

A high-throughput framework to detect synapses in electron microscopy images

Saket Navlakha¹, Joseph Suhan², Alison L. Barth^{2,3} and Ziv Bar-Joseph^{1,*}

¹Machine Learning Department, School of Computer Science, ²Department of Biological Sciences and ³Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA 15213, USA

ABSTRACT

Motivation: Synaptic connections underlie learning and memory in the brain and are dynamically formed and eliminated during development and in response to stimuli. Quantifying changes in overall density and strength of synapses is an important pre-requisite for studying connectivity and plasticity in these cases or in diseased conditions. Unfortunately, most techniques to detect such changes are either low-throughput (e.g. electrophysiology), prone to error and difficult to automate (e.g. standard electron microscopy) or too coarse (e.g. magnetic resonance imaging) to provide accurate and large-scale measurements.

Results: To facilitate high-throughput analyses, we used a 50-year-old experimental technique to selectively stain for synapses in electron microscopy images, and we developed a machine-learning framework to automatically detect synapses in these images. To validate our method, we experimentally imaged brain tissue of the somatosensory cortex in six mice. We detected thousands of synapses in these images and demonstrate the accuracy of our approach using cross-validation with manually labeled data and by comparing against existing algorithms and against tools that process standard electron microscopy images. We also used a semi-supervised algorithm that leverages unlabeled data to overcome sample heterogeneity and improve performance. Our algorithms are highly efficient and scalable and are freely available for others to use.

Availability: Code is available at http://www.cs.cmu.edu/~saketrn/detect_synapses/

Contact: zivbj@cs.cmu.edu

1 INTRODUCTION

The mammalian brain can contain hundreds of millions of neurons, each with thousands of specialized connections called *synapses* that enable indirect communication between cells. Estimates for the number of synapses in the mammalian brain ranges into the trillions. Synapses are essential for the transfer of information across neuronal ensembles, and individual synapses can be modulated by patterns of incoming neural activity, a phenomenon thought to underlie learning and memory.

Changes in the relative strength and number of synapses can be regulated by a myriad of factors, including developmental age (Cowan *et al.*, 1984; Huttenlocher and Dabholkar, 1997; Stoneham *et al.*, 2010), sensory experience (Klintsova and Greenough, 1999), drug addiction (Van den Oever *et al.*, 2012), estrus cycle (Cooke and Woolley, 2005) and brain pathology. For example, in a form of autism linked to mutation of the Fragile X gene, spine density in the neocortex is elevated

(Hinton *et al.*, 1991; Pfeiffer and Huber, 2009), a feature that has also been observed in mice carrying the same genetic mutation (Nimchinsky *et al.*, 2001). Rett syndrome, another neurodevelopmental disorder, is characterized by smaller brain size caused by deficits in synaptogenesis (Glaze, 2004; Johnston *et al.*, 2005; Na and Monteggia, 2011) that results in fewer spines. Similarly, in Alzheimer's disease and other dementias, cognitive deficits are associated with reduced synapse density in the hippocampus, a brain structure critical for learning (Clare *et al.*, 2010). Understanding how connectivity across neurons can change is thus an important question that drives contemporary neuroscience research.

Because synapse distribution is a useful and diagnostic criterion to evaluate circuit function in learning and disease, there have been a variety of methods used to estimate synaptic connectivity or overall synapse numbers. Electrophysiological methods to estimate connectivity and the number of inputs per cell can be informative (e.g. Lefort *et al.*, 2009; Yassin *et al.*, 2010), but these approaches are low-throughput and can typically only capture tens or hundreds of connections in reasonable amounts of time (Walz, 2007). MRI-based techniques can be used to study network function at the level of brain regions or voxels, but they do not provide enough spatial resolution to estimate neural connectivity (Sporns, 2010). Anatomically, synapse densities are measured via light-microscopy to identify specialized substructures called spines that stud the dendrites of neurons or using electron microscopy (EM) to identify ultrastructural features that correspond to pre- and post-synaptic elements. Traditional approaches have used cumbersome manual detection to count synapses in these images (e.g. White *et al.*, 1986; Knott *et al.*, 2002; da Costa *et al.*, 2009; Morshedi *et al.*, 2009) and were thus constrained to small-scale measurements or required the use of specialized transgenic animals (Feng *et al.*, 2012; Kim *et al.*, 2012) limiting their usage for studying plasticity and development in wild-type mice.

Since the early 1990s, bioimage informatics has emerged as an important area in the analysis of biological images (Peng, 2008). Imaging datasets are usually much larger than other high-throughput biological datasets (e.g. confocal microscopy data can range in the hundreds of gigabytes for a single imaging session). Accurately identifying elements of interest (molecules, cells, synapses, etc.) within these massive datasets requires the development of sophisticated and efficient computational models. This often involves a classification-based strategy in which a (small) manually labeled training set is used to learn a general model that can be used to analyze a larger collection of images automatically. The key computational challenges involve the reliability and speed at which the analysis is done, as well as dealing with the heterogeneity of biological structures and noise

*To whom correspondence should be addressed.

present in each image. Electron microscopy data suffer particularly from these problems and often contain undesired variation in intensity and contrast within and across samples and preparations. This presents a major computational challenge because model parameters learned from one sample may not generalize to other samples. Although reconstruction and segmentation of conventional EM images has helped answer important questions about brain structure and function (Bock *et al.*, 2011; Denk *et al.*, 2012), these approaches have yet to reach the point of full automation (Merchan-Perez *et al.*, 2009; Jain *et al.*, 2010; Cardona *et al.*, 2012), which has limited their scale and accuracy (Kreshuk *et al.*, 2011; Morales *et al.*, 2011).

2 APPROACH

To aid in identifying and quantifying synapses in EM images, we used a 50-year-old experimental technique (Bloom and Aghajanian, 1966, 1968) to *selectively stain* for synapses in any animal model (Fig. 1A). Unlike conventional EM, this protocol uses ethanolic phosphotungstic acid (EPTA) to pronounce electron opacity at synaptic sites by targeting specific proteins in contact zones. This technique typically leaves non-synaptic membranes (e.g. plasma membranes, neurotubules and vesicles) unstained, though considerable variation can exist from sample to sample. We dissected brain tissue in the mouse somatosensory cortex and performed EM experiments using the EPTA method. We used mice from different ages (P14, P17 and P75) and isolated the same cortical region in each animal to gauge variance in sample quality.

To demonstrate the advantages of this protocol for large-scale synapse identification, we developed a machine-learning framework to detect synapses in these images in a *high-throughput* and *fully-automated* manner. We describe a two-step approach. First, we use a highly accurate first-pass filtration step to reduce the search space of possible synapses by 1–2 orders of magnitude, thereby significantly reducing false positives. Second, we train a classifier to recognize synapses using texture- and shape-based features extracted from small image patches around potential objects of interest. We show that our approach is highly accurate and that it outperforms correlation-based (Roseman, 2004) techniques and an automated technique designed to detect synapses in conventional EM images (Morales *et al.*, 2011).

