Vision-Based Recognition of Mice Home-Cage Behaviors

H. Jhuang E. Garrote N. Edelman T. Poggio A. Steele* T. Serre†
McGovern Institute for Brain Research, Massachusetts Institute of Technology
*Broad Fellows in Brain Circuitry Program, Division of Biology, California Institute of Technology
corresponding authors: † serre@mit.edu *steelea@caltech.edu

Abstract

We describe a trainable computer vision system enabling the automated analysis of complex mouse behaviors. We also collect and manually annotate a very large video database used for training and testing the system. Our system performs on par with human scoring, as measured from the ground-truth manual annotations. Our video-based software should complement existing sensor based automated approaches and help develop an adaptable, comprehensive, high-throughput, fine-grained, automated analysis of mouse behavior.

1. Introduction

Automated quantitative analysis of mouse behavior will play a significant role in comprehensive phenotypic analysis - both on the small scale of detailed characterization of individual gene mutants and on the large scale of assigning gene functions 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, it allows for addressing an entirely new set of questions about mouse behavior such as conducting experiments on time scales that are orders of magnitude larger than traditionally assayed.

Most previous automated systems [3, 5] rely on the use of sensors like infrared beams or tracking techniques to monitor behavior. These approaches are limited in the complexity of the behavior that they can measure. 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, 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.

A few computer-vision systems for the recognition of mice behaviors have been recently described (a commercial system CleverSys, Inc, and [2, 7]). They have not been tested yet in a real-world lab setting using long uninterrupted video sequences which contain potentially ambiguous behaviors.

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.

2. System overview

Our system consists of two stages: (1) a feature computation stage, and (2) a classification stage. In the feature computation stage, a 310 dimensional feature descriptor is computed for each frame of an input sequence based on the motion and the position of the mouse. In the classification stage, a classifier is trained from the feature descriptors and labels of video sequences. The outputs are a sequence of labels, one for each frame of the sequence. The system is illustrated in Figure 1.

2.1. Feature computation stage

The feature computation stage takes as input a video sequence and outputs for each frame a feature vector of 310 dimensions. This comes from the concatenation of 300 motion features and 10 position- and velocity-based features, which are normalized separately before concatenation. A background subtraction procedure is first applied to an input video to compute a foreground mask for pixels belonging to the animal based on the instantaneous location of the animal in the cage (Figure 1(A)). The background subtraction procedure is adapted from our previous work for the recognition of human actions [4]. A bounding box centering on the animal is derived from the foreground mask (Figure 1(B)). Two types of features are then computed: position- and velocity-based features as well as motion features. Position- and velocity-based features are computed directly from the foreground mask (Figure 1(C)), and motion-features are computed on the bounding-box within a hierarchical architecture (Figure 1(D)).

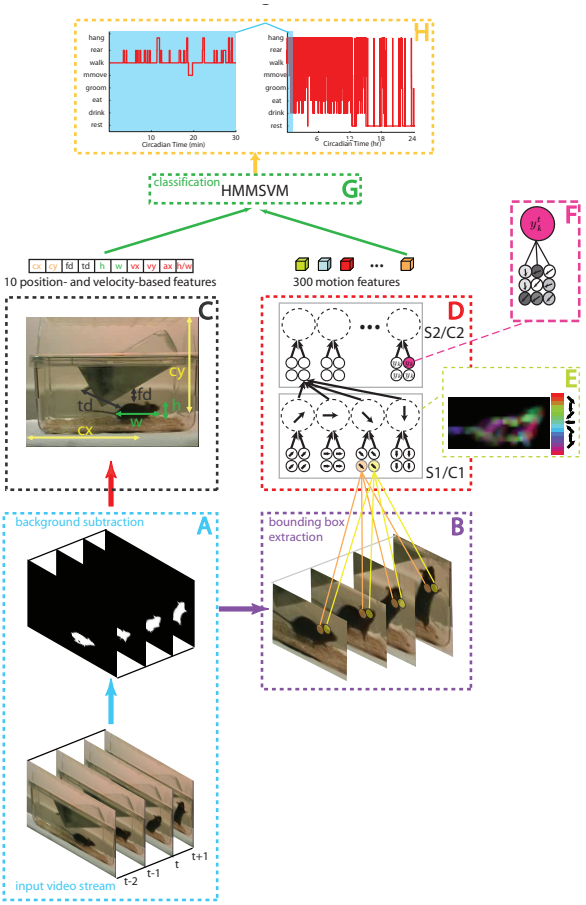

Figure 1. Overview of the proposed system for recognizing the home-cage behavior of mice. The system consists of a feature computation stage (A-F) and a classification stage (G). See text for detail.

