A Methodology for Quality Control in Cell Nucleus Segmentation

Pascal Bamford and Brian Lovell,
Cooperative Research Centre for Sensor Signal and Information Processing,
Department of Computer Science and Electrical Engineering,
University of Queensland, Australia.
P.Bamford@cssip.uq.edu.au, B.Lovell@cssip.uq.edu.au

Abstract

In order to achieve the very high accuracy rates required in unsupervised automated biomedical applications, it is often necessary to complement a successful segmentation algorithm with a robust error checking stage. The better the segmentation strategy, the less severe the error checking decisions need to be and the fewer correct segmentations that are discarded. These issues are dealt with in this paper in order to achieve 100% accuracy on a data set of 19946 cell nucleus images using an established segmentation scheme with a success rate of 99.47%. The method is based upon measuring changes in the final segmentation contour as the one parameter that governs its behaviour is varied.

1. Introduction

Machine vision systems for the unsupervised automation of otherwise manual tasks usually require image processing components with exceptionally high accuracy rates. This is especially true in the biomedical domain where failures result in mis-diagnoses. The fact that research is still continuing on the development of a cervical cancer screening machine despite being initiated in the 1950's is perhaps a good indication of the size of the leap required to go from an algorithm obtaining 'good' results on a small test data set, to obtaining high levels of accuracy in a real environment. The main difficulty with this application has been identified as the robust segmentation of cells and cell nuclei. In fact, the segmentation stage has been reported to be 'the key to a working machine' [4]. Many algorithms have been proposed in the past with varying success, but just as important as a high accuracy rate is knowing when a failure has occurred, '... an erroneously segmented cell is much worse than a rejected cell' [11].

Authors in the past have included artefact and incorrect segmentation rejection schemes in their algorithms. MacAulay used a post-processing step after segmentation to remove potential artefacts based on shape and appearance that was capable of detecting some of the incorrectly segmented nuclei [9]. Nordin describes an algorithm that is able to report a failure at various levels of segmentation, as well as a separate artefact rejection stage [11]. McKenna used a neural network to pre-select potential nuclei in scenes for subsequent segmentation. It was pointed out that a post-processing stage would also be necessary to filter out 'erroneously detected objects' [10].

A common trait in these techniques is the use of a separate process to view the output of the segmentation and to use shape and appearance measurements to classify the results as 'pass' (looks like a cell) or 'fail' (doesn't look like a cell). We have proposed a segmentation scheme that not only employs an algorithm with much better performance than previously reported [3], but also enables a confidence measure in the resulting segmentation to be given.

2. The Segmentation System

For a full explanation of the underlying segmentation technique, the reader is referred to [3]. An active contour method was used and is summarised here only to introduce the development of the subsequent stages.

2.1. Active Contour Implementation

The use of active contours in bio-medical applications is well established and global minimum searching methods have been found to be particularly useful in the presence of the many artefacts usually associated with these images [5][6][7]. Here, a dynamically programmed search method was implemented that was based upon a suggestion in [8]. A search space is first set up within the image, bounded by two concentric circles centralised upon a point found by an initial rough segmentation. This search space is then sampled by discretising both circles and a number of radii joining them (figure 1).

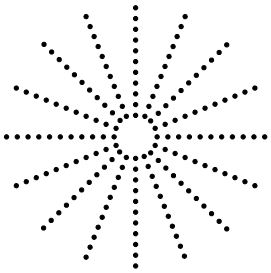


Figure 1. Discrete search space

Every possible contour that lies upon the points of the search space is then considered and an associated cost function is calculated. This cost represents the traditional snake property of a balance between the contour's smoothness and how much of it lies upon regions of high image gradient. This balance is controlled by a single regularisation parameter, $\lambda \in [0, 1]$. By choosing a high value of λ, the smoothness term dominates, which may lead the contour to ignore important image edges. A low value of λ will allow sharp corners to develop in the contour as it attempts to follow all high gradient edges, which may not necessarily be on the desired objects edge. Once every contour has been evaluated, the single contour with least cost is chosen as the solution.

