Detecting individual body parts improves mouse behavior classification

Christoph M. Decker, Fred A. Hamprecht
Heidelberg University, HCI/IWR
{christoph.decker, fred.hamprecht}@iwr.uni-heidelberg.de

***Abstract*—We experiment with interactive machine learning for mouse behavior classification, following the pioneering work JAABA [1]. Here, we describe a simple image processing pipeline that allows extracting individual body parts from single mouse top view video. Our experiments show that behavior classification accuracy increases substantially when transitioning from whole-body descriptors to features computed from individual body parts, their position and motion.**

I. Introduction

The study of animal behavior and its pathologies provides important cues to biologists and medical researchers alike. Rodents exhibit complex behavior and are among the most popular research animals. To further increase the throughput and/or complexity of rodent behavioral experiments, there is strong demand for automated behavior classification solutions with as little user input as possible. Work such as JAABA [1] is driven by the vision that users can define any behavior of interest and train a classifier to detect it with minimal effort. Once training is completed, large data sets can be analyzed automatically without further supervision.

We decide to work with top (rather than side) view video for a number of reasons: Firstly, in experiments such as the open field test the variability of the animal's appearance due to location and orientation is smaller. Secondly, self-occlusions of the animal's body parts become less likely. And lastly, the rotational invariance of motions and behaviors can be exploited during feature computation.

II. Related Work

Several approaches for (semi-) automated mouse behavior classification have been published in recent years. For side view video, [2] introduced a novel spatio-temporal interest point detector. An approach to model the behavior by a large set of postures detected from shape and position was implemented in the software HomeCageScan [3], [4]. In [5], a Hidden Markov Model (HMM) Support Vector Machine was trained on position-based and spatio-temporal motion features. In the domain of top view video, [6] trained different variants of HMMs on position and silhouette features. Using the output of their own tracking software, [7] inferred the behavior from sequences of basic behavioral elements which they detected with a classifier based on motion, optical flow, shape and position features. Additionally to shape and motion features, [8], [9] used a depth camera to extract the elevation angle of the mouse as a feature.

All of these methods come with a fixed number of predefined classes, making it impossible for the experimenter to look for custom behavior patterns in the data. To the best of our knowledge, the only approach published to date which allows the experimenter to train user-defined behavior classes with minimal guidance is JAABA [1] by Kabra et al. They train a GentleBoost classifier on sparse user annotations and achieve strong results on adult and larval Drosophila with up to 15 behavior classes. For mouse behavior analysis, an animal was represented in terms of an ellipse and all features were extracted from this representation.

In this work, we aim at providing more expressive features to our classifier by detecting individual body parts and describing their shape, movement and relative position.

III. Body Part Prediction & Segmentation

To take the first step towards a more detailed body part representation from which we can deduce strong features, we use the "Pixel Classification" workflow of the open source interactive learning and segmentation toolkit ilastik [10]. This software learns a random forest classifier [11] on sparse user labels input via a graphical user interface to distinguish between different user-defined pixel classes (here: body parts). The class decision is based on a set of color, edge and texture features computed at different scales[1]. After successful training we can export a probability map $P_{t,c}(x,y)$ for each frame t which indicates the propensity for each and every pixel (x,y) to belong to one of the body part classes or to background (c denotes the class index).

Starting from these probability maps, we propose the sequence of steps summarized in figure 1 to obtain exactly one connected component per body part per frame.

A. Background Subtraction

As we deal with video from a static camera and the mouse typically only covers a small region of every frame, we use a standard median filter method [12], [13] to exclude the background from further processing to reduce misdetections and computational cost.

B. Temporal Consistency Enforcement

Given standard video frame rates and ordinary single mouse behavior, it is valid to assume that the displacement of body

[1] We only select a subset of the 37 appearance features available in *ilastik* for our body part prediction.

