JAABA: interactive machine learning for automatic annotation of animal behavior

Mayank Kabra[1,4], Alice A Robie[1,4], Marta Rivera-Alba[1,2], Steven Branson[1,3] & Kristin Branson[1]

We present a machine learning–based system for automatically computing interpretable, quantitative measures of animal behavior. Through our interactive system, users encode their intuition about behavior by annotating a small set of video frames. These manual labels are converted into classifiers that can automatically annotate behaviors in screen-scale data sets. Our general-purpose system can create a variety of accurate individual and social behavior classifiers for different organisms, including mice and adult and larval *Drosophila*.

Quantitative measurement of the behavior of model organisms is an important tool for understanding genetics, evolution, development and the nervous system, and it has led to insights into human diseases and behaviors. Many applications involve large-scale screens in which the behavior of thousands of animals must be compared. The scale of these experiments necessitates high-throughput, automated approaches. For both large screens and smaller-scale experiments, accurate, detailed quantification of behavior will result in richer information about the effects of a manipulation and will enable discovery of subtle behavioral differences undetectable through existing methods.

Video of behaving animals contains a wide breadth of detailed information about the behavior of the animals. Recently, computer-vision techniques have emerged for automatically tracking the animals[1–5], transforming video data into trajectories of their positions over time.

We present our general-purpose, open-source software for allowing biologists to encode their intuition about the structure of behavior and to transform the trajectories output by these trackers into higher-order, scientifically meaningful statistics of behavior. Our system, the Janelia Automatic Animal Behavior Annotator (JAABA), uses state-of-the-art machine learning methods to allow users to create a variety of automatic behavior classifiers. These classifiers input the animals' trajectories computed by a tracking system, and they output time series indicating whether each animal is performing a given behavior in each video frame.

In this paper, we quantitatively show that JAABA: (i) works well across a wide range of behaviors and multiple species and organism body plans (adult *Drosophila*, mice and *Drosophila* larvae), (ii) is applicable to the diversity of behavior in screen-scale experiments, (iii) is usable by biologists who are experts in behavior but not computer science and (iv) produces behavior statistics that can be used to understand subtle behavior differences between populations of animals.

The input to JAABA is video of the animals behaving as well as their trajectories. JAABA's user interface (**Fig. 1**) allows users to observe this video and add labels to frames in which they are certain of the animals' behavior. Users label a selected animal in a selected frame as either performing the given behavior (for example, 'Touch') or not ('None'). These labels are the medium of communication through which users transmit their intuition to the underlying machine learning system (**Fig. 1**). The machine learning algorithm searches for the classifier function that inputs the trajectories and best reproduces these manual labels. Quickly, within 15–40 s, a new behavior classifier is trained, and visualizations of it and its performance are returned to the user. The user can investigate the classifier's performance on any frame for the current animal or another animal in the same video or another video to understand the classifier's current state and to find frames in which the classifier is either predicting incorrectly or has low confidence. The user can then label these frames, retrain the classifier and repeat (**Supplementary Video 1**). The final output of JAABA is the last classifier trained, which can be used to automatically annotate new (tracked) videos with high throughput.

JAABA has been adapted to work with the outputs of several tracking systems (Online Methods). Our fly behaviors were based on the trajectories output by Ctrax[1], which consisted of the five-dimensional ellipses fit to each fly (wing extension and double wing flick also use tracked wing positions). Mouse tracking output was of a similar form, and the larva tracking output consisted of an 11-point skeleton and larva area.

JAABA first transforms these trajectories into a novel, efficient, general-purpose representation amenable to machine learning. A suite of 'per-frame' features are computed from the trajectory data, describing the state of the animal in the current frame: for example, the instantaneous speed of the animal (**Fig. 1** and **Supplementary Fig. 1a**). From these, JAABA computes a general set of 'window' features that provide temporal context from a window around the current frame (**Fig. 1** and **Supplementary Fig. 1b**). These novel window features are fast, yet they encode information in a manner

[1]Howard Hughes Medical Institute, Janelia Farm Research Campus (HHMI JFRC), Ashburn, Virginia, USA. [2]Development, Evolution and the Environment Laboratory, Instituto Gulbenkian de Ciência, Oeiras, Portugal. [3]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to K.B. (bransonk@janelia.hhmi.org).

