AUTOMATED HOME-CAGE BEHAVIORAL PHENOTYPING OF MICE

Hueihan Jhuang[1], Estibaliz Garrote[1], Xinlin Yu[3], Vinita Khilnani[3], Tomaso Poggio[1], Andrew D. Steele[3]* and Thomas Serre [1,2]*

[1]Department of Brain and Cognitive Sciences, McGovern Institute, Massachusetts Institute of Technology

[2]Current address: Department of Cognitive, Linguistic & Psychological Sciences, Brown Institute for Brain Sciences, Brown University

[3]Broad Fellows in Brain Circuitry Program, Division of Biology, California Institute of Technology

*Corresponding authors: TS serre@mit.edu or AS steelea@caltech.edu

ABSTRACT

Neurobehavioral analysis of mouse phenotypes requires the monitoring of mouse behavior over long periods of time. Here, we describe a trainable computer vision system enabling the automated analysis of complex mouse behaviors. We provide software and an extensive manually annotated video database used for training and testing the system. Our system performs on par with human scoring, as measured from ground-truth manual annotations of thousands of clips of freely behaving mice. As a validation of the system, we characterized the home-cage behaviors of two standard inbred and two non-standard mouse strains. From this data we were able to predict in a blind test the strain identity of individual animals with high accuracy. Our video-based software will complement existing sensor based automated approaches and enable an adaptable, comprehensive, high-throughput, fine-grained, automated analysis of mouse behavior.

INTRODUCTION

Automated quantitative analysis of mouse behavior will play a significant role in comprehensive phenotypic analyses – both on the small scale of detailed characterization of individual gene mutants and on the large scale of assigning gene function across the entire mouse genome[1]. One key benefit of automating behavioral analysis arises from inherent limitations of human assessment: namely cost, time, and reproducibility. Although automation in and of itself is not a panacea for neurobehavioral experiments[2], it allows for addressing an entirely new set of questions about mouse behavior and to conduct experiments on time scales that are orders of magnitude larger than traditionally assayed. For example, reported tests of grooming behavior span time scales of minutes[3, 4] whereas an automated analysis will allow for analysis of this behavior over hours or even days and weeks.

Indeed, the significance of alterations in home-cage behavior has recently gained attention as an effective means to detect perturbations in neural circuit function – both in the context of disease detection and more generally to measure food consumption and activity parameters[5-10]. Previous automated systems (e.g., ref.[8, 9, 11, 12] and Supplementary Note) rely mostly on the use of simple detectors such as infrared beams to monitor behavior. These sensor-based approaches tend to be limited in the complexity of the behavior that they can measure, even in the case of costly commercial systems using transponder technologies[13]. While such systems can be used effectively to monitor locomotor activity and perform operant conditioning, they cannot be used to study home-cage behaviors such as grooming, hanging, jumping, and smaller movements (termed 'micro-movements' below). Visual analysis is a potentially powerful complement to these sensor-based approaches for the recognition of such fine animal behaviors.

Advances in computer vision and machine learning over the last decade have yielded robust computer vision systems for the recognition of objects[14, 15] and human actions (see ref. [16] for review). In fact, the use of vision-based approaches is already bearing fruit for the automated tracking[17-19] and recognition of behaviors in insects[20, 21]. Several open-source and commercial computer-vision systems for the recognition of mouse behavior have also been developed (see ref. [22, 23] and Supplementary Note). However, these systems are not widely used, exhibit similar limitations to sensor-based approaches, or are cost prohibitive.

In this paper, we describe a trainable, general-purpose, automated and potentially high-throughput system for the behavioral analysis of mice in their home-cage. Developed from a computational model of motion processing in the primate visual cortex[24, 25], the computer system is trained with labeled examples with manually annotated behaviors of interest and used to analyze automatically new recordings containing hours of freely behaving animals. As a proof of concept, we trained the system on common mouse behaviors and demonstrate that the resulting system performs on par with humans for the scoring of these behaviors. Using the resulting system, we analyze the home-cage behavior of several mouse strains, including the commonly used strains C57BL/6J, DBA/2J, the BTBR strain that displays autistic-like behaviors, and a wild-derived strain CAST/EiJ. We characterize differences in the behaviors of these strains and use these profiles to predict the strain type of an animal.

RESULTS

Our system (available as Supplementary Software) consists of three separate modules: (1) a video database, (2) a feature computation module, and (3) a classification module.

Video database

We video-recorded a large database of video sequences of singly housed mice in their home-cages from an angle perpendicular to the side of the cage (see Fig. 1 for examples of video frames) using a consumer grade camcorder. In order to create a robust recognition system we varied the lighting conditions by placing the cage in different positions with respect to the overhead lighting. In addition, we used many mice of different size, gender, and coat color. We considered eight behaviors of interest, which included: drinking, eating, grooming, hanging, rearing, walking, resting, and micro-movements of the head. Several investigators were trained to score the mouse behavior using two different scoring techniques.

The first set of annotations denoted the *'clipped database'* included only clips scored with very high stringency, seeking to annotate only the best and most exemplary instances of particular behaviors. A pool of eight annotators (*'Annotator group 1'*) manually hand-scored more than 9,000 short clips, each containing a unique annotation. To avoid errors, this database was then curated by one of the annotators who watched all 9,000 clips again, retaining only the most unambiguous assessments, leaving 4,200 clips (26,2360 frames corresponding to about 2.5 hours) from twelve distinct videos (recorded at twelve separate sessions) to train and tune the feature computation module of the proposed system, as described below.

The second set of annotations, called the *'full database'* involved labeling every frame (with less stringency than in the *'clipped database'*) for twelve unique videos (different from the twelve videos used in the *'clipped database'*) corresponding to over 10 hours of continuously annotated video. Again two sets of annotators were used to correct mistakes and make sure the annotation style was consistent throughout the whole database. This database was used to train and test the classification module of the computer system. The distribution of behavior labels for the *'clipped database'* and the *'full database'* is shown in Supplementary Fig. S1a-b and Supplementary Fig. S1c-d respectively.

Computation and evaluation of motion and position features

The architecture used here to pre-process raw video sequences (Fig. 2a-b) and extract motion features (Fig. 2c) is adapted from previous work for the recognition of human actions and biological motion[25]. To speed up the system, the computation of the motion features was limited to a sub-window centered on the animal (Fig. 2b), whose location can be computed from the foreground pixels obtained by subtracting off the video background (Fig. 2a). For a static camera as used here, the video background can be well approximated by a median frame in which each pixel value corresponds to the median gray-value computed over all frames for that pixel location (day and night frames under red lights were processed in separate videos).

The computation of motion features is based on the organization of the dorsal stream of the visual cortex, which has been linked to the processing of motion information (see ref.[26] for a recent review.) Details about the implementation are provided in the Supplementary Methods. A hallmark of the system is its hierarchical architecture: The first processing stage corresponds to an array of spatio-temporal filters tuned to four different directions of motion and modeled after motion-sensitive (simple) cells in the primary visual cortex (V1)[27] (S1/C1 layers, Fig. 2d). The architecture then extracts space-time motion features centered at every frame of an input video sequence via a multiple processing stages, whereby features become increasingly complex and invariant with respect to 2D transformations as one moves up the hierarchy. These motion features are obtained by combining the response of V1-like afferent motion units that are tuned to different directions of motion (Fig. 2e, also see Supplementary Methods for details).

The output of this hierarchical pre-processing module consists of a dictionary of about 300 space-time motion features ($S2/C2$ layers, Fig. 2e) that are obtained by matching the output of the $S1/C1$ layers with a dictionary of motion-feature templates. This basic dictionary of motion-feature templates corresponds to discriminative motion features that are learned from a training set of videos containing labeled behaviors of interest (the *'clipped database'*), via a feature selection technique.

To optimize the performance of the system for the recognition of mouse behaviors, several key parameters of the model were adjusted. The parameters of the spatio-temporal filters in the first stage (e.g., their preferred speed tuning and direction of motion, the nature of the non-linear transfer function used, the video resolution, etc) were adjusted so as to maximize performance on the *'clipped database'*.

In order to evaluate the quality of these motion features for the recognition of high-quality unambiguous behaviors we trained and tested a multi-class Support Vector Machine (SVM) on *single isolated* frames from the *'clipped database'* using the all-pair multi-class classification strategy. This approach does not rely on the temporal context of measured behaviors beyond the computation of low-level motion signals and classifies each frame independently. On the *'clipped database'*, we find that such a system leads to 93% accuracy (as the percentage of correctly predicted clips, chance level 12.5% for 8-class classification), which is significantly higher than the performance of a representative computer vision system[23] (81%) trained and tested in the same conditions (see Supplementary Methods). Performance was estimated based on a leave-one-video-out procedure, whereby clips from all except one video are used to train the system while performance is evaluated on the clips from the remaining video. The procedure was repeated for all videos; we report the overall accuracy. This suggests that the representation provided by the dictionary of motion-feature templates is suitable for the recognition of the behaviors of interest even under conditions where the global temporal structure (i.e., the temporal structure beyond the computation of low-level motion signals) of the underlying temporal sequence is discarded.

