Who Moved My Cheese? Automatic Annotation of Rodent Behaviors with Convolutional Neural Networks

Zhongzheng Ren Adriana Noronha Annie Vogel Ciernia Yong Jae Lee
University of California, Davis
{zzren,abnoronha,ciernia,yongjaelee}@ucdavis.edu

Abstract

In neuroscience, understanding animal behaviors is key to studying their memory patterns. Meanwhile, this is also the most time-consuming and difficult process because it relies heavily on humans to manually annotate the videos recording the animals. In this paper, we present a visual recognition system to automatically annotate animal behaviors to save human annotation costs. By treating the annotation task as a per-frame action classification problem, we can fine-tune a powerful pre-trained convolutional neural network (CNN) for this task. Through extensive experiments, we demonstrate our model not only provides more accurate annotations than alternate automatic methods, but also provides reliable annotations that can replace human annotations for neuroscience experiments.

1. Introduction

Rodent models are now commonly used in neurobiology to answer basic questions about brain function and to study disease mechanisms. One of the most commonly used measures of rodent cognitive function is to examine measures of learning and memory. Mice and rats are capable of learning and remembering a wide variety of species relevant information including the encoding and retrieval of spatial relationships and object identity. There are a variety of behavioral tests that neurobiologists utilize to probe memory formation and retrieval in rodents. Two of the most common tasks examine memory for an object's location or identity. The Object Location Memory (OLM) and Novel Object Recognition memory (NOR) tasks have been used to examine memory enhancements and impairments in a wide variety of genetically modified rodent models [6].

In both tasks an adult rodent is given a specified amount of time to explore and learn about two identical objects. The rodent is later brought back for a subsequent memory test. During the test one of the two objects is either moved to a new location (OLM) or replaced with a new object (NOR). The test rodent is given an opportunity to explore both the familiar and moved/new object during the testing session.

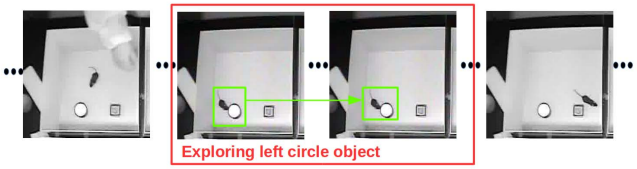

Figure 1. An example recording session of rodent behavior. The annotator (machine or human) needs to label when the rodent begins to explore an object and the duration of exploration. Due to the strict neuroscience definition of 'rodent exploration', the labeling task can be very challenging. By formulating the task as per-frame classification and fine-tuning powerful convolutional neural network models, our approach significantly outperforms previous tracking-based methods.

This task relies on the rodent's *innate preference for novelty* and a rodent that remembers the original training session will preferentially explore the moved/new object compared to the familiar object during the test.

The most common way of measuring the rodent's behavior is by: (1) recording a video of the rodent in action, and (2) for an experienced human scorer to watch the video and manually record the amount of time the rodent spends exploring each object. To produce reliable data, typically a single video is annotated by 2-3 humans, in order to remove personal annotation biases. In practice, each human annotator often ends up labeling the same video multiple times in order to fix annotation mistakes. Moreover, training a human scorer requires lots of time and practice; for example, in [6] a beginner was trained by repeatedly annotating seven videos until their annotations matched those of previous experienced humans, and in [5] it took $\sim$350 hours for a team of experts to annotate 88 hours of video. Due to these limitations, *performing large-scale experiments is extremely difficult, and replacing labor costs in annotation has become a critical problem*.

To alleviate expensive annotation costs, researchers have proposed to build automatic systems. The most common approach is to track the animal in the videos by tracking its nose [31, 1, 27], body [31, 1, 27, 5, 12], and rear key-

978-1-5090-4822-9/17 $31.00 © 2017 IEEE
DOI 10.1109/WACV.2017.147

points [31, 1, 27]. However, existing tracking based methods have failed to replace human annotators for two main reasons: First, there is a very strict definition for each rodent behavior pattern and two closely-related behaviors can have only subtle differences (e.g., a rodent digging close to an object vs. exploring an object). Second, the raw video data obtained for the neuroscience experiments are often of poor quality (e.g., the frames are gray-scale, low resolution, and with illumination artifacts), and last relatively long (5-15 minutes). There are frequent occlusions between the rodent and objects, and fast movements of the rodent cause motion blur. These issues can easily lead to drifting for tracking methods.

Main idea and contributions. In this paper, we pose the annotation task as a *per-frame action classification* problem. The key insight is that classifying a rodent's behavior patterns for the OLM and NOR neuroscience experiments does not require knowing what the rodent is doing at every timestep – instead it only requires knowing what the rodent is doing *when interacting with the objects* in the scene. By posing the problem as one of frame-level classification instead of object tracking, we are not constrained by the difficulty of precisely tracking the various rodent parts, and thus can circumvent issues with drifting. Furthermore, we can leverage existing pre-trained convolutional neural network image classification models and fine-tune them for this task. Our experimental results demonstrate that our approach performs significantly better than tracking based methods, and are reliable enough to replace human annotations in neuroscience experiments. We release our models, source code, and a new dataset for studying rodent behavior.

