Journal of Neuroscience Methods xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Learning to recognize rat social behavior: Novel dataset and cross-dataset application

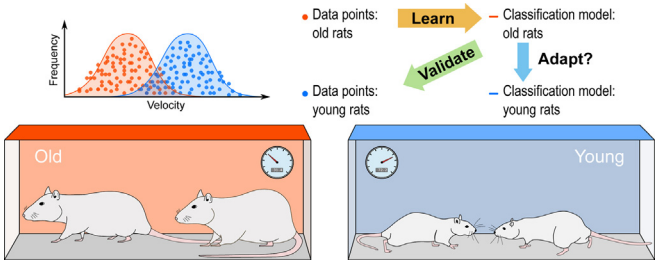
Malte Lorbach[a,c,*], Elisavet I. Kyriakou[b,c], Ronald Poppe[a], Elsbeth A. van Dam[c], Lucas P.J.J. Noldus[c], Remco C. Veltkamp[a]

[a] Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands
[b] Department of Cognitive Neuroscience, Radboud University Medical Centre, Kapittelweg 29, 6525 EN Nijmegen, The Netherlands
[c] Noldus Information Technology BV, Nieuwe Kanaal 5, 6709 PA Wageningen, The Netherlands

HIGHLIGHTS

- Using video analysis to measure rodent social behavior receives growing attention.
- Developing and validating automated measuring methods requires annotated datasets.
- We introduce the first, publicly available rat social interaction dataset, RatSI.
- Cross-dataset validation of automated methods ensures validity in practice.
- Validity may be expanded by developing novel dataset adaptation techniques.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 31 January 2017
Received in revised form 4 May 2017
Accepted 5 May 2017
Available online xxx

Keywords:
Social behavior
Rodents
Automated behavior recognition
Dataset

ABSTRACT

Background: Social behavior is an important aspect of rodent models. Automated measuring tools that make use of video analysis and machine learning are an increasingly attractive alternative to manual annotation. Because machine learning-based methods need to be trained, it is important that they are validated using data from different experiment settings.
New method: To develop and validate automated measuring tools, there is a need for annotated rodent interaction datasets. Currently, the availability of such datasets is limited to two mouse datasets. We introduce the first, publicly available rat social interaction dataset, RatSI.
Results: We demonstrate the practical value of the novel dataset by using it as the training set for a rat interaction recognition method. We show that behavior variations induced by the experiment setting can lead to reduced performance, which illustrates the importance of cross-dataset validation. Consequently, we add a simple adaptation step to our method and improve the recognition performance.
Comparison with existing methods: Most existing methods are trained and evaluated in one experimental setting, which limits the predictive power of the evaluation to that particular setting. We demonstrate that cross-dataset experiments provide more insight in the performance of classifiers.
Conclusions: With our novel, public dataset we encourage the development and validation of automated recognition methods. We are convinced that cross-dataset validation enhances our understanding of

* Corresponding author.
E-mail address: m.t.lorbach@uu.nl (M. Lorbach).

http://dx.doi.org/10.1016/j.jneumeth.2017.05.006

Please cite this article in press as: Lorbach, M., et al., Learning to recognize rat social behavior: Novel dataset and cross-dataset application. J Neurosci Methods (2017), http://dx.doi.org/10.1016/j.jneumeth.2017.05.006

rodent interactions and facilitates the development of more sophisticated recognition methods. Combining them with adaptation techniques may enable us to apply automated recognition methods to a variety of animals and experiment settings.

1. Introduction

Social interaction is an important component of psychiatric research as well as neurological testing of animal models in behavioral neuroscience (Urbach et al., 2010). As part of the emotional screening of a model it relates to aspects such as anxiety, stress, play and sexual behavior (File and Seth, 2003). Moreover, abnormal social behavior can be indicative of a psychopathology (Peters et al., 2015) and can therefore inform us of the onset or progression of conditions such as schizophrenia (Wilson and Koenig, 2014), Huntington's (Urbach et al., 2014) and Alzheimer's disease (Lewejohann et al., 2009) as well as Rett syndrome (Veeraragavan et al., 2016). Including social behavior in rodent models therefore increases their predictive power and value for the transition to clinical trials and treatments for humans (Peters et al., 2015; Richardson, 2015).

Whether we seek to enhance our understanding of social behavior or include it in a rodent model, we need to objectively measure and quantify it. Traditionally, this involves annotating the interactions among rodents in hours of either live observations or video recordings of social interaction tests. While this can be done manually, it is time-consuming and subjective. Subjectivity may be reduced by a meticulously defined ethogram and thorough training of the human annotators at the cost of additional work.

An attractive alternative to manual scoring are automated measuring tools (Schaefer and Claridge-Chang, 2012; Steele et al., 2007; Egnor and Branson, 2016; Noldus et al., 2001). Such tools track the locations of the rodents in video recordings and provide quantitative measures such as the distance traveled and the time spent in proximity (Spruijt et al., 1992; Sams-Dodd, 1995; Dell et al., 2014). Recent advances in video analysis have made the tracking of rodents more robust and accurate (Hong et al., 2015; Pérez-Escudero et al., 2014). This allows us to take the next step and consider the automated recognition of specific *interactions* such as approaching and following. Although the interaction categories that can currently be handled automatically are not as fine-grained and large in quantity as the categories that humans are able to annotate, automated methods can still support manual annotation and reduce labor. For example, by providing a first segmentation into these broader categories with high accuracy, the human effort can be reduced to annotating fine-grained behaviors only in the relevant video segments instead of the full length of the video.