In addition to the motion features described above, we computed an additional set of features derived from the instantaneous location of the animal in the cage (Fig. 2f). Position- and velocity-based measurements were estimated based on the 2D coordinates *(x, y)* of the foreground pixels (Fig. 2a) for every frame. These included the position and the aspect ratio of the bounding box around the animal (indicating whether the animal is in a horizontal or vertical posture), the distance of the animal to the feeder as well as its instantaneous velocity and acceleration. Fig. 2f illustrates some of the key features used (see Supplementary Table S1 for a complete list).

Classification module

The reliable phenotyping of an animal requires more than the mere detection of stereotypical non-ambiguous behaviors. In particular, the present system aims at classifying every frame of a video sequence even for those frames that are ambiguous and difficult to categorize. For this challenging task, the temporal context of a specific behavior becomes an essential source of information; thus, learning an accurate temporal model for the recognition of actions becomes critical (see Supplementary Fig. S2 for an illustration). Here we used a Hidden Markov Support Vector Machine (SVMHMM, Fig. 2g)[28, 29], which is an extension of the Support Vector Machine classifier for sequence tagging. This temporal model was trained on the *'full database'* as described above, which contains manually labeled examples of about 10 hours of *continuously scored* video sequences from twelve distinct videos.

Assessing the accuracy of the system is a critical task. Therefore, we made two comparisons: 1) between the resulting system and commercial software (HomeCageScan 2.0, CleverSys, Inc) for mouse home-cage behavior classification and 2) between the system and human annotators. The level of agreement between human annotators sets a benchmark for the system performance since the system relies entirely on human annotations to learn to recognize behaviors. In order to evaluate the agreement between two sets of labelers, we asked a set of four human annotators ('*Annotator group 2'*) independent from '*Annotator group 1'* to annotate a subset of the *'full database'.* This subset (denoted *'set B'*) corresponds to many short random segments from the *'full database'*; each segment is about 5-10 min in length and they add up to a total of 1.6 hours of annotated video. Supplementary Fig. S1d shows the corresponding distribution of labels for *'set B'* and confirms that *'set B'* is representative of the '*full database'* (Supplementary Fig. S1c).

Performance was estimated using a leave-one-video-out procedure, whereby all but one of the videos was used to train the system while performance was evaluated on the remaining video. The procedure was repeated $n=12$ times for all videos and the performance. We found that our system achieves 76.6% agreement with human labelers on *'set B'* (averaged across frames), a result substantially higher than the HomeCageScan 2.0 (60.9%) system and on par with humans (71.6%), as shown in Table 1. For all of the comparisons above, the annotations made by the *'Annotator group 1'* were used as ground truth to train and test the system because these annotations underwent a second screening and were therefore more accurate than the annotations made by the *'Annotator group 2'*. The second set of annotations made by the *'Annotator group 2'* on *'set B'* was only used for measuring the agreement between independent human annotators. It is therefore possible for a computer system to appear more 'accurate' than the second group of annotators, which is in fact what we observed for our system. Table 1 also shows the comparison between the system and commercial software on the *'full database'*.

Fig. 3 shows the confusion matrices between the computer system and *'Annotator group 1'* (Fig. 3a), between *'Annotator group 1'* and *'Annotator group 2'* (Fig. 3b), and between the HomeCageScan system and *'Annotator group 1'* (Fig. 3c). A confusion matrix is one way to visualize the agreement between two entities, where each entry (x,y) of the matrix represents the probability that the first entity (say *'Annotator group 1'*) will label a specific behavior as x and the second entity (say the computer system) as y. For instance, two entities with perfect agreement would exhibit a 1 value along every entry along the diagonal and 0 everywhere else. In Fig. 3a for example, the matrix value along the fourth row and fourth column indicates that the computer system correctly classifies 92% of the 'hanging' behaviors as they labeled by a human observer while 8% of the behaviors are incorrectly classified as 'eating' (2%), rearing (5%) or others (less than 1%). These numbers are also reflected in the color codes used, with red/blue corresponding to better/worse levels of agreement. We also observed that adding the position- and velocity-based features led to an improvement in the system's ability to discriminate between visually similar behaviors and for which the location of the animal in the cage provides critical information (see Supplementary Methods and Supplementary Fig S3). For example, drinking (vs. eating) occurs at the water bottle spout while hanging (vs. rearing) mice have at least two limbs off the ground. Examples of automated scoring of videos by the system are available as Supplementary Movie 1 and Supplementary Movie 2.

How scalable is the proposed approach to new behaviors? How difficult would it be to train the proposed system for the recognition of new behaviors or environments (e.g., outside the home-cage and/or using a camera from a different view-point?) The main goal of the present study is to build a system that generalizes well to many different laboratory settings. For this reason, we collected and annotated a large dataset of videos. Sometimes, however, it might be advantageous to train a more 'specialized' system very quickly from very few training examples.

To investigate this issue, we systematically evaluated the performance of the system as a function of the amount of training videos available for training. Fig. 4a shows that a relatively modest amount of training data (i.e., as little as 2 minutes of labeled video for each of the eleven training videos) is indeed sufficient for robust performance. Additionally, in such cases where generalization is not required, an efficient approach would be to train the system on the first few minutes of a video and then let the system complete the annotation on the rest of the video. Fig. 4b shows that by using a representative set of only 3 minutes of video data, the system is already able to achieve 90% of its peak level. Near peak performance can be achieved by using 10 minutes of a single video for training.

Characterizing the home-cage behavior of mouse strains

To demonstrate the applicability of this vision-based approach to large-scale phenotypic analysis, we characterized the home-cage behavior of four strains of mice, including the wild-derived strain CAST/EiJ, the BTBR strain, a potential model of autism[4], as well as two of the most popular inbred mouse strains C57BL/6J and DBA/2J. We video recorded *n=7* mice of each strain during one 24-hour session, encompassing a complete light-dark cycle. An example of an ethogram containing all the eight behaviors obtained over a 24-hour continuous recording period for one of the CAST/-EiJ (wild-derived) strains is shown in Fig. 2h. One obvious feature was that the level of activity of the animal decreased significantly during the day (12-24 hr) as compared to night time (0-12hr). In examining the hanging and walking behaviors of the four strains, we noted a dramatic increase in activity of the CAST/EiJ mice during the dark phase, which show prolonged walking (Fig. 5a) and a much higher level of hanging activity (Fig. 5b) than any of the other strains tested. As compared to the CAST/EiJ mice, the DBA/2J strain showed an equally high level of hanging at the beginning of the dark phase but this activity quickly dampened to that of the other strains C57BL/6J and BTBR. We also found that the resting behavior of this CAST/EiJ strain differed significantly from the others: while all four strains tended to rest for the same total amount of time (except BTBR which rested significantly more than C57BL/6J), we found that the CAST/EiJ tended to have resting bouts (a continuous duration with one single label) that lasted almost three times longer than those of any other strain (Fig. 6a-b).

As BTBR mice have been reported to hyper-groom[4], we next examined the grooming behavior of BTBR mice. In the study of McFarlane et al.[4], grooming was scored manually during a 10-min session starting immediately after a 10-min habituation period following the placement of the animal in the new environment. Under the same conditions, our system detected that the BTBR strain spent approximately 150 seconds grooming compared to the C57BL/6J mice, which spent a little more than 85 seconds grooming (Fig. 6c). This behavioral difference was reproduced by two more human observers ('H' and 'A') who scored the same videos (Fig. 6c). Using annotator 'H' as ground truth, the frame-based accuracy of the system vs. annotator 'A' was 90% vs. 91.0%. This shows that the system can reliably identify grooming behaviors with nearly the same accuracy as a human annotator. Note that in the present study the C57BL/6J mice were approximately 90 days old (+/- 7 days) while the

BTBR mice were approximately 80 days old (+/-7 days). In the McFarlane *et al.* study younger mice were used (and repeated testing was performed), but our results essentially validate their findings.

Prediction of strain-type based on behavior

We characterized the behavior of each mouse with a 32-dimensional vector called the 'pattern of behavior', corresponding to the relative frequency of each of the eight behaviors of interest, as predicted by the system, over a 24-hour period. To visualize the similarities/dissimilarities between patterns of behaviors exhibited by all twenty-eight individual animals (7 mice × 4 strains) used in our behavioral study, we performed a Principal Component Analysis (PCA). Fig. 6d shows the resulting twenty-eight data-points, each corresponding to a different animal, projected onto the first three principal components. Individual animals tend to cluster by strains even in this relatively low dimensional space, suggesting that different strains exhibit unique patterns of behaviors that are characteristic of their strain-types. To quantify this statement, we trained and tested a linear SVM classifier directly on these patterns of behaviors. Fig. 6e shows a confusion matrix for the resulting classifier that indicates the probability with which an input strain (along the rows) was classified as each of the 4 strains (along the columns). The higher probabilities along the diagonal and the lower off-diagonal values indicate successful classification for all strains. Using a leave-one-animal-out procedure, we found that the resulting classifier was able to predict the strain of all animals with an accuracy of 90%.

