AUTOMATIC GRAPH INTERPRTATION

Christopher Williamson
Physics and Biophysics Department
Massey University

Donald Bailey
Image Analysis Unit
Massey University

ABSTRACT

This project seeks to extract numerical data from a two dimensional graph. The motivation is to obtain soil density information from well logs, where the original numerical data is unavailable. The first step is to locate the grid lines. This may be achieved by a Hough transform of the image containing the graph. In the Hough transform, the grid lines appear as a regular pattern of peaks, with each peak corresponding to a grid line. The parameters of each grid line are obtained, enabling the grid lines to be removed from the image. The image now contains only the graph line, which may be scanned to extract the data values. The resultant data is processed to eliminate noise and to interpolate between breaks in the graph. Finally the coordinates are transformed from pixel units to the graph units, using the grid information extracted previously. This procedure also corrects for axis rotation in the captured image frame. With optimisation this method allows for fast and simple extraction of data, usable under a large range of applications.

INTRODUCTION

Graphical representation of soil density information is readily obtainable in the form of well logs but expensive in numerical form. Thus, there is a need for the reconstruction of the numerical data form, but with the considerable volumes of data involved, this is a time consuming process if performed manually. To aid the reconstruction process, image processing techniques may be applied to provide a timely and automated approach to the problem. By use of a piece-wise approach, varying formats of graphical input are able to be reconstructed, and the algorithm can be refined for specific applications, providing increased speed performance. In this report five main elements comprise the process of reconstruction. These include, initial preprocessing to simplify and enhance the image, grid detection and parameterisation, interpretation of detected grid lines, removal of grid, and data extraction.

PREPROCESSING REQUIREMENTS

As a captured image of the graph, contains a range of grey scale values, it is desirable to binarise the image into information points and redundant background data. It is first noted that the histogram of the input image consists of two distinct peaks, corresponding to the background and data regions. Thus a value in the lower part of the valley between these will serve as a threshold level. The mean intensity value of the input image is considered as a first approximation to the threshold. Mean values are then calculated of the intensities above and below this threshold. The midway point between these two means may then be taken as an improved value of the threshold. To refine the threshold value, this process maybe repeated until the change is less than some criteria.

To simplify the detection of grid and data values, the image is thinned down to its skeleton. This also has the advantage of reducing most of the noise to small isolated spots, which with a graph consisting of long joined stretches, may be removed by consideration of their lengths. Figures 1 and 2 illustrate part of a typical input image and the result after preprocessing.

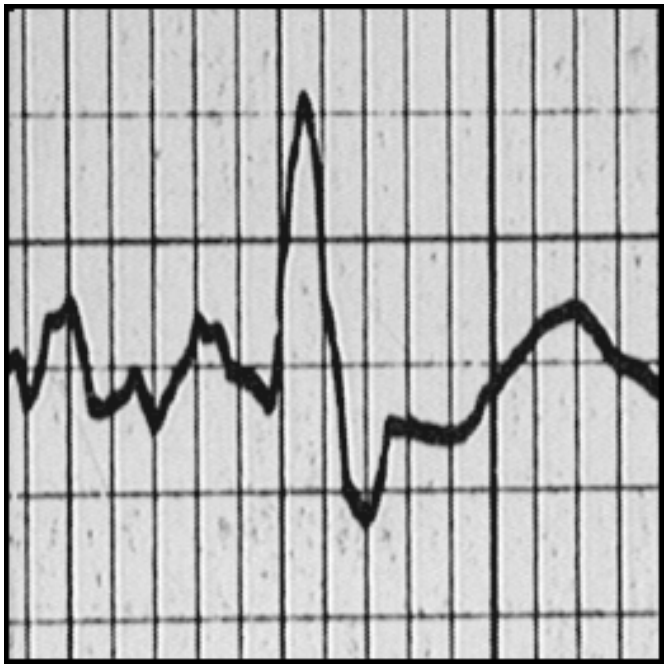

Figure 1: A section of a typical graph to be interpreted.

