# Predicting distance maps using Basic Image Features

Tom Wyatt, word count 4,941

7th March 2012

## Abstract

Complete wiring diagrams of nervous sytems, called connectomes, would contain invaluable information for neuroscientists. The segmentation of electron micrograph images of nervous tissue is a rate limiting step in the acquisition of such connectomes. Here, the field of connectomics is broadly reviewed along with a closer review of the segmentation problem. A method is proposed which uses distance maps to improve segmentation. The ability of random forests to use Basic Image Features to predict such distance maps is assessed.

## 1 Introduction

The nervous systems of animals are formed from vast networks of cells called neurons. Within this network, neurons, which are connected by synapses, send and receive signals to each other via electrochemical waves. They can thus coordinate the actions of animals by detecting physical stimuli, processing information and activating responses. The field of neuroscience seeks to further our understanding of all aspects of this network and has made inspiring progress since its beginnings in the neuron diagrams of Santiago Ramón y Cajal (see figure 1) which supported the formation of the neuron doctrine in the 19th century.

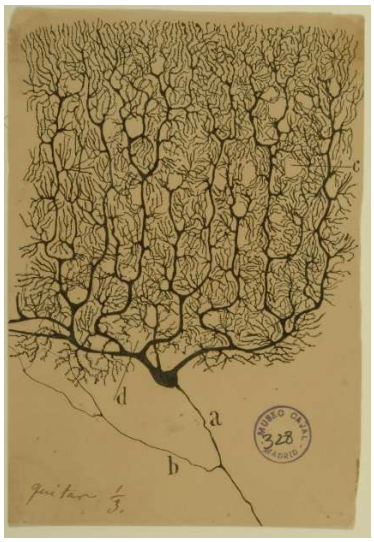


Figure 1: A drawing of a Purkinje neuron by Santiago Ramón y Cajal (1899).

This report is concerned with a branch of neuroscience, called connectomics, which strives to create ever more detailed maps of entire nervous systems. Proponents of connectomics argue that such maps, which they call connectomes, contain invaluable information for neuroscientists [1, 2]. Obtaining connectomic data, however, is a difficult task, even for the simplest of nervous systems. This is foremost due to the incredible complexity of nervous systems - neurons are capable of making thousands of connections and do so via long branching filaments, called axons and dendrites, pictured in figure 1. With an electron microscope (EM) it is possible to acquire connectomic data at the most detailed level since they are powerful enough to resolve all axons and dendrites, along with even smaller intracellular structures such as vesicles, which help identify synapses.

To turn the EM images into wiring diagrams, the

numerous filaments of each neuron must be traced throughout the image volume and the points where they form synapses with other neurons must be identified. Currently, this is the rate limiting step in connectomics. With training the work can be done manually but the required labour quickly becomes impractical as connectome size increases. Manual segmentation takes roughly $200 - 400$ hours per mm to trace neurons [3] and a typical neurite density is 4.5 km per cubic mm [4].

To circumvent this problem an automated or at least semi-automated method is required. Many techniques from the field of computer vision have been developed and applied to EM image segmentation, with some degree of success. To date, however, no method has been found which is as accurate as manual segmentation.

In this report I will describe an approach I explored which attempts to improve on existing techniques by using Basic Image Features (BIFs) to predict a distance map of the segmentation. It has been proposed that this predicted distance transform could help in creating accurate segmentations.

The rest of this report is structured as follows. Firstly, I will describe the biological question at hand by broadly reviewing the field of connectomics, discussing its importance to neuroscience and explaining the achievements of the field to date. In section 3 I will focus on image segmentation, reviewing some of the approaches taken to tackle the problem along with the metrics which have been developed to measure performance. Finally, I will explain the novel method of image segmentation that I explored and describe the results which were obtained.

# 2 Connectomics

## 2.1 Connectomics rationale

The field of connectomics has frequently been criticised. One criticism is that it is descriptive, whereas true science should be hypothesis driven. It is worthwhile, therefore, to briefly discuss the rationale behind the connectomics approach.

The rationale can be understood via analogous approaches in the field of genetics. In 1988 Francis Crick summarised a view still held by many today when he advised "if you want to understand function, study structure" [5]. Indeed genetics was revolutionised by the discovery made by himself, Rosalind Franklin and James Watson regarding the structure of DNA in 1953 [6]. In 2003, 50 years after their discovery, the Human Genome Project published data detailing the entire DNA sequence of a single person [7]. Among other things, this data allowed genes to be examined as part of a complete, interacting network, rather than as isolated objects. The Human Genome Project is today widely considered a huge success and the developments it lead to have had huge ramifications.

Likewise, structure is studied in connectomics in the hope that it will give insight into function. Indeed it could be argued that it is unlikely that the full function of the brain could be understood without knowing the detailed structure. It is also argued, in analogy with whole genome sequencing, that it is important to be able to examine the system as a whole because investigation into isolated fragments can provide only limited insight into an interacting network. A related argument is that entire networks will inevitably be pieced together as different regions of interest are investigated but that a concerted effort is much more efficient and will produce a more consistent result.