2. Related work

Rodent memory research. Object Location Memory (OLM) and Novel Object Recognition (NOR) tasks were introduced in [30] to study animal memory and behavior. These tasks have since then become a popular way for studying rodent memory and brain structures [41, 2, 26]. OLM requires the hippocampus for encoding, consolidation, and retrieval [16, 29] while NOR needs several different regions such as the insular cortex [2] and ventromedial prefrontal cortex [2]. Others [9, 5, 27, 19, 12, 32] study rodent behaviors using hand-engineered pipeline systems.

Object tracking. Tracking algorithms (e.g., [11, 17, 20, 21, 44]) can be used to study both OLM and NOR tasks since the absolute location of the objects in the video is fixed (i.e., knowing the precise location of the rodent's keypoints is often sufficient to know whether the rodent is exploring an object or not). However, existing attempts [1, 31, 12, 27, 19] to replace human annotations have largely been unsuccessful. Apart from the usual limitation of tracking algorithms requiring human supervision to initialize the track, the rodent's fast motion and ambiguous appearances frequently cause drifting, which results in inaccurate predictions of the rodent's behavior patterns.

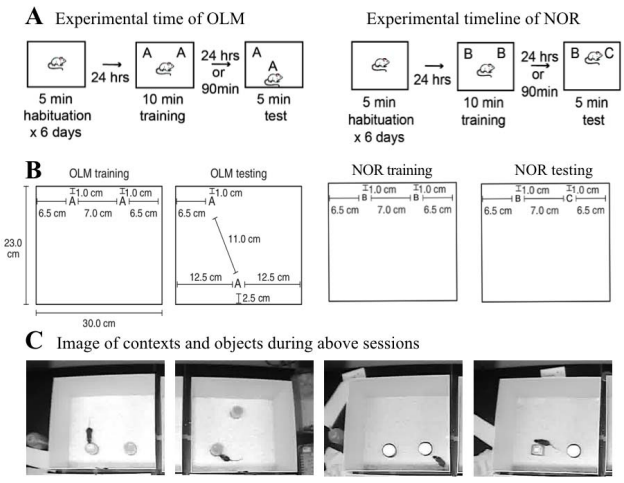

Figure 2. (**A**) The neuroscience experimental timeline of Object Location Memory (OLM) and Novel Object Recognition Memory (NOR). (**B**) Object placement and experimental room settings. (**C**) Images from our Rodent Memory (RM) dataset.

Convolutional neural networks. Convolutional Neural Networks [25, 24, 34, 37, 39, 18] (CNN) have recently revolutionized the field of computer vision. They have been used for various tasks including image classification [24, 34, 37, 18], object detection [14, 13], human pose estimation [43, 38], and video classification [39]. In particular, it has been shown that a CNN trained on one task (e.g., ImageNet for image classification) can be *fine-tuned* for use on another related task [10, 14, 45]. This is especially useful when training data for the new task is limited. For animal behavior recognition, recent work uses deep neural networks to study fruit fly egg-laying [36] and marmoset head position [40]. For rodent behavior research, we are the first to leverage the transferability of CNNs to make use of powerful pre-trained classification models.

3. Neuroscience background

Novel Object Recognition (NOR) and Object Location Memory (OLM) have been widely used in the study of the neurobiological mechanisms underlying long-term memory formation. In this section, we first explain the detailed neuroscience settings of our experiments. We then introduce the precise criteria used to define rodent exploration.

3.1. Neuroscience experimental setting

There are three main steps in NOR and OLM experiments as shown in Figure 2A. First, a five-minute daily habituation is applied for six days, during which the experimental rodents are allowed to explore the empty chamber

freely. 24 hours after habituation, the rodents are put into the same chamber again with two identical objects. The rodents are allowed to explore the objects for 10 minutes. After 24 hours or 90 minutes (depending on testing conditions), the rodents will explore the same chamber with one novel object (NOR) or one of the objects moved to a different location (OLM). This session will last for 5 minutes.

Both OLM and NOR use identical chambers (a $23 \times 30 \times 23$ cm white open rectangular field). Circular tins filled with cement to brim are used in OLM and NOR. Square candle holders filled with cement are used as the new object for NOR. The specific arrangements are shown in Figure 2B. Example video frames from recorded sessions for OLM and NOR are shown in Figure 2C left and right, respectively.

The main measure for both tasks is time spent in exploration of the two objects during testing. (During training, the animals are supposed to show similar preferences for the two objects.) Animals that remember the original training experience will preferentially explore the *changed* object during testing. When combined correctly, these two tasks can allow users to address a variety of experimental questions involving manipulations of different brain regions and molecular targets.

3.2. Criteria for rodent exploration

The following criteria [6] are used to define a rodent's exploration behavior:

Exploration: Interaction time of a rodent with the object when the rodent's nose is within 1 cm of the object and is pointing directly at the object so that an elongated line from eyes to nose would hit the object. No part of the rodent's body should be touching the object except the nose.

The following **do not** count as exploration:

1. The rodent is not approaching the object (e.g., if the rodent re-orientates itself and the nose accidentally comes close to the object (Figure 3A)).