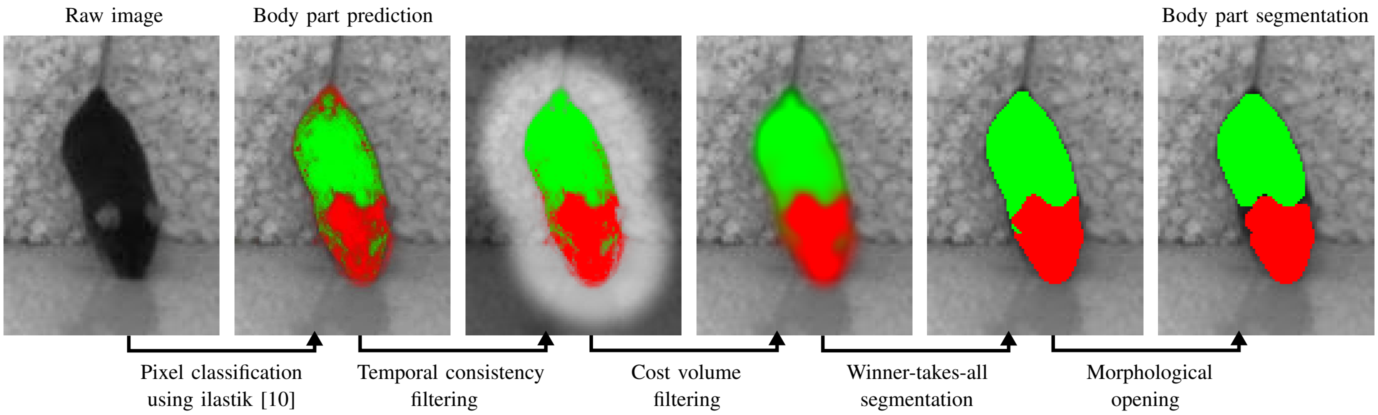
Raw image
Body part prediction
Body part segmentation
Pixel classification using ilastik [10]
Temporal consistency filtering
Cost volume filtering
Winner-takes-all segmentation
Morphological opening

Fig. 1. From raw video to a meaningful body part segmentation with classes "head" (red) and "body" (green) through the proposed processing pipeline. The probability of the background class and its labeling are not shown here for cleaner visualization.

parts between successive frames is small. We exploit this observation by shrinking the allowed region for finding a body part based on its previous detection.

In practice, we apply a separate mask to each class probability map, which is 1 inside the allowed region and 0 elsewhere. For computational reasons, we set the allowed region to be the bounding ellipse of the respective body part's connected component detected in the previous time step dilated by δ pixels. These masks can be multiplied by a factor $\gamma > 0$ to adjust the strength of the influence of the previous detection on the current probabilities. To allow a body part to move slightly further than δ without sharply truncating it at the boundaries of the constraint region, we smooth the masks with a Gaussian kernel. The effect of the temporal consistency enforcement on a probability map is shown exemplarily in figure 2.

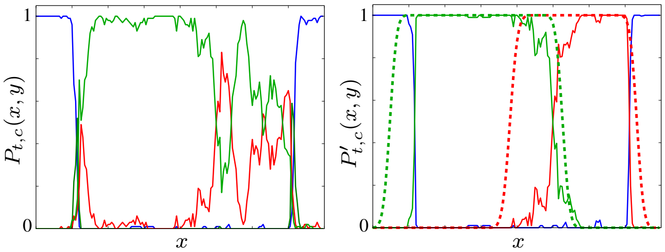

Fig. 2. Cut through a probability map $P_{t,c}(x, y)$ with classes "head" (red), "body" (green) and "background" (blue) along the x-axis before (left) and after (right) applying the temporal consistency enforcing masks represented by the dashed lines (right).

C. Regularization

Since ilastik's "Pixel Classification" workflow only provides body part probabilities for each pixel position independently, the probability maps typically are too noisy to directly extract a clean segmentation with a single contiguous region for every body part. To regularize, we recur to cost volume filtering [14] as a computationally efficient drop-in replacement for MAP estimation in a multilabel Markov random field.