RECEIVED 16 JUNE; ACCEPTED 6 NOVEMBER; PUBLISHED ONLINE 2 DECEMBER 2012; DOI:10.1038/NMETH.2281

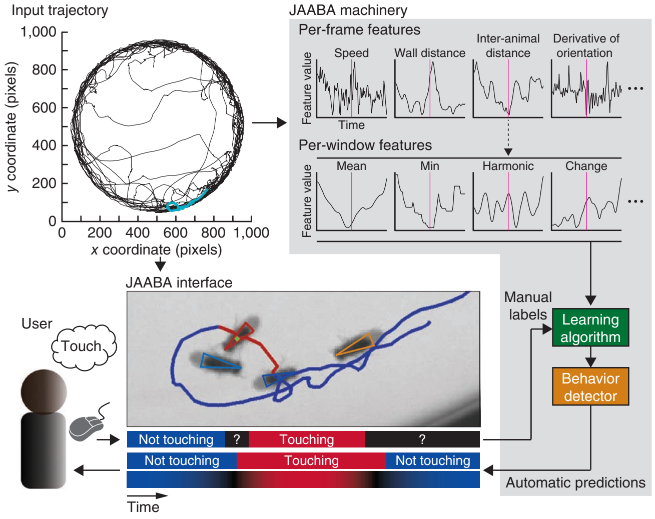

Figure 1 | JAABA overview. Input trajectory (top left): the *x*,*y* position of a fly over 1,000 s of video is plotted in black (cyan portion corresponds to other plots). This video and trajectory are input to JAABA. JAABA interface (bottom): the users interact with the JAABA interface to encode their definition of the behavior (for example, touching). Users are shown video of the animal (image) overlaid with its tracked position (bold triangle) and its future and past trajectory (bold line). They encode their definition by labeling frames (top timeline) in which they are certain the animal is (red) or is not (blue) performing the behavior. Unlabeled frames are indicated in black. These labels are passed to the JAABA machinery, which creates a new behavior detector. Visualizations of this behavior detector are returned in the middle and bottom timelines, which show the detector's predictions (color) and confidence (saturation). JAABA machinery (gray shading): the underlying JAABA machinery inputs the animals' trajectories. 'Per-frame' feature time series are computed from the input trajectories (examples in first row). Each plot corresponds to an example per-frame feature (full set in **Supplementary Note**). The pink line indicates the frame shown. 'Window'-feature time series are computed from the inter-animal distance per-frame feature (full set in **Supplementary Note**) (second row). This window feature–based representation and the manual behavior labels are input into the learning algorithm (green), which finds the automatic behavior detector (orange) that inputs the window features and best predicts the input labels.

compatible with machine learning algorithms (**Supplementary Fig. 1c**). Second, the highly processed representation of the animals' trajectories and the manually entered behavior labels are input into the GentleBoost learning algorithm[6], which we have optimized for learning speed (**Supplementary Fig. 1d**). Details of the algorithm and software are discussed in the Online Methods and **Supplementary Note**.

This interactive framework—in which, iteratively, the user labels then retrains the classifier—is a practical adaptation of active learning[7], and it is enabled by the speed of JAABA's underlying machinery. It allows the user to label frames that are most informative to the learning algorithm and for which the user is confident of the correct behavior class (**Supplementary Fig. 2**). Besides allowing accurate classifiers to be trained efficiently, this interactive framework, combined with JAABA's user-friendly interface and visualizations of the underlying classifier, makes JAABA accessible to users without in-depth knowledge of machine learning.

To demonstrate that JAABA could successfully be used by novice users with little expertise in machine learning, 12 volunteers with expertise in fly behavior but no prior experience using JAABA trained fly chase behavior detectors. On their own laptops, after a 15-min introduction, the volunteers were able to independently train chase detectors with an average ground-truth error rate of 2.4% (**Supplementary Fig. 3a** and Online Methods).

To show that JAABA was general purpose and accurate, we trained and measured the accuracy of a diverse set of single-animal and social behavior detectors for adult *Drosophila*, mice and *Drosophila* larvae. We trained 15 behavior detectors for adult *Drosophila* to describe a variety of locomotion and social behaviors we observed in flies in an open-field environment (Online Methods): walk, crabwalk, chase, back up, jump, stop, touch, copulation, wing extension[1,2], wing grooming, tail pivot, center pivot, righting, attempted copulation and double wing flick (**Fig. 2a**, **Supplementary Video 2** and **Supplementary Note**). With this set, ~90% of frames were assigned a behavior. We trained two detectors for *Drosophila* larvae: head cast and crawl[5,8] (**Fig. 2b**). We trained two detectors for mice: walk and follow[3] (**Fig. 2c**; see Online Methods and **Supplementary Table 1** for training methodology). These behaviors were not mutually exclusive, and each behavior detector classified whether an animal was or was not performing a given behavior. For a few behaviors, to enforce exclusivity, the output of one classifier was used as the input of another.

To measure the accuracy of these automatic behavior detectors in replicating human annotations, for each behavior, we manually labeled whether the animal was performing the behavior in a randomly chosen subset of 6,000–20,000 frames with three levels of confidence: important, unimportant and unknown (Online Methods and **Supplementary Table 2**). The error rates measured (**Fig. 2**) demonstrate that JAABA could be used to create a diverse set of accurate behavior classifiers.