The automated recognition of interactions typically involves applying classification algorithms to a quantified representation (features) of the visual information in the video (Hong et al., 2015; Kabra et al., 2012; Burgos-Artizzu et al., 2012; Giancardo et al., 2013). The features are derived from the tracked animals and may include velocity and distance. In order to distinguish between the different interactions, the parameters in the classification algorithms are determined using labeled feature examples. In this training phase, the classifier learns the similarities among the examples and thereby creates a *model* of each interaction. For instance, it may learn that whenever a rat approaches another, it moves at a certain velocity while the distance between the two decreases. It is important how the classifier learns such models. A classifier that simply "remembers" the feature values will not perform well on unseen examples which have slightly different values. Instead, it must generalize from the empirical examples to the inherent variations of the interaction classes.

Generally, there are two types of variation in the examples of an interaction. First, two animals will perform the same interaction slightly differently every time, for instance, at a slightly different velocity or from a different starting point. We consider this the natural variation of an interaction. Second, there is a systematic bias in the natural variation that depends on the tested population and the environment in which the interactions are observed. Rats from the tested population, which is characterized by the genetic background, the age and possibly the progress of a condition or its treatment, could for example move slower than rats from another population. The environment, which is often created by the researcher to study specific behaviors, comprises factors such as the available space and the presence of hiding places or novel objects that may allow or prevent interactions to be performed in certain ways.

As a consequence, the models learned by the classifier depend on the distribution of training examples with respect to the systematic bias. If the bias changes due to modifications to the animal population or the environment (Schneider and Levine, 2014), the models could lose their effectiveness.

Therefore, when we evaluate the performance of a trained classifier, we typically use test examples that follow the same distribution as the training examples. Both training and test examples are usually taken from a dataset of video recordings of one specific experiment (Hong et al., 2015; Kabra et al., 2012; Burgos-Artizzu et al., 2012; Giancardo et al., 2013; Eyjolfsdottir et al., 2014; Kuehne et al., 2016). That ensures that the bias is kept constant during evaluation and that we obtain a plausible measure of the performance.

This evaluation scheme becomes critical when we apply the trained classifier in practice. Beyond the tested experiment setting, the evaluation is of limited value as it cannot predict the classifier's performance in another setting. Given the difficulty of precisely replicating experiment settings (Crabbe et al., 1999) as well as appeals to increase experiment heterogeneity (Richter et al., 2009), we argue for an evaluation of interaction classifiers across settings and therefore across datasets. Only with cross-dataset evaluation can we be confident about the performance of the classifier in practice (van Dam et al., 2013) and judge to which settings we can apply it without retraining.

We argue that there is a need for datasets for at least two purposes: to train classifiers and to evaluate them across experiment settings. Currently, there are only two rodent social behavior datasets publicly available for researchers and both focus on mice: the Caltech Resident-Intruder Mouse dataset (CRIM13) (Burgos-Artizzu et al., 2012) and the Mice Behavior Analysis dataset (MBADA) (Giancardo et al., 2013).

Given the increasing interest in rats for studying social behavior (Veeraragavan et al., 2016; Homberg et al., 2016), we introduce the first rat social interaction dataset (RatSI).[1] It contains 2.25 h of annotated video recordings of two interacting rats in an open-field arena, including accurate 3-point tracking of the animals. The

[1] http://www.noldus.com/innovationworks/phenorat-dataset.

Please cite this article in press as: Lorbach, M., et al., Learning to recognize rat social behavior: Novel dataset and cross-dataset application. J Neurosci Methods (2017), http://dx.doi.org/10.1016/j.jneumeth.2017.05.006

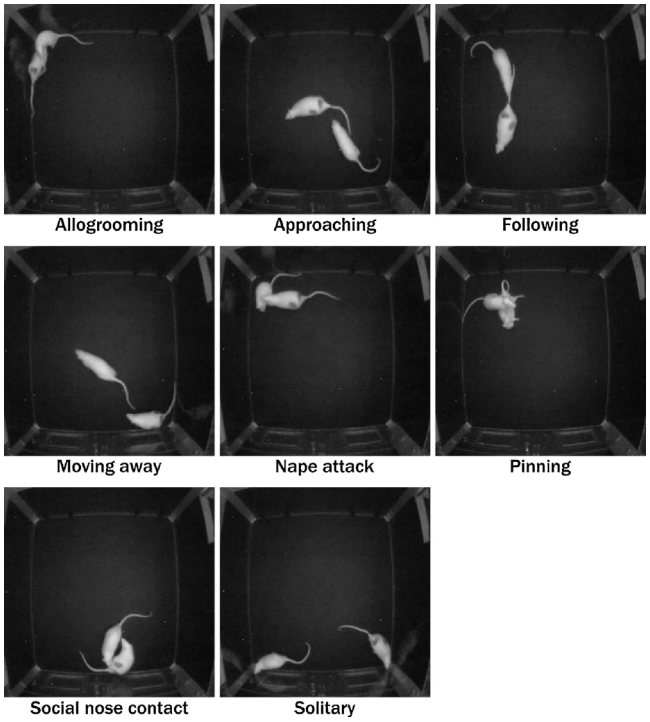


Fig. 1. Example frames of each behavior in RatSI dataset.

dataset can be used to develop novel interaction classifiers and to validate existing ones.

