Rat behavior: human versus automatically generated annotation

Elsbeth A. van Dam and Lucas P.J.J. Noldus, Noldus Information Technology BV, Wageningen, The Netherlands.

The automatic recognition of rodent behavior has long since been on the wish list of laboratory researchers in the field of neuroscience and pharmacology. Rats and mice are used as models for human diseases and their behavior is studied in order to develop new drugs for neurological and psychiatric disorders. Accurate measurement of behavior is crucial to advance research in these fields. Labelling of behaviors is mostly done by humans. This is labor-intensive, and as is known from previous studies (List et al. 2004, Jhuang et al. 2010) error-prone and subject to individual interpretation.

In building automated behavior annotation systems these imperfect manual annotations serve as ground truth for evaluating the performance of the systems. Since we don't want to mimic the human errors in the system, we are left with a performance rate that cannot exceed the human performance. This poses a problem when we want to learn from the performance figures and use them to improve the system.

Human annotation errors can be incidental, like missed events or wrong key strokes due to concentration loss. Other well-known errors made by humans are bias and observer drift (observation influenced by context) (Lehner, 1979). Furthermore, human annotators are subjective in their interpretation of behavior and the precise timing of events (start and stop times and the tolerance for short pauses in behavior events). Automated systems do not suffer from these kinds of errors. They introduce their own mistakes, like errors caused by poor sensor quality or input artifacts, or by the inability to recognize behaviors that are not in the training dataset. Although outliers are detectable, in most cases unseen behavior is composed of normal movements and the automated system will switch between the behaviors that it is trained to recognize. These fundamental differences between human and machine annotation raise fundamental questions on the way we can evaluate the output of automated behavior recognition systems.

Apart from these method-specific errors, there is the inherent ambiguity in the behaviors itself that poses problems for evaluation. Although some rat behaviors are clearly defined like rearing (standing on the hind legs to explore the environment above it) and grooming, others are less explicit and have overlapping definitions, such as sniff, root, dig, gnaw and eat. Rats are often performing a mixture of these or something in between. Therefore, in the data itself there are events that are more unmistakable than others. It is only our artificial behavior classification that is discrete, not the continuous stream of behavior itself. Many studies circumvent this issue by making clipped datasets of unambiguous events and leave out the transitions or the ambiguous events (see Poppe 2010). But systems designed for real world applications need to do action detection as well as classification and therefore be validated on the entire video sequences. In this study we use three alternative ways to compare human to automatic annotation. From the results we get a reliable insight in the performance quality of the system and in how we can improve it.

We built a system for automated rat behavior recognition (ABR),as an extension of our video tracking system EthoVision (www.noldus.com/ethovision) that can recognize rat behaviors *'drink', 'eat', 'sniff', 'groom', 'jump', 'rear wall', 'rear unsupported', 'rest', 'twitch'* and *'walk'* from a continuous stream of video. ABR uses an overhead camera view, and does not need on-site training: the only input needed is the size of the cage and the animal.

For recognition, the system generates multiple features per frame, based on the shape and motion of the tracked animal and the temporal context. These features are then classified using a simple pipeline of dimension reduction, a normal density based classifier, followed by post processing of the resulting behavior probabilities. The final annotation is the class with the highest probability. In order to train the ABR system a dataset of >74 hours video was recorded of single housed rats in a home cage with infrared lighting. Subsets of these recordings were annotated by annotation experts, leading to a data set of >250,000 frames in 13 behavior classes.