Application of the system to additional mouse behaviors

We next asked whether the proposed system could be extended to the recognition of additional behaviors beyond the eight standard behaviors described above. We collected a new set of videos for an entirely new set of behaviors corresponding to animals interacting with "low profile" running wheels (Fig. 7a). The "*wheel-interaction database*" contains thirteen fully annotated one-hour long videos taken from six C57BL/6J mice. Here we consider four actions of interest: "running on the wheel" (defined by having all 4 paws on the wheel, with the wheel rotating), "interacting with the wheel but not running" (any behavior on the wheel other than running), "awake but not interacting with the wheel", and "resting outside the wheel". Using the same leave-one-video-out procedure and accuracy formulation as used before for the *"full database"*, the system achieves 93% accuracy. The confusion matrix shown in Fig. 7b indicates that the system can discriminate between visually similar behaviors such as "interacting with the wheel but not running" and "running on the wheel" (see also

Supplementary Movie 3 for a demonstration of the system scoring the wheel-interaction behaviors). In order to understand how many annotated examples are required to reach this performance, we repeated the same experiment, each time varying the number of training examples available to the system. Fig. 4c suggests that satisfactory performance can be achieved with only 2 minutes of annotation for each training video, corresponding to 90% of the performance obtained using 30 minutes of annotations. Fig. 4d shows that training with very short segments collected from a single video seems sufficient for robust performance on the "*wheel-interaction database*" but, unlike for the eight standard home-cage behaviors, the system performance increases linearly with the number of training examples. This might be due to the large within-class variation of the action "awake but not interacting with wheel", which combines all of the actions that are performed outside the wheel, such as walking, grooming, eating, and rearing, within one single category.

Discussion

Here we describe a trainable computer vision system capable of capturing the behavior of a single mouse in the home-cage environment. As opposed to previous proof-of-concept computer vision studies[22, 23], our system has been used in a "real-world" application, characterizing the behavior of several mouse strains and discovering strain-specific features. Moreover, we demonstrate that this system adapts well to more complex environments and behaviors that involve additional objects placed in the home-cage. We provide open-source software as well as large annotated video databases with the hope that it may further encourage the development and benchmarking of similar vision-based systems.

Genetic approaches to understand the neural basis of behavior require cost effective and high-throughput methodologies to find aberrations in normal behaviors[30]. From the manual scoring of mouse videos (see *'full database'* above), we have estimated that it requires approximately 22 person-hours to manually score every frame of a one-hour video with high stringency. Thus, we estimate that the 24-hour behavioral analysis conducted above with our system for the twenty-eight animals studied would have required almost 15,000 person-hours of manual scoring. An automated computer-vision system permits behavioral analysis that would simply be impossible using manual scoring by a human experimenter. By leveraging recent advances in graphics processing hardware and exploiting the high-end graphical processing units (GPU) available on modern computers, the current system runs in real-time.

In principle, our approach should be extendable to other behaviors such as dyskinetic movements in the study of Parkinson's disease models or seizures for the study of epilepsy as well as social behaviors involving two or more freely behaving animals. In conclusion, our study shows the promise of learning-based and vision-based techniques in complementing existing approaches towards a quantitative phenotyping of complex behavior.

METHODS

Mouse strains and behavioral experiment

All experiments involving mice were approved by the MIT and Caltech committees on animal care. For generating training and testing data we used a diverse pool of hybrid and inbred mice of varying size, age, gender, and coat color (both black and agouti coat colors). In addition, we varied the lighting angles and used both 'light' and 'dark' recording conditions (with a 30 Watt bulb dim red lighting to allow our cameras to detect the mice but without substantial circadian entrainment effects.) A JVC digital video camera (GR-D93) with frame rate 30 fps was connected to a PC workstation (Dell) via a Hauppauge WinTV video card. Using this setup we collected more than 24 distinct MPEG-2 video sequences (from one to several hours in length) used for training and testing the system. For processing by the computer vision system, all videos were down-sampled to a resolution of 320 × 240 pixels. The throughput of the system could thus be further increased by video recording 4 cages at a time using a two-by-two arrangement with a standard 640x480 pixel VGA video resolution.

Videos of the mouse strains (*n=28* videos) were collected separately for the validation experiment, using different recording conditions (recorded in a different mouse facility). All mouse strains were purchased from the Jackson Laboratory (Bar Harbor, Maine), including C57BL/6J (stock 000664), DBA/2J (000671), CAST/EiJ (000928), and BTBR *T+tf*/J (002282). Mice were singly housed for 1-3 days before being video recorded. On the recommendation of Jackson Laboratories, the CAST/EiJ mice (*n=7*) were segregated from our main mouse colony and housed in a quiet space where they were only disturbed for husbandry 2-3 times per week. This may have influenced our behavioral measurements as the other three mouse strains were housed in a different room. The mice used for the running wheel study were 3-month-old C57BL/6J males also obtained from Jackson labs.

Data annotation

Training videos were annotated using a freeware subtitle-editing tool, Subtitle Workshop by UroWorks (available at http://www.urusoft.net/products.php?cat=sw&lang=1). A team of eight investigators (*'Annotator group 1'*) was trained to annotate eight typical mouse home-cage behaviors. The four annotators in the *'Annotator group 2'* were randomly selected from the '*Annotator group 1*' pool. Behaviors of interest included: drinking (defined by the mouse's mouth being juxtaposed to the tip of the drinking spout), eating (defined by the mouse reaching and acquiring food from the food bin), grooming (defined by the fore- or hind-limbs sweeping across the face or torso, typically as the animal is reared up), hanging (defined by grasping of the wire bars with the fore-limbs and/or hind-limbs with at least two limbs off the ground), rearing (defined by an upright posture and forelimbs off the ground), resting (defined by inactivity or nearly complete stillness), walking (defined by ambulation) and micro-movements (defined by small movements of the animal's head or limbs). For the *'full database'* to be annotated, every hour of videos took about 22 hours of labor for a total of 264 hours of work. For the *'clipped database'* it took approximately 110 hours (9 hrs/hr of video) to manually score 9,600 clips of a single behavior (corresponding to 5.4 hours of clips compiled from around 20 hours of video). We performed secondary screening to remove ambiguous clips, leaving 4,200 clips for which the human-to-human agreement is very close to 100%. This second screening took around 25 hours for the 2.5 hour long *'clipped database'*. Supplementary Fig. S1a-b and Supplementary Fig. S1c-d shows the distribution of labels for the *'clipped database'* and the *'full database'*, respectively.

Training and testing the system

The evaluation on the *'full database'* and *'set B'* shown in Table 1 was obtained using a leave-one-out cross-validation procedure. This consists in using all but one of the videos to train the system and using the left out video to evaluate the system; repeating this procedure ($n = 12$) times for all videos. System predictions for all the frames are then concatenated to compute the overall accuracy as: (total # frames correctly predicted by the system)/(total # frames) and the human-to-human agreement as: (total # frames correctly labeled by '*Annotator group 2*')/(total # frames). Here a prediction or label is considered 'correct' if/when it matches the annotations generated by the '*Annotator group 1*'. Such a procedure provides the best estimate of the future performance of a classifier and is standard in computer vision. This guarantees that the system is not just recognizing memorized examples but generalizing to previously unseen examples. For the *'clipped database'*, a leave-one-video-out procedure is used whereby clips from all except one video are used to train the system while testing is performed on clips of the remaining video. This procedure is repeated ($n = 12$) times for all videos. A single prediction is obtained for each clip (each clip has a single annotated label) via voting across frames, and predictions for all the clips of all videos are then concatenated to compute the overall accuracy as (# total clips corrected predicted by the system) / (# total clips).

In addition to measuring the overall performance of the system as above, we also used a confusion matrix to visualize the system's performance on each behavioral category in Fig. 3 and Fig. 6e. A confusion matrix is a common visualization tool used in multi-class classification problems[23]. Each row of the matrix represents a true class, and each column represents a predicted class. Each entry (x,y) in the confusion matrix is the probability that an instance of behavior x (along the rows) will be classified as instance of behavior y (along the columns), as computed by (# frames annotated as type x and classified as type y)/(# frames annotated as type x). Here the frame predictions are obtained by concatenating the predictions for all videos, as described above. The higher probabilities along the diagonal and the lower off-diagonal values indicate successful classification for all behavioral types. For example, in Fig. 3a, 94% of the frames annotated as 'rest' are classified correctly by the system, and 5% are misclassified as 'groom'.

Comparison with the commercial software

In order to compare the proposed system with available commercial software, the HomeCageScan 2.0 (CleverSys Inc), we manually matched the thirty-eight output labels from the HomeCageScan to the eight behaviors used in the present work. For instance, actions such as 'slow walking', 'walking left' and 'walking right' were all re-assigned to the 'walking' label to match against our annotations. With the exception of very few behaviors (e.g., 'jump', 'turn' and 'unknown behavior'), we were able to match all HomeCageScan output behaviors to one of our eight behaviors of interest (see Supplementary Table S2 for a listing of the correspondences used between the labels of the HomeCageScan and our system). It is possible that further fine-tuning of HomeCageScan parameters could have improved upon the accuracy of the scoring.