To demonstrate the practical value of the dataset, we use it to train a basic classifier for rat social behavior recognition. We then evaluate the trained classifier on another validation dataset. We also give an example of how a systematic bias can influence the classifier performance. Considering the animal age as the bias, we investigate how we can adapt the classifier so as to be applicable across datasets.

We continue the article with a description of the RatSI dataset. In Section 3 we introduce the recognition method. We present the evaluation results in Section 4 and conclude in Section 5.

2. Materials: RatSI dataset

We compiled the dataset from videos and behavior annotations of a study on a rat model for Spinocerebellar ataxia type 17 (SCA17) (Kelp et al., 2013; Kyriakou et al., 2016).

2.1. Video acquisition

The dataset comprises nine videos of a social interaction test in a controlled open-field environment with two rats. The videos are recorded from a top-view perspective in a 90×90 cm Noldus PhenoTyper® 9000 cage[2] with standard top unit (image resolution 704×576, 25 fps) without bedding and accessories. Each recording captures 15 min of interactions between different rat pairs. Fig. 1 shows examples of the captured interactions.

The recorded experiments are part of a larger social interaction study adopting the following protocol. Three days before the recordings, the rats were individually introduced to the cage arena for 20 min. Twenty-four hours before the test, the rats were isolated to stimulate a desire for social interaction. Each rat was then put in the recording cage together with another, unfamiliar rat. The recordings started with the introduction of the second animal.

2.2. Animals

Naive male rats, 9 months, of two genotypes were used: SCA17 (Kelp et al., 2013) ($n = 8$) and wild-type-like (Sprague Dawley, $n = 10$). Animals were housed in pairs under reversed day-light cycle conditions and water and food were available ad libitum. Subjects were housed in type IV cages according to EU welfare regulations except for the 24 h isolation period prior to social testing where the animals were housed in type III cages. Testing was performed during the animals' active (dark) phase. All experiments were performed after approval of the Ethical Committee for Animal Experiments of the Radboud University Nijmegen Medical Center for compliance to ethical standards and use of laboratory animals according to EU-guidelines.

2.3. Annotation of interactions

Every video frame was annotated by an expert with one of nine interaction labels (Peters et al., 2016), described in Table 1. The annotations are non-overlapping. Note that the interactions occur with very different frequencies which leads to a non-uniform distribution of the prior occurrence probabilities. In particular, the animals perform solitary behavior in the majority (58.6%) of the frames. Such a skewed distribution is common for behavioral datasets (Burgos-Artizzu et al., 2012; Giancardo et al., 2013).

The annotated interactions are related to either the trajectories of the animals such as *Approaching* and *Following*, or a contact category such as *Allogrooming* and *Nape attacking*. To distinguish between the fine-grained contact interactions automatically, we require additional information from features other than the animal trajectories (Lorbach et al., 2015), for example image features. What features are best suited for this task is yet an open research question (Robie et al., 2017). To facilitate such research we make the annotations of all interactions available online. Here we use a restricted set of annotations in which we have merged *Allogrooming*, *Nape attacking*, *Pinning* and *Social nose contact* into one common *Contact* class. The *Contact* class groups interactions that are not easily distinguished by the classifier on basis of only trajectory features. If a fine-grained categorization is required in the behavior analysis, the interactions classified as *Contact* can be annotated manually afterwards.

2.4. Tracking and features

The animal locations and body point positions have been tracked throughout the videos using Noldus EthoVision[3] XT 12 with a customized rat identification algorithm. The algorithm uses appearance differences (here reinforced by black markers) to distinguish and maintain the identities up to a few errors which we correct manually afterwards. Note that the identification algorithm is still under development to facilitate marker-less identification and is therefore not included in the official EthoVision XT 12 version. We track three points on the rat body: the nose, the center of body mass, and the tail-base (see Fig. 2a for an illustration). Compared to tracking only the center point, three-point tracking yields a more detailed pose representation and improves the recognition accuracy (Dell et al., 2014; Lorbach et al., 2015; Decker and Hamprecht, 2014).

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

[3] http://www.noldus.com/ethovision.

Please cite this article in press as: Lorbach, M., et al., Learning to recognize rat social behavior: Novel dataset and cross-dataset application. J Neurosci Methods (2017), http://dx.doi.org/10.1016/j.jneumeth.2017.05.006

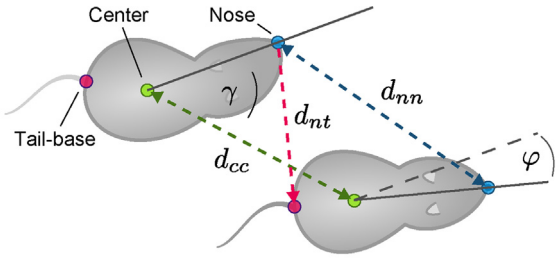
Table 1
Description of the behavior classes, their prior probability regarding the frame count p and the number of events m.