To further improve classification and adjust for varying experimental conditions (different sample, different microscope, different person performing experiments, etc.), we developed a model that classifies synapses in images using both labeled and unlabeled data. This type of approach is known as *semi-supervised learning* (Zhu, 2005), as it combines ideas from supervised learning (classification) and unsupervised learning (clustering). This technique can be used to build more robust classifiers in cases (such as ours) where large imaging datasets can easily be collected but where it is much harder to manually annotate these images. By integrating unlabeled data in the learning phase, a new sample can help fine-tune parameters of a model built from a previous sample, which can improve accuracy without requiring users to manually annotate images in the new sample. Indeed, we show that a classifier learned only on labeled data from our P14 samples and tested on our P75 samples performs

worse than a semi-supervised classifier that also leverages unlabeled data from P75.

3 METHODS

First, we describe the experiments we performed to gather and process mouse brain tissue for EM imaging with EPTA, and then we describe our machine-learning framework to automatically detect synapses in these images.

3.1 Experimental approach and data collected

We gathered tissue from the mouse somatosensory cortex because it is a well-characterized anatomical area in the cortex (Fox, 2008) and thus serves as an important benchmark for the validity of our experimental approach. To prepare the tissue for EM analysis, we extracted, fixed (using 2.5% glutaraldehyde buffered with phosphate buffered saline) and sectioned 50- μm thick tissue from wild-type C57bl6 mice at ages P14, P17 and P75 with two animals per time point. This range of tissue was collected to determine whether our experimental procedure and subsequent computational analysis remained robust to natural variation in tissue samples within and across time points and to variation in the image acquisition process. Each mouse whisker is somatotopically mapped to a single neocortical column. To ensure that the same cortical region was identified in each sample, we applied a mitochondrial stain (cytochrome oxidase) to each section to visually identify layer 4 (called the *barrel*) of the D1 column/whisker. The barrel was extracted using a dissecting light microscope.

Tissue was prepared for transmission electron microscopy in a series of steps. First, the tissue was washed with three changes (5 min each) of distilled water, followed by an incubation in 0.02% NaOH for 10 min. This latter step was absent from the original procedure of Bloom and Aghajanian (1966, 1968), but we found that it helped increase the contrast of synapses (Fig. 1). Second, the tissue was dehydrated with an ascending series of EtOH (25, 50, 70, 80, 90 and 100%), followed by fixation with 1% phosphotungstic acid (PTA) in 100% EtOH. Third, a small amount, 7 μl , of 95% ethanol, was added to each 1000 μl of PTA stain used, and the PTA was washed from the sample with two changes of 100% ethanol. Fourth, propylene oxide was used as a transitional solvent (the first change of propylene oxide was on ice), and then the specimen was infiltrated with Spurr embedding resin, which was polymerized at 60°C for 48 h. Finally, 100 nm sections were cut using a diamond knife on an ultramicrotome, which were picked up using 50 or 75 mesh copper grids. The specimen was observed using a transmission electron microscope, and images were taken digitally.

We took roughly 130 images per animal (each image is of size 5 μm by 5 μm) covering a total surface area of $\sim 3000 \mu\text{m}^2$ per sample and with a total of six samples (see Supplementary Information for additional details about image acquisition). Images were taken from a single plane, and therefore no correction was necessary to account for double-counting synapses across serially sectioned images. There was significant sample-to-sample variability in the images, but synapses were typically dark with most other biological structures washed away (Fig. 1A). In total, we collected >800 images each containing between 0 and 12 synapses. The raw images were provided as input to the machine-learning algorithms described later in the text.

3.2 Strategies for effectively detecting synapses

As described earlier in the text, electron microscopy images are inherently noisy owing to variations in the samples (e.g. different age), in the manual processing steps and in the image acquisition process. To overcome these issues, we developed a pipeline that uses object segmentation, background information, normalization and alignment to obtain a better feature set that can be used for effective classification.

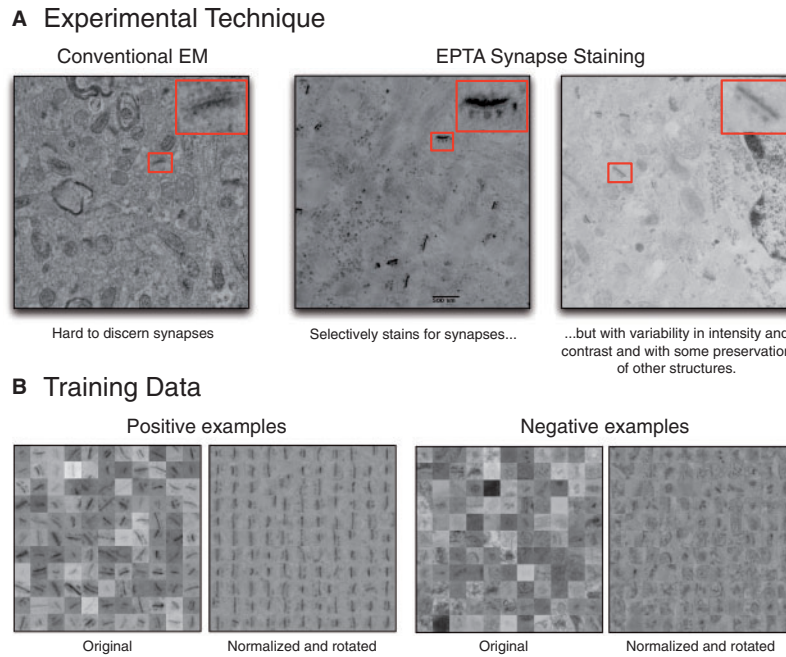


Fig. 1. Experimental technique and training data collected. **(A)** Comparison of conventional and EPTA-stained EM images shows a marked difference in clarity of synaptic structures, albeit high sample-to-sample variability. **(B)** Subset of positive and negative examples taken from multiple images across all samples. The original examples exhibit high variance and noise; normalization and alignment reduces this heterogeneity

1. Reducing the search space of possible synapses via segmentation One popular technique to search for objects in an image is the sliding window approach (Szeliski, 2010). In this approach, non-overlapping windows of a pre-defined size are slid across the image, and each window is classified as synapse-containing (positive) or not (negative). One drawback to this technique is that it can double-count synapses that lie in multiple windows or miss synapses that lie adjacent to each other. A more thorough approach is to use overlapping windows, but this greatly increases computational complexity when using large images, and subsequent post-processing steps still need to account for double-counting synapses. Further, this technique creates a large imbalance of negative-to-positive examples (windows), which may lead to many false positive predictions.

