Automated Recognition of Social Behavior in Rats: The Role of Feature Quality

Malte Lorbach[1,2], Ronald Poppe[1], Elsbeth A. van Dam[2], Lucas P.J.J. Noldus[2], and Remco C. Veltkamp[1]

[1] Utrecht University, Department of Information and Computing Sciences, Utrecht, The Netherlands
[2] Noldus Information Technology, Wageningen, The Netherlands

Abstract. We investigate how video-based recognition of rat social behavior is affected by the quality of the tracking data and the derived feature set. We look at the impact of two common tracking errors – animal misidentification and inaccurate localization of body parts. We further examine how the complexity of representing the articulated body in the features influences the recognition accuracy. Our analyses show that correct identification of the rats is required to accurately recognize their interactions. Precise localization of multiple body points is beneficial for recognizing interactions that are described by a distinct pose. Including pose features only leads to improvement if the tracking algorithm can provide that data reliably.

Keywords: social behavior, action recognition, tracking quality

1 Introduction

We investigate the automated recognition of social interactions between rats. Rat social behavior is of interest for biologists who look for indicators for neurological and psychiatric disorders such as Huntington's disease. Such indicators can be abnormalities in how often and how long the animals engage in specific social interactions. Currently, these studies involve laborious and error-prone manual coding of interactions and thus automating the coding is desired.

Video-based recognition of rat interactions typically requires three problems to be solved, namely: tracking and identifying the animals in the presence of occlusions, deriving meaningful features from these tracks, and classifying the features into interaction categories. Previous work on recognizing interactions has mainly focused on these steps in isolation, in particular by assuming perfect tracking when computing features. The effects of mistaken identities and noisy tracking on the classification have received less attention. As a consequence, we yet lack the ability to trace back recognition errors to either tracking or classification.

With this paper we aim at unraveling the links between feature quality and recognition accuracy. We derive trajectory features from tracking data with varying degrees of common errors, and compare the performance using off-the-shelf

This is the post-print version of Lorbach, M., Poppe, R., van Dam, E.A., Noldus, L.P., Veltkamp, R.C.: Automated Recognition of Social Behavior in Rats: The Role of Feature Quality. In: Murino, V., Puppo, E. (eds.) Image Analysis and Processing ICIAP 2015, Lecture Notes in Computer Science, vol. 9280, pp. 565574. Springer International Publishing (2015). The final publication is available at Springer via `http://dx.doi.org/10.1007/978-3-319-23234-8_52`.

classifiers. This work can be seen as a thorough investigation of the factors involved in automated rat social behavior analysis.

The remainder of this work is structured as follows. In Section 2 we discuss related work. Sections 3 and 4 introduce our data set and the analysis pipeline. The results are presented in Section 5 and discussed in Section 6. We conclude in Section 7.

2 Rodent Action Recognition

Action recognition has been applied not only to rodents [4] but also to humans. In contrast to human action recognition, the recognition of rodent actions is characterized by confined spaces, less articulated, similar looking animals, and a combination of need-driven and playful behavior. The common procedure in rodent action recognition is to split the recognition into three tasks: tracking the position of the animals, deriving features from those tracks, and classifying the actions using the features.

Different tracking solutions have been presented. Some require that the animals are uniquely marked [9] or have an implanted RFID tag [12]. Others attempt to identify the animals based on their thermal [6] or visual appearance [10]. Recently, the use of depth cameras has been proposed to enhance the visual segmentation in contact situations [8]. A pronounced difference of the solutions is whether only one [1,2],[12] or more body parts [3],[6] are tracked. Tracking multiple body parts has been shown to improve solitary behavior recognition [5].

The location data obtained by the tracking algorithm is used to derive a feature set. This set often comprises individual features such as velocities and accelerations [7], and pairwise features such as the distance between animals and their relative orientation [2]. In addition, one may add features derived directly from the image data. Exploiting spatio-temporal interest points in a bag-of-words setup has been shown to yield only minor improvements over a trajectory-only feature set [2].

At the classification level, differences can be found in the way temporal information is considered. If the video frames are considered samples that have to be assigned a class label, then temporal information may be included by collecting statistical values across neighboring frames using a sliding window [4],[7]. To model temporal information in a more structural way, for example, to incorporate transition probabilities between interactions, one can deploy hidden Markov models [1]. The classification problem is then formulated so as to find the optimal temporal segmentation of the video into labeled action segments.