2.2. Segmentation Accuracy

A data set of 19946 Pap stained cervical cell images was available for testing. These images were of the order of 128x128 pixels, quantised to 256 gray levels and each contained a single nucleus.

The single parameter that affects the behaviour of the algorithm, λ, was chosen to be 0.7 after measuring its effect on segmentation accuracy using a small sub-set of the images. This sub-set was made up of 141 known 'difficult' images from previous studies [3][2], augmented by a random sample of 269 images from the remaining data set. This careful data selection was necessary as previous experience showed that for the majority of images, the resulting segmentation was fairly insensitive to the choice of λ, making the choice of optimum value difficult. However, more demanding images require specific values to achieve correct segmentation. The results of this trial are shown by the graph of figure 2.

With λ set at 0.0, the smoothness constraint is completely ignored and the point of greatest gradient is chosen along each search space radius. Previous studies [2] have shown that for approximately 65% of images, all points of greatest gradient actually lie upon the nucleus cytoplasm border (figure 3(a)), so these cell images will be correctly segmented.

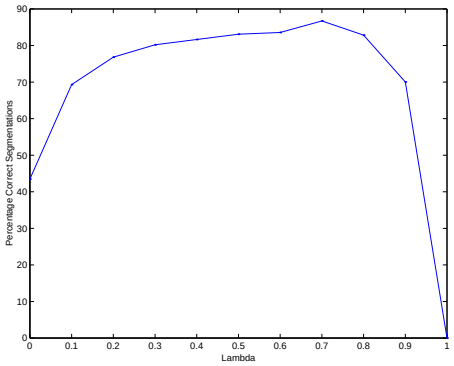
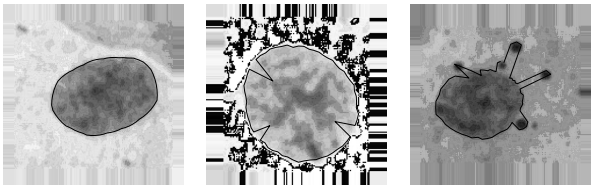


Figure 2. Plot of percentage of correct segmentations against λ for a set of images consisting of known 'difficult' images and randomly selected images.

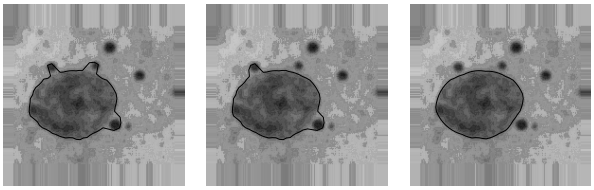


(a) (b) (c)

Figure 3. $\lambda = 0.0$. a) Largest gradients occur on the nucleus border, b) darkly stained chromatin generates largest gradients, c) dark artefacts generate largest gradients.

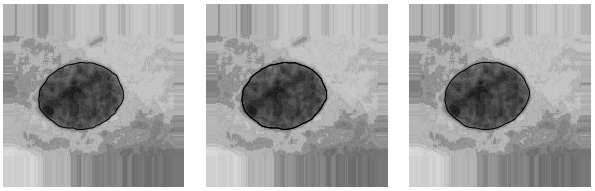
For the remaining 35% of images, a large gradient due to an artefact or darkly stained chromatin will draw the contour away from the desired border (figures 3(b)&(c)). As λ increases, the large curvatures present in these configurations become less probable (figure 4).

The graph shows a value of $\lambda = 0.7$ as the most suitable for these particular images. Every image in the data set was then segmented at $\lambda = 0.7$ and the results verified by eye. Of the 19946 images, 99.47% were found to be correctly segmented. Three main classes of failure were identified. Eighty seven of the failures were due to the nuclei lying close to the cytoplasm boundary. As the background cytoplasm boundary contrast is much greater than that of the nucleus cytoplasm boundary, the contour tended to lie upon the former very low image energy area (high gradient edges). Fourteen of the failures were caused by the inappropriate choice of λ for that individual image (they all



(a) (b) (c)

Figure 4. The effect of increasing λ. a) $\lambda = 0.1$, b) $\lambda = 0.2$, c) $\lambda = 0.5$.



