Automating Cell Segmentation Evaluation with Annotated Examples

Pascal Bamford

Cooperative Research Centre for Sensor Signal and Information Processing,
Department of Information Technology and Electrical Engineering,
The University of Queensland
E-mail: P.Bamford@cssip.uq.edu.au

Abstract

Previously the development of a cell nucleus segmentation algorithm had been evaluated by eye by the author. This is an impractical method when attempting to evaluate and compare many algorithms and parameter sets on very large data sets. For this work, a dataset of 20,000 cell nucleus images was annotated by hand by three non-expert assistants. This paper concentrates on comparing the previous interactive approach to evaluating a segmentation algorithm to automated techniques using this annotated data.

1. Introduction

We have previously reported work on using a dynamic-programming algorithm for cell nucleus segmentation [1]. In that work, almost 20,000 cell images were segmented and the output judged as either a *pass* or a *fail* by the author (where any deviation from the perceived boundary was declared a fail). Then, an attempt was made to tune the algorithm parameter over a subset of the data using the same evaluation method.

This is clearly an extremely time consuming and subjective process. In fact, of the arguments encouraging more evaluation in computer vision [2], it seems that finding a better method than *eye-balling* results over the large datasets required to develop real systems is the most compelling! This is especially so if many algorithms and parameter sets are to be compared thoroughly.

The ultimate method of evaluating segmentation algorithms is to use the final outcome of the complete vision system as the performance metric [3]. Unfortunately this is very difficult except in the simplest of cases. This is due to the fact that there may be many processes, each with their own sources of variability and complex interactions, between the segmentation output and final measure [4], thus requiring very large datasets in order to perform a robust experiment. Attempts at evaluation therefore either evaluate individual components in isolation or consider the final outcome of the system [6].

Image segmentation is a module that is generally evaluated in isolation. This is either done by eye, via annotated examples or via some other goodness measure that does not rely on *ground truth* (e.g. inter-region contrast). This latter method is generally the only available option to those working in *general recovery* [2] work exemplified in [7]. Of the methods that employ annotated examples (Zhang's *empirical discrepancy methods* [8]), either the segmentation masks are pixel-wise compared or features extracted from those masks are compared. The latter method has been criticized as it is possible to obtain good agreement for a feature where the masks do not agree well [9] (trivial example: area). Also extracted features can be very sensitive to small differences in masks, complicating the detection of *significant* differences [5].

The difficulties associated with empirical discrepancy evaluation have been summarized to be [9]

- difficulties in defining measures/metrics,
- standardizing evaluation protocols, but mostly
- determining and acquiring ground truth data.

It is well known that image segmentation is a highly application-dependent task. Previous approaches to evaluation, which are briefly summarized in the following two sections, seem to show that this task is also more application dependent that one may expect. Selected techniques are then applied to the task of verifying results previously obtained by eye for cell segmentation [1].

2. Error Measures and Metrics

A framework for evaluating segmentation methods has recently been proposed where the measure was trained using examples of failure [10]. The error value measured edge-detection type errors (*bits* - false positive edges, and

holes - false negative edges) that were assembled into patterns and then rated by human observers. The individual errors were weighted by a number of parameters and the measure trained to match the observers' score. This measure was classified as belonging to a group termed *low error models*, i.e. suited only to problems where the segmentation is already very near the final solution (and was tested upon synthetic images).

We investigated the failure modes for a number of algorithms in [1] and found that they generally either failed quite dramatically or performed an *acceptable* job (little or no perceived delineation error). Thus we are initially more interested in employing a measure that is capable of measuring *large* differences.

