Expectation Maximization driven Geodesic Active Contour with Overlap Resolution (EMaGACOR): Application to Lymphocyte Segmentation on Breast Cancer Histopathology

Hussain Fatakdawala*, Ajay Basavanhally*, Jun Xu*, Gyan Bhanot†, Shridar Ganesan†,
Michael Feldman‡, John Tomaszewski‡ and Anant Madabhushi*†
*Dept. of Biomedical Engineering, Rutgers University, Piscataway, NJ, USA 08854
Email: hussainf@eden.rutgers.edu, anantm@rci.rutgers.edu
†The Cancer Institute of New Jersey, NJ, USA 08903, Email: ganesash@umdnj.edu
‡Dept. of Surgical Pathology, Philadelphia, PA, USA 19104.

Abstract

The presence of lymphocytic infiltration (LI) has been correlated with nodal metastasis and tumor recurrence in HER2+ breast cancer (BC), making it important to study LI. The ability to detect and quantify extent of LI could serve as an image based prognostic tool for HER2+ BC patients. Lymphocyte segmentation in H & E-stained BC histopathology images is, however, complicated due to the similarity in appearance between lymphocyte nuclei and cancer nuclei. Additional challenges include biological variability, histological artifacts, and high prevalence of overlapping objects. Although active contours are widely employed in segmentation, they are limited in their ability to segment overlapping objects. In this paper, we propose a segmentation scheme (EMaGACOR) that integrates Expectation Maximization (EM) based segmentation with a geodesic active contour (GAC). Additionally, a novel heuristic edge-path algorithm exploits the size of lymphocytes to split contours that enclose overlapping objects. For a total of 62 HER2+ breast biopsy images, EMaGACOR was found to have a detection sensitivity of over 90% and a positive predictive value of over 78%. By comparison, EMaGAC (model without overlap resolution) and GAC (Randomly initialized geodesic active contour) model yielded corresponding sensitivities of 57.4% and 26.7%, respectively. Furthermore, EMaGACOR was able to resolve over 92% of overlaps. Our scheme was found to be robust, reproducible, accurate, and could potentially be applied to other biomedical image segmentation applications.

1. Introduction

Breast cancer (BC) is the most common cancer diagnosis in women in the United States with an estimated incidence of 180,000 and mortality of over 40,000 in 2008 [1]. The diagnosis and prognosis of BC is typically based on examination of breast biopsy tissue specimens by a pathologist who attempts to identify image derived features to recognize patterns and changes in phenotypes that are characteristic of BC. Certain kinds of phenotypic changes in tissue pathology, such as lymphocytic infiltration (LI), may be related to patient survival and outcome and may aid in prescribing appropriate therapy [2]. LI has been correlated with nodal metastasis and tumor recurrence in HER2+ BC, thus making it necessary to detect and quantify lymphocyte patterns in BC histopathology [2].

The visual detection of lymphocytes in histopathology images is complicated due to the similarity in appearance between lymphocyte nuclei and cancer nuclei. These two classes of nuclei are often confused with each another during manual segmentation, which may introduce error and negatively affect consistency in determining extent of LI. This in turn implies that a clinician's ability to predict survival and disease outcome will be affected by inter- and intra-clinician variability. Hence, there exists the need for accurate automated detection of lymphocyte nuclei in BC histopathology images that involves minimal manual intervention. Additional challenges in segmentation include the variability between images due to inconsistencies in histological staining, fixation and digitization procedures. Furthermore, LI is characterized by a high density of lymphocytes, which makes overlap among lymphocyte nuclei and other structures highly prevalent in BC images. Hence, the ability to accurately estimate the true extent of LI is contingent on an algorithm's ability to resolve such overlaps.

2. Previous related work

Manual detection and semi-automated segmentation of nuclei and glands were employed in [3] to distinguish low and high grades of BC using textural and nuclear architectural features.

Segmentation of structures in breast histopathology images have been attempted using fuzzy c-means clustering [4] and adaptive thresholding [5]. Thresholding tends to work only on uniform images and does not produce consistent results if there is considerable variability across image sets. Watershed algorithms [6] tend to be limited by the same problem and require prominent 'necks' to segment adjacent or overlapping objects. Active contours are widely used in image segmentation [7], however they are often unable to resolve multiple objects segmented as a single object and inclusion of other irrelevant objects from the background further detracts from the final result. Contour models alone are insufficient in providing satifactory segmentation as random initialization limits their ability to target specific objects of interest. More recently, probabilistic models have been employed to drive segmentation models [8], [9] but require manual training that limits automation and application. For example, a Bayesian classifier in conjunction with a template matching scheme was used to grade prostate and breast cancer histopathology [8] and distinguish LI in HER2+ BC [9]. A supervised classifier requires manual training and a representative annotated sample data set to model the entire data. Data sets for training are difficult to define due to variability across images. Furthermore, such models

978-0-7695-3656-9/09 $25.00 © 2009 IEEE
DOI 10.1109/BIBE.2009.75

may not be generalizable and have limited application. The limitations in the applicability of the aforementioned segmentation methods for detecting lymphocyte nuclei in BC histopathology images include,

- variability in image sets due to artifacts from staining, fixation and digitization,
- inability to resolve overlap between objects,
- limitations in obtaining representative annotated sample data sets for training probabilistic models.

