Automatic Head Tracking of The Common Marmoset.

Hjalmar K. Turesson[1 *], Thamiris Botelho Ribeiro Conceição[2], Sidarta Ribeiro[3]

1 York University, Toronto, Canada
2 Universidade Federal do Rio Grande do Norte, Natal, Brazil
3 Instituto do Cérebro, Universidade Federal do Rio Grande do Norte, Natal, Brazil

*** turesson@yorku.ca**

Abstract

New technologies for manipulating and recording the nervous system allow us to perform unprecedented experiments. However, the influence of our experimental manipulations on psychological processes must be inferred from their effects on behavior. Today, quantifying behavior has become the bottleneck for large-scale, high throughput, experiments. The method presented here addresses this issue by using deep learning algorithms for video-based animal tracking. Here we describe a reliable automatic method for tracking head position and orientation from simple video recordings of the common marmoset (*Callithrix jacchus*). This method for measuring marmoset behavior allows for the indirect estimation of gaze within foveal error, and can easily be adapted to a wide variety of similar tasks in biomedical research. In particular, the method has great potential for the simultaneous tracking of multiple marmosets to quantify social behaviors.

Introduction

Recent technological developments allow us to manipulate [3, 35, 40] and record [7, 9, 53] the nervous system with historically unprecedented precision and scale. Yet, the psychological relevance of our sophisticated manipulations and large-scale recordings can only be inferred from their effect on behavior. Today, quantifying behavior has become the bottleneck for high-throughput experiments [5, 48]. Common practice is to apply standard tests designed to measure psychological constructs such as anxiety and memory, for example, by using the Elevated Plus Maze [34] or the Morris Water Maze [28]. Such testing requires animals to be individually handled, making data acquisition labor intensive, increasing costs and reducing experimental throughput. Alternatively, various simple detectors (e.g. capacitance sensors or photo-beams) can be arranged to automatically acquire data at specific sites (e.g. drinking [12] and feeding [10] stations). This type of automatization allows for high throughput, but fails to capture complex or subtle behaviors such as social interactions and gaze behaviors. A more promising approach is to record high-dimensional data from sensor arrays (e.g. video) and extract relevant information using computer vision algorithms. This approach has the potential to provide a better characterization of behavior, capable of automatically capturing complex and subtle behaviors, while simultaneously reducing both cost and labor intensity [39]. Raw video frames are composed of a high number of pixels whose values do not straightforwardly correlate with an animal's behavior. In

order to reliably measure behavior in the face of such variability, many computer vision systems rely on explicitly-designed dimensionality reduction methods that extract specific features (e.g. to achieve illumination invariance [13]). The result of this preprocessing can then be used to measure or classify behavior [19, 54]. However, explicitly designing feature extraction methods that are robust under a wide range of conditions is difficult and time-consuming [2]. Deep learning algorithms, in contrast, automatically learn a hierarchy of increasingly abstract representations, running from simple feature detectors (e.g. edge detectors) to the final classification. This strategy obviates the need for manually designed feature extraction and has proven successful in a great variety of tasks [2, 24]. A particular group of deep learning algorithms, convolutional neural networks (CNNs), are inspired by the local connectivity pattern of visual cortex [55], and are optimized for data organized in arrays like those produced by a digital camera. The recent development of CNNs have led to impressive improvements on visual recognition tasks [23], making them the most popular deep learning method in computer vision.

Most work on the automatic measurement of behavior has been done on fruit flies [4, 43–45], zebra fish [32, 33] and mice [29, 31]. To the best of our knowledge, no previous work on non-human primates has been published, despite their importance as a research model in several areas of life sciences. In particular, the common marmoset (*Callithrix jacchus*) is becoming an increasingly important primate model in biomedical [16, 30, 38, 46, 56], genetic [18, 20, 37], and neuroscience [22, 42] research. The growing popularity of marmosets arises from their similar disease susceptibility profile to humans [8, 47], their relative ease of handling, their high fecundity and speedy development [36], and the recent development of key tools for genetic manipulations [37, 41] and neuroscience experiments [30].

In contrast to rodents, primates in general have to orient their gaze precisely which makes gaze direction informative of what they pay attention to [26, 42]. Further, a striking feature of the gaze behavior of small-headed primates such as marmosets is their rapid head movements. When shifting their gaze, marmosets tend to move their head in quick jerks, similar to how bigger primates move their eyes when making saccades. Although marmosets can and do make saccadic eye movements, they are limited to the central 10° [26], with the result that head movements contribute substantially to the final gaze shift [15, 25]. This might be explained by these primates' small head and relatively large eye [11, 49]. The lesser head size results in lower rotational inertia which in turn, lowers the muscular force required to produce rapid head movements. Since rotational inertia decreases faster with size than muscle force does, head movement become gradually more favorable with smaller head sizes. Thus, the combination of a small fovea and a big contribution of head movement to the gaze direction makes head tracking of marmosets remarkably informative, allowing gaze estimation without eye measurements. For marmosets, head orientation alone indicates marmosets' gaze direction within ±10°. By tracking a marmoset's head position and direction, we get a rich source of information about how it moves and what captures its interest. We here present a video-based method that uses a CNN to track head position and orientation (thus approximating gaze direction) in the common marmoset. To our knowledge, this is the first report of automatic gaze tracking of freely moving non-human primates.

