Automated Cellular Detection in Histopathological Breast Cancer Images

Sophia Sanchez
Department of Computer Science, Stanford University
353 Serra Mall, Stanford, CA 94305

Abstract - **Breast cancer is one of the leading types of cancer, representing 1 in 4 new cancer diagnoses in women worldwide. Cellular proliferation and analysis of histopathological slides represents one of the strongest prognostic factors in breast carcinoma. Currently, this analysis is performed manually by a trained pathologist, but automated cell detection may provide a more accurate and scalable alternative. The current research relied on a series of 12 benign and malignant breast cancer cell slides with associated ground truth binary images and cell counts. Each image was run through a cell detection algorithm consisting of Gaussian filtering, dilation reconstruction, Otsu thresholding to generate a binary image, masking to remove non-cellular artifacts, and watershed segmentation to count the resulting cells. This method achieved an overall error rate of 1.57% relative to the ground truth dataset. This research in biological image analysis provides a promising avenue for automated cellular detection and improved prognostics.**

Keywords: Cell detection, breast cancer, watershed, dilation reconstruction

I. INTRODUCTION

Breast cancer is the second most common cause of cancer death for women, and breast cancer alone makes up 1 in 4 of all new cancer diagnoses in women worldwide [1]. In the United States alone, 12% of women will develop breast cancer over the course of their lifetime, and 41,000 patients die from breast cancer per year [2, 3]. The identification and tracking of breast cancer after diagnosis are critical components for therapeutic decision-making.

Cell proliferation represents the strongest prognostic factor in breast carcinoma [4]. The current gold standard to evaluate cell proliferation is manual cell count and cell analysis by a trained pathologist [5]. However, this established technique is subjective and prone to error [5]. In particular, the heterogeneity of breast cancer cells combined with the high inter-observer error rate associated with current methodology provides the opportunity for computational image analysis techniques to improve upon the status quo. One of the most critical steps in computer-aided diagnostics (CAD) is accurate automated cell identification and cell count of histological slides. This serves as a foundational component for quantitative analyses, including cell morphology. As such, automated cell identification and count provides an opportunity to improve breast cancer prognostics and therapeutic decision-making.

II. BACKGROUND

A number of efforts have been made to automate cell detection in breast cancer cell slides. These efforts have included the use of diverse methods such as Support Vector Machines (SVMs) [6], Laplacian of Gaussian (LoG) filtering [7], and radial symmetry-based voting [8].

However, cell detection algorithms and techniques have been hindered by a number of key problem areas. First, histopathology images tend to have significant background artifacts and noise, including non-cellular objects, blurring, and poor contrast between cell and non-cell objects [9]. Removing these artifacts to improve the signal-to-noise ratio remains a significant challenge for cell detection. Second, cells often cluster or overlap in histopathology images, making cell wall detection or nucleus identification methods more difficult. Lastly, nuclei and cells exhibit a wide variety of morphologies, such that differences in sizes, shapes, and textures must be taken into consideration by a cell detection algorithm. This task proves especially difficult with malignant cell slides, which may have a higher frequency of mitotic cells, increased frequency of highly undifferentiated cells, and cells with an atypical shape.

As such, cell imaging techniques that perform well when tested using a benign cell dataset may perform poorly when analyzing malignant cells [9]. When taken in conjunction, slide artifacts, cell clustering, and malignant cell morphologies pose significant challenges to the task of automated cellular detection in breast cancer histopathology images.

III. DATASET

The dataset consists of 58 Hematoxylin and Eosin (H&E) histopathology images of breast tissue published by the Center for Bio-image Informatics, University of California, Santa Barbara [10]. This image set consists of 27 cancerous cell images and 32 benign images cut from 10 H&E stained breast cancer biopsies. All 58 images are stored in 24-bit RGB format with a resolution of 896 x 768. Of these images, 6 benign and 6 malignant images were additionally analyzed in an approximately 200 x 200 pixel subsection of the original slide. This subset of 12 images served as the dataset for the present study. The ground truth cell outlines and cell counts were determined by a pathologist (fig. 1).

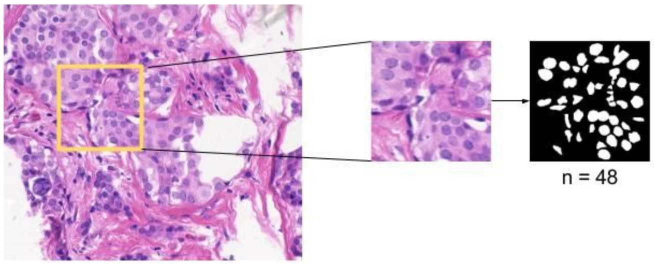

Fig. 1. Example histopathology image of breast tissue, 200 x 200 pixel subsection, and ground truth cell outlines and count.