Most recognition systems are trained using a subset of the data. Exceptions are rule-based classifiers [3],[12], and the Janelia Automatic Animal Behavior Annotator (JAABA) [7]. The latter pursues an active learning approach in which the user trains a classifier by iteratively annotating a number of action events.

Tracking, feature extraction, and classification clearly depend on each other. Despite advances in all three areas, it has not been analyzed systematically how errors in one task affect the final classification.

3 Rat Social Behavior Data Set

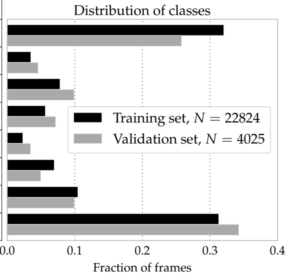
The Rat Social Behavior Data Set (RSBD), which we use throughout our analyses, was obtained in a study on play behavior of young rats [11]. In 40 sessions, two male Sprague Dawley rats, 5-6 weeks old, were placed together in a Noldus PhenoTyper 9000 cage (90 cm × 90 cm) and were recorded by an infrared camera at 25 fps from a top-view perspective for about 30 min. The actions of one focal animal were labeled using Noldus Observer XT 10. From the 14 original labels, we removed actions that are too subtle to be captured by trajectory features (e.g., biting and kicking). The remaining classes capture seven interactions and one class that covers all solitary actions. Short descriptions of the classes are given in Table 1.

For our analyses we chose five videos from which we randomly selected ten events of each interaction per video (400 segments in total). Every segment includes a 0.6 s margin before and after the interaction. In total this yields 12.6 min of footage. We further chose four of the five videos at random to be used for training our system. The remaining video was considered a validation set and has never been used other than for the results presented in this paper.

Rat interactions have different temporal properties. Their durations have different means and often large inner-class variances as we can see in Table 1. This difference leads to a highly unbalanced data set. Note that our data is unbalanced regarding the number of frames but balanced in the number of interaction events.

Table 1. Left: mean, standard deviation, minimum, and maximum of the durations (in s) of the selected interactions in the training set. Right: the distribution of classes in both training and validation set.

	μ	σ	min	max
Allogrooming (alo): grooming fur of other rat	6.21	7.12	0.32	30.48
Approaching (app): moving towards other rat	0.40	0.26	0.08	1.40
Moving away (awy): moving away from other rat	0.68	0.59	0.08	5.00
Following (fol): following other rat	0.89	0.95	0.08	6.32
Nape attacking (nap): attacking other rat's neck area	0.45	0.49	0.04	3.80
Pinning (pin): keeping other rat lying on its back	1.57	1.28	0.28	6.00
Social nose contact (snc): inspecting other body	0.82	1.05	0.04	7.72
Solitary actions (sol): all non-social behaviors	4.48	8.97	0.08	49.24



4 Feature Quality in Social Behavior Recognition

Our goal is to highlight how the feature quality influences the recognition performance. We vary the quality in two ways. First, we incrementally correct two types of tracking errors. Second, we derive three feature sets from those tracks capturing the rat's articulated body at varying degrees of detail.

4.1 Eliminating Systematic Tracking Errors in RSBD

The video tracking system used in this experiment (Noldus EthoVision XT 11) tracks three points on the animal bodies: the nose point, the center of gravity, and the tail-base. We incrementally eliminate two types of tracking errors and thus introduce three data set versions. We denote the initial, uncorrected version as `RSBD`. In the first step (denoted as `RSBD-ID`), we corrected identity swaps. Identities were not changed during fast, close-contact situations. In those situations, the positions provided by the tracker are occasionally wrong and thus identity assignment becomes arbitrary. In the second step (`RSBD-ID+Loc`), we additionally corrected all body point locations. This decreases the amount of noise in the positions, eliminates body part confusions (swaps of nose and tail-base points), and yields reliable orientation values.

4.2 Extracting Features from the Data Set Versions

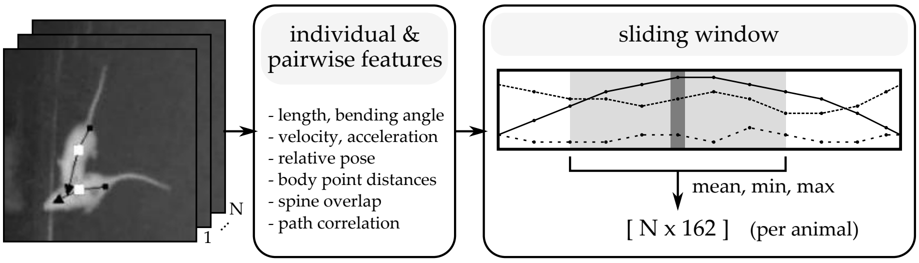


Fig. 1. From nose, center, and tail-base points we derive individual and pairwise features that describe the pose, motion, and distances of the animals in each of the N frames. In a window centered at the current frame various statistics are computed.