To circumvent these problems, instead of using sliding windows, we apply a filtration step to reduce the number of windows to test by leveraging the fact that the experimental procedure outlined earlier in the text is designed specifically to stain synapses (there may be other biological structures, such as mitochondria and membranes, that also appear in the image, but rarely are there synapses that do not appear as dark). To avoid the challenge of selecting thresholds to segment the image (which may not generalize across samples owing to differences in intensity and enhancement), we adopt the contrast-limiting adaptive histogram equalization algorithm (Zuiderveld, 1994). This method enhances the contrast of each window T in the image to approximately match a flattened histogram by mapping each pixel value $v \in [0, 255]$ to its value in the cumulative distribution $f(v)$ computed in T :

$$f(v) = \frac{1}{n} \sum_{i=0}^v \text{hist}(i) = f(v-1) + \frac{1}{n} \text{hist}(v), \quad (1)$$

where $\text{hist}(\cdot)$ represents the original histogram of the image and n is the number of pixels in the window. We also limit the enhancement by clipping the histogram at a scalar value of 0.20 before enhancement. The window is then rescaled to $[0, 255]$. The equalization is performed in each local window of the image, and then the windows are combined using

bilinear interpolation to eliminate artificially induced boundaries. By only considering small windows of the image, this technique prevents the over-amplification of noise or artifacts that only appear in localized regions of the image.

Next, we binarize the equalized image using a single, *sample-independent* threshold, which was determined manually to be 10% (i.e. only the top 10% of pixels values are kept). The final segmentation is produced by computing connected components (segments) in the binary image. Compared with a non-overlapping sliding window approach that would produce roughly 300 windows to test per 1016×1024 -sized pixel image on average, our segmentation produces 25–35 candidate segments per image and an overall ratio of negative-to-positive examples of 9:1. We also allow for optional filtration of segments that are too small or too large to be synapses, as defined by the user. This step is only designed to produce segments and not to normalize the image, which we do separately later in the text.

To validate that the segmentation step preserves synapses, we looked at two samples (P14 and P75) and manually checked what percentage of the first 100 synapses encountered were correctly segmented. We found that only 1% (1 in each sample) of the synapses were lost, and these were always due to two synapses touching each other, which caused them to be merged into a single segment.

2. Using background cues to augment synapse identification One key indicator to decide whether a candidate segment is a synapse is the physical context in which it lies. We consider a local 75×75 -pixel window around the centroid of each segment to capture information about the object *and* the neighborhood surrounding the object. This is important because elongated synapse-like segments may also appear within mitochondria, but such segments are always surrounded by an oval-like contour marking the boundary of the mitochondria, which can be a useful cue for classification. Similarly, dark circular spots in the nucleus may also be stained (see Fig. 1A), but their neighborhoods typically contain other such spots, which also serve as strong discriminators. These are not

steadfast rules (synapses can certainly lie adjacent to mitochondria), but local neighborhoods are nonetheless a strong visual cue used by EM experts when annotating images (Arbelaez *et al.*, 2011).

3. Handling experimental variation across samples The aforementioned technique produces a set of windows (in which each candidate segment is embedded), but the actual pixel values within these windows may vary significantly from sample-to-sample and image-to-image. Figure 1B exemplifies the discrepancy that can exist across both negative and positive examples in the original images. These issues may be caused by a variety of factors (e.g. consistency of chemical reactions during sample preparation, age of the sample, skill of the experimenter, differential illumination in the microscope, etc.) that are extremely difficult to overcome experimentally and thus must be accounted for computationally. To reduce the effect of these differences, we normalize each candidate window W by computing (Arbelaez *et al.*, 2011): $W' = \frac{W_{xy} - \mu(W)}{\sigma(W)}$, where W_{xy} is the current value at pixel (x, y) and μ and σ are the mean and standard deviation of the pixel intensities in W . Figure 1B shows that this significantly reduces the variance across windows, which helps features dependent on pixel intensities to be compared in an equal setting.

3. Adjusting for synapse heterogeneity Synapses in EM images may be angled at any 2D orientation, which may add undesirable variation in training examples. To create a more invariant set of synapses, we applied the generalized Hough transformation (Duda and Hart, 1972) to automatically rotate the segment (within the candidate window) such that its major axis points vertically. Briefly, this is done by computing the Hough transform matrix \mathcal{H} , where entry i, j of \mathcal{H} corresponds to the number of points in the segment that fall along a line parameterized in polar coordinates as $r_i = x \cos(\theta_i) + y \sin(\theta_i)$. We find the element (r^*, θ^*) in \mathcal{H} that corresponds to the (peak) line for which the most segment points lie. The corresponding θ^* is used to compute the angle of rotation. We then cropped the image to 60×60 pixels to remove the effects of the interpolation (this size is still large enough to fit almost all synapses). The outcome (Fig. 1B) shows much greater uniformity for both synapses and non-synaptic structures compared with the original images.

3.3 Supervised learning framework and features used to detect synapses

After the processing steps aforementioned, each image is reduced to a set of candidate segments defined by a (normalized and aligned) square window around the centroid of the segment. Next, we build an accurate and robust classifier to discriminate between positive (synapses) and negative candidates using texture- and shape-based features.

Texture is a common cue used by humans when manually segmenting structures from electron micrographs, and its use has become popular in many image processing tasks today (Arbelaez *et al.*, 2011; Varma and Zisserman, 2003). In their seminal article, Leung and Malik (2001) defined texture by convolving an image with a bank of 48 Gaussian filters and used the filter responses at each image location to define a 'texton'. Textons were clustered and used to represent and classify images in a reduced dimension. Recently, more effective filter banks have been proposed to represent texture based on modeling joint distributions of intensity values over small and compact neighborhoods of the image (as opposed to the entire image), an approach we also adopt here.

The maximum response (MR8) is one such technique that is derived from a set of 38 filters: 6 orientations \times 3 scales \times 2 oriented filters + 2 isotropic filters. By recording only the maximum response across orientations, the number of responses is reduced to 8 (Varma and Zisserman, 2003). Each pixel is now represented as an 8D vector of responses at its (x, y) location. For each dimension d , we compute a normalized histogram with 17 bins (larger bin sizes yielded marginal gain in performance but increased computational complexity) composed of responses for d over all pixels in the window. Thus, the texture of the window is

represented as a 8×17 -sized vector in \mathbb{R}^n . We used the default parameters for MR8 (Varma and Zisserman, 2003).

Synapses also have a characteristic shape (typically long and elongated) that we also attempt to capture by extracting the following 10 shape descriptors for each segment. These features operate on the binary segment only (they ignore the intensity values of the pixels) and therefore contribute different information than texture alone.