Allogrooming	Grooming another rat's fur	$p = 0.047$	$m = 105$
Approaching	Moving towards another rat in a straight line	$p = 0.075$	$m = 355$
Following	Chasing another, moving rat within a tail length distance	$p = 0.093$	$m = 259$
Moving away	Moving away from another rat in a straight line	$p = 0.044$	$m = 387$
Nape attacking	Snout or oral contact directed at neck region, possibly with biting/pulling fur in that region	$p = 0.01$	$m = 85$
Pinning	Actively restrain another rat on its back	$p = 0.006$	$m = 8$
Social nose contact	Non-incidental nose-body contact (e.g. inspection)	$p = 0.103$	$m = 506$
Solitary	Any activity not directed at another rat	$p = 0.586$	$m = 484$
Other	Any interaction not covered by another category	$p = 0.036$	$m = 196$

The feature set that we derive from each animal's trajectory is described in Fig. 2. The set is based on previous work in the field (Kabra et al., 2012; Burgos-Artizzu et al., 2012; Eyjolfsdottir et al., 2014). Static pose information is represented by the distances between the three body points (d_{cc}, d_{nn}, d_{nt}), the head orientation in relation to the other rat's position (γ), and the relative orientation of the pair (φ). Dynamic information is captured by two body point velocities (v_c, v_n) as well as the change of distance and orientation between consecutive video frames. For details on the features we refer to Appendix A.

In the considered interactions, the two rats often take on different roles. For example, one rat approaches while the other is *being approached*. This asymmetry is information that the classifier cannot use because the role is unknown beforehand and thus not encoded in the features. In fact, the order of the rats in the feature vector is arbitrary (i.e., first features of rat A, then of rat B or vice versa). To make the classifier invariant to the order, we aggregate the features across animals. We take the minimum, the maximum and the absolute difference of all features except those that are already invariant to the order (center and nose point distances and the relative orientation). The final feature vector of one frame has 24 elements.

As the final step we reduce feature noise that may have been introduced during the tracking and propagated through above computations. We smooth the sequence of feature values over time using a moving average over five surrounding frames (two before and two after).



(a)

	d/dt	Unify	Description
d_{cc}	x	-	Distance between center points
d_{nn}	x	-	Distance between nose points
d_{nt}	x	x	Distance between nose and tail
v_c	-	x	Center point velocity
v_n	-	x	Nose point velocity
$\cos(\gamma)$	x	x	Relative position
$\lvert\varphi\rvert$	x	-	Relative orientation

(b)

Fig. 2. Features extracted from tracked body points. Asymmetric features are unified to one common value per rat pair.

3. Method: rat interaction recognition

We now turn to the recognition method and its evaluation in a cross-dataset classification task. Our interaction classifier models interactions as Gaussian distributions of the features. To be able to capture more interaction variations, we allow the classifier to model each interaction using multiple Gaussian distributions. The distributions are combined in Gaussian Mixture Models (GMM).

During the training phase, the classifier determines the parameters of the models using the Expectation Maximization algorithm (Dempster et al., 1977). This yields a set of n model parameters $\{\theta_1, \ldots, \theta_n\}$ per interaction class. In addition to the parameters of the Gaussian distributions, we need to determine the number of distributions in each mixture model and an optional constraint that constrains the covariance matrices of the Gaussians to be diagonal. The latter simplifies the models and decreases the time needed for training. We find the settings that yield the highest accuracy automatically using cross-validation.

To predict the discrete interaction label $\hat{y}$ of a feature vector x extracted from an unseen video frame, the classifier computes the probability of the data point given the model parameters for each class, θ_i, and returns the class with the highest probability:

$$\hat{y} = arg\,\max_i p\left(x|\theta_i\right). \quad (1)$$

Note that we intentionally neglect the information of how often a particular interaction has occurred during training (the *prior* probability) to prevent biased predictions in test sets with different interaction occurrence ratios.

3.1. Validation dataset

The evaluation of our recognition method is performed on another Validation dataset. The Validation set is similar to RatSI as it also contains videos from an open-field social interaction test and the same interactions are annotated by an expert (Peters et al., 2016). The experiments however were performed in a different laboratory. The rats are also younger (5 weeks instead of 9 months) and thus smaller, quicker and they engage frequently in dynamic playing interactions. The Validation set contains 400 annotated segments from five videos (50 per interaction class) with a total duration of 12.5 min. The interactions occur with different frequencies and durations than in RatSI. The locations of the rats were tracked with Noldus EthoVision XT 11. Tracking and identity errors were corrected manually.

3.1.1. Animals

One group of ten naive wild-type-like (Sprague Dawley) males, 5 weeks, were used in an social interaction test with the same protocol as described in Section 2.1. The experiments were performed in adherence to the legal requirements of Dutch legislation on laboratory animals (WOD/Dutch "Experiments on Animals Act") and were reviewed and approved by an Animal Ethics Committee ("Lely-DEC").

Please cite this article in press as: Lorbach, M., et al., Learning to recognize rat social behavior: Novel dataset and cross-dataset application. J Neurosci Methods (2017), http://dx.doi.org/10.1016/j.jneumeth.2017.05.006

3.2. Experiments

We perform three experiments to evaluate three aspects of our interaction classifier. First, we assess whether the classifier is able to recognize interactions in the same experiment setting as it has been trained on (*Within-data*). Second, we assess whether the classifier generalizes to other settings by evaluating its performance on the Validation dataset (*Cross-dataset*). Third, we examine whether we can neutralize the differences between the two experiment settings by adapting the distribution of the feature values (*Adaptation*). We use the restricted annotation set for our experiments.

3.2.1. Within-dataset

The within-dataset evaluation is performed in a 3-fold cross-validation scheme. That is, we split the dataset into three parts (three videos each) and then train the classifier on two parts and measure its performance on the remaining part. This is repeated such that we evaluate the performance on all three parts once. We automatically determine the best classifier settings by performing a cross-validated model selection on the two training parts (with four training videos, two test videos and three repetitions).