Statistical analysis

To detect differences among the four strains of mice, ANOVAs were conducted for each type of behavior independently and Tukey's post-hoc test was used to test pair-wise significances. All post-hoc tests were Bonferroni corrected for multiple comparisons. For the grooming behavior, a one-tailed Student's T test was used since only two groups (C57BL/6J and BTBR) were being compared and we had predicted that BTBR would groom more than C57BL/6J.

Mouse strain comparisons

Patterns of behaviors were computed from the system output by segmenting the system predictions for a 24-hour video into four non-overlapping 6-hour long segments (corresponding to the first and second halves of the day and night periods) and calculating a histogram for the eight types of behaviors for each video segment. The resulting 8-dimensional (one dimension for each of the eight behaviors) vectors were then concatenated to obtain a single 32-dimensional vector (8 dimensions × 4 vectors) for each animal. To visualize the data, we performed a Principal Component Analysis directly on these 32-dimensional vectors.

In addition, we conducted a pattern classification analysis on the patterns of behaviors by training and testing an SVM classifier directly on the 32-dimensional vectors. This supervised procedure was conducted using a leave-one-animal out approach, whereby twenty-seven animals were used to train a classifier to predict the strain of the remaining animal (CAST/EiJ, BTBR, C57BL/6J or DBA/2J). The procedure was repeated twenty eight times (once for each animal). Accuracy measures for the four strain predictions was then computed as (# animals correctly classified) / (# animals).

ACKNOWLEDGMENT

This project was sponsored by the McGovern Institute for Brain Research. ADS, LY, and VK were funded by the Broad Fellows Program in Brain Circuitry at Caltech. HJ was founded by the Taiwan National Science Council (TMS-094-1-A032). We wish to thank Jessi Ambrose, Andrea Farbe, Cynthia Hsu, Alexandra Jiang, Grant Kadokura, Xuying Li, Anjali Patel, Kelsey von Tish, Emma Tolley, Ye Yao and Eli T. Williams for their efforts in annotating the videos for this project. We are grateful to Susan Lindquist for allowing us to use her facilities and mice to generate much of the training video database and to Walker Jackson for assistance with operating HCS. We wish to thank Piotr Dollar for providing source code. The algorithm and software development and its evaluation was conducted at the Center for Biological & Computational Learning, which is in the McGovern

Institute for Brain Research at MIT, as well as in the Dept. of Brain & Cognitive Sciences, and which is affiliated with the Computer Sciences & Artificial Intelligence Laboratory (CSAIL).

CONTRIBUTIONS

H.J., T.P., A.D.S. and T.S. designed research; H.J., E.G., A.D.S. and T.S. conducted research; H.J., E.G., X.Y., V.K., A.D.S. and T.S analyzed data; H.J., T.P., A.D.S. and T.S. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

REFERENCES

1. Auwerx, J. & El, A. The European dimension for the mouse genome mutagenesis program. *Nat. Genet.* 36, 925-927 (2004).
2. Crabbe, J.C., Wahlsten, D. & Dudek, B.C. Genetics of mouse behavior: interactions with laboratory environment. *Science* 284, 1670-1672 (1999).
3. Greer, J.M. & Capecchi, M.R. Hoxb8 Is Required for Normal Grooming Behavior in Mice. *Neuron* 33, 23-34 (2002).
4. Mcfarlane, H.G. et al. Autism-like behavioral phenotypes in BTBR T1tf/J mice. *Genes Brain Behav.* 7, 152-163 (2008).
5. Roughan, J.V., Wright-Williams, S.L. & Flecknell, P.A. Automated analysis of postoperative behaviour: assessment of HomeCageScan as a novel method to rapidly identify pain and analgesic effects in mice. *Lab. Anim.* 43, 17-26 (2008).
6. Chen, D., Steele, A.D., Lindquist, S. & Guarente, L. Increase in Activity During Calorie Restriction Requires Sirt1. *Science* 310, 1641 (2005).
7. Steele, A.D., Jackson, W.S., King, O.D. & Lindquist, S. 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.* 104, 1983-1988 (2007).
8. Goulding, E.H. et al. A robust automated system elucidates mouse home cage behavioral structure. *Proc. Natl Acad. Sci.* 105, 20575-20582 (2008).
9. Dell'Omo, G. et al. Early behavioural changes in mice infected with BSE and scrapie: automated home cage monitoring reveals prion strain differences. *Eur. J. Neurosci.* 16, 735-742 (2002).
10. Steele, A.D. et al. Heat shock factor 1 regulates lifespan as distinct from disease onset in prion disease. *Proc. Natl Acad. Sci.* 105, 13626-13631 (2008).
11. Jackson, W.S., Tallaksen-greene, S.J., Albin, R.L. & Detloff, P.J. Nucleocytoplasmic transport signals affect the age at onset of abnormalities in knock-in mice expressing polyglutamine within an ectopic protein context. *Hum. Mol. Gen.* 12, 1621-1629 (2003).
12. Noldus, L.P., Spink, A.J. & Tegelenbosch, R.A. EthoVision: a versatile video tracking system for automation of behavioral experiments. *Behav. Res. Meth. Ins. C.* 33, 398-414 (2001).
13. Rudenko, O., Tkach, V., Berezin, V. & Bock, E. Detection of early behavioral markers of Huntington's disease in R6/2 mice employing an automated social home cage. *Behav. Brain Res.* 203, 188-199 (2009).
14. Dalal, N. & Triggs, B. Histograms of oriented gradients for human detection. *Proc. IEEE Comp. Vision and Patt. Recogn.* (2005).

15. Viola, P. & Jones, M. Robust Real-time Object Detection. *Proc. IEEE Int. Conf. on Comp. Vision* (2001).
16. Moeslund, T.B., Hilton, A. & Kruger, V. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.* 104, 90-126 (2006).
17. Veeraraghavan, A., Chellappa, R. & Srinivasan, M. Shape-and-behavior encoded tracking of bee dances. *IEEE Trans. Pattern Anal. Mach. Intell.* 30, 463–476 (2008).
18. Fry, S.N., Rohrseitz, N., Straw, A.D. & Dickinson, M.H. TrackFly: virtual reality for a behavioral system analysis in free-flying fruit flies. *J. Neurosci. Methods* 171, 110-117 (2008).
19. Khan, Z., Balch, T. & Dellaert, F. MCMC-Based Particle Filtering for Tracking a Variable Number of Interacting Targets. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1805-1819 (2005).
20. Branson, K., Robie, A.A., Bender, J., Perona, P. & Dickinson, M.H. High-throughput ethomics in large groups of Drosophila. *Nat. Methods* 6, 451-457 (2009).
21. Dankert, H., Wang, L., Hoopfer, E.D., Anderson, D.J. & Perona, P. Automated monitoring and analysis of social behavior in Drosophila. *Nat. Methods* 6, 297-303 (2009).
22. Xue, X. & Henderson, T.C. Feature fusion for basic behavior unit segmentation from video sequences. *Robot. Auton. Syst.* 57, 239-248 (2009).
23. Dollar, P., Rabaud, V., Cottrell, G. & Belongie, S. Behavior recognition via sparse spatio-temporal features. *Proc. IEEE Int. Workshop on VS-PETS* (2005).
24. Giese, M.A. & Poggio, T. Neural Mechanisms for the Recognition of Biological Movements. *Nat. Rev. Neurosci.* 4, 179-192 (2003).
25. Jhuang, H., Serre, T., Wolf, L. & Poggio, T. A biologically inspired system for action recognition. *Proc. IEEE Int. Conf. on Comp. Vision* (2007).
26. Born, R.T. & Bradley, D.C. Structure and Function of Visual Area MT. *Annu. Rev. Neurosci.* 28, 157-189 (2005).
27. Simoncelli, E.P. & Heeger, D.J. A model of neuronal responses in visual area MT. *Vision Res.* 38, 743-61 (1998).
28. Altun, Y., Tsochantaridis, I. & Hofmann, T. Hidden Markov Support Vector Machines. *Proc. Int. Conf. on Mach. Learn.* (2003).
29. Joachims, T., Finley, T. & Yu, C.-n.J. Cutting-Plane Training of Structural SVMs. *Mach. Learn.* 76, 27-59 (2009).
30. Tecott, L.H. & Nestler, E.J. Neurobehavioral assessment in the information age. *Nat. Neurosci.* 7, 462-466 (2004).

Figure Legends

Figure 1. Home-cage behaviors for training the system. Snapshots taken from representative videos for the eight home-cage behaviors of interest.