From the tracked body points, we compute, per animal, a number of pose, motion, and distance variables as well as their derivatives. The values are aggregated over time computing mean, minimum, and maximum values in a sliding window of 0.52 s. Figure 1 illustrates the extraction pipeline. The variants of the feature set are created by varying how detailed the articulated body is captured by the features.

Feature Set Variants We compare three sets with ascending number of features. In the first set (`CP`), we only exploit the position (x, y) of the animals' center-of-gravity. This corresponds to the approach taken by a number of previous works, e.g., [2],[12]. Features include velocity and acceleration, the distance between the animals, and the correlation between the animals' paths. In the second set (`CP+Ori`), we add orientation information (x, y, φ) which allows us to compute the velocity vector with respect to the animal's orientation and the relative orientation between the animals. The third set (`Full`) exploits all three tracked body points $((x_0, y_0), (x_1, y_1), (x_2, y_2))$ and additionally incorporates pose features such as body length and bending, several body point distances and the degree of body overlap.

To facilitate the generalization of the features to other setups and rats, we standardize the features of all sets with respect to size of the specific rat. That is, distances, velocities, and accelerations are scaled to animal length units.

4.3 Experiment Setup

To analyze the links between feature quality and recognition accuracy, we examine the effects of tracking errors on the accuracy alone (using the `Full` feature set) as well as in combination with the different feature sets. We assess the recognition accuracy in terms of the overall and the per-class classification performance. We mainly look at the F1 score and, if appropriate, at precision, recall and confusions between specific classes. When averaging the F1 score, we average across classes. Compared to averaging across frames, the class average puts higher weight on short or rare events and thus represents unbalanced data sets like ours better.

For overall performance measures, we apply a 5-fold cross-validation scheme where each fold corresponds to one of the five videos in the data set. If we look at per-class performance, we train on the four training videos and test on the validation video. When we compare different tracking errors, we train and test using data of the same error level.

To find a suitable classifier for the analyses, we compare six off-the-shelf classifiers and then stick to one classifier for the remaining experiments. We compare the following classifiers: Linear Discriminant Classifier (LDC); Linear Discriminant Analysis with subsequent One-vs-All Quadratic Discriminant Classifier (LDA+QDC); Support Vector Machines with Gaussian (SVM-RBF) and linear (SVM-Lin) kernels; LDA with k-Nearest-Neighbors (LDA+kNN); and Random Forest (RF). Where applicable, classifier parameters are found empirically by optimizing the F1 score in the same cross-validation scheme as described above.

5 Results

5.1 Tracking errors

The comparison of the classifiers (Tab. 2) shows that all six classifiers perform comparably on all three data set versions. Given the range of classifiers tested, this emphasizes that feature quality, rather than the classifier, largely determines the performance. The remaining experiments are conducted with the simplest of the classifiers: LDC. We further see in Table 2 that fewer tracking errors lead to higher average accuracy. With each additional error eliminated, the average per-class F1 score increases by approximately 0.12.

Looking at the F1 scores per interaction (Tables 3, 4, and 5) and confusions (Figures 2, 3, and 4), we notice that not all interactions are affected by tracking

Table 2. The average per-class F1 scores achieved by the six classifiers on the three data set versions with increasing degree of tracking quality

		`RSBD`		`RSBD-ID`		`RSBD-ID+Loc`	
Classifier	Parameters	μ_{F1}	σ	μ_{F1}	σ	μ_{F1}	σ
LDC	–	0.51	0.05	0.63	0.05	0.75	0.03
LDA+QDC	–	0.50	0.04	0.62	0.05	0.74	0.03
SVM-RBF	$C = 1, \gamma = .00625$	0.51	0.04	0.65	0.03	0.74	0.02
SVM-Lin	$C = 0.001$	0.50	0.04	0.63	0.04	0.74	0.03
LDA+kNN	$k = 10$	0.48	0.04	0.61	0.04	0.73	0.02
RF	$n = 100, d_{\max} = 16$	0.52	0.05	0.68	0.04	0.76	0.02

errors in the same way. The accuracies are generally high for solitary actions and approaches (in which the animals are separated by definition). Contact interactions are not recognized well in the `RSBD` version but improve gradually as errors are corrected. Let us look at each correction step separately.

