UNIVERSITY OF
BIRMINGHAM

Medical Image Processing with Quadtrees

School of Computer Science
University of Birmingham

Josh Wainwright

Supervisor: Iain Styles

Date: September 2014

Abstract

This study deduces that reionization began at a redshift of $z = 17.82(+3.06, -2.4)$ and ended at a redshift of $z = 7 \pm 1.8$. This is calculated by directly applying the dynamics of star formation and the ionization rate of neutral hydrogen in the Inter-galactic Medium. A photometry strategy consisting of 3 multi-band surveys is proposed in order to observe Lyman Break Galaxies across redshifts 6–17. The surveys will locate 100.5 ± 37.0, 138.7 ± 100.6, 358.1 ± 158.6 galaxies in redshift ranges 6–8.5, 8.5–10 and 10–17 respectively. These surveys will be completed by the James Webb Space Telescope and Euclid which are planned for launch in the coming decade. A follow up spectroscopy survey will be used to confirm the redshift and properties of 24, 4 and 48 galaxies in these 3 surveys respectively. The spectroscopy will be carried out using James Webb Space Telescope and a combination of single and multi-slit spectroscopy. It is shown that the use of known gravitational lenses, located between redshift 0.5–0.7, is very beneficial for discovering high redshift candidates as it can increase the depth of surveys by up to 3 magnitudes.

Table of Contents

Part I

Introduction

1 Medical Imaging

In various scientific fields, viewing and imaging objects smaller than the human eye can naturally observe is an ability eagerly sought. Microscopy in medical fields has allowed us to learn about the nature of tissues, micro-organisms and cells and the way they work together and to develop preventative measures and cures for injuries and diseases.

The humble microscope, used as far back as the 1500's, is able to show us a world that is not usually visible, but in recent times, as our understanding has grown, we have desired to see beyond what ordinary microscopes are capable and have invented machines that let us catch a glimpse of some of the smallest structures, molecules. However, when the objects to be viewed get this small, of the order of a few tens of nanometers, the light that is used to view them become the limiting factor.

2 Sub-Diffraction-Limit Imaging

Imaging objects becomes more difficult as they get smaller because of the wavelength of light. Once two objects are separated by a distance of an order similar to that of the wavelength (λ) of the light used to view them, it is no longer possible to resolve these two objects apart, instead all that can be seen is a blur of the two objects together.

There have been several techniques developed for distinguishing objects apart on smaller and smaller scales. Many of these involve using different wavelengths of light. For example, instead of being limited by visible light, $\lambda \approx 5 \times 10^{-7}$m, x-ray radiation ($\lambda \approx 10^{-10}$m) or even electrons ($\lambda \approx 10^{-11}$m) can be used to resolve smaller scales in x-ray and electron microscopy respectively. These, however, have the issue that, because the smaller wavelengths imply higher energies, there is the danger of destroying the sample.

The minimum distance that two objects can be resolved at is given by Abbe's criterion,

$$d = \frac{\lambda}{2NA} \tag{1}$$

$$= \frac{\lambda}{2n \sin\theta}, \tag{2}$$

where NA is the numerical aperture of the microscope, the range of angles that the microscope's lens will let light through properly. For $\lambda \approx 500\,\text{nm}$ (in the middle of the visible range), and $NA = 1.5$, the maximum resolving distance is $d = 160\,\text{nm}$, this is the diffraction limit. This is an order of magnitude larger than the objects that need to be resolved. A few attempts to avoid this limit using exotic types of lenses have been developed [Fang et al., 2005], but these are currently far more expensive to use than traditional imaging equipment.

2.1 Image Manipulation

Instead of trying to avoid the diffraction limit using shorter wavelengths of light or other particles, other techniques employ different methods of actually capturing the image, or cleaver manipulation of the images that are produced, to get around the limitations of the diffraction problem.

For example Stochastic Optical Reconstruction Microscopy (STORM) [Rust et al., 2006] and PhotoActivation Localisation Microscopy (PALM) [Owen et al., 2010] use a technique where the objects to be imaged are molecules of a fluorescent dye. These are attached to the object of interest, a cell or sample of tissue for example. The type of dye molecule used allows the fluorescence to be switched on and off, allowing some markers to be imaged separately to others, effectively increasing the distance between points. Once an image is captured, the point spread function (PSF) of the point is used to locate the single marker, the "on" markers are changed and the image retaken. When many of these images are taken, they can be combined to provide accurate information on the original location of the markers and hence the shape and dimensions of the object.

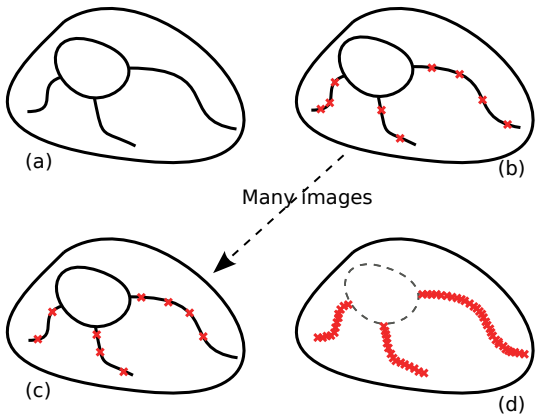


Figure 1: *STORM imaging. (a) The actual structure to be imaged is too small for regular microscopy. In stages (b) through (c), many images are taken, each with a different subset of the fluorescent markers activated. When the images are combined, (d), the points add up and reveal the nature of the object.*

3 Benchmarking

Throughout the project, a set of files will be used to test the algorithms that are developed; their correctness and effectiveness, speed and resource use. These files contain real data formated in the same way as would be expected for data given to the plugin in general use. The files that will be used are detailed in Table 1.

Note that `palm-3-small.txt` is a subset of `palm-3.txt` which is used for simply checking correctness of algorithms. A summary of the columns that are included in the files, used and unused fields, is included in Appendix A.

File Name	Size	Points
`palm-1.txt`	12 MiB	65572
`palm-2.txt`	6.4 MiB	36672
`palm-3.txt`	5.8 MiB	33342
`palm-3-small.txt`	176 KiB	1000
`uniform.txt`	22 MiB	2000000

Table 1: *These files containing sample data are used for benchmarking throughout the project.*

Part II

Data Structures

The way in which data is represented in memory has a large effect on the speed, efficiency and effectiveness of any algorithms that are performed on that data. There are almost always a number of tradeoffs that must be considered when choosing or designing a data structure to hold data: speed of access vs. speed of search or traversal, storage space used vs. time to insert a datum, etc.