(a) (b) (c)

Figure 5. Example of an image that is stable over a range of λ. (a) $\lambda = 0.1$, (b) $\lambda = 0.5$, (c) $\lambda = 0.7$.

subsequently produced correct segmentations with different values of λ.) The remaining four images were found to fail at all attempts. The failures due to the presence of the background in the nucleus images are preventable through careful design of a prior cell-finding stage [1]. Here, the cytoplasm background boundary is known and can therefore be prevented from appearing in the nucleus images. The detection of the other classes of failure is therefore the major issue.

3. Development of an Error Checking Framework

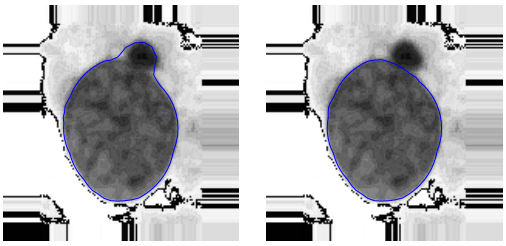
Despite the exceptionally high accuracy rate that the global minimum searching contour method achieves, there is still a possibility of sample contamination from the few failures that do occur. In order to prevent this, the need would still exist for a human to view the output of this stage, undermining its utility in a practical system. The remainder of this paper therefore concerns itself with the development of a framework that further increases the accuracy of a potential system.

3.1. Lambda Sensitivity

For the majority of relatively simple images with little ambiguity in the true location of the nuclear boundary, the final segmentation can be fairly insensitive to λ over a wide range of values (figure 5).

By contrast, 'difficult' images (even for humans) produce very different contours depending upon the choice of λ (figures 4 and 6).

These images usually contain artefacts near or on the nuclear boundaries that make the 'true' border hard to find. These examples show that no single value of λ is capable of accurately segmenting all of the images. Therefore, rather than segment the images at one value of λ and use a post-process to reject possible failures, we are interested in viewing the output of the algorithm for various values of λ in order to detect stability as a measure of confidence in the resulting segmentation.



(a) (b)

Figure 6. Example of an image that is not stable over a range of λ. (a) $\lambda = 0.5$, (b) $\lambda = 0.7$

3.2. Error Checking

The graph of figure 2 shows monotonically increasing segmentation accuracy for $0.0 < \lambda < 0.7$. In fact, from the data it was observed that the set of correct segmentations at λ_1 was a strict subset of the set of correct segmentations at λ_2 where $\lambda_1 < \lambda_2 < 0.7$. Therefore, by segmenting any image at the highly probable value of $\lambda = 0.7$ for success and again at $\lambda = 0.0$, a similarity between the two contours indicates a high level of stability (figure 5). This image is then classified as a 'very easy' image to segment and for convenience labelled 'level 0'. Lack of similarity leads to a comparison of the original contour at $\lambda = 0.7$ with a contour at $\lambda = 0.1$. Similarity leads to a classification of level 1 and so on.

This classification method suggests a means to discard bad segmentations. For example, if we keep only level 0 cell images, we discard approximately a third of the data set, but achieve a 100% correct segmentation rate on those retained [2].

4. Fine Tuning

In order to pursue this method, the data set was split into two sets: $\mathcal{F}$, Those images that been incorrectly segmented at $\lambda = 0.7$ (105 images) and $\mathcal{C}$, those that had been correctly segmented (19841 images). Statistics were then measured for each level by comparing the segmentation at $\lambda = 0.7$ with those at $\lambda = 0.0, 0.1, ..., 0.6$ for every image in both sets.

As the contours to be compared were the result of the same algorithm and indeed the same search space in the image, the comparison between any two contours is trivial. The distance between each chosen point on each of the search space radii (figure 1) for each contour was calculated and the maximum absolute deviation (MAD) evaluated.

A cumulitative plot of the percentage of the set $\mathcal{F}$ against MAD for level zero (comparison between contours at $\lambda = 0.7$ and $\lambda = 0.0$) is shown in figure 7.

