 Modeling rat behavior as switching states

The confusion matrices in Figure 4.5 present the result of the RNN-VAE model on the rat behavior dataset, calculated from the sequences of the aligned six body-point coordinates per frame. Figure 4.5, panel (A) shows the confusions at event level, panel (B) shows confusions at sub-event level. It is clear that the recognition works well for some of the state behaviors and is less successful for other behaviors. Half of the drinking frames are detected as sniffing, and most of the eating samples are seen as sniffing or grooming. Eating is executed in three different ways: at the feeder, in which case it overlaps with sniffing, or away from the feeder in a sitting pose or off the floor, in a way that it also overlaps with the grooming-snout pose. Nearly all behaviors are confused with sniffing, which is due to overlap in both pose and movement intensity of the very wide distribution of sniffing poses. For a human annotator, it is the context of more explicit behavior that determines the decision. The confusion in resting behavior is because the sequences in the test data are very short compared to the few very long resting periods in the training data, and in different poses. In the detailed results of the rearings, the middle part of the rearing ('high') is confused differently than the upward and downward movements, which can be due to our observation that rearing events contain a relatively large amount of transitional samples.

Overall we identify four types of confusion. First, the features can be sub-optimal, i.e. incomplete, insufficient or just noisy and incorrect. Next, point behaviors may not be detected. Furthermore, confusion is likely when the relevant context is not picked up. Finally, not all confusions are errors. Transitional samples between states get labeled but are in fact ambiguous ground truth.



(A) (B)

Figure 4.5: Confusion matrices with classification results on the rat dataset using the supervised RNN-VAE, per behavior on Panel **(A)** and per sub-behavior on Panel **(B)**. The overlaps differ per subbehavior.

4.4.2 Modeling artificial behaviors

The first set of artificial data contains only state behaviors, without structure. Both models can recognize the behaviors equally well, as shown in confusion matrices in Figure 4.6, panels (A) and (B). The confusion that we see is grouped according to the behavior definitions of the dataset. As expected, classes b01, b02 and b05 are well-separated. The models have difficulty with 2 of the 3 confusion groups: confusion group 1 with poses that are alike (classes b03 and b04), and confusion group 2 with uncommon event distributions (classes b06 for long events and b07 for short events). Confusion group 3 with class-specific periodicity (classes b08, b09 and b10) is handled correctly. We conclude that both models can learn state behaviors that have no specific dynamical structure, except for behaviors with class-specific event durations (confusion group 2).

The results on the artificial dataset with composite behavior are presented in Figure 4.6, panels (C) and (D). This artificial dataset was inspired by the analysis of confusions made in classifying the rat dataset, and contains state behaviors, point behaviors and transitions, as well as state sequences with ambiguous subbehaviors. The behavior definitions overlap in the same way that the rat behaviors do, see the definitions in 4.3.2. In the confusion matrix, we see the confusions that we expect, even with a big enough dataset. Again, classes b01 and b02 are well separated. In both models, point behavior b11 (equivalent to 'rear') is confused with state behavior b12 ('sniff'), but also with b15 ('jump'), which is most likely due to the overlap with the transitional poses that comprise most of the b11 context samples. In confusion group 4, behavior b13 ('groom') was

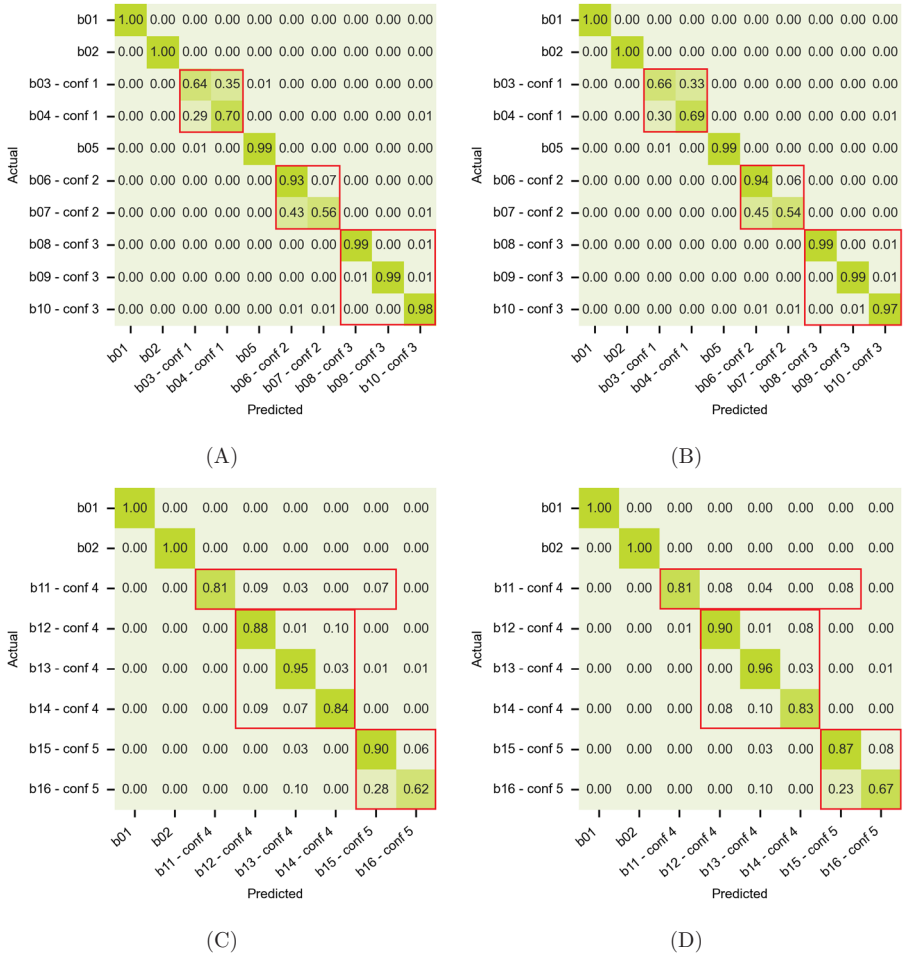


Figure 4.6: Confusion matrices with classification results on the artificial datasets for **(A)** state behaviors and model RNN-VAE, **(B)** state behaviors and model Transformer, **(C)** composite behaviors and model RNN-VAE, **(D)** composite behaviors and model Transformer. Confusion groups are outlined in red.

defined as an unordered sequence of substates corresponding to different grooming poses, one of which is overlapping with state behaviors b12 ('sniff') and b14 ('eat'). See supplementary Figure4.7 for the sub-event level confusion matrix. The models did not use the surrounding context of substates to infer behavior b13. Nor could the models solve confusion in confusion group 5, namely find the conditional context of behavior b15 ('jump') that separates it from b16 behavior ('walk').

4.5 Discussion

Currently known automated systems for the recognition of animal behavior from video suffer from lack of robustness with respect to animal treatment and environment setup. In order to be useful in behavioral research, systems must recognize the behaviors of control

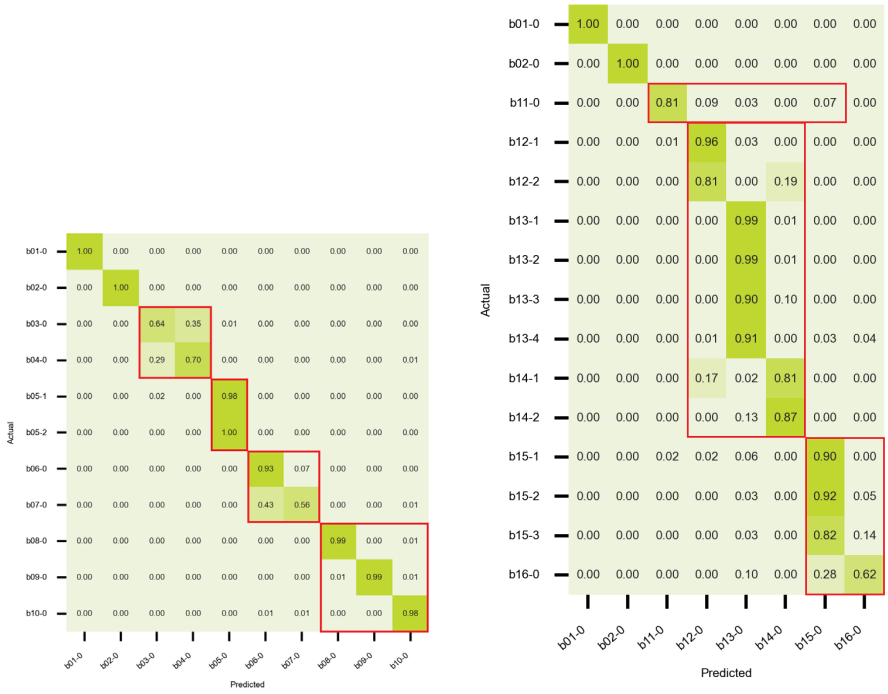
and treated animals regardless of compound effects on appearance, behavior execution and behavior sequence. Careful analysis of miss-detections in rat behavior recognition lead us to distinguish behaviors into four types of behavior constituents, namely state events, point events and pose transitions, and sequences thereof. To study the performance of recognition models on the different types of dynamics, we created artificial time-series and present results for the most advanced recognition systems.

The classification results on the artificial dataset show that, even with sufficient amount of data with absent noise and ideal annotation quality, and with supervised classification and hyperparameter tuning, the networks are not capable of classifying the composite rodent behaviors, or behavior-specific event durations. Therefore, the solution towards more robust rodent behavior classification is not only to train on more data or to avoid input noise. We also need to improve on how to break down the composition. If models can learn to compress time-series into segments that correspond to behavior constituents, they can analyse segment properties and sequences regardless of the temporal scale of the segments. The usual way of segmenting data into equidistant samples and segments of equal duration is therefore not the best way to segment behavior, and adding the attention mechanism of the Transformer is not enough to overcome this.

Although rodents can switch goal poses instantaneously, they can only change their actual pose in a continuous manner. This makes certain samples more informative than others. For instance, a rat can be walking towards the feeder and suddenly decide to drink first. It takes intermediate positions to change a walking pose into drinking pose. Such pose changes while changing from one behavior to another are often not informative for the behaviors themselves. This is generally true for recordings of intentional agents. How to infer the agent's goal poses is unsolved so far, but if we can discard the uninformative transitional samples we can reduce confusion. One possible way to achieve this is to predict future poses, and take as start and stop pose of the transition the frames that are difficult to predict. Although this seems a good approach, it is very difficult to steer the predictions from the data itself given the amount of variation and valid, possible projections.

With the data compressed into behavior segments and transitions, we would be able to normalize and augment the behavior execution and the behavior sequence which would make classifiers more robust. Breaking down composite behaviors will furthermore increase the detection accuracy of difficult behaviors, for it allows to highlight short yet necessary constituents.

We showed that adding more training data is not sufficient to make progress for several ethologically relevant behaviors, and we argue that understanding the composite nature of animal behavior is necessary to move the field forward. We believe that discarding uninformative pose transitions will reduce confusions and that detection and evaluation of segment sequences will pick up more relevant context. Future research can focus on this direction.



(A) (B)

Figure 4.7: Sub-behavior level confusion matrices with classification results of model RNN-VAE on the artificial datasets for **(A)** state behaviors and **(B)** composite behaviors. For the composite behaviors, only some of the sub-behaviors overlap with other behaviors. Confusion groups are outlined in red.

Chapter 5

Practical solutions for recognition of new behaviors

The following chapter describes two different practical solutions to classify new behaviors with help of the ABR features. The first is a supervised method for the classification of scratching behavior from videos with a framerate of 100 frames per second. Once trained, the classifier can be deployed on new videos recorded in a comparable setup. The second solution describes an active learning approach, for the efficient and accurate annotation of more difficult and subtle behaviors. This AI-assisted annotation method learns to take over from the human annotator, but leaves decisions about ambiguous cases to the user.

5.1 Robust scratching behavior detection in mice from generic features and a lightweight neural network in 100 fps videos

This subchapter is based on this publication: E.A. van Dam, M.H. Roosken and L.P.J.J. Noldus (2022). Robust scratching behavior detection in mice from generic features and a lightweight neural network in 100 fps videos. *Volume 2 of the Proceedings of the Joint Meeting of the 12th International Conference on Measuring Behavior and the 6th Seminar on Behavioral Methods Held Online, 18–20 May 2022*, p. 301-305. `http://doi.org/10.6084/m9.figshare.20066849`

5.1.1 Introduction

Quantification of rodent scratching behavior is important because scratching is used in animal models for skin diseases and stress. Scratching behavior consists of a rapid, repetitive movement of the paw against the body, mostly the hind paw against the neck. Since scratching instances are usually rare and short, it is difficult to annotate them manually. Also, in the field of automated behavior recognition, short and infrequent events pose a challenge. A scratching bout usually lasts less than a second, hence only a small fraction of input is positive. This hinders the training of classification algorithms, which need sufficient examples for their models to converge. Furthermore, the behavior is unevenly distributed: it occurs frequently in one recording and does not occur at all in others. Another important consideration for classification of this behavior is the high speed of paw moments. This may cause part of the movement to be lost on recordings with a low sample rate. The average frequency of paw movement during scratching in our records is roughly 20 Hz. Given this information, Nyquist's Sampling Theorem implies that the sample rate of the recordings should be at least 40 Hz to prevent distortions. This means

that in case of inferring scratching behavior from video, the frame rate should be higher than the CCTV standard video frame rate of 25 or 30 frames per second (fps). Although some authors report results using the standard frame rate, e.g. Akita et al. (2019), several others reported methods based on higher frame rates. For instance, Nie et al. (2012) calculate the frame-to-frame difference on videos with 240 fps and use a short-pulse detection filter to detect the scratches. More recently, Kobayashi et al. (2021) applied a convolutional recurrent neural network directly to the video input of 60 fps, on a sliding window of 20 frames. Both methods were trained and tested on videos from a single dataset and are not designed to work out of the box on footage recorded in other circumstances, e.g with different cage, background, camera height and light. That severely limits the practical applicability of these methods. In this work we aim for automatic recognition of scratching behavior of mice in footage from multiple datasets using the features described in Van Dam et al. (2013), which are derived from tracked body-point locations and optical flow. From this, a normalized 2D motion profile map of the animal movement over time is created. The final set of 169 features is the result of sliding window statistics and 1D log-Gabor responses in the temporal direction. Classification of these features enables robust behavior recognition of rodent behavior across datasets.