Zhang [8] reviewed a number of simple measures of which only two are applicable here: the number and position of misclassified (segmented) pixels. Also reviewed were methods for measuring over- and under-segmentation. These errors, and those of Roman-Roldin [10], are of less interest in this work as a well-formed mask is a pre-requisite to accepting the segmentation. Thus we assume that a mask for the object of interest is the final output of the segmentation stage (including all pre- and post-processing). A simple method of error checking, for this application, is then to evaluate the Euler number of the mask image. If it is not equal to one, then the mask is rejected outright and a failure assigned to that segmentation - the failure need no longer be quantitatively evaluated. This may be seen as first-step goodness measure that requires no ground truth.

More recently Chalana [9] employed the Hausdorff distance and average distance between human and computer boundaries. The Hausdorff distance is an attractive metric for this application as segmentation algorithms generally tend to fail in one localized position around the nuclear border - an error that may become masked when using normalized or average values [11]. The Hausdorff distance is defined to be the maximum of the set of shortest distances between corresponding points of two shapes.

3. Evaluation Protocols

Algorithms are generally evaluated using either the raw measures to establish how close competing algorithms (and different parameters sets) get to the annotated (observers') data or by thresholding the measures in order to obtain percentage success rates which are then compared. Chalana [9] used both methods and a number of observers' data to determine whether the computer boundary differed from the observers' boundaries as much as the observers' boundaries differed from one another. This was evaluated using a modified Williams' index and a *percent statistic*. The Williams' index, I', divides the average number of agreements (inverse disagreements, $D_{j,j'}$) between the computer ('observer' 0) and $n-1$ human observers (j) by the average number of agreements between human observers (eq. 1).

$$I' = \frac{\frac{1}{n}\sum_{j=1}^{n}\frac{1}{D_{0,j'}}}{\frac{2}{n(n-1)}\sum_{j}\sum_{j':j'\neq j}\frac{1}{D_{j,j'}}} \quad (1)$$

If the upper value of the confidence interval of the result is greater than one, then it is concluded that the computer is a reliable member of the group of observers. The percent statistic measures the percentage of cases where the computer boundary lies within the inter-observer range.

Finally Everingham [12] has recently suggested an interesting method for combining large sets of results into an ROC type curve, where only the best performing results contribute to the final output.

4. Method

Although no *ground truth* exists for cell segmentation, the images do not necessarily require expert annotation. Figure 1 shows an example image and three corresponding non-expert annotations. This is quite a deviation from

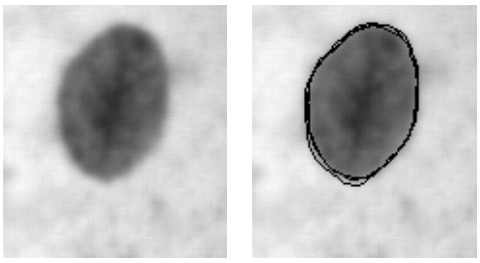

Figure 1. Example from dataset and corresponding observers' boundary

other imaging modalities where inter- and intra- observer variance can be very high, but valid (e.g. ultra-sound [9]). It is however desirable to obtain a number of interpretations so that inter-observer variability may be nonetheless investigated.

A Wacom PL400 pen-and-tablet was used to input the data. This device enabled almost immediate use by the observers to delineate the cell nuclei. The observers were instructed to draw a continuous line between the nucleus and background (cytoplasm), i.e. on the *transition region* [13] of the edge. Due to the low pass filtering effect of the optics used to capture the images, this covers a number of pixels and is readily identifiable in the majority of examples. However the exact area for delineation was not overly

specified and left to the individual. Three observers were employed to annotate the entire 20,000 image dataset. The nucleus images are of the order of 128x128 pixels. The PL400 LCD screen has square pixels of pitch 0.264mm. By displaying the images at the native screen resolution therefore produced cell nuclei of approximately 1-2cm diameter on screen. This was found to be too fiddly and handshake became a problem. Thus the images were first upsampled to twice the original dimensions using a nearest-neighbour algorithm. The pen line thickness on the screen was made equal to one pixel at the original image resolution (i.e. four pixels on screen). This also assisted to reduce handshake.