- (1) **Area:** the number of pixels in the segment.
- (2) **Perimeter:** the number of pixels in the boundary of the segment.
- (3) **Major axis:** the number of pixels constituting the major axis of the ellipse that has the same normalized second central moments (covariance) as the segment.
- (4) **Minor axis:** same as above but for the minor axis of the ellipse.
- (5) **Orientation:** the angle between the x -axis and the major axis of the ellipse.
- (6) **Eccentricity:** the ratio of the distance between the foci of the ellipse and its major axis length.
- (7) **Convex area:** the number of pixels in the convex hull of the segment.
- (8) **Solidity:** the proportion of pixels in the convex hull that are also in the segment.
- (9) **Diameter:** the diameter of the circle with the same area as the segment.
- (10) **Extent:** the ratio of pixels in the segment to pixels in the smallest bounding box of the segment.

The final feature we use is the histogram of oriented gradients (HoG) descriptor proposed by Dalal and Triggs (2005). We used nine orientation bins, cell and block sizes of 10×10 and 6×6 , respectively, and a value of 0.2 for clipping the L2-norm. Intuitively, this 334D feature describes the appearance of an object by concatenating the distributions of intensity gradients (edge directions) in different subregions of the window. This descriptor has been shown to outperform other popular feature sets (e.g. PCA-SIFT and generalized Haar wavelets) in a variety of object detection tasks (Dalal and Triggs, 2005; Skibbe *et al.*, 2011). In total, each window is represented by a 480D feature vector in \mathbb{R}^n . All features are scaled to lie in $[0, 1]$.

These features were then used to build a support vector machine (SVM) classifier (Chang and Lin, 2011) using a radial basis function kernel, and we performed a grid search to optimize parameters of the model. We also learned a Random Forest (Breiman, 2001) classifier and an AdaBoost ensemble model (Freund and Schapire, 1995) using 100 trees/learners, respectively. An overview of the supervised algorithm is shown in Figure 2.

3.4 Semi-supervised learning

The supervised algorithm described earlier in the text only uses labeled data to train a classifier. We now show how unlabeled data can be used to further improve the classifier using *co-training*. The co-training algorithm (Blum and Mitchell, 1998) assumes that features can be split into two different and fairly independent sets (in our case, texture and shape). Blum and Mitchell (1998) proved that for a classification problem with two such feature sets, the target concept can be learned based on a few labeled and many unlabeled examples, provided that the views are *compatible* and *uncorrelated*. The compatibility condition requires that all examples are identically labeled by the two classifiers (one for each of the feature sets). The uncorrelated condition means that for any pair of features, the two sets of features are independent, given the label. In real applications, these two conditions are rarely satisfied simultaneously. For example, in our task, compatibility may be hindered due to noise and imaging artifacts. Still, co-training has proven useful in several real-world applications (Zhu, 2005).

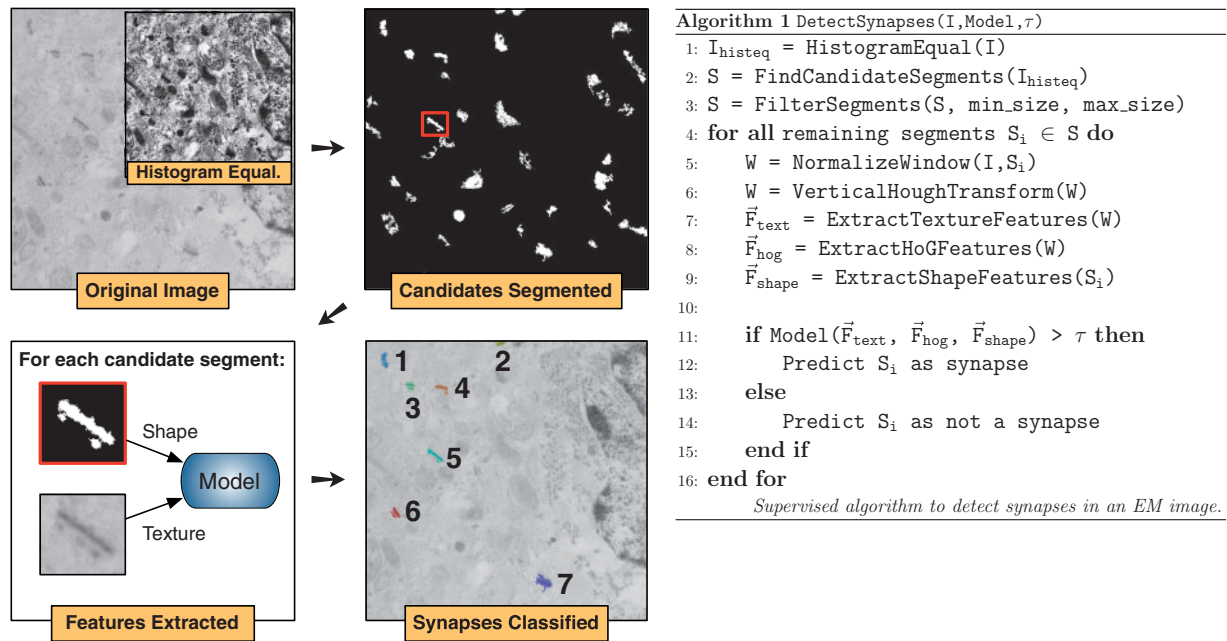


Fig. 2. Main steps of the supervised synapse-detection algorithm and corresponding pseudocode. High-contrast objects are automatically segmented using histogram equalization. Shape-based features are extracted from these objects, as well as texture-based features from a small window surrounding the object. A model is learned from these two features types and is used to classify synapses

In a co-training algorithm, two separate classifiers are trained with the labeled data using the two subfeature sets, respectively. Each classifier is applied to the unlabeled examples, and it outputs a list of positive (putative synapses) and negative examples, ranked by confidence of assignments. We consider all the predictions in each list above a given threshold and any example for which both independent classifiers agree on its label is added to the labeled dataset. We iterate this process once and then retrain a final single classifier using all features from the original set of labeled examples and the new (predicted) set of labeled examples obtained during this iterative process.

In addition to our attempts to normalize each image, this approach provides another way to account for variability in never-before-seen samples without requiring explicit annotation, which can be cumbersome to obtain in practice.

3.5 Testing and comparing with other approaches

For experiments using supervised learning, we manually labeled 11% (59) of the 520 images from P14 and P75 distributed equally across each sample. To do so, we performed the first-pass segmentation described earlier in the text and labeled the resulting windows (segments) as positive or negative. A total of 230 synapses were identified (each showing a clear post-synaptic density and elongated shape) along with 2062 negative examples. The MR8, HoG and shape descriptors were extracted from each example and stored as a 480D feature vector. We performed 10-fold cross-validation and report precision, recall and area under the ROC and precision-recall curves. The confidence in each prediction was measured based on the distance from the test vector to the decision boundary in feature space (for SVM) and based on the proportion of tree agreements for the Random Forest classifier. For supervised learning, we did not use any unlabeled data.