Our automatic chase behavior detector replicated human annotation with an error rate of 0.6%, comparable with the 2.4% interannotator disagreement rate observed. It outperformed the chase behavior detector distributed with Ctrax[1], which achieved an error rate of 10.4%. The CADABRA chase definition[2] resulted in no chase detections in our videos (error rate of 50%) (Online Methods). Part of the difference in performance is due to the static Ctrax and CADABRA classifiers being tuned for different environments, which illustrates the necessity of adapting behavior definitions to the environment.

To show that JAABA could be used to create behavior detectors that perform well across a large, phenotypically diverse screen, we examined the performance of a subset of the fly behavior classifiers on data generated as part of an ongoing neural activation screen of the Rubin *GAL4* collection[9]. This data set consisted of 14,524 16-min videos of 2,144 different lines (in total, 168 full days of video of 290,480 flies). The activation of small subsets of neurons caused a wide variety of phenotypes (**Supplementary Video 3**). We generalized the classifiers to include this variability by examining their predictions on a variety of videos and iteratively adding labels for misclassified or low-confidence frames and retraining, thus making extensive use of JAABA's interactivity. To approximate performance across all *GAL4* lines, we measured the classifiers' accuracy for 16 lines that spanned the behavior

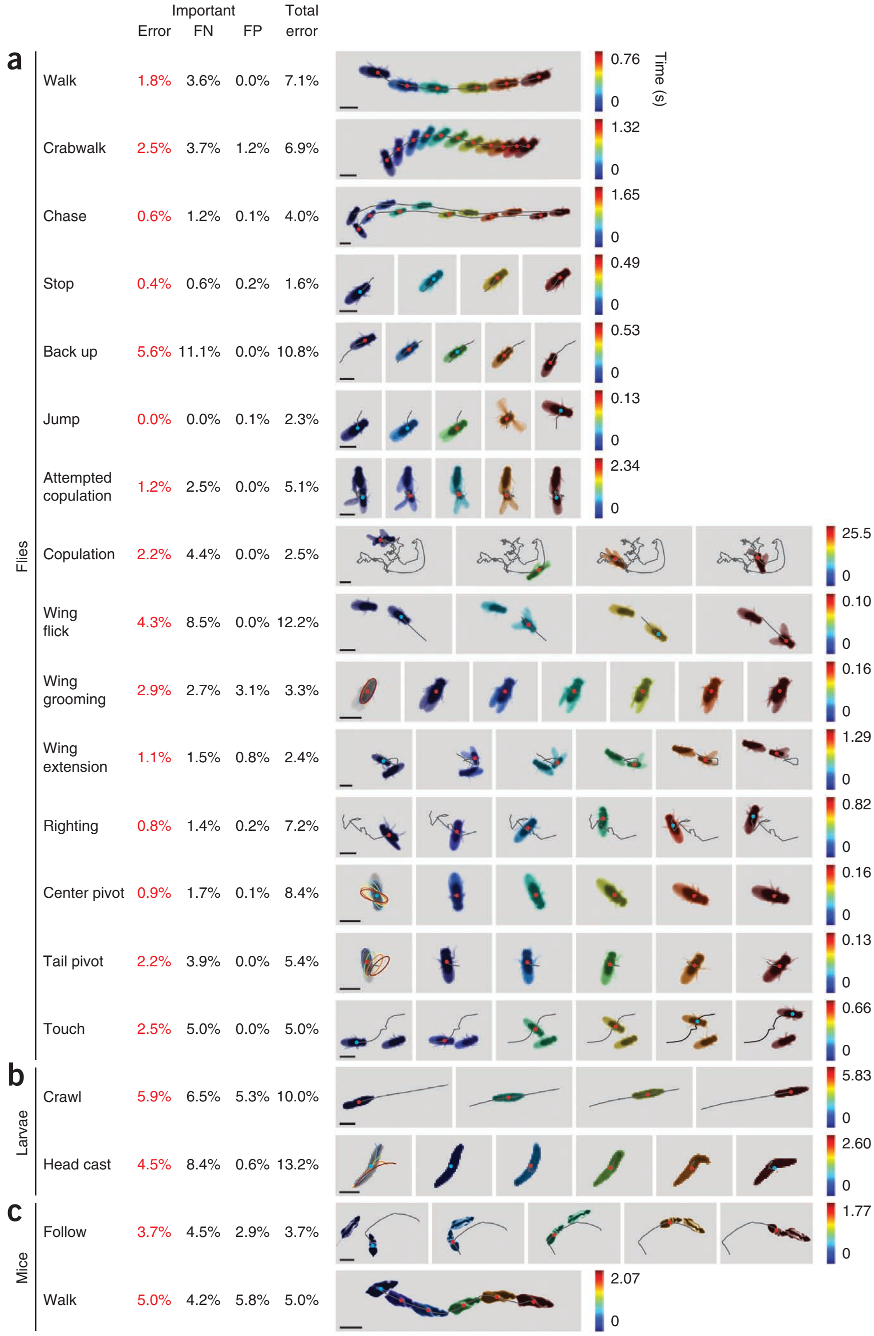

Figure 2 | Behavior detector accuracy across behaviors and organism types. (**a**–**c**) Each row corresponds to a different behavior. The first three columns with numeric values describe the total error rate, false negative (FN) rate and false positive (FP) rate on ground-truth data labeled as important. The fourth column indicates the total error rate on all labeled ground-truth data (important and unimportant). Using measured inter-annotator confusion rates, we determined that assessing error on important frames best measures the classifier's performance (Online Methods). The average error rate over all behaviors was 1.9% for flies (**a**), 5.2% for larvae (**b**) and 4.4% for mice (**c**). Complete results are shown in **Supplementary Table 2**. Right, selected frames from example intervals describing each behavior. For walk, crabwalk and chase, we overlaid the animals' positions in each frame. Otherwise, each frame is drawn separately. For social behaviors (such as chase, touch and follow), we also show the animal that the labeled animal is interacting with. Animal color indicates time. Scale bars, 2 mm for flies and larvae; 5 cm for mice. The main animal's trajectory is plotted in gray. Its centroid position is indicated with a circle whose color indicates whether it is predicted as performing (red) or not performing (blue) the behavior by the automatic detectors. For wing grooming, tail pivot, center pivot and head cast, the first frame also indicates the animal's position in every selected frame.