Some data structures may be naïvely chosen based on one of these, at the expence of the other. For example, consider storing the pixel information for a sparse image generated by a number of single pixel points—a linear array might be selected. This would give extremely good access and modification times, both $O(1)$, but insertion and deletion are very slow. There would also be a large amount of wasted space since every pixel that is black, i.e., does not have any points in, would need to have an array index with the same entry, 0.

To decide the most approriate data structure, a number of different approaches are implemented and tested under different types of data and for a range of different operations performed on them.

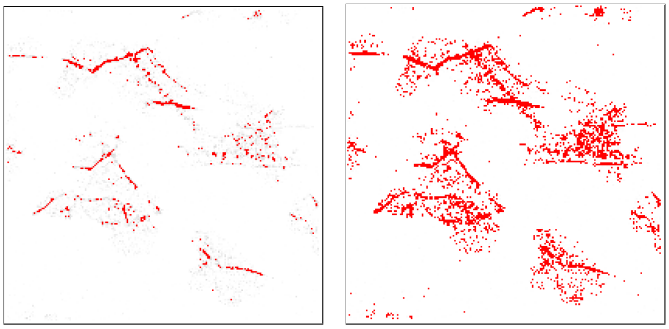
4 Simple Grid Method

The simplest method for analysing the distribution of points is to use a regular grid of cells and place the points into the cells one at a time. Once all points have been added, the number of points per cell can be treated as a grey scale brightness value. This gives a simple pixel image, with brightness as a function of density of the points, in the `pnm` image format. A thresholding filter can then be applied to remove the points that are isolated and leave the denser areas corresponding to clusters.

Though the resolution of this method can be easily changed by altering the size of the cells and the grid, it performs badly when presented with data that is even slightly noisy. If the clusters themselves have a density that is not significantly above the background noise level, the thresholding step is prone to either exclude much of the real data, or to increase the size of the clusters by including too much noise. These two effects can be seen clearly in Figure 2, where `palm-1.txt` was used with a cell size of 200. The range of the data is from 0 to 41000 for both the x and the y axes, thus the images are 205 by 205 pixels. This data took 495 ms to generate.

There are steps that can be taken to improve the approach of this simple grid when handling outlying points caused by noise.

1. First the algorithm is modified to include a thresholding step before writing the data to a file. This means that the pixels can be adjusted with greater accuracy and any arbitrary level can be chosen to threshold at.
2. Next, once an image has been generated, the number of points that contributed to each pixel is no longer of interest and so the image can be converted to a binary image. This is an image with just two possible values, the first represents white space, where there are no points, the seconds is black where there were points and so is of interest.
3. Once a binary image has been generated, erode and dilate filters can be applied to remove remaining outliers and to try to close the gaps in the structures that have been identified so that they are more solid.



(a) **(b)**

Figure 2: *Setting a low threshold, (a), means that many of the points in the clusters are lost. Setting it higher, (b), includes too many of the points deemed to be noise. Pixels are cells that ended with points, red are points that would be kept by the threshold process, black would be removed.*

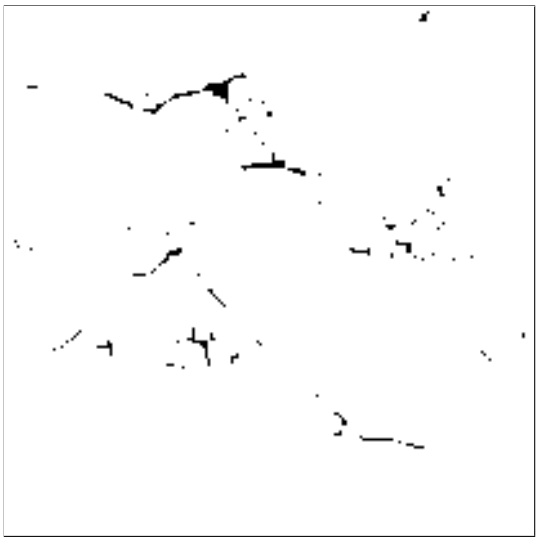


Figure 3: *Using a `close`-ing algorithm can help to emphasise the structure in the data and, at the same time, remove the isolated points representing noise.*

These steps lead to significantly better isolation of the interesting parts of the image, as can be seen in Figure 3, however, much of the detail of the structure is lost in this process.

5 Quadtrees

Since the simple grid method described in section 4 performs slowly and does not offer good cluster analysis, a different approach is needed. The chosen method is to use a quadtree data structure.

Quadtrees are a type of recursive abstract data type in the form of a tree where every node has exactly zero or four chil-

dren. A node with zero children is a leaf and contains some information, value or quantity. A node with four children is not a leaf and cannot hold information.

Quadtrees are often used in image processing since the four children of the root node can naturally represent the four quadrants of the image; upper left, upper right, lower left and lower right. Since each of these children is also a quadtree, the image can be subdivided to any arbitrary depth. From this point, information about the image can be "seen" more easily by the computer and statistics calculated.

5.1 Quadtree Definitions

It is useful to define a few terms that shall be used in the context of quadtrees. Many of these are identical to the definitions more commonly applied to binary trees.

Node (Cell) A leaf in the tree which holds a number of points.

Root The node at the topmost position in the tree.

Child One of the four nodes that are beneath a given node for which there is a direct path of length 1 between this node and it.

Height The length of the longest path from the root node to one of the leaf nodes.

Depth The length of the path from a given node up to the root.

Completeness A quadtree is 'complete' when each of the root node's four children have the same height. In this case, the number of leaves in the tree is a maximum.

Quadtree Code (Code) The unique binary number assiged to a node by adding its position with respect to its siblings to the code of its parent.

5.2 Code Orderings

In order to identify a node uniquely in the tree, each node is given a code that is built up from it's parent code plus some value that identifies it among it's siblings. The root node is usually chosen to have an empty code so that the first four children are given the first level codes.

The choice of what order to label the children is important if the order in which the nodes are placed is important. For spatial indexing, for example, each node represents a quadrant in two dimensional space, so being able to traverse the children in a sensible and predictable way is essential.

5.2.1 Morton Order (Z-Order)

Perhaps the most natural order to give to the values in a spatial quadtree is to number them from 1 to 4, left to right, top to bottom. This can be made more appropriate for a computer to use by numbering from 0 to 3. This is called Morton Order [Morton, 1966] or Z-order because of the resulting path that would be followed by traversing the nodes in order, Figure 4a. This has several useful features.

1. First, the numbers can be converted to base 2:
   - 0 becomes 00
   - 1 becomes 01
   - 2 becomes 10 and
   - 3 becomes 11.
2. This has advantages since binary is very efficient for computers to work with and allows certain tricks to be employed (see Morton order coordinates).
3. Also, this numbering system is easily extendible to any depth of tree that can be imagined.
   (a) The root, as mentioned before, is given no value,
   (b) each of the children are numbered 00 through 11.
   (c) the children of these children are numbered 00 to 11 with the parent as a prefix. So the children of node 00 are 0000, 0001, 0010 and 0011. Likewise, the children of 11 are 1100, 1101, 1110 and 1111.
   (d) The children are always numbered in the same order. If starting at the top and going top to bottom and left to right, this is maintained for all children.