5.1.2 Methods

Data

The dataset consists of nine trials recorded at two different labs, see Figure 5.1. The trials are roughly 30 minutes long and were recorded with a Basler USB-3 IR camera (acA1920-155um) with 100 frames per second, resolution 1920×1080 pixels. The first setup combines three cages. Both setups use home cages, black mice, and sawdust covering. The videos were recorded for other behavioral research studies that are covered by approval of authorized ethical committees.

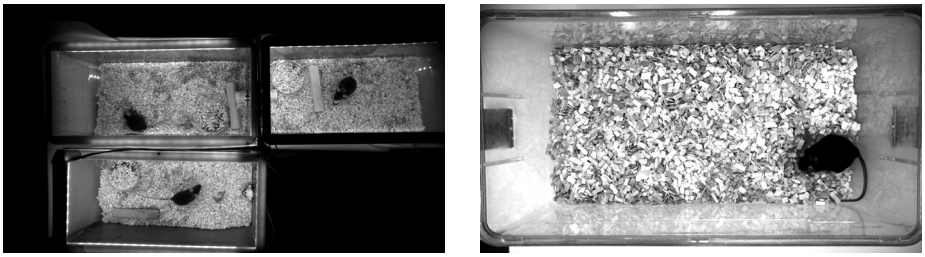


Figure 5.1: Stills from the two recording setups.

Figure 5.2: Four equally spaced frames from a single scratching paw movement in video 4. This set of consecutive paw movements had an average frequency of 17.9 Hz.

Features

The input data for the model consists of the 169 features designed by Van Dam et al. (2013) with a small adjustment to log-Gabor filters because of the higher frame-rate. Notice that the log-Gabor filters are calculated in the temporal dimension and respond to the periodic movement of the scratching, each filter at a specific frequency. As such, the feature array for a single time step is sufficient to classify that frame. This means that there is no direct need to employ RNNs, so that we can use a lighter, faster architecture for the model. The features were calculated using EthoVision XT 17, a video-tracking system developed by Noldus Information Technology (`http://www.noldus.com/ethovision`).

Network

The model is a simple Multi-Layer Perceptron (MLP) consisting of linear layers and activation functions. The network topology consists of one hidden layer with 75 units. This is left fixed for performance reasons.

Loss function

For the training loss we used the focal loss, which is defined as follows:

$$\ell(\hat{p}|y_0) = -(1 - p(y = y_0))^\gamma \log(\hat{p}_{y_0})$$

Here, $\hat{p}$ is an array of predictions, y_0 is the ground truth, $p(y = y_0)$ is the proportion of class y_0 in the train data and $\gamma \geq 0$ is a hyperparameter. The focal loss is a rescaled version of the familiar categorical cross-entropy (i.e. the log loss), with class weights $w_i = (1 - p(y = y_0)^\gamma$. It is biased towards classes that are rare in the training dataset. A larger value of γ increases the relative class weight of rare classes, which in our case is scratching behavior. Both L2-regularization and dropout are applied to all hidden layers of the network in order to reduce the chance of overfitting.

Hyperparameter optimization

For optimization of hyperparameters, we used the Optuna framework (Akiba et al., 2019). This framework attempts to find the best set of hyperparameters by efficiently sampling from the hyperparameter space. We ran Optuna to optimize the following: learning rate, batch size, number of epochs, L2 factor, dropout probability, activation (sigmoid, relu, tanh) and γ. We ran Optuna with 500 trials and pruning enabled. As our trial objective, we took the minimum of the precision and the recall. This forces Optuna to always improve the lowest of the two values. We computed this objective by nine-fold cross-validation over the nine videos, averaging over folds.

Cross-validation

We used cross-validation to evaluate the best model as found by Optuna. The nine videos were used as folds for cross-validation. On each fold, we recorded the f1, precision, recall, FPR (false positive rate) and FNR (false negative rate). The model was trained five times per video to compute the mean and standard deviation. Finally, the combined predictions on on all nine folds were used to compute total performance metrics of the classifier.

5.1.3 Results

Hyperparameter optimization

Optuna ran for 500 trials, of which 145 finished and the rest was pruned (due to unpromising results). The best value was found after 454 trials, having a score of 0.783. The optimization history is shown in Figure 5.3.

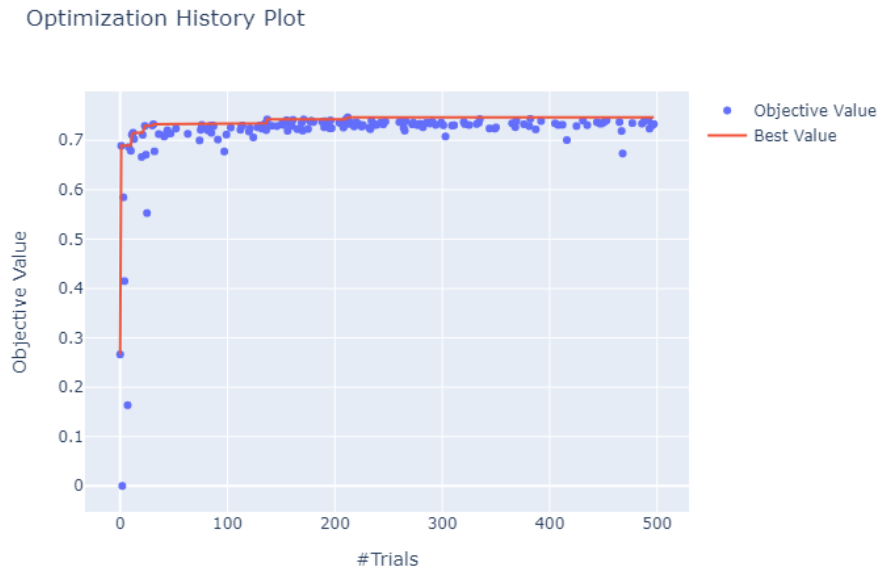


Figure 5.3: Optuna optimization history.

The optimal values for the hyperparameters are as follows: learning rate = 7.2×10^{-4}, batch size = 512, number of epochs = 6, activation function = sigmoid, L2 factor = 1.1×10^{-5}, dropout probability = 0.16, $\gamma = 0.097$

Cross-validation

The results of cross-validation, with means and standard deviation, are shown in Table 5.1. Note that videos 4 and 8 contain no scratching behavior. The metrics are all computed based on the number of frames correctly classified (rather than the number of events). The predictions for videos 3 and 5, respectively, are shown in Figure 5.4 and Figure 5.5.

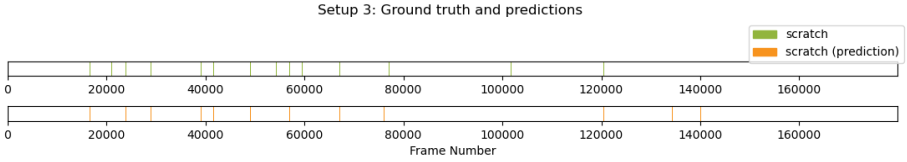


Figure 5.4: Predictions compared to ground truth for video 3.

Table 5.1: Table with cross-validation results f1, precision, recall, false positive rate (FPR) and false negative rate (FNR) over the nine videos, as well a total score over all videos. The number before the ± sign indicates the mean and the one after the standard deviation.

Video	f1	precision	recall	FPR	FNR
Video 1	0.725 ± 0.003	0.796 ± 0.022	0.666 ± 0.012	0.0032 ± 0.0005	0.3342 ± 0.0116
Video 2	0.763 ± 0.007	0.775 ± 0.019	0.752 ± 0.012	0.0010 ± 0.0001	0.2484 ± 0.0120
Video 3	0.813 ± 0.005	0.832 ± 0.014	0.795 ± 0.014	0.0024 ± 0.0003	0.2053 ± 0.0142
Video 4	-	-	1.000 ± 0.000	0.0012 ± 0.0004	-
Video 5	0.909 ± 0.001	0.901 ± 0.005	0.917 ± 0.005	0.0138 ± 0.0009	0.0828 ± 0.0051
Video 6	0.902 ± 0.002	0.909 ± 0.009	0.894 ± 0.011	0.0099 ± 0.0012	0.1058 ± 0.0113
Video 7	0.666 ± 0.015	0.720 ± 0.035	0.621 ± 0.015	0.0014 ± 0.0002	0.3795 ± 0.0154
Video 8	-	-	1.000 ± 0.000	0.0002 ± 0.0001	-
Video 9	0.772 ± 0.004	0.785 ± 0.015	0.761 ± 0.013	0.0027 ± 0.0003	0.2393 ± 0.0133
Total	0.872 ± 0.001	0.878 ± 0.005	0.867 ± 0.005	0.0036 ± 0.0002	0.1335 ± 0.0049

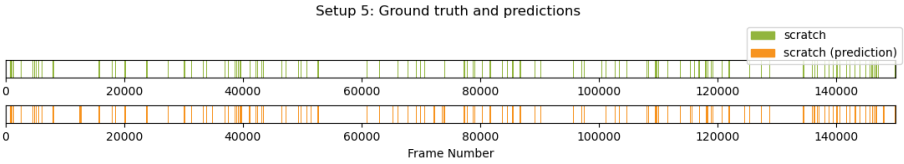


Figure 5.5: Predictions compared to ground truth for video 5.

5.1.4 Conclusion

In this study we present a robust rodent scratching behavior classifier that works out-of-the-box for top-view videos recorded at 100 fps. We used generic features derived from earlier work and optimized a small classification network for the detection. The detector will be used in behavioral studies at multiple labs so we will have more validation data in the future.

5.2 Fast annotation of rodent behaviors with AI assistance: Human observer and SmartAnnotator collaborate through active learning

This subchapter is based on the following publication: E.A. van Dam, T.J. Daniels, L. Ottink, M.A.J. van Gerven and L.P.J.J. Noldus (2024). Fast Annotation of Rodent Behaviors with AI Assistance: Human Observer and SmartAnnotator Collaborate through Active Learning. *Proceedings of Measuring Behavior 2024, Aberdeen, 15-17 May 2024* p. 230-235. https://doi.org/10.6084/m9.figshare.25897855

5.2.1 Introduction

Automated annotation of rodent behavior from video recordings is an essential tool for behavioral research. It speeds up and improves rodent behavior analysis with more readily available and consistent behavior annotations. Since early 2013, commercially available solutions as well as open source projects exist for a specific set of behaviors (Van Dam et al., 2013, 2022; Isik and Unal, 2023; Segalin et al., 2021). Although great strides have been made and the most common rodent behaviors can be detected in video streams under specific recording conditions, many ambiguous or rare behaviors that are also relevant in behavioral research (such as epileptic seizures, stereotypic variants of behaviors, whisking), or for which a different definition is used, are still scored by hand. Developing generic, robust automatic solutions is costly since it requires a large set of precise and consistently labeled video footage that contains the same variation as the variation in the footage that the solution will be applied to (Van Dam et al., 2020). This refers not only to the appearance of the animals and environment (fur color, cage, lighting), but also to the way the behaviors are executed. The speed of walking, the length of a grooming session, the height of a rearing and angle to the camera, the behaviors before and after a scratching event, etc., vary with for instance age, time of day, mood, motor skills and drug treatment. An automated classifier that was optimized with use of machine learning can only reliably recognize what has been seen in training data, so all variations in the deployment data set must occur in the training set as well. If this is not the case, the classifier will suffer from selection bias. To collect and label a sufficient amount of such training data is difficult to achieve, especially for rare and subtle behaviors, or for behaviors that are very specific to the research question at hand. To meet the demand for faster annotation of behaviors for which there is no generic solution available, and to tackle the challenges in behavior classification, we developed a novel AI-assisted annotation tool, SmartAnnotator (Figure 5.6). This tool helps the researcher to annotate behaviors, by training a classifier through active learning. Here, active learning refers to the ability of an AI system to interactively query a human user to label new data points to maximally improve learning performance. Instead of playing a video from start to end and scoring behavior by hand, the researcher is presented short video clips to annotate. Simultaneously, a classifier is trained in the background on these annotations, and infers behaviors on unlabeled video clips, until the entire video is annotated. Importantly, SmartAnnotator selects video clips that were given an uncertain label by the model and asks the user to label these clips. In other words, SmartAnnotator selects events whose annotation will maximally improve behavior classification. This exploits both human expertise and AI to increase annotation accuracy. This interactive approach is much more efficient than labeling all data from start to end and avoids observer drift, as well as unavoidable decision delay in manual scoring. Hence, it reduces annotation time while increasing the quality of the labels. Furthermore, unlike previously described tools for interactive behavior annotation (Kabra et al., 2013; Lorbach et al., 2019), SmartAnnotator is cloud-based, so annotation can be done in any

web browser and resources are scalable.

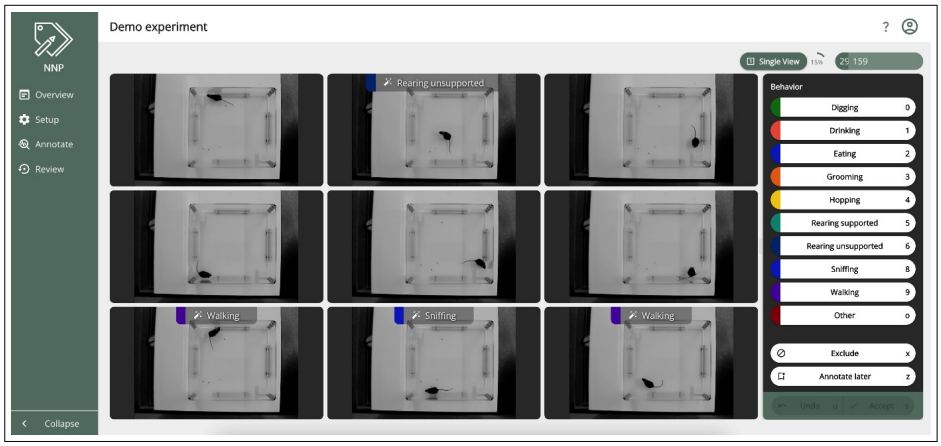


Figure 5.6: A screen capture of the SmartAnnotator tool. Multiple selected clips are presented to the user for labeling.