Sceptics would argue that the above is a misleading analogy. A connectome is more malleable than a genome as it must constantly develop as memories form. Also, whereas 99.9% of the genome is shared by the entire human population [8], it is unlikely that such homology would be found between connectomes. In my opinion, these are important criticisms but can only be addressed after further connectomic research.

## 2.2 Obtaining connectomes

The first step in obtaining connectomic data is EM imaging. Any EM technique must involve taking thin slices of the tissue, the thickness of which will dictate the resolution in direction perpendicular to the imaging plane. One method, called serial section transmission electron microscopy (SSTEM), painstakingly re-

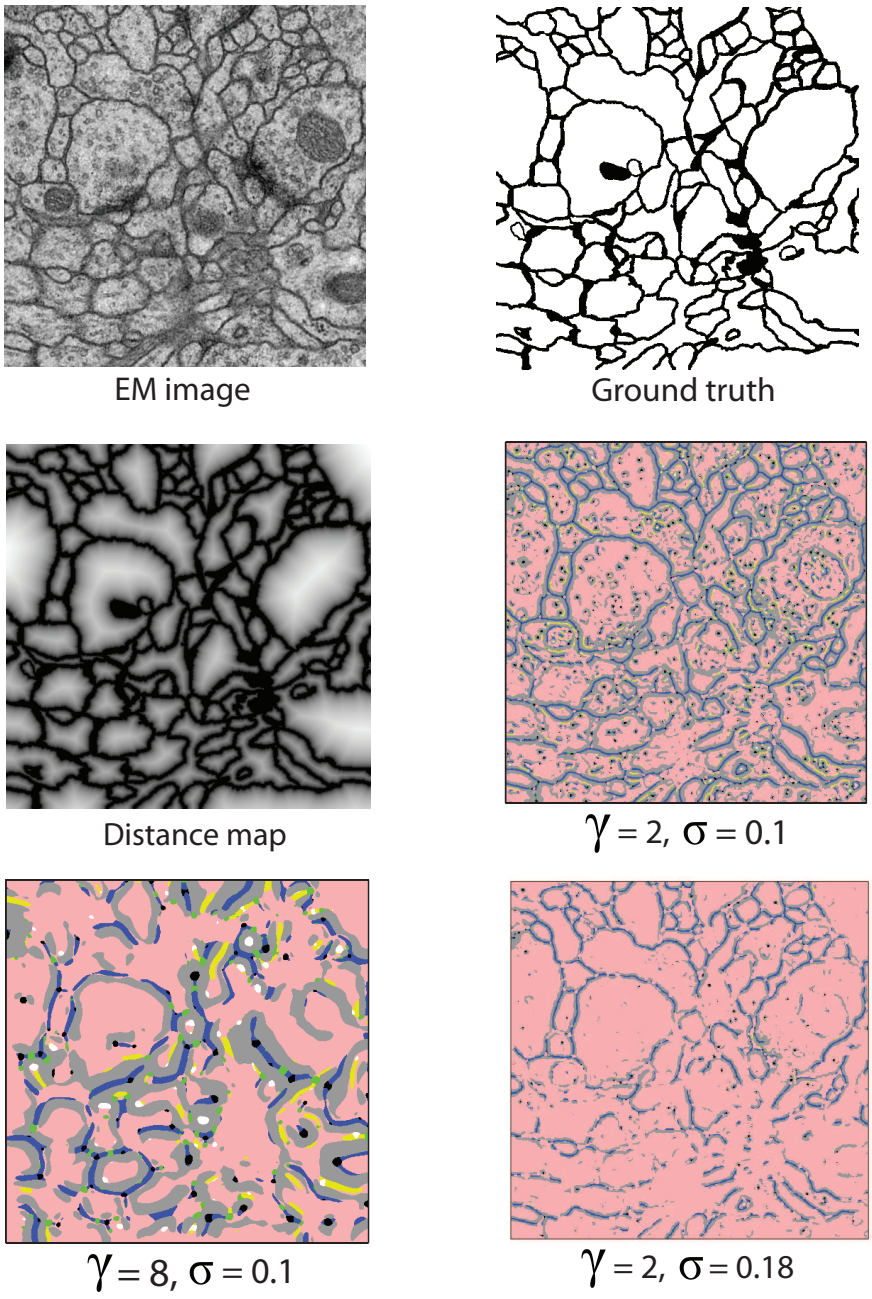


Figure 2: Top left is a raw EM image of a Drosophila Brain. Top right is the corresponding ground truth. Middle left is a distance map of this ground truth. The final three images are BIFs of the raw EM image with varying values of the key parameters.

moves slices from the tissue sample for imaging with a transmission electron microscope (TEM) [9]. Using a diamond knife, slices as small as 45 nm have been achieved [9]. Unfortunately this is not small enough to reliably detect the smallest of neuronal processes when they run in the imaging plane, since these can be as small as 50 nm in diameter. It is also hard to avoid creating folds, tears and other distortions in the tissue during the sectioning process.

An alternative to SSTEM is to scan the surface of the tissue sample using a scanning electron microscope (SEM) and then, after each image is taken, remove a thin slice. This method, called serial block-face scanning electron microscopy (SBFSEM), avoids imaging the distortions created during sectioning [10]. It is made possible by using low electron beam energies which, at around 3 kV, don't penetrate further than 25 nm [10]. This allows good depth discrimination. Diamond knife microtomes are able to remove layers as thin as 25 nm in SBFSEM and even smaller increments can be taken using focused ion-beam milling [11].