3.2.2. Cross-dataset

For the cross-dataset validation, we determine the GMM settings and train the classifier using the same 3-fold cross-validated model selection scheme. Since the performance is now evaluated on the Validation dataset, we use all RatSI videos for training.

3.2.3. Adaptation

To examine whether some of the differences in the experiment settings can be neutralized, we aim to remove the systematic bias (as introduced in Section 1) from the feature values.

We employ a simple technique that scales the values of each feature such that the fifth-percentile value is −1 and the 95th-percentile value is 1. Using the percentiles instead of the minimum and maximum values increases the tolerance against outliers and skewed class priors. After independently scaling the training and Validation sets, we repeat the cross-dataset experiment.

To illustrate how training sets with different properties (e.g. experiment setting, number of examples) can affect the performance, we repeat all three experiments in reverse order, i.e., using the Validation set for training, and RatSI for validation.

3.2.4. Performance metric

The performance is measured per class by the F1-score. The F1-score is the harmonic mean of the precision (true positive predictions divided by total number of positive predictions) and recall scores (true positive predictions divided by the number of actual occurrences). The class scores range from 0, with no correct predictions, to 1 for the correct prediction of all examples. To obtain a single measure of performance for the classifier, we average the F1-scores over all interaction classes leading to a final score in the range from 0 to 1. Averaging over classes as opposed to the total number of frames (equivalent to the ratio of correct frames) assigns equal importance to all interaction classes and prevents the score from being biased by the most-occurring interactions. Hence it is better suited for behavior datasets with interactions that occur with different frequencies.

4. Results

We report the performance of our interaction recognition method in Fig. 3. In the within-dataset experiment, we achieve a F1-score of 0.52 (±0.03) on RatSI and 0.68 (±0.06) on Validation. When trained on RatSI and evaluated on Validation, the level of accuracy is maintained (0.69). After adapting the features, the score even slightly improves to 0.72.

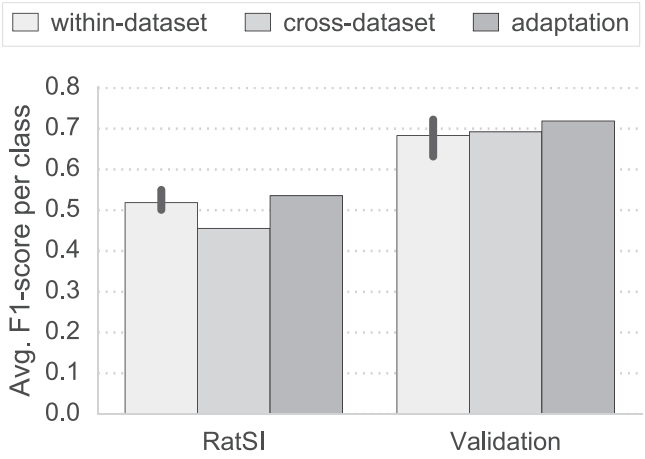


Fig. 3. Recognition performance (average F1-score) with s.e. for cross-validated *within-dataset* experiment.

In reversed training direction (Validation → RatSI), the F1-score of 0.52 drops by 11.5% to 0.46 in the cross-dataset experiment. The drop is compensated fully by applying the feature adaptation (0.54).

The results show that RatSI is a suitable dataset for training social interaction classifiers. The score achieved on the Validation set (0.69) is in the same order as reported in related work on similar datasets (Burgos-Artizzu et al., 2012; Giancardo et al., 2013; Eyjolfsdottir et al., 2014). Note the relatively low performance for *Moving away* of 0.26 (see Table 2) which is partly caused by confusions with the *Solitary* class. These occur because incidental movements away from another animal are typically classified as *Moving away*, whereas the human annotator only decided for *Moving away* if the event succeeded another interaction such as *Contact*. Our frame-based classifier does not take such context information into account yet.

While the classifier trained on RatSI generalizes well to Validation, training on the Validation dataset is not optimal as is evident from the declined performance on RatSI. This illustrates the necessity to validate classifiers on other datasets. The decline in accuracy is presumably caused by the limited size of the Validation set (12.5 min compared to 135 min in RatSI). It further contains interaction variations that are more specific to young rats such as *Following* at high velocity. The high velocity does not translate well to the older, slower rats in RatSI, leading to a biased classifier and consequently to a decreased accuracy for *Following*: from 0.53 to 0.25.

A simple feature adaptation technique however is able to compensate for this age difference and restores the accuracy to the level of the classifier trained on the same dataset. This is a promising result as it demonstrates that classifiers are not necessarily bound to one experiment setting. With more elaborate techniques we may

Table 2
Per interaction recognition performance for within-dataset (w), cross-dataset (c) and adaptation (a) experiments

Class	RatSI			Validation		
	w	*c*	*a*	*w*	*c*	*a*
Approaching	0.43	0.35	0.41	0.61	0.59	0.62
Contact	0.58	0.57	0.65	0.94	0.95	0.96
Following	0.53	0.25	0.51	0.58	0.66	0.66
Moving away	0.26	0.24	0.24	0.44	0.49	0.53
Solitary	0.80	0.87	0.86	0.84	0.77	0.84

Please cite this article in press as: Lorbach, M., et al., Learning to recognize rat social behavior: Novel dataset and cross-dataset application. J Neurosci Methods (2017), http://dx.doi.org/10.1016/j.jneumeth.2017.05.006

be able to handle more pronounced variations such as different species.