AI-assisted annotation, nevertheless, also has its challenges. It relies on the relevance and quality of the features that it can extract from the videos, the segmentation of the data into clips, and on how reliably the tool can detect similarities in the data. Furthermore, the learning and processing needs to be fast, since it is an interactive process that must be user-friendly. Yet the advantage over generically trained classifiers that are deployed out-of-the-box is that with AI-assisted annotation, a dedicated classifier can be trained on a specific set of features with a specific interpretation of the behavior by the user. Also, the algorithm does not need to account for unseen variance since it works on the entire set of videos from one behavioral experiment at once. Recent work (Luxem et al., 2022; Weinreb et al., 2024) on automated behavior detection furthermore highlights the use of behavioral clusters, i.e. short pieces of behavior of a certain type that are learned by clustering a latent representation (embedding) of the data, learned by a self-supervised auto-encoder. It has also been shown (Van Dam et al., 2023) that there is not always a straightforward one-to-one relation between such clusters and specific behaviors of interest. Therefore, we have not applied classification on the embedding at this moment but used the behavioral clusters only for segmentation of our data. We have used SmartAnnotator to annotate specific target behaviors in a set of videos: 'stretched attend' in a mouse dataset, and 'unsupported rearing' in a rat dataset. 'Stretched attend' was chosen since it is relatively easy to recognize manually from videos and does not have a large variation in event duration. It is more difficult to recognize for automated recognition tools, since elongated body postures also occur during 'walking', 'jumping' and 'rearing wall' events. 'Unsupported rearing' was selected as a challenging example that is difficult to detect automatically from 2D-videos since event durations are typically short and contain a relatively large number of transitional postures that are equal to postures of other behaviors (Van Dam et al., 2023). For both behaviors, this can easily lead to a large number of false positive detections. We show that our method improves annotation accuracy and reduces the amount of data that needs manual labeling compared to a supervised classification of these behaviors. This demonstrates the benefit of active learning and our SmartAnnotator tool in finding particular or rare behaviors in a larger dataset, that would otherwise easily have been misclassified.

5.2.2 Methods

Annotation through active learning using SmartAnnotator

Using SmartAnnotator, instead of playing a video from start to end and simultaneously scoring behavior, the researcher is presented short video clips to annotate, based on their similarity in the videos. The similarity is derived from low-level behavior features that are precalculated from the experiment videos by EthoVision XT (`http://www.noldus.com/ethovision`). While the researcher is labeling clips, a classifier is trained on these annotations, infers the labels of similar clips, and clips with low certainty are presented to the researcher to ensure high annotation accuracy. Once all clips are labeled, the researcher can inspect the clips and edit the labels if necessary. The low-level behavior features were designed for automatic mouse and rat behavior recognition. They are described in Van Dam et al. (2013) and combine spatial body shape features, movement features, multi-scale temporal window features, and environment proximity features based on location. Together these features form a low-level behavior profile over time, independent of species, gender, age and appearance. The features have proven their richness and robustness in the Mouse and Rat Behavior Recognition modules of EthoVision XT. Also, these features were successfully used as input for a scratch behavior classifier (Van Dam et al., 2022). In order to create an AI-assisted annotation, the following steps are performed: 1) Cluster experiment data (low-level behavior features) and use these for data segmentation to create events in the data, 2) Let the human label n events, and 3) Train a behavior classifier on all the labeled data available. We first transformed the data using MiniRocket (Dempster et al., 2021), and then used a linear classification model with one fully connected layer. Steps 2 and 3 are repeated until a user-defined accuracy threshold is reached. 4) Apply the best classifier resulting from tuning to all unlabeled data, and label the certain events, 5) Pick n uncertain events and present them to the human for labeling. Steps 3-5 are repeated until all data are labeled. For the experiments in this study, we made use of a look-up table (so called oracle) with access to ground truth labels, as a replacement of the human labeler.

Annotation using supervised classification

To determine the benefits of an active learning approach using our SmartAnnotator tool for the annotation of 'stretched attend' and 'unsupported rearing', we compared it to supervised classification of those behaviors, where the classifiers were given access to the manually labeled ground truth of the entire training dataset at once. To this end, we trained a classifier for each of the two behaviors, consisting of two main parts: first, a variational autoencoder of three 1D-CNN layers to learn an embedding of the features that can be used to reconstruct the input, and second a classification head of two linear layers to estimate the behaviors from the embedding. For supervised classification, we used the same datasets as for the active learning approach. For the mouse dataset we made a training split including four out of five videos and a validation split with the remaining data. The training split contained 85.1% of the 'stretched attend' samples in the dataset. For the rat dataset we made a training split including 12 out of 14 videos and a validation split including data of the remaining two videos. In this case, the training split contained 75.1% of the 'unsupported rearing' samples. In both the active learning scenario and the supervised scenario, we optimized the classification towards high recall at the cost of low precision, because in post-processing it is easier to correct for false positives than for false negatives.

Datasets and behaviors

The mouse dataset that we used for annotation of 'stretched attend' consists of 5×5 minutes of video with annotated behaviors 'stretched attend' (121 events), 'walk' (110) and 'other' (172). Since we focused on the annotation of 'stretched attend' (Figure 5.7 pane A), the 'walk' behavior events were also considered as 'other' (resulting in 282 events for 'other'). The recordings were made for other purposes at Utrecht University, and were given to us with permission to use for our research. The rat dataset that we used for annotation of 'unsupported rearing' consists of 14×5 minutes of video, with annotated behaviors 'drink', 'eat', 'groom', 'jump', 'unsupported rearing', 'rearing supported', 'rest', 'sniff', 'walk', and 'other'. Since we focused on the annotation of 'unsupported rearing' (Figure 5.7 pane B), the rest of the behavior events were also considered as 'other', resulting in 31 events for 'unsupported rearing' and 912 events for 'other'. The dataset was reused from previous work and described in Van Dam et al. (2013). With respect to ethical permissions we remark that no animals were handled for his work.

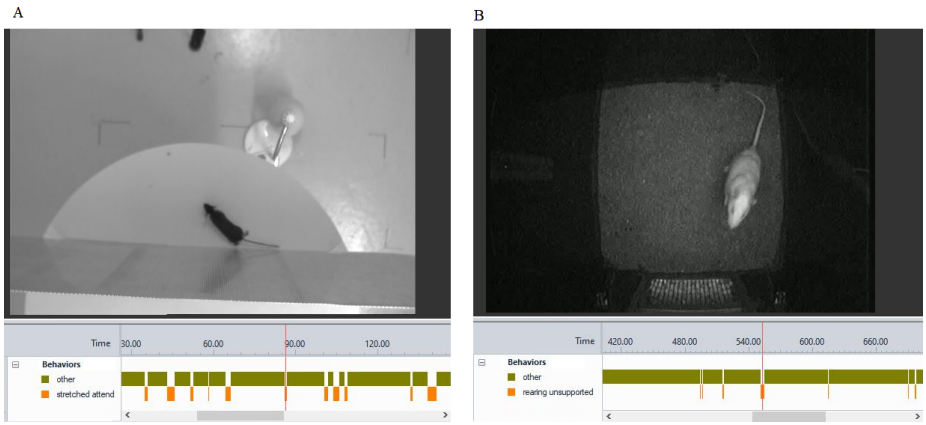


Figure 5.7: A. Screen capture of one of the 'stretched attend' videos with event logs. B. Screen capture of one of the 'unsupported rearing' videos.

5.2.3 Results

We evaluated the benefit of active learning using SmartAnnotator with two examples of specific behaviors that are easily misclassified by generic automatic tools: 'stretched attend' and 'unsupported rearing'. We analyzed precision, recall and f1-scores of the annotation using the active learning approach as well as the supervised classification approach (Table 5.2). We performed a Wilcoxon rank sum test to test for differences between the active learning and supervised approach. The active learning approach results in similar recall compared to supervised classification, for both 'stretched attend' ($p = 0.104$) and 'unsupported rearing' ($p = 0.762$; Table 5.2).

The precision, however, is higher in the active learning approach ($p < 0.001$ for both behaviors), and with that also the f1-scores ($p < 0.001$ for both behaviors; Table 5.2). The standard deviation, however, of precision is quite high, especially for 'unsupported rearing'. One explanation might be the different training data between runs because of the setup of the active learning process. Additionally, the active learning approach requires much less manually labeled data compared to supervised classification. For 'stretched attend', 18.3% manual labels was required for the active learning method (75 out of 403 events),

Table 5.2: Results of the active learning and supervised approach in classifying 'stretched attend' and 'unsupported rearing'. Reported are the true number of events of the behavior, the total number of events in the dataset, the mean percentage of manual labels (events labeled by the human) required, and precision, recall and f1-scores. The results are the mean of ten runs for each behavior and for both methods. $^{*}p < 0.001$ for the comparison between the active learning and the supervised approach.

	classification method	n events behavior	% manual labels	precision (mean ± std)	recall (mean ± std)	f1-score (mean ± std)
Stretched attend	Active learning	121 (of 403)	18.3	0.73 (± 0.13)*	0.88 (± 0.08)	0.79 (± 0.09)*
	Supervised	121 (of 403)	85.1	0.44 (± 0.04)	0.84 (± 0.03)	0.57 (± 0.04)
Unsupported rearing	Active learning	31 (of 943)	25.4	0.41 (± 0.32)*	0.75 (± 0.05)	0.46 (± 0.27)*
	Supervised	31 (of 943)	75.1	0.04 (± 0.01)	0.75 (± 0.10)	0.08 (± 0.01)

as opposed to the 85.1% (343 out of 403 events) we used for the supervised method. For 'unsupported rearing', 25.4% (239 out of 943 events) manual labels was required for active learning while we used 75.1% (708 out of 943 events) manual labels for the supervised method (Table 5.2). Overall, these results indicate that the active learning method using SmartAnnotator recalls a high number of specific behavior instances, while also reducing the number of false positives (reflected in the higher precision) compared to supervised classification of these two behaviors.

This is also reflected in Figure 5.8, where we plotted the annotation of the target behaviors by the active learning approach across the dataset, compared to ground truth. Via visual inspection we can see that the pattern over time is recovered well for both 'stretched attend' and 'unsupported rearing', and that most of the events have been retrieved.

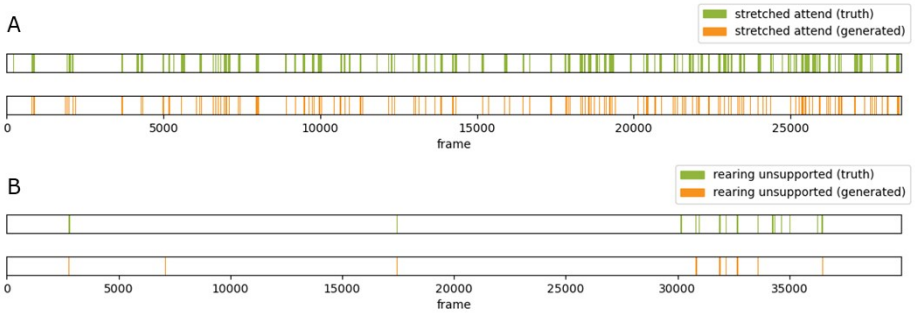


Figure 5.8: The generated (predicted) annotations using the active learning approach, compared to ground truth, of the run that resulted in statistics closest to the mean (Table 5.2). Annotations of videos in the dataset are appended. A. Generated annotation of 'stretched attend' (precision = 0.74, recall = 0.9, f1 = 0.81). B. Generated annotation of 'unsupported rearing' (precision = 0.44, recall = 0.77, f1 = 0.56). Here, we plotted the first 40000 frames (out of 95750), as in that portion most of the 'unsupported rearing' events occur.

5.2.4 Discussion

In this study, we demonstrate the benefit of an active learning approach of our SmartAnnotator tool in annotating specific or rare behaviors that are otherwise easily misclassified: 'stretched attend' and 'unsupported rearing'. The results in this study indicate

that SmartAnnotator increases annotation accuracy and reduces the number of required manual annotations which are otherwise labor-intensive and thereby decreases annotation time. While automatic annotation was already available for the most commonly observed rodent behaviors, more ambiguous or rare behaviors are still scored by hand. We demonstrate the advantage of active learning in annotating such behaviors. These results are also promising considering annotation of specific behaviors that have a clear definition but are not commonly observed. By letting the model pick uncertain events and ask the human for a label of these events, the model can be specifically trained to recognize the target behavior, even if the behavior only rarely occurs in the dataset. In such situations, we argue that the interactive way of annotating behaviors using SmartAnnotator is useful, as the algorithm can automatically annotate a large portion of the data but ask feedback from the human about parts where it is uncertain, and thereby ask for examples of a difficult behavior. In the current study, we present results for 'stretched attend' and 'unsupported rearing'. Considering that these two behaviors are difficult to automatically classify precisely, our annotations using the active learning approach are quite accurate. For a more difficult behavior like 'unsupported rearing', the model needs a relatively large number of manual labels, as is reflected in the percentage of manual labels (Table 5.2). We expect that for less difficult behavior types, the tool will need fewer manual labels to reach an accurate annotation, and furthermore yield higher precision and recall. Besides, the tool can be used to annotate multiple behaviors in the same dataset. In our current active learning approach, event segmentation is based on framewise clustering of the temporal features from EthoVision XT. This leads to far more events than would be annotated by a human observer. One of the potential improvements that is to be explored is to combine cluster traversals into coarser events, to investigate whether this could increase annotation accuracy even further. In short, the results in the current study demonstrate the benefit of applying an active learning approach to classify ambiguous or rare rodent behaviors. These are easily misclassified by conventional automatic classification and are therefore currently still often scored by hand. Tools like SmartAnnotator increase annotation accuracy of such difficult behaviors and reduce annotation time. Through the development of this AI-assisted approach we contribute a hybrid AI solution, where combining the skills of humans and machines yields better performance than using one of both.