Once images have been collected the next step is to retrieve useful information from them. The most important task is to produce a wiring diagram. This involves tracing all axons and dendrites through the volume and identifying where they make connections with other neurons. On top of this, it is often useful to identify the neurons' classes and to measure the size and shape of synapses to approximate connection strength and to predict if synapses are excitatory or inhibitory.

The post-imaging processing can be broken down into 5 steps. The first step is registration. This is the alignment and stitching of the separate images to create a uniform 3D volume. Next is segmentation. This is where cellular membranes are traced in the 2D images to divide the image into separate cells. The next step is linkage, where links are chosen between the segmented areas in adjacent 2D images in order to trace neurons through the volume. Since mistakes are inevitably made, the fourth stage is proof reading. This final stage is annotation, which requires synapses and other features to be labelled and the neuronal class of cells to be identified. It is the segmentation stage which this report will focus on.

## 2.3 Connectomic results to date

The crowning achievement of connectomics to date is undoubtably the completed connectome of *Caenorhabditis elegans* [12]. This simple worm contains only 302 neurons so was a natural starting point. The nervous system was imaged using SSTEM and then manually segmented [12]. In all, the project took over a decade to complete, the most time consuming factor being the manual reconstruction. Importantly, this project was a proof of principle: it proved SSTEM could be used to produce a complete connectome. The data has since been improved [13] and used extensively to measure network properties such as small-worldness [13] and to formulate hypotheses regarding the neural basis of certain behaviours such as touch avoidance [14] and feeding [15].

Since the completion of the C. Elegans connectome, over 20 years ago, no other connectome has been completed. This is because manual reconstruction of EM data for a more complicated organism is not feasible and automated techniques are not yet at a high enough standard. Nevertheless, partial connectomes of various animals have been obtained to answer specific questions.

Recently a section of a mouse retina was reconstructed using SBFSEM in a study related to the retina's detection of directed motion [16]. By combining the connectomic data with 2 photon functional calcium imaging it was possible to discover an asymmetry in the number of connections between inhibitory starburst amacrine cells and direction selective ganglion cells, dependent on ganglion cell preferred direction. Crucially, this result supported several models of direction detection processes while ruling out others. A similar study used the same combination of techniques to image the mouse primary visual cortex [17]. They were able to verify a controversial prediction made via physiological evidence that inhibitory interneurons in this part of the nervous system receive convergent input from neighbouring excitatory neurons.

Connectome data has also been used to discount assumptions that were used for many years in inferring synaptic connectivity from light microscopy. The assumption which came to be called "Peters' rule"

stated that synaptic connectivity could be estimated by multiplying the number of "potential synapses" by the "connectivity fraction" [18]. Here the number of "potential synapses" is defined as the number of times a dendrite comes within a certain distance of another cell's axon and can be counted from light microscopy images. A section of rat hippocampus was reconstructed from SSTEM images to show that Peters' rule in this case was not a good assumption [19].

Other studies apply dense reconstruction to important questions but focus strongly on developing the approaches which will be required to obtain connectomic data on larger scales. In [20], SSTEM was used to image a *Drosophila* first instar larval brain neuropile and explored which network motifs appeared most frequently. At the same time they were demonstrating how the software package TrakEM2 [21] could be used to dramatically reduce the effort required in the image processing stages.

Finally, there has been research into formulating what kind of information could be obtained in the future. Seung, for instance, in [2], outlines possible ways in which procedural and spatial memories could "read" from pairwise connectivity data and speculates how this could be extended improved with dense data.

# 3 Image Segmentation

## 3.1 Overview

As mentioned earlier, the image segmentation stage is currently by far the most labour intensive stage in obtaining connectomic data. Since manual segmentation is so time consuming, the key to finding the connectomes of more complex animals is in improving the automation of segmentation. The high degree of accuracy with which humans are able to segment the images proves that the images themselves contain enough information for the task and suggests it should be possible for a computer to do the same. The problem ultimately falls in the remit of the field of computer vision.

A huge variety of techniques have been applied to the problem. In general, the problem comes in three stages. Firstly information must be extracted from the image. Secondly this information must be used to predict the positions of boundaries. Lastly these predictions must be turned into segmentations. The work presented in section 4 focusses on the first two stages.

A key differentiating factor between methods is to what degree machine learning is used. Machine learning is the use of training data, for which there is a known ideal answer, in automatic tuning of the technique. To this end, large databases of EM images have been created along with the human segmented images which are called the ground truths [22].

Some early attempts used no machine learning at all. In [23] for example, information was extracted by convolving a manually crafted filter with the image and used this with a classifier which had been manually designed to process the results.

Most modern approaches, however, use some degree of machine learning. Boosted Edge Learning [24] has been shown to an effective technique. This trains a tree classifier (described in section 3.4) using information from a large bank of features, including filter responses and gradient information, all over large range of scales. A method referred to as $P_b$ has also been shown to be effective [25]. This looks for discontinuities in brightness, colour and texture and uses this information to train a logistic regression model.