There has been considerable work in improving the implementational performance of the Hausdorff metric for the more general problem of comparing shapes under transformation [14]. Here, we have implemented a rapid and simple routine to obtain the maximum distance between corresponding points, d_{MAX}, on two binary masks by

1. Obtaining the distance transforms, A_{DT} and B_{DT}, of the perimeter of the mask images, A and B.
2. Obtaining the pixel-wise maximum of A_{DT} and B_{DT} to produce AB_{DT}.
3. Obtaining the XOR of the mask images, AB_{XOR}
4. Using AB_{XOR} to mask AB_{DT} to produce AB_{MASK}
5. Obtaining d_{MAX} as the maximum value in AB_{MASK}.

These steps are illustrated in figure 2. The average distance between the masks was also computed. Chalana [9] used an iterative technique to evaluate the average distance between two curves, which yielded an average curve as a result. Here we implemented a rapid method of evaluating the average distance, d_{AV}, as the average value in AB_{MASK}. These two measures can be obtained very rapidly using operations for which implementations are widely available.

The above data and measures were then used to confirm the results reported in [1]. This was done by comparing the algorithm performance against its (regularisation) parameter, λ. The Williams' index was first computed, using d_{MAX} as the discrepency measure $D_{j,j'}$, over half of the data over the full range of permissible [1] values of λ ($\in [0, 1]$) at increments of 0.1. Figure 3 shows an example distribution of the Williams' index for $\lambda = 0.2$. This plot shows a group of (normally distributed) values near the observers' boundaries with a mean value near 1.0. In addition, there are a number of out-lying counts between 0.0 and roughly 0.5. These correspond to the failed segmentations. Thus rather than compare performance versus λ using summary statistics, as in [9], the Williams' index was thresholded in order to determine the percentage of correct segmentations at that threshold. However the selection

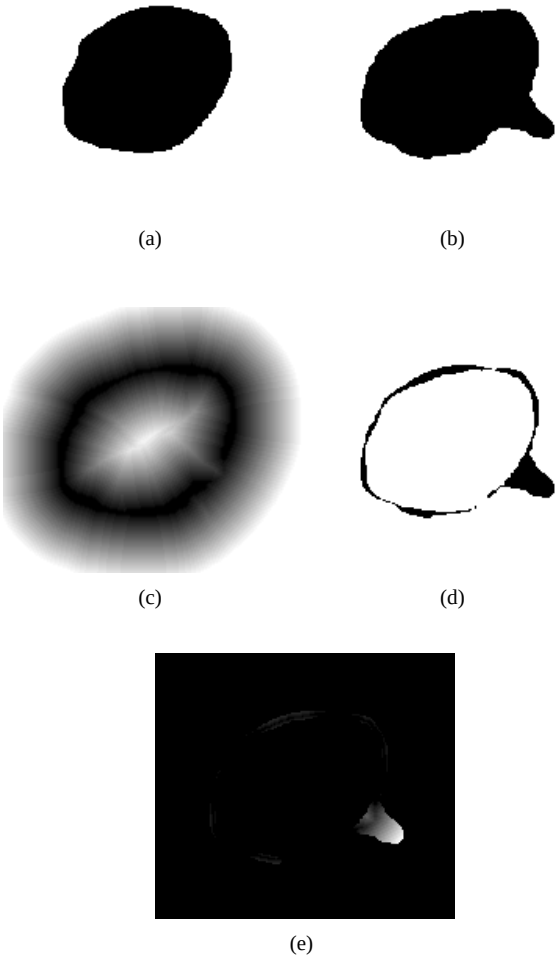

Figure 2. Steps for deriving the maximum distance between two binary masks. (a) and (b) are the two masks to be compared. (c) is the pixel-wise maximum of the distance transforms (DT) of the perimeters of (a) and (b). (d) is the result of the XOR operation on the two masks. (e) shows the final masked DT, the maximum value of which is the desired value, corresponding to the length of the protrusion in the mask (b).