2. The rodent is on top of the object (even if it is looking down at the object) (Figure 3B).

3. The rodent is looking over the object (Figure 3C).

4. The rodent is engaged in a repetitive behavior like digging close to the object or biting the object (Fig. 3D).

The following special cases are also **excluded**:

1. Animals that do not explore more than 3 seconds total for both objects are excluded from analysis.

2. Animals that have discrimination indexes (see Eqn. 2) greater than 20 are considered to have a significant location/object bias and are also excluded from analysis.

Figure 3. Examples of ambiguous behaviors. The actions shown in these images are all considered to be *non*-exploration. When recognizing actions, the false positive predictions are most likely to happen in these cases since their appearances look very similar to those of exploration.

Due to the strict definition of exploration vs. non-exploration, annotating rodent exploration—be it manual or automatic—can be quite challenging. Figure 3 shows examples of ambiguous cases that are non-exploration behaviors that could easily be confused to be exploration.

4. Approach

In this section, we first present our Rodent-Memory dataset, which we use to train and test our models. We then describe our approach for transferring a pre-trained CNN classification model to per-frame rodent behavior classification. Finally, we introduce a new human annotation tool for providing ground-truth annotations and visualizing the machine generated prediction results.

4.1. Rodent Memory dataset

We build a new dataset called Rodent Memory (RM) for both computer vision and neuroscience researchers to use. It contains 80 videos from our neuroscience experiments and frame-level annotations of rodent behaviors. The 5 behavioral categories are constructed based on the object types, their placements, and rodent actions. For NOR data, the 5 categories are: exploring (C0) nothing, (C1) left circle object, (C2) left square object, (C3) right circle object, and (C4) right square object; for OLM data, they are: exploring (C0) nothing, (C1) left object, (C2) right object, (C3) top object, and (C4) bottom object. See Figure 2C for examples of the different configurations of the objects, and Table 1 for dataset statistics.

4.2. Per-frame classification of rodent behaviors

Neuroscientists have used commercial tracking software [31, 1], which track the rodent's keypoints (nose, body, and rear) to automatically annotate the rodent's behavior. However, in practice, these programs fail to save human annotation costs due to frequent errors. According to our observations, their three main failure modes are: (1) The tracker is prone to drift due to the low-quality videos used in recording the neuroscience experiments and the fast motion of the rodents; (2) When using these commercial programs, the user has to define an interest donut-shaped "object" region; a rodent behavior counts as exploration if the rodent's nose

	videos	frames	C0	C1	C2	C3	C4
NOR train	36	450720	429916	7314	3815	5139	4536
NOR test	12	152776	144707	2148	1750	2382	1789
OLM train	24	227716	220039	1829	1557	1645	2646
OLM test	8	77108	73516	777	491	1025	1299

Table 1. Rodent Memory dataset statistics. See Sec. 4.1 for details.

is in the region while the body and rear points are outside the region. Because of this specific region setting, the tracking-based methods often mislabel the non-exploration examples shown in Fig. 3 as exploration (for example, in Figs 3A, C, D, the rodent's nose keypoint will be in the region while the body and rear keypoints will be outside of it); (3) Because of the low resolution of the video, sometimes the tracker confuses the nose and the rear.

Since we only care about the time instance in which a rodent is exploring an object (i.e., we can ignore all other frames), it is not necessary to track every movement of the rodent. Thus, unlike previous methods that track the rodent, we instead opt to treat the rodent behavior annotation task as a *per-frame classification* problem. This allows us to avoid drifting issues associated with tracking.

For this, we leverage CNNs, which have been proven to produce robust visual representations that are more invariant to translation and scale changes of the pattern-of-interest than traditional hand-crafted features (e.g., SIFT [28] or HOG [7]). Since state-of-the-art deep networks (e.g., AlexNet [24]) have a large number of parameters which can easily lead to overfitting, instead of training a CNN from scratch, we take a pre-trained CNN trained on ImageNet [8] and fine-tune it for our rodent behavior classification task. The CNN is fine-tuned to detect each behavior category, which is equivalent to detecting a specific configuration of the objects and rodent (e.g., rodent exploring the left circle object). Due to the invariance properties of CNNs, the model will be able to detect the configuration regardless of its specific location within the frame. Furthermore, the model can disambiguate more ambiguous cases (as shown in Figure 3), since they will explicitly be labeled as non-exploration during training. We next describe our network architecture.

4.3. Architecture details

A typical CNN is composed of a stack of convolutional layers and fully-connected layers. Because of the local connectivity and shared weights in the convolutional layers, a CNN drastically reduces the amount of parameters compared with more traditional (fully-connected) neural networks, which helps reduce overfitting and allows for efficient processing on GPUs. In this paper, we use abbreviations Ck, Fk, P, D, C to represent a convolutional layer with k filters (Ck), a fully-connected layer with k neurons (Fk), a down sampling max-pooling layer (P) with kernel size 3 and stride 2, a dropout layer (D) [35], and a soft-max classifier (C). We consider two network architectures: AlexNet [24] and C3D [39].