Chapter 6

General discussion

Here the research chapters are summarized and contributions are highlighted and reflected on. After that, a general conclusion is formulated and directions for future research are pointed out.

6.1 Chapter summaries and reflections

6.1.1 Chapter 2: An automated system for the recognition of various specific rat behaviors

Summary

Automated measurement of rodent behavior is crucial to advance research in neuroscience and pharmacology. Rats and mice are used as models for human diseases; their behavior is studied to discover and develop new drugs for psychiatric and neurological disorders and to establish the effect of genetic variation on behavioral changes. Such behavior is primarily labelled by humans. Manual annotation is labour intensive, error-prone and subject to individual interpretation. In Chapter 2, a system for automated behavior recognition (ABR) is presented that recognises the rat behaviors 'drink', 'eat', 'groom', 'jump', 'rear unsupported', 'rear wall', 'rest', 'sniff', 'twitch' and 'walk'. The ABR system needs no on-site training; the only inputs needed are the sizes of the cage and the animal. This is a major advantage over other systems that need to be trained with hand-labelled data before they can be used in a new experimental setup. ABR uses an overhead camera view, which is more practical in lab situations and facilitates high-throughput testing more easily than a side-view setup. ABR has been validated by comparison with manual behavioral scoring by an expert. For this, animals were treated with two types of psychopharmaca: a stimulant drug (Amphetamine) and a sedative drug (Diazepam). The effects of drug treatment on certain behavioral categories were measured and compared for both analysis methods. Statistical analysis showed that ABR found similar behavioral effects as the human observer.

The main results from this chapter are the following:

- We presented an automated system for the recognition of the most relevant rat behaviors that performs on par with human annotation (ABR).
- ABR can be deployed real-time on a continuous video stream.
- Automated annotation with ABR is repeatable, objective and consistent. To our knowledge, ABR is the only system that can recognise behaviors across different setups. It therefore outperforms other known systems by offering equal reliability

without the need for on-site training that requires labour-intensive hand-labelled data.

- For both Amphetamine and Diazepam, ABR finds similar effects on behavior as can be found using human annotation.

Reflection

During the development of ABR around 2010, most emphasis was put on design of the feature generation to make them generically usable for the detection of diverse behaviors. This was accomplished by including diverse aspects of body shape and movement into the features and by making all distance and direction features relative to the size and orientation of the animal. Descriptive statistics at multiple timescales ensured that the features are suitable to detect behaviors of diverse durations. Using such generalized features has the benefit that it simplifies the classification task and that less training data is needed than for more complicated methods that can learn from raw input data. Nonetheless, the drawback is that it requires preprocessing. In the case of ABR, it requires reliable tracking of the animal's contour and body points (nose and tail). Also, in normalizing and generalizing information before training takes place, we might throw away informative features and not reach the best solution for the specific setup at hand. There is a balance to be found between broad usage and peak performance, which corresponds to a trade-off between robustness and flexibility. Robustness suggests that minor changes are ignored, while flexibility implies that classification can be adjusted to identify subtle differences. How can these properties be combined? Human annotators let interpretation of behavior depend on how the surrounding behaviors are performed (for instance older individuals behave slower), but that contradicts with consistent annotation across experiment trials. Out-of-the-box classifiers that apply to diverse experiment setups should learn to do what humans do: adjust their interpretations when animals behave differently than seen before. For humans, this takes agreement on the ethogram and time to learn to annotate accordingly. For automated classifiers, it might need a similar tuning phase to adjust to out-of-domain input, for instance by unsupervised or semisupervised transfer learning.

6.1.2 Chapter 3: Deep learning improves automated rodent behavior recognition within a specific experimental setup

Summary

Traditional automated systems rely on tracking input and feature preprocessing and can benefit from advances in AI. In this chapter, we explore whether it is possible to classify rat behaviors directly from the video frames in an end-to-end manner, using deep learning. In our experiments we use the Multi-Fiber network (MF-Net), which is an ensemble of lightweight networks. MF-Net performs well and generally more efficient than other networks on important benchmark datasets for human activity recognition. We show that when using this network in conjunction with data augmentation strategies, within-setup dataset performance improves over the conventional ABR module that we described in Chapter 2. In order to be useful in behavioral research, systems must perform independent of animal treatment and laboratory setup. Therefore, we also compare the results of the end-to-end method to the ABR module across experimental setups. For within-setup performance on a clipped dataset we show that MF-Net with end-to-end input outperforms both handcrafted ABR and MF-Net with tracking region-based input, provided sufficient data augmentation. For cross-setup performance on continuous video, we show that MF-Net with end-to-end input does not outperform ABR. We argue that the end-to-end model has difficulty recognizing environment cues and is not robust to differences

in behavior sequences observed. We conclude that deep learning networks give us good performance on fixed setups with known behavior, but that more research is necessary to reach adaptive and flexible human-like performance that is independent of the setup and behavior performance.

The main results from this chapter are the following:

- Deep networks improve recognition when trained and applied in equal setups.
- Improvements do not transfer to other setups or to animals behaving differently, for instance due to treatment.
- We presented two new video augmentations: video cutout and dynamic illumination change.
- The network performs better on end-to-end input than on region-based input based on tracking.
- For deployment on continuous video, it is better to train on noisier, continuous videos than on a subset of clips.

Reflection

The outcomes of Chapter 3 are good news for those who need to analyse large rodent behavior datasets in a constant setup, such as the long-term monitoring of well-being in home cages or the frequent and standardized screening of drugs that effect merely the frequency and duration of behaviors instead of how the behaviors are executed qualitatively. In these cases, it is possible to train behavior classifiers directly on the videos, without the need for intermediate steps such as animal tracking and pose estimation. This method requires a large set of manually annotated data and is resource intensive during the training phase, but has stable within-dataset accuracy for behaviors 'groom', 'rear', 'sniff' and 'walk'. It is very fast to deploy. For unknown drug effects or more subtle behaviors the method is not suitable and retraining is needed. Future research may be to increase accuracy by training on much larger and more diverse input datasets, and to add a floor plan to the input data, to let the network find relations to walls, feeder, spout and other objects more easily.

6.1.3 Chapter 4: Disentangling rodent behaviors to improve automated behavior recognition

Summary

Developments in deep learning have enabled progress in object detection and tracking, but rodent behavior recognition struggles to exceed 75–80% accuracy for ethologically relevant behaviors, for both conventional machine learning systems and deep learning systems. We distinguish three aspects of behavior dynamics that are difficult to automate, namely 1) behaviors defined by poses of minimal duration, surrounded by transitional poses that depend on the surrounding behaviors (for example 'rear'), 2) behaviors that consist of a sequence of different subbehaviors, some of which are similar to another behavior (for example 'groom-snout' and 'eat'), and 3) behaviors characterized by a fixed sequence of few short poses of minimal duration and a larger part being ambiguous (for example 'jump' and 'walk'). We isolate these aspects in an artificial dataset and reproduce effects with the state-of-the-art behavior recognition models. These newer models use self-supervised learning to first generate a lower-dimensional representation of the data before classification. Artificial datasets have the advantage that there is no limit to the amount of

labeled training data and that the noise can be regulated. The classification results on the artificial dataset show that, even with sufficient amount of data with absent noise and ideal annotation quality, and with supervised classification and hyperparameter tuning, the networks are not capable of classifying the composite rodent behaviors. Therefore, the solution towards more robust rodent behavior classification is not only to train on more data or to avoid input noise. We also need to improve on how to break down the composition automatically. If models can learn to ignore transitional poses and compress time-series into segments that correspond to behavior constituents, they can analyse segment properties and sequences regardless of the temporal scale of the segments. The usual way of segmenting data into equidistant samples and segments of equal duration is therefore not the best way to segment behavior, and adding the attention mechanism of the Transformer is not enough to overcome this.

The main results from this chapter are the following:

- We distinguish three aspects of behavior dynamics that are difficult to automate.
- We isolate these aspects in an artificial dataset and reproduce effects with two state-of-the-art behavior recognition models, namely a RNN-VAE and a Transformer model. Both models learn a self-supervised embedding first by reconstructing the input and then learn to classify the behaviors from the embedding.
- Adding more training data is not sufficient to make progress for several ethologically relevant behaviors, therefore we argue that understanding the composite nature of animal behavior is necessary to move the field forward.

Reflection

This study does not present a method that is able to model the dynamics of behavior, but instead tries to pinpoint the difficulties in behavior classification in a both precise and generic manner, by trying to understand how the shortcomings of current approaches are related to the structure of the data. For this, a broad palette of behaviors was evaluated. Mostly, publications that report results of rodent behavior classification are trained on two or three behaviors and avoid classification of more subtle and fine-grained behavior categories. The results in this work supports the idea that both the conventional and non-hierarchical deep learning methods have reached their limits for the full ethogram, even with sufficient amount of correct input data. Hopefully in the future, adding intermediate levels of subevents and key poses allows for modeling multiple hierarchical levels. So far, to our knowledge, hierarchical network architectures have not been applied to rodent behavioral timeseries, other than by modeling multiple input streams, multiple resolutions or timescales. Furthermore, it might be needed to incorporate a latent level that models the decision-making strategies of an animal. It would be interesting to see how good humans perform on classification tasks of comparable difficulty, without their common knowledge of animals and their behavior.

6.1.4 Chapter 5: Practical tools for the recognition of rodent behaviors

Robust scratching behavior detection in mice from generic features and a lightweight neural network in 100 fps videos

Summary The generic applicability and robustness of the ABR features is shown by using them to classify 'scratch' behavior of mice in two different datasets, from high-speed video recordings. The method performs well on home-cage videos recorded from top view in infrared light with a frame-rate of 100 frames per second. Although scratching events are infrequent and usually very short, our method detects events with a precision of 0.878

and recall of 0.867. Due to the use of highly generic, normalized features followed by a relatively small neural network, the method does not need a large training set and is fast to deploy.

The main results from this chapter are the following:

- We presented an automated system for the recognition of 'scratch' behavior of mice.
- The system is accurate and fast to deploy.
- The system performs in multiple recording setups without additional training or tuning.

Fast annotation of rodent behaviors with AI assistance: Human observer and SmartAnnotator collaborate through active learning

Summary AI-assisted behavior annotation saves time compared to manual annotation. Although automated systems for rodent behavior annotation exist, specific behaviors are still scored by hand, as results are not equally accurate across behaviors. With active learning, we can tailor the annotation towards the needs within a research experiment and reduce the annotation effort. In this chapter, the benefit of active learning is presented on two particular and ambiguous behaviors: 'stretched attend' and 'unsupported rearing', from the low-level features from the ABR system. Classification metrics and number of required manual labels can be improved even more for less ambiguous behaviors.

The main results from this chapter are the following:

- Active learning is beneficial for the annotation of behaviors for which no classifier or supervised training dataset is available. This can be either new behaviors or behaviors that for which classification has not successful so far, such as rare or subtle behaviors.
- We demonstrate this using the SmartAnnotator tool for the annotation of behaviors 'stretched attend' and 'unsupported rearing', from the input of the ABR features.
- The proposed solution is a hybrid AI solution: combining the skills of human and AI yields better performance than using one of both. The SmartAnnotator increases annotation accuracy and decreases manual annotation effort.

Reflection

This chapter describes two methods for adapting classification to new behaviors. The first method is useful in stable, standardized setups: Create a new, frame-accurately annotated dataset that contains representative videos with respect to the animal's appearance and behavior, and train a supervised classifier on it. For many behaviors, this will produce a classifier that can be used on unseen videos as long as they contain behaviors executed in the same way as the behaviors in the training set. When classification accuracy is below the desired accuracy (for instance because the behavior is characterized by on of the aspects that were mentioned in Chapter 4), the model can be tuned to favor false positives over false negatives. The researcher can review and interpret the false positives, which is less effort than viewing the entire dataset.

In case of research datasets in changing setups or with changing execution of the behavior due to drug treatment, the effort of frame-accurate annotation and training a supervised classifier will most likely not pay off. In these cases, the second method can be used. The SmartAnnotator acts as an AI assistant that takes over the annotation from the researcher by asking labels for specific fragments and by proposing annotations for the

others. The small study presented in Chapter 5.2 implies that this is possible, although it was demonstrated here in small datasets and on two behaviors only. Yet, it is a promising hybrid solution that gives the human control over the interpretation of the behaviors and is transparent during the process. As such, it is a nice example of explainable and responsible AI.

6.2 General conclusions and future directions

This thesis presented an automated rodent behavior recognition system (ABR) that performs on par with human annotation for many, although not for all behaviors. The raw, handcrafted ABR features serve as a generic embedding of behavior that can be used as input for classification of diverse behaviors of both rats and mice. Classification can be trained either in a fully supervised or in a self-supervised manner, or in an active learning scenario. With active learning, the human annotator teams up with AI to annotate together by prioritizing the labeling of more informative samples. To optimize the learning process this way makes the annotation process more efficient with less annotation work for the human annotator. At the same time, the precision is improved by giving the most difficult samples to the human to decide.