In the purest applications of machine learning, the features themselves are also learnt. This avoids the effort of hand crafting filters for each particular application and has been shown to be very successful using convolutional neural networks [26].

## 3.2 Metrics

Being able to quantify the performance of a given segmentation is of fundamental importance because it allows techniques to be compared and provides a quantity to be maximised in machine learning.

Here I will describe three popular metrics, all formulated as measures of error in relation to a ground truth. The first is pixel error. This takes the binary image prediction and simply calculates the fraction of pixels that disagree with the corresponding ground truth pixels. This has been criticised as a poor mea-

sure of true segmentation performance. One misclassified pixel, for instance, could lead to the merging of two sections but would not be strongly penalised.

A perhaps more appropriate measure of performance is Rand error [27]. If the $ith$ pixel of the ground truth is in segment number $T_i$ and the $ith$ pixel of the predicted segmentation is in segment $S_i$ then the Rand error $R(T,S)$ is given by:

$$R(T,S) = \binom{N}{2}^{-1} \sum_{i \neq j} \left| \delta(S_i, S_j) - \delta(T_i, T_j) \right|,$$

where $\delta(X,Y) = 1$ if $X = Y$ and $\delta(X,Y) = 0$ if $X \neq Y$. This gives the fraction of pixel pairs for which there is disagreement (between ground truth and the predicted segmentation) as to whether or not the pixels are in the same segment. Minor translations of boundaries are penalised much less harshly than in pixel error and splits and mergers lead to an appropriately bad score. A criticism of this metric is that it weighs a topological error by the area (or ultimately volume) of the segments in question. It is, however, simple to implement and can be easily estimated for large images by taking a random sample of pixel pairs.

A metric introduced more recently is the warping error [28]. The warp error was devised to penalise only topological errors such as splits, mergers or the introduction of holes. It completely ignores errors in the location of the boundaries. A precise definition of the warping error requires ideas from the field of digital topology which are beyond the scope of this report, however, the principle can still be understood. First the "best warping" of the predicted segmentation onto the ground truth is produced. This is a transformation which can make any changes at all to the boundary shapes as long as no topological properties are changed (i.e. no splits, mergers etc). In this way the predicted segmentation is warped until every last pixel is as close to the ground truth as possible. The warping error is then defined as the pixel error between this best warping and the ground truth. Therefore, the warping error, when multiplied by the number of pixel in the image, is an upper bound to the number of topological errors made in the prediction. A key criticism is that the final number is hard to interpret. Some topological errors lead to more than one wrong pixel in the best warping and so are more heavily weighted, yet it is not clear precisely when this occurs.

## 3.3 Basic Image Features

Basic Image Features (BIFs) [29] are a set of 7 types of local image structure and have been shown to be useful tools in computer vision. They've been used successfully in object recognition [30] and texture classification [31] and are therefore a natural choice for image segmentation.

In the BIF regime, structure is identified by considering symmetry. Mathematically, a symmetry is defined by a transformation (such as a reflection or rotation) which leaves an object unchanged. For instance, a black circle can be rotated about its centre or reflected in any line which passes through its centre and remains unchanged. Other basic structures in images, such as gradients and straight lines, can similarly be defined in terms of symmetries.

BIFs are identified by measuring the filter response of an image to set of 2D Gaussian derivative filters. These filters are given by

$$G_\sigma(x) := \left(2\pi\sigma^2\right)^{-\frac{1}{2}} e^{-\frac{x^2}{2\sigma^2}},$$

$$G_\sigma^{(n)}(x) := \frac{d^n}{dx^n} G_\sigma(x).$$

The 2D filters are then given by

$$G_\sigma^{(m,n)}(x,y) := G_\sigma^{(m)}(x)\, G_\sigma^{(n)}(y).$$

The filter response at pixel $i$ is the inner product of the image centred on that pixel, $I(i)$, and the Gaussian filter:

$$j_{mn}(i) = \left\langle G_\sigma^{(m,n)} \mid I(i) \right\rangle.$$

These are then normalised, defining $s_{mn}$ by

$$s_{mn} := \sigma^{m+n} j_{mn}.$$

To define BIFs, these scale-normalised filters are combined in a way which maximises their sensitivity to symmetries. We make the definitions:

$$\lambda = (s_{20} + s_{02})$$

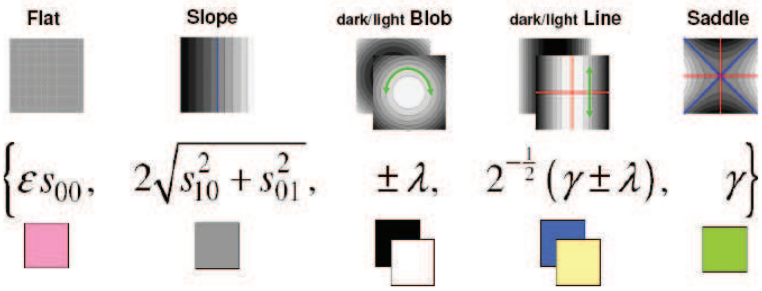


Figure 3: Basic Image Features

$$\gamma := \sqrt{\frac{1}{4}\left(s_{20}-s_{02}\right)^2 + s_{11}^2}.$$

With these definitions the BIF definitions, along with the shapes which they are designed to be sensitive to, are shown in figure 3. Each pixel in an image can be assigned one of the 7 BIFs according to which filter it is most sensitive too. Three examples of EM images, colour coded by its BIF classifications, are shown in the figure 2. The difference between them is the value of the parameters $\sigma$ and $\epsilon$ that was used. The $\sigma$ parameter effectively changes the size of the area of the image that the filter looks at. The $\epsilon$ parameter dictates the fractional contrast threshold, below which the area is defined as flat. These parameters will play a crucial role in my application of BIFs to the segmentation problem.