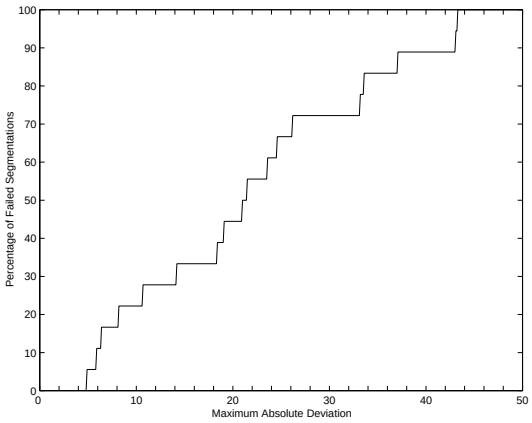


Figure 7. A plot of the percentage of the elements of $\mathcal{F}$ (incorrect segmentation at $\lambda = 0.7$ on the test data set) against measured MAD for level 0.

This graph shows that for level zero, a MAD threshold of 4.84 pixels would detect every failed segmentation. In a similar manner, it is possible to establish thresholds for each level so that the detection of every failed segmentation is guaranteed (table 1).

Level	0	1	2	3	4	5	6
Threshold	4.85	3.20	2.45	0.79	0.79	0.79	0.79

Table 1. Minimum MAD thresholds for the detection of every element in $\mathcal{F}$ (incorrect segmentation at $\lambda = 0.7$ on the test data set) for levels 0 - 6.

The thresholds decrease with increasing level. This is expected as closer values of λ are compared at higher levels. The values then taper to a limit of 0.79 pixels as this is the distance between two adjacent radial points on the discrete search space.

In order to establish the effect of setting such thresholds on $\mathcal{C}$, the percentage of $\mathcal{C}$ that would be discarded against MAD threshold for level zero is shown in figure 8.

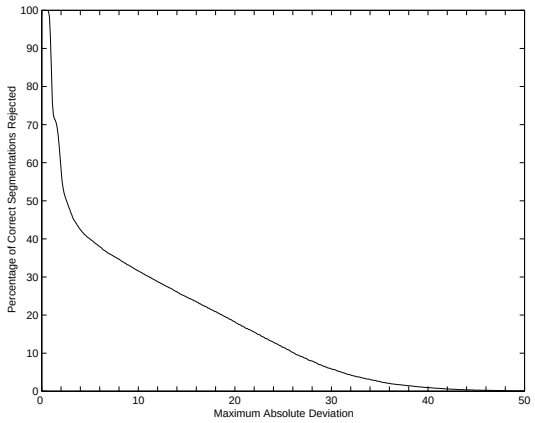


Figure 8. A plot of the percentage of elements of $\mathcal{C}$ (correct segmentations at $\lambda = 0.7$ on the test data set) rejected against MAD threshold for level 0.

Therefore, by setting a threshold of 4.84 pixels and rejecting any segmentation with a greater MAD, 40.2% of the correct segmentations would be discarded. This procedure may be repeated for each level, using the thresholds previously calculated. The percentage of C that falls above the threshold for each level (i.e. a good segmentation being discarded) against MAD is shown in figure 9.

Although a harsher threshold is used at level 1 than at level 0, fewer correct segmentations are discarded. This is due to the absence of any smoothness constraint at $\lambda = 0.0$ which leads to the wild deviations such as those shown in figure 3. However, the small smoothness contribution at $\lambda = 0.1$ corrects many of these deviations resulting in the large drop in average MAD (table 2).

Level	0	1	2	3	4	5	6
Average MAD	8.90	2.78	1.91	1.56	1.35	1.22	1.07

Table 2. Average MAD for levels 0 - 6.

Therefore by running at level 2, it is possible to detect every failure and only discard 10.78% of the correct segmentations.

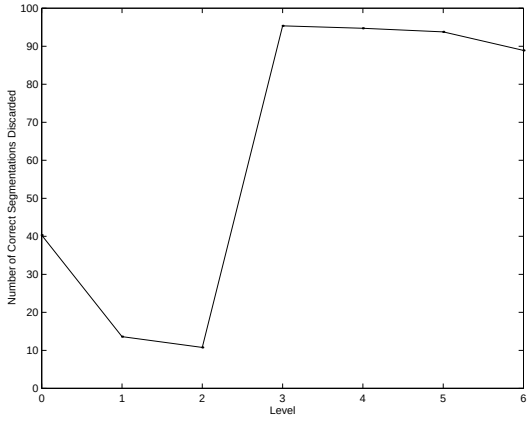


Figure 9. Plot of the percentage of elements of $\mathcal{C}$ (correct segmentations at $\lambda = 0.7$ on the test data set) rejected at each level using the thresholds of table 1.