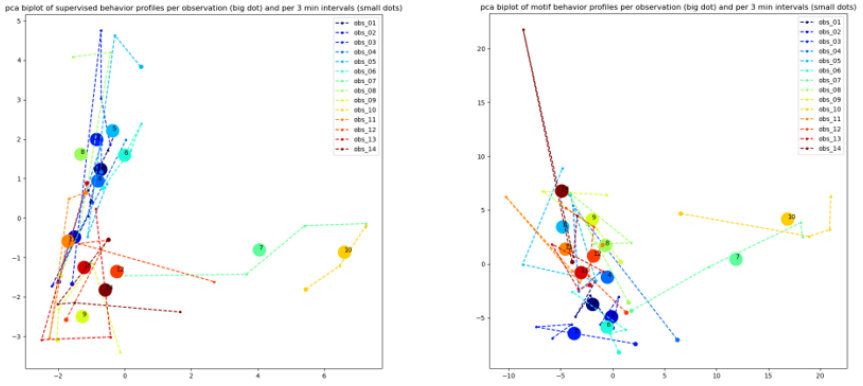
Deep networks improve recognition when trained and applied in identical setups, but improvements do not transfer to other setups or to animals behaving differently, for instance due to treatment. There are three aspects of behavior dynamics that are difficult to automate for all models and by using artificial data it was shown that adding more training data is not sufficient to make progress for these categories. This indicates that we need network architectures that can model the composite nature of animal behavior to move the field forward. Whether this modeling is best done while learning the representation of the data or during classification of the representation is an open research question.

During this project, in research not described in this thesis, I tried to segment the data stream based on the hypothesis that as long as the data stream is predictable the same behavior is conducted. The rationale behind this is that rodents are intentional agents that exhibit goal-driven behavior. Their consecutive goals can be viewed as states, that are linked together through state transitions when their goals change. If we can predict the behavior within a certain behavioral state, it is perhaps possible to infer state changes from failing predictions. If this is possible we can model the data as a consecutive set of states and state transitions, and build a dictionary of states, which would yield a representation of behavior at a higher level than that of frames or frame windows and a step closer to an ethogram that human annotators use. To infer the states from the data has been tried before by using autoregressive Hidden Markov Models (AR-HMMs) (Markowitz et al., 2018). However, predicting the continuous rodent behavior data turned out to be very difficult to steer, because the complexity of the task is not constant and the networks were easily overfitted if they are too large or fail to learn if they are too small. The impossibility to predict continuous video data at the lowest level of frames and pixels was discussed recently by Yann LeCun[1], during a discussion whether it is possible to train a generative model for video analogue to the successfully Large Language Models (LLMs) trained on text. LeCun argues that the world is incredibly more complicated and richer in terms of information than text, and therefore it is far more complicated to predict video than it is to predict text. Text is discrete, whereas video is high-dimensional and continuous. The proposed solution for this has been to model a representative embedding. However, he explains, this has not worked out in the last ten years, despite many attempts.

[1]Lex Fridman Podcast #416, March 7, 2024 - Video Prediction `https://www.youtube.com/watch?v=5t1vTLU7s40&t=1066s`

What should be done instead, according to LeCun, is do the prediction in representation space, which allows the system to learn an abstract representation of the world where what can be modeled and predicted is preserved and the rest is viewed as noise and eliminated by the encoder. This lifts the level of abstraction of the representation. I believe that the same is true for the prediction of multivariate timeseries recorded from a behavioral state-switching intentional agent such as a rodent. It is not feasible to build a generative model for the low level timeseries as is, but it must be possible to learn a more abstract representation that is better suited to classify behaviors from. This requires a new type of modeling, and it would be interesting to try the newest Joint Embedding Predictive Architectures for video (V-JEPA) (Bardes et al., 2024) on the rodent datasets.

Another interesting direction to explore in future work is to loosen the focus on specific behaviors to annotate, but instead see what a completely unsupervised approach can tell about the behavioral differences between animal groups or individuals. In a preliminary experiment I trained an embedding on all the ABR features from an experiment with fourteen animals that were treated with different compounds. After training the embedding, the embedding of the entire dataset was clustered into 30 prototypes. Next, all timeseries were segmented based on the clustering, and behavior profiles were created per 3-minute interval by counting the cluster types and cluster transitions. After a Principal Components Analysis (PCA) over all the data, the intervals and the mean interval per animal were projected on the two main axes of the PCA in Figure 6.1. The distances between the animals are very similar, which suggests that the same behavioral effects of compounds can be retrieved from fully unsupervised data analysis as can be found by analysis of the manually annotated behavior frequencies. More research is needed to confirm these results in other datasets and to see whether this can be used for subtle behavioral effects as well.



(A) (B)

Figure 6.1: Biplots with the preliminary results of animal similarities based on either manual annotation Panel **(A)** and unsupervised annotation Panel **(B)**. The plots suggest that the same behavioral effects can be found with fully unsupervised data analysis as with manual behavior annotation analysis.

Furthermore, with the fast growing number of solutions that can recognize rodent behaviors automatically, it would be useful to have a representative benchmark to compare the methods. Ideally, this benchmark should contain recordings from multiple research labs, multiple experiments, multiple strains and include different types of behaviors, in-

cluding behaviors that are difficult to classify, such as the compound behavioral categories mentioned in Chapter 4. However, the difficulty with such a benchmark is the amount of annotated data that is needed for training and evaluation of the methods, as well as the amount of work to host and monitor the benchmark.

Finally, future work can be to extend the project to other species and multiple subjects. Although the generated ABR features depend on video tracking and are specific for rodents recorded from a camera with a top-view perspective, the behavior recognition strategies are independent of this and can be applied to other streams of data, such as audio with ultrasonic vocalizations, or physiological data like heart rate and temperature, or brain signals from EEG or fMRI. The active learning strategy can perhaps even be combined with the end-to-end processing to avoid the tracking altogether, if synchronized video is available and the datasets is sufficiently large to cover the variance.

To conclude, there are many behavioral research applications were automated annotation is eagerly waited for and that can benefit from the work in this thesis. The exciting progress in artificial intelligence algorithms, hardware and deployment options will certainly make more automation possible in the near future.

6.3 Ethics: Animal testing, automated behavior recognition and AI

The work described in this thesis combines three ethically sensitive themes, namely animal testing, automated behavior recognition and artificial intelligence. All three themes can contribute to greater science output and thus to increasing prosperity and well-being. Although major investments are already made in replacing animal testing (`https://www.animalfreeinnovationtpi.nl/`), tests are still indispensable for safety checks on food and medicines, for example. In scientific research, animals are used to understand human physiology and to learn how organisms detect and interact with each other and their environment. Studies of brain function, such as memory and social behavior, still need animal models (Homberg et al., 2021). Automated behavior annotation can increase the throughput of such studies and make them more reliable, thus contributing to the guiding principles for more ethical use of animals in testing: the 3R's described by Russell et al. (1959), namely Replacement, Reduction and Refinement. More data per animal with 24/7 observation can reduce the number of animals needed and automated home-cage monitoring can contribute to stress reduction (Refinement) by detecting behavioral changes without subjecting animals to unnecessary handling or exposure to novel environments (Fuochi et al., 2024). With artificial intelligence, automated behavior recognition can be improved.

Like all technology, animal testing, automated behavior monitoring and artificial intelligence can be used for malicious purposes that are harmful for the well-being of animals and humans. Behavior monitoring of people can harm privacy and freedom. The use of AI can accelerate this even further and make it more easily accessible to people or institutions (governments, companies) or, if you like, to other AIs with bad intentions (see Tegmark (2018) for different future scenarios of what AIs that supersede human intelligence could mean for society). It is therefore important to regulate its use. Animal testing must be assessed for necessity, the number of animals required must be minimized, and the tests must be conducted in such a way that they minimize the amount of discomfort, pain, and permanent damage and maximize welfare. These conditions are checked by ethics committees in most countries, and animal testing is not possible without prior approval of the experiment. The same regulations and law enforcement are necessary for the use of automated monitoring tools, which must be limited to well-defined purposes that are non-biased, transparent, with respect for privacy and security, to prevent social manipulation

and misinformation. The recently introduced EU AI Act[2] regulates the use of artificial intelligence in Europe and tries to find the right balance between ethical values and technological innovation. It is also up to us, AI researchers and AI engineers, to ensure that we create ethical, usable, and beneficial applications for society. I strongly believe that building behavioral research tools that combine the processing capacity and scalability of AI with human expertise and guidance contributes to this.

[2] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Bibliography

Aggarwal, J. K. and Ryoo, M. S. (2011). Human activity analysis: A review. *Acm Computing Surveys (Csur)*, 43(3).

Aitchison, J. and Greenacre, M. (2002). Biplots of compositional data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 51(4):375–392.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 2623–2631.

Akita, S., Tsuichihara, S., and Takemura, H. (2019). Detection of rapid mouse's scratching behavior based on shape and motion features. In *Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 925–928. IEEE.

Aristotle (2000). *De Partibus Animalium, De Motu Animalium en De Incessu Animalium.* Historische uitgeverij Groningen. Originally written 384-322 BC, Dutch translation: Over Dieren, by R. Ferwerda.

Arroyo-Araujo, M., Graf, R., Maco, M., Dam, E. A. van., Schenker, E., Drinkenburg, W., Koopmans, B., Boer, S. F. de., Cullum-Doyle, M., Noldus, L. P. J. J., Loos, M., Dommelen, W. van., Spooren, W., Biemans, B., Buhl, D. L., and Kas, M. J. (2019). Reproducibility via coordinated standardization: a multi-center study in a Shank2 genetic rat model for autism spectrum disorders. *Scientific Reports*, 9(1):11602.

Baran, S. W., Bratcher, N., Dennis, J., Gaburro, S., Karlsson, E. M., Maguire, S., Makidon, P., Noldus, L. P. J. J., Potier, Y., Rosati, G., et al. (2022). Emerging role of translational digital biomarkers within home cage monitoring technologies in preclinical drug discovery and development. *Frontiers in Behavioral Neuroscience*, 15:758274.

Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., and Ballas, N. (2024). Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*.

Bateson, M. and Martin, P. (2007). *Measuring behaviour: an introductory guide. Chapter7: How good are your measures - observer reliability.* Cambridge University Press, third edition.

Batty, E., Whiteway, M., Saxena, S., Biderman, D., Abe, T., Musall, S., Gillis, W., Markowitz, J., Churchland, A., Cunningham, J. P., et al. (2019). BehaveNet: nonlinear embedding and Bayesian neural decoding of behavioral videos. In *Advances in Neural Information Processing Systems*, volume 32.

Benjamini, Y., Fonio, E., Galili, T., Havkin, G. Z., and Golani, I. (2011). Quantifying the buildup in extent and complexity of free exploration in mice. *Proceedings of the National Academy of Sciences*, 108:15580–15587.

Braak, C. J. F. ter. and Šmilauer, P. (2012). *Canoco reference manual and user's guide: software for ordination, version 5.0*. Microcomputer power, Ithaca, USA.

Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., and Ramesh, A. (2024). Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators. Accessed: 2024-09-15.

Burgos-Artizzu, X. P., Dollár, P., Lin, D., Anderson, D. J., and Perona, P. (2012). Social behavior recognition in continuous video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1322–1329.

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the Kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308.

Casarrubea, M., Magnusson, M. S., Anguera, M. T., Jonsson, G. K., Castañer, M., Santangelo, A., Palacino, M., Aiello, S., Faulisi, F., Raso, G., Puigarnau, S., Camerino, O., Di Giovanni, G., and Crescimanno, G. (2018). T-pattern detection and analysis for the discovery of hidden features of behaviour. *Journal of Neuroscience Methods*, 310:24–32.

Chen, Y., Kalantidis, Y., Li, J., Yan, S., and Feng, J. (2018). Multi-fiber networks for video recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 352–367.

Costacurta, J., Duncker, L., Sheffer, B., Gillis, W., Weinreb, C., Markowitz, J., Datta, S. R., Williams, A., and Linderman, S. (2022). Distinguishing discrete and continuous behavioral variability using warped autoregressive HMMs. In *Advances in Neural Information Processing Systems*, volume 35, pages 23838–23850.

Crabbe, J. C., Wahlsten, D., and Dudek, B. C. (1999). Genetics of mouse behavior: interactions with laboratory environment. *Science*, 284(5420):1670–1672.

Dam, E. A. van., Harst, J. E. van der., Braak, C. J. F. ter., Tegelenbosch, R. A. J., Spruijt, B. M., and Noldus, L. P. J. J. (2013). An automated system for the recognition of various specific rat behaviours. *Journal of Neuroscience Methods*, 218(2):214–224.

Dam, E. A. van., Noldus, L. P. J. J., and Gerven, M. A. J. van. (2020). Deep learning improves automated rodent behavior recognition within a specific experimental setup. *Journal of Neuroscience Methods*, 332:108536.

Dam, E. A. van., Noldus, L. P. J. J., and Gerven, M. A. J. van. (2023). Disentangling rodent behaviors to improve automated behavior recognition. *Frontiers in Neuroscience*, 17:1198209.

Dam, E. A. van., Roosken, M. H., and Noldus, L. P. J. J. (2022). Robust scratching behavior detection in mice from generic features and a lightweight neural network in 100 fps videos. In *Volume 2 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods*, pages 301–305.

Dankert, H., Wang, L., Hoopfer, E. D., Anderson, D. J., and Perona, P. (2009). Automated monitoring and analysis of social behavior in Drosophila. *Nature Methods*, 6(4):297–303.

Darwin, C. R. (1872). *The Expression of the Emotions in Man and Animals*. John Murray, London.

Datta, S. R. (2019). Q&A: Understanding the composition of behavior. *BMC Biology*, 17(1):44.

Dempster, A., Schmidt, D. F., and Webb, G. I. (2021). MiniRocket: A very fast (almost) deterministic transform for time series classification. In *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 248–257.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

DeVries, T. and Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Dollár, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *Proceedings of the IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, volume 2005, pages 65–72.

Drai, D., Kafkafi, N., Benjamini, Y., Elmer, G., and Golani, I. (2001). Rats and mice share common ethologically relevant parameters of exploratory behavior. *Behavioural Brain Research*, 125(1-2):133–140.

Drew, K. and Glick, S. (1988). Characterization of the associative nature of sensitization to amphetamine-induced circling behavior and of the environment dependent placebo-like response. *Psychopharmacology*, 95:482–487.