Figure 2. Overview of the proposed system for monitoring the home-cage behavior of mice. The computer vision system consists of a feature computation module (a-f) and a classification module (g). (a) A background subtraction procedure is first applied to an input video to compute a foreground mask for pixels belonging to the animal vs. the cage. (b) A sub-window centered on the animal is cropped from each video frame based on the location of the mouse (see Supplementary Method). Two types of features are then computed: (c) space-time motion features as well as (f) position- and velocity-based features. In order to speed-up the computation, motion-features are extracted from the sub-window (b) only. These motion features are derived from combinations of the response of V1-like afferent motion units that are tuned to different directions of motion (d, e). (f) Position- and velocity-based features are derived from the instantaneous location of the animal in a cage. These features are computed from a bounding box tightly surrounding the animal in the foreground mask. (g) The output of this feature computation module consists of 310 features per frame that are then passed to a statistical classifier, an SVMHMM (Hidden Markov Model Support Vector Machine), to reliably classify every frame of a video sequence into a behavior of interest. (h) Ethogram of sequence of labels predicted by the system from a 24-hr continuous recording session for one of the CAST/EiJ mice. The red panel shows the ethogram for 24-hours, and the light blue panel provides a zoomed version corresponding to the first 30 minutes of recording. The animal is highly active as it was just placed in a new cage prior to starting the video recording. The animal's behavior alternates between 'walking', 'rearing' and 'hanging' as it explores the new cage.

Figure 3. Confusion matrix of the system. Confusion matrices evaluated on the doubly annotated *'set B'* to compare the agreement between (a) the system and human scoring, (b) human to human scoring, and (c) the CleverSys system to human scoring. Each entry *(x,y)* in the confusion matrix is the probability with which an instance of a behavior *x* (along rows) is classified as type *y* (along column), and which is computed as (# frames annotated as type *x* and classified as type *y*) / (# frames annotated as type *x*). As a result, values sum to a value of 1 in each row. The higher probabilities along the diagonal and the lower off-diagonal values indicate successful classification for all categories. Using

the annotations made by the '*Annotator group 1*' as ground truth, the confusion matrix was obtained for measuring agreement between ground truth (row) with system (computer system), with the '*Annotator group 2'* (human) and with baseline software (CleverSys commercial system). For a less cluttered visualization, entries with values less than 0.01 are not shown. The color bar indicates the percent agreement, with more intense shades of red indicating agreements close to 100% and lighter shades of blue indicating small percentages of agreement.

Figure 4: Training the system with varying numbers of examples. For this leave-one-video-out experiment, the system accuracy is computed as a function of the amount of video data (in minutes/video) used for training the system. For each leave-one-out trial, the system is trained on a representative set of videos and tested on the full length of the left-out video. A representative set consisting of x 1-minute segments is manually selected such that the total time of each of the 8 behaviors is roughly the same. (a) Average accuracy and standard error across the 12 leave-one-out runs on the '*full database'.* (b) Average accuracy and standard error across the 13 leave-one-out runs on the '*wheel-interaction database'.* We also perform the training/testing on the same video: a representative set of video segments is selected from the first 30 minute of each video for training; testing is done on the remaining of the video from the 30th minute to the end of the same video. System accuracy is computed as a function of the amount of video data (in minute) used for training. (c) Average accuracy and standard error across the 12 runs on the '*full database*'. (d) Average accuracy and standard error across the 13 runs on the '*wheel-interaction database'*.

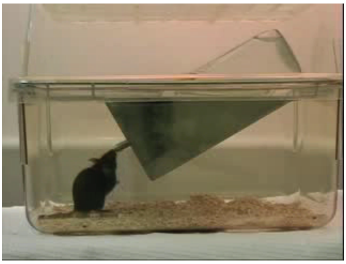
Figure 5. Walking and hanging behaviors for the four mouse strains. Average time spent for (a) the 'walking' and (b) 'hanging' behaviors for each of the four strains of mice ($n=7$ animals for each strain) over a 20-hour period. The plots begin at the onset of the dark cycle, which persists for 11 hours (indicated by the gray region), followed by 9 hours of the light cycle. Every 15 minute of the 20-hour period, we computed the total time one mouse spent walking or hanging within a one-hour temporal window centered at that current time point. The CAST/EiJ (wild-derived) strain is much more active than the three other strains as measured by their walking and hanging behaviors. Shaded areas around the curves correspond to 95% confidence intervals and the darker curve corresponds to the mean. The colored bars indicate the duration when one strain exhibits a statistically significant difference (*p<0.01 by ANOVA with Tukey's post test) with other strains.

Figure 6. Behavioral characterization of four mouse strains. (a) Average total resting time for each of the four strains of mice over 24 hours (*n=7* animals for each strain). (b) Average duration of resting bouts (defined as a continuous duration with one single behavior). While all strains tend to spend roughly the same total amount of time resting (a), the CAST/EiJ tends to rest for longer stretches. Mean +/- SEM are shown, *p<0.01 by ANOVA with Tukey's post test. (c) Total time spent grooming exhibited by the BTBR strain as compared to the C57BL/6J strain within 10th-20th minute after placing the animals in a novel cage. Here we show that using the computer system we were able to match manual scoring by two experimenters and reproduce previously published results[4] demonstrating the propensity of the BTBR strain to groom more than a control C57BL/6J. Mean +/- SEM are shown, *p<0.05 by Student's T test, one-tailed. p = 0.04 for System and p =0.0254 for human 'H', p = 0.0273 for human 'A'). (d-e) Characterizing the genotype of individual animals based on the patterns of behavior measured by the computer system. The pattern of behaviors for each animal is a 32-dimensional vector, corresponding to the relative frequency of each of the eight behaviors of interest, as predicted by the system, over a 24-hour period. (d) To visualize the similarities/dissimilarities between patterns of behaviors exhibited by all twenty-eight individual animals (seven mice for each of the × four strains) used in our behavioral study, we performed a Principal Component Analysis (PCA) on the patterns of behaviors. The result shows that animals tend to cluster by strain (with the exception of 2 BTBR mice that tended to behave more like DBA/2J). (e) Confusion matrix for an SVM classifier trained on the patterns of behavior using a leave-one-animal-out procedure. The SVM classifier is able to predict the genotype of individual animals with an accuracy of 90% (chance level is 25% for this 4-class classification problem). The confusion matrix shown here indicates the probability for an input strain (along the rows) to be classified, based on its pattern of behavior, as each of the four alternative strains (along the columns). The higher probabilities along the diagonal and the lower off-diagonal values indicate successful classification for all categories. For example, the value of 1.0 for the C57BL/6J strain means that all C57BL/6J animals were correctly classified as such.

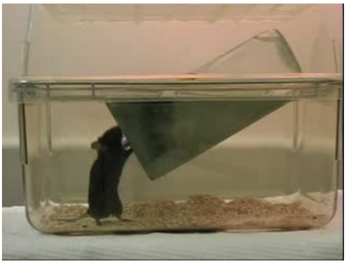
Figure 7: Extension of the system to wheel running and investigatory behaviors. (A) Snapshots taken from the *"wheel-interaction database"* for the four types of interaction behaviors of interest: resting outside of the wheel, awake but not interacting with the wheel, running on the wheel, and interacting with (but not running on) the wheel. (B) Confusion matrices for the system (column) vs. human scoring (row).

Table 1: Accuracy of the System. Shown is a comparison between the performance of the proposed system, a leading commercial software (HomeCageScan 2.0 by CleverSys system) and human annotators. Training and testing of the system was based on a leave-one-video-out procedure. This consists in using all but one of the videos to train the system and using the left out video to evaluate the system; repeating this procedure (*n=12*) times for all videos. Here a prediction or label is considered 'correct' if/when it matches the annotations generated by the '*Annotator group 1*'. Accuracies are reported as averaged across frames/ across behaviors (underlined numbers, computed as the average of the diagonal entities in the Fig. 3 confusion matrix; chance level is 12.5% for a 8-class classification problem). The somewhat better performance of the proposed system averaged across frames vs. behaviors suggest that it does better on the most common behaviors (as expected from the training procedure).