3. Novel contributions of this work

In this paper, we present a new segmentation scheme — Expectation Maximization driven Geodesic Active Contour with Overlap Resolution (EMaGACOR) — for detecting lymphocyte nuclei in BC histopathology images. (Figure 1). EMaGACOR is able to overcome the drawbacks associated with probabilistic segmentation schemes, namely the need for annotated training data and defining representative data sets. We attempt to avoid these issues by using an Expectation Maximization (EM) algorithm to initialize a geodesic active contour model (GAC). In this paper, we use the Magnetostatic Active Contour (MAC) described in [7] as the GAC model. The EM algorithm effectively replaces the Bayesian framework and hence eliminates the need for representative data sets for training and reduces the effect of data set variability on segmentation results. Initialization of the active contour using EM allows the model to focus on relevant objects of interest. The EM algorithm provides an initial segmentation in the form a likelihood scene. The centroids of the objects detected in this scene are used as seed points to initialize the contour which enhances the performance of the active contour model in capturing target objects. As mentioned earlier, overlapping/touching objects are characteristic of our problem and we further process the contour result by splitting the contour between high concavity points. While Chetverikov et al. [10] and Yang et al. [11] have previously described related concavity detection algorithms, our methodology involves a modified approach involving the use of vectors to allow continuous concavity detection on closed contours. A novel contour splitting scheme is devised where the concavity points are connected by an *edge-path* algorithm that defines paths through relevant edge points within the contour while simultaneously ensuring an optimal split. The algorithm incrementally defines a path by including single edge points at a time to ensure that the split represents an edge or a potential overlap boundary. An intelligent heuristic rule based on object size is used to determine the need for a split. This enables us to distinguish between multiple objects identified as a single entity by the active contour. The last module of the system involves object class segregation based on texture based clustering in order to distinguish lymphocyte nuclei from other similar structures in the image.

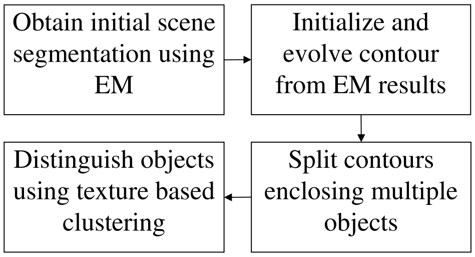
Obtain initial scene segmentation using EM
Initialize and evolve contour from EM results
Split contours enclosing multiple objects
Distinguish objects using texture based clustering

Figure 1: Flow chart depicting proposed EMaGACOR model

In summary, our segmentation method (EMaGACOR) improves lymphocyte detection in BC histopathology images by,

- eliminating the need for training data sets that are difficult to define due to image variability,
- enhancing the performance of active contours by providing an appropriate initialization via the EM algorithm,
- resolving the issue of contours enclosing overlapping/touching objects by splitting contours in favor of obtaining better object detection.

The rest of the paper is organized as follows: Section 4 describes the overall methodology of our segmentation model and Section 5 explains the experiments performed to validate the improvements in segmentation using our method. In Section 6 we discuss and summarize our results. Concluding remarks are presented in section 7.

4. Methods

4.1. Data Description and Notation

Hematoxylin and Eosin (H & E) stained breast biopsy cores were scanned into a computer using a high resolution whole slide scanner at 40x optical magnification at The Cancer Institute of New Jersey. A total of 62 images representing HER2+ breast cancer exhibiting LI were used in our analysis. The ground truth for detection of lymphocytes was obtained via manual detection performed by an expert from The Cancer Institute of New Jersey.

An image is defined as $\mathcal{C} = (C, f)$ where C is a 2D grid representing pixels $c \in C$, with $c = (x, y)$ representing the Cartesian coordinates of a pixel or a point and f is a function of c that assigns pixel values corresponding to intensities in R, G, and B channels.

4.2. EM based segmentation

The EM algorithm is used to determine the probability of each pixel c belonging to one of K classes, ω_k, $k \in \{1, 2, \ldots, K\}$ in an image scene. The EM algorithm attempts to identify the individual Gaussian distribution from a mixture of K normal class densities. For the application considered in this paper, we set $K = 3$ representing $\omega_k \in \{\text{lymphocyte nuclei, stroma, cancer nuclei}\}$. The EM algorithm will compute the posterior class conditional probability $P(\omega_k|f(c))$ of each pixel c belonging to class ω_k given the prior probability $p(f(c)|\omega_k)$. The algorithm is run iteratively, and comprises of two steps: the Expectation (E-step) and the Maximization step (M-step). The E-step calculates $P(\omega_k|f(c))$ based on the current parameters of Gaussian mixture model while the M-step recalculates or updates the model parameters at each iteration i, $\gamma_k^i = \{\mu_k^i, \Sigma_k^i, p_k^i\}$ where μ_k^i and Σ_k^i are the mean and covariance of each Gaussian component, respectively, and $p_k^i = p^i(f(c)|\omega_k)$, also referred to as the mixture coefficients in the Gaussian mixture model. After convergence, the EM algorithm will assign each pixel c, a $1 \times K$ probability vector