Drew, K. L. and Glick, S. D. (1990). Role of D-1 and D-2 receptor stimulation in sensitization to amphetamine-induced circling behavior and in expression and extinction of the Pavlovian conditioned response. *Psychopharmacology*, 101:465–471.

Duda, R. O., Hart, P. E., et al. (2001). *Pattern classification*. John Wiley & Sons, second edition.

Dunne, F., O'Halloran, A., and Kelly, J. P. (2007). Development of a home cage locomotor tracking system capable of detecting the stimulant and sedative properties of drugs in rats. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 31(7):1456–1463.

Eyjolfsdottir, E., Branson, K., Yue, Y., and Perona, P. (2016). Learning recurrent representations for hierarchical behavior modeling. *arXiv preprint arXiv:1611.00094*.

Faraway, J. J. (2006). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394.

Fonio, E., Benjamini, Y., and Golani, I. (2009). Freedom of movement and the stability of its unfolding in free exploration of mice. *Proceedings of the National Academy of Sciences*, 106(50):21335–21340.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition.* Academic Press Professional, second edition.

Fuochi, S., Rigamonti, M., O'Connor, E. C., De Girolamo, P., and D'Angelo, L. (2024). Big data and its impact on the 3Rs: a home cage monitoring oriented review. *Frontiers in Big Data*, 7:1390467.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning.* The MIT Press. `http://www.deeplearningbook.org`.

Goodwin, N. L., Choong, J. J., Hwang, S., Pitts, K., Bloom, L., Islam, A., Zhang, Y. Y., Szelenyi, E. R., Tong, X., Newman, E. L., et al. (2024). Simple Behavioral Analysis (SimBA) as a platform for explainable machine learning in behavioral neuroscience. *Nature Neuroscience*, pages 1–14.

Graving, J. M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B. R., and Couzin, I. D. (2019). DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *eLife*, 8:e47994.

Greenacre, M. J. (2012). Biplots: the joy of singular value decomposition. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(4):399–406.

Grieco, F., Bernstein, B. J., Biemans, B., Bikovski, L., Burnett, C. J., Cushman, J. D., Dam, E. A. van., Fry, S. A., Richmond-Hacham, B., Homberg, J. R., Kas, M. J. H., Kessels, H. W., Koopmans, B., Krashes, M. J., Krishnan, V., Logan, S., Loos, M., McCann, K. E., Parduzi, Q., Pick, C. G., Prevot, T. D., Riedel, G., Robinson, L., Sadighi, M., Smit, A. B., Sonntag, W., Roelofs, R. F., Tegelenbosch, R. A. J., and Noldus, L. P. J. J. (2021). Measuring behavior in the home cage: Study design, applications, challenges, and perspectives. *Frontiers in Behavioral Neuroscience*, 15:735387.

Gupta, S. and Gomez-Marin, A. (2019). A context-free grammar for Caenorhabditis elegans behavior. *bioRxiv preprint bioRxiv:10.1101/708891.*

Harst, J. E. van der., Baars, A.-M., and Spruijt, B. M. (2003a). Standard housed rats are more sensitive to rewards than enriched housed rats as reflected by their anticipatory behaviour. *Behavioural Brain Research*, 142(1-2):151–156.

Harst, J. E. van der., Fermont, P. C. J., Bilstra, A. E., and Spruijt, B. M. (2003b). Access to enriched housing is rewarding to rats as reflected by their anticipatory behaviour. *Animal Behaviour*, 66(3):493–504.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., and Lange, F. P. de. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, 119(32):e2201968119.

Homberg, J. R., Adan, R. A., Alenina, N., Asiminas, A., Bader, M., Beckers, T., Begg, D. P., Blokland, A., Burger, M. E., Dijk, G. van., et al. (2021). The continued need for animals to advance brain research. *Neuron*, 109(15):2374–2379.

Huang, G., Liu, Z., Maaten, L. van der., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708.

Isik, S. and Unal, G. (2023). Open-source software for automated rodent behavioral analysis. *Frontiers in Neuroscience*, 17:1149027.

Jhuang, H., Garrote, E., Yu, X., Khilnani, V., Poggio, T., Steele, A. D., and Serre, T. (2010). Automated home-cage behavioural phenotyping of mice. *Nature Communications*, 1(6):1–9.

Jiang, Z., Crookes, D., Green, B. D., Zhao, Y., Ma, H., Li, L., Zhang, S., Tao, D., and Zhou, H. (2019). Context-aware mouse behavior recognition using hidden markov models. *IEEE Transactions on Image Processing*, 28(3):1133–1148.

Kabra, M., Robie, A. A., Rivera-Alba, M., Branson, S., and Branson, K. (2013). JAABA: interactive machine learning for automatic annotation of animal behavior. *Nature Methods*, 10(1):64–67.

Keller, G. B. and Mrsic-Flogel, T. D. (2018). Predictive processing: A canonical cortical computation. *Neuron*, 100(2):424–435.

Kim, T., Ahn, S., and Bengio, Y. (2019). Variational temporal abstraction. In *Advances in Neural Information Processing Systems*, pages 11570–11579.

Kobayashi, K., Matsushita, S., Shimizu, N., Masuko, S., Yamamoto, M., and Murata, T. (2021). Automated detection of mouse scratching behaviour using convolutional recurrent neural network. *Scientific Reports*, 11(1):658.

Kramida, G., Aloimonos, Y., Parameshwara, C. M., Fermüller, C., Francis, N. A., and Kanold, P. (2016). Automated mouse behavior recognition using VGG features and LSTM networks. In *Proceedings of the Visual Observation and Analysis of Vertebrate and Insect Behavior Workshop (VAIB)*, pages 1–3.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). HMDB: A large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 2556–2563.

Lauer, J., Zhou, M., Ye, S., Menegas, W., Schneider, S., Nath, T., Rahman, M. M., Di Santo, V., Soberanes, D., Feng, G., Murthy, V. N., Lauder, G., Dulac, C., Mathis, M. W., and Mathis, A. (2022). Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nature Methods*, 19(4):496–504.

Le, V. A. and Murari, K. (2019). Recurrent 3D convolutional network for rodent behavior recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1174–1178.

List, T., Bins, J., Vazquez, J., and Fisher, R. B. (2005). Performance evaluating the evaluator. In *Proceedings of the IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 129–136.

Lorbach, M., Poppe, R., and Veltkamp, R. C. (2019). Interactive rodent behavior annotation in video using active learning. *Multimedia Tools and Applications*, 78:19787–19806.

Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th International Joint Conference on Artificial Intelligence*, volume 2, pages 674–679.

Luxem, K., Mocellin, P., Fuhrmann, F., Kürsch, J., Miller, S. R., Palop, J. J., Remy, S., and Bauer, P. (2022). Identifying behavioral structure from deep variational embeddings of animal motion. *Communications Biology*, 5(1):1267.

Ma, Q., Zheng, J., Li, S., and Cottrell, G. W. (2019). Learning representations for time series clustering. In *Advances in Neural Information Processing Systems*, volume 32, pages 3781—3791.

Markowitz, J. E., Gillis, W. F., Beron, C. C., Neufeld, S. Q., Robertson, K., Bhagat, N. D., Peterson, R. E., Peterson, E., Hyun, M., Linderman, S. W., Sabatini, B. L., and Datta, S. R. (2018). The striatum organizes 3D behavior via moment-to-moment action selection. *Cell*, 174(1):44–58.e17.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., and Bethge, M. (2018). DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21:1281–1289.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized linear mixed models*, volume 325. Wiley, New York, second edition.

Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., and Legg, S. (2023). Levels of AGI: Operationalizing progress on the path to AGI. *arXiv preprint arXiv:2311.02462*.

Nascimento Alvarez, J. Do., Fukushiro, D. F., Tatsu, J. A. O., De Carvalho, E. P., De Castro Gandolfi, A. C., Tsuchiya, J. B., Carrara-Nascimento, P. F., Lima, M. L., Bellot, R. G., and Frussa-Filho, R. (2006). Amphetamine-induced rapid-onset sensitization: role of novelty, conditioning and behavioral parameters. *Pharmacology Biochemistry and Behavior*, 83(4):500–507.

Newell, A. (2022). *Learning to solve structured vision problems*. PhD Thesis, Princeton University.

Nie, Y., Ishii, I., Tanaka, A., and Matsuda, H. (2012). Automatic scratching analyzing system for laboratory mice: SCLABA-Real. In *Human-Centric Machine Vision*. Intech Open.

Noldus, L. P. J. J., Spink, A. J., and Tegelenbosch, R. A. J. (2001). EthoVision: a versatile video tracking system for automation of behavioral experiments. *Behavior Research Methods, Instruments, & Computers*, 33:398–414.

OpenAI (2022a). ChatGPT. `https://openai.com/index/chatgpt/`. Accessed: 2024-09-15.

OpenAI (2022b). Dall-E. `https://openai.com/index/dall-e-2/`. Accessed: 2024-09-15.

Pereira, T. D., Tabris, N., Matsliah, A., Turner, D. M., Li, J., Ravindranath, S., Papadoyannis, E. S., Normand, E., Deutsch, D. S., Wang, Z. Y., et al. (2022). SLEAP: A deep learning system for multi-animal pose tracking. *Nature Methods*, 19(4):486–495.

Perez, M. and Toler-Franklin, C. (2023). CNN-based action recognition and pose estimation for classifying animal behavior from videos: A survey. *arXiv preprint arXiv:2301.06187*.

Pérez-Escudero, A., Vicente-Page, J., Hinz, R. C., Arganda, S., and De Polavieja, G. G. (2014). idTracker: tracking individuals in a group by automatic identification of unmarked animals. *Nature Methods*, 11(7):743–748.

Roser, M. (2023). AI timelines: What do experts in artificial intelligence expect for the future? *Our World in Data*. https://ourworldindata.org/ai-timelines. Accessed: 2024-09-15.

Rousseau, J. B. I., Lochem, P. B. A. van., Gispen, W. H., and Spruijt, B. M. (2000). Classification of rat behavior with an image-processing method and a neural network. *Behavior Research Methods, Instruments, & Computers*, 32(1):63–71.

Russell, W. M. S., Burch, R. L., Hume, C. W., et al. (1959). *The principles of humane experimental technique*, volume 238. Methuen & Co, London.

Segalin, C., Williams, J., Karigo, T., Hui, M., Zelikowsky, M., Sun, J. J., Perona, P., Anderson, D. J., and Kennedy, A. (2021). The mouse action recognition system (MARS): a software pipeline for automated analysis of social behaviors in mice. *eLife*, 10:e63720.

Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576.

Soomro, K., Zamir, A. R., and Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Spruijt, B. M., Hol, T., and Rousseau, J. (1992). Approach, avoidance, and contact behavior of individually recognized animals automatically quantified with an imaging technique. *Physiology & Behavior*, 51(4):747–752.

Spruijt, B. M. and Visser, L. de. (2006). Advanced behavioural screening: automated home cage ethology. *Drug Discovery Today: Technologies*, 3(2):231–237.

Steele, A. D., Jackson, W. S., King, O. D., and Lindquist, S. (2007). The power of automated high-resolution behavior analysis revealed by its application to mouse models of Huntington's and prion diseases. *Proceedings of the National Academy of Sciences*, 104(6):1983–1988.

Sun, J. J., Kennedy, A., Zhan, E., Anderson, D. J., Yue, Y., and Perona, P. (2021). Task programming: Learning data efficient behavior representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2876–2885.

Sun, J. J., Marks, M., Ulmer, A. W., Chakraborty, D., Geuther, B., Hayes, E., Jia, H., Kumar, V., Oleszko, S., Partridge, Z., et al. (2023). MABe22: A multi-species multi-task benchmark for learned representations of behavior. In *Proceedings of the International Conference on Machine Learning*, pages 32936–32990.

Tegmark, M. (2018). *Life 3.0: Being human in the age of artificial intelligence*. Penguin Books.

Tinbergen, N. (1951). *The study of instinct*. Clarendon Press/Oxford University Press.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6459.

Vanderschuren, L. J., Schoffelmeer, A. N., Mulder, A. H., and De Vries, T. J. (1999). Dopaminergic mechanisms mediating the long-term expression of locomotor sensitization following pre-exposure to morphine or amphetamine. *Psychopharmacology*, 143:244–253.

Wahlsten, D., Metten, P., Phillips, T. J., Boehm, S. L., Burkhart-Kasch, S., Dorow, J., Doerksen, S., Downing, C., Fogarty, J., Rodd-Henricks, K., et al. (2003). Different data from different labs: lessons from studies of gene-environment interaction. *Journal of Neurobiology*, 54(1):283–311.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 20–36.

Weinreb, C., Pearl, J. E., Lin, S., Osman, M. A. M., Zhang, L., Annapragada, S., Conlin, E., Hoffmann, R., Makowska, S., Gillis, W. F., et al. (2024). Keypoint-MoSeq: parsing behavior by linking point tracking to pose dynamics. *Nature Methods*, 21(7):1329–1339.

Wiltschko, A. B., Johnson, M. J., Iurilli, G., Peterson, R. E., Katon, J. M., Pashkovski, S. L., Abraira, V. E., Adams, R. P., and Datta, S. R. (2015). Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135.

Wiltschko, A. B., Tsukahara, T., Zeine, A., Anyoha, R., Gillis, W. F., Markowitz, J. E., Peterson, R. E., Katon, J., Johnson, M. J., and Datta, S. R. (2020). Revealing the structure of pharmacobehavioral space through motion sequencing. *Nature Neuroscience*, 23(11):1433–1443.

Würbel, H. (2002). Behavioral phenotyping enhanced-beyond (environmental) standardization. *Genes, Brain and Behavior*, 1(1):3–8.

Xu, H., Zhou, P., Tan, R., Li, M., and Shen, G. (2021). LIMU-BERT: Unleashing the potential of unlabeled data for IMU sensing applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, pages 220–233. ACM.