profiles observed and found an average error rate of 0.6% (Online Methods, **Fig. 3** and **Supplementary Table 3**). We also confirmed by eye the classifiers' performance on lines automatically predicted to perform each behavior most (top 2%). Thus, we showed, for the first time to our knowledge, the successful use of machine learning to create behavior classifiers that accurately annotated a large set of phenotypically diverse data.

To show that the behavior statistics returned by JAABA could be used to detect and understand subtle behavioral differences between populations of animals, we performed three types of comparisons for adult flies: six strains of wild type, four starvation conditions and four age conditions. For each of these, we trained a logistic regression classifier to distinguish these subtly different strains or conditions on the basis of only the fraction of time the flies perform each of ten behaviors (**Supplementary Note**). We also investigated which behaviors the strains or conditions differed on, to better understand how they varied.

The six strains of wild-type *Drosophila melanogaster* (Berlin, Oregon-R, Dickinson and three populations of Canton-S) could all be reliably distinguished from each other with, on average, 98% accuracy. Several interesting differences between the strains were found (**Supplementary Fig. 4**), including that Berlin flies performed the most touches and chases, whereas Dickinson flies performed the most jumps and righting maneuvers. Among the four starvation conditions compared (0, 6, 12 and 24 h starved), 0-h starved flies could be reliably distinguished from 12- and 24-h-starved flies with 93% and 92% accuracy, respectively. As expected, increased starvation corresponded with increased activity (less stopping, more walking); however, not all behaviors increased with starvation: jumping, righting and backing up were not significantly affected (**Supplementary Fig. 5**). The four age conditions (2, 6, 13 and 20 d old) could all be reliably distinguished from each other with, on average, 91% accuracy (**Supplementary Fig. 6**).

As demonstrated with these analyses, applying multiple automatic behavior detectors provided insight into how the individual elements of the animals' behaviors differed as well as a powerful tool for detecting effects of experimental manipulations.