IV. METHODS

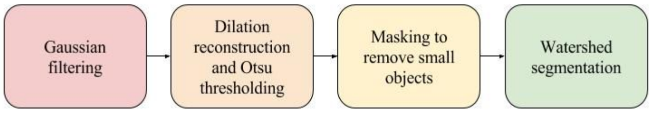
Gaussian filtering
Dilation reconstruction and Otsu thresholding
Masking to remove small objects
Watershed segmentation

Fig. 2. Summary of image processing and segmentation processing.

Each image was passed through a series of filters and transforms in order to identify and count the cells (fig. 2). All images were processed using the Python packages numpy, scikit-image, scipy, and matplotlib [11,12,13,14]. First, a Gaussian filter was applied in order to attenuate high frequencies and reduce noise. Next, dilation reconstruction was applied. Morphological reconstruction by dilation serves as a way to isolate connected regions of an image by marking local maxima and connecting pixels less than or equal to that local seed value to the seed. Thus, this method is particularly useful for demarcating clustered cells.

After reconstruction, Otsu thresholding was used to generate a binary image of the cells. This method automatically performs clustering-based image thresholding by calculating the threshold value that maximizes the inter-class variance [15]. To find this optimal value, an exhaustive search is performed using the following formula:

$$\sigma_w^2(t) = \omega_0(t)\sigma_0^2(t) + \omega_1(t)\sigma_1^2(t)$$

where weights ω_0 and ω_1 are the probabilities of the two classes separated by some threshold t and σ_0^2 and σ_1^2 are the variances of those classes. Third, masking was used to remove small artifacts from the binary image that were unlikely to represent cells. The threshold for masking was determined via heuristic methods based on the known magnification of the histopathological images and the minimum expected size of a cell.

Lastly, a watershed algorithm was used in order to identify individual cells [16, 17]. Specifically, a distance transform was applied to the binary image to compute distances to the background. This transform represents the distance from every pixel to the nearest nonzero-valued pixel. Next, the maximum distances are chosen to be markers, and the cells are separated along a watershed line by flooding the resulting basins, allowing for single-cell demarcation. These four processing steps provide a methodology for automated cell detection in breast cancer cell images.

V. RESULTS

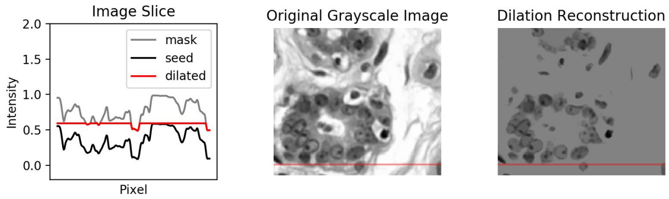

Fig. 3. Summary of dilation reconstruction in a representative histopathological image. Intensity by pixel (left), original grayscale image (center), and image after dilation reconstruction (right). The red line represents an example image slice, from which mask and seed intensities are calculated.

After Gaussian filtering, each image underwent dilation reconstruction (fig. 3). After the masking process, by which small, non-cellular artifacts were removed, Otsu thresholding was applied in order to generate the binary image. Finally, the watershed algorithm was used to delineate cells from one another. This process was conducted for both benign (fig. 4) and malignant (fig. 5) cell images.

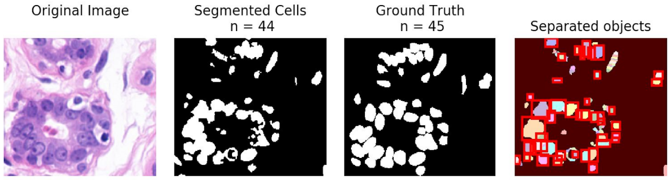

Fig. 4. Binary image after Otsu thresholding (2nd from left) and watershed (rightmost) in a representative histopathological image of benign cells. Identified cells are marked by red boxes (rightmost). The original image (leftmost) and ground truth image (2nd from right) were provided in the dataset [10]. Predicted and ground truth cell counts are listed below the respective binary images.