5. Conclusion

We introduced the first publicly available rat social interaction dataset, RatSI. The dataset is suitable for training rat interaction recognition methods as well as for validating methods trained on other datasets. The dataset can be used to study the temporal aspects of rat interactions and how these may improve the recognition performance. We encourage the development of new automated methods and the use of the presented method for comparison.

We further illustrated the importance of cross-dataset evaluations considering the different experiment settings encountered in practice. We showed that behavior variations induced by the experiment setting, for example the animal age and its effect on the velocity, can lead to reduced performance.

Through the performed cross-dataset evaluation, we were able to identify and neutralize the behavioral variation from our validation dataset, and could thus improve the classification performance. The fact that we were able to achieve this improvement with a simple scaling technique demonstrates the potential of cross-dataset application of interaction classifiers.

Developing more sophisticated methods for adapting to behavior variations will not only enhance our understanding of rodent interactions, it could also enable us to apply automated measuring tools across species and to longitudinal studies of diseases.

Funding

This work was supported by the EC FP7 Marie Curie ITN PhenoRat [GA no. 317259].

Acknowledgements

We thank Suzanne Peters for providing the data of the Validation set.

Appendix A. Trajectory features

The features that we introduced in Section 2.4 are derived from the tracked body point locations over time. Each feature is computed for every frame of a given video. We enumerate the animals and indicate the identity in a subscript together with specific body point (c for center point, n for nose point, t for tail-base point). For example, the center point of rat 1 measured in frame t is $\vec{p}_{1,c}(t)$. For the sake of clarity we omit the frame identifier (t) unless it is necessary to distinguish between values of different frames.

A.1 Distance

We measure three distances between the two animals, namely between the center points, between the nose points, and between the nose point and the tail-base point. All distances are Euclidean distances, indicated by $||\cdot||^2$.

$$d_{cc} = \left|\left|\vec{p}_{1,c} - \vec{p}_{2,c}\right|\right|^2 \tag{A.1}$$

$$d_{nn} = \left|\left|\vec{p}_{1,n} - \vec{p}_{2,n}\right|\right|^2 \tag{A.2}$$

$$d_{nt} = \left|\left|\vec{p}_{1,n} - \vec{p}_{2,t}\right|\right|^2 \tag{A.3}$$

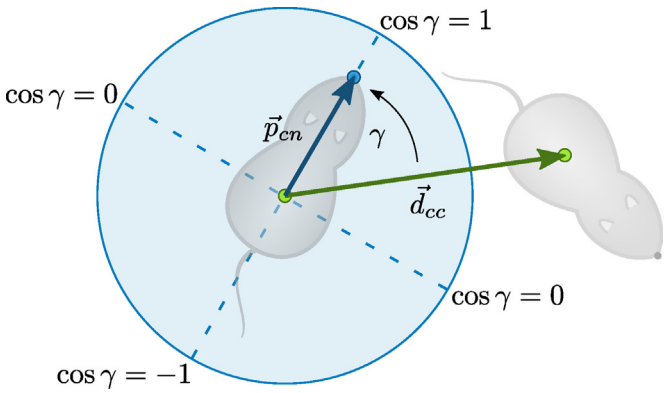


Fig. A4. The relative position of one rat with respect to the head orientation of the other.

A.2 Velocity

The velocities of the center and the nose points are estimated by the positional difference between two consecutive frames. To standardize velocity across different video frame rates, we divide by the time interval covered by the two frames: $\delta = 1/\text{fps}$, where fps is the video frame rate:

$$v_c(t) = \left|\left|\vec{p}_c(t) - \vec{p}_c(t-1)\right|\right|^2/\delta \tag{A.4}$$

$$v_n(t) = \left|\left|\vec{p}_n(t) - \vec{p}_n(t-1)\right|\right|^2/\delta. \tag{A.5}$$

A.3 Relative orientation

We measure the relative orientation between the rats as the angle between their head directions. The head vector of rat j is $\vec{p}_{j,cn}$, $j \in \{1, 2\}$, pointing from the center point to the nose point. The relative orientation is the absolute angle between the head vectors of the two rats:

$$\varphi = \left|\angle\left(\vec{p}_{1,cn}, \vec{p}_{2,cn}\right)\right|. \tag{A.6}$$

A.4 Relative position

The relative position captures where in an animal's environment the other animal is (e.g., in front, behind, next to). We designed this feature to be invariant to the distance between the animals and to be symmetric with respect to the side (left/right). It is calculated as $\cos\gamma$, where γ is the angle between the animal's head vector $\vec{p}_{j,cn}$ and the line connecting both animals' center points $\vec{d}_{cc}$ as illustrated in Fig. A4.

References

Burgos-Artizzu, X.P., Dollár, P., Lin, D., Anderson, D.J., Perona, P., 2012. Social behavior recognition in continuous video. Proc. Conf. Comput. Vis. Pattern Recognit., 1322–1329, http://dx.doi.org/10.1109/CVPR.2012.6247817.

Crabbe, J.C., Wahlsten, D., Dudek, B.C., 1999. Genetics of mouse behavior: interactions with laboratory environment. Science 284 (5420), 1670–1672, http://dx.doi.org/10.1126/science.284.5420.1670.