For the semi-supervised approach, we selected all the labeled images from one sample (sample A, e.g. P14) and trained two classifiers using texture and shape features, respectively. Each classifier was then applied to the $\approx 90\%$ of unlabeled images in sample B (P75), and all predictions

in either the positive or negative set in which both classifiers agreed (i.e. both predicted the same label with confidence above a given threshold) were added to the training set. A new single classifier C' was then built using the known labeled examples from sample A as well as the high-confidence examples predicted by the co-trained classifiers. We also learned a baseline classifier C that was only trained on the known labels from sample A. Both classifiers were then tested for accuracy on the true labeled examples from sample B.

We compared our supervised approach against a correlation-based technique (Roseman, 2004) that classifies a test window W as synapse-containing with respect to training set T if $\max_i \text{corr}(W, T_i) > t$; i.e. if the correlation coefficient between W and any positive example in the training set is $> t$, where $t \in [0, 1]$. We also compare against Espina (Morales *et al.*, 2011), a tool designed to detect synapses in conventional EM images (see Supplementary Information).

4 RESULTS AND DISCUSSION

4.1 Validating against conventional EM

One potential concern with using the EPTA method is that some synapses may be washed away alongside other non-synaptic structures. To validate the correctness of our experimental procedure, we tested whether EPTA-stained images preserve roughly the same density of synapses that appear in conventional EM images of the same region. First, we isolated tissue corresponding to the D2 barrel of the mouse somatosensory cortex at P75. In one hemisphere, we performed standard EM chemistry and in the other hemisphere, we applied our EPTA stain. We then asked an expert EM biologist (J.S.) and an expert neuroscientist (A.L.B.) to manually annotate 26 conventional EM images for high- and medium-confidence synapses, and we compared density ratios versus our automated algorithm on the EPTA-stained tissue.

Table 1. EPTA-stained EM images preserve synapse density

Technique	Number of images	Number of synapses	Ratio	Time
Expert 1 (Standard EM, manual counting)				0.77 min/image
High confidence only	26	71	2.73 ± 1.66	
High and medium confidence	26	88	3.38 ± 1.90	
Expert 2 (Standard EM, manual counting)				0.58 mins/image
High confidence only	24	65	2.71 ± 1.55	
High and medium confidence	24	89	3.71 ± 1.85	
Algorithm (EPTA EM, automatic counting)				0.06 mins/image
Threshold = 0.7	275	745	2.71 ± 2.00	
Threshold = 0.5 (default)	275	893	3.25 ± 2.19	

Note: We manually counted high and medium confidence synapses in 24–26 conventional EM images and used our algorithm to automatically count synapses in 275 EPTA-stained EM images of the same region. Different classifier thresholds allow us to closely approximate the true number of synapses.

EPTA appears to conserve synaptic density compared with conventional EM (Table 1). In particular, the two experts found an average of 3.55 high and medium-confidence synapses per image within the conventional EM data versus 3.25 using the EPTA-stained images (an average difference of only 8%). We can even more closely approximate the number of high-confidence-only synapses by increasing the stringency of the classifier (a difference of <1%). Because synaptic density may slightly vary according to pial depth of the specimen even within D2, some variation may be expected across hemispheres.

If the EPTA stain were selectively staining for only asymmetric (excitatory) synapses and not symmetric (inhibitory) synapses, then we would expect a roughly 20% difference (Micheva and Beaulieu, 1995) between the number of synapses counted using EPTA and conventional EM. However, the close correspondence between the two methods suggests that we may be capturing both. This further demonstrates the validity of the EPTA stain as synapse-preserving (Bloom and Aghajanian, 1966) and provides a way of choosing an appropriate threshold for classification.

4.2 Detecting synapses using supervised and semi-supervised learning

As aforementioned, we used texture- and shape-based features to train several different classifiers (SVM, Random Forest and AdaBoost). We next compared the performance of these classifiers as well as a template-matching algorithm previously suggested for this task (Roseman, 2004) (Fig. 3). The SVM trained on both texture and shape features (MR8, HoG and Shape) performed best with an accuracy ranging from 54.8 to 81.3% on the positive set and 99.6 to 96.8% on the negative set depending on the classifier threshold used. Specifically, the default threshold of 0.5 allowed us to recall 67.8% of the true synapses with a precision of 83.3% (Fig. 3B). This significantly outperformed all other classifiers, as well as the correlation-based template-matching algorithm. These results suggest that the EPTA stain facilitates the high-throughput and automated detection of synapses compared with conventional EM. Such a performance would enable the large-scale use of this method to study experience-dependent plasticity in the brain and to detect abnormal changes in synaptic density owing to neurological disorders.

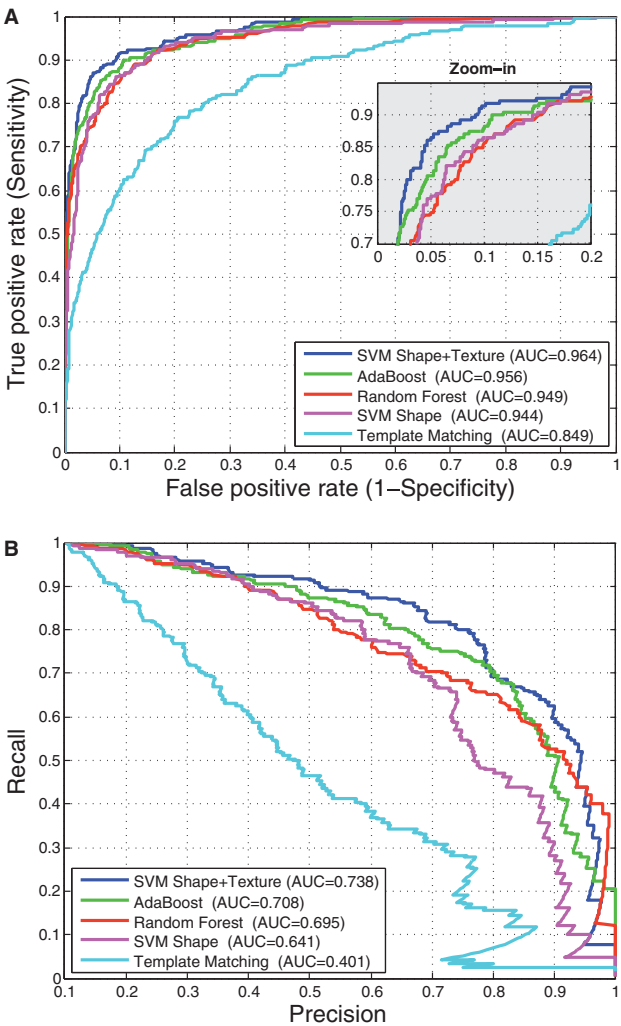


Fig. 3. Cross-validation accuracy of all methods. AUC of the ROC curve (top) and precision-recall curve (bottom) show that the SVM trained using both texture and shape descriptors outperforms all other methods

The shape descriptors by themselves worked well for identifying positive examples but predicted many false positives owing to similar contours in, for example, nucleus regions, which are better discriminated using texture. The HoG features especially benefited from the vertical alignment of the synapses (a 7–10% decrease in false negatives with similar false positives comparing classifiers trained with and without alignment).