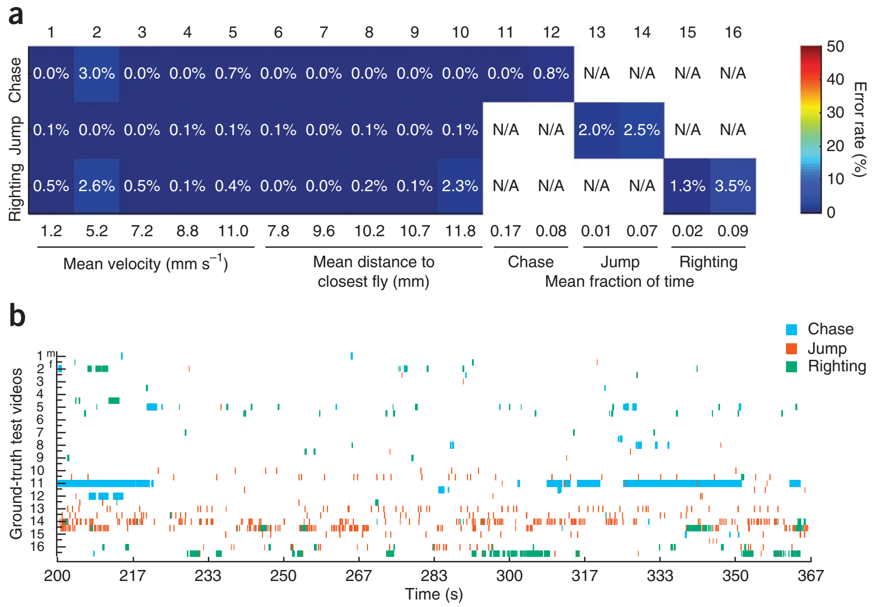

Figure 3 | Behavior classifiers generalize over a phenotypically diverse data set. (**a**) A matrix of the total ground-truth error rate for three behavior classifiers (rows) and 16 representative *GAL4* lines from the *TrpA* activation screen (columns). Each error rate was calculated from a separate set of 2,000 frames of ground-truth labels. The classifiers generalized successfully with a mean error rate of 0.6%. For full results, see **Supplementary Table 3**. Videos 1–5 were sampled to span the range of observed average speed of the flies, with the corresponding average speed in each video shown below its column. Videos 6–10 were similarly selected to span the range of observed average closest fly distance, with the corresponding mean distance shown below its column. For each of the three behaviors, two lines were selected that overperformed the given behavior. The fraction of time spent performing the respective behavior is shown below the columns. N/A is shown for behavior-line combinations for which accuracy was not measured. For details on how videos were selected, see the **Supplementary Note**. (**b**) Ethograms of flies from the ground-truth test set. Results of the behavior detectors chase (blue), jump (orange) and righting (green) for a 167-s period (5,000 frames) are shown for 32 flies from the 16 ground-truth videos. For each video, a male (m, major tick) and female (f, minor tick) were randomly selected.

The changes measured in these basic locomotion and social behaviors can represent an overall behavioral phenotype that reflects an internal physiological state such as age or starvation.

JAABA is freely available for download at http://jaaba.sourceforge.net/ (**Supplementary Software**). JAABA operates on the output of tracking systems including the three trackers for flies, larvae and mice demonstrated in this paper as well as several published trackers[2,4,5,8], and it is thus complementary to these systems. Several other systems also perform automatic behavior classification[1–4,8,10–12]. The interactive, general-purpose, learning-based approach used by JAABA allows it to surpass these approaches in its ability to train accurate classifiers for a wide variety of complex behaviors; to generalize across screen-scale, phenotypically diverse data sets; and to be usable by members of the biology community without computer science expertise. We have also found that, for behaviors that were less well defined, this interactive framework allowed users to explore behavior definitions by first labeling canonical examples then solidifying their definition on boundary cases. We envision that JAABA will enable biologists to convert high-dimensional tracker output into meaningful, interpretable statistics.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank A. Miller and R. Egnor (HHMI JFRC) for creating the mouse behavior detectors; R. Egnor and S. Ohayon (Caltech) for providing their mouse data before publication; P. Keller for help in preparing the manuscript and helpful discussions; H. Kelstrup and C. Mirth (Instituto Gulbenkian de Ciência) for providing their larva data; G. Rubin (HHMI JFRC) for providing *GAL4* lines before publication; G. Card and U. Heberlein (HHMI JFRC) for sharing flies; the Janelia Fly Olympiad Team (particularly J. Hirokawa, L. Umayam, W. Korff, M. Phillips and S. Roberts) and the Janelia Fly Facility (particularly J. MacMahon and T. Laverty) for helping in the *TrpA* neural activation fly screen; B. Arthur and J. Liu for help in developing the JAABA software; U. Heberlein, P. Perona, M. Zlatic, T. Jovanic, E. Hoopfer, B. Afonso, M. Gershow and C. Schusterreiter for helpful discussions; S. Albin, Y. Aso, R. Azanchi, B. Baker, G. Card, C. Dan, E. Hoopfer, V. Jayaraman, W. Korff, G. Ophir, Y. Pan and A. Seeds for participating in our study of JAABA's usability; and W. Karsh for acronym help.

AUTHOR CONTRIBUTIONS

K.B. conceived of the project with help from M.K., A.A.R. and S.B. M.K. and K.B. designed and wrote the software with help from A.A.R., S.B. and M.R.-A. A.A.R. created and tested the fly behavior detectors with help from K.B. A.A.R. developed and performed the fly experiments with help from K.B. and M.K. M.R.-A. developed the larva tracker and larva behavior detectors with help from K.B. and M.K. K.B., A.A.R. and M.K. wrote the paper with help from M.R.-A.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nmeth.2281. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Branson, K., Robie, A.A., Bender, J., Perona, P. & Dickinson, M.H. *Nat. Methods* **6**, 451–457 (2009).
2. Dankert, H., Wang, L., Hoopfer, E.D., Anderson, D.J. & Perona, P. *Nat. Methods* **6**, 297–303 (2009).
3. de Chaumont, F. *et al. Nat. Methods* **9**, 410–417 (2012).
4. Swierczek, N.A., Giles, A.C., Rankin, C.H. & Kerr, R.A. *Nat. Methods* **8**, 592–598 (2011).
5. Luo, L. *et al. J. Neurosci.* **30**, 4261–4272 (2010).
6. Friedman, J., Hastie, T. & Tibshirani, R. *Ann. Stat.* **28**, 337–407 (2000).
7. Settles, B. Active learning literature survey (Computer sciences technical report 1648) ⟨http://research.cs.wisc.edu/techreports/2009/TR1648.pdf⟩ (Univ. Wisconsin-Madison, 2009).
8. Gomez-Marin, A., Stephens, G.J. & Louis, M. *Nat. Commun.* **2**, 441 (2011).
9. Jenett, A. *et al. Cell Rep.* **2**, 991–1001 (2012).
10. Martin, J.R. *Behav. Processes* **67**, 207–219 (2004).
11. Robie, A.A., Straw, A.D. & Dickinson, M.H. *J. Exp. Biol.* **213**, 2494–2506 (2010).
12. Aggarwal, J.K. & Ryoo, M.S. *ACM Comput. Surv.* **43**, 16 (2011).