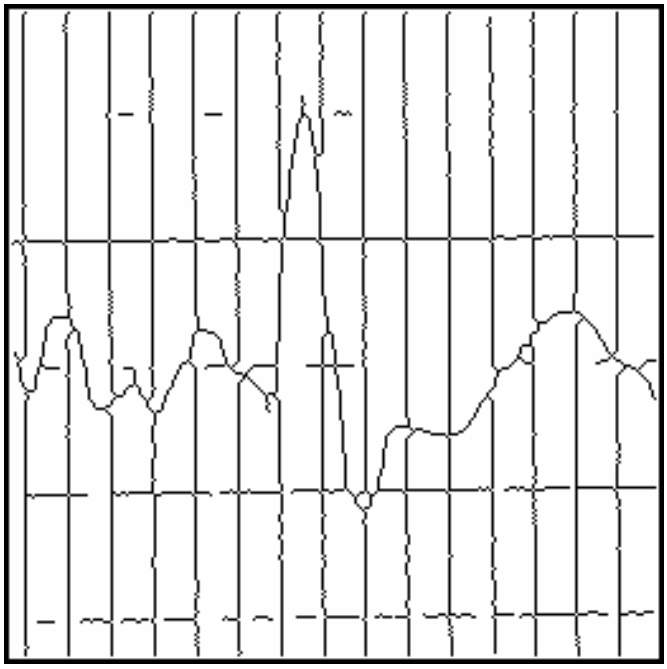

Figure 2: Resultant image after binarisation and thinning.

GRID DETECTION

The image after preprocessing contains the graph data superimposed on approximately straight grid lines. As some of the grid lines appear very faint in the original image, and large sections of the grid may be missing, a search and delete approach is impractical. For this

reason, we have chosen a technique of parameterising distinct grid lines and later interpreting for those missed.

To parameterise the grid lines, a well known technique, the Hough transform [1] is available. This method has the advantage of being insensitive to noise, and is unaffected by the graph line (unless it itself is very straight). A possible set of parameters for describing an individual grid line are ρ and θ, from the equation

$$\rho = x\cos\theta + y\sin\theta$$

where θ is the angle between the x axis and the line, and ρ the perpendicular distance from the line to the origin. For each point detected in the input image, the parameters for every possible line through that point are determined. Plotting these give a sinusoid in the parameter space. Parameters from successive input points increment those points covered in parameter space. Where a number of input pixels are collinear, peaks or maxima occur as the common parameters receive multiple votes. Figure 3a shows the output after performing the Hough transform on the preprocessed image, with θ varying from 0 to 360 degrees. The two sets of vertical maxima seen in this image corresponding to the distinct vertical and horizontal lines of the grid. The blurring about the 3rd maxima down of the left hand set of peaks, results from the flat regions of the graph giving a low level of correlation. A possibility of peak splitting may occur within the image, due to thinning processes reducing a grid line to a combination of two neighbouring lines. This is the result of the unevenness of the original grid lines. A blurring of the parameter image by box averaging, ensures split peaks are merged without loss of generality.

The parameters of each grid line may be determined by locating the peaks in parameter space. This peak detection process [2] labels each pixel in parameter space according to its associated peak, which is defined by the lack of a valley between the pixel and the peak. To reduce the effects of noise, adjacent peaks are merged if the valley between them is less than a specified tolerance. As can be seen in figure 3a, additional smaller secondary peaks exist beside the main peaks. These peaks contain no useful information and may be eliminated by considering only narrow slices about the two vertical columns of strong peaks. The width of the slices can be determined from the location of the secondary peaks. As the graph has limited resolution, a weighted average of the parameter values in the area about each peak is used to improve the accuracy of the resultant parameters. Figures 3b and 3c show the peak detection for the horizontal and vertical grid line parameters respectively.