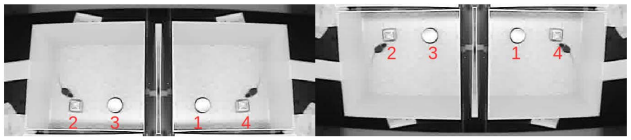

Figure 4. Data augmentation via horizontal/vertical flipping and 180-degree rotation. From left to right: original frame, vertical flipping, horizontal flipping, and 180-degree rotation. We update the labels accordingly.

We transfer AlexNet [24] into use by replacing its last 1000-dimensional classification layer with a 5-dimensional classification layer. Using the above abbreviations, our network architecture is (kernel size in parentheses): C96(11)-P-C256(5)-P-C384(3)-C384(3)-C256(3)-P-F4096-D-F4096-D-C. Except the last classification layer, all convolutional and fully-connected layers are followed by a ReLu [24] non-linearity. The input image size is $227\times227\times3$ $(width\times height\times color)$. We randomly initialize the weights of the last fully-connected layer. We initialize the remaining layers using the weights pre-trained on ImageNet [8] and fine-tune them using our RM dataset.

We also transfer C3D [39], which simultaneously learns spatial and temporal features by performing 3D convolutions, and has been shown to outperform alternate 2D CNNs for video classification tasks. In the C3D architecture, the convolutional filters move in three directions – along the horizontal and vertical spacial axes as well as along the temporal axis. Our C3D network architecture is: C64-P-C128-P-C256-C256-P-C512-C512-P-C512-C512-P-F4096-D-F4096-D-C. All kernel sizes are set to $3\times3\times3$. The model takes consecutive frames with size $128\times171\times d\times3$ $(width\times height\times depth\times color)$ as input. We replace its last classification layer with a 5-dimensional classification layer. It is pre-trained on Sports-1M [23] and fine-tuned on our RM dataset.

Data augmentation. We increase our training data by employing data augmentation, which helps reduce overfitting. Because of the specific placement of the objects in the scene for OLM and NOR (left/right and top/down, see Figure 2B), we adopt three specific data augmentation techniques: horizontal flipping, vertical flipping, and 180-degree rotation. After augmentation, we change the class labels of the new augmented data accordingly. For example, in Figure 4, the first picture is the original one. With horizontal flipping and 180-degree rotation (second and last pictures of Figure 4, respectively), the left square (class 2) and right circle (class 3) become left circle (class 1) and right square (class 4), respectively. After vertical flipping, the labels remain the same (third picture of Figure 4).

Post-processing. After we obtain the test-frame predictions and their probabilities, we use a temporal gaussian filter (with sigma equal to five frames) to smooth out the pre-

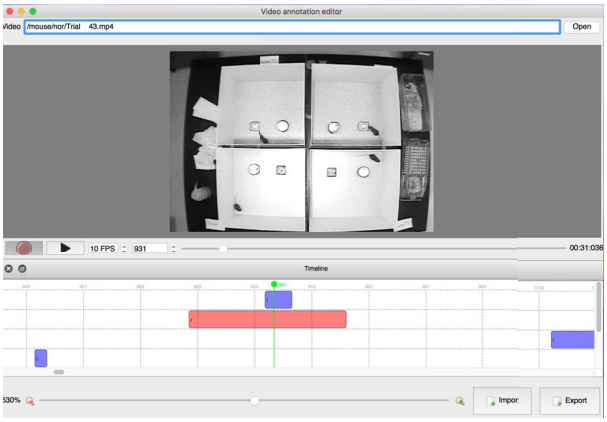

Figure 5. A screen shot of our ground-truth annotation tool. The imported video can be played in adjustable frame rates. The human-generated or machine-generated annotation is shown at the bottom, one row for each of the four rooms. The green line indicates the current temporal position and the blue/red box indicates the rodent is exploring the left/right object.

dictions. This helps reduce noise and outlier predictions, and results in 0.1%-0.7% higher classification accuracy for NOR and OLM in our experiments.

4.4. Annotation tool for ground-truth generation

As part of this work, we create an annotation tool based on [33] to manually-annotate our RM dataset with ground-truth labels (for training and evaluating our models, and visualize our model's outputs). With a carefully designed interface, the program allows ground-truth annotations to be generated efficiently, and allows the machine-generated results to be more easily checked by researchers (for correcting any machine-generated errors).

As shown in Figure 5, it is common to have several neuroscience experiments conducted simultaneously over several rooms (4 in our experiments). The main reason for this is because a single video can capture all rooms, saving video recording and storage costs as well as experimental costs. Prior to our work, annotators used a crude labeling interface (with a simple spacebar click to record the start and end of an action) and thus had to watch the same video four different times to label each room. Our annotation tool allows changeable playing frame rates and start points, which means all rooms can be labeled and checked at the same time and frames of no interest can be easily skipped. In addition, the tool allows a user to choose a different color for each behavior category for easier visualization.

Our annotation tool can also be used to visualize the machine-generated prediction results, which were lacking in previous commercial softwares (e.g., [31, 1] only generate a cumulative exploration time and do not produce per-frame outputs). Thus, it was not easy for researchers to correct the errors in the machine-generated results. Our tool can significantly reduce re-checking efforts while generating more accurate frame-level ground truth labels to begin with. Another important feature is that we can visualize the results according to their prediction probabilities, which means the human annotator can focus on the low-confidence predictions (which are more likely to have errors).