Results

We set out to develop a video-based method for automatic tracking of head position and gaze direction of marmosets. To this end, we first recorded marmosets engaged in a 15-minute vocal learning experiment using a video camera mounted above the test box (**Fig 1**). We then annotated a subset of the video with head position and orientation,

and trained a CNN to predict this from the raw video frames. Head position was indirectly estimated from the model's spatial activation pattern [57] (see Materials and Methods), while orientation was explicitly estimated by the classifier.

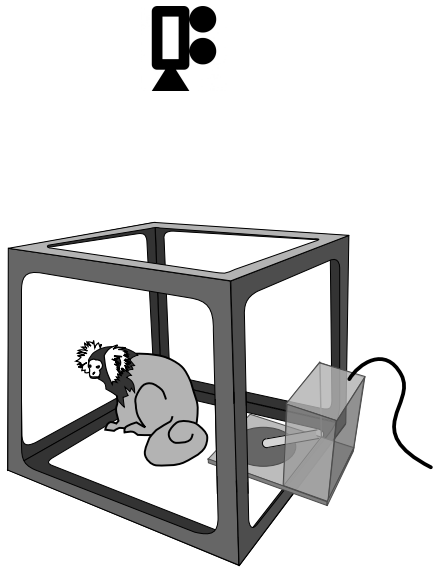

Fig 1. Experimental setup. Video was recorded by a camera placed above the subject while it was taking part in a vocal learning experiment. The device to the right of the subject is the reward dispenser with spout and reward collection dish (dark gray circle).

There were two sources of uncertainty in the input data. First, the human estimate of head direction is not completely consistent and will vary somewhat from frame to frame even when the true direction is constant. Second, head direction angles (0 to 360°) were binned into 28 classes resulting in a loss of precision. To provide an estimate of the first source of uncertainty, two investigators annotated the same 1,361 frames and the differences between the estimated positions and directions were measured. Median differences between the investigators were 8.5 pixels in position and 7.2° in direction. We refer to these differences as the inter-human disagreement. The inter-human disagreement also provides a reference to which the model's performance can be compared. We operationalized model error as the difference between a human scorer's estimate of head position and orientation, and that of the model.

Fig 2A shows an example of predicted head direction and position in comparison with the human annotation in a sequence of video frames. Note that although the manually marked position is shown, this is only to evaluate the model's performance and no explicit position information was used in training the model.

To test tracking performance we annotated 2,000 video frames with head position and direction, allowing us to compare the model's predictions with the inter-human disagreement. Randomly selecting the frames from the training set is likely to overestimate performance since adjacent frames in these videos often are highly correlated. Therefore, to test the model, we used labeled frames from five videos (400 frames from each video) excluded from both the training and validation data. The five videos included all three subjects. **Fig 2B** shows the predicted head position and orientation, as well as the associated errors, in one of the five videos. The head position error is reported in pixels (640 × 480-pixel frames) and the direction error in degrees from 0 to 180°.

Fig 2C shows the position and direction error on all the test data, compared with the inter-human disagreement. The median position error was 27 pixels. To put this error in anatomical perspective, it can be compared to the marmosets' head size. The width of the head, measured as the inter-tuft-distance, was on average 47 pixels (20 to 80 pixels depending on the distance to the camera). Thus, the median position error is about 57% of the average head width. In comparison, the inter-human disagreement is around 9 pixels, that is, around 20% of the average head width. The median head direction error was 10.4°, which was close to the median inter-human disagreement of 7.2°.

