THE IMPACT OF SELECTION BIAS WHEN EXAMINING INDIVIDUAL CELLS FROM WHOLE-FIELD IMAGES

P. A. Appleby[1*], S. Shabir[2], J. Southgate[2] and D. Walker[1]

ABSTRACT

Imaging populations of cells is important in a wide range of experimental work. Although averaging of the whole-field image is sometimes appropriate, in many cases it is the response of individual cells that is of interest. Extracting this information is a challenging image segmentation problem where single cells must be delineated from their neighbours. As an example, we examine data from whole-field immunofluorescence images of normal human urothelial (NHU) cells in monolayer culture. These cells are loaded with ratiometric calcium-binding fluorescent dyes and stimulated by application of ATP, resulting in changes in the cytosolic calcium concentration. Typically, a region of interest (ROI) around each cell is defined by hand. These ROIs are then used to calculate time profiles of cytosolic calcium concentration for individual cells. Although straightforward, this process is time consuming and introduces a strong selection bias towards cells that experience a large elevation of cytosolic calcium. Selection bias is of particular importance to us as population-level behaviours such as wound closing are believed to be strongly dependent on the context-dependent cued response of individual cells, which gives rise to population heterogeneity. In order to eliminate this bias, we have developed an algorithm designed to automatically extract ROIs from images of urothelial cells. This algorithm is specialised to the nature of the images we process but is fast and achieves a high success rate. We compare the results of this algorithm to those generated by picking out cells by hand. We show that, far from being homogeneous, a wide range of intrinsically different cell responses are identified and these cells would be likely be missed using more conventional methods of image analysis.

[1]Department of Computer Science, Kroto Research Institute, Broad Lane, University of Sheffield, Sheffield, UK
[2]Jack Birch Unit for Molecular Carcinogenesis, Department of Biology, University of York, York, UK
[*]email: p.appleby@sheffield.ac.uk

INTRODUCTION

Analysis of cell imaging data is central to a wide range of experimental work. If the cell population is homogeneous the simplest approach to analysing this data is to examine the image as a whole, for example by calculating the average intensity of all the pixels in the image. Averaging images is fast, guarantees that all cells in the image are included, and reduces the impact of noise. The main drawback is that there is no distinction between individual cells. In other words, this technique rests on an assumption of homogeneity that permits the information derived from the cells to be averaged in a meaningful way. However, even genetically identical cell populations may display heterogeneous behaviours, for example reflecting cell cycle stage in non-synchronised populations, or relative position in the colony. A common approach to dealing with heterogeneity to define a set of regions-of-interest (ROIs) around individual cells in the image. The use of ROIs removes the background pixels and distinguishes individual cells, although at the expense of higher noise due to the smaller number of pixels captured by each ROI. The simplest way of defining a set of ROIs is to identify cells by eye and draw a ROI around each cell by hand. This method is accurate but time consuming and also introduces an element of human error in judging the boundary of a cell. A better solution is to automate the process which, whilst not eliminating these effects entirely, removes subjectivity and at the same time is faster and captures a greater proportion of cells in the image. A common approach is to identify cell boundaries based on intensity, texture or gradient features and use this information to segment the image. Examples include seeded-watershed methods (Vincent and Soille, 1991), and watershed and mean shift (Cheng, 1995; Yang et al., 2005a, 2005b). A combination of region-based detection, which uses pre-segmentation to estimate the intensity distributions present in the image and level set segmentation (Osher and Sethian, 1998) to localise individual cells, and edge-detection has been shown to work well with phase contrast microscopy (Li et al, 2008). Another popular set of techniques is based on particle filtering (Smal et al., 2006; Godinez et al., 2007; Docuet and Ristic, 2002), an approach which has been successfully applied to tracking of cells imaged using fluorescence microscopy (Smal et al., 2007).

Although progress has been made using these approaches, a method of robustly automating the extraction of ROIs remains elusive, in particular with cells that are not well separated (Bahnson et al., 2005), that move between frames, or that divide (Kirubarajan et al., 2001). There is also a great deal of variation in the image characteristics derived from different experiments. Perhaps the greatest problem with any method of defining ROIs is selection bias. When working by eye there is a natural tendency to focus on the brightest cells in the image. With an automated algorithm there is typically a bias towards cells which can easily be delineated from their neighbours and the background. This selection bias presents a problem. If the population is intrinsically heterogeneous then selection bias will lead to certain subgroups being preferentially selected over others. Entire sub-populations could even be missed if those cells respond poorly to the stimulation protocol used. A lack of rigor during the selection process could therefore lead to misleading results that overstate the importance of the sub-groups within the larger population, which could have significant consequences for the conclusions drawn from the data. Our own experiments involve imaging normal human urothelial cells grown in culture and the analysis of the wound closing behaviour that is observed in response to scratch wounding. In one class of experiments, stimulation by application of agonists such as ATP results in changes in