This method of numbering is simple and so acceptable for the standard uses of quadtrees, but it was found to be difficult to work with in a spatial context when information about neighbouring cells is needed. The steps required to calculate the neighbours of any given cell are reasonably complex and so would add computational and time complexity to calculations performed on the tree.

Morton Order Coordinates

Another useful feature of the Morton ordering is the simple conversion from quadtree code to Cartesian coordinate notation. The steps to convert to coordinate form are:

1. Ensure the code is in binary format with two bits for each level of the tree,
2. *de-interleave* the bits of the code (starting with the first being given to the y-axis, assign bits to the y and x axes building up a binary value for each),
3. convert the resulting two binary value to decimal to give a standard decimal (x, y) coordinate.

This method means that it is very easy to calculate an arbitrary number of nodes in any direction by simply converting the code for a node to coordinates, adding or subtracting the number of positions to move in the x and y directions and then converting back to quadtree code representation by following the algorithm above in the reverse order.

Of course, this method has no knowledge of the structure of the quadtree being used and so only provides the code of the node that would occupy the space at the given coordinate. That node might not exist—the tree might not extend deep enough, so the node in that position is larger than expected; or the tree might be deeper at that location meaning the node is smaller. In these cases, there are a number of options as to how to find the correct node for the code:

- The code can be shortened and/or lengthened and the resulting adapted code checked to see if it is in the tree. This trial and error method can be fastest when there is a limit on the depth range allowed when searching (see Section 7.1).

- The tree can be traversed to find the nearest node to the expected code reference.

5.2.2 Hilbert Order

One of the reasons the Z-order above becomes difficult to work with is that the resulting path from traversing the nodes in-order has to make large jumps and so cells which are numbered next to each other may, in fact, not be near each other in the image.

A number of routes exist that avoid this jumping around the image. These are based on space filling curves which have the property of being a simple recursive pattern that visits every point in a 2D space exactly once. These curves were first discovered in the early 1900's and described mathematically by D. Hilbert [Hilbert, 1970]. One of the curves that Hilbert found, the Hilbert Curve, is particularly useful since it can be represented in the simplest level in a two by two square which is then recursively repeated for each quadrant of that first square—exactly as the quadtree does.

The path that the traversal of points follows becomes fairly complicated, Figure 4b. This means, again, that the calculation of neighbours becomes difficult.

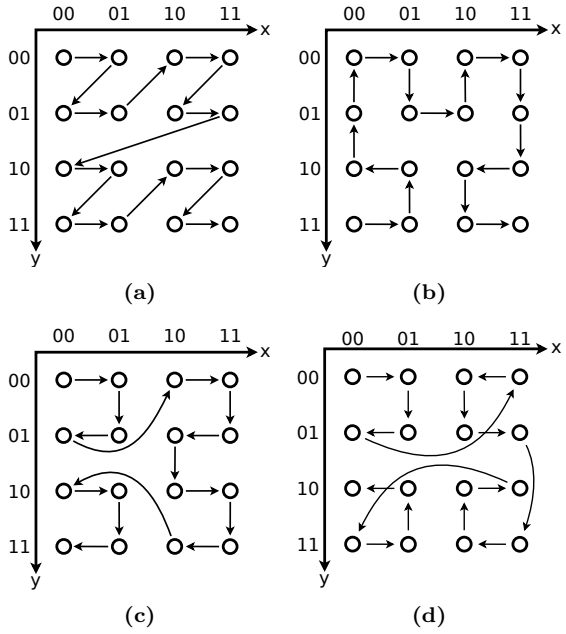


Figure 4: *A selection of possible tree traversal orderings. Each of theses is, more formally, a space filling curve which is a direct, unique mapping from a 2D grid to a 1D array. (a) shows Z-order, (b) shows Hilbert order, (c) shows Gray code order and (d) shows the modified Gray code order developed for this project.*

5.2.3 Gray Codes

The Gray Code [Gray, 1953], developed by Frank Gray in 1953, was originally designed to reduce the error rate produced by mechanical electronics. The code is a variation on binary where each step when counting up changes only a single bit at a time. This meant that electromechanical apparatus was less likely to make a mistake or generate errors since the actions required to count from one to two required only a single bit change, rather than two, as would be required for binary counting. When using just two bits, i.e., counting from zero to three, the steps are very similar to binary, (00, 01, 11, 10).

The path that this follows is shown in Figure 4c. This does not seem to provide any benefits since there is now more jumping around the image space than with Z-order and the neighbours are just as difficult to calculate as for Hilbert Order. However, by using a different arrangement of the sub-trees, as the Hilbert curve does, the leaf nodes group themselves in a very ordered fashion. When arranged as in Figure 4d and 5, each cell is arranged such that

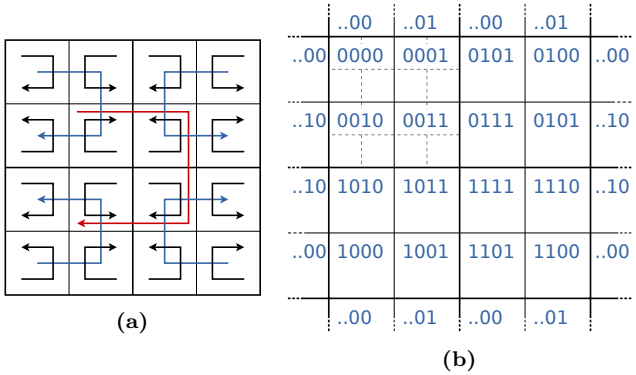


Figure 5: *The tree traversal that was developed for this project is a variation of the Gray code order discussed in Section 5.2.3. Instead of all sub-quadrants having the same orientation, each is reflected in either the x- or y-axis, (a). This has the advantage that neighbouring nodes have codes which differ by exactly one bit, (b).*