ONLINE METHODS

Tracking. JAABA operates on the output of any tracking algorithm. In this work, three different tracking algorithms were used for flies, larvae and mice. Additionally, JAABA has been adapted to work with the output of several published tracking systems[2,4,5,8].

Flies were tracked using Ctrax[1]. The output of this tracker was an ellipse fit to each fly over time, which was represented by the centroid, orientation and major- and minor-axis lengths. For two of the behaviors, wing extension and double wing flick, we performed a second step of tracking to fit the positions of the wings using custom MATLAB (MathWorks) software (**Supplementary Note**). For each fly, the relative angles of the wings and the areas of each wing are output by the tracker.

Larvae were tracked with custom MATLAB software (**Supplementary Note**). The output of this tracker was the 11-point skeleton fit to each larva's body and its area (in total a 23-dimensional vector).

Mice were tracked using the MoTr tracking software (S. Ohayon and R. Egnor, Caltech and HHMI JFRC). The output of this tracker was also an ellipse fit to each mouse over time.

JAABA user interface. To train an accurate behavior detector that works well on large, diverse data sets, users must provide training labels that sufficiently cover the behavior space. In addition, to ensure that the classifier is working well across the data set, users need to visualize the classifier's performance across large amounts of data. To enable users to sift through multiple lengthy movies to find and label informative frames, JAABA provides many intuitive ways of navigating across time, between flies and between movies (**Supplementary Software**). For example, to aid users in finding incorrect or low-confidence predictions, JAABA includes tools for navigating to predicted bouts of the behavior, low-confidence frames, errors in training data and cross-validation errors (**Supplementary Note**). Users can also quickly navigate to parts of the movie in which a selected per-frame feature has a high or low value. For instance, jumps can be located by navigating to frames in which the animal's speed is high. To aid users in deciding to examine a different animal or video, JAABA provides tables of statistics of each animal: for example, information about how the classifier's predictions on the animal have changed as the classifier changed. Users can obtain feedback on how well their current behavior classifier will generalize to novel data by examining the cross-validation error.

JAABA algorithm overview. The goal of JAABA's algorithm is to create automatic behavior classifiers. From the input trajectory data, a suite of 'per-frame' features are computed to describe the state of the animal in the current frame, such as the instantaneous speed of the animal. Per-frame features have limited ability to discriminate behavior classes on the basis of a single threshold (**Supplementary Fig. 1a**). From these per-frame features, JAABA computes a general set of 'window' features that provide temporal context from a window around the current frame. This transformation improves their ability to discriminate the behavioral classes (**Supplementary Fig. 1b**). This feature computation is detailed below.

Even with these more powerful features, a single thresholding rule is rarely sufficient to accurately predict the behavior class. JAABA uses the GentleBoost[6] learning algorithm to combine many such thresholding rules into a more accurate classifier. These rules are added iteratively by the boosting algorithm with the goal of moving examples away from the classifier's decision boundary, increasing the separation between the two classes (**Supplementary Fig. 1c**). Both the training and generalization error rate decrease as separation increases (**Supplementary Fig. 1d**).

JAABA uses an interactive framework in which, iteratively, the user selects and labels frames that will be informative to the current classifier, then trains a new classifier (**Supplementary Fig. 2**). Examples on which the classifier predicts incorrectly or has low confidence often provide new information to the training set, as these examples are likely dissimilar to all the current training data. These data points lie in regions that are sparsely populated by the current training set and are close to the decision boundary of the classifier. Combined with the tools for navigating within and between videos, the classifier's raw scores can thus be used to find informative frames to label.

JAABA feature computation. JAABA inputs the animal trajectories described above, computed by automatic tracking systems. Per-frame features are simple, engineered features that can be computed from the positions of single or multiple animals in 1–3 consecutive frames.