5. Conclusions

By analysing the modes of failure of a highly successful cell nucleus segmentation algorithm, an error checking framework was implemented that was capable of detecting every failure. The algorithm parameter was first optimised for the data set. It was then noticed that different values of the algorithm parameter obtained different solutions for difficult images, but simple images generated stable solutions. Therefore, by varying the algorithm parameter this stability could be detected. A decision to reject or accept the segmentation was then made, based upon measured thresholds for each level. For the data set of 19946 images, it was found that by comparing the resulting contours at values of the algorithms parameter $\lambda = 0.7$ and $\lambda = 0.2$, and rejecting the segmentation if the maximum absolute deviation (MAD) between the contours was greater than 2.45 pixels, every failure could be detected whilst only discarding 10.78% of the correct segmentations. In this study, only values of λ with a resolution of 0.1 have been considered. It is possible that by increasing this resolution in the region of interest (i.e. near 'level 2' operation) and repeating the exercise, a further increase in the performance could be achieved. Naturally, the parameters and results that have been reported are optimised not only for one type of image but also for the hardware configuration that was used to capture them. Future work will involve the incorporation of the proposed system into a *Cytometer* (an automatic imaging system) using the same methodology to achieve optimal performance for that hardware. This will also allow much more extensive analysis of the proposed methods through the accessability of a greater amount of data. This result has great potential for implementation in an unsupervised cancer screening device where only a sample of cells is required.

Finally, The 'rejected' cells have simply been labelled as such. These could be interpreted as having been 'flagged' by the algorithm as problematic and requiring processing by a higher level (e.g. to invoke a different algorithm etc.) By achieving such high accuracy rates and confidence in the segmentation stage, the following feature extraction and classification processes can only become more robust.

References

[1] P. Bamford and B. Lovell. A water immersion algorithm for cytological image segmentation. In *Proceedings of the APRS Image Segmentation Workshop*, pages 75–79, University of Technology Sydney, Sydney, December 1996.

[2] P. Bamford and B. Lovell. Improving the robustness of cell nucleus segmentation. In P. H. Lewis and M. S. Nixon, editors, *Proceedings of the Ninth British Machine Vision Conference, BMVC '98*, pages 518–524, University of Southampton, England, UK, September 1998.

[3] P. Bamford and B. Lovell. Unsupervised cell nucleus segmentation with active contours. *Signal Processing Special Issue: Deformable Models and Techniques for Image and Signal Processing*, 71(2):203–213, December 1998.

[4] E. Bengtsson. The measuring of cell features. *Analytical and Quantitative Cytology*, 9(3):212–217, June 1987.

[5] L. D. Cohen and I. Cohen. Finite-element methods for active contour models and balloons for 2-D and 3-D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1131–1147, 1993.

[6] C. A. Davatzikos and J. L. Prince. An active contour model for mapping the cortex. *IEEE Transactions on Medical Imaging*, 14(1):65–80, 1995.

[7] D. Geiger, A. Gupta, L. Costa, and J. Vlontzos. Dynamic programming for detecting, tracking, and matching deformable contours. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(3):294–302, 1995.

[8] S. R. Gunn. *Dual Active Contour Models for Image Feature Extraction*. PhD thesis, University of Southampton, May 1996.

[9] C. E. MacAulay. *Development, implementation and evaluation of segmentation algorithms for the automatic classification of cervical cells*. PhD thesis, University of British Columbia, August 1989.

[10] S. J. McKenna. *Automated analysis of papanicolaou smears*. PhD thesis, University of Dundee, October 1994.

[11] B. Nordin. *The development of an automatic prescreener for the early detection of cervical cancer: algorithms and implementation*. PhD thesis, Uppsala University, 1989.