### 3.4 Classification and regression by trees

Many approaches to the problem of segmentation will contain a classification or regression problem. These apply situations where the value of a set of variables, called predictors, $X_1, X_2 \ldots X_N$ are known and a prediction is required of an outcome $Y$. For example, pixels may nee to be classified as membrane or not membrane, given intensities of surrounding pixels.

Here I will focus on tree classifiers. These are particularly appropriate when the predictors interact in complicated, non-linear ways making the creation of a global model infeasible.

A classification tree consists of branches which repeatedly split at nodes until the point where they stop, which is called the leaf node. At each node a question is asked about one or more of the predictors and the branch which is followed to the next node depends on the answer. This is repeated until the leaf node is reached, which contains the predicted outcome $Y$.

A simple method for growing regression trees, using training data, is to make the split which minimises the quantity $S$, given by

$$S = \sum_{leaves} \sum_{i} (y_i - \overline{y})^2$$

where $y_i$ are the values that $Y$ takes in the data. $\overline{y}$ is given by

$$\overline{y} = \frac{1}{n}\sum_i y_i$$

where $n$ is the number of data points. The point at which dividing stops can be defined as a value by which $S$ must be reduced by for further divisions to be worthwhile.

Numerous approaches have been tried in order to improve the predictive power of classification trees. It has been found that growing an ensemble (or "forest") of slightly different trees and combining the predictions they make can lead to significant improvement. One such method is bagging [32]. Here a boot-strap sample of the data is taken to grow each tree. That is, for a data set of size $n$, a set of $n$ samples is taken *with replacement*, in order to grow the tree. A method called random split selection has also been shown to improve results [33]. Here, the split at each

node is chosen at random from the best few splits. The improvement found in such methods has been attributed to the randomness which is introduced and averaged over.

One of the most successful such methods is the random forest technique [34]. Again, an ensemble of trees are grown and again each tree uses a bootstrap sample of the data. This time, however, randomness is introduced by choosing at random at each node a total of $m$ of the $N$ predictors with which to base the split on, where $m$ is a parameter to be chosen (or tuned). At the predicting stage each tree gets a unit vote and the most popular result is chosen. Random forests have been shown to be more accurate, more robust to noise and outliers and faster to implement than other similar classifiers [34]. Another advantage is the possibility of internal estimates of error, i.e. estimates which don't require new data. This is because the bootstrap sampling means that for each tree roughly 1/3 of the data was not involved in growing the tree and so can be used to estimate error. This is referred to as out of bag error.

A criticism of random forest and similar classifiers is that the prediction mechanism is not easily understandable. In many cases it is desirable to know which of the predictors are important and to find if any are dependent on others. In random forest a few, rough methods have been developed to gain this kind of insight. By randomly permuting the values of the $p$th predictor between out of bag data points and finding what the resulting trees vote, it is possible to find how important that predictor is by observing the reduction in the number of trees which voted for the correct class. To some degree, it is also possible to find dependence between predictors by measuring error when predictions are based on only one predictor and then measuring how things improve when other predictors are introduced.

The random forest classifier has been used in a divers range of fields and in particular has was used in [35] and [36] in segmenting EM images.

# 4 Predicting Distance Maps

## 4.1 Method

The method I will describe here incorporates distance maps in an attempt to improve the segmentation of EM images. A distance map of a binary image is the image whose every pixel has a value equal to the distance to the nearest object. The distance map of a ground truth labeling is shown in figure 2. Since the object in the original image was membrane, each pixel value is the distance to the nearest membrane.

I explored a method of predicting a distance map from EM images using random forest regression on BIF data at a range of values for $\sigma$ and $\epsilon$. The rationale behind trying to predict a distance map is that a genuine distance map has a rigid structure. At each point there must be a direction in which the gradient is 1 and the gradient must be $-1$ in the opposite direction. The exception to this is for pixels with distance zero and pixels which lie on the morphological skeleton of the segment area, these have two directions with gradient $-1$.

By predicting a distance map, along with an estimate for its uncertainty, it would be possible to make further predictions as to what the genuine distance map was. One can imagine an algorithm which focussed on altering pixels with high uncertainties, so as to make the distance map obey the rules outlined above.

BIFs were calculated at a range of $\sigma$ and $\epsilon$ values to provide the random forest with maximal information with which to predict the distance map. This is intuitively reasonable: since $\sigma$ controls the scale over which the BIFs take in information a range of distance information must be contained. Ultimately this idea can only be tested through testing.