Per-frame features can be broadly grouped into locomotion, landmark-based, appearance-based and social categories (**Fig. 1**). Locomotion features describe the movement of a single animal, such as the speed or change in orientation of the animal. Landmark-based features describe the position or movement of an animal in relation to landmarks in the environment, such as the distance to the arena wall. Appearance-based features describe the pose of the animal in a single frame, such as the area of the ellipse (flies, mice) or the bend of the spine (larvae). Social features describe the position or motion of an animal relative to other, nearby animals, such as inter-animal distance or difference in orientation between the given animal and the nearest other animal. Descriptions of the 63 per-frame features developed for flies and mice, 32 wing-based per-frame features and 77 per-frame features developed for larvae are in the **Supplementary Note**.

In addition to trajectory-based per-frame features, predictions from existing behavior detectors can be used as a per-frame feature. For example, the output of the jump behavior detector for flies was used as an input for the righting behavior detector.

We have engineered large, inclusive sets of per-frame features for use with flies, larvae and mice. However, adapting JAABA to new types of organisms, environments or behaviors may require development of new per-frame features. For example, to detect behaviors such as antennal grooming, we may require spatiotemporal features computed directly from the video[13]. JAABA is structured so as to make it painless to add new types of per-frame features.

To provide temporal context information, we use window features that describe the distribution of per-frame features in time windows around the current frame. We use multiple window sizes and temporal offsets relative to the current frame. Within each window we compute the mean, s.d., minimum, maximum and other functions (**Fig. 1** and **Supplementary Fig. 1b**). Details of the 11 types of window features are in the **Supplementary Note**. All the window features can be computed efficiently using

doi:10.1038/nmeth.2281

convolution or image morphology methods. In our MATLAB implementation, window features were computed at a rate of 0.4 μs per feature per frame. This equated to 3–10 ms per frame for our classifiers (dual quad-core 2.67-GHz machine).

The total number of window features computed depends on the behavior. Our learning algorithm examined 15,088 window features for the fly wing extension behavior, 9,225 for the fly chase behavior, 4,602 for the fly righting behavior, 2,337 for the fly jump behavior, 22,991 for the larva head-cast detector and 3,720 for the mouse follow detector.

JAABA learning algorithm. We use the GentleBoost learning algorithm to train the classifier[6]. Boosting algorithms such as GentleBoost combine many weak rules to learn an accurate classifier. Our weak rules are decision stumps, which threshold a single, selected feature. In each boosting iteration, the learning algorithm adds the weak rule that best increases the separation of the two classes in the training data (**Supplementary Fig. 1c**).

We have made several modifications to the standard GentleBoost algorithm to improve training speed (**Supplementary Note**). These improvements resulted in training being basically independent of the number of training examples, requiring between 15 and 40 s (dual quad-core 2.67-GHz machine), depending on the number of window features used.

JAABA implementation. The software was implemented in MATLAB. We freely provide both the source code and compiled executables for Windows, Mac and Linux (http://jaaba.sourceforge.net/; **Supplementary Software**).

Flies. Groups of 20 *Drosophila melanogaster* (10 males and 10 females) were recorded at 30 frames per s (f.p.s.) for 1,000 s with a digital camera (Basler A622f) in a 12.7-cm-diameter arena developed for high-throughput screening of walking fruit flies adapted from Simon and Dickinson[14]. *GAL4* lines from an ongoing neural activation screen were used (A.A.R., unpublished data). The *GAL4* driver-line collection was provided by G. Rubin[9]. These lines were crossed to +;*UAS-dTRPA1/(CyO)*;+ (X and third chromosome back-crossed to Canton-S from M. Heisenberg for six generations). The control line used for these experiments was pBDPGAL4U (a *GAL4* insertion without a promoter in *w^{1118};attp2*)[15]. Wild-type strains used in this paper include stocks maintained by the Janelia Fly Core Facility: Canton-S from the Rubin laboratory (CS-GR), Canton-S from the Heisenberg laboratory via the Anderson laboratory (CS-MH) and the Dickinson laboratory stock (DL), a laboratory population derived from 200 wild-caught females. Additionally, wild-type strains maintained by the Heberlein laboratory—Canton-S (CS-UH), Oregon-R (OR) and Berlin (WTB)—were used. See **Supplementary Note**.

Larvae. Groups of ten mid-third instar larvae were recorded for 50 min at 30 f.p.s. with a digital camera (Sony DCR-HC52) on a 60-mm Petri dish. Three species of *Drosophila* were tested independently: *D. melanogaster* (Oregon-R), *D. virilis* and *D. willistoni* (H. Kelstrup, C. Mirth and M.R.-A., unpublished data). See **Supplementary Note**.

Mice. Four mice (two males and two females) were housed in a 0.6 × 0.6 × 0.6–m polycarbonate population cage and filmed at 30 f.p.s. for 5 d with a digital camera (Basler A622f). The mice were C57Bl/6J (Jackson Labs) (S. Ohayan and R. Egnor, unpublished data). Behavior classifiers were trained and ground truth established by A. Miller and R. Egnor. See **Supplementary Note**.