We evaluated ABR in 3 different ways:
1) Frame-wise comparison to various videos annotated by experts. For testing robustness, test videos were recorded with different video resolution, rat strain, lighting and background. The annotations were revised to make them frame-accurate. Frame-accurate annotation is not the normal way of annotating in behavioral research; it is very laborious and the usual timing discrepancy does not influence research results. It takes over 2 hours to annotate 10 minutes of video (25 frames per second). Revising the annotation we also repaired obvious mistakes, such as missed events, invalid durations or wrong key strokes due to concentration loss while scoring. Still, it is not a perfect ground truth due to the inherent ambiguities in the data and overlapping behavior definitions. With this frame-wise comparison we show that the system performs equally well as humans and other known systems (Jhuang et al.2010).
2) Comparison to human annotation by comparing effects found in a treatment-effect study. For this, a group of 4 rats was treated with two types of psychopharmaca: a stimulant drug (Amphetamine) and a sedative drug (Diazepam). The experiments have been performed in adherence to the legal requirements of The Netherlands concerning research on laboratory animals (Wod/Dutch 'Experiments on Animals Act') and have been approved by an Animal Ethics Committee ('Lely-DEC'). This dataset consists of 6.7 hours of video annotated by an expert. The annotation is not frame-accurate. As long as both methods are internally consistent in their scoring, the different latency will not affect the results. Results in table 1 show that both methods agree on most of the effects. However, there are also some disagreements. These pointed to systematic errors in both human and automatic annotation.
3) Comparison to human annotation by measuring the correlation of behavior frequency and duration in 5 min intervals. We used the same dataset as was used in the previous comparison. Interval comparison is another good alternative to frame wise comparison if annotation is not frame accurate. For both methods and treatments, figure 1 shows the difference in frequency results per behavior. The absolute frequencies of ABR are much higher due to more fragmented annotation of ABR. However, since the frequency effects are similar, we can conclude that this annotation difference is not affecting research results.

	Amphetamine		Diazepam	
	ABR	Hum	ABR	Hum
drink	-	-	none	-
eat	none	-	-	-
explore	+	+	-	none

groom	-	-	-	-
jump	none	none	-	none
rear	+	+	-	-
rest	-	-	+	+
walk	+	+	-	none

*Table 1. Significant treatment effects (increase(+), decrease(-) or no effect(*none*)) found by both ABR and human annotation resulting from a two-tailed Wilcoxon signed-rank test on the log durations per 5-min intervals (p <= 0.05) for treated animals and their control group. There are 11 agreements and 5 disagreements. Close inspection of the disagreements revealed both human and ABR mistakes.*