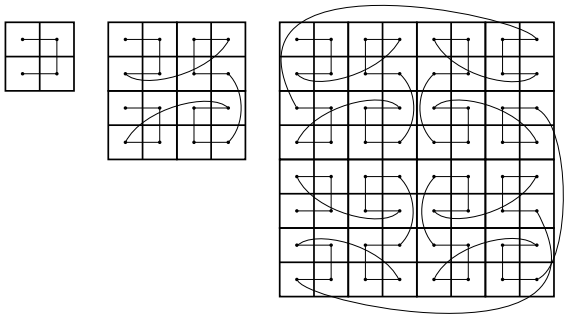


Figure 6: *modgray-Steps*

Since the quadtree is a recursive data structure, it is necessary to be able to maintain the correct orientation of child trees with respect to their parents at the construction stage. It turns out that this is easy to achieve and thus adapting it for each of the arrangements discussed above is simply a matter of adjusting the next part of the code that is added for each of the four children when creating them. Pseudo code to achieve this for the Morton order and the modified Gray code order is shown in Listing 1.

```
// Constructor
Quadtree(max_x, max_y, code)

// Child creation, Z-Order
t_l = new Quadtree(100, 100, this.code + "00");
t_r = new Quadtree(100, 100, this.code + "01");
b_l = new Quadtree(100, 100, this.code + "11");
b_r = new Quadtree(100, 100, this.code + "10");

// Child creation, Gray Code
```

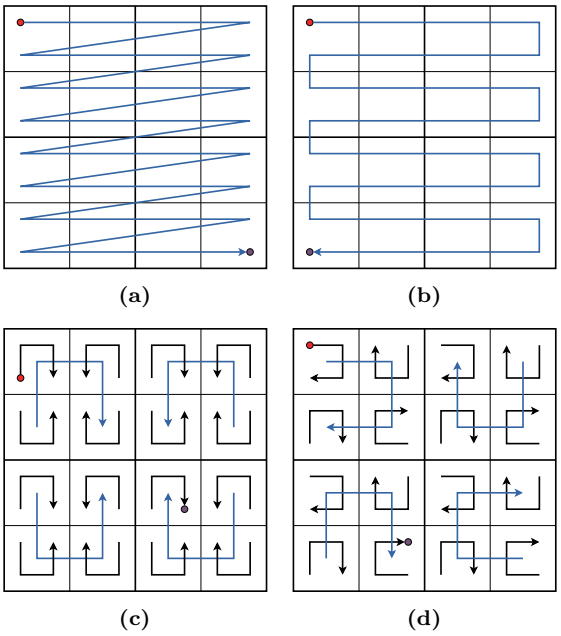


Figure 7: *modgray-2-Alternatives*

```
t_l = new Quadtree(100, 100, this.code + "00");
t_r = new Quadtree(100, 100, this.code + "01");
b_l = new Quadtree(100, 100, this.code + "10");
b_r = new Quadtree(100, 100, this.code + "11");

// Child creation, Modified Gray Code
if (pos == "tl") ) {
   new_code[0] = code + "00";
   new_code[1] = code + "01";
   new_code[2] = code + "10";
   new_code[3] = code + "11";

} else if (pos == "tr")) {
   new_code[0] = code + "01";
   new_code[1] = code + "00";
   new_code[2] = code + "11";
   new_code[3] = code + "10";

} else if (pos == "bl")) {
   new_code[0] = code + "10";
   new_code[1] = code + "11";
   new_code[2] = code + "00";
   new_code[3] = code + "01";

} else if (pos == "br")){
   new_code[0] = code + "11";
   new_code[1] = code + "10";
   new_code[2] = code + "01";
   new_code[3] = code + "00";
}
t_l = new Quadtree(100, 100, this.code + new_code[0]);
t_r = new Quadtree(100, 100, this.code + new_code[1]);
b_l = new Quadtree(100, 100, this.code + new_code[2]);
b_r = new Quadtree(100, 100, this.code + new_code[3]);
```

Listing 1: *Code to generate children of the current quadtree while maintaining the correct ordering.*

5.3 Hash Table Implementation

In order to speed up the subsequent operations applied to the quadtree structure, it can be converted to a simpler, one dimensional data structure. The method discussed above, to number the cells in a logical fashion based on the code of the parent, can be viewed as a way to map the two dimensional image to a single unique binary code, and hence to a one dimensional format. Thus, the quadtree can be easily converted to an array structure by simply using the key code as the array index.

As mentioned in Part II, this has the potential to have a very poor space usage if the image is not densely populated (in which case it is likely that few clusters would be identifiable anyway). Instead, an implementation is created that makes use of the *Hash Table* data structure [Cormen et al., 2001]. This allows the data structure to increase in size dynamically as more space is needed, but also provides linear time complexity for access and modification. The quadtree code is used as the key for the hash table, and the data stored within the cell with that code is the value.

Using this data structure means that several operations are significantly sped up. For example, for the raw quadtree format, the steps required to check if a given code is present involve traversing every node in the tree, thus linear $O(n)$. However, for a hash table, the hash function is used to get a hash of the code to be checked and the appropriate location checked. If the codes match, then the code does appear in the tree, a total of two operations, constant $O(1)$.

Also, since the structure is defined recursively, it is not possible to directly access a given location of the data. Instead it must be arrived at by starting at the root, and, for every level, deciding which of the children the destination exists in. This would be a time consuming step for a number of operations, but the biggest effect would be when checking the neighbours of cells since for every neighbour of every node, the tree must be traversed. This is not an issue for the hash table since the single dimensional nature means that data can be accessed directly anywhere.

Little spatial information should be lost during the conversion from quadtree to hash table since this is all contained in the quadtree code assigned to the node during the quadtree generation step.

Part III

Cluster Analysis

Cluster analysis is the grouping of a set of objects or items in a spatially or informationally logical way such that the items that are placed in the same group are more similar to each other than they are to the objects in the other groups in the set. These groups shall be called *clusters*. When dealing with images, the clustering that is of interest is based on spatial location, i.e., clusters should be composed of objects that are close together in the image and clusters should be separated by regions of emptiness or background level noise.

One of the primary reasons for chosing the quadtree method over the simple grid methods was that the simple act of placing the objects, in this case coordinates of data points, into the quadtree starts the process of analysing the data. Since the points end up in a tree structure with the number of points closely separated being on the lowest levels of the tree, the data is already clustered in a way.

There are a number of alternative methods of identifying clusters in images.

6 Rolling Ball Analysis

The accessible surface area (ASA) algorithm, also known as the "Rolling Ball Method", is a technique used in image processing for describing the outer limit of a cluster of points. It is derived from biological molecules analysis where it describes the surface area of a molecule that is accessible to a solvent.

The rolling ball method can be used to analyse a cluster of points by imagining a solid ball that sits against one of the outer-most points. From here it is "rolled" around the cluster such that it is always touching at least one point. Once the ball has reached the point it started at, the line that the ball traced is reduced in size by the radius of the ball. This line then represents the outer limit of the cluster.

The size of the ball must be chosen depending on the average separation of the points within the cluster.