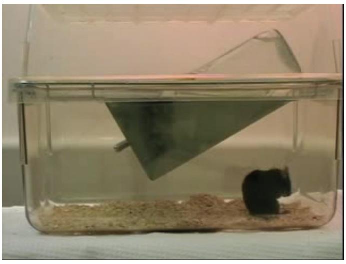
	Our system	CleverSys commercial system	Human ('*Annotator group 2*')
'*set B*' (1.6 hours of video)	76.6 % / 74.3%	60.9 % / 64.0%	71.6 % / 75.7%
'*full database*' (over 10 hours of video)	77.6 % / 74.4%	61.0 % / 65.8%	



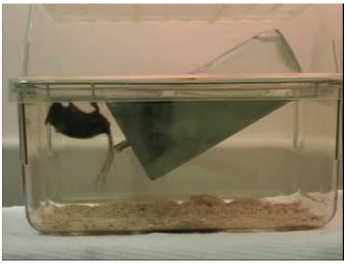
drink



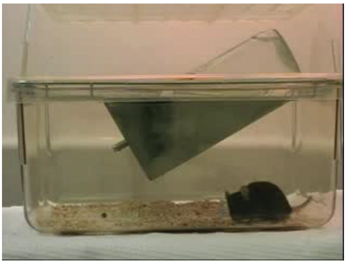
eat



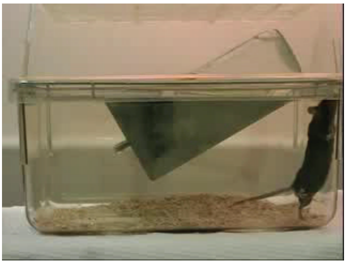
groom



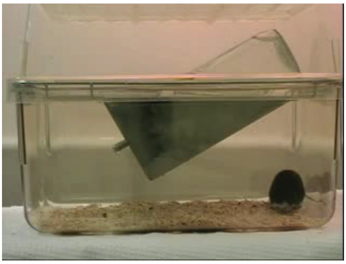
hang



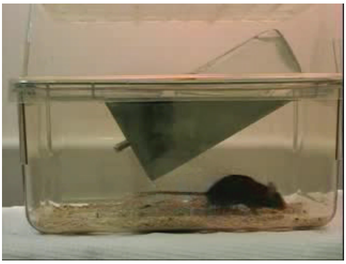
micro-movement



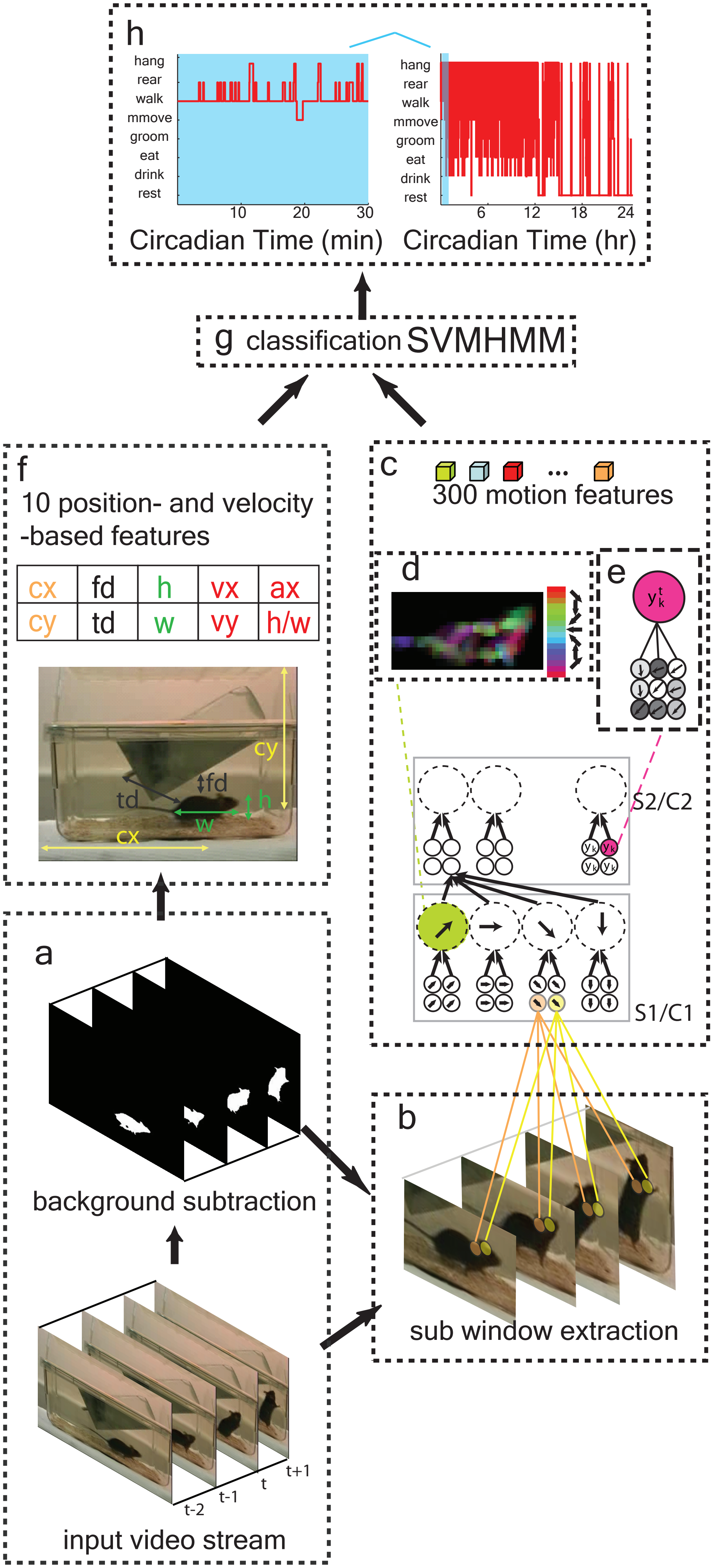
rear

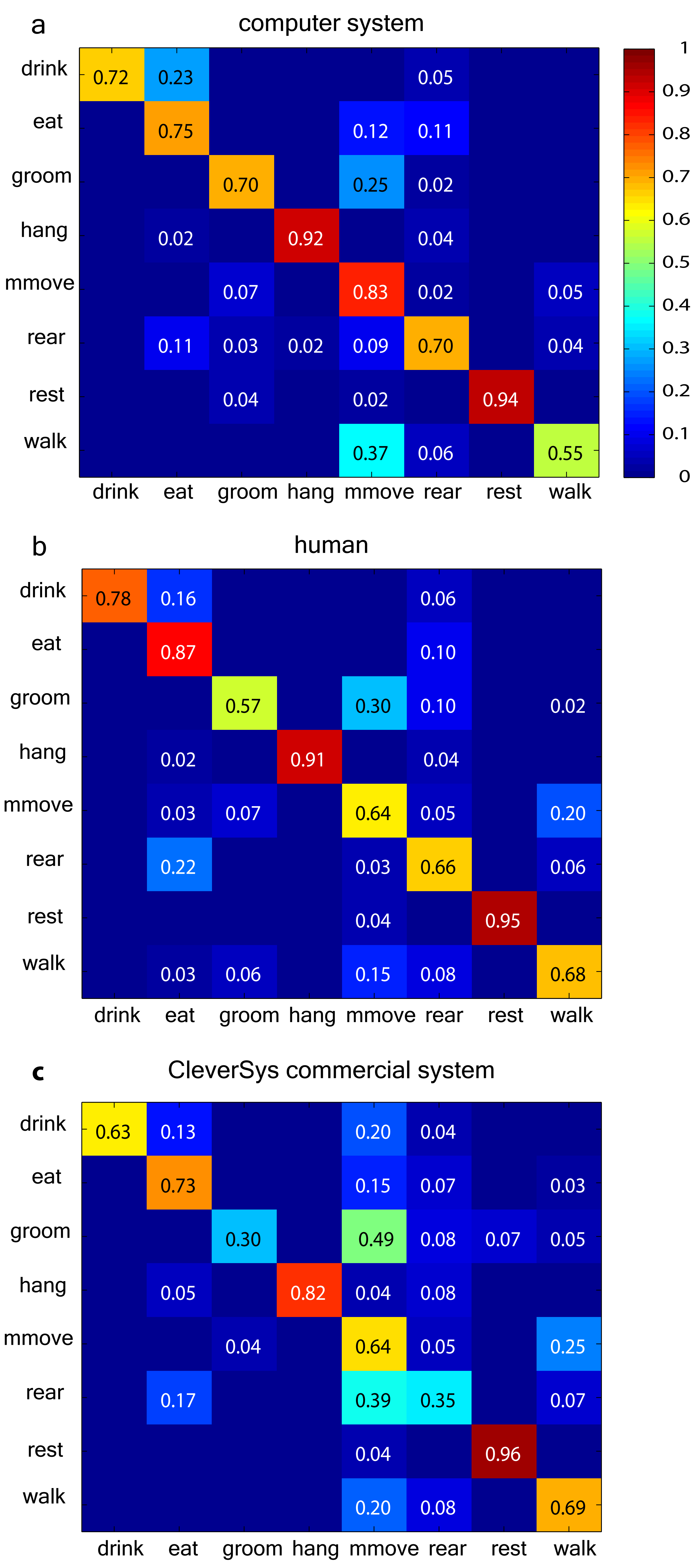


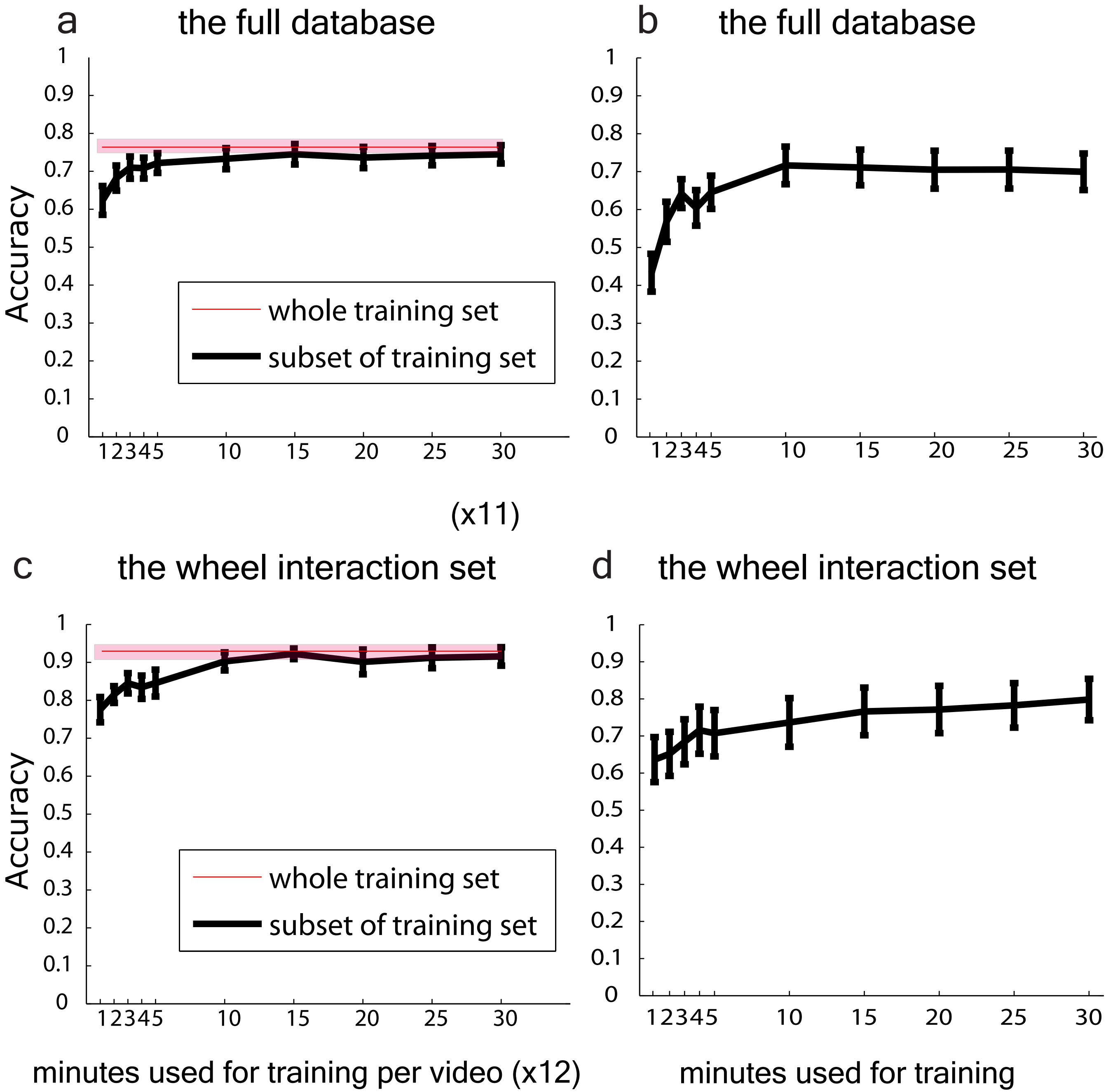
rest

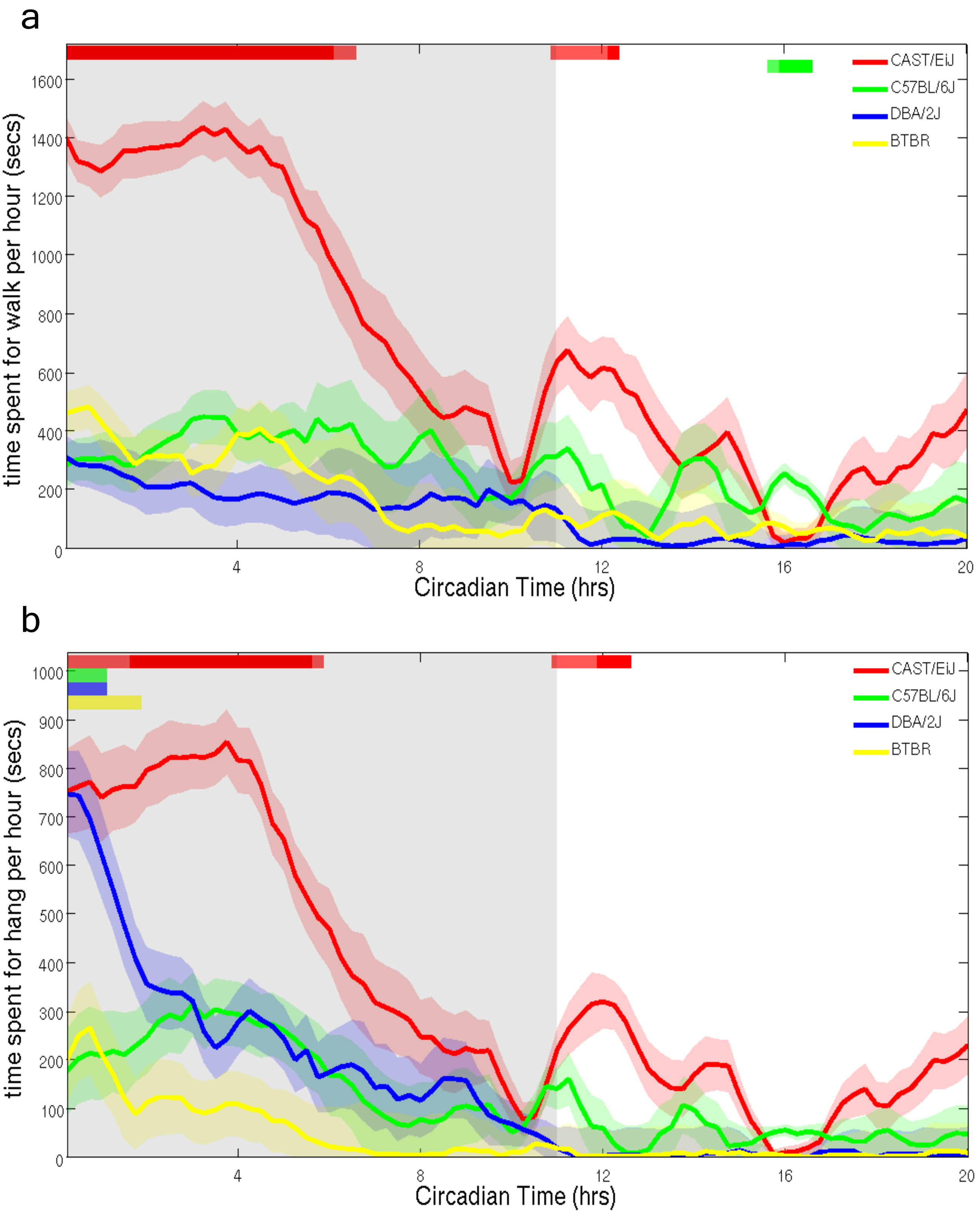


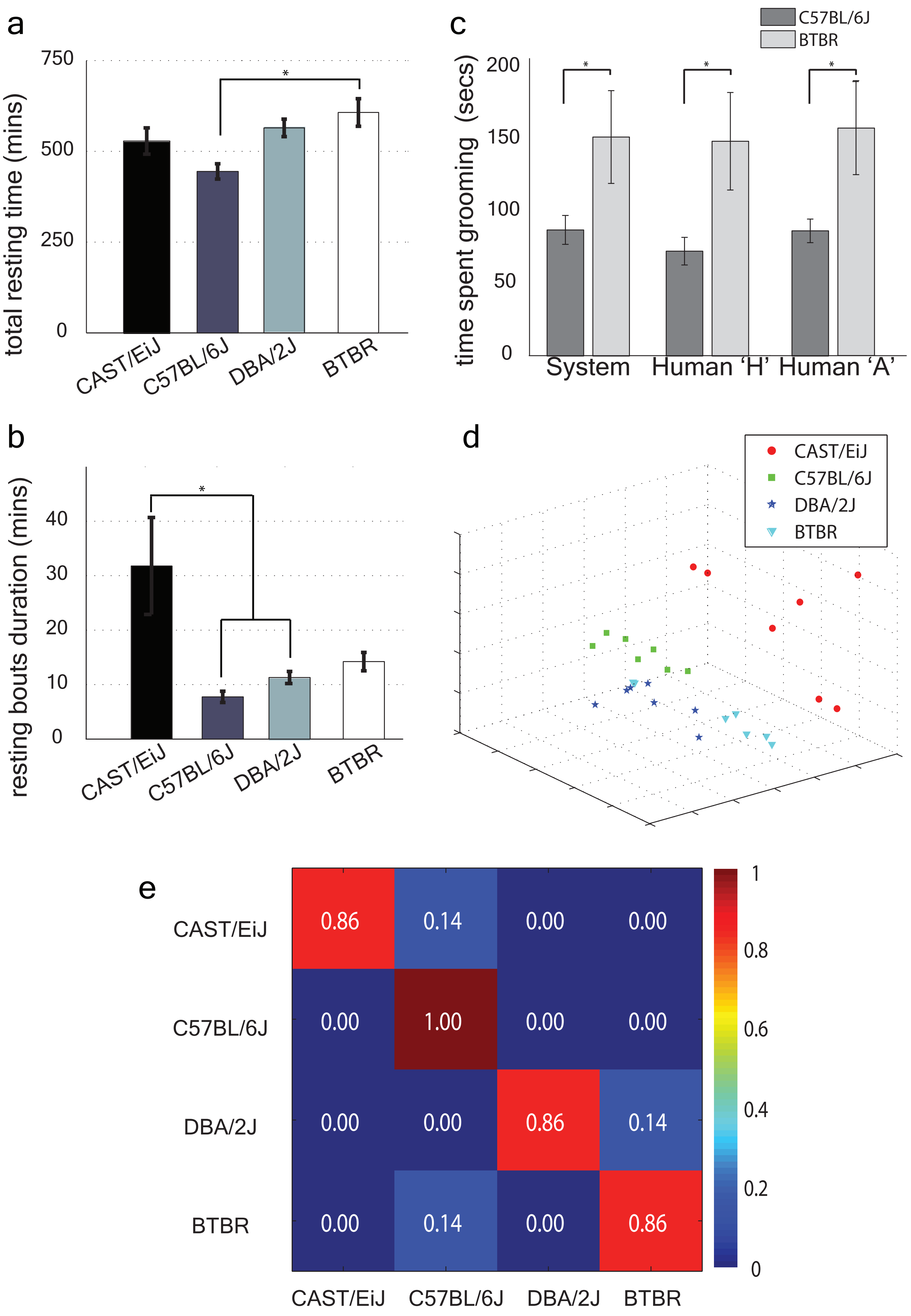
walk

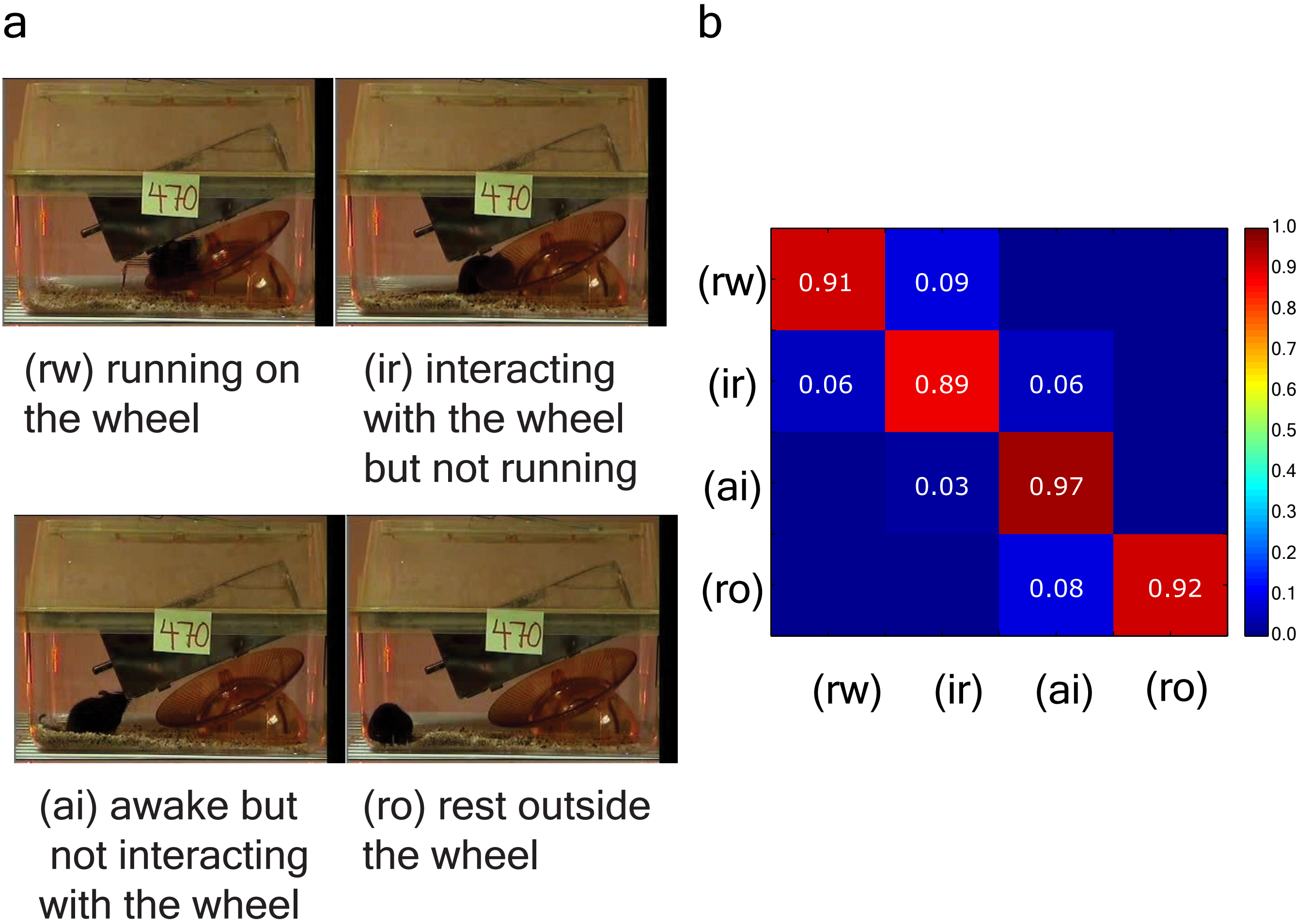

h
hang
rear
walk
mmove
groom
eat
drink
rest
10
20
30
Circadian Time (min)
hang
rear
walk
mmove
groom
eat
drink
rest
6
12
18
24
Circadian Time (hr)
g classification SVMHMM
f
10 position- and velocity -based features
cx
fd
h
vx
ax
cy
td
w
vy
h/w
cy
fd
td
h
w
cx
c
300 motion features
