Clustering-based Active Learning in Unbalanced Rodent Behavior Data

Malte Lorbach*†, Ronald Poppe*, Elsbeth A. van Dam†, Lucas P.J.J. Noldus† and Remco C. Veltkamp*
*Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands
†Noldus Information Technology, Wageningen, The Netherlands

***Abstract*—For the objective measurement of animal behavior from video, automated recognition systems are frequently employed. These systems rely on action models learned from labeled example videos. Manually labeling videos of animal behavior however is time consuming and error-prone. We propose to reduce the labeling effort by selecting suitable training instances from the unlabeled corpus and learn the action models iteratively in interaction with the user. Due to the typical imbalance of behavior datasets, a random selection strategy would fail to sample enough minority class examples. To address the imbalance we first find potential action prototypes by clustering the unlabeled data using a Dirichlet process Gaussian mixture model. We then sample instances from the prototypes and obtain a more balanced training set. We evaluate our system on two rat interaction datasets with different class priors and demonstrate an increased learning rate that is superior to the baseline.**

1. Introduction

Automatically recognizing rodent actions in videos plays an integral part in quantifying rodent behavior for research on, e.g., neurodegenerative diseases such as Huntington's disease. The automation has enabled researchers to classify rodent actions in large datasets with reduced manual efforts and improved reproducibility [1].

Typically, the recognition system is trained by learning action models from a training set of labeled examples. The creation of such a training set however is expensive and error-prone. A human expert needs on average 1 hour to annotate 5 minutes of video [2], [3]. Moreover, it is difficult to estimate in advance how much training data is required for an optimal recognition accuracy as it depends heavily on the difficulty of the task. Finally, rodent behavior datasets as in Figure 1 are often unbalanced with respect to the different behaviors. This imbalance poses a challenge to select suitable training videos as all behaviors must be present to a sufficient amount.

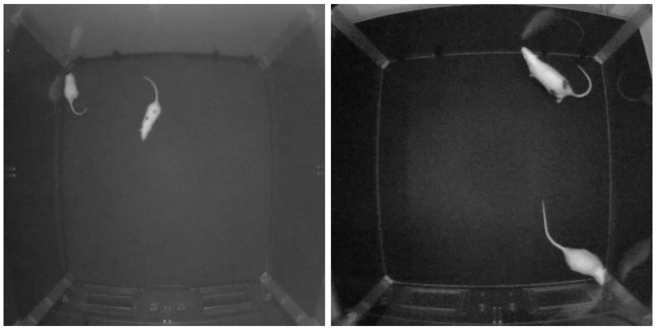


Figure 1. The animals in our two datasets YRPB (left) and PRSCA (right) are at different ages. The age difference has a strong effect on the occurrence of the various behaviors.

We address the reduction of the labeling effort by formulating the task of learning action models as an active learning problem [4]. In active learning, the training set is initially unlabeled and the learner iteratively queries an oracle, e.g. a human expert, to label selected data points. Since we deal with unbalanced data however, querying random points is inadequate. Figure 2 illustrates this aspect: by embedding the data in a 2-dimensional, neighborhood-preserving space [5], we can visualize how the large amount of *contact* instances (blue circles) hides most of the structure that, if balanced, is clearly present. The learner would clearly benefit from balancing queries among action classes.

Contribution: we propose a sample selection algorithm that explores the input space and searches for the inherent structure in the unlabeled data before making a selection. We uncover that structure by clustering the data into potential action prototypes. The clustering is obtained by fitting a mixture of Gaussians following a Dirichlet Process. We then exploit the clusters and balance the queries among them. The result is a training set with a more homogeneous class distribution. To evaluate the effectiveness we compare different selection strategies based on the obtained clustering with a random baseline method.

We continue by discussing common learning schemes in related work dealing with the recognition of rodent behavior from video.