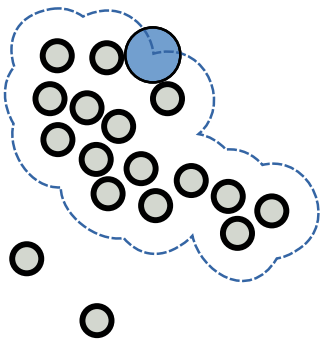


Figure 8: *The rolling ball method for cluster detection provides a way of identifying clusters, as inspired by molecular biology. A very simple implementation can be fast but is not particularly successful at finding clusters unless the data points are very dense and there is no noise.*

A simple approximation of this technique can be acheived by using the same `open`-ing and `close`-ing processes as are described in Section 4.

7 Quadtree Traversal

Whether the quadtree is stored in memory as a recursive quadtree data structure, or as hash table, Section 5.3, the most important and computationally intensive step is extracting the clusters at the correct depth and disregarding those data points that can be attributed to noise.

The quadtree numbering system chosen lends itself very well to analysis based on spatial location and the proximity of neighbours to a given node being examined.

7.1 Algorithm Description

The propagation algorithm is performed as follows:

1. Choose a starting location that has not yet been checked.
2. For this given starting location, the n neighbours are checked for validity based on the conditions of the analysis (the way these neighbours are selected is discussed in Section 7.3).
3. If a node fails the check, then it is recorded as having done so and shall not be checked again. If it passes, then it is itself propagated.
    - To be "propagated" means the perform this propagation algorithm using that node as the starting location.
4. When all nodes have been checked, return to step 1.
5. The algorithm completes either when either;
    - every one of the nodes in the image have been checked and included or ignored, or
    - the cluster has ended, so all the neighbours of all checked nodes have failed the validity tests and no further starting locations were found or specified.

This process is shown graphically in Figure 9. Here, and from now on, blue nodes are nodes that either need to be checked or are currently being checked; red are nodes that have been checked and have been included in the cluster; question marks (?) represent the neighbours of the blue nodes that will have to be checked to determine if it is included or not and crosses (x) represent neighbours that have been checked and excluded.

When discovering clusters via this propagation technique, care must be taken to avoid a run-away situation where every node in the tree gets included. This would happen when looking at the neighbours of a node and blindly including them. Since every internal node has exactly four neighbours, the propagation would terminate only when reaching the edge nodes.

Instead, the depth of the node must be considered. Again, the simplest method is not sufficient. If the propagation is limited to a given node depth, even if this is not the same as the deepest node, the size of any clusters that are identified will be limited, as shown in Figure 10a. Since the depth to consider is not able to change, when the neighbours of the blue node are checked, no correct neighbours are found and so

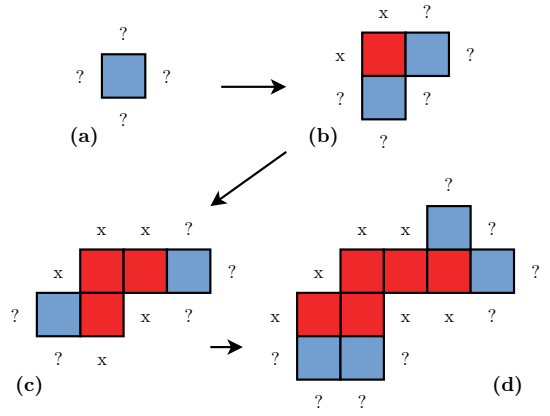


Figure 9: *For a starting location, a), in an image, no information is known and so the neighbours are checked. Some of these are found to be part of the cluster, others are rejected. The ones that are included are then, themselves, checked and so on. As the cluster grows, b), c) and d), the number of checked nodes increases.*

the process terminates. When able to view the larger structure of the nodes, however, it is clear that the structure continues beyond the gap.

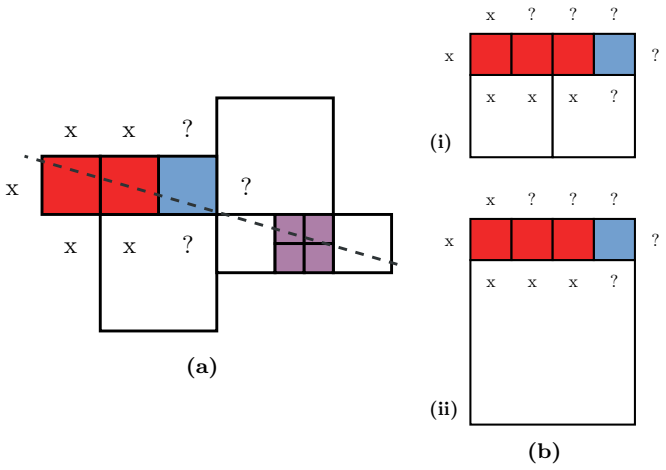


Figure 10: *Considerations regarding quadtree levels that will be accepted when propagating a node in a quadtree. (a) If the depth range that specifies how far up the tree to look for valid neighbours is too small then a cluster might be terminated too soon. (b)(i) a depth range of two and, (ii) a depth range of three.*

To avoid this, a certain amount of leniency must be given when deciding what constitutes a neighbour. Given an appropriate value, this would allow both of the larger white cells in Figure 10a to be included.

The term *depth range* shall define the levels that are to be considered when choosing neighbours with respect to a target depth. Since clusters are being considered as areas of increased density of points, all cells with a depth greater than the target depth shall be allowed, so the purple cells in Figure 10a would be included when the target depth is the same as the depth of the included red cells. A depth range of zero is equivalent to the situation above where only cells of a given depth are considered. A depth range of 1 would mean that the white square in Figure 10b (i) would be included but not in (ii), whereas a depth range of three would include both and so on.

7.2 Clustering Start Locations

In order for the algorithm to proceed correctly, a good initial node, a starting location, must be chosen. Since the clusters to be found are regions of high point density, it makes sense to start the clustering algorithm at the point in the image with the highest point density. This should ensure that the most defined cluster is always found with subsequent clusters being less dense, and so less well defined.

When starting at the highest density node, i.e., the node which is at the deepest level in the tree, propagating this node and then terminating; the cluster shown in Figure 11 is found. The file `palm-1.txt` was used and generated this data in 457 ms. This shows that the algorithm works correctly. Altering the parameters that are used to generate the quadtree affects the size of the nodes that are included in the cluster and the depth that is searched.