d
e
y_k^t
S2/C2
y_k
y_k
y_k
y_k
S1/C1
a
background subtraction
t-2
t-1
t
t+1
input video stream
b
sub window extraction


a
computer system
drink 0.72 0.23 0.05
eat 0.75 0.12 0.11
groom 0.70 0.25 0.02
hang 0.02 0.92 0.04
mmove 0.07 0.83 0.02 0.05
rear 0.11 0.03 0.02 0.09 0.70 0.04
rest 0.04 0.02 0.94
walk 0.37 0.06 0.55
drink eat groom hang mmove rear rest walk
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
b
human
drink 0.78 0.16 0.06
eat 0.87 0.10
groom 0.57 0.30 0.10 0.02
hang 0.02 0.91 0.04
mmove 0.03 0.07 0.64 0.05 0.20
rear 0.22 0.03 0.66 0.06
rest 0.04 0.95
walk 0.03 0.06 0.15 0.08 0.68
drink eat groom hang mmove rear rest walk
c
CleverSys commercial system
drink 0.63 0.13 0.20 0.04
eat 0.73 0.15 0.07 0.03
groom 0.30 0.49 0.08 0.07 0.05
hang 0.05 0.82 0.04 0.08
mmove 0.04 0.64 0.05 0.25
rear 0.17 0.39 0.35 0.07
rest 0.04 0.96
walk 0.20 0.08 0.69
drink eat groom hang mmove rear rest walk


a
the full database
b
the full database
Accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
whole training set
subset of training set
1 2 3 4 5
10
15