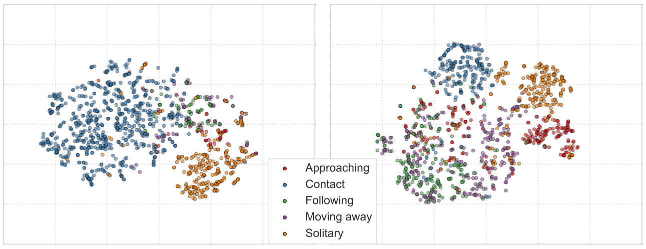


Figure 2. 2-dimensional embeddings of a subset of YRPB. Left: random sample; Right: balanced sample. This figure is best viewed in color.

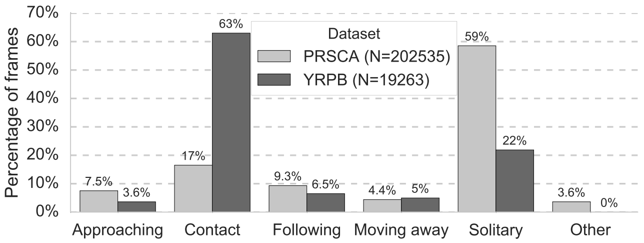


Figure 3. The datasets are unbalanced regarding the occurrence of each action. The young rats (YRPB) spend much more time in contact than the older rats (PRSCA).

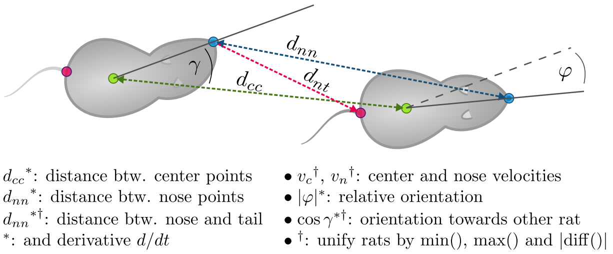

d_{nn}
γ
d_{cc}
d_{nt}
φ
• $d_{cc}{}^*$: distance btw. center points
• $d_{nn}{}^*$: distance btw. nose points
• $d_{nn}{}^{*\dagger}$: distance btw. nose and tail
• *: and derivative d/dt
• $v_c{}^\dagger$, $v_n{}^\dagger$: center and nose velocities
• $|\varphi|^*$: relative orientation
• $\cos\gamma^{*\dagger}$: orientation towards other rat
• †: unify rats by min(), max() and |diff()|

Figure 4. The features are derived from three tracked body points: the tail base, the center point and the nose point. In total we derive 24 features.

2. Learning to recognize rodent behavior

The automated recognition of rodent behaviors involves classifying video frames into behavior categories based on features extracted from video. The features may be derived from the trajectories of the moving animals [3], [6]–[8] as well as from optical flow [2] and dense trajectories [9].

The classifier needs to be trained for a particular dataset. The most common training approach is supervised learning [2], [3], [6], [8], [10]–[12]. Supervised learning requires to obtain a training set of videos labeled by experts. This set needs to be sufficiently large to facilitate cross-validating the model parameters and evaluating the accuracy. In practice, it can be surprisingly challenging to create a high quality training set that captures the natural variability of also rarely occurring actions.

Although recent work has advanced our understanding of rodent behavior recognition, few have addressed the problems arising from the supervised learning approach. One example that avoids the creation of a large training set is the Janelia Automatic Animal Behavior Annotator [7]. By adopting an active learning approach, the system's classifier is interactively trained by the user, who himself selects the video sequences to label. We extend the idea by selecting the examples automatically in a data-driven approach.

Our algorithm combines the ideas of a data-driven active learning approach [13] and the properties of the Dirichlet Process that facilitate the discovery of rare classes [14]. To our knowledge we are the first to apply clustering-based active learning to rodent behavior recognition.

3. Datasets

Our investigations are based on two datasets of rat social behavior. They comprise continuous top-view video recordings of two rats interacting in a 90×90 cm cage. Example frames are shown in Figure 1. Each frame is annotated by an expert with one of six different action labels: *approaching*, *contact*, *following*, *moving away*, *solitary*, or *other*. The *other* class contains undefined behavior for which we do not learn a model. The occurrence of actions is not uniformly distributed but is highly skewed towards a majority class (see Fig. 3 for the distribution of the class priors).