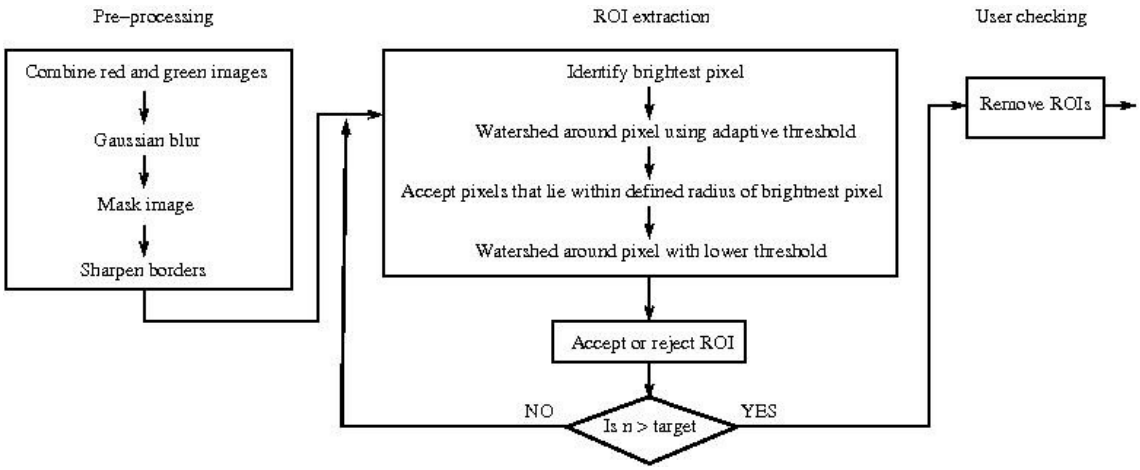

Figure 1: flow chart showing the operation of the custom algorithm we use to automate ROI extraction. n refers to the number of ROIs identified, which is compared to a target number specified by the user. Details of each step are given in the main text.

fluorescence over time that correspond to changes in the cytosolic calcium concentration. The issue of selection bias is of particular importance to this work as the subsequent wound closing behaviour is believed to depend, in part, on the intrinsic heterogeneity of the population.

EXPERIMENTAL AND COMPUTATIONAL SETUP

At the beginning of the experiment we describe here the urothelial cells are seeded onto glass coverslips and loaded with two fluorescent dyes, fluo4-AM and fura red-AM. The cells are then placed in a laminar flow perfusion chamber and imaged at one frame per second. The cells are stimulated with 100 micromolar ATP, which triggers an elevation in cytosolic calcium concentration. The two fluorescent dyes generate two channels of information, one red and one green. When the cytosolic calcium concentration rises the intensity of the green channel increases and the intensity of the red channel decreases. The time course of the intracellular calcium found by taking the ratio of these two channels. ROIs are identified either by eye or by using an algorithm that combines information from the red and green channels. For this algorithm we use the 1^{st} red frame, in which cells which initially take up a lot of dye appear very bright, and the 70^{th} green frame, in which cells that respond strongly to the application of ATP appear very bright. In both cases, the ROIs are then used as a mask to examine the time course of the ratio of the intensity of the red and green channels within each ROI in the full set of 120 ratioed images. An outline of the automated algorithm is shown in Figure 1. The algorithm and can be divided into three phases: pre-processing, ROI extraction, and user checking. During the pre-processing step the 1^{st} red and 70^{th} green images are combined and smoothed using Gaussian blurring. The image is then masked by setting all pixels below a specified threshold to zero. The mean brightness of a region around each pixel is calculated and pixels whose brightness is below a specified fraction of this mean brightness are set to zero. This creates clearer borders between cells by identifying pixels where the local intensity gradient is zero. In the ROI extraction step the pixel with the highest intensity in the image is identified. The region around this pixel is selected using a water-shed with an adaptive threshold under the constraint that each