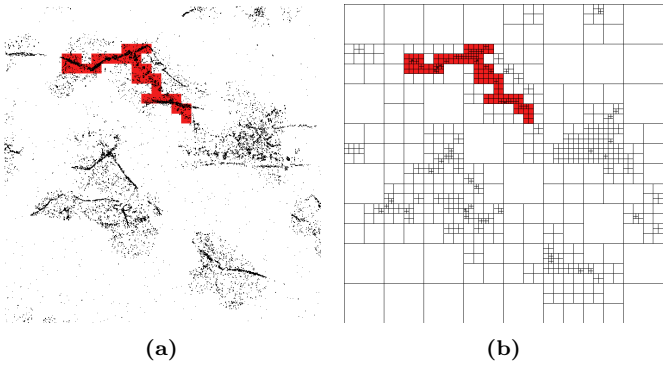


Figure 11: *single-cluster*

In order to find other clusters, the algorithm must be restarted with a new starting location. This is chosen as the deepest node in the tree that is not already included in a cluster. There are a number of different way to terminate this process of finding new clusters:

- Perform a set number of iterations. This performs well if the clusters to be located can be easily counted, but in the general case, this is not possible or desired. If the number of iterations is set to a high value, of the order of 50, the algorithm will continue to locate "clusters" even if they do not exist and will eventually simply report the background noise as a cluster. To prevent this, a limit can be set on the depth a starting location must be in order to be valid.

- Continue iterating until a depth limit is reached. This is a more general form of the previous case, but for this case, the limit on the depth of a starting location will always be reached.

- Allow the user to make a number of starting point location selections manually. Since ImageJ allows multi-point region of interest (ROI) selections, the user could be asked to place a new ROI at the places they consider a reasonable place to find a cluster. The algorithm would then convert the locations of each of these ROI points into the relevant quadtree code and begin propagating from that node and terminate when all of the ROI's had been used.

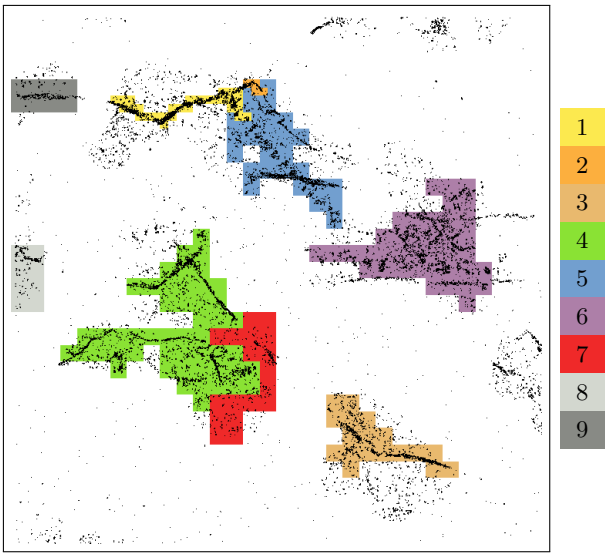


Figure 12: *Initial versions of the clustering algorithm included too many nodes, making the clusters too large. This image shows how each node cluster that is started is independent of the previous ones. The clusters were identified in the order they are listed on the right.*

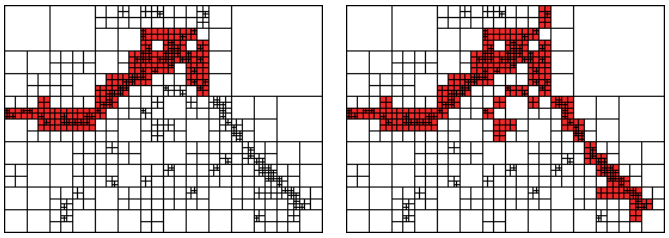
7.3 Choosing Neighbours

Some care must be taken when deciding what constitutes a neighbour of a node and what does not. As mentioned above, when detecting clusters using propagation, the neighbours of a node are checked and, if they are valid, are themselves propagated. For this reason, a poor choice of neighbours means that either the propagation will

- be cut short too early, and so not all of the clusters will be located, or
- include too many nodes, in which case the clusters will not represent the actual data.

The first neighbours that must be considered, named *rook's case* neighbours by [Abel and Mark, 1990], are the four nodes that lie to the north, east, sound and west of the current node. These are the nodes that are directly in contact with the node and so, if they are valid, represent a direct continuation of the cluster.

However, if the choice of neighbours is limited to these four, some major structures are missed. Figure 13a shows how this arrangement misses a large portion of the cluster, simply because the propagation could not consider the nodes across the boundary. If, in addition to the rook's case, the four *diagonal* neighbours are included, giving a total of eight, the results are much more complete, as shown in Figure 13b.



(a) **(b)**

Figure 13: *A comparison between different neighbour sets. Figure (a) shows how some of the cluster is lost when using just the rook's case neighbours, whereas, using all eight neighbours, Figure (b) more of the cluster is included.*

Part IV
ImageJ Plugin

ImageJ is an public domain, Java based image manipulation program written and maintained by developers at the National Institute of Health. It is widly used in medical and biological research and has an open API to allow extension via macros, plugins and scripts.

Since they offer significantly betting integration, meaning speed and efficiency improvements, a plugin is chosen to inegrate this project into ImageJ over macros, which suffer performace loss when more than a few steps are involved, and scripts which do not offer such tight integration with the rest of the program.

The plugin shall allow a user to load a data set, as gathered from STORM, PALM or similar imaging techniques discussed in Section 2, choose the format of the data (the coloumns that are of interest etc.) analyse the data for clusters and receive detailed information regarding the clusters that were found.

8 ImageJ Plugin

[NIH, 2014]

8.1 Displaying the QuadTree

The first version of the plugin simply allowed the user to visualise the data, once it had been processed and entered into a quadtree. Though this provided little benefit to the researcher producing data, it can be a useful tool to get an insight into the process that a program is using to analyse one's data so that the results can be better interpreted. For this reason, this version served as a foundation for later versions of the plugin so that, when loading data, a user gets the opportunity to view the data before proceeding with the analysis.

Again, earlier versions of the program displayed this data using the built-in GUI classes in the AWT [Zukowski, 1997] and Swing [Loy et al., 2002] libraries included in the standard Java distribution. This effectively prohibited any further actions being performed on the image once it had been generated since what was displayed was only modifiable by the JVM via compiled code. It was also very memory intensive since, in many cases, many thousands of separate objects (data points represented by zero length lines, quadtree cells by boxes, etc.) and so was slow to draw initially and redraw with any subsequent move or resize of the window.

The code used to generate this view of the data was modified to make use of the easy image generation functions present in ImageJ. Now, instead of many different objects being manipulated for each view of the data, a single array with a value for each pixel is needed. For the cases where the user wishes to view the clusters that have been found, but not have them affect the image, the image is created with a number of different *slices* in the image *stack*, Figure 14. Slices are ImageJ's representation of images with two or more layers, or alternative views, each of which resides in a stack of slices. Each slice in a stack must have the same dimensions.