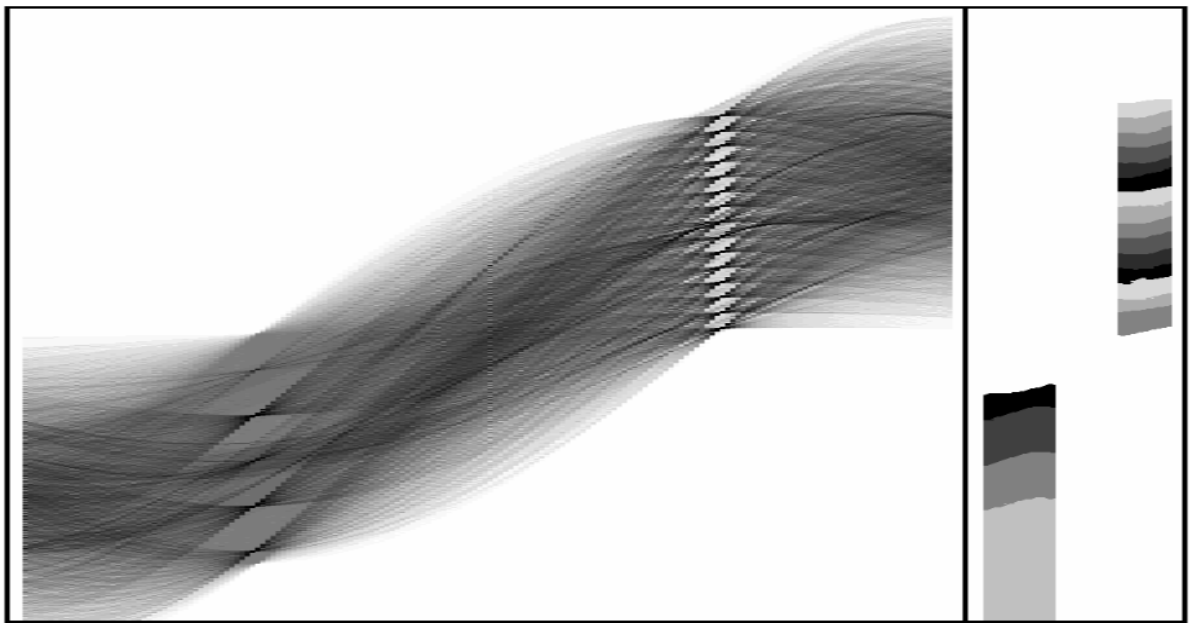

Figure 3: a) (left) Image of parameter space obtained from Hough transform, ρ vs θ
b) (centre) and c) (right) Peaks detected for the horizontal and vertical grid lines, respectively

GRID INTERPRETATION

A number of fainter grid lines may not be detected, and extra erroneous lines from the graph itself maybe. We need to determine which of the points found correspond to grid lines and interpolate for those missing. To obtain an estimate of the average grid line spacing it is required, and hence assumed, that most of peaks found are of neighbouring grid lines.

The method is as follows:

a) A first approximation to the average grid spacing is obtained. Since it is assumed that most spacings between peaks are between consecutive grid lines, those which vary greatly from the mean spacing are omitted.

b) The starting average is taken to construct an estimate of the grid and this is visualised as being placed beside the detected peaks. Figure 4 shows this pictorially

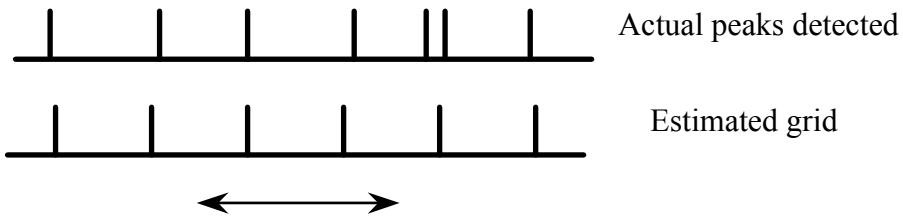

Figure 4: Estimated grid slides over actual peaks until position of least error is obtained.

c) The best fit of the estimated grid to the actual peaks is determined from the least mean displacement between corresponding points. Only points closer than half the currently estimated average are considered. Use of least mean rather than a least mean square calculation, retains directional information, enabling a systematic approach to locating the position of best fit.

d) Points not closest to their corresponding estimated points are removed. Remaining points which vary significantly from the estimate are also removed.

e) A new average spacing is calculated from all remaining data, using the old average to identify spacings of multiple average width.

f) Steps b) and c) are repeated using those data points remaining, and the new average determined.

f) Additional points are added where grid lines have not been detected, using the final average spacing.

GRID REMOVAL

By removing all points in the thinned image corresponding to the parameters detected, the grid lines may be removed. However, since the grid lines are not perfectly straight, a line of single pixel width will not remove the grid completely. To guarantee a clean removal of grid lines under all cases, a width of two pixels out from the calculated grid line needs to be removed. As this width is considerable, a simple masking technique is impractical since it removes too much of the graphed data. A more complicated search and deletion of points has to be used instead. The calculated grid line is traversed, removing at each point only that pixel which is closest to the calculated line, within a specified maximum distance from the line. Where the graph is steep and becomes merged with the grid line, this process follows a grid line to the graph, then deletes part of the graph itself, until the grid line is located on the other side of the graph.

This problem may be reduced by restricting the rate at which consecutive deleted points can vary sideways. For reasonable lengths, the actual grid line only varies back an forth between neighbouring pixels. Hence a restriction is made to only search within this pair (the pair itself must still be enclosed by the predefined deletion width for the line) until the line has remained perfectly straight for a minimum of three consecutive pixels. If the line deviates more rapidly, then the algorithm is restrained from following the line. If no line is located for a count of three positions, the search returns to the full width until another point is encountered. This approach may leave small sections of grid lines, but as these are small they are removed as noise. A classification of objects in the image by length, and removal of short sections, eliminates noise points and the small grid sections that have been left.

Missing sections of the curve, created by the grid removal process are replaced using a simple linear interpolation. Figure 5 shows the resultant graph image.