We obtain the trajectories of the animals using a customized version of Noldus EthoVision XT [15]. This version maintains the identity of the animals up to about five errors per five minutes of video, which are then corrected manually. From the trajectories we compute 24 features derived from 12 base features per animal. The features are combined across the animals by computing the min, max, and absolute difference between the respective values. Refer to Figure 4 for an overview about the features.

The datasets differ considerably in their quality. YRPB comprises about 14 min of selected sequences with revised, frame-correct annotations. PRSCA is a larger dataset of nine videos with 15 min each that is adopted without further modification.

4. Active learning for rodent behavior recognition

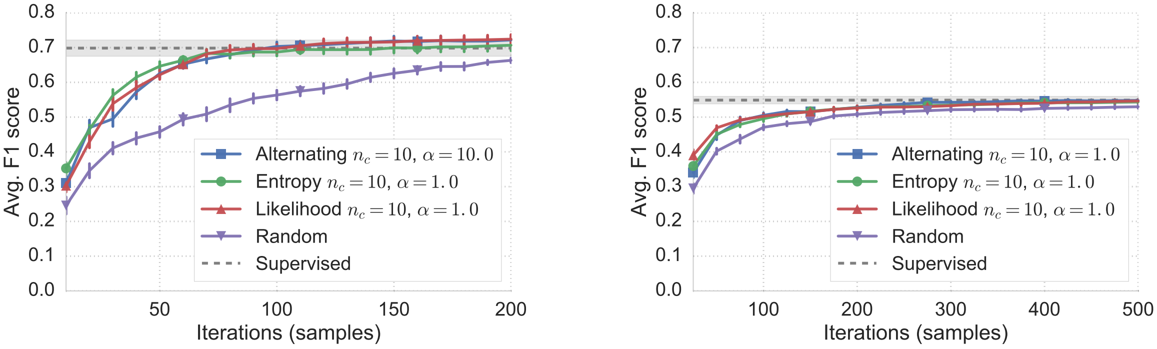
We now turn to our active learning framework. We first formulate the learning task and then introduce the proposed algorithm.

The learner has access to a large pool of unlabeled data $\mathcal{U}$ and a pool of labeled instances $\mathcal{L}$ which may initially contain a few random instances. The learner then iteratively selects one instance x_i from $\mathcal{U}$ and queries an oracle about its true label y_i. The instance is removed from $\mathcal{U}$ and added to $\mathcal{L}$. After each iteration or after a number of iterations, a classifier is trained using the available labeled instances. The task of the learner is to achieve a level of accuracy in classification using fewer examples than a random learner.

4.1. Cluster-based query strategy

The key component of our active learning framework is the sample selection process. We attempt to balance the selection among the true action classes by exploring the structure in the unlabeled data. Under the assumption that instances of the same class are close in feature space, we cluster the data into potential action prototypes. We fit a Gaussian mixture model (GMM) while adding mixture components following a Dirichlet process (DP) [16]. As the DP is able to expand the mixture model theoretically to an infinite number of Gaussians, it enables us to discover an unknown number of clusters covering also small classes.

The process of selecting a query instance x_i is divided into two steps: we first select a cluster and then an instance



(a) Dataset: YRPB ($N_{\text{YRPB}} = 18987$)

(b) Dataset: PRSCA ($N_{\text{PRSCA}} = 195015$)

Figure 5. Learning curves of each strategy with its best scoring parameter set. Scores at each iteration are averaged across three cross-validation folds and five repetitions each. Error bars give the standard error of the mean. The score of the supervised classifier with access to all labels is given for reference.

from that cluster. We explore three selection strategies incorporating different characteristics of the clusters:

- `Alternating`: The clusters are selected one-by-one in round-robin sequence. Instances are then picked at random.
- `Likelihood`: Clusters are selected in round-robin sequence; instances are picked with probability according to the likelihood that the cluster model has generated the instance. This strategy is less likely to select ambiguous instances where models overlap.
- `Entropy`: Clusters are selected with probability according to the entropy of their known labels. Heterogeneous clusters are therefore selected more often. Instances are then selected randomly.