D. Segmentation

To transform the post-processed class probabilities to a meaningful segmentation, we use a simple winner-takes-all strategy. We hence create a binary indicator array whose elements $B_{t,c}(x, y) \in \{0, 1\}$, $\sum_c B_{t,c}(x, y) = 1$, indicate to which class c a pixel (x, y) in frame t belongs. Each of the C channels now contains one or multiple candidate connected components for the respective body part.

To remove remaining excrescences and misdetections, we next apply an opening operator with a small circular structuring element.

Finally, we select the single largest (in terms of pixel count) candidate in every body part channel.

IV. Behavior Classification

Once this representation of the mouse is found, we use it to derive information about its pose and movement by means of a large collection of spatio-temporal shape and motion features. These features are then fed into a random forest classifier which can be trained on an arbitrary number of behavior classes defined by the experimenter.

A. Features

To cover as many aspects of any custom behavioral element as possible, we propose to use an exhaustive set of per-frame features which describe both shape and motion of the whole body, every individual part and also the relation between each pair of parts. Additionally, we use window features [1] to encode the temporal evolution of the per-frame features in small temporal windows around the current frame. Table I shows all features we use for behavior classification. Some feature descriptions refer to properties introduced in figure 3.

B. Behavior Prediction

The classifier can rely on a high-dimensional feature vector $\vec{f}_t$ for every frame t, which contains six features per body, ten per body part, eight for each pair of parts and additionally

TABLE I

FEATURES FOR BEHAVIOR CLASSIFICATION

Category	Feature	Computed for
Shape	Volume and temporal change thereof	whole body and every part
	Major and minor radius of bounding ellipse	
	Radii ratio and temporal change thereof	
	Internal distance $\vec{d}_{i,j}$	every pair of parts i and j[2]
Motion	Axial speed $v^{\parallel}_{i,j}$ and perpendicular speed $v^{\perp}_{i,j}$ of part i relative to $\vec{d}_{i,j}$	
	Change of internal distance $d_{i,j}$ over time	
	Rotation velocity of $\vec{d}_{i,j}$	
	Angular acceleration of $\vec{d}_{i,j}$	
	Current velocity (i. e. displacement between $t - dt$[3] and t)	every part
	Displacement between $t - dt$ and $t + dt$ (leaving current position out)	
	Acceleration magnitude and angle (change of velocity vector between t and $t + dt$)	
Window	Mean	every per-frame feature
	Median	
	Standard deviation	

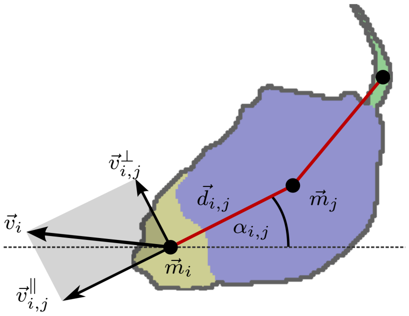

Fig. 3. Schematic drawing to illustrate some important properties and features: Center of mass $\vec{m}_i$, distance $\vec{d}_{i,j}$ between parts i and j, angle $\alpha_{i,j}$ between $\vec{d}_{i,j}$ and the x-axis, total speed $\vec{v}_i$ and its perpendicular and axial components $\vec{v}^{\perp}_{i,j}$ and $\vec{v}^{||}_{i,j}$ with respect to $\vec{d}_{i,j}$.

three window features for each of the aforementioned per-frame features.

To train the random forest classifier, the experimenter has to add as many behavior patterns as necessary and annotate a few frames for each behavior class (e. g. in the first minutes of the video in question). Once training is finished, the classifier can be used to predict with which probability an unseen frame belongs to each of the behavior classes.

To finally infer the predicted behavior class, we smooth the probabilities over time with a Gaussian kernel followed by a winner-takes-all segmentation, resulting in a single most probable behavior for every frame.