	GT Avg. length	Avg. length [6]	Accuracy [6]
NOR	13.5	4.2	68.7%
OLM	17.3	4.3	63.6%

Table 2. Comparison between the annotations obtained with [6] vs. ground-truth (GT) annotations obtained with our annotation tool.

Comparison to [6]. In [6], a timer program was used to record the rodent's total exploration time. While watching a video, an annotator presses/releases a key when an exploration begins/ends. We compare the annotations collected using this previous program, with the ground-truth annotations[1] collected with our interface.

Table 2 shows the result. We can see that: (1) It is unavoidable for a human to miss some exploration frames using the previous annotation tool because some amount of time is needed to make a judgement on what the rodent is doing. This is why the average exploration time (*Avg. length*, which is the average duration of an exploration in terms of number of frames) is shorter using the previous annotation tool, since the annotator often misses ground-truth exploration behaviors. (2) In [6], one cannot play or stop the video at any time. Therefore, when the annotator makes a mistake, he/she must re-annotate from scratch to fix the error. Due to this, it is almost impossible for a human annotator to go through the video without making a single mistake. Thus, the accuracy of the previous annotations, with respect to the ground-truth annotations, is less than 70%.

5. Experiments

In this section, we evaluate: (1) Frame-wise classification performance of our automatic annotation framework; (2) Qualitative results of both successful and failure predictions; and (3) Whether our automatic annotator can replace human annotators for neuroscience experiments.

Implementation details. We fine-tune the AlexNet [24] and C3D [39] pre-trained models available in the Caffe Model Zoo [22]. The frames are cropped into 4 images, so that each contains only one room. The images are resized to AlexNet: 256×256 and C3D: 128×171. During training, AlexNet: 227×227 and C3D: 114×114 crops are randomly sampled from an image. We use SGD with mini-batch size of 60, weight-decay of 0.00001, and momentum of 0.9. The learning rate starts from 10^{-4} and is divided by 10 every 1000 iterations. The fully-connected layers have a dropout [35] rate of 0.5. Due to the disproportionately

[1]Ours is ground-truth, since we carefully annotated and confirmed every frame in all training and testing videos.

Figure 6. Running screenshot of AnyMaze. The tracker often drifts due to varying illumination conditions, gray-scale and low-resolution frames, and long-term tracking requirements (a video can be several minutes long).

	NOR	OLM
AnyMaze[2]	78.34%	74.25%
AlexNet	**93.17%**	**95.34%**
C3D (d=3)	89.30%	91.98%
C3D (d=5)	87.54%	83.03%
C3D (d=9)	77.74%	73.39%

Table 3. Rodent exploration classification accuracy on RM dataset. Our best CNN-based model (AlexNet) significantly outperforms the commercial AnyMaze tracking software.

large training data of the C0 class (which can make learning difficult), we randomly sample ∼10% of C0 frames during training.

Baseline method. AnyMaze [1] is a tracking-based commercial automatic annotator, previously used for OLM and NOR experiments [42, 15, 4, 3]. It is designed specifically for neuroscience behavior experiments, and in particular, to detect and track rodents under various illumination conditions and experimental conditions.

Since AnyMaze is a commercial software, we cannot know its technical details. However, from our observation of its outputs, it appears to be a detection-based tracking program. It detects the rodent in the room first and then tracks three body points (head, body, and rear) inside the detected bounding box. If the tracker drifts significantly, it will try to reinitialize the tracker by detecting the rodent.

Before running AnyMaze, a human annotator must manually provide the *object region* in the first frame (since the camera is static, the first frame's annotation is used for all frames). The room boundary and the color of the rodent also needs to be manually set. When the rodent's nose keypoint is inside the object region and the body and rear keypoints are outside of it, it counts as exploration.

Evaluation metrics. We evaluate all methods using the ground-truth annotations obtained using our new annotation tool from Sec. 4.4. We quantitatively evaluate per-frame classification accuracy, and the total exploration time using a neuroscience behavioral metric.

5.1. Rodent exploration classification accuracy

We compare our CNN-based framework to the baseline AnyMaze tracking approach. We evaluate per-frame classification accuracy on our RM testing set.

Table 3 shows the result. First, both our fine-tuned AlexNet and C3D networks significantly outperform the commercial AnyMaze software baseline. Since the behavior classification task only requires knowing whether the rodent is exploring an object or not, there is no need to track all of the rodent's movements. By taking advantage of this fact, our frame-level classification approach performs much better than the tracking-based AnyMaze baseline, which often fails due to drifting. Figure 6 shows typical failure cases of AnyMaze. The tracker often drifts, which leads to misclassification of the rodent's exploration behavior.