Novice user study. We presented a tutorial on our JAABA system at an informal meeting at our institute and asked for volunteers to train their own chase detectors and label ground-truth data fly videos. Our presentation was approximately 15 min long. We explained how to download and install JAABA and our sample data set, the basic usage of the JAABA GUI and how to annotate ground truth for the final classifier.

Twelve volunteers responded, and the accuracy of their classifiers on their own ground-truth data sets is shown in **Supplementary Figure 3a**. These volunteers downloaded and installed JAABA on their own computers and trained their own chase detectors using two videos we collected, one containing pBDPGAL4U controls and the other a *GAL4* line that we identified as chasing more frequently. For all data used, we had previously tracked the flies and set all parameters for JAABA. The volunteers each also labeled the same 2,000 frames of ground-truth data from one control video, which we selected using our chase classifier to bias random selection toward positive examples. Although different users used different labeling strategies, the interactive system allowed all users to create accurate classifiers, with final error rates ranging between 0.0% and 6.1%.

All volunteers labeled the same ground-truth data, which allowed us to examine inter-annotator disagreement rates. The volunteers annotated ground-truth labels with three levels of certainty—important, unimportant or unknown—to reflect their confidence. We compared inter-annotator disagreement for important and unimportant frames, and we found that inter-annotator confusion rates were low for important frames (1.0%) and high for unimportant frames (18.6%) (**Supplementary Fig. 3b**). This motivated our choice of using important frames to assess classifier performance. Unless otherwise noted, error rate refers to error on important frames.

Training behavior classifiers. Our training methodology is detailed in the **Supplementary Note**. The training sets were of varying sizes, as chosen while interacting with JAABA. For flies, we trained the behavior classifiers using varying combinations of pBDPGAL4U control flies and a hand-picked set of *GAL4* lines observed to have different phenotypic properties, detailed in the **Supplementary Note** and **Supplementary Table 1**. For larvae, we trained the classifiers on two videos from each of the three drosophilid species. For mice, we trained the classifiers on one 30-min video.

Measuring classifier accuracy. To quantitatively evaluate the performance of the classifiers, we randomly selected animals and intervals of video outside the training data and then manually labeled whether a selected animal was performing the behavior in the selected frames. For flies and larvae, we labeled with three levels of certainty. Frames labeled 'important' were those for which we were absolutely certain of the behavior label. A high error rate on these frames would cause us not to use the classifier. Frames labeled 'unimportant' were those in which we had an opinion on what the correct behavior label was but either were not confident

doi:10.1038/nmeth.2281

in this label (for example, frames at transitions between behaviors) or thought correct classification would be difficult for any classifier (for example, frames containing input tracking errors). Frames labeled 'unknown' were those in which we were unable to determine the correct behavior label, either because of limited resolution or the ambiguity of our behavior definition.

For each behavior, we randomly selected N total frames consisting of N/M short intervals of M frames distributed over all the frames, animals and videos in the ground-truth data. The interval length M varied by behavior and was chosen according to the distribution of bout lengths for the behavior. As most behaviors occurred rarely, we chose intervals to label randomly, but we increased the weight of intervals containing predicted positive frames (**Supplementary Note**).

For the ground-truth results on flies shown in **Figure 2**, we labeled a total of 6,000 frames of data over five videos not in the training set (**Supplementary Table 2**). For most behaviors, all videos contained pBDPGAL4U control flies. For behaviors not frequently performed by control flies (wing extension, copulation, attempted copulation and double wing flick), we annotated ground truth on novel videos from *GAL4* lines in the behavior's training set. For larvae, we annotated ground truth on 6,000 frames selected randomly from three videos, one for each of the three species examined. For mice, we annotated ground truth on 20,000 and 12,000 frames of data for following and walking behaviors, respectively, selected from the 30 min of video after the 30 min of video used for training.

Comparison to Ctrax and CADABRA. We compared the fly chase behavior detector trained using JAABA to the detector trained using Ctrax and the manually tuned threshold definition used by CADABRA. For Ctrax, we used the behavior classifier distributed with the Ctrax software. For the CADABRA definition, we ignored restrictions of the definition depending on wing positions; thus, the definition used was more permissive. Both of these classifiers were tuned for different environments: Ctrax for a 24.5-cm-diameter walking arena and CADABRA for a 4 × 5–cm chamber. We tested these static behavior definitions on the control ground-truth data set. On this data set, JAABA obtained an error rate of 0.6%, Ctrax an error rate of 10.4% and CADABRA detected no chase events (error rate of 50%).

13. Dollar, P., Rabaud, V., Cottrell, G. & Belongie, S. in *Visual Surveillance and Performance Evaluation of Tracking Surveillance (VS-PETS)* 65–72 (IEEE, 2005).
14. Simon, J.C. & Dickinson, M.H. *PLoS ONE* **5**, e8793 (2010).
15. Pfeiffer, B.D. *et al. Proc. Natl. Acad. Sci. USA* **105**, 9715–9720 (2008).

doi:10.1038/nmeth.2281