V. EXPERIMENTS

To quantitatively evaluate the proposed features, we have used two grayscale videos M1 and M2 of different mice from different labs, see figure 4. M1 was recorded in the laboratory of Prof. Andreas Draguhn in the Institute of Physiology and Pathophysiology of Heidelberg University, Germany, and contains 9000 frames with a spatial resolution of 1280 × 720 pixels recorded at 30 fps (five minutes of video). M2 was recorded by Prof. Roian Egnor in her laboratory at Janelia Farm Research Campus, Virginia, USA, and consists of 5000 frames recorded at 29 fps (about three minutes of video) with a spatial resolution of 1024 × 768 pixels.

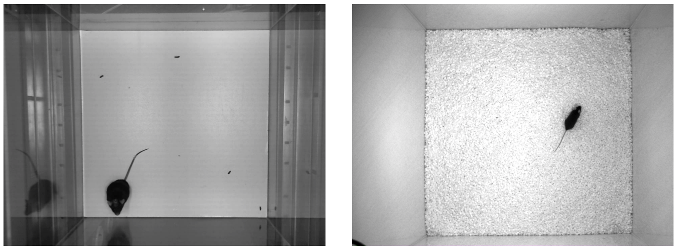

Fig. 4. Exemplary frames from data sets M1 (left) and M2 (right). Note the mirroring of the mouse in M1 which is suppressed by the proposed processing pipeline.

The four behavior classes we use are “locomotion”, “micromovement”[4], “immobility” and “rearing”. The numbers of frames contained in the respective training and test sets of M1 and M2 are listed in table II.

TABLE II

NUMBER OF TRAINING LABELS AND TEST SET FRAMES IN M1 AND M2

	M1		M2	
	Training	Testing	Training	Testing
Immobility	23	712	6	54
Micromovement	27	4140	19	712
Locomotion	24	1422	26	1667
Rearing	21	596	25	890
Total	95	6870	85	3323

Training the pixel classifier for body part predictions as described in section III takes about 30 minutes for each data set. Once trained, the classifier is ready to be applied to arbitrary length video of a similar-looking mouse. As the tails are only dragged in both data sets and thus do not add information for behavior classification, we only operate on the two individual parts “head” and “body”, leaving us with 34 per-frame and 102 window features.

The parameters described in section III have been tuned manually until the body part segmentation looked sensible: For the dilation δ in the temporal consistency enforcement in III-B, the maximum pixel distance the mouse might move between successive frames is decisive. The standard deviation of the Gaussian kernel used for cost volume filtering in III-C and the size of the opening operator used to clean up the

[2] Most of these features are invariant under permutation of i and j, but axial and perpendicular speed have to be computed for both combinations.

[3] Computing the motion features for different values of dt might be useful to capture even more information about the current behavior.

[4] The compound behavior class “micromovement” corresponds to small in-place movements such as sniffing, grooming and the like, for the discrimination of which we did not have enough training frames in the annotated data.

TABLE III

Confusion matrices for behavior classification on data sets M1 and M2. Transitioning from reduced (whole-animal) to full (body part) features never deteriorates, and often improves, accuracy.

	M1								M2							
	reduced feature set				full feature set				reduced feature set				full feature set			
Actual \ Predicted	I	M	L	R	I	M	L	R	I	M	L	R	I	M	L	R
Immobility (I)	**84**	13	00	03	**85**	15	00	00	**02**	89	00	09	**20**	69	00	11
Micromovement (M)	10	**60**	03	27	08	**80**	03	09	01	**65**	10	24	01	**74**	11	14
Locomotion (L)	00	03	**87**	10	00	06	**87**	07	00	07	**91**	02	00	07	**91**	02
Rearing (R)	01	20	07	**72**	01	18	07	**74**	00	14	03	**83**	00	09	02	**89**

segmentation in III-D depend on the size of the smallest body part to be detected.

As table II reveals, the behavior classifier needs only little training: About 25 frames per behavior class have been annotated with a user effort of around ten minutes per video.

A. Results

Table III shows the per-frame confusion matrices for M1 and M2 when comparing the classifier's prediction output to manually annotated ground truth. Creating the latter has taken about five hours for a total of eight minutes of video (14000 frames).