Decker, C., Hamprecht, F.A., 2014. Detecting individual body parts improves mouse behavior classification. In: Workshop on Visual Observation and Analysis of Vertebrate and Insect Behavior (VAIB), URL http://homepages.inf.ed.ac.uk/rbf/VAIB14PAPERS/decker.pdf.

Dell, A.I., Bender, J.A., Branson, K., Couzin, I.D., de Polavieja, G.G., Noldus, L.P.J.J., Pérez-Escudero, A., Perona, P., Straw, A.D., Wikelski, M., Brose, U., 2014. Automated image-based tracking and its application in ecology. Trends Ecol. Evol. 29 (7), 417–428, http://dx.doi.org/10.1016/j.tree.2014.05.004.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B 39 (1), 1–38, URL http://www.jstor.org/stable/2984875.

Egnor, S.E.R., Branson, K., 2016. Computational analysis of behavior. Annu. Rev. Neurosci. 39 (1), 217–236, http://dx.doi.org/10.1146/annurev-neuro-070815-013845.

Please cite this article in press as: Lorbach, M., et al., Learning to recognize rat social behavior: Novel dataset and cross-dataset application. J Neurosci Methods (2017), http://dx.doi.org/10.1016/j.jneumeth.2017.05.006

Eyjolfsdottir, E., Branson, S., Burgos-Artizzu, X.P., Hoopfer, E.D., Schor, J., Anderson, D.J., Perona, P., 2014. Detecting social actions of fruit flies. Proc. Conf. Comput. Vis. (ECCV) 8690, 772–787, http://dx.doi.org/10.1007/978-3-319-10605-2_50.
File, S.E., Seth, P., 2003. A review of 25 years of the social interaction test. Eur. J. Pharmacol. 463 (1–3), 35–53, http://dx.doi.org/10.1016/S0014-2999(03)01273-1.
Giancardo, L., Sona, D., Huang, H., Sannino, S., Managò, F., Scheggia, D., Papaleo, F., Murino, V., 2013. Automatic visual tracking and social behaviour analysis with multiple mice. PLoS ONE 8 (9), E74557, http://dx.doi.org/10.1371/journal.pone.0074557.
Homberg, J., Olivier, J., VandenBroeke, M., Youn, J., Ellenbroek, A., Karel, P., Shan, L., Van, B., Ooms, S., Balemans, M., Langedijk, J., Muller, M., Vriend, G., Cools, A., Cuppen, E., Ellenbroek, B., 2016. The role of the dopamine D1 receptor in social cognition: studies using a novel genetic rat model. Dis. Models Mech. 9 (10), 1147–1158, http://dx.doi.org/10.1242/dmm.024752.
Hong, W., Kennedy, A., Burgos-Artizzu, X.P., Zelikowsky, M., Navonne, S.G., Perona, P., Anderson, D.J., 2015. Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning. Proc. Natl. Acad. Sci. U. S. A. 112 (38), E5351–E5360, http://dx.doi.org/10.1073/pnas.1515982112.
Kabra, M., Robie, A.A., Rivera-Alba, M., Branson, S., Branson, K., 2012. JAABA: interactive machine learning for automatic annotation of animal behavior. Nat. Methods 10 (1), 64–67, http://dx.doi.org/10.1038/nmeth.2281.
Kelp, A., Koeppen, A.H., Petrasch-Parwez, E., Calaminus, C., Bauer, C., Portal, E., Yu-Taeger, L., Pichler, B., Bauer, P., Riess, O., Nguyen, H.P., 2013. A novel transgenic rat model for Spinocerebellar Ataxia Type 17 recapitulates neuropathological changes and supplies in vivo imaging biomarkers. J. Neurosci. 33 (21), 9068–9081, http://dx.doi.org/10.1523/JNEUROSCI.5622-12.2013.
Kuehne, H., Gall, J., Serre, T., 2016. An end-to-end generative framework for video segmentation and recognition. Proc. Conf. Appl. Comput. Vis. (WACV), 1–8, http://dx.doi.org/10.1109/WACV.2016.7477701.
Kyriakou, E.I., van der Kieft, J.G., de Heer, R.C., Spink, A., Nguyen, H.P., Homberg, J.R., van der Harst, J.E., 2016. Automated quantitative analysis to assess motor function in different rat models of impaired coordination and ataxia. J. Neurosci. Methods 268, 171–181, http://dx.doi.org/10.1016/j.jneumeth.2015.12.001.
Lewejohann, L., Hoppmann, A.M., Kegel, P., Kritzler, M., Krüger, A., Sachser, N., 2009. Behavioral phenotyping of a murine model of Alzheimer's disease in a seminaturalistic environment using RFID tracking. Behav. Res. Methods 41 (3), 850–856, http://dx.doi.org/10.3758/BRM.41.3.850.
Lorbach, M., Poppe, R., van Dam, E.A., Noldus, L.P.J.J., Veltkamp, R.C., 2015. Automated recognition of social behavior in rats: the role of feature quality. Proc. Conf. Image Anal. Process. (ICIAP), 565–574, http://dx.doi.org/10.1007/978-3-319-23234-8_52.
Noldus, L.P.J.J., Spink, A.J., Tegelenbosch, R.A.J., 2001. EthoVision: a versatile video tracking system for automation of behavioral experiments. Behav. Res. Methods Instrum. Comput. 33 (3), 398–414, http://dx.doi.org/10.3758/BF03195394.