The correction of identity swaps (`RSBD` $\rightarrow$ `RSBD-ID`) leads to two major improvements. Firstly, the confusion of *following* with *moving away* is largely resolved although some confusion persists. The F1 score for *following* increases from 0.27 to 0.56 and for *moving away* from 0.47 to 0.72. Secondly, virtually all *nape attacks* that had been mistaken as *following* are now corrected. Consequently, the recall of *nape attacking* improves from 0.39 to 0.56. Notably, precision stays at a low level of 0.29.

Correcting the body point locations (`RSBD-ID` $\rightarrow$ `RSBD-ID+Loc`) increases the precision of *nape attacking* from 0.29 to 0.46, and the recall of *pinning* from 0.3 to 0.72. Confusions remain between these two classes and also between *following* and *approaching*. A number of small improvements across all classes eventually leads to higher average F1 scores at both frame level (+0.07) and class level (+0.11).

5.2 Feature Set Variants

Figure 5 shows the F1 scores of the combinations of data set versions and feature sets. There is an upwards trend across the data set versions irrespective of which feature set is used. In `RSBD`, the F1 score remains at approximately 0.5 for all feature sets. In both `RSBD-ID` and `RSBD-ID+Loc`, the F1 score increases with richer feature sets. The standard deviation of the accuracy decreases by approximately 50% using `Full` on `RSBD-ID+Loc` compared to `RSBD-ID`.

6 Discussion

On the overall performance level, we have seen that eliminating tracking errors leads to better classification. This pattern occurred for all tested classifiers, which suggests that the effect is indeed inherent to the underlying data and not to the classifier. We further showed that orientation and pose features are important

Table 3. Per-class results: `RSBD`

	Prec.	Recall	F1	#
alo	0.86	0.71	0.78	1038
app	0.55	0.80	0.65	184
awy	0.48	0.46	0.47	398
fol	0.33	0.24	0.27	288
nap	0.29	0.39	0.33	139
pin	0.24	0.35	0.28	200
snc	0.40	0.61	0.48	399
sol	0.98	0.87	0.92	1379
μ_{frames}	0.72	0.67	0.69	4025
μ_{classes}	0.52	0.55	**0.52**	8

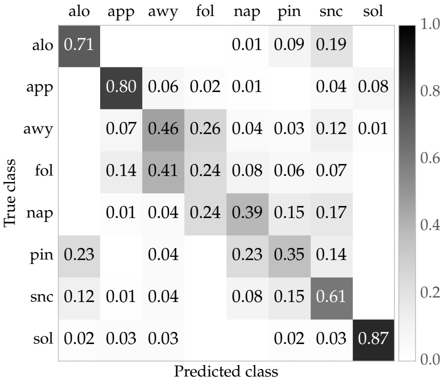


Fig. 2. Confusion matrix: `RSBD`

Table 4. Per-class results: `RSBD-ID`

	Prec.	Recall	F1	#
alo	0.85	0.83	0.84	1038
app	0.62	0.87	0.72	184
awy	0.75	0.70	0.72	398
fol	0.65	0.50	0.56	288
nap	0.29	0.56	0.39	139
pin	0.65	0.30	0.41	200
snc	0.50	0.70	0.58	399
sol	0.98	0.88	0.93	1379
μ_{frames}	0.80	0.76	0.77	4025
μ_{classes}	0.66	0.67	**0.65**	8

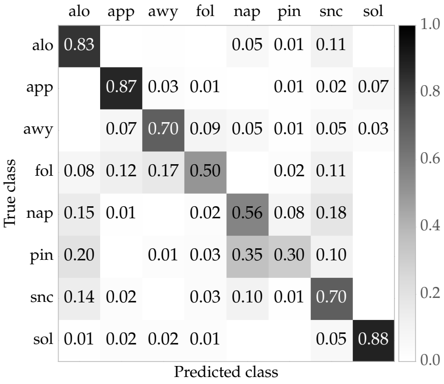


Fig. 3. Confusion matrix: `RSBD-ID`

Table 5. Per-class results: `RSBD-ID+Loc`

	Prec.	Recall	F1	#
alo	0.92	0.92	0.92	1038
app	0.67	0.83	0.74	184
awy	0.73	0.87	0.79	398
fol	0.89	0.59	0.71	288
nap	0.46	0.69	0.55	139
pin	0.74	0.72	0.73	200
snc	0.63	0.67	0.65	399
sol	0.97	0.90	0.93	1379
μ_{frames}	0.85	0.84	0.84	4025
μ_{classes}	0.75	0.77	**0.76**	8