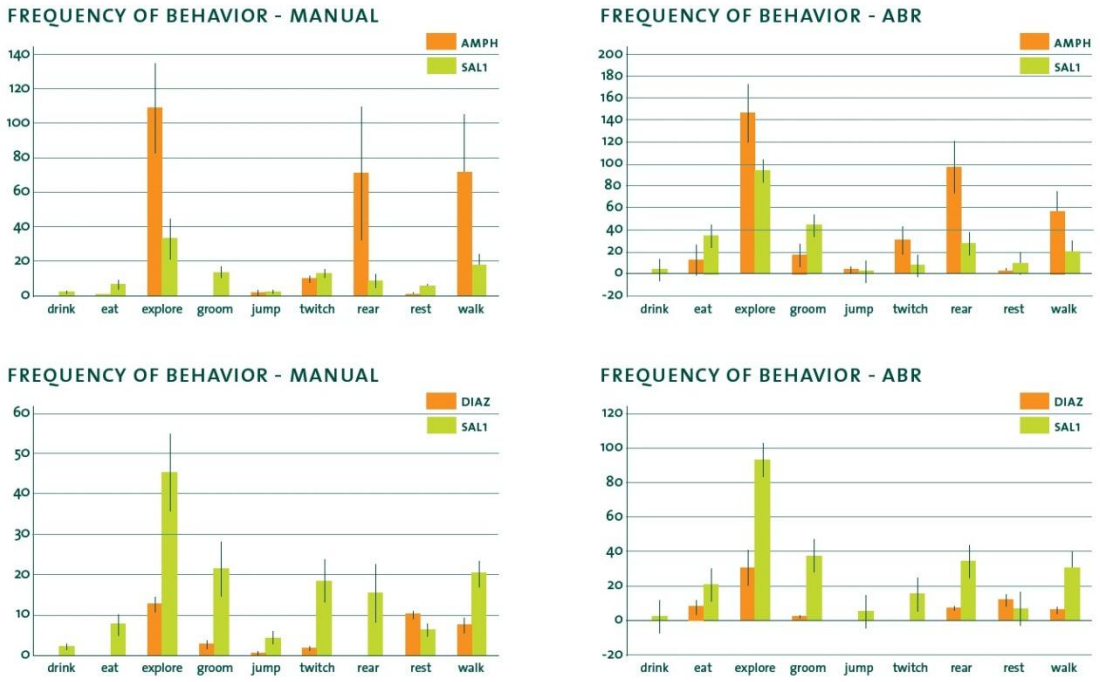

Figure 1 Differences in behavior frequencies between treated animals and their control group, for both treatments and methods. For both the Amphetamine treatment and the Diazepam treatment, the methods (manual scoring and ABR) agree on the direction and strength of the treatment effect on all behaviors frequencies except 'twitch'. However, the annotations differ in absolute frequencies.

There are errors that don't influence research results and errors that do. The first are tolerable, the latter are not. To our knowledge there is no error measure that takes this into account. We need to have an error measure that accounts for the fact that some mistakes are worse than others. It should be tolerant for differences in latency and on behavior boundaries. It should punish confusion between unambiguous classes more than confusion between overlapping behavior classes. Well-known reliability measures used in behavioral research, as Cohen's kappa (Cohen 1960), were developed for the comparison of human annotations and are mainly good at tolerating differences in scoring latencies between observers. They cannot deal with tolerable differences between human and automatic annotations, such as frequency

differences. One way to circumvent this is to do treatment-effect comparison as mentioned above, but this is restricted to the effects that can be measured. For frame-wise comparison a possible solution could be to tolerate multiple labels at event boundaries and at ambiguous segments, and to add a certainty to both annotations and use it in the error measure. Whether or not such measures are allowable and sufficient is to be discussed by the behavior recognition community.

Often the human annotation is used as the ground truth and it is said that the automatic annotation system cannot be expected to perform better than the inter-observer agreement, which is around 75% for a 10 class-problem (Jhuang et al. 2010). However, our investigations show that the types of mistakes made by an automated system are fundamentally different than the mistakes of human annotators. This leads to the conclusion that the most commonly used evaluation method for behavior recognition can and should be improved. When we want to improve automated systems, we need to have better, more accurate performance measures that tolerate acceptable mistakes and punish intolerable errors, or else we cannot measure performance beyond the 75% of the human annotation skills. Distinguishing between types of errors and identifying them is a first step. We already use human annotation to improve our automated systems. We can also use the automatic annotation to trace errors in the human annotation.

Acknowledgements

We would like to thank Rob Ottenhoff (Noldus Information Technology, Netherlands), Leen Raeymakers (Janssen Research and Development, Beerse, Belgium), Johanneke van der Harst (Delta Phenomics, Netherlands) and Cajo ter Braak (Wageningen University, Netherlands) for their contribution. This research was partially financed by grants from Agentschap NL (NeuroBasic-PharmaPhenomics) and NWO (SenseWell).

References

1. J. Cohen, A coefficient of agreement for nominal scales. *Educat. Psychol. Measurem.* 1960; 20: 37–46.
2. H. Jhuang, E. Garrote, X. Yu, V. Khilnani, T. Poggio, A. D. Steele, and T. Serre. Automated home-cage behavioural phenotyping of mice. *Nat Commun*, 1(6):68, Sep 2010.
3. P.N. Lehner. Handbook of ethological methods. New York: Garland STPM Press 1979.
4. T. List, J. Bins, J. Vazquez, and R. Fisher. Performance evaluating the evaluator. *Visual Surveillance and Performance Evaluation of Tracking and Surveillance* 0:129–136, 2005.
5. R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976 – 990, 2010.
6. T. Sharpe and J. Koperwas. Behavior and sequential analysis: principles and practice. Sage Publications Inc, 2003.