20
25
30
(x11)
c
the wheel interaction set
d
the wheel interaction set
whole training set
subset of training set
minutes used for training per video (x12)
minutes used for training


a
time spent for walk per hour (secs)
1600
1400
1200
1000
800
600
400
200
0
4
8
12
16
20
Circadian Time (hrs)
CAST/EiJ
C57BL/6J
DBA/2J
BTBR
b
time spent for hang per hour (secs)
1000
900
800
700
600
500
400
300
200
100
0
4
8
12
16
20
Circadian Time (hrs)
CAST/EiJ
C57BL/6J
DBA/2J
BTBR


a
total resting time (mins)
750
500
250
0
*
CAST/EiJ
C57BL/6J
DBA/2J
BTBR
b
resting bouts duration (mins)
40
30
20
10
0
*
CAST/EiJ
C57BL/6J
DBA/2J
BTBR
c
C57BL/6J
BTBR
time spent grooming (secs)
200
150
100
50
0
*
*
*
System
Human 'H'
Human 'A'
d
● CAST/EiJ
■ C57BL/6J
★ DBA/2J
▽ BTBR
e
CAST/EiJ
C57BL/6J
DBA/2J
BTBR
0.86
0.14
0.00
0.00
0.00
1.00
0.00
0.00
0.00
0.00
0.86
0.14
0.00
0.14
0.00
0.86
CAST/EiJ
C57BL/6J
DBA/2J
BTBR
1
0.8
0.6
0.4
0.2
0



(rw) running on the wheel

(ir) interacting with the wheel but not running

(ai) awake but not interacting with the wheel

(ro) rest outside the wheel