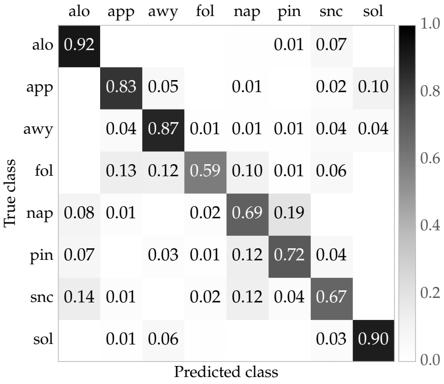


Fig. 4. Confusion matrix: `RSBD-ID+Loc`

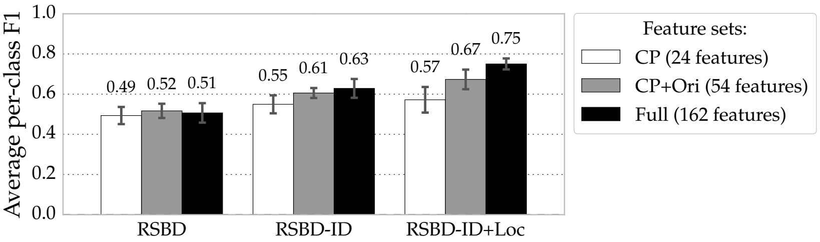


Fig. 5. The average per-class F1 score using three different feature sets, tested on all three data set versions

for the recognition. If those features are correct, they lead to better classification. If they are not, that is, if the tracking algorithm fails to provide stable pose information, we induce the risk of overfitting to the noise in the features. As a consequence, the classification accuracy stagnates or even decreases. A potential way to overcome this limitation is to include more training data, which are particularly expensive to obtain. Moreover, when we trained the classifier with corrected data but used uncorrected data to test it, we failed to achieve competitive performance ($\mu_{\text{classes}} = 0.42$, $\sigma = 0.05$, 5-fold cross-validation). For that reason, we do not benefit from corrected, clean features as long as we cannot guarantee that we can generate them without expensive, manual intervention.

6.1 Difference between Interactions

On the class level, we observed that the classes are affected differently by tracking errors and the choice of features. By which type of tracking error an interaction is most affected is determined by its characteristics. Interactions such as *following* and *pinning* rely more on the identity assignment than, for example, *solitary actions*. Because most of our interactions are indeed sensitive to the correct role assignment, we see large gains in F1 score after correcting identity swaps. Clearly, maintaining the correct identities is a necessity for social behavior recognition.

Another characteristic of the interactions is how important the relative pose is for the recognition. *Nape attacking*, *pinning*, and *following* events have a very distinct relative pose while it is less relevant for other interactions. For example, for *social nose contact* the pose can be different in every event because the class includes the inspection of all body parts. Therefore, we expect that the more an interaction is defined by the pose, the better it should be recognized if correct pose features are provided. We find supporting evidence in the results. *Nape attacking* (+0.16), *pinning* (+0.32), and *following* (+0.15) benefit most from the correction of body part locations and thus pose. Accordingly, adding uncorrected orientation and pose features results in only a small improvement (`RSBD-ID`: `CP` → `Full` = +0.08). We conclude that the accuracy of social behavior recognition can be improved by incorporating reliable orientation and pose features.

6.2 Unresolved Confusions

There are some confusions that persist even with perfect tracking. The predominant confusions occur between *following* and *appoaching*, and among the four classes *allogrooming*, *nape attacking*, *pinning*, and *social nose contact*.

There are two reasons for the confusions. First, *approaching* often evolves into *following* but the transition is not clearly defined. As a result, the predictions around the transition point become arbitrary. We see the same effect to a lesser degree for *awy* → *sol* and *sol* → *app*. Second, the four confused interactions can be very ambiguous in their appearance. The classifier cannot separate the classes properly and hence makes mistakes.

As for solving the transition ambiguity, we need to find more clues to when one behavior changes into another. A potential direction is to explicitly learn the temporal structure of the transitions and to incorporate the other rat's reaction.