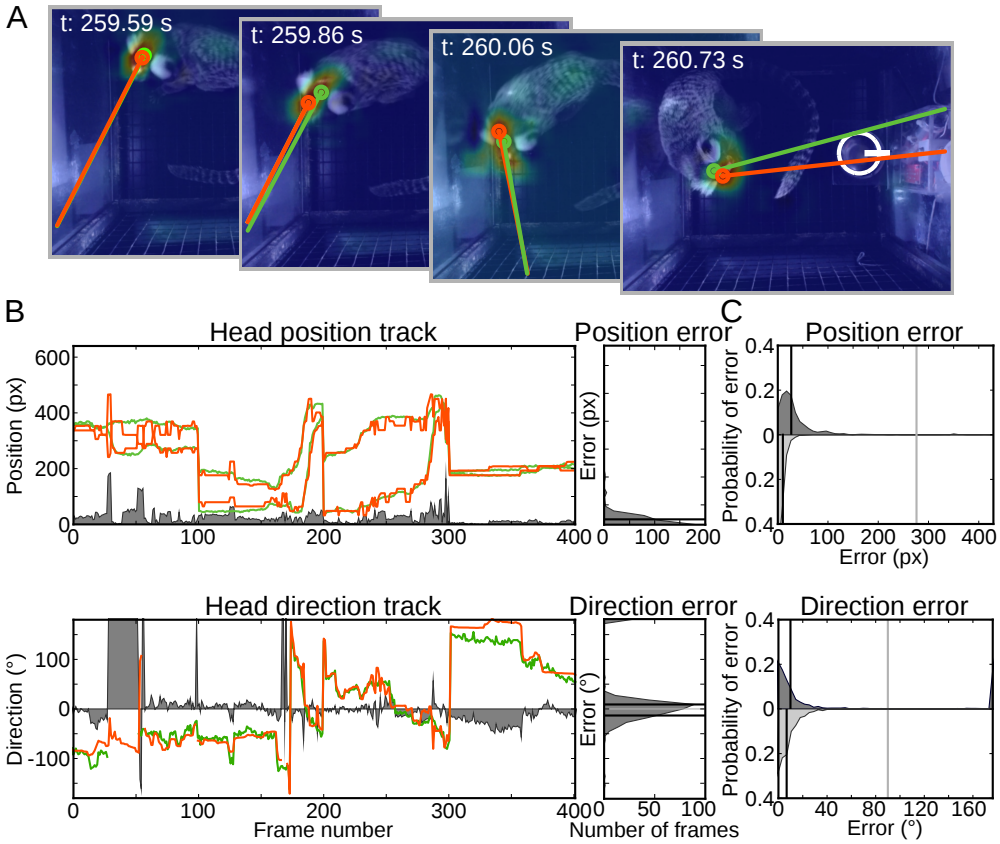

Fig 2. Head tracking. **A**, Four video frames illustrating the tracking of head position and direction. The time of each frame is shown on the upper left corner. The orange circle and line show the predicted head position and direction, respectively. The green circle and line show the corresponding manual measurements. The class activity map (see Materials and Methods) is superimposed on the video frame as a heat map with more discriminative image regions shown in warmer colors; this allowed head position to be implicitly inferred from the model. Note that because position was indirectly inferred from the model, the manually estimated positions were used only to evaluate the model's performance. The white circle and horizontal bar in the leftmost frame show the location of the reward spout and collection dish. **B**, Left panels show head tracking on 400 frames (from one video) of the test data. Orange lines show predicted head position coordinates in pixels (upper panel), and head direction in degrees (lower panel). Green lines show human annotation, and gray shows the model's error. The right panels show histograms of the error from the left panels (left and right panels share y-axis). **C**, Histograms of position (top) and direction (bottom) errors across all test data (2,000 frames from five videos). The model errors are shown as dark gray histograms oriented upwards; inter-human disagreements are shown, for comparison, as light gray histograms oriented downwards. Black and light gray lines show median and chance error, respectively.

Given information about head position and orientation, we investigated where subjects spent most time and what they looked most at. To provide a coarse-grained picture of the most visited locations during an experimental session we calculated two-dimensional histograms of head positions representing the horizontal plane of the test box (see **Fig 3A** for an example). With similar granularity, we estimated the points of regard through combining the head's position and direction and represented it as histograms of locations along the perimeter of the test box (see **Fig 3A** for an example).

As expected, the reward dispenser was central to the subject's activities. The preferred locations were either close to the dispenser or, to a lesser extent, in the upper left corner, and most time was spent looking towards the dispenser. To assess how this changed over time, we aligned the data around the first 25 rewards and computed positions and points of regard relative to the spout. **Fig 3B** shows how the subject looks and moves in the direction of the spout around a second after a reward was delivered.