We compare the cluster-based strategies with the `Random` baseline that samples instances disregarding the clustering. For classification we follow a Bayesian approach and fit a GMM to the labeled instances. We use one mixture component per class with a diagonal covariance matrix, which we found to work best in previous experiments. Note that any type of classifier is applicable at this point. In particular we have compared the GMM to an SVM but have found the SVM to be prone to overfitting despite regularization and sensitive to its parameterization.

5. Empirical evaluation

We evaluate our approach by computing the learning curves in terms of the F1 score averaged over classes. The F1 score is the harmonic mean between precision and recall. Averaging the score over classes instead of frames prevents the score from being dominated by the majority class.

For comparison we vary two parameters of the DP, namely: the maximum number of clusters $|\mathcal{C}|_{\max}$ and the concentration parameter α. We relate the number of clusters to the number of action classes, $|\mathcal{C}|_{\max} = n_c \cdot n_{\text{actions}}$, and vary $n_c \in \{3, 5, 10\}$. The parameter α controls the chance of creating new clusters so that a higher value of α leads to more clusters. We vary $\alpha \in \{1, 10\}$.

For cross-validation we split the dataset into three folds such that folds consist of entire videos. Each learning experiment is repeated five times and we average the results over repetitions and folds.

Due to space constraints we cannot show all learning curves. Instead we compute the area under the learning curve (AUC) as a performance metric. We divide the area by the number of iterations so that the maximum score is 1.

5.1. Results

The learning curves in Figure 5 show that the three cluster-based strategies need fewer instances than the baseline method to achieve high classification accuracy. They even score higher than the supervised method on YRPB.

Between datasets, the learner appears more effective and efficient for YRPB than for PRSCA, where the difference to the baseline method is smaller and the final accuracy lower.

Among the methods and their parameterization the differences in AUC are relatively small (Table 1). While α has only a marginal effect on the performance, limiting the number of clusters too much leads to lower performance as clusters become larger and more likely to be heterogeneous. Note that internally the parameters serve as an upper bound to the number of clusters and therefore values higher than reported do not have an effect on the score.

5.2. Discussion

The active learner works well on YRPB, but it does not achieve a similar improvement over the baseline on PRSCA. It appears that PRSCA is the more challenging dataset. We believe that the main reason lies in the amount of label noise present. If we analyze the distribution of labels in the clusters after the last iteration, we find more heterogeneous clusters than in YRPB. With too much heterogeneity, we lose the ability to balance the sampling across actions and the active learner is no better than the random learner.

Dealing with label noise is challenging; in particular if the labeling has not even taken place yet as in active learning. A simple approach could be to exclude a particularly

Table 1. Area under learning curve for a range of algorithm parameterizations. Maximum is score is 1, best score per strategy in bold.

YRPB

n_c	3		5		10	
α	1.0	10.0	1.0	10.0	1.0	10.0
Alternating	0.550	0.557	0.593	0.608	0.621	**0.625**
Likelihood	0.572	0.570	0.603	0.603	**0.628**	0.615
Entropy	0.581	0.569	0.610	0.588	**0.625**	0.609
Random	0.517					

PRSCA

n_c	3		5		10	
α	1.0	10.0	1.0	10.0	1.0	10.0
Alternating	0.471	0.469	0.494	0.489	**0.497**	0.491
Likelihood	0.477	0.475	0.490	0.495	**0.496**	0.494
Entropy	0.481	0.481	0.479	0.491	**0.493**	0.490
Random	0.472					

noisy cluster as candidate for sampling. Eventually though, we may want to perform probabilistic inference taking the noisy oracle into account.

Most of the variation in the performance is due to the parameterization of the clustering rather than the strategies exploiting them. Evidently, most of the work is done by the former. On the one hand, we could make better use of the information that we obtain from the clusters. On the other hand, being the backbone of the current learner it is crucial to improve the clustering, for instance by considering temporal dependencies.