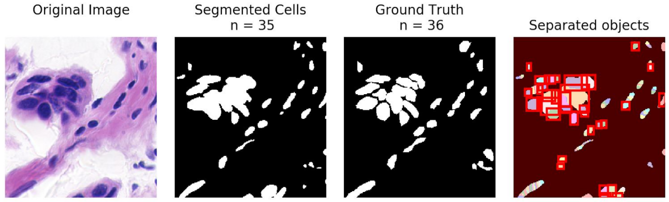

Fig. 5. Binary image after Otsu thresholding (2nd from left) and watershed (rightmost) in a representative histopathological image of malignant cells. Identified cells are marked by red boxes (rightmost). The original image (leftmost) and ground truth image (2nd from right) were provided in the dataset [10]. Predicted and ground truth cell counts are listed below the respective binary images.

In the benign cell images, 359 cells were identified, while the ground truth cell count across all benign cell images was 355 (fig. 6). This represents an error rate of 1.13%. In the malignant cell images, 269 cells were identified, while the ground truth cell count across all malignant cell images was 283. This represents an error rate of 4.95%. When combining the two cell types, 628 cells were identified, with a ground truth of 638 cells, and an error rate of 1.57%.

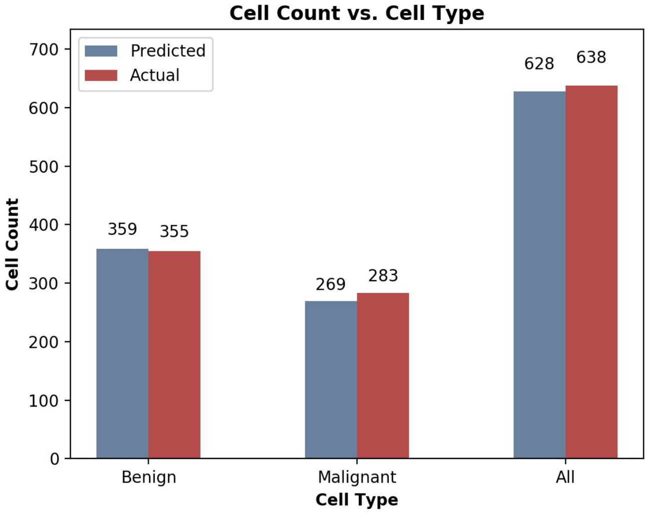

Fig. 6. Summary of predicted and actual cell count in benign and malignant cell images. Total predicted and actual cell counts are shown on the far right.

VI. DISCUSSION

Overall, the proposed cell detection method effectively identified cells in histopathological breast cancer images. Although the error rate for malignant cells (4.95%) was slightly higher than for benign cells (1.13%), this difference was not statistically significant. Moreover, the overall error rate (1.57%) was much lower than known histopathological cell identification error rates (up to 26%) for widely available packages, such as ImageJ and Farsight [17].

One potential limitation of these findings is the dataset used to evaluate the error rate. The current study uses a limited dataset of 12 images, of which only 6 represent malignant cell slides. Although the error rate for these images is quite low, a larger dataset would provide a more rigorous evaluation of the algorithm's effectiveness. Future iterations of the study should test the cell detection method on a more substantial dataset, including more images and different types of cancerous tissue.

Moreover, the error rates reported do not necessarily reflect the true sensitivity and specificity. In other words, as evidenced in the rightmost panel of fig. 4. and fig.5, the watershed algorithm occasionally misidentified sub-sections of a cell as a whole cell, erroneously increasing the cell count (false positive), or failed to identify a cell as such (false negative). A more robust analysis might include the sensitivity and specificity, although the lack of centroid coordinates for each cell in the ground truth dataset increases the difficulty of this approach.

Another potential avenue might be to analyze pixel concordance; in other words, calculating the percentage of pixels in the generated binary image that were correctly identified as cell mass. This approach would require comparing each pixel in the generated image to the corresponding image in the ground truth dataset. One difficulty with this analysis, however, is that the cropped images of the provided histological slides and ground truth binary images were not necessarily cropped at the same pixel coordinates. These discrepancies in the dataset cropping could be fixed algorithmically by locating the optimal bounding box and cropping both the original and binary ground truth images using identical coordinates. However, this approach was beyond the scope of the current research.

Another limitation of the current study is the application of the watershed method, which tended to overly segment cell bodies. For example, there are many many cell bodies that were not labeled as objects in the image because they were overly segmented (fig. 5, rightmost panel). This over-segmentation of cells caused them to appear small enough that they did not pass through the secondary threshold, which removed artifacts too small to plausibly be cells at the given magnification. Over-segmentation and sensitivity to false edges is a known problem with the watershed algorithm.