Second, we find that C3D does not perform as well as AlexNet, even though it encodes (short-range) temporal information. This is likely due to the fixed temporal depth of the network. Since the length of a rodent's exploration of an object can vary from as short as 3 frames (0.1 seconds) to as long as 45 frames (1.5 seconds), setting the temporal depth to be bigger than the actual exploration time can cause the network to miss short explorations (which occur very often). Empirically, we find that shorter temporal depths leads to more accurate results. Thus, AlexNet (which produce per-frame classifications, i.e., depth of 1 frame) outperforms any variant of the C3D network. Another possibility is that C3D is pre-trained on Sports-1M [23] which consists of various human sport actions. Thus, the domain differences between humans and rodents can cause the network to underperform. In contrast, AlexNet is pre-trained on ImageNet [8] with more diverse objects, so it can generalize better.

Figure 7 shows the confusion matrix of our AlexNet network. Overall, our model is able to accurately differentiate the four exploration categories. Incorrect predictions are mostly due to mis-classifying the non-exploration category (C0) with one of the exploration categories (C1-4). This is because the most ambiguous actions belong to C0 (as we show in the qualitative results in Sec. 5.2). For NOR testing, our model makes more mistakes when classifying C2 (it often mis-classifies it as C4). This is mainly due to one testing video, in which the two objects are placed very close to each other. If the rodent is roaming in between the two objects, it is difficult to distinguish which side is its head (see Figure 9, Body, 2nd, 4th, and 5th columns).

5.2. Qualitative results

We next show qualitative results of our predictions. Figures 8 and 9 show successful and failure examples, respectively, according to our model's prediction confidence. (The confidence decrease from left to right.)

[2]Since AnyMaze does not provide frame-level predictions and only cumulative exploration timings, we manually inspect all testing video frames to compute its classification accuracy.

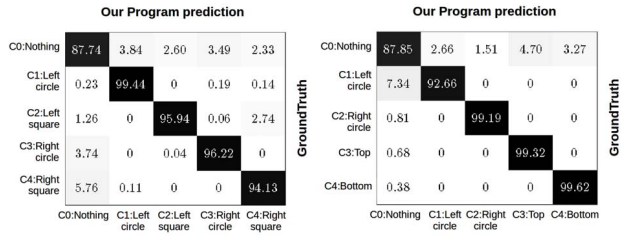

Figure 7. Confusion matrix of our AlexNet predictions. (Left) NOR testing, (Right) OLM testing.

Successful cases. Our model is simple yet powerful to detect the exploration frames among the huge number of frames (only 5 percent of all frames are exploration). From the qualitative results, we can see that:

- The higher our model's confidence, the more accurate it is in predicting the rodent's behavior.
- Our model is robust to clutter (such as a human hand in the scene), various illumination conditions, and deformation of the rooms and objects due to changes in the camera viewpoint.
- Some actions such as passing or circling the object can easily be confused to be exploration by tracking-based methods. However, our model can differentiate those behaviors since it has been trained (with lots of data) to only classify frames in which the rodent's nose is very close to and pointing to the object.
- Many of the lower confidence predictions are on frames in which the body of the rodent is curled. This is because when the rodent is climbing or digging the object (which are considered as non-exploration), it will also curl its body. Therefore, this is an ambiguous pattern for our model.

Failure cases. While our model produces accurate predictions in most cases, it does make mistakes:

- If the rodent is heading to the object and the distance is very close (but farther than 1cm), our model can misclassify the frame as exploration (Fig. 9, row 1).
- Because the video frames are gray scale and low resolution, it can be difficult to tell the difference between the front-end and rear-end of the rodent. This can be problematic when the rodent is moving away from the object but its rear stays very close to the object, which can be misclassified as exploration (Fig. 9, row 2).
- If the rodent is just passing rather than exploring the object, it can sometimes be hard even for humans to differentiate these frames (Fig. 9, row 3).
- The rodent's head reaches a little into the object, which should be counted as non-exploration (Fig. 9, row 4).

Annotator	T_1	T_2	DI
Ours	**0.79**	**0.952**	**0.845**
AnyMaze	0.67	0.659	0.599

Table 4. Pearson's correlation coefficient between the automatic (our approach / AnyMaze) and ground-truth annotations.

5.3. Application to neuroscience experiments

Finally, we evaluate whether our automatically generated annotations can replace human annotations for neuroscience behavior experiments. For this, we compute a standard neuroscience metric used to measure how long a rodent explores one object versus another object. We compare the score obtained using our model's predictions to that obtained using human annotations.

Denote $n1, n2, fr$ as the number of frames annotated (predicted) as exploring the first object (unchanged object), second object (moved/new object), and the video frame rate. Then, the total exploration time of each object is:

$$T_1 = n1 * fr \qquad T_2 = n2 * fr \tag{1}$$

and the Discrimination Index (DI) is:

$$DI = (n1 - n2)/(n1 + n2) \tag{2}$$

We compare the DI and total exploration times computed by different methods across all testing videos. Table 4 shows the Pearson's correlation coefficient of these scores between our program / AnyMaze and the ground-truth annotations. We can see that our prediction has a stronger correlation with the ground-truth human annotations for both total exploration times as well as DI.

In [6], a trainee annotator becomes qualified when his/her annotations and the ground-truth annotations (on 8 test videos) have a Pearson's correlation coefficient higher than 0.7. Our automatic annotator meets this criterion, and thus is good enough to replace a qualified human annotator.