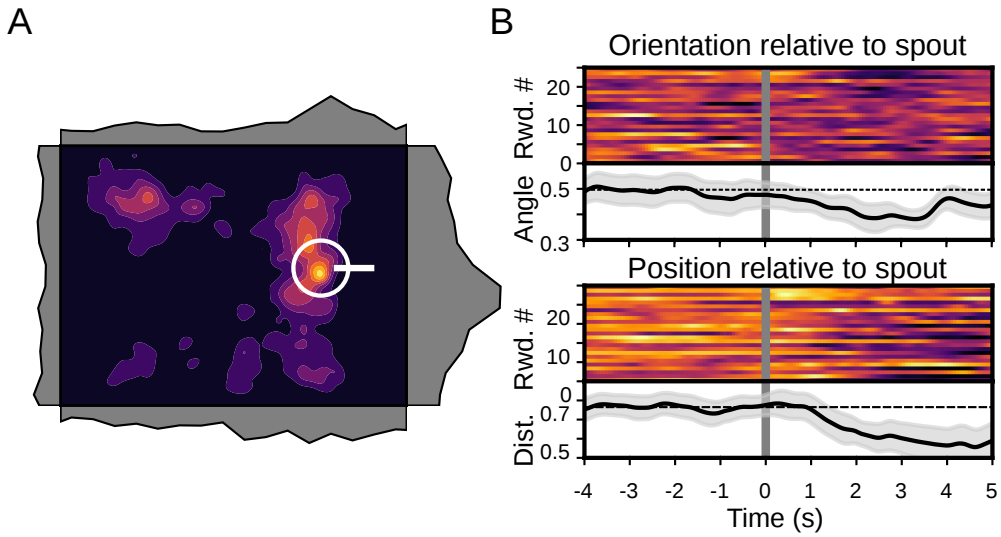

Fig 3. Head positions and points of regard. A, The density of head positions is shown as a heat map spanning the horizontal plane of the test box. The white horizontal bar and circle on the right shows the locations of the reward spout and collection dish. Along the perimeter of the heat map (i.e. the test box), points of regard are shown as gray histograms. Note that the subject spent most time by the reward dispenser and in the upper left corner looking towards the spout (indicated by the peak in the right edge histogram). **B**, Head position and gaze direction around the time of reward delivery. The top panels show the head direction relative the reward spout (i.e. angular distance to spout) from 4 seconds before to 5 seconds after reward delivery. The top portion of the panel shows the normalized angular distances of the first 25 rewards (one row per reward) in a session (warmer colors represent greater distance), and the bottom portion shows the average (black trace) with 95% confidence intervals (light gray field). The dashed line indicates the baseline (i.e. the pre-reward average). The x-axis shows time in seconds, and the vertical gray bar at zero seconds indicates the time of reward delivery. The lower panels show the normalized spatial distance to the spout; otherwise, the figure layout follows the top panels.

Discussion

Here we provided a demonstration of how deep learning can be used to track marmoset behavior. We trained a CNN on video data to predict a marmoset's head position and direction, giving an estimate of the marmoset's location and where it was looking. The tracking performance was good, especially for head direction for which the disagreement between the model and human was close to the inter-human disagreement (**Fig 2C**). For position, the disagreement between human and model was greater but still comparable to human performance. This good performance was not unexpected since deep learning algorithms applied to other complex tasks have been shown to perform at [27], or above

human levels [14]. Since the model was limited by the width of the orientation bins and the accuracy of the human-scored training data, a median orientation error slightly greater than the uncertainty in the data is probably close to optimal.

The method presented here is flexible and can easily be adapted to similar tasks without any changes to the programs used (available at `https://github.com/kalleknast/head-tracker`). To apply the model to video recorded under other conditions than those reported here, all that is needed is to label data and train the model. In the current work we annotated a total of 20,127 frames, taking an estimated 20 hours. However, a subset of those (5,000 frames) was used to select model architecture and to evaluate performance. Thus, 15,000 frames (15 hours of work) should be sufficient to train the model. Since annotating video is the main time-sink when applying the model on new data, it is worth considering how to minimize this effort. One way to make the process more efficient would be to selectively label the more informative frames. Video frames are often highly correlated, meaning that many of them do not contribute much when optimizing the model, and are thus redundant. Through the selective sampling of less correlated frames the model could probably be successfully trained using less labeled data [43]. This strategy should minimize the unnecessary labeling of uninformative frames, but the net amount of saved time remains to be determined.

Although our experiments were performed with isolated individuals, we believe that the general method can be extended to recordings of multiple animals. Such simultaneous tracking of multiple animals would provide a way to quantify social behaviors. Social behaviors are, in general, difficult to measure since they are complex and thus, require complex analysis of high dimensional data such as video. Earlier, this could only be reliably done by manually scoring videos, making this type of analysis low-throughput [39] and subject to unconscious or conscious bias [6]. However, the recent developments in deep learning hold the promise of automatizing this task. Marmosets, like humans, look directly at regions of social interest, particularly faces [26]. Thus, by tracking their gaze in a social environment, we get a measure of their social interactions. Gaze and position tracking should allow automatic detection of such behaviors as play and grooming. That the method presented here provides a first step in that direction.

The recordings reported here were done in a restricted environment that essentially enabled only two-dimensional movement. The tracking was further constrained to only concern the horizontal plane, thus decreasing the complexity of the signal processing implemented. However, marmosets are not ground dwelling and in most naturalistic environments tracking in three dimensions would be necessary. While three-dimensional tracking requires multiple cameras, the tracking algorithm would not require much modification. Successful, multi-camera, three-dimensional tracking has been reported for *drosophila* [44, 45]. However, that method relies on background subtraction to localize the target, and thus, requires a visually constant and unobstructed environment which precludes extension to primate home-cage conditions. The utilization of deep learning algorithms provides a promising alternative.