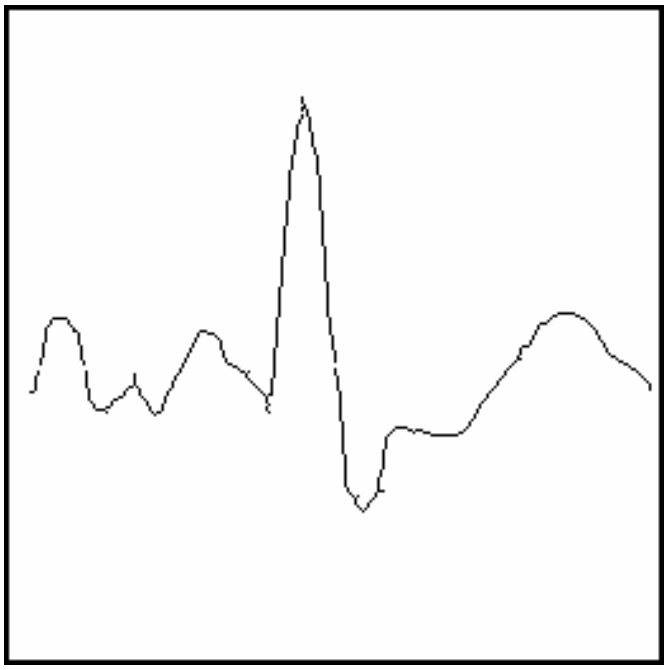

Figure 5: Resultant graph image after noise and grid removal.

A noise free graph remains in the image. This is expanded and masked with the original preprocessed image, extracting only the true curve. Those interpolated sections will also be corrected back to their true form. To obtain adequate coverage of the original curve, some grid lines also become attached. These are removed by stripping off the outer layer pixels since the grid lines are only a couple of pixels wide. To reobtain the curve shape, the outer layer is replaced.

EXTRACTION OF DATA

At the writing of this paper the data extraction step had not been completed. The method, though, is as follows.

Misalignment between the graph axes and the frame buffer is inevitable. A new set of coordinates is defined parallel to the axes of the graph as shown in figure 6. The transformation between image coordinates (x y) and the true graph coordinates (X Y) is given by

$$X = ((x - x_{origin})\cos\theta + (y - y_{origin})\sin\theta) / X_{spacing}$$
$$Y = ((y - y_{origin})\cos\theta - (x - x_{origin})\sin\theta) / Y_{spacing}$$

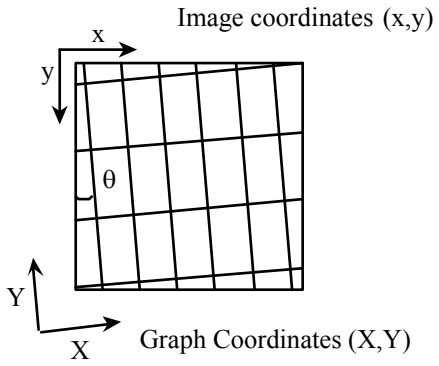

Figure 6: Correspondence between image coordinates and graph coordinates.

By scanning the thinned graph (figure 5), each point is corrected by finding the centre of the true graph line in the original image. The coordinates of this point are then transformed into graph coordinates. As the graph is traversed, this results in a list of data points. Since these data points will be irregularly spaced, they are resampled at the desired data resolution to produce an output data file.

To ensure compatibility of results between various runs, all data values are referenced to a relative origin point. The intersection of leftmost and bottommost grid lines will serve as the location of this origin. This, however, requires that the graph be positioned before capturing to within a discrepancy of less than the smallest grid spacing, to obtain consistent results. Alternatively, a marker may be positioned in a common area of the graph before capturing, but this may be unreliable if the marker becomes surrounded by noise. Additional user input of actual grid spacing will be required to correct for scaling factors between the true graph and image size, resulting from the digitisation process.

CONCLUSION

Currently the described process requires considerable time to interpret a single graph. This is largely due to the extensive approach taken in various stages to allow for a variation in initial graph quality. Considerable speed improvement is expected by code restructuring and where consistent input quality is provided, results are expected to be obtainable in reasonable time. Future extensions include the ability to interpret graphs with logarithmic scales, dashed graph lines, and multiple graph lines in the same image.

REFERENCES

[1] Gonzalez R.C. and Wintz P.,"Digital Image Processing", Addison-Wesley, Reading, Massachusetts, USA (1987)

[2] Bailey D.G., "Raster Based Region Growing", 6th New Zealand Image Processing Workshop, Lower Hutt (1991)