6. Conclusion

We have presented a simple yet effective approach to automatically annotate rodent's behaviors in neuroscience experimental videos that study rodents' long-term memory. Our results demonstrate that our model significantly outperforms a previous tracking-based commercial software designed specifically for this task. We also show that our approach produces annotations that can be used to replace a qualified human annotator.

Today, most neuroscience experiments rely heavily on extensive human labeling. We hope this paper will open up possibilities for using computer vision techniques to automate such arduous labeling tasks. Importantly, this direction would allow researchers to scale up their experiments, which may lead to more robust and novel findings.

Acknowledgements. We gratefully acknowledge NVIDIA for donating the GPUs used in this research.

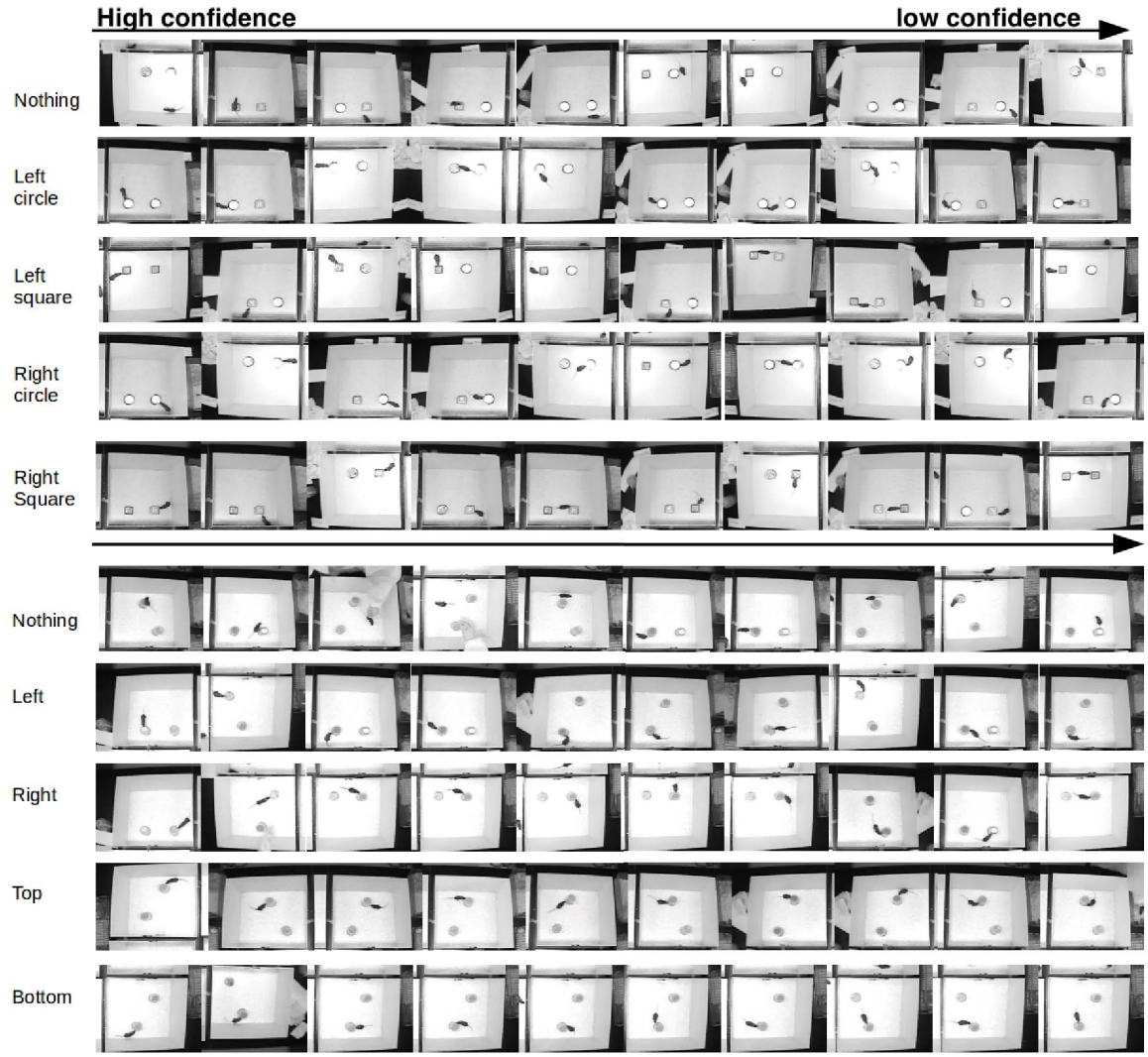

Figure 8. Example successful predictions. NOR: first 4 rows, OLM: next 4 rows. In each row, the examples are ordered according to our model's prediction confidence (from left to right, high to low confidence). See text for details.

Figure 9. Examples of the most common failure cases in which our model falsely predicts the frames as exploration. See text for details.

References

[1] Any-maze behavioural tracking software. In *http://www.anymaze.co.uk/index.htm*.