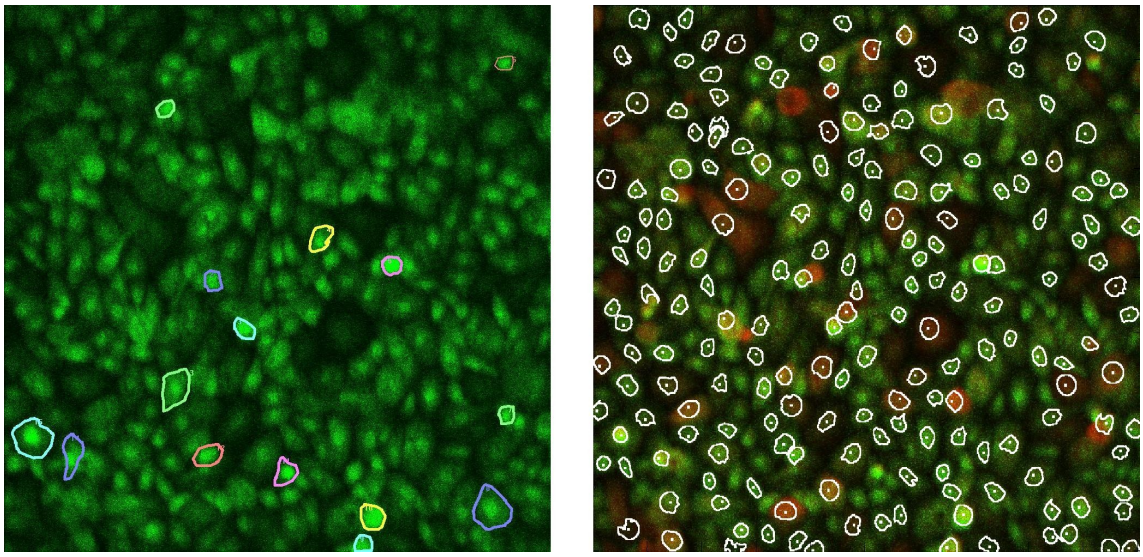

Figure 2: Comparison of ROIs identified by eye and using the automated algorithm. **(A)** ROIs identified by hand, shown overlayed onto the 70th green image. 15 cells are identified, which is typical for an experiment of this nature. **(B)** ROIs identified using the algorithm, shown overlayed onto the combined red 1st - green 70th image. 200 cells are identified covering a much greater proportion of the image.

pixel is within a maximum radius of the initial pixel. These pixels become the ROI. A second water-shed is then applied using a lower intensity threshold. These pixels become the "halo" region and are set to zero, removing them from the image and helping to separate the cells from one another. The ROI is rejected if it is too large or too small in which case all pixels in the region are set to zero. This process is repeated until a number of ROIs specified by the user has been identified. In the user checking step the set of ROIs are displayed and individual ROIs can be selected and deleted by the user, as required.

RESULTS

We compare the time course of cytosolic calcium in the urothelial cells using ROIs selected by eye with those identified by the automated algorithm outlined above. A set of 15 ROIs defined by hand is shown in Fig. 1A. This is representative of the number of ROIs typically defined during analysis of this kind of experiment. The ROIs are shown overlayed onto the 70th green image, which was used for this process as many cells appear very bright in this image and so can easily be identified. However, this introduces a strong bias towards cells that respond strongly to the ATP. The result of applying the custom algorithm is shown in Fig. 1B. Once the process is automated a much larger number of cells is identified, and in this example case 200 are shown. The algorithm uses information combined from the 1st red and 70th green images. This novel approach enables the algorithm to identify both cells which respond readily to ATP and cells that initially take up lots of dye. Furthermore, the adaptive threshold used in the algorithm can identify cells in the image across a wide range of intensities. Fig 2 shows the calcium transients generated from these two sets of ROIs. For the hand-defined ROIs the time profiles are all qualitatively very similar, with all 15 cells experiencing a large change in cytosolic calcium. This strongly suggests that the cells are largely

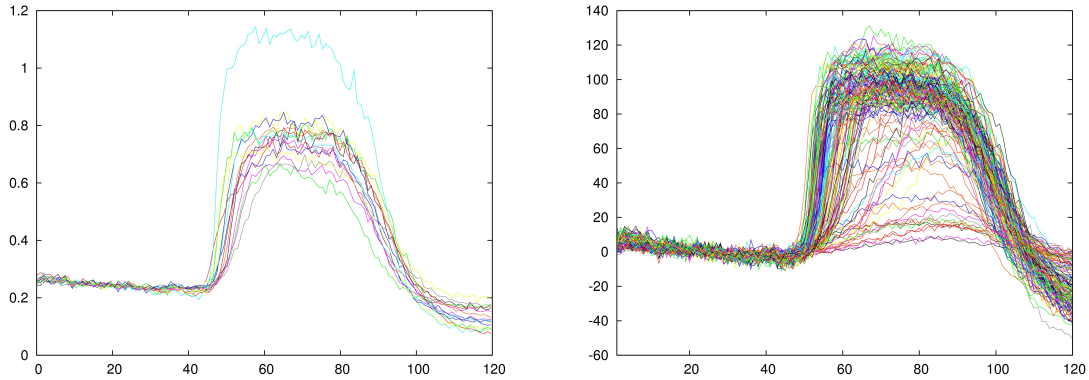

Figure 2: Time course of cytosolic calcium concentration extracted from the images using the ROIs shown in Fig. 1. **(A)** The 15 ROIs identified by hand all experience a large elevation following application of ATP. The conclusion would be that the cell population is largely homogeneous. **(B)** The set of 200 ROIs identified using the custom algorithm provides a much more accurate impression of the range of responses elicited in the cells and shows that the population has a high degree of heterogeneity.

