Part I

Introduction

1 Medical Imaging

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 damaging the sample. When imaging biological samples, this can be unreasonable.

2.1 Image Manipulation

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 the STORM method [Rust et al., 2006] uses a technique where the objects to be imaged are molecules of a fluorescent dye. 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.

Even using exotic types of lenses to reduce or remove the problem of diffraction [Fang et al., 2005].

3 Benchmarking

Throughout the project, a set of files wil 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 four files that will be used are detailed in Table 3.

File Name	Size	No. of 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

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

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 greyscale birghtness value. This gives a simple pixel image with brightness as a function of density in the pnm image format. A thresholding filter can then be applied to remove the points that are isolated and leave the denser areas coresponding 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 1, where palm-1.txt was used with a cell size of 200.

(a)   (b)
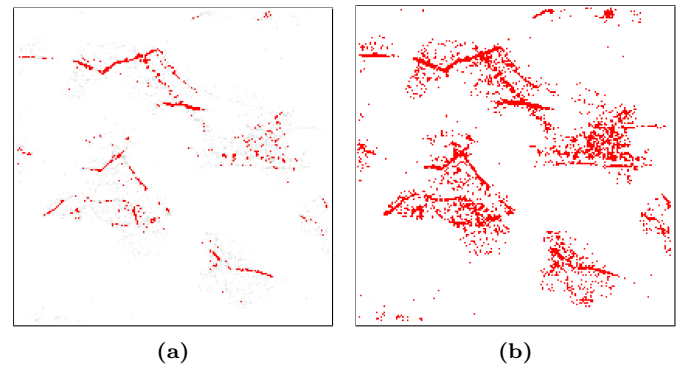
Figure 1: *Setting a low threshold, (a), means that many of the points in the clusters are lost. Setting it higher, (b), includes too many of the points deemed to be noise.*

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 children. 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 based on some quality of the image. From this point, information about the image can be "seen" more easily by the computer and statistics calculated.

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 in 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.

5.1 Morton Code (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 2a. 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.
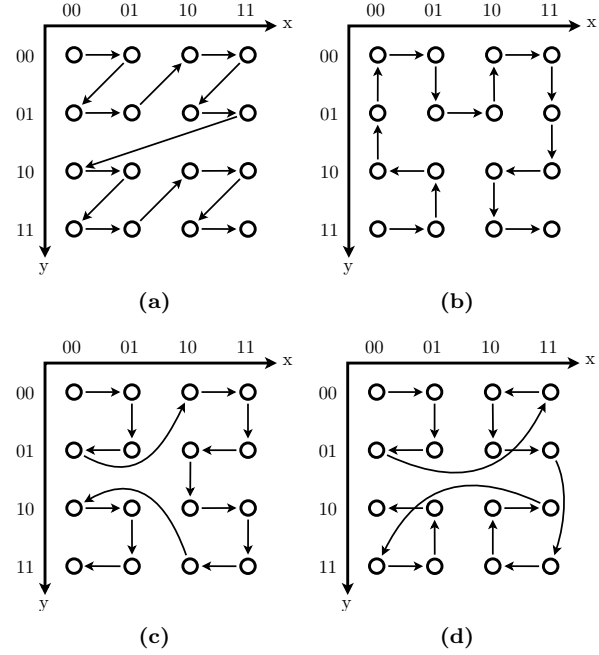

Figure 2

5.2 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 2b. This means, again, that the calculation of neighbours becomes difficult.

5.3 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 apperatus 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. couting from zero to three, the steps are very similar to binary.

0   00

1   01

2   11

3   01

The path that this follows is shown in Figure 2c. This does not seem to add 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 arangement of the sub-trees, as the Hilbert curve does, the leaf nodes group themselves in a way in a very ordered fasion. When arranged as in Figures 2d and **??**
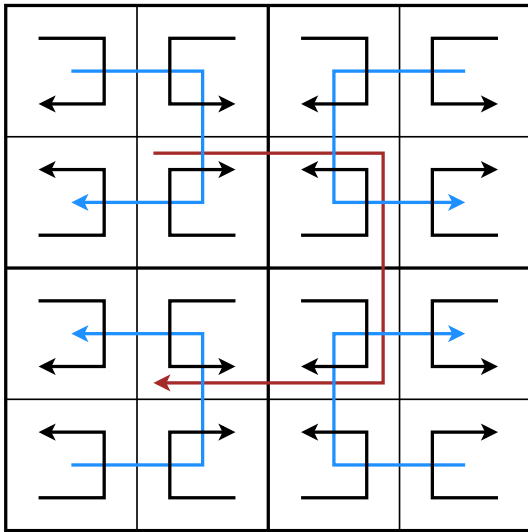
Figure 3: *modgray-2-Levels*

References

[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.

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

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

[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.

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 fol-
lows.

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	?	no
Localisation	?	no
Accuracy	?	no
Xw	?	no
Yw	?	no
Xwc	?	no
Ywc	?	no