[2] I. Balderas, C. J. Rodriguez-Ortiz, P. Salgado-Tonda, J. Chavez-Hurtado, J. L. McGaugh, and F. Bermudez-Rattoni. The consolidation of object and context recognition memory involve different regions of the temporal lobe. In *Learning Memory*, 2008.

[3] B. T. Baune, F. Wiede, A. Braun, J. Golledge, V. Arolt, and H. Koerner. Cognitive dysfunction in mice deficient for tnf- and its receptors. In *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 2008.

[4] J. C. Brenesa, M. Padillaa, and J. Fornaguera. A detailed analysis of open-field habituation and behavioral and neurochemical antidepressant-like effects in postweaning enriched rats. In *Behavioural Brain Research*, 2009.

[5] X. P. Burgos-Artizzu, P. Dollar, D. Lin, D. J. Anderson, and P. Perona. Social behavior recognition in continuous video. In *CVPR*, 2012.

[6] A. V. Ciernia and M. A. Wood. Examining object location and object recognition memory in mice. In *Current Protocols in Neuroscience*, 2014.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[9] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *ICCCN*, 2005.

[10] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014.

[11] J. Gao, H. Ling, W. Hu, and J. Xing. Transfer learning based visual tracking with gaussian processes regression. In *ECCV*, 2014.

[12] L. Giancardo, D. Sona, H. Huang, S. Sannino, F. Manag, D. Scheggia, F. Papaleo, and V. Murino. Automatic visual tracking and social behaviour analysis with multiple mice. In *PLoS ONE*, 2013.

[13] R. B. Girshick. Fast R-CNN. In *ICCV*, 2015.

[14] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[15] T. D. Gould, D. T. Dao, and C. E. Kovacsics. The open field test. In *Mood and Anxiety Related Phenotypes in Mice*, 2009.

[16] J. Haettig, D. P. Stefanko, M. L. Multani, D. X. Figueroa, S. C. McQuown, and M. A. Wood. Hdac inhibition modulates hippocampus-dependent long-term memory for object location in a cbp-dependent manner. In *Learning Memory*, 2011.

[17] S. Hare, A. Saffari, and P. H. S. Torr. Struck: Structured output tracking with kernels. In *ICCV*, 2011.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[19] 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. In *Proceedings of the National Academy of Sciences*, 2015.

[20] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV*, 2008.

[21] J. S. S. III and D. Ramanan. Self-paced learning for long-term tracking. In *CVPR*, 2013.

[22] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014.

[23] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[25] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. In *Neural Computation*, 1989.

[26] R. H. Lima, J. I. Rossato, C. R. Furini, L. R. Bevilaqua, I. Izquierdo, and M. Cammarota. Infusion of protein synthesis inhibitors in the entorhinal cortex blocks consolidation but not reconsolidation of object recognition memory. In *Neurobiology of Learning and Memory*, 2009.

[27] 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 *ICIAP*, 2015.

[28] D. G. Lowe. Distinctive image features from scale-invariant keypoints. In *IJCV*, 2004.

[29] D. G. Mumby. Perspectives on object-recognition memory following hippocampal damage: lessons from studies in rats. In *Behavioural Brain Research*, 2011.

[30] T. Murai, S. Okuda, T. Tanaka, and H. Ohta. In *Physiology & Behavior*, 2007.

[31] L. P. J. J. Noldus, A. J. Spink, and R. A. J. Tegelenbosch. Ethovision: A versatile video tracking system for automation of behavioral experiments. In *Behavior Research Methods, Instruments, & Computers*, 2001.

[32] S. Ohayon, O. Avni, A. L. Taylor, P. Perona, and S. E. R. Egnor. Automated multi-day tracking of marked mice for the analysis of social behavior. In *J Neurosci Methods*, 2013.

[33] R. Ribeiro and H. Cachitas. Python video annotator, gui for video analysis. In *http://umsenhorqualquer.github.io/pythonVideoAnnotator/*.

[34] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. In *JMLR*, 2014.

[36] U. Stern, R. He, and C.-H. Yang. Analyzing animal behavior via classifying each video frame using convolutional neural networks". In *Scientific reports*, 2015.

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[38] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014.

[39] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

[40] H. K. Turesson, T. B. R. Conceicao, and S. Ribeiro. Automatic head tracking of the common marmoset. In *bioRxiv*, 2016.

[41] A. Vogel-Ciernia, D. P. Matheos, R. M. Barrett, E. Kramr, S. Azzawi, Y. Chen, C. N. Magnan, M. Zeller, A. Sylvain, J. Haettig, Y. Jia, A. Tran, R. Dang, R. J. Post, M. Chabrier, A. Babayan, J. I. Wu, G. R. Crabtree, P. Baldi, T. Z. Baram, G. Lynch, and M. A. Wood. The neuron-specific chromatin regulatory subunit baf53b is necessary for synaptic plasticity and memory. In *Nature Neuroscience*, 2013.

[42] A. A. Walf and C. A. Frye. The use of the elevated plus maze as an assay of anxiety-related behavior in rodents. In *Nature Protocols*, 2007.

[43] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[44] F. Xiao and Y. J. Lee. Track and segment: An iterative unsupervised approach for video object proposals. In *CVPR*, 2016.

[45] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *NIPS*. 2014.