English summary

Research into animal behavior has since long been of great importance to our understanding of biological and psychological phenomena. Given their genetic similarities to humans, rats and mice serve as models for human disease, allowing researchers to determine the effectiveness and safety of treatments for psychiatric and neurological disorders. The behavior of transgenic rodents provides valuable insights into the genetic underpinnings of brain disorders and the function of specific proteins and genes. Objective recording of behavior is crucial for reliable and reproducible research. Long-term observation is important to collect as much data as possible per individual, thus reducing the total number of laboratory animals required for the experiments. It is also important to monitor the welfare of laboratory animals. However, manually annotation of animal behavior by human observers is time-consuming (and therefore expensive), difficult and error-prone, and therefore unsuitable as input for systems that must be able to work continuously, without intervention.

This thesis describes how rodent behavior can be automatically annotated and investigates how this can be made more robust, generic and flexible. The recognized behavior is limited to the behavior of individually housed rodents (rat or mouse) that are observed with a camera placed above the cage, with constant background and infrared lighting. **Chapter 2** describes the traditional approach for activity recognition, namely classifying generic, hand-defined features computed from videos. It presents the Automated Behavior Recognition (ABR) system for the recognition of specific rat behaviors most commonly annotated by hand: 'drink', 'eat', 'groom', 'jump', 'rear unsupported', 'rear wall', 'rest', 'sniff', 'twitch' and 'walk'. The system is validated on an unseen video by comparison with manual behavior annotation by an expert. In addition, the effects of two medications on the behavioral categories are measured and compared for both annotation methods. The measured effects are similar for both treatments. **Chapter 3** investigates whether more generic classification of rodent behavior is possible by using deep learning to infer rat behavior directly from the video frames in an end-to-end manner, without prior tracking and without generation of predefined features. Performance is evaluated within and across experimental setups. It shows that using a 3D-convolutional network in combination with data-augmentation strategies improves recognition within the setup compared to the traditional ABR system. However, it also shows that improvements do not occur for videos that were recorded in different experimental setups. Finally, possible causes and treatments are discussed.

Chapter 4 explores the main reasons why rodent behavior recognition generally does not reach 100% accuracy and performs poorly for certain behaviors. Three aspects are distinguished in the dynamics of behavior that are difficult to automate. These aspects are isolated in an artificial dataset and the results are reproduced on the artificial data, using state-of-the-art behavioral recognition models. These newer models use self-supervised learning to first generate a lower-dimensional representation of the data before classification.

The final research chapter, **Chapter 5**, elaborates on the practical solutions and tools that can help behavioral researchers annotate new behaviors for which no previously

trained classifier is available. Certain research experiments require flexible classification, for example in the case of new setups or new behavioral effects. The first part of the chapter, **Chapter 5a** proves the robustness and generic applicability of the ABR features by using them to classify mouse scratching behavior in two different datasets, from high-speed video recordings. In the second part, **Chapter 5b**, the possibility of combining manual annotation with AI assistance into a hybrid active learning solution is explored, such that the annotation process is on the one hand more efficient than fully manual annotation and on the other hand, that the end result is more accurate than fully automated annotation. The advantage of active learning is presented on the behaviors 'stretched attend' and 'unsupported rearing'.

The dissertation concludes with a discussion in **Chapter 6**, which highlights the contributions per research chapter and explains the shortcomings and future solutions. Furthermore, it provides a glimpse into new ways of behavioral analysis that take advantage of completely unsupervised detection of behavioral effects that allow behavioral researchers to interactively investigate their data and the behavioral effects between experimental groups. The discussion chapter ends with a reflection on the ethical implications of research into automated rodent behavior recognition.

In conclusion, there are many behavioral research applications where robust and flexible automated annotation of behavior is eagerly awaited. The work in this thesis meets this need with an automatic system and provides directions for further development of behavioral recognition. The exciting advances in artificial intelligence algorithms, hardware and deployment options will certainly enable more and more flexible automation in the near future.

Nederlandse samenvatting

Onderzoek naar diergedrag is sinds lange tijd van groot belang voor ons begrip van biologische en psychologische verschijnselen. Gezien hun genetische overeenkomsten met mensen dienen ratten en muizen als modellen voor ziekten bij de mens, waardoor onderzoekers de effectiviteit en veiligheid van behandelingen van psychiatrische en neurologische aandoeningen kunnen bepalen. Het gedrag van transgene knaagdieren biedt waardevolle inzichten in de genetische onderbouwing van hersenaandoeningen en de functie van specifieke eiwitten en genen. Voor betrouwbaar en reproduceerbaar onderzoek is het objectief vastleggen van het gedrag cruciaal. Lange-termijnobservatie is belangrijk om per individu zoveel mogelijk data te verzamelen, waardoor het totaal aantal proefdieren dat nodig is voor de experimenten wordt verminderd. Het is eveneens belangrijk om het welzijn van proefdieren te monitoren. Echter, het handmatig vastleggen van het diergedrag door menselijke waarnemers is tijdrovend (en dus kostbaar), moeizaam en foutgevoelig, en daarom ongeschikt als invoer voor systemen die volcontinu en zonder interventie moeten kunnen werken.

Dit proefschrift beschrijft hoe het gedrag van knaagdieren automatisch kan worden geannoteerd en onderzoekt hoe dit robuuster, generieker en flexibeler gemaakt kan worden. Het herkende gedrag beperkt zich tot het gedrag van individueel-gehuisveste knaagdieren (rat of muis) die worden geobserveerd met een camera die boven de kooi is geplaatst, en waarvan de achtergrond- en infraroodverlichting constant is. **Hoofdstuk 2** beschrijft de traditionele aanpak in patroonherkenning, namelijk door classificatie van generieke, handgedefinieerde features (kenmerken) die worden berekend uit video's. Het presenteert het Automated Behavior Recognition (ABR)-systeem voor de herkenning van specifieke rattengedragingen die het meest met de hand worden geannoteerd: 'drinken', 'eten', 'lopen', 'niet-ondersteund oprichten', 'oprichten tegen de muur', 'rusten', 'snuffelen', 'springen', 'uitschudden' en 'wassen'. Het systeem wordt gevalideerd op een ongeziene video door vergelijking met handmatige gedragsscores door een expert. Daarnaast worden voor beide annotatiemethodes de effecten van twee medicamenten op de gedragscategorieën gemeten en vergeleken. Voor beide behandelingen komen de gemeten effecten overeen. **Hoofdstuk 3** onderzoekt of generiekere classificatie van knaagdiergedrag mogelijk is door met behulp van deep learning het gedrag van ratten rechtstreeks uit de videoframes af te leiden op een end-to-end manier, zonder tracking vooraf en zonder vooraf gedefinieerde features te genereren. De prestaties worden geëvalueerd binnen en tussen experimentele opstellingen. Het laat zien dat het gebruik van een 3D-convolutioneel netwerk in combinatie met data-augmentatiestrategieën de herkenning binnen de setup verbetert ten opzichte van het traditionele ABR-systeem. Het laat echter ook zien dat verbeteringen niet optreden voor video's in verschillende experimentele opstellingen. Tenslotte worden mogelijke oorzaken en behandelingen besproken.

Hoofdstuk 4 gaat dieper in op de belangrijkste redenen waarom de herkenning van knaagdiergedrag in het algemeen geen 100% nauwkeurigheid bereikt en slecht presteert voor bepaalde gedragingen. Er worden drie aspecten onderscheiden in de dynamiek van gedrag die moeilijk te automatiseren zijn. Deze aspecten worden geïsoleerd in een kunstmatige dataset en met de kunstmatige data worden de resultaten gereproduceerd, met de

modernste gedragsherkenningsmodellen. Deze nieuwere modellen maken gebruik van self-supervised learning om eerst een lager-dimensionale representatie van de data te genereren voordat ze worden geclassificeerd.

Het laatste onderzoekshoofdstuk, **Hoofdstuk 5**, gaat dieper in op de praktische oplossingen en tools die gedragsonderzoekers kunnen helpen om nieuw gedrag te annoteren waarvoor voorheen geen getrainde classifier beschikbaar is. Bij bepaalde onderzoeksexperimenten is flexibelere classificatie nodig, bijvoorbeeld in geval van nieuwe opstellingen of nieuwe gedragseffecten. Het eerste deel van het hoofdstuk, **Hoofdstuk 5a** bewijst de robuustheid en generieke toepasbaarheid van de ABR-features door ze te gebruiken om 'scratch'-gedrag van muizen te classificeren in twee verschillende datasets, uit high-speed video-opnames. In het tweede deel, **Hoofdstuk 5b**, wordt de mogelijkheid onderzocht om handmatige annotatie te combineren met AI-assistentie tot een hybride active learning oplossing, zodanig dat het annotatieproces enerzijds efficiënter is dan volledig handmatige annotatie en anderzijds dat het eindresultaat nauwkeuriger is dan volledig geautomatiseerde annotatie. Het voordeel van active learning wordt gepresenteerd aan de hand van de gedragingen 'gestrekte attentie' en 'niet-ondersteund oprichten'.

Het proefschrift wordt afgesloten met een discussiehoofdstuk **Hoofdstuk 6**, waarin de contributies per onderzoekshoofdstuk worden belicht en waarin de tekortkomingen en toekomstige oplossingen worden uitgelegd. Het geeft bovendien een kijk op nieuwe manieren van gedragsanalyse die profiteren van volledig unsupervised detectie van gedragseffecten die gedragsonderzoekers in staat stelt hun gegevens en de gedragseffecten tussen experimentgroepen interactief te onderzoeken. Het discussiehoofdstuk eindigt met een reflectie op de ethische implicaties van onderzoek naar de herkenning van knaagdiergedrag.

Concluderend kunnen we stellen dat er veel onderzoekstoepassingen zijn waarbij met spanning wordt uitgekeken naar robuuste en flexibele geautomatiseerde annotatie van gedrag. Het werk in dit proefschrift komt tegemoet aan deze behoefte met een automatisch systeem en geeft aanwijzingen voor de verdere ontwikkeling van gedragsherkenning. De spannende ontwikkelingen op het gebied van artificial intelligence-algoritmen, hardware en uitrolmogelijkheden zullen in de nabije toekomst zeker meer en flexibelere automatisering mogelijk maken.

Acknowledgements

First of all, I sincerely thank my supervisor Professor Marcel van Gerven, for his efficient and effective guidance, his valuable advice and his constant support throughout this project. I am also very grateful to Professor Lucas Noldus for his encouraging enthusiasm, advice and thorough reviews. I thank him again in his capacity as my employer for more than 20 years at Noldus Information Technology, for allowing me to carry out this PhD program partly during working hours, and for his future-oriented, open attitude and continued commitment to innovation.

I would like to thank the colleagues from Noldus Information Technology who formed the basis of this project: Ruud Tegelenbosch for setting the requirements for automated rodent behavior recognition systems and for initiating the contact with Janssen for the collection of the data, and Andrew Spink for organizing the grants that helped funding this research. I furthermore want to thank Professor Sabine Hunnius for her enthusiasm and for her help in writing the project proposal for this PhD project in 2017. The research was mostly financed by Noldus Information Technology and partly by a grant from Agentschap NL (NeuroBasic-Pharma-Phenomics) and ICTRegie (SenseWell). The publication on the SmartAnnotator tool was part of the project Dutch Brain Interface Initiative (DBI^2) with project number 024.005.022 of the research program Gravitation, which is (partly) funded by the Dutch Research Council (NWO).

I want to thank Leen Raeymakers (Janssen Research and Development, Belgium) for recording and annotating the ABR dataset, and I am grateful for the biotechnical assistance of Niek van Stipdonk (Delta Phenomics, The Netherlands) in recording the ABR validation dataset and the stretched-attend videos used in Chapter 5.2. Finally, I thank Markus Koester of the open science portal (opnME) of Boehringer Ingelheim (BI), Germany, and Heike Schauerte of the GenPharm Team (BI) for providing the data for the unsupervised approach that is discussed in Chapter 6.2.

I thank my former colleague Roos Ottenhoff for her many useful ideas and discussions on ABR feature generation. I also thank Pavel Paclik (PR Sys Design, Delft, The Netherlands, `www.perclass.com`) and David Tax (Pattern Recognition Laboratory(PRL), Delft University of Technology, The Netherlands) for their help with ABR classification. I thank Marco Loog (PRL) for his help in writing the ABR publication.

I would like to thank my co-authors and colleagues Timon, Loes, Marco and Malte, and co-authors Johanneke, Berry and Cajo for their valuable contributions and constructive feedback. I thank all fellow PhD students of the Artificial Cognitive Systems group at the Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, especially Gabi, Josh, Luca and Gianluigi, for their support and advice.

Additionally, I want to thank colleagues and former colleagues from Noldus Information Technology for their support and interest: Albert, Annabel, Arjan, Arthur, Ben, Egon, Federico, Georgios, Guus, Harold, Jeroen, Kevin, Leon, Martin, Melita, Mike, Olga, Reinko, Richard, Robert, Rogier, Romain, Sanne, Simone, Veli, Wibo and all others.

Lastly, I would like to thank friends an family. I thank my dear friends Jeanette who talked me in to it, and Gwen who talked me through it, and my dearest friend Merel with whom I have been sharing everything important and not so important since I was

15 years old. I thank long-term friends Liesbeth and Sandra as well as my family who have supported me by both asking and not asking: my siblings Joosje, Sander and Jaap and my mother Heleen, from whom I inherited the initial self-confidence that everything is achievable. I deeply thank our children Janna, Marit and Roemer for their support and patience. I aimed to set a good example, although they might now reconsider a few times more before pursuing their own PhD. I hope they too will be given the opportunity in life to pursue what they want to contribute. And finally, for being there always, for his never-ending support, warm care, belief in me, putting-things-into-perspective jokes, love of my life Frank. I am glad to be back.

Research data management

Data Management

Research Data Management was performed according to the FAIR principles, ensuring that the data is Findable, Accessible, Interoperable, and Reusable. However, the data is not readily available because they are proprietary to either Noldus Information Technology or Boehringer Ingelheim. The data and documentation necessary to reproduce the results are available upon request and after permission from Noldus Information Technology, with restriction to academic use. The artificial data that was used in Chapter 4 is publicly available at `https://github.com/ElsbethvanDam/artificial_behavior_data`. All data remain available for at least ten years after termination of the studies. Requests to access the proprietary datasets should be directed to info@noldus.com.