Pérez-Escudero, A., Vicente-Page, J., Hinz, R.C., Arganda, S., de Polavieja, G.G., 2014. idTracker: tracking individuals in a group by automatic identification of unmarked animals. Nat. Methods 11 (7), 743–748, http://dx.doi.org/10.1038/nmeth.2994.
Peters, S.M., Pothuizen, H.H.J., Spruijt, B.M., 2015. Ethological concepts enhance the translational value of animal models. Eur. J. Pharmacol. 759, 42–50, http://dx.doi.org/10.1016/j.ejphar.2015.03.043.
Peters, S.M., Pinter, I.J., Pothuizen, H.H.J., de Heer, R.C., van der Harst, J.E., Spruijt, B.M., 2016. Novel approach to automatically classify rat social behavior using a video tracking system. J. Neurosci. Methods 268, 163–170, http://dx.doi.org/10.1016/j.jneumeth.2016.02.020.
Richardson, C.A., 2015. The power of automated behavioural homecage technologies in characterizing disease progression in laboratory mice: a review. Appl. Anim. Behav. Sci. 163, 19–27, http://dx.doi.org/10.1016/j.applanim.2014.11.018.
Richter, S.H., Garner, J.P., Würbel, H., 2009. Environmental standardization: cure or cause of poor reproducibility in animal experiments? Nat. Methods 6 (4), 257–261, http://dx.doi.org/10.1038/nmeth.1312.
Robie, A.A., Seagraves, K.M., Egnor, S.E.R., Branson, K., 2017. Machine vision methods for analyzing social interactions. J. Exp. Biol. 220 (1), 25–34, http://dx.doi.org/10.1242/jeb.142281.
Sams-Dodd, F., 1995. Automation of the social interaction test by a video-tracking system: behavioural effects of repeated phencyclidine treatment. J. Neurosci. Methods 59 (2), 157–167, http://dx.doi.org/10.1016/0165-0270(94)00173-E.
Schaefer, A.T., Claridge-Chang, A., 2012. The surveillance state of behavioral automation. Curr. Opin. Neurobiol. 22 (1), 170–176, http://dx.doi.org/10.1016/j.conb.2011.11.004.
Schneider, J., Levine, J.D., 2014. Automated identification of social interaction criteria in *Drosophila melanogaster*. Biol. Lett. 10 (10), E20140749, http://dx.doi.org/10.1098/rsbl.2014.0749.
Spruijt, B.M., Hol, T., Rousseau, J., 1992. Approach, avoidance, and contact behavior of individually recognized animals automatically quantified with an imaging technique. Physiol. Behav. 51 (4), 747–752, http://dx.doi.org/10.1016/0031-9384(92)90111-E.
Steele, A.D., Jackson, W.S., King, O.D., Lindquist, S., 2007. The power of automated high-resolution behavior analysis revealed by its application to mouse models of Huntington's and Prion diseases. Proc. Natl. Acad. Sci. U. S. A. 104 (6), 1983–1988, http://dx.doi.org/10.1073/pnas.0610779104.
Urbach, Y.K., Bode, F.J., Nguyen, H.P., Riess, O., von Hörsten, S., 2010. Neurobehavioral tests in rat models of degenerative brain diseases. In: Rat Genomics, vol. 597 of Methods in Molecular Biology. Humana Press, pp. 333–356, http://dx.doi.org/10.1007/978-1-60327-389-3_24.
Urbach, Y.K., Raber, K.A., Canneva, F., Plank, A.C., Andreasson, T., Ponten, H., Kullingsjö, J., Nguyen, H.P., Riess, O., von Hörsten, S., 2014. Automated phenotyping and advanced data mining exemplified in rats transgenic for Huntington's disease. J. Neurosci. Methods 234, 38–53, http://dx.doi.org/10.1016/j.jneumeth.2014.06.017.
van Dam, E.A., van der Harst, J.E., ter Braak, C.J.F., Tegelenbosch, R.A.J., Spruijt, B.M., Noldus, L.P.J.J., 2013. An automated system for the recognition of various specific rat behaviours. J. Neurosci. Methods 218 (2), 214–224, http://dx.doi.org/10.1016/j.jneumeth.2013.05.012.
Veeraragavan, S., Wan, Y.W., Connolly, D.R., Hamilton, S.M., Ward, C.S., Soriano, S., Pitcher, M.R., McGraw, C.M., Huang, S.G., Green, J.R., Yuva, L.A., Liang, A.J., Neul, J.L., Yasui, D.H., LaSalle, J.M., Liu, Z., Paylor, R., Samaco, R.C., 2016. Loss of MeCP2 in the rat models regression, impaired sociability and transcriptional deficits of Rett syndrome. Hum. Mol. Genet. 25 (15), 3284, http://dx.doi.org/10.1093/hmg/ddw178.
Wilson, C.A., Koenig, J.I., 2014. Social interaction and social withdrawal in rodents as readouts for investigating the negative symptoms of schizophrenia. Eur. Neuropsychopharmacol. 24 (5), 759–773, http://dx.doi.org/10.1016/j.euroneuro.2013.11.008.

Please cite this article in press as: Lorbach, M., et al., Learning to recognize rat social behavior: Novel dataset and cross-dataset application. J Neurosci Methods (2017), http://dx.doi.org/10.1016/j.jneumeth.2017.05.006