To further validate our classifier with respect to human annotation, we took 30 unlabeled images from our P14 sample and asked an independent party (a technician in the Barth lab familiar with the EPTA protocol) to manually annotate all high-confidence synapses. We then applied our supervised classifier to these images using the default classifier threshold of 0.5 and obtained an accuracy of 87.25% (89/102) on the set of positive predictions (66.6% recall) and 96.28% (1164/1209) on the negative predictions. These percentages are similar to those obtained using cross-validation.

Next, we applied semi-supervised learning to detect synapses in new samples for which training data is not available (Table 2). Here, we trained a single classifier C on one sample (either P14 or P75) and used co-training to learn another classifier C' using unlabeled examples from the other sample. We found that the accuracy of C' increases amongst the positive class by 8–12% and increases the AUC by 1–4% compared with the baseline classifier C (see Section 3). We varied the number of unlabeled examples to transfer into C' and found that using the top 0.5% of positive examples (with a corresponding number of negative examples to maintain the same ratio of positive-to-negative) improved accuracy but that including the top 1.5% led to loss in performance.

To explore how semi-supervised learning can account for the variance between mice at the same age, we trained on positive and negative examples from one P75 mouse and used the examples from the other P75 mouse to test (and vice versa). The baseline without semi-supervised learning had an average precision-recall AUC of 68.89% (78.81 and 58.97% individually) and an average ROC AUC of 95.80% (98.48 and 93.12% individually). Then, we used co-training to train a new classifier and

Table 2. Semi-supervised learning boosts cross-sample accuracy

Train/Test	Co-training	Accuracy		AUC	
		Positive (%)	Negative (%)	Prec-recall (%)	ROC (%)
P75/P14	No	66.36	98.20	73.65	96.91
P75/P14	Yes (0.5%)	72.90	98.60	75.75	97.14
P75/P14	Yes (1.5%)	74.77	96.91	73.06	96.65
P14/P75	No	48.78	98.96	60.50	90.38
P14/P75	Yes (0.5%)	60.16	98.21	64.23	92.89
P14/P75	Yes (1.5%)	60.98	97.55	63.80	92.83

Note: We trained a classifier using images from one sample and tested it on images from another sample (first column). This was done without co-training as a baseline and with co-training using two thresholds for selecting the number of unlabeled examples to include (second column). The third and fourth columns show accuracy on the positive and negative classes, and the last two columns show the precision-recall and ROC AUCs, respectively. Bold cells indicate best performance.

improved the average precision-recall AUC to 73.06% (82.45 and 63.66%) and the ROC AUC to 97.15% (98.78 and 95.52%). Both these results demonstrate the power of using semi-supervised learning on unlabeled data, which is often plentiful but ignored in bioimage applications.

Our approach is also scalable and efficient: using unoptimized MATLAB code, we can process a single image in 3.4s on a standard desktop machine using a single processor.

4.3 Development changes in the mouse barrel cortex

Next, we used our method to study developmental changes in synapse density and size in a defined area of the mouse brain, the representation of a single specified (D1) whisker in layer 4 of somatosensory cortex. We performed additional experiments and imaged the D1 barrel in two P17 mice following the same

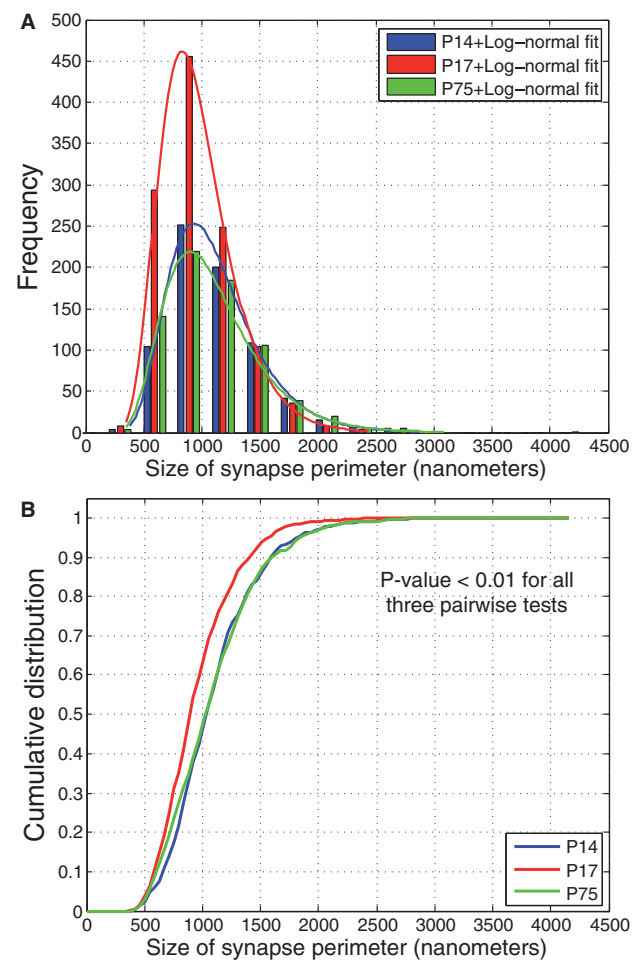


Fig. 4. Developmental changes and log-normal distribution of synaptic strength. (A) Histogram of synapse sizes (as measured by length of the perimeter of the synapse) from two P14 (blue), two P17 (red) and two P75 (green) samples. There is significant growth in number of synapses from P14 to P17, followed by pruning by adulthood (P75). Further, all three distributions closely approximate a log-normal curve, which matches electrophysiological data on synaptic connection strengths (Song *et al.*, 2005). (B) Cumulative distribution of synapse sizes suggest a shift in synapse strength over time ($P < 0.01$, Kolmogorov–Smirnov test)

procedure as outlined in Section 3. We used our classifier to count synapses in images from both time points and, consistent with previous findings (De Felipe *et al.*, 1997), synapse density appeared to rapidly increase from P14 to P17 (2.73 synapses per image at P14 versus 4.25 at P17; Fig. 4A). Further, we found that synapse density decreases in adults to about the same level as P14 (2.78 synapses per image at P75), which is also consistent with prior work on pruning (White *et al.*, 1997). Thus, our pipeline can be used to better understand the rates of developmental processes and can be used to define the precise timelines of their occurrence.