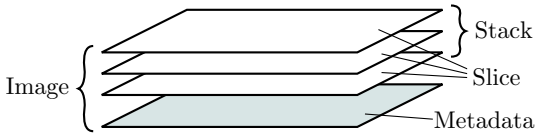


Figure 14: *An "image" in ImageJ can be composed of a number of layers, each of which is easily viewable and can be saved on its own. These layers are used to display different parts of the generated image: data points, located clusters, quadtree structure, etc.*

8.2 Column Picker

8.3 Results Table

In addition to the image of the data with the clusters, the plugin also calculates a number of statistics for each of the clusters that are found. These include the number of points that are included in the cluster, the area of the cluster, and it's perimeter. The area and perimeter values are given as a fraction where the whole image represents 1. Thus, to find the actual area for the real life object, the area of the cluster should be multiplied by the area of the image, and for the perimeter, the length of one side of the image should be multiplied by the perimeter.

To calculate each of these values, each node that exists in a cluster is examined and its contribution to the overall value added.

8.3.1 Cluster Area

To calculate the area of the clusters, each node in each cluster is examined. Since the size of the node must be known, it is calculated from the quadtree code for that node. The formula is shown in Equation 4,

$$a = \frac{1}{4^d} \tag{3}$$

$$a_i = 4^{-l_i/2}, \tag{4}$$

where a_i is the area of the node i, d is the depth in the quadtree and l_i is the length of the quadtree code of that node. For every node in the cluster, where there are n nodes, this value is summed to give the total cluster area, A:

$$A = \sum_{i=0}^{n} a_i. \tag{5}$$

8.3.2 Cluster Perimeter

Similarly to the cluster area, the perimeter is given as a fractional value of the length of one side of the whole image, as calculated for each node from the quadtree code, as shown in Equation 7,

$$p = \frac{1}{2^d} \tag{6}$$

$$p_i = 2^{-l_i/2}, \tag{7}$$

where the symbols have the same meaning as above.

However, this simply gives the length of one side of the node for any node. In order to calculate the perimeter of the cluster, it is not enough to simply sum these values, as for the cluster

area, since not all nodes contribute to the perimeter. Instead, for each node, it must be decided whether it contributes to the perimeter and how much (1, 2, 3 or 4 sides), and then increase the total perimeter by this many times the length of one side. The total perimeter, P, then is given by Equation 8,

$$P = \sum_{i=0}^{n} s * p_i, \tag{8}$$

where $s \in \{0..4\}$. This is demonstrated in Figure 15 where the perimeter is simple to calculate in case a) as the nodes are all the same, but the size of each node must be taken into account in case b).

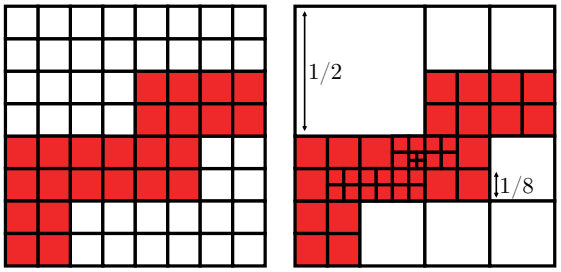


Figure 15: *When calculating the perimeter of a cluster using the nodes that contribute, the size of each node must be taken into account. In case (a), the process is simple since all nodes are the same size and a fractional value of 3.5 is calculated. For case (b), the steps are 25 lengths of size $^1/_8$ and 6 lengths of size $^1/_{16}$ which gives the same result, 3.5.*

Within the cluster, *holes* occur where a node, or number of nodes, is surrounded on all sides by the same cluster. An example can be seen in Figure 13. Unfortunately, these are included in the calculation of the perimeter and so, where holes exist in a cluster, the actual perimeter is slightly smaller than that calculated.

Part V
Further Considerations

Some further considerations—what went well, what could have been better, improvements, next steps etc.

9 Possible Improvements

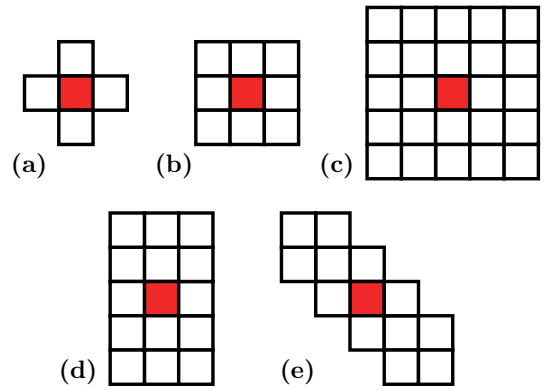


Figure 16: *kernel-Shapes*

A Data File Structure

The data files that are produced from the initial analysis of the images have a standard format.

1. Tab separated fields.
2. Single header line with names of fields.
3. One or more item of data, separated by newlines.

The columns that represent fields in the file are as follows.

Header	Meaning	Used?
`Channel Name`	Wavelength channel that was used to capture data. First value, I, is the incident wavelength of the light used to excite the dye and the second, E, is the wavelength emitted that was imaged.	no
`X`	x-coordinate of the point	no
`Y`	y-coordinate of the point	no
`Xc`	centered, normalised x-coordinate of point	yes
`Yc`	centered, normalised y-coordinate of point	yes
`Height`	the height of the fitted gaussian peak used to extract the point from the original image	not yet
`Area`	area of the point	not yet
`Width`	full width half maximum of the point	not yet
`Phi`	?	no
`Ax`	?	no
`BG`	?	no
`I`	?	no
`Frame`	?	no
`Length`	?	no
`Valid`	?	no
`Z`	?	no
`Zc`	?	no
`Photons`	?	no
`Lateral Localisation Accuracy`	?	no
`Xw`	?	no
`Yw`	?	no
`Xwc`	?	no
`Ywc`	?	no

B Class Diagram

Figure 17 is a class diagram showing the structure of the ImageJ plugin that was created for this project.