Materials and Methods

Dataset Description

The subjects were three captive-born adult common marmosets (two females and one male), housed at the Instituto do Cérebro, Universidade Federal do Rio Grande do Norte. The marmosets were housed socially in two wire mesh enclosures ($1.20 \times 1.50 \times 2.45$ m), enriched with tree branches, ropes, plants, hammocks and

nesting tubes. The animals were fed twice daily with fresh and dried fruit, nuts, egg and chicken, and had *ad libitum* access to water. The colony was maintained outdoors protected by a roof while still allowing daily sunbaths in natural light. The animals were housed in compliance with SISBIO permit #18394, and the experiment was approved by the ethics committee of Universidade Federal do Rio Grande do Norte with CEUA permit #028/2013. No animal was sacrificed at the end of the experiment.

We recorded video of marmosets taking part in a vocal learning experiment. During an experimental session, an individual marmoset was kept in a test box measuring approximately $30 \times 30 \times 30$ cm (**Fig 1**) for 15 minutes. The subjects were trained to voluntarily leave their home-cage and enter the test box cage in return for preferred food items. The recordings were part of a vocal learning experiment where the subjects were rewarded for making *phee* calls [50]. Sound was monitored online via a microphone streaming data to a computer running a *phee*-detection program [51]. Two seconds after the end of a *phee* call, a dispenser delivered a liquid reward (Yakult, Tokyo, Japan) via a spout to a reward collection dish placed below (**Fig 1** and **2A**). We recorded a total of 17 hours and 30 min of video over a period of six months. Video was recorded at 15 frames per second with a C920 Logitech HD Pro Webcam (Logitech, Lausanne, Switzerland) placed 30 cm above the cage. The recording software was custom written in Python 3.5 utilizing the GStreamer open-source multimedia framework (`https://gstreamer.freedesktop.org`) and is available at `https://github.com/kalleknast/video-tools`.

Subsequent to the data collection, the videos were annotated with the marmosets' head position and orientation in the horizontal plane. In order to reduce effort and improve annotation consistency, we wrote software to aid the annotation process. Annotation did not require any previous training since it only consisted of marking the location of the marmoset's forehead followed by the base to the two tufts. From these three points, the head position and angle in the horizontal plane were calculated. However in some instances, frames could not be assigned any meaningful head direction in the horizontal plane. For example, sometimes the subject looked out of the horizontal plane (i.e. up or down), or its head was not clearly visible. Those frames were annotated with "*angle-does-not-apply*". We annotated 18,127 frames from 15 different videos (1208 $\pm$413 frames per video). These frames were split into a training (15,127 frames) and a validation set (3,000 frames). The validation set was used to tune the model architecture (e.g. to find the optimal number of classes, see below). Another 2,000 frames were annotated for the test data. Test data was taken from five videos, each contributing 400 frames that were sampled in four randomly spaced batches of 100 consecutive frames. Since the temporal correlation between consecutive video frames is high, using randomly drawn frames from the same videos as used for training data, would probably overestimate the model's ability to generalize to new data. Thus, in order fairly assess the model's performance, the videos used for test data were different from those used for validation and training data.

Video annotation was done by two investigators. In order to estimate their agreement and to get a reference to compare the model's performance to, both investigators annotated the same 1,361 video frames. There was a good agreement about both position (median 8.5 pixels , 640×480-pixel frames), and orientation (median 7.2°, 0 to 180° range). It took approximately one hour to label 1,000 video frames (3.6 s per frame), and around 20 h to annotate the complete dataset.

Classification Algorithm

Our goal was to find a model able to predict both the head's position and its direction. In order to achieve this, we took advantage of the fact that the units a CNN have spatially restricted receptive fields [55], and thus, implicitly carry information about the

location of the source of their activation. That is, the units act as object localizers despite not receiving any explicit information about the location of the object in the image. However, most CNN architectures incorporate one or more layers of fully-connected units in the final stages, thus losing the position information. Here, we instead followed [57] and replaced the fully-connected layer with a global average pooling layer, averaging over filter features while retaining spatial information [21, 57]. This layer is, in turn, connected to a softmax classification layer. In order to find the discriminative image regions, the softmax output is mapped back to the last convolutional layer resulting in a class activity map [57]. Head position was approximated as the maximum of the class activity map (see **Fig 2A**). Following this strategy we built a CNN with six convolutional layers and two max pool layers. Filter sizes in the convolutional layers decreased from 11×11 to 3×3. The global average pooling had a size of 1024 and connected to a softmax layer with the number of units equal to the number of classes. We used cross-entropy loss regularized by the Euclidean (L2) norm of the weights the regularization parameter set to 0.0001.Weights were initialized with random values drawn from a normal distribution with zero mean and a standard deviation set to $\sqrt{2/n_w}$ where n_w is the number of weights in a layer [14]. The model was optimized with the Adam algorithm (52). Thus, model was trained to predict a head direction class given a whitened and scaled down (from 640×480 to 160×120 pixels) gray-scale (averaged RGB color) video frame as input.