Finally, to demonstrate the utility of the EPTA stain beyond simply detecting synapses, we used shape features to characterize the *strength* of each synapse based on its well-established correlation with spine size (Lee *et al.*, 2012). Modulation in synaptic strength is an important facet of neuroplasticity and development (Wen and Barth, 2011) with larger contact areas likely to transmit more current to the post-synaptic neuron. We used our classifier-predicted synapses in all samples (P14, P17 and P75) and plotted the distribution of the perimeter of all synapses detected (larger perimeter \rightarrow larger size \rightarrow stronger synapse). Structurally similar synapses may appear to have different sizes when projected onto 2D because the image cross-sections may be taken at any angle. Exactly correcting for this bias would involve 3D reconstruction of synapses from stacks of EM images, which is a separate and interesting problem in its own right (Jain *et al.*, 2010; Kreshuk *et al.*, 2011; Merchan-Perez *et al.*, 2009; Morales *et al.*, 2011).

We found that the distribution of synapse size at all ages fits a log-normal distribution ($P < 0.01$ by Shapiro–Wilk and Anderson–Darling tests; Fig. 4A), which is consistent with previous electrophysiological data (Song *et al.*, 2005). We also observed that there was a significant shift in the distribution of synapse size toward the emergence of small synapses during the period of rapid increase in synapse density at P17, without a concomitant loss of large synapses within this short developmental time window (Fig. 4B). These data suggest that developmental processes may preferentially enable the addition of new synapses without necessarily augmenting the size of already-existing synapses. The data analysis pipeline that we have established can thus be used to generate and test specific hypotheses about synapse growth and maturation in the developing neocortex in a high-throughput and statistically robust manner.

5 DISCUSSION AND CONCLUSIONS

Synaptic density and strength are dynamically modulated in the brain and are important facets for understanding neural circuitry and function. Experience-dependent plasticity, circuit development and neuropathologies have all been linked to changes in the number and strength of synapses in the brain, and thus their characterization is a useful parameter to facilitate our understanding of network function. Although current experimental techniques for studying these conditions have many advantages [e.g. electrophysiology provides a wealth of data not extractable from images, and conventional EM can be used to resolve synapses to extract pre- and post-synaptic partners and thus potentially allow for reconstruction of microcircuits (Denk *et al.*, 2012)], several interesting questions can also be answered by

sampling from brain regions of interest over several conditions or time points and generating high-confidence large-scale statistics of synaptic structures present. But to do so requires both clarity in the data and robust algorithms to explore this data.

We used an old experimental technique for selectively staining synapses in EM images. This technique does not require specialized animal models for enhancing synapses, and we validated it on new tissue of the mouse somatosensory cortex and against conventional EM. We collected >800 images from three time points and developed a fully automated and high-throughput machine-learning framework that detected thousands of synapses in these images. We also used semi-supervised learning to learn models that can adapt to variability in new samples by leveraging unlabeled data. Such an approach is suitable for several other bio-image classification problems that also face issues of sample heterogeneity. Our approach is general and scalable enough to handle large datasets and is freely available for others to use.

For future work, it would be a great interest to perform immunocytochemistry using EPTA against GABA and AMPA receptors to separately classify symmetric and asymmetric synapses. Accuracy could also be improved by detecting synapses in 3D, though this would require many more images and more sophisticated computational techniques that can automatically segment, align and reconstruct synapses across serial sections. The 2D sampling-based strategies (e.g. considering sections separated by $10\mu\text{m}$) may certainly miss synapses, but if the same procedure is applied to each sample at each time point, the relative number of synapses per image or unit area can still be compared in a fair manner. Further, to remove some biases in 2D analysis caused by larger synapses, previous works have proposed formulas to adjust counts based on the average size of synaptic profiles observed in the sample (Coggeshall and Lekan, 1996; Mayhew, 1996; Huttenlocher and Dabholkar, 1997), which we can also use.

ACKNOWLEDGEMENTS

S.N. would like to thank Fernando Amat and Gustavo Rohde for advice on developing the image processing algorithms and Patrick Beukema for manual annotation of EPTA images.

Funding: National Institutes of Health (1RO1 GM085022) and National Science Foundation (DBI-0965316) awards to Z.B.J.

Conflict of Interest: none declared.

REFERENCES

- Arbelaez, P. *et al.* (2011) Experimental evaluation of support vector machine-based and correlation-based approaches to automatic particle selection. *J. Struct. Biol.*, **175**, 319–328.
- Bloom, F.E. and Aghajanian, G.K. (1966) Cytochemistry of synapses: selective staining for electron microscopy. *Science*, **154**, 1575–1577.
- Bloom, F.E. and Aghajanian, G.K. (1968) Fine structural and cytochemical analysis of the staining of synaptic junctions with phosphotungstic acid. *J. Ultrastruct. Res.*, **22**, 361–375.
- Blum, A. and Mitchell, T. (1998) Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT)*, COLT'98, pp. 92–100, ACM, New York, NY, USA.
- Bock, D.D. *et al.* (2011) Network anatomy and in vivo physiology of visual cortical neurons. *Nature*, **471**, 177–182.
- Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