The data used was obtained in [20]. The 30 images were split into 25 to train the random forest and 5 to test it. This made it possible to test for over-fitting. Since each of the images was 512 by 512 pixels in size, the total amount of data generated was huge so just a random selection was used to train each random forest.

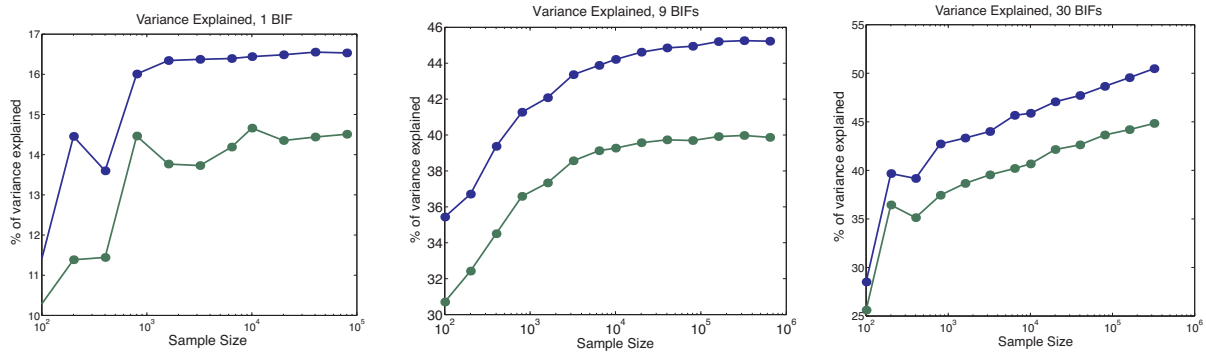


Figure 4: The variation in the performance of random tree models in predicting distance maps as a function of the amount of data the model was trained with.

## 4.2 Results

The chosen values of $\sigma$ and $\epsilon$ were$\{1, 2, 4, 8, 14\}$ and $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.3\}$ respectively. This gives a total of 30 predictors from which to predict the distance map. I also trained random forests using just 9 and just 1 predictors using values for $\sigma$ and $\epsilon$ of $\{2, 4, 8\}$and $\{0.005, 0.1, 0.3\}$, and $\{2\}$ and $\{0.1\}$ respectively. These values were not at all tuned, but selected by visually inspecting a large range of values and taking values which seemed to represent the range of qualitative effects produced. Tuning these values could be a useful way to improve the method.

The main measure of the predictive power of the random forest which I have used is

$$v = 1 - \frac{\delta^2}{\Sigma^2},$$

where $\Sigma^2$ is the variance of the data and $\delta^2$ the mean square error of the predictions. This represents the total variance explained by the random forest model.

Figure 4 shows $v$ as a function of the amount of data points used to train the random forest. For all graphs the predictions of the test data are significantly worse than the predictions of the training data. I do not believe this is due to over-fitting of data since the performance is worse even when very small amounts of data are used. The difference is more likely to be due to qualitative differences in the test images themselves, in relation to the training images.

Using just 1 BIF, the model extracts the maximum possible amount of information from a relatively small amount of data, as shown by the plateau starting at around 1000 data points. When 9 BIFs are used we see a similar plateau but a lot more data is needed and a large increase in performance is attained.

The plateaus in both the 1 and 9 BIF cases are not the type of behaviour you see in cases which are over-fitted, where performance continually increases for the training data and plateaus or decreases for the test data. This is interesting: why can the model not continually improve its predictions by seeing more data? The answer is simple for the 1 BIF case: there are only 7 values the predictor can take. Therefore, the best it can do when given one of these values is predict the average distance for this value in the training data it saw. In the 9 BIF case however there are $7^9 = 40353607$ different combinations the predictors can make and the performance begins to plateau after the model has trained on only around 100000 data points. The only explanation is that there is a high level of redundancy in the data. It was discovered that indeed the BIFF class was frequently the same across a large proportion of the 9 BIFs. This shows that more careful choice of $\sigma$ and $\epsilon$ values is important.

In the case of 30 BIFs no plateau is observed. In the final data point 320000 data points were used to train the random forest. This took several days

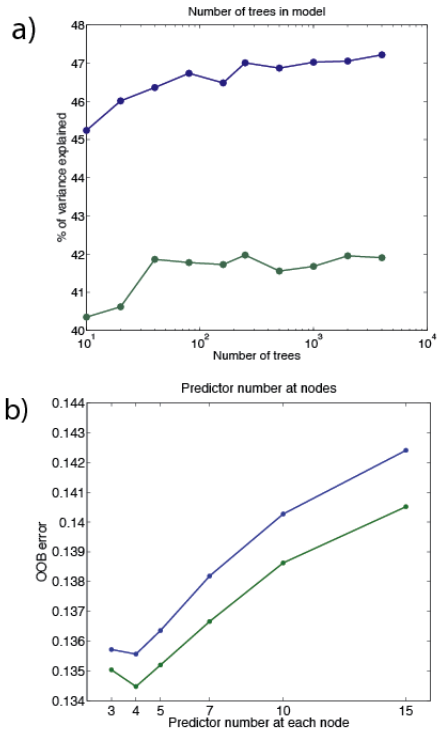


Figure 5: a) the variation in random forest model with the number of trees used and b) the affect that the number of predictors used at each node has on error. Note the different measure of performance in b).

of computation so no further data could be collected due to time restrictions. Computational times should not yet be a real problem, since the model was not trained on a particularly powerful computer.