of such a threshold is at this stage an arbitrary process - what constitutes a *correct* segmentation? Therefore algorithm performance was plotted over the natural range of the Williams' index, i.e. $\in [0, 1]$. At a value of 0, the computer boundary is infinitely far from those of the observers. At a value of 1 the computer boundary is as close to the observers as they are to each other. Values above 1 are of less

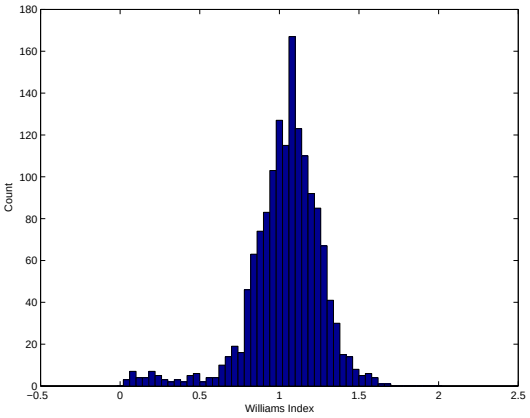

Figure 3. Distribution of the Williams index for $\lambda = 0.2$

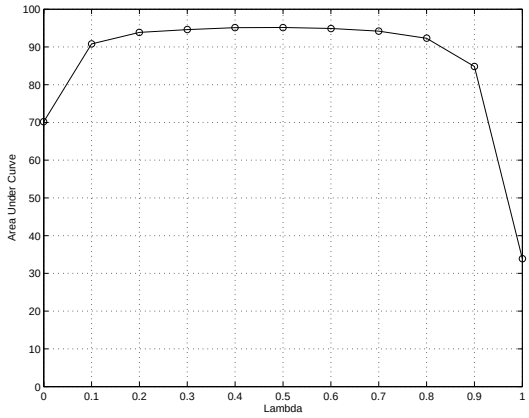

Figure 5. Plot of area under curves of figure 4 against λ.

interest - this can be understood to occur when one of the observers disagrees with the other two *more* than the computer does. Therefore all values greater than 1 are treated as a correct segmentation. Figure 4 is a convenient normalised and bounded representation to view segmentation algorithm performance against multiple observers in a single output. As a measure of overall performance at each

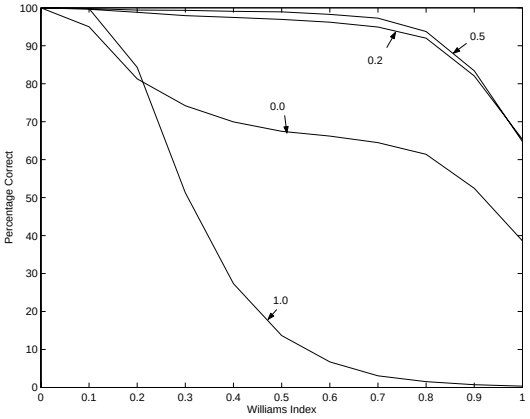

Figure 4. Plot of cumulative percentage success segmentation versus Williams Index for values of λ.

value of λ, the area under this curve was computed and is represented in figure 5. The maximum value of these area values is 95.16% which occurs at $\lambda = 0.5$.

5. Conclusion

We have considered methodologies for evaluating cell segmentation using annotated examples in an automated fashion. We found that recent approaches to segmentation evaluation have concentrated on low error models where the measures and metrics for segmentation error, in addition to the evaluation procedures, were unsuitable for this application. The shape of the graph in figure 5, the optimal value of algorithm parameter λ and the overall performance rate all agree well to the results reported in [1]. However, they were attained using a far more satisfactory and repeatable method.