Head direction measured in degree of head rotation in the horizontal plane was binned in $n - 1$ classes and the frames labeled as *angle-does-not-apply* were assigned to an nth class. In order to select the optimal number of classes, and thus, the size of direction bins, we trained multiple models configured to predict from five to 33 classes (steps of four), and compared the direction error on the validation set. The direction error decreased with an increasing number of classes until $n = 29$, where it reached the inter-human disagreement. Thus, head direction angles were converted into labels through binning 0 to 360° into 28 bins, each $\approx 12.9°$ wide. For the final model, we merged the training and validation sets and trained with 18,127 exemplars for 80 epochs.

We used Python 3.5 to implement the head tracker. The core model was written using GPU-enabled TensorFlow 0.10.0 [1], while NumPy, SciPy [52] and Matplotlib [17] were used for the other functionality. The software is available at `https://github.com/kalleknast/head-tracker`. Computations were done on a Lenovo ThinkStation P900 running Ubuntu 16.04 with a 4 GB NVIDIA Quadro K2200 graphics card.

Data Analysis

We evaluated the model's performance on both position and direction. The position error was measured as the distance in pixels between the human-assigned position and the coordinates of the peak of the class activity map. Similarly, the direction error was the angular distance between the human-assigned label and the model's prediction. The predicted head direction was calculated as the weighted average of the softmax output, excluding the nth class (*angle-does-not-apply*). The nth class was treated differently: disagreement was considered 180° error and agreement 0° error. **Fig 2B** demonstrates an example of this. The lower left panel shows a 25 frame long period of disagreement resulting in 180° error. The same period is evident in the associated error histogram (lower right panel) as a small peak at 180°.

To investigate the subjects preferred positions and points of regards during the experimental sessions, we combined the head position and direction data. Preferred position was estimated by computing two-dimensional histograms representing the density of head positions in the horizontal plane of the test box (see **Fig 3A**). This density map was smoothed with a Gaussian kernel (10×10-pixel standard deviation)

for visual clarity. Points of regard were estimated by combining head position and direction and representing the resulting locations along the perimeter of the test box as histograms (see **Fig 3A**). For reward-dependent behavior we aligned the data to the time points of the 25 first rewards in a session (from four seconds before to five seconds after reward). From this reward-aligned data we calculated the distance in pixels between the head position and the spout, and the angular distance between the point of regard and the spout (i.e. how many degrees the head would need to be rotated in order to be oriented straight towards the spout). The time points where the model predicted *angle-does-not-apply* were interpolated from neighboring values. For both measures, we smoothed the 25 traces with a Gaussian kernel (13 ms standard deviation) and computed averages and 95% confidence intervals.

Acknowledgments

We thank Stephen Shepherd for comments on the manuscript, and William Wood, Josy Pontes, Janaina Figueiredo, Bruno Soares, Maria Bernardete de Souza and the Primatology Center of UFRN for help with various aspects of housing & specialized marmoset care. Research equipment and HKT's and TBRC's stipends were funded by CNPq/BJT grant # 402422/2012-0 to SR and HKT. SR further thanks the FAPESP Research, Innovation and Dissemination Center for Neuromathematics (grant # 2013/07699-0, S. Paulo Research Foundation). We thank Taeksoo Kim for unknowing guidance on the implementation of the CNN model.

References

1. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. Technical report, mar 2016.

2. Y. Bengio, A. Courville, and P. Vincent. Representation learning: a review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–828, aug 2013.

3. E. S. Boyden, F. Zhang, E. Bamberg, G. Nagel, and K. Deisseroth. Millisecond-timescale, genetically targeted optical control of neural activity. *Nature neuroscience*, 8(9):1263–8, sep 2005.

4. K. Branson, A. A. Robie, J. Bender, P. Perona, and M. H. Dickinson. High-throughput ethomics in large groups of Drosophila. *Nature methods*, 6(6):451–7, jun 2009.

5. D. Brunner, E. Nestler, and E. Leahy. In need of high-throughput behavioral systems. *Drug discovery today*, 7(18 Suppl):S107–12, sep 2002.

6. G. M. Burghardt, J. N. Bartmess-LeVasseur, S. A. Browning, K. E. Morrison, C. L. Stec, C. E. Zachau, and T. M. Freeberg. Perspectives - Minimizing Observer Bias in Behavioral Studies: A Review and Recommendations. *Ethology*, 118(6):511–517, jun 2012.

7. G. Buzsáki, E. Stark, A. Berényi, D. Khodagholy, D. R. Kipke, E. Yoon, and K. D. Wise. Tools for probing local circuits: high-density silicon probes combined with optogenetics. *Neuron*, 86(1):92–105, apr 2015.

8. R. Carrion and J. L. Patterson. An animal model that reflects human disease: the common marmoset (Callithrix jacchus). *Current opinion in virology*, 2(3):357–62, jun 2012.