Two additional parameters in random forest regression are the number of trees used and the number of predictors used at each node. The affect these had on the performance of the model was examined and the results are shown in figures 5 a) and 5 b). The number of trees used had little effect. Performance slowly increases with number of trees but the improvement is unlikely to be worth the additional computational burden.

The affect of changing the number of predictors used at each node is highly significant. Note the use of OOB error rather than variance explained. The standard choice for this parameter in regression is $m = \frac{D}{3}$ where $D$ is the total number of predictors available. Accordingly all previous models used a value of $m = 10$. The graph, however, predicts that much better results could be obtained by using $m = 4$ and so further models were trained at this value. The reason for this value is unclear and further work would be required to explain it.

Using $m = 4$ and training with 640000 samples, 51.9% of the variance of training data was explained. This is encouraging, the model is already out-performing a simpler version of distance map prediction from BIF data which explained at best around 48.6% (unpublished data). Since there is no sign of performance plateauing, it is fair to assume further improvement is possible.

The distance map that is predicted is shown in figure 6 (bottom). The results look good visually. In the figure, a measure of uncertainty has been overlaid in red. This uncertainty was estimated as the variance in the distances predicted by all trees. The figure reveals that the variance is highest around structures inside that cell which are not membrane. This is expected since these structures can look similar to membrane. The high uncertainty around these structures allows us to effectively detect them and could be very useful in extending this method to improving segmentations.

# 5 Conclusion

The data that connectomes provide has already be shown to provide immensely useful data to neuroscientists, as shown by the discoveries made using the connectome of *Caenorhabditis elegans* and the partial connectomes of the rat hippocampus and mouse retina, among others. Connectome data of more complex animals seems not far from our grasp. It will certainly not be sufficient data to explain all workings of any nervous system, but I believe it will be necessary.

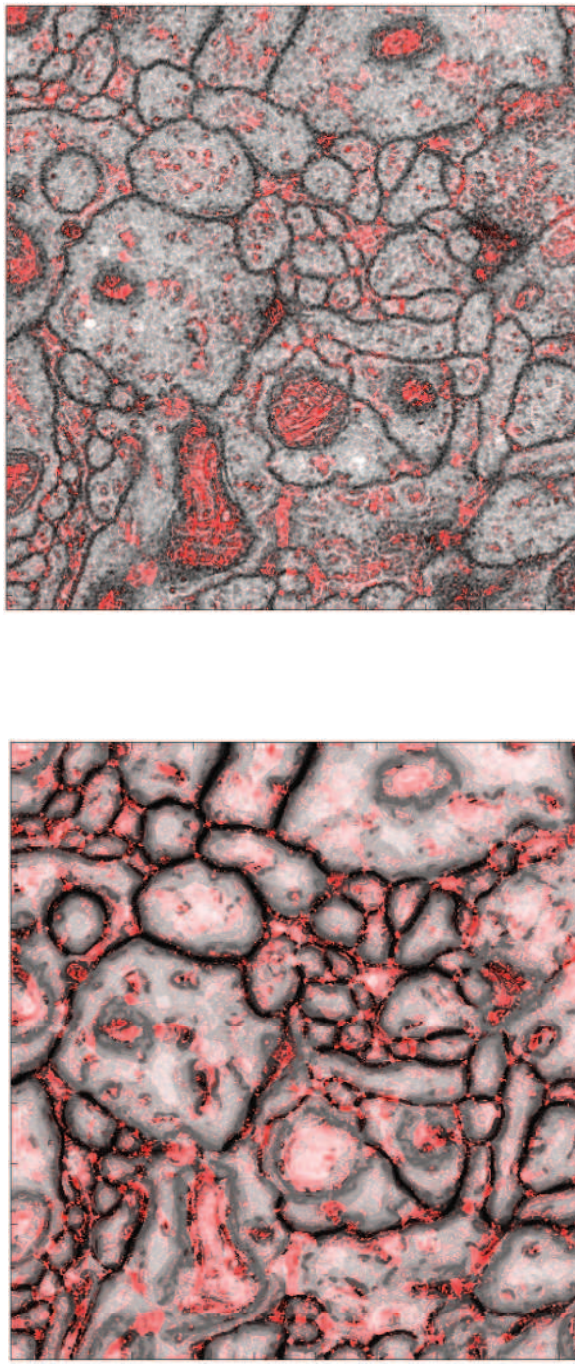

Figure 6: Top shows one of the raw test EM images. Bottom shows a distance map predicted by the random forest. In both the uncertainty in the predictions is overlaid in red.

Daunting challenges must be faced in order to acquire this data, not least the fast and accurate automatic segmentation of EM images. The results presented here describe just one part of a method which could be used to improve on current segmentation approaches. It has been shown that BIF data can be used with random forest regression in order to predict a distance map to greater accuracy than achieved before. Whether or not the prediction will be accurate enough to use in an effective segmentation from the processed distance map and whether the idea of using a distance map at all will be successful, are both still open questions.