This work represents a very early part of a greater project to thoroughly evaluate cell segmentation methods using annotated examples. The next stage is to attempt to compare other algorithms for this task. In addition, individual modules may be evaluated in isolation. For example marker extraction algorithms may be evaluated using the same data but measures that detect whether an inner marker is completely within the desired object. This work will then be expanded to include the original scene images from which the nucleus images were captured, representing a different segmentation task. Also, methods that use annotated data to obtain edge and region models, enabling the improvement or design of algorithms [15] will eventually be investigated. Finally, once a small number of discrepancies in the observers' data have been fixed (i.e. where the Williams' index is significantly greater than one!), this dataset will be made publicly available.

6. Acknowledgement

The author wishes to sincerely thank Wacom Co., Ltd for financially assisting the purchase of the PL400 pen-and-tablet.

References

[1] P. Bamford and B. Lovell, "Unsupervised cell nucleus segmentation with active contours," *Signal Processing*, vol. 71, no. 2, pp. 203–13, 1998.

[2] R. C. Jain, T. O. Binford, M. A. Snyder, Y. Aloimonos, A. Rosenfeld, T. S. Huang, K. W. Bowyer, and J. P. Jones, "Ignorance, myopia, and naivete in computer vision systems," *CVGIP: Image Understanding*, vol. 53, no. 1, pp. 112–28, 1991.

[3] B. McCane, "On the evaluation of image segmentation algorithms," in *DICTA'97 and IVCNZ'97*, 1997, pp. 455–9.

[4] R. M. Haralick, "Performance characterization in computer vision," in *Proceedings of 5th International Conference on Computer Analysis of Images and Patterns (CAIP'93)*, D. Chetverikov and W. G. Kropatsch, Eds. Washington Univ. Seattle WA USA, 1993, pp. 1–9.

[5] C. MacAulay, "Development, implementation and evaluation of segmentation algorithms for the automatic classification of cervical cells," PhD, University of British Columbia, 1989.

[6] M. Greiffenhagen, D. Comaniciu, H. Niemann, and V. Ramesh, "Design, analysis, and engineering of video monitoring systems: an approach and a case study." Visualization Dept. Siemens Corp. Res. Inc. Princeton NJ USA, 2001.

[7] K. Cho, P. Meer, and J. Cabrera, "Performance assessment through bootstrap," *IEEE-Transactions-on-Pattern-Analysis-and-Machine-Intelligence*, vol. 19, no. 11, pp. 1185–98, 1997.

[8] Y. J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern-Recognition*, vol. 29, no. 8, pp. 1335–46, 1996.

[9] V. Chalana and Y. Kim, "A methodology for evaluation of boundary detection algorithms on medical images," *IEEE Transactions on Medical Imaging*, vol. 16, no. 5, pp. 642–652, 1997.

[10] R. Roman-Roldan, J. F. Gomez-Lopera, C. Atae-Allah, J. Martinez-Aroza, and P. L. Luque-Escamilla, "A measure of quality for evaluating methods of segmentation and edge detection," *Pattern Recognition*, vol. 34, no. 5, pp. 969–80, 2001.

[11] A. Hammoude, "An empirical parameter selection method for endocardial border identification algorithms," *Computerized Medical Imaging and Graphics*, vol. 25, pp. 33–45, 2001.

[12] M. Everingham, H. Muller, and B. Thomas, "Evaluating image segmentation algorithms using the pareto front," *Lecture Notes in Computer Science*, vol. 2353, pp. 34–48, 2002.

[13] J. J. Gerbrands, "Segmentation of noisy images," PhD, Delft University, The Netherlands, 1988.

[14] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the hausdorff distance," *IEEE-Transactions-on-Pattern-Analysis-and-Machine-Intelligence*, vol. 15, no. 9, pp. 850–63, 1993.

[15] M. Brejl and M. Sonka, "Object localization and border detection criteria design in edge-based image segmentation: automated learning from examples," *IEEE Transactions on Medical Imaging*, vol. 19, no. 10, pp. 973–85, 2000.