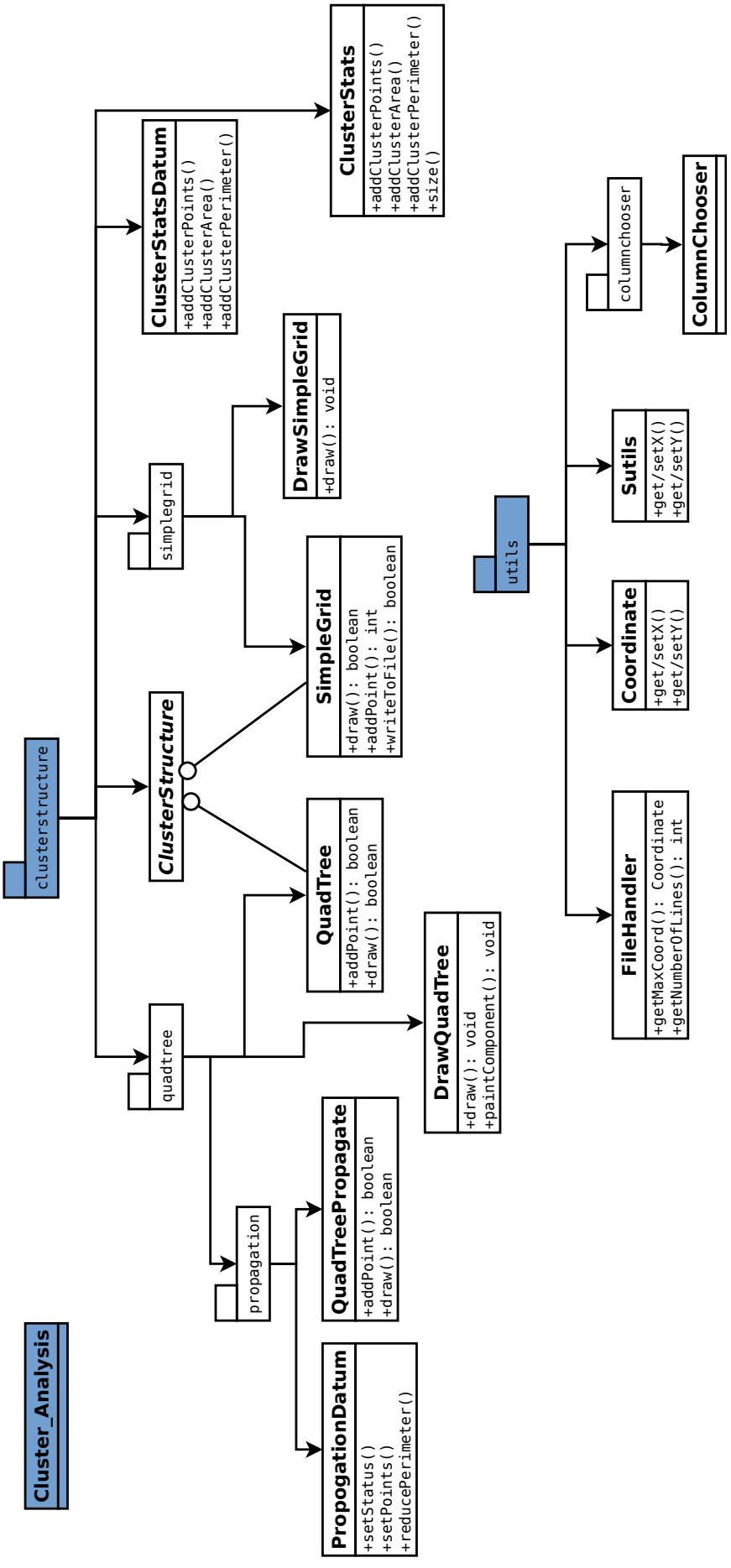


Figure 17: *Class diagram for the Cluster Analysis ImageJ plugin written for this project.*

References

[Abel and Mark, 1990] Abel, D. J. and Mark, D. M. (1990). A comparative analysis of some two-dimensional orderings. *International Journal of Geographical Information System*, 4(1):21–31.

[Asano et al., 1997] Asano, T., Ranjan, D., Roos, T., Welzl, E., and Widmayer, P. (1997). Space-filling curves and their use in the design of geometric data structures. *Theoretical Computer Science*, 181(1):3–15.

[Bailer, 2006] Bailer, W. (2006). Writing imagej plugins—a tutorial. *Upper Austria University of Applied Sciences, Austria*.

[Cormen et al., 2001] Cormen, T. H., Leiserson, C. E., Rivest, R. L., Stein, C., et al. (2001). *Introduction to algorithms*, volume 2. MIT press Cambridge.

[Fang et al., 2005] Fang, N., Lee, H., Sun, C., and Zhang, X. (2005). Sub–diffraction-limited optical imaging with a silver superlens. *Science*, 308(5721):534–537.

[Gray, 1953] Gray, F. (1953). Pulse code communication. US Patent 2,632,058.

[Hess et al., 2006] Hess, S. T., Girirajan, T. P., and Mason, M. D. (2006). Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. *Biophysical journal*, 91(11):4258–4272.

[Hilbert, 1970] Hilbert, D. (1970). Über die stetige abbildung einer linie auf ein flächenstück. In *Gesammelte Abhandlungen*, pages 1–2. Springer.

[Hjaltason and Samet, 2002] Hjaltason, G. R. and Samet, H. (2002). Speeding up construction of pmr quadtree-based spatial indexes. *The VLDB Journal—The International Journal on Very Large Data Bases*, 11(2):109–137.

[Loy et al., 2002] Loy, M., Eckstein, R., Wood, D., Elliott, J., and Cole, B. (2002). *Java swing*. " O'Reilly Media, Inc.".

[Mokbel et al., 2002] Mokbel, M. F., Aref, W. G., and Kamel, I. (2002). Performance of multi-dimensional space-filling curves. In *Proceedings of the 10th ACM international symposium on Advances in geographic information systems*, pages 149–154. ACM.

[Morton, 1966] Morton, G. (1966). A computer oriented geodetic data base and a new technique in file sequencing. *IBM, Ottawa, Canada*.

[NIH, 2014] NIH (2014). *ImageJ API Documentation*. US National Institutes of Health, 1.48t edition.

[Owen et al., 2010] Owen, D. M., Rentero, C., Rossy, J., Magenau, A., Williamson, D., Rodriguez, M., and Gaus, K. (2010). PALM imaging and cluster analysis of protein heterogeneity at the cell surface. *Journal of biophotonics*, 3(7):446–454.

[Rasband, 1997] Rasband, W. (1997). ImageJ, US National Institutes of Health. *Bethesda, Maryland, USA*, 2012.

[Rust et al., 2006] Rust, M. J., Bates, M., and Zhuang, X. (2006). Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature methods*, 3(10):793–796.

[Samet, 1990] Samet, H. (1990). Applications of spatial data structures. *Computer Graphics, Image Processing, and GIS*.

[van Oosterom, 1999] van Oosterom, P. (1999). Spatial access methods. *Geographical information systems*, 1:385–400.

[Williamson et al., 2011] Williamson, D. J., Owen, D. M., Rossy, J., Magenau, A., Wehrmann, M., Gooding, J. J., and Gaus, K. (2011). Pre-existing clusters of the adaptor lat do not participate in early t cell signaling events. *Nature immunology*, 12(7):655–662.

[Zukowski, 1997] Zukowski, J. (1997). *Java AWT reference*, volume 3. O'Reilly.