Motion features The use of motion features is taken from our previous work, which models the organization of the dorsal stream (motion pathway) in the visual cortex and was applied for the recognition of human actions [4]. The model computes features for the space-time volume centering at every frame of an input video sequence via a hierarchy of processing stages, whereby features become increasingly complex and invariant with respect to $2D$ transformations as they move up the hierarchy. In the S_1/C_1 stage, motion signals are extracted from an input video sequence (Figure 1(E)). In the S_2/C_2 stage, feature vectors are computed from the similarity between the motion present in the current sequence and that in the training sequences (Figure 1(F)). The S_1 stage consists of an array of spatio-temporal filters (9 pixels $\times$ 9 pixels $\times$ 9 frames) tuned to 4 motion directions equally spaced between 0^0 and 360^0. An input gray-value sequence is convolved with each of the four filters, resulting in four S_1 maps centering at each frame. In the C_1 stage, a C_1 map is obtained by computing a local maximum over an 8×8 grid at every 4 pixels of a S_1 map. The S_2 stage matches the C_1 maps of the current frame with 300 stored templates that were extracted from training sequences (see below). At every position of the C_1 map, we perform a template matching (normalized dot product) between each of the 300 templates and the C_1 patch, with the same size as the template and centering at the current position. This stage generates 300 S_2 maps for each frame. The C_2 stage computes a global maximum (scalar) of each S_2 map. Finally, we obtain a 300-dimensional C_2 feature vector for each frame.

Learning the dictionary of motion templates In order to train a set of motion templates that are useful for discriminating between behavior categories, we manually collected a set of $4,200$ clips with the best and most exemplary instances of each behavior (each clip contains one single behavior). This set contains different mice (differing in coat color, size, gender, etc) recorded over 12 separate day or night sessions. We first draw $12,000$ motion templates, each as a patch of a random C_1 map computed from the clips. These templates are of sizes $n \times n$ pixels $(n = 4, 8, 12)$. We then do a feature selection as in [4], retaining the most representative 300 motion templates.

Position- and velocity-based feature computation In addition to the motion features, we compute an additional set of features derived from the instantaneous location of the animal in the cage (Figure 1(C)). We perform a background subtraction technique to obtain a foreground bounding box centering at the animal. For a static camera as used here, the background can be well approximated by a median frame in which each pixel value is the median value across all the frames at the same pixel location. Position- and velocity-based measurements are estimated for each frame based on the $2D$ coordinates (x, y) of the bounding box. These include the position and the aspect ratio of the bounding box (indicating whether the animal is in a horizontal or vertical posture), the distance of the animal from the feeder as well as the instantaneous velocity and acceleration. The 10 position- and velocity-based features are illustrated in Figure 1(C).

2.2. Classification stage

Existing systems for the recognition of mouse behavior focus on recognizing highly-exemplary instances of behaviors present in short clips (less than 100 frames). Performing a reliable phenotyping of an animal requires more than the mere detection of stereotypical non-ambiguous behaviors. In particular, the proposed system aims at classifying every frame of a video sequence even for those frames

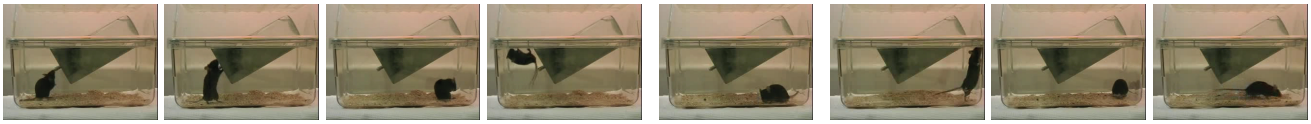

Figure 2. Snapshots taken from representative videos for the eight home-cage behaviors of interest.

whose underlying actions are difficult to categorize. For this challenging task, the temporal context of a specific behavior is an essential source of information for learning an accurate model of the behavior. In order to learn the temporal context, we use a Hidden Markov Support Vector Machine(SVMHMM) [6], which is an extension of the SVM for sequential tagging.