- Cardona, A. *et al.* (2012) TrakEM2 software for neural circuit reconstruction. *PLoS One*, **7**, e38011.
- Chang, C.-C. and Lin, C.-J. (2011) LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 27:1–27:27.
- Clare, R. *et al.* (2010) Synapse loss in dementias. *J. Neurosci. Res.*, **88**, 2083–2090.
- Coggeshall, R.E. and Lekan, H.A. (1996) Methods for determining numbers of cells and synapses: a case for more uniform standards of review. *J. Comp. Neurol.*, **364**, 6–15.
- Cooke, B.M. and Woolley, C.S. (2005) Gonadal hormone modulation of dendrites in the mammalian CNS. *J. Neurobiol.*, **64**, 34–46.
- Cowan, W.M. *et al.* (1984) Regressive events in neurogenesis. *Science*, **225**, 1258–1265.
- da Costa, N.M. *et al.* (2009) A systematic random sampling scheme optimized to detect the proportion of rare synapses in the neuropil. *J. Neurosci. Methods*, **180**, 77–81.
- Dalal, N. and Triggs, B. (2005) Histograms of oriented gradients for human detection. In: Schmid, C., *et al.* (eds), *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2, IEEE Computer Society, Washington, DC, USA, pp. 886–893.
- De Felipe, J. *et al.* (1997) Inhibitory synaptogenesis in mouse somatosensory cortex. *Cereb. Cortex*, **7**, 619–634.
- Denk, W. *et al.* (2012) Structural neurobiology: missing link to a mechanistic understanding of neural computation. *Nat. Rev. Neurosci.*, **13**, 351–358.
- Duda, R.O. and Hart, P.E. (1972) Use of the hough transformation to detect lines and curves in pictures. *Commun. ACM*, **15**, 11–15.
- Feng, L. *et al.* (2012) Improved synapse detection for mGRASP-assisted brain connectivity mapping. *Bioinformatics*, **28**, 25–31.
- Fox, K. (2008) *Barrel Cortex*. Cambridge University Press, Cambridge.
- Freund, Y. and Schapire, R.E. (1995) A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the Second European Conference on Computational Learning Theory (EuroCOLT)*. pp. 23–37. Springer-Verlag, London, UK.
- Glaze, D.G. (2004) Rett syndrome: of girls and mice—lessons for regression in autism. *Ment. Retard Dev. Disabil. Res. Rev.*, **10**, 154–158.
- Hinton, V.J. *et al.* (1991) Analysis of neocortex in three males with the fragile X syndrome. *Am. J. Med. Genet.*, **41**, 289–294.
- Huttenlocher, P.R. and Dabholkar, A.S. (1997) Regional differences in synaptogenesis in human cerebral cortex. *J. Comp. Neurol.*, **387**, 167–178.
- Jain, V. *et al.* (2010) Machines that learn to segment images: a crucial technology for connectomics. *Curr. Opin. Neurobiol.*, **20**, 653–666.
- Johnston, M.V. *et al.* (2005) Rett syndrome and neuronal development. *J. Child Neurol.*, **20**, 759–763.
- Kim, J. *et al.* (2012) mGRASP enables mapping mammalian synaptic connectivity with light microscopy. *Nat. Methods*, **9**, 96–102.
- Klintsova, A.Y. and Greenough, W.T. (1999) Synaptic plasticity in cortical systems. *Curr. Opin. Neurobiol.*, **9**, 203–208.
- Knott, G.W. *et al.* (2002) Formation of dendritic spines with GABAergic synapses induced by whisker stimulation in adult mice. *Neuron*, **34**, 265–273.
- Kreshuk, A. *et al.* (2011) Automated detection and segmentation of synaptic contacts in nearly isotropic serial electron microscopy images. *PLoS One*, **6**, e24899.
- Lee, K.F. *et al.* (2012) Examining form and function of dendritic spines. *Neural Plast.*, **2012**, 704103.
- Lefort, S. *et al.* (2009) The excitatory neuronal network of the c2 barrel column in mouse primary somatosensory cortex. *Neuron*, **61**, 301–316.
- Leung, T. and Malik, J. (2001) Representing and recognizing the visual appearance of materials using three-dimensional textures. *Int. J. Comput. Vision*, **43**, 29–44.
- Mayhew, T.M. (1996) How to count synapses unbiasedly and efficiently at the ultrastructural level: proposal for a standard sampling and counting protocol. *J. Neurocytol.*, **25**, 793–804.
- Merchan-Perez, A. *et al.* (2009) Counting synapses using FIB/SEM microscopy: a true revolution for ultrastructural volume reconstruction. *Front. Neuroanat.*, **3**, 18.
- Micheva, K.D. and Beaulieu, C. (1995) An anatomical substrate for experience-dependent plasticity of the rat barrel field cortex. *Proc. Natl Acad. Sci. USA*, **92**, 11834–11838.
- Morales, J. *et al.* (2011) Espina: a tool for the automated segmentation and counting of synapses in large stacks of electron microscopy images. *Front. Neuroanat.*, **5**, 18.
- Morshedi, M.M. *et al.* (2009) Increased synapses in the medial prefrontal cortex are associated with repeated amphetamine administration. *Synapse*, **63**, 126–135.
- Na, E.S. and Monteggia, L.M. (2011) The role of MeCP2 in CNS development and function. *Horm. Behav.*, **59**, 364–368.
- Nimchinsky, E.A. *et al.* (2001) Abnormal development of dendritic spines in FMR1 knock-out mice. *J. Neurosci.*, **21**, 5139–5146.
- Peng, H. (2008) Bioimage informatics: a new area of engineering biology. *Bioinformatics*, **24**, 1827–1836.
- Pfeiffer, B.E. and Huber, K.M. (2009) The state of synapses in fragile X syndrome. *Neuroscientist*, **15**, 549–567.
- Roseman, A.M. (2004) FindEM—a fast, efficient program for automatic selection of particles from electron micrographs. *J. Struct. Biol.*, **145**, 91–99.
- Skibbe, H. *et al.* (2011) SHOG: spherical hog descriptors for rotation invariant 3d object detection. In *Proceedings of the 33rd International Conference on Pattern Recognition, DAGM'11*. pp. 142–151. Springer-Verlag, Berlin, Heidelberg.
- Song, S. *et al.* (2005) Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biol.*, **3**, e68.
- Sporns, O. (2010) *Networks of the Brain*. MIT Press, Cambridge.
- Stoneham, E.T. *et al.* (2010) Rules of engagement: factors that regulate activity-dependent synaptic plasticity during neural network development. *Biol. Bull.*, **219**, 81–99.
- Szeliski, R. (2010) *Computer Vision: Algorithms and Applications*. 1st edn. Springer-Verlag New York, Inc, New York, NY.
- Van den Oever, M.C. *et al.* (2012) The synaptic pathology of drug addiction. *Adv. Exp. Med. Biol.*, **970**, 469–491.
- Varma, M. and Zisserman, A. (2003) Texture classification: are filter banks necessary? In *Proceedings of the 2003 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2, pages II–691–8 vol. 2.
- Walz, W. (2007) *Patch-Clamp Analysis: Advanced Techniques*. Humana Press, Neuromethods Series, Totowa, New Jersey, USA.
- Wen, J.A. and Barth, A.L. (2011) Input-specific critical periods for experience-dependent plasticity in layer 2/3 pyramidal neurons. *J. Neurosci.*, **31**, 4456–4465.
- White, E.L. *et al.* (1997) A survey of morphogenesis during the early postnatal period in PMBSF barrels of mouse Sml cortex with emphasis on barrel D4. *Somatosens. Mot. Res.*, **14**, 34–55.
- White, J.G. *et al.* (1986) The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **314**, 1–340.
- Yassin, L. *et al.* (2010) An embedded subnetwork of highly active neurons in the neocortex. *Neuron*, **68**, 1043–1050.
- Zhu, X. (2005) Semi-supervised learning literature survey. *Technical report*. Computer Sciences, University of Wisconsin-Madison, Madison.
- Zuiderveld, K. (1994) *Graphics Gems IV. Chapter Contrast Limited Adaptive Histogram Equalization*. Academic Press Professional, Inc, San Diego, CA, pp. 474–485.