All research data has been structured in a standardized way that is described in accompanying text files. The documentation includes specifications on:

- experimental setup
- data variables
- formatting of the raw data
- providing of analysis scripts or pipelines
- specification of version numbers for the software used

Ethical Approval

The rodent datasets that are used in this thesis are listed in Table 6.1. This PhD research followed the applicable laws and ethical guidelines. The ABR videos where recorded in 2009 at Janssen Pharmaceutical Research & Development in Beerse, Belgium according to the local regulations. The ABR validation experiments were performed in 2011 by Delta Phenomics in Utrecht, Netherlands, in adherence to the legal requirements of Dutch legislation on laboratory animals (Wod/Dutch 'Experiments on Animals Act') and were approved by an Animal Ethics Committee ('Lely-DEC'). The data for the unsupervised approach mentioned in the future outlook of the general discussion was derived based on a crowd sourcing call via Boehringer Ingelheim's open innovation portal, opnMe.com.

Table 6.1: Datasets used

Chapter	Dataset	Recorded at	Owned by
2-5	RBR dataset	Janssen Pharmaceutical R&D	Noldus IT
2	Validation dataset	Delta Phenomics	Noldus IT
5.1	Scratching dataset	Two customers of Noldus IT	Noldus IT
5.2	Stretched Attend dataset	Utrecht University	Noldus IT
6	OpnME dataset	Boehringer Ingelheim	Boehringer Ingelheim

For these videos, the maintenance and handling of animals were carried out in compliance with (i) the ethical guidelines established by German National Animal Welfare Laws within the framework of the European Union Directive 2010/63/EU and (ii) the Guide for the Care and Use of Laboratory Animals produced by the National Research Council and the Association for Assessment and Accreditation of Laboratory Animal Care International (AAALAC). The study protocol was approved by the responsible German authority (Regierungspräsidium Tübingen).

Curriculum vitae

Elisabeth (Elsbeth) Aafje van Dam was born in Doetinchem, The Netherlands, on the 15th of June, 1969. After graduation from Ludger College in Doetinchem in 1987, she started her educational career by attending the Dutch Circus School in Arnhem where she learned to juggle, balance the rope and view the world from an upside down perspective. In 1988, she started to study Mathematics at the University of Utrecht and after a year combined this with studying Cognitive Artificial Intelligence. After moving to Nijmegen in 1990, she continued her studies at the Radboud University, combining Mathematics with Philosophy until 1992. She continued with a Free Doctoral program consisting of courses in Philosophy of Language, Philosophy of Science, Philosophy of Mind, Cognitive Science, Functional Psychology and Programming. In 1994, she received her Master degree with a thesis entitled *The world and I: an attempt to integrate the subjectivity of experience into an objective description of reality.*

Her working career started in 1996 at Human Inference BV in Arnhem, a company specialized in name and address matching in large customer databases using what was then called 'fuzzy logic'. The first year she worked as a helpdesk employee and test engineer, after that as a software engineer and finally as a lead developer in 2003. In 2004, she decided to pursue a job with more societal impact, aimed towards academic customers and she joined Noldus Information Technology as a System Engineer working on EthoVision, the company's flagship for the acquisition and analysis of behavior data from video footage. Soon, she picked up the work on efficient high-throughput analysis software in the NeuroBasic project. In 2007, the opportunity arose to become a Research Engineer, working on algorithm development for rodent behavior recognition. She specialized in Computer Vision and Machine Learning techniques that are robust and fast enough to work with real life data in daily practice. In this capacity, she coached many visiting students at bachelor, master and PhD level and joined several collaborative research projects to work and give advice on behavior recognition, in various species (rodents, human adults, toddlers and babies, livestock) and with various sensors (video, wearables, audio). In 2018, she enrolled at Radboud University as an external PhD student herself. Not long after, in 2019, she became AI team lead at Noldus IT and responsible for the development of the new generation of behavior recognition tools based on the latest advances in AI.

The common thread throughout her life has been to understand how living beings with brains manage to perceive, interpret and respond to the world around them. Her driving force is the realization that semantic perception, including the idea of the self, emerges exclusively from the interaction of a brain with a complex world, and hence that we humans are not god-created special beings, but that we are part of the world around us and are as precious as everything else in nature.

Elsbeth is married to Frank Derks. They have three wonderful children: Janna (born in 1996), Marit (born in 1997) and Roemer (born in 2005). In her spare time, she likes to visit family and friends, camp, cycle, hike, read, play the accordion and grow vegetables.

List of publications

1. Arroyo-Araujo, M., Graf, R., Maco, M., Dam, E. A. van, Schenker, E., Drinkenburg, W., Koopmans, B., Boer, S. F. de, Cullum-Doyle, M., Noldus, L. P. J. J., Loos, M., Dommelen, W. van, Spooren, W., Biemans, B., Buhl, D. L., and Kas, M. J. (2019). Reproducibility via coordinated standardization: a multi-center study in a Shank2 genetic rat model for autism spectrum disorders. *Scientific Reports*, 9(1):11602.

2. Brouwer, A.-M., Dam, E. A. van, Erp, J. B. F. van, Spangler, D. P., and Brooks, J. R. (2018). Improving real-life estimates of emotion based on heart rate: a perspective on taking metabolic heart rate into account. *Frontiers in human neuroscience*, 12:284.

3. Brouwer, A.-M., Hogervorst, M. A., Erp, J. B. F. van, Grootjen, M., Dam, E. A. van, and Zandstra, E. H. (2019). Measuring cooking experience implicitly and explicitly: Physiology, facial expression and subjective ratings. *Food Quality and Preference*, 78:103726.

4. Dam, E. A. van, Harst, J. E. van der, Braak, C. J. F. ter, Tegelenbosch, R. A. J., Spruijt, B. M., and Noldus, L. P. J. J. (2013). An automated system for the recognition of various specific rat behaviours. *Journal of Neuroscience Methods*, 218(2):214–224.

5. Dam, E. A. van, Lorbach, M., and Tegelenbosch, R. A. J. (2018). Fast development of customized rodent behavior recognition with semi-automated labeling. In *Proceedings of Measuring Behavior 2018*, page 2.

6. Dam, E. A. van, Noldus, L. P. J. J., and Gerven, M. A. J. van. (2020). Deep learning improves automated rodent behavior recognition within a specific experimental setup. *Journal of Neuroscience Methods*, 332:108536.

7. Dam, E. A. van, Noldus, L. P. J. J., and Gerven, M. A. J. van. (2021). Deep learning systems for automated rodent behavior recognition systems suffer from observer bias: Time to raise the bars. In *Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods*.

8. Dam, E. A. van, Roosken, M. H., and Noldus, L. P. J. J. (2022). Robust scratching behavior detection in mice from generic features and a lightweight neural network in 100 fps videos. In *Volume 2 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods*, pages 301–305.

9. Dam, E. A. van, Noldus, L. P. J. J., and Gerven, M. A. J. van (2023). Disentangling rodent behaviors to improve automated behavior recognition. *Frontiers in Neuroscience*, 17:1198209.

10. Dam, E. A. van, Daniels, T., Ottink, L., Gerven, M. A. J. van, and Noldus, L. P. J. J. (2024). Fast annotation of rodent behaviors with AI assistance: Human observer and SmartAnnotator collaborate through active learning. *Measuring Behavior 2024*.

11. Grieco, F., Bernstein, B. J., Biemans, B., Bikovski, L., Burnett, C. J., Cushman, J. D., Dam, E. A. van, Fry, S. A., Richmond-Hacham, B., Homberg, J. R., Kas, M. J. H., Kessels, H. W., Koopmans, B., Krashes, M. J., Krishnan, V., Logan, S., Loos, M., McCann, K. E., Parduzi, Q., Pick, C. G., Prevot, T. D., Riedel, G., Robinson, L., Sadighi, M., Smit, A. B., Sonntag, W., Roelofs, R. F., Tegelenbosch, R. A. J., and Noldus, L. P. J. J. (2021). Measuring behavior in the home cage: Study design, applications, challenges, and perspectives. *Frontiers in Behavioral Neuroscience*, 15:735387.

12. Kapidis, G., Dam, E. A. van, Poppe, R., Noldus, L. P. J. J., Veltkamp, R. C., Grant, R., Allen, T., Spink, A., Sullivan, M., et al. (2018a). Action detection from egocentric videos in daily living scenarios. In *Measuring Behavior 2018*, pages 405–407.

13. Kapidis, G., Poppe, R. W., Dam, E. A. van, Veltkamp, R. C., and Noldus, L. P. J. J. (2018b). Where am I? comparing CNN and LSTM for location classification in egocentric videos. In *2018 IEEE international conference on pervasive computing and communications workshops (PerCom Workshops)*, pages 878–883.

14. Kapidis, G., Poppe, R., Dam, E. A. van, Noldus, L. P. J. J., and Veltkamp, R. (2019a). Egocentric hand track and object-based human action recognition. In *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld / SCALCOM / UIC / ATC / CBDCom / IOP / SCI)*, pages 922–929.

15. Kapidis, G., Poppe, R., Dam, E. A. van, Noldus, L. P. J. J., and Veltkamp, R. (2019b). Multitask learning to improve egocentric action recognition. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4396–4405.

16. Kapidis, G., Poppe, R., Dam, E. A. van, Noldus, L. P. J. J., and Veltkamp, R. C. (2020). Object detection-based location and activity classification from egocentric videos: A systematic analysis. In Chen, F., Garc´ıa-Betances, R. I., Chen, L., Cabrera-Umpi´errez, M. F., and Nugent, C., editors, *Smart Assisted Living: Toward An Open Smart-Home Infrastructure, Computer Communications and Networks*, pages 119–145. Springer International Publishing, Cham.

17. Lorbach, M., Poppe, R., Dam, E. A. van, Noldus, L. P. J. J., and Veltkamp, R. C. (2015). Automated recognition of social behavior in rats: The role of feature quality. In Murino, V. and Puppo, E., editors, *Image Analysis and Processing — ICIAP 2015*, pages 565–574, Cham. Springer International Publishing.

18. Lorbach, M., Poppe, R., and Dam, E. A. van (2016a). Transfer learning for rodent behavior recognition. In *Proceedings of Measuring Behavior 2016*, pages pp. 461–489.

19. Lorbach, M., Poppe, R., Dam, E. A. van, Noldus, L. P. J. J., and Veltkamp, R. C. (2016b). Clustering-based active learning in unbalanced rodent behavior data. In *Proceedings of the International Workshop on Visual observation and analysis of Vertebrate And Insect Behavior (VAIB)*.

20. Lorbach, M., Kyriakou, E. I., Poppe, R., Dam, E. A. van, Noldus, L. P. J. J., and Veltkamp, R. C. (2018). Learning to recognize rat social behavior: Novel dataset and cross-dataset application. *Journal of Neuroscience Methods*, 300:166–172.

21. Palmero, C., Dam, E. A. van, Escalera, S., Kelia, M., Lichtert, G. F., Noldus, L. P. J. J., Spink, A. J. and Wieringen, A. van (2018). Automatic mutual gaze detection in face-to-face dyadic interaction videos. In *Proceedings of Measuring Behavior 2018*

Donders Graduate School

For a successful research Institute, it is vital to train the next generation of scientists. To achieve this goal, the Donders Institute for Brain, Cognition and Behaviour established the Donders Graduate School in 2009. The mission of the Donders Graduate School is to guide our graduates to become skilled academics who are equipped for a wide range of professions. To achieve this, we do our utmost to ensure that our PhD candidates receive support and supervision of the highest quality.

Since 2009, the Donders Graduate School has grown into a vibrant community of highly talented national and international PhD candidates, with over 500 PhD candidates enrolled. Their backgrounds cover a wide range of disciplines, from physics to psychology, medicine to psycholinguistics, and biology to artificial intelligence. Similarly, their interdisciplinary research covers genetic, molecular, and cellular processes at one end and computational, system-level neuroscience with cognitive and behavioural analysis at the other end. We ask all PhD candidates within the Donders Graduate School to publish their PhD thesis in de Donders Thesis Series. This series currently includes over 700 PhD theses from our PhD graduates and thereby provides a comprehensive overview of the diverse types of research performed at the Donders Institute. A complete overview of the Donders Thesis Series can be found on our website: `https://www.ru.nl/donders/donders-series`

The Donders Graduate School tracks the careers of our PhD graduates carefully. In general, the PhD graduates end up at high-quality positions in different sectors, for a complete overview see `https://www.ru.nl/donders/destination-our-former-phd`. A large proportion of our PhD alumni continue in academia (¿50%). Most of them first work as a postdoc before growing into more senior research positions. They work at top institutes worldwide, such as University of Oxford, University of Cambridge, Stanford University, Princeton University, UCL London, MPI Leipzig, Karolinska Institute, UC Berkeley, EPFL Lausanne, and many others. In addition, a large group of PhD graduates continue in clinical positions, sometimes combining it with academic research. Clinical positions can be divided into medical doctors, for instance, in genetics, geriatrics, psychiatry, or neurology, and in psychologists, for instance as healthcare psychologist, clinical neuropsychologist, or clinical psychologist. Furthermore, there are PhD graduates who continue to work as researchers outside academia, for instance at non-profit or government organizations, or in pharmaceutical companies. There are also PhD graduates who work in education, such as teachers in high school, or as lecturers in higher education. Others continue in a wide range of positions, such as policy advisors, project managers, consultants, data scientists, web- or software developers, business owners, regulatory affairs specialists, engineers, managers, or IT architects. As such, the career paths of Donders PhD graduates span a broad range of sectors and professions, but the common factor is that they almost all have become successful professionals. For more information on the Donders Graduate School, as well as past and upcoming defences please visit: `http://www.ru.nl/donders/graduate-school/phd/`

DONDERS INSTITUTE

Radboud University

Radboudumc