The temporal dependency between frames may be useful in two ways. The clustering of difficult, ambiguous action classes could benefit from the fact that there should be more frame transitions within a cluster (same action) than across clusters (transition to another action). Similarly, we could query short sequences while avoiding sequences that cross cluster borders as they are more likely to contain transitions between actions that are inherently difficult to label.

In summary, the evaluation shows that using clustering to balance the selected samples among true action classes increases the learning rate.

6. Conclusion

We have investigated a clustering-based active learning framework for rodent behavior recognition. We have demonstrated the effectiveness of our approach to use clustering for balancing the sample selection among action classes. The limitation of the algorithm as it relies on clusters with largely homogeneous distribution of action labels is the subject of future investigations. Potential directions of improvement address both label noise and temporal information.

Furthermore, to stimulate queries of short sequences instead of single frames, we would like to explicitly incorporate the expected labeling cost in the sampling. Exploiting the temporal dependency of frames would allow for an estimation of that cost. Eventually these enhancements should lead to an increased learning rate and thus accelerate the training of rodent behavior classifiers.

Acknowledgments

This work was funded by the EC under the 7th Framework Programme (FP7) as part of the Marie Curie Initial Training Network "PhenoRat" (GA no. 317259).

References

[1] S. E. R. Egnor and K. Branson, "Computational Analysis of Behavior," *Annual Review of Neuroscience*, vol. 39, no. 1, pp. 217–236, 2016.

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

[3] L. Giancardo, D. Sona, H. Huang, S. Sannino, F. Managò, D. Scheggia, F. Papaleo, and V. Murino, "Automatic Visual Tracking and Social Behaviour Analysis with Multiple Mice," *PLoS ONE*, vol. 8, no. 9, p. E74557, 2013.

[4] B. Settles, "Active learning literature survey," University of Wisconsin-Madison, Technical Report 1648, 2009.

[5] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[6] X. P. Burgos-Artizzu, P. Dollár, D. Lin, D. J. Anderson, and P. Perona, "Social behavior recognition in continuous video," in *Proc. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1322–1329.

[7] 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*, vol. 10, no. 1, pp. 64–67, 2012.

[8] E. Eyjolfsdottir, S. Branson, X. P. Burgos-Artizzu, E. D. Hoopfer, J. Schor, D. J. Anderson, and P. Perona, "Detecting Social Actions of Fruit Flies," in *Proc. Conf. Computer Vision (ECCV)*, vol. 8690, 2014, pp. 772–787.

[9] H. Kuehne, J. Gall, and T. Serre, "An end-to-end generative framework for video segmentation and recognition," in *Proc. Conf. Applications of Computer Vision (WACV)*, 2016, pp. 1–8.

[10] W. Hong, A. Kennedy, X. P. Burgos-Artizzu, M. Zelikowsky, S. G. Navonne, P. Perona, and D. J. Anderson, "Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning," *Proc. National Academy of Sciences*, vol. 112, no. 38, pp. E5351–E5360, 2015.

[11] 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, no. 6, pp. 1–9, 2010.

[12] M. Lorbach, R. Poppe, E. A. van Dam, L. P. J. J. Noldus, and R. C. Veltkamp, "Automated Recognition of Social Behavior in Rats: The Role of Feature Quality," in *Proc. Conf. Image Analysis and Processing (ICIAP)*, 2015, pp. 565–574.

[13] S. Dasgupta and D. Hsu, "Hierarchical sampling for active learning," in *Proc. Conf. Machine Learning (ICML)*, 2008, pp. 208–215.

[14] T. S. F. Haines and T. Xiang, "Active Rare Class Discovery and Classification Using Dirichlet Processes," *International Journal of Computer Vision*, vol. 106, no. 3, pp. 315–331, 2013.

[15] L. P. J. J. Noldus, A. J. Spink, and R. A. J. Tegelenbosch, "EthoVision: A versatile video tracking system for automation of behavioral experiments," *Behavior Research Methods, Instruments, & Computers*, vol. 33, no. 3, pp. 398–414, 2001.

[16] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, pp. 209–230, 1973.