9. K. Chung, J. Wallace, S.-y. Kim, S. Kalyanasundaram, A. S. Andalman, T. J. Davidson, J. J. Mirzabekov, K. A. Zalocusky, J. Mattis, A. K. Denisin, S. Pak, H. Bernstein, C. Ramakrishnan, L. Grosenick, V. Gradinaru, and K. Deisseroth. Structural and molecular interrogation of intact biological systems. *Nature*, 497(7449):332–7, may 2013.

10. K. S. Gannon, J. C. Smith, R. Henderson, and P. Hendrick. A system for studying the microstructure of ingestive behavior in mice. *Physiology & behavior*, 51(3):515–21, mar 1992.

11. A. N. Goss, J. White, and G. C. Townsend. Craniofacial growth in young marmosets (Callithrix jacchus). *Laboratory animals*, 17(4):303–6, oct 1983.

12. 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 of the United States of America*, 105(52):20575–82, dec 2008.

13. G. H. Granlund and H. Knutsson. *Signal Processing for Computer Vision*. Springer US, Boston, MA, 1995.

14. K. He, X. Zhang, S. Ren, and J. Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034. IEEE, dec 2015.

15. S. A. Heiney and P. M. Blazquez. Behavioral responses of trained squirrel and rhesus monkeys during oculomotor tasks. *Experimental brain research*, 212(3):409–16, jul 2011.

16. B. Horvat, B. K. Berges, and P. Lusso. Recent developments in animal models for human herpesvirus 6A and 6B. *Current opinion in virology*, 9:97–103, dec 2014.

17. J. D. Hunter. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

18. J. C. Izpisua Belmonte, E. M. Callaway, S. J. Caddick, P. Churchland, G. Feng, G. E. Homanics, K.-F. Lee, D. A. Leopold, C. T. Miller, J. F. Mitchell, S. Mitalipov, A. R. Moutri, J. A. Movshon, H. Okano, J. H. Reynolds, D. Ringach, T. J. Sejnowski, A. C. Silva, P. L. Strick, J. Wu, and F. Zhang. Brains, genes, and primates. *Neuron*, 86(3):617–31, may 2015.

19. H. Jhuang, E. Garrote, J. Mutch, X. Yu, V. Khilnani, T. Poggio, A. D. Steele, and T. Serre. Automated home-cage behavioural phenotyping of mice. *Nature communications*, 1(6):68, sep 2010.

20. T. Kaiser and G. Feng. Modeling psychiatric disorders for developing effective treatments. *Nature medicine*, 21(9):979–88, sep 2015.

21. D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, pages 1–13, dec 2014.

22. N. Kishi, K. Sato, E. Sasaki, and H. Okano. Common marmoset as a new model animal for neuroscience research and genome editing technology. *Development, growth & differentiation*, 56(1):53–62, jan 2014.

23. A. Krizhevsky, G. E. Hinton, and I. Sutskever. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.*, pages 1–9, 2012.

24. Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015.

25. R. A. McCrea and G. T. Gdowski. Firing behaviour of squirrel monkey eye movement-related vestibular nucleus neurons during gaze saccades. *The Journal of physiology*, 546(Pt 1):207–24, jan 2003.

26. J. F. Mitchell, J. H. Reynolds, and C. T. Miller. Active vision in marmosets: a model system for visual neuroscience. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 34(4):1183–94, jan 2014.

27. V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–33, feb 2015.

28. R. Morris. Morris water maze. *Scholarpedia*, 3(8):6315, 2008.

29. S. Ohayon, O. Avni, A. L. Taylor, P. Perona, and S. E. Roian Egnor. Automated multi-day tracking of marked mice for the analysis of social behaviour. *Journal of neuroscience methods*, 219(1):10–9, sep 2013.

30. H. Okano, K. Hikishima, A. Iriki, and E. Sasaki. The common marmoset as a novel animal model system for biomedical and neuroscience research applications. *Seminars in fetal & neonatal medicine*, 17(6):336–40, dec 2012.

31. T. P. Patel, D. M. Gullotti, P. Hernandez, W. T. O'Brien, B. P. Capehart, B. Morrison, C. Bass, J. E. Eberwine, T. Abel, and D. F. Meaney. An open-source toolbox for automated phenotyping of mice in behavioral tasks. *Frontiers in behavioral neuroscience*, 8(October):349, 2014.

32. S. D. Pelkowski, M. Kapoor, H. A. Richendrfer, X. Wang, R. M. Colwill, and R. Creton. A novel high-throughput imaging system for automated analyses of avoidance behavior in zebrafish larvae. *Behavioural brain research*, 223(1):135–44, sep 2011.

33. J. Rihel, D. A. Prober, A. Arvanites, K. Lam, S. Zimmerman, S. Jang, S. J. Haggarty, D. Kokel, L. L. Rubin, R. T. Peterson, and A. F. Schier. Zebrafish behavioral profiling links drugs to biological targets and rest/wake regulation. *Science (New York, N.Y.)*, 327(5963):348–51, jan 2010.