To improve the separability of ambiguous interactions, we may want to increase the diversity of the features. The four confused interactions are ambiguous because they are close-contact situations for which the animal's trajectories and poses appear similar. However, differences may arise if we incorporate which animal is on top or below (e.g., by exploiting 3D trajectories) and capture fine-grained motion with image features (e.g., optical flow or histogram of gradients).

7 Conclusion

In this paper we investigated the effects of feature quality on video-based recognition of rat social behavior. We looked at the impact of two types of tracking errors – misidentification and inaccurate localization – as well as the type of features that are derived from the tracking data.

From the analysis of the classification accuracy across interaction classes, we observed that although correcting tracking errors improves the classification, each class is affected differently. Correctly identifying the animals is required to recognize virtually all interactions, whereas correctly tracking body parts has a larger impact on classes that are defined by a distinct relative pose. Hence, including orientation and pose features is advantageous under the condition that the tracking algorithm can provide them reliably.

We have further found that perfect tracking alone is insufficient for recognizing ambiguous behavior. Exploiting temporal context and reaction patterns alongside with features that go beyond 2D trajectories are directions that seem worth pursuing in the future.

Acknowledgments. We would like to thank Suzanne Peters and Johanneke van der Harst (Delta Phenomics) for their annotation work. 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). This work was funded by the European Commission under the 7$^{\text{th}}$ Framework Programme

(FP7) as part of the Marie Curie Initial Training Network "PhenoRat" (GA no. 317259).

References

1. Arakawa, T., Tanave, A., Ikeuchi, S., Takahashi, A., Kakihara, S., Kimura, S., Sugimoto, H., Asada, N., Shiroishi, T., Tomihara, K., Tsuchiya, T., Koide, T.: A male-specific QTL for social interaction behavior in mice mapped with automated pattern detection by a hidden Markov model incorporated into newly developed freeware. J. Neurosci. Meth. 234, 127–134 (2014)
2. Burgos-Artizzu, X.P., Dollár, P., Lin, D., Anderson, D.J., Perona, P.: Social behavior recognition in continuous video. In: Proc. CVPR. pp. 1322–1329 (2012)
3. de Chaumont, F., Coura, R.D.S., Serreau, P., Cressant, A., Chabout, J., Granon, S., Olivo-Marin, J.C.: Computerized video analysis of social interactions in mice. Nat. Methods 9(4), 410–417 (2012)
4. 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.: An automated system for the recognition of various specific rat behaviours. J. Neurosci. Meth. 218(2), 214–224 (2013)
5. Decker, C., Hamprecht, F.A.: Detecting individual body parts improves mouse behavior classification. In: Proc. of the Workshop on Visual observation and analysis of Vertebrate and Insect Behavior. Stockholm, Sweden (2014)
6. Giancardo, L., Sona, D., Huang, H., Sannino, S., Managò, F., Scheggia, D., Papaleo, F., Murino, V.: Automatic visual tracking and social behaviour analysis with multiple mice. PLoS ONE 8(9), e74557 (2013)
7. Kabra, M., Robie, A.A., Rivera-Alba, M., Branson, S., Branson, K.: JAABA: Interactive machine learning for automatic annotation of animal behavior. Nat. Methods 10(1), 64–67 (2012)
8. Matsumoto, J., Urakawa, S., Takamura, Y., Malcher-Lopes, R., Hori, E., Tomaz, C., Ono, T., Nishijo, H.: A 3D-video-based computerized analysis of social and sexual interactions in rats. PLoS ONE 8(10), e78460 (2013)
9. Ohayon, S., Avni, O., Taylor, A.L., Perona, P., Roian Egnor, S.: Automated multiday tracking of marked mice for the analysis of social behaviour. J. Neurosci. Meth. 219(1), 10–19 (2013)
10. Pérez-Escudero, A., Vicente-Page, J., Hinz, R.C., Arganda, S., de Polavieja, G.G.: idTracker: Tracking individuals in a group by automatic identification of unmarked animals. Nat. Methods 11(7), 743–748 (2014)
11. Peters, S.M., Pinter, I., de Heer, R.C., van der Harst, J.E., Spruijt, B.M.: Automated classification of rat social behavior. In: Proc. of Measuring Behavior. Wageningen, The Netherlands (2014)
12. Weissbrod, A., Shapiro, A., Vasserman, G., Edry, L., Dayan, M., Yitzhaky, A., Hertzberg, L., Feinerman, O., Kimchi, T.: Automated long-term tracking and social behavioural phenotyping of animal colonies within a semi-natural environment. Nat. Commun. 4, Article No. 2018 (2013)