homogeneous in their response to ATP. However, when the ROIs identified by the automated algorithm are analysed, a much wider range of responses is observed. Some cells undergo large elevations in cytosolic calcium and others respond little or not at all. In fact, the population appears to form a continuum with the middle and lower responding cells being equally represented in terms of cell numbers as the strongly responding cells. This removes the subjectivity of selection and demonstrates that the cell population is not homogeneous, an important observation that would be missed if the analysis were carried out using the ROIs defined by hand.

CONCLUSION

Here we have examined the impact of selection bias when analysing fluorescence microscopy data using images generated from our own experiments on urothelial cells. We have compared a conventional “by hand” method of identifying ROIs with a custom algorthim designed to minimise selection bias. When ROIs are identified by hand very few cells are identified and a strong selection bias is introduced towards cells that experience a large elevation of cytosolic calcium. With the automated algorithm a much larger proportion of cells are captured. We find that a broad range of responses is present within the cell population with approximately equal representation of low, mid and highly responding cells. The algorithm we have used is specialised for our images and may not generalise well to other cases. However, it is not the performance of this algorithm that is our primary concern; it is the impact that selection bias has on the conclusions drawn from the data. If the data we have shown here were analysed using only the ROIs identified by hand the population would appear to be largely homogeneous. Such a finding could be used as a justification for whole-field analysis of the images, which would further mask the heterogeneity in the responses of the cells by averaging the entire population. Selection bias when identifying ROIs can therefore strongly influence conclusions drawn from a data set, in particular when the population is intrinsically heterogeneous.

REFERENCES

Vincent, L., Soille, P., 1991. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Trans. Pattern Anal. Mach. Intell. 13 (6), 583–598.

Cheng, Y., 1995. Mean shift, mode seeking, and clustering. IEEE Trans. Pattern Anal. Mach. Intell. 17 (8), 790–799.

Yang, F., Mackey, M.A., Ianzini, F., Gallardo, G., Sonka, M., 2005a. Cell segmentation, tracking, and mitosis detection using temporal context. In: Duncan, J.S., Gerig, G. (Eds.), Proceedings of the Medical Image Computing and Computer-Assisted Intervention, vol. I, pp. 302–309.

Yang, X., Li, H., Zhou, X., Wong, S., 2005b. Automated segmentation and tracking of cells in time-lapse microscopy using watershed and mean shift; Proc. Int. Symposium on Intelligent Signal Processing and Communication Systems, pp. 533–536.

Osher, S., Sethian, J.A., 1988. Fronts propagating with curvature dependent speed: algorithms based on Hamilton–Jacobi formulations. J. Comput. Phys. 79, 12–49.

Li, K., Miller, E.D., Chen, M., Kanade, T., Weiss, L.E., Campbell, P.G., 2008. Cell population tracking and lineage construction with spatiotemporal context, Medical Image Analysis 12, pp. 546–566.

Smal, I., Niessen, W., Meijering, E., 2006. Bayesian tracking for fluorescence microscopic imaging. In: Proceedings of the IEEE International Symposium on Biomedical Imaging, pp. 550–553.

Godinez, W.J., Lampe, M., Wörz, S., Müller, B., Eils, R., Rohr, K., 2007. Tracking of virus particles in time-lapse fluorescence microscopy image sequences. In: Proceedings of the IEEE International Symposium on Biomedical Imaging, pp. 256–259.

Doucet, A., Ristic, B., 2002. Recursive state estimation for multiple switching models with unknown transition probabilities. IEEE Trans. Aerosp. Elec. Sys. 38, 1098–1104.

Smal, I., Draegestein, K., Galjart, N., Niessen, W., Meijering, E., 2007. Rao-blackwellized marginal particle filtering for multiple object tracking in molecular bioimaging. In: Proceedings of the International Conference on Information Processing in Medical Imaging, pp. 110–121.

Bahnson, A., Athanassiou, C., Koebler, D., Qian, L., Shun, T., Shields, D., Yu, H., Wang, H., Goff, J., Cheng, T., Houck, R., Cowsert, L., 2005. Automated measurement of cell motility and proliferation. BMC Cell Biol. 6 (19).

Kirubarajan, T., Bar-Shalom, Y., Pattipati, K.R., 2001. Multiassignment for tracking a large number of overlapping objects. IEEE Trans. Aerosp. Electron. Syst. 37 (1), 2–21.