SVMHMM combines the advantage of SVM and HMM by discriminatively training models that are similar to hidden Markov models. Here we use a first-order transition model. Given an input sequence $\mathbf{X} = (\mathbf{x}_1 \dots \mathbf{x}_T)$ of feature vectors, the model predicts a sequence of labels $\mathbf{y} = (y_1 \dots y_T)$ according to the following linear discriminant function:

$$\mathbf{y} = \text{argmax}_{\mathbf{y}} \sum_{t=1}^{T} [\mathbf{x}_t \cdot \mathbf{w}_{y_t} + \mathbf{w}_{tr}(y_{t-1}, y_t)] \qquad (1)$$

$\mathbf{x}_t$ is the motion + position feature described above for the t-th frame of a video sequence, and y_t is the label (one behavior of interest) for the t-th frame. $\mathbf{w}_{y_t}$ is an emission weight vector for the label y_t and $\mathbf{w}_{tr}$ is a transition weight vector for the transition between the label y_{t-1} and y_t. These weight vectors are learned from 11 labeled videos (see Sec. 3.1 for training videos, and [6] for the training procedure). Each training video is split into non-overlapping 1 minute segments, each as a training example $\mathbf{X}$.

In the classification stage, the SVMHMM model takes as input a sequence of feature vectors of an input video and outputs a predicted label for each frame (Figure 1(G)). The resulting time sequence of labels can be further used to construct ethograms of the animal behavior. For example, the right panel of Figure 1(H) shows the ethogram of an animal for 24 hours, and the left panel provides a zoom-in version corresponding to the first 30 minutes of recording.

3. Experiments and results

3.1. Video dataset

Currently, the only public dataset for mouse behavior is a set of clips and is limited in the scope [2]. Each clip is no longer than 1 minute in length and contains one single action. In order to train and test our system on a real-world lab setting where mice behaviors are continuously observed and scored over hours or even days, we collected a dataset, *full database*. This set contains 12 continuously labeled videos: each frame is labeled with a behavior of interest. Each video is $30 - 60$ minutes in length, resulting in a total of over 10 hours of data. These videos contain different mice recorded at different times. We annotate 8 types of common mouse behaviors:

- drinking: the mouse's mouth being juxtaposed to the tip of the drinking spout
- eating: the mouse reaches and acquires food from the food bin
- grooming: the fore- or hind-limbs sweeps across the face or torso, typically as the animal is reared up
- hanging: grasping of the wire bars with the forelimbs and/or hind-limbs with at least two limbs off the ground
- micro-movements: small movements of the animal's head or limbs
- rearing: an upright posture and forelimbs off the ground
- resting: inactivity or nearly complete stillness
- walking: ambulation

These behaviors are shown in Figure 2.

3.2. Data Annotation

The videos were annotated using a freeware subtitle editing tool, Subtitle Workshop from UroWorks. A team of 8 investigators ('Annotators group 1') was trained to annotate mouse home cage behaviors. Two annotators of the 'Annotators group 1' further performed a secondary screening on these annotations to correct mistakes and make sure the annotation style is consistent throughout the whole database. In order to evaluate the agreement between two independent labelers, we consider a small subset of the *full database*, denoted as *doubly annotated subset*. It consists of many short video segments which are randomly selected from the *full database*. Each segment is $5-10$ minutes long and they add up to a total of about 1.6 hours of dataset. The *doubly annotated subset* has a second set of annotations made by the 'Annotators group 2', consisting of 4 annotators randomly selected from the 'Annotators group 1'.