# References


[1] J. W. Lichtman and J. R. Sanes, *Curr. Opin. Neurobiol.* **18,** 346-353 (2008).

[2] H. S. Seung, *Neuron* **62,** 17-29 (2009).

[3] M. Helmstaedter, K. L. Briggman and W. Denk, *Curr. Opin. Neurobiol.* **18,** 633-641 (2008).

[4] V. Braitenberg and A. Schuz, *Statistics and Geometry of Neuronal Connectivity*, Heidelberg (1998).

[5] F. Crick, *What Mad Pursuit: A Personal View of Scientific Discovery,* Basic Books (1988).

[6] J. D. Watson and G. H. C. Crick, *Nature* **171**, 737-738 (1953).

[7] http://www.ornl.gov/sci/ techresources/Human_Genome/project/ hgp.shtml.

[8] D. A. Wheeler M.Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y. J. Chen, V. Makhijani, G. T. Roth et al., *Nature* **452,** 872-876 (2008).

[9] K. M. Harris, E. Perry, J. Bourne, M. Feinberg, L. Ostroff and J. Hurlburt, *J. Neurosci.* **26,** 12101-12103 (2006).

[10] W. Denk and H. Horstmann, *PLoS Biol* **2,** e329 (2004).

[11] J. A. Heymann, M. Hayles, I. Gestmann, L. A. Giannuzzi, B. Lich and S. Subramaniam, *J. Struct. Biol.* **155,** 63-73 (2006).

[12] J. G. White, E. Southgate, J. N. Thomson and S. Brenner, *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences* **314**, 1-340 (1986).

[13] L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, D. B. Chklovskii, *Comp. Biol.* **7** (2001).

[14] M. Chalfie, J. E. Sulston, J. G. White, E. Southgate, J. N. Thomson, and S. Brenner, *J. Neurosci.* **5,** 956–964(1985).

[15] L. Avery, and H. R. Horvitz, *Neuron* **3,** 473–485 (1989).

[16] K. L. Briggman, M. Helmstaedter and W. Denk, *Nature* **471,** 183-188 (2011).

[17] D. D. Bock, W. A. Lee, A. M. Kerlin, M. L. Andermann, G. Hood, A. W. Wetzel, S. Yurgenson, E. R. Soucy, H. S. Kim and R. C. Reid, *Nature* **471,** 177-182 (2011).

[18] A. Peters, and M. L. Feldman, *J. Neurocytol.* **5,** 63–84 (1976).

[19] Y. Mishchenko, T. Hu, J. Spacek, J. Mendenhall, K. M. Harris and D. B. Chklovskii, *Neuron* **67,** 1009-1020 (2010).

[20] A. Cardona, S. Saalfeld, S. Preibisch, B. Schmid, A. Cheng, J. Pulokas, P. Tomancak, V. Hartenstein, *PloS Biology* **8** (2010).

[21] http://www.ini.uzh.ch/˜acardona/ trakem2.html.

[22] http://www.eecs.berkeley.edu/ Research/Projects/CS/vision/bsds/.

[23] J. Canny, *IEEE Trans Pattern Anal. Mach. Intell.* **8,** 679-698 (1986).

[24] P. Dollar, Z. Tu and S. Belongie, *Supervised learning of edges and object boundaries* (2006).

[25] D. R. Martin, C. C. Fowlkes and J. Malik, *IEEE Trans. Pattern Anal. Mach. Intell.* **26**, 530-549 (2004).

[26] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, *Proc. IEEE* **86**, 2278-2324 (1998).

[27] W. Rand, *Journal of the American Statistical Association*, (1971).

[28] V. Jain, B. Bollmann, M. Richardson, D. R. Berger, M. N. Helmstaedter, K. L. Briggman, W. Denk, J. B. Bowden, J. M. Mendenhall, W. C. Abraham, K. M. Harris, N. Kasthuri, K J. Hayworth, R. Schalek, J. C. Tapia, J. W. Lichman and H. S. Seung, *IEEE* (2010).

[29] L. D. Griffin, M. Lillholm, M. Crosier and J. van Sand, *SSVM*, 343–355, (2009).

[30] M. Lillholm and L. Griffin, *IEEE* (2008).

[31] M. Crosier and L. D. Griffin, *Int. J. Comput. Vis.* **44,** 447-460 (2010).

[32] L. Breiman, *Machine Learning* **24,** 123-140 (1996).

[33] T. Dietterich, *Machine Learning*, 1–22. (1998).

[34] L. Breiman, *Machine Learning* **45,** 5-32 (2001).

[35] B. Andres, U. Kothe, M. Helmstaedter, W. Denk and F. Hamprecht, *Pattern Recognition,* 142-152 (2008).

[36] V. Kaynig, T. Fuchs and J. Buhmann, *CVPR*, (2010).

[37] S. C. Turaga, K. L. Briggman, M. Helmstaedter, W. Denk and H. S. Seung, arXiv:0911.5372v1 (2009).

[38] D. B. Chklovskii, S. Vitaladevuni and L. K. Scheffer, *Curr. Op. Neurobiol.* **20,** 667-675 (2010).