Future iterations of the present research might consider a two-step modified watershed algorithm to identify cells. In this process, an initial pass at object identification is made using k-means clustering, and automated thresholding is then used in conjunction with the original watershed algorithm for more fine-grained object identification [18]. This approach has enjoyed notable success with medical images, and could provide an avenue for significant improvements to the present breast cancer cell detection algorithm.

Lastly, although the current study focused on cell detection and cell count, future research could also benefit from investigating cell morphology. Given cell and nucleus outlines, a future algorithm could look at the shape and size of cell walls and nuclei, as well as detect mitosis in cells, factors that are highly relevant when calculating tumor proliferation and breast cancer severity scores [19].

When taken in conjunction, the present findings and current research suggest that algorithmic image analysis may provide a superior alternative to the current standard of manual cell count by a trained pathologist. Indeed, automated cellular detection in histopathological breast cancer images paves the way for improved prognostics and therapeutic decision-making by providing a reliable, infinitely scalable tool for physicians. Such an approach is well-poised to help advance the field of breast cancer medicine and improve breast cancer outcomes.

REFERENCES

[1] American Cancer Society. Global Cancer Facts & Figures 3rd Edition. Atlanta: American Cancer Society; 2015.

[2] Tao, Z., Shi, A., Lu, C., Song, T., Zhang, Z., & Zhao, J. (2015). Breast cancer: epidemiology and etiology. Cell biochemistry and

biophysics, 72(2), 333-338.

[3] Siegel, R., Ma, J., Zou, Z., & Jemal, A. (2014). Cancer statistics, 2014. CA: a cancer journal for clinicians, 64(1), 9-29.

[4] F. Tavassoli, P. Devilee Pathology and genetics tumours of the breast and female genital organ International Agency for Research on Cancer Press (2003)

[5] Demir C, Yener B. Automated cancer diagnosis based on histopathological images: a systematic survey. Technical Report, Rensselaer Polytechnic Institute; 2005.

[6] Lu C, Mandal M. Toward automatic mitotic cell detection and segmentation in multispectral histopathological images. IEEE J. Biomed. Health Inform. 2014 Mar;18(2):594–605.

[7] Al-Kofahi Y, Lassoued W, Lee W, Roysam B. Improved automatic detection and segmentation of cell nuclei in histopathology images. IEEE Trans. Biomed. Eng. 2010
Apr;57(4):841–852.

[8] Qi X, Xing F, Foran DJ, Yang L. Robust segmentation of overlapping cells in histopathology specimens using parallel seed detection and repulsive level set. IEEE Trans. Biomed. Eng. 2012 Mar;59(3):754–765.

[9] Xing, F., & Yang, L. (2016). Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review. *IEEE reviews in biomedical engineering*, *9*, 234-263.

[10] Gelasca, E. D., Obara, B., Fedorov, D., Kvilekval, K., & Manjunath, B. S. (2009). A biosegmentation benchmark for evaluation of bioimage analysis methods. *BMC bioinformatics*, *10*(1), 368.

[11] Walt, S. V. D., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. Computing in Science & Engineering, 13(2), 22-30.

[12] Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., ... & Yu, T. (2014). scikit-image: image processing in Python. PeerJ, 2, e453.

[13] Jones, E., Oliphant, T., & Peterson, P. (2014). {SciPy}: open source scientific tools for {Python}.

[14] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing In Science & Engineering, 9(3), 90-95.

[15] Otsu, N. (1979). A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics, 9(1), 62-66.

[16] Vincent, L., & Soille, P. (1991). Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Transactions on Pattern Analysis & Machine Intelligence, (6), 583-598.

[17] Schmitz, C., Eastwood, B. S., Tappan, S. J., Glaser, J. R., Peterson, D. A., & Hof, P. R. (2014). Current automated 3D cell detection methods are not a suitable replacement for manual stereologic cell counting. Frontiers in neuroanatomy, 8.

[18] Ng, H. P., Ong, S. H., Foong, K. W. C., Goh, P. S., & Nowinski, W. L. (2006, March). Medical image segmentation using k-means clustering and improved watershed algorithm. In Image Analysis and Interpretation, 2006 IEEE Southwest Symposium on (pp. 61-65). IEEE.

[19] Elston, C. W., & Ellis, I. O. (1991). Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long‐term follow‐up. Histopathology, 19(5), 403-410.