3.3. Training and Testing the system

The evaluation as shown in Table 1 was obtained using a leave-one-out cross-validation procedure, *i.e.*, training the system on all but one of the videos and test on the left out video; repeating this procedure (n=12) times for all videos. The system accuracy is computed as: (total # frames correctly predicted by the system)/(total # frames) and the human-to-human agreement as: (total # frames correctly labeled by 'Annotators group 2')/(total # frames). Here a prediction or label is considered 'correct' if/when it matches the annotations by the 'Annotators group 1'.

3.4. Comparison with a commercial software and with human performance

Using the annotations of the 'Annotators group 1' as ground truth, we compared the performance of the system against a commercial software HCS (HomeCageScan 2.0, CleverSys, Inc) for mouse home cage behavior classification and against human manual scoring ('Annotators group 2'). Table 1 shows the comparison. Overall we found that our system achieves 76.6% agreement with human labelers ('Annotators group 1') on the *doubly annotated subset*, a result significantly higher than the HCS and on par with humans ('Annotators group 2'). For the *full database*, we only compared the system against the HCS, and our system also outperforms the HCS by 17%. Two online videos demonstrating the automatic scoring of the system are at `http://techtv.mit.edu/videos/5561` and `http://techtv.mit.edu/videos/5562`.

	Our system	HCS	Human ('Ann. group 2')
doubly annotated set	76.6%	60.9%	71.6%
full databse	77.6%	61.0%	

Table 1. Accuracy of our system, human annotators and HomeCageScan 2.0 CleverSys system.

3.5. Discussion

A common source of disagreement between human annotators and between humans and the system is the precise boundary between 2 actions (when one action starts and the other ends). Furthermore, some of our behaviors of interest can be very short (10-20ms), thus making it hard to allow for longer tolerances in the precise locations of these boundaries. The disagreement also comes from the ambiguity of actions themselves. For example, a mouse standing against the back side of a cage (rearing) looks very similar to a mouse reaching for the food bin (eating). In both cases, the head of the mouse appears to touch the food bin as seen from the front side of the cage. Small movements of an animal's limbs (micro-movement) are sometimes associated with a slow movement of the whole body, blurring the boundary between walking and micro-movements. The videos at `http://techtv.mit.edu/videos/5563` and `http://techtv.mit.edu/videos/5564` show annotations from two humans simultaneously and illustrate the confusions described above. We believe that errors from the system result from inconsistencies in the annotations produced by multiple annotators and the inherent ambiguity between certain actions.

4. Conclusion

We have applied a biological model of motion processing in the dorsal stream to the recognition of human and animal actions. It has also been suggested that analysis of shape in the ventral stream of the visual cortex may also be important for the recognition of actions. Future work will extend the present approach to integrate shape/contour, motion and sensor information. Another important future direction is to extend the study of single mouse behavior to multiple mice behaviors and social behaviors.

5. Acknowledge

This research was sponsored by grants from: DARPA (IPTO and DSO), NSF-0640097, NSF-0827427, IIT, and the McGovern Institute for Brain Research. ADS were funded by the Broad Fellows Program in Brain Circuitry at Caltech. HJ was funded by the Taiwan National Science Council (TMS-094-1-A032).

References

[1] J. Auwerx and A. El. The European dimension for the mouse genome mutagenesis program. *Nature Genetics*, 36(9):925–927, 2004.

[2] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *visual surveillance and Performance Evaluation of Tracking and Surveillance*, 2005.

[3] E. H. Goulding, A. K. Schenk, P. Juneja, A. W. Mackay, J. M. Wade, and L. H. Tecott. A robust automated system elucidates mouse home cage behavioral structure. *Proceedings of the National Academy of Sciences*, 105(52), 2008.

[4] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. *IEEE International Conference on Computer Vision*, 2007.

[5] L. P. Noldus, A. J. Spink, and R. A. Tegelenbosch. EthoVision: a versatile video tracking system for automation of behavioral experiments. *Behavior Research Methods, Instruments, Computers*, 33(3):398–414, August 2001.

[6] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large Margin Methods for Structured and Interdependent Output Variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.

[7] X. Xue and T. C. Henderson. Feature fusion for basic behavior unit segmentation from video sequences. *Robotics and Autonomous Systems*, 57(3):239–248, March 2009.