With a reduced feature set, containing only shape, motion and window features of the whole body[5], classification achieves an average prediction accuracy of 75.7% on M1. For M2, table II reveals that there have been too few training and testing frames for the class "immobility" to get meaningful results. This circumstance presumably explains the strong confusion with "micromovement". Disregarding "immobility", our method achieves an average accuracy of 79.8% which is consistent with the result on M1.

The comparison between training the classifier on aforementioned reduced feature set and training it on our full set of features extracted from the individual body parts as listed in table I highlights the main point of this contribution. As can be seen from table III, the average prediction accuracy is increased by more than 5% to 81.7% and 84.8% on M1 and M2 respectively, once the full set of features is used. Especially the confusion of "micromovement" and "rearing" decreases significantly by more than 10% in both data sets.

Note that the accuracy which can be reached by a classifier trained on user input has an upper bound dictated by the limited inter-observer agreement of humans. In [5], this inter-observer agreement for mouse behavior classification with eight classes is reported to be only 72%[6]. Hence, any classifier reaching at least comparable accuracy could replace human observation.

VI. Conclusion

We have presented an approach to improve automated mouse behavior recognition from top view video. The key ingredient of our method is the segmentation into distinct body parts. This representation gives us access to a much richer set of shape and motion features which we then use to classify custom user-defined behaviors. We show that the additional features derived from the segmentation can lead to a significant increase in classification accuracy.

References

[1] M. Kabra, A. A. Robie, M. Rivera-Alba, S. Branson, and K. Branson, "JAABA: interactive machine learning for automatic annotation of animal behavior," *Nature Methods*, 2012.

[2] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. IEEE, 2005, pp. 65–72.

[3] Y. Liang, V. Kobla, X. Bai, and Y. Zhang, "Unified system and method for animal behavior characterization in home cages using video analysis," 2007, US Patent 7,209,588.

[4] A. D. Steele, W. S. Jackson, O. D. King, and S. Lindquist, "The power of automated high-resolution behavior analysis revealed by its application to mouse models of huntington's and prion diseases," *Proceedings of the National Academy of Sciences*, vol. 104, no. 6, pp. 1983–1988, 2007.

[5] H. Jhuang, E. Garrote, X. Yu, V. Khilnani, T. Poggio, A. D. Steele, and T. Serre, "Automated home-cage behavioural phenotyping of mice," *Nature Communications*, vol. 1, p. 68, 2010.

[6] J. Verbeek, "Rodent behavior annotation from video," Intelligent Systems Laboratory, University of Amsterdam, Tech. Rep., 2005.

[7] E. A. van Dam, J. E. van der Harst, C. J. ter Braak, R. A. Tegelenbosch, B. M. Spruijt, and L. P. Noldus, "An automated system for the recognition of various specific rat behaviours," *Journal of Neuroscience Methods*, vol. 218, no. 2, pp. 214 – 224, 2013.

[8] J. P. Monteiro, H. P. Oliveira, P. Aguiar, and J. S. Cardoso, "Depth-map images for automatic mice behavior recognition," in *1st PhD Students Conference in Electrical and Computer Engineering, Porto, Portugal*, 2012.

[9] J. P. Monteiro, "Automatic behavior recognition in laboratory animals using kinect," Master's thesis, Universidade do Porto, 2012.

[10] C. Sommer, C. Straehle, U. Koethe, and F. A. Hamprecht, "ilastik: Interactive learning and segmentation toolkit," in *8th IEEE International Symposium on Biomedical Imaging*, 2011.

[11] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[12] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1337–1342, 2003.

[13] M. Piccardi, "Background subtraction techniques: a review," in *IEEE International Conference on Systems, Man and Cybernetics*, vol. 4. IEEE, 2004, pp. 3099–3104.

[14] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 3017–3024.

[5] As we did not compute motion features for the whole body, we used those computed for the part "body".

[6] On a four class problem the inter-observer agreement will presumably be higher due to less confusions.