34. O. E. Rodríguez-Sierra, S. Goswami, H. K. Turesson, and D. Pare. Altered responsiveness of BNST and amygdala neurons in trauma-induced anxiety. *Translational psychiatry*, 6(7):e857, jul 2016.

35. B. L. Roth. DREADDs for Neuroscientists. *Neuron*, 89(4):683–94, feb 2016.

36. A. B. Rylands. *Marmosets and tamarins: Systematics, behavior, and ecology.* Oxford Univeristy Press, Oxford, UK, 1 edition, 1993.

37. E. Sasaki. Prospects for genetically modified non-human primate models, including the common marmoset. *Neuroscience research*, 93:110–5, apr 2015.

38. C. A. Scanga and J. L. Flynn. Modeling tuberculosis in nonhuman primates. *Cold Spring Harbor perspectives in medicine*, 4(12):a018564, dec 2014.

39. A. T. Schaefer and A. Claridge-Chang. The surveillance state of behavioral automation. *Current opinion in neurobiology*, 22(1):170–6, feb 2012.

40. O. Shalem, N. E. Sanjana, and F. Zhang. High-throughput functional genomics using CRISPR-Cas9. *Nature reviews. Genetics*, 16(5):299–311, may 2015.

41. H. Shen. Precision gene editing paves way for transgenic monkeys. *Nature*, 503(7474):14–5, nov 2013.

42. S. G. Solomon and M. G. P. Rosa. A simpler primate brain: the visual system of the marmoset monkey. *Frontiers in neural circuits*, 8:96, aug 2014.

43. U. Stern, R. He, and C.-h. Yang. Analyzing animal behavior via classifying each video frame using convolutional neural networks. *Scientific reports*, 5:14351, 2015.

44. A. D. Straw, K. Branson, T. R. Neumann, and M. H. Dickinson. Multi-camera real-time three-dimensional tracking of multiple flying animals. *Journal of the Royal Society, Interface / the Royal Society*, 8(56):395–409, mar 2011.

45. A. D. Straw, S. Lee, and M. H. Dickinson. Visual control of altitude in flying Drosophila. *Current biology : CB*, 20(17):1550–6, sep 2010.

46. B. A. 't Hart, Y. van Kooyk, J. J. G. Geurts, and B. Gran. The primate autoimmune encephalomyelitis model; a bridge between mouse and man. *Annals of clinical and translational neurology*, 2(5):581–93, may 2015.

47. S. D. Tardif, K. G. Mansfield, R. Ratnam, C. N. Ross, and T. E. Ziegler. The marmoset as a model of aging and age-related diseases. *ILAR journal / National Research Council, Institute of Laboratory Animal Resources*, 52(1):54–65, 2011.

48. L. H. Tecott and E. J. Nestler. Neurobehavioral assessment in the information age. *Nature neuroscience*, 7(5):462–6, may 2004.

49. D. Troilo, H. C. Howland, and S. J. Judge. Visual optics and retinal cone topography in the common marmoset (Callithrix jacchus). *Vision research*, 33(10):1301–10, jul 1993.

50. H. K. Turesson and S. Ribeiro. Can vocal conditioning trigger a semiotic ratchet in marmosets? *Frontiers in Psychology*, 6, oct 2015.

51. H. K. Turesson, S. Ribeiro, D. R. Pereira, J. P. Papa, and V. H. C. de Albuquerque. Machine Learning Algorithms for Automatic Classification of Marmoset Vocalizations. *PloS one*, 11(9):e0163041, 2016.

52. S. van der Walt, S. C. Colbert, and G. Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering*, 13(2):22–30, mar 2011.

53. J. Viventi, D.-H. Kim, L. Vigeland, E. S. Frechette, J. a. Blanco, Y.-S. Kim, A. E. Avrin, V. R. Tiruvadi, S.-W. Hwang, A. C. Vanleer, D. F. Wulsin, K. Davis, C. E. Gelber, L. Palmer, J. Van der Spiegel, J. Wu, J. Xiao, Y. Huang, D. Contreras, J. a. Rogers, and B. Litt. Flexible, foldable, actively multiplexed, high-density electrode array for mapping brain activity in vivo. *Nature neuroscience*, 14(12):1599–605, dec 2011.

54. A. B. Wiltschko, M. J. Johnson, G. Iurilli, R. E. Peterson, J. M. Katon, S. L. Pashkovski, V. E. Abraira, R. P. Adams, and S. R. Datta. Mapping Sub-Second Structure in Mouse Behavior. *Neuron*, 88(6):1121–35, dec 2015.

55. D. L. K. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–65, mar 2016.

56. J.-W. Yun, J.-B. Ahn, and B.-C. Kang. Modeling Parkinson's disease in the common marmoset (Callithrix jacchus): overview of models, methods, and animal care. *Laboratory animal research*, 31(4):155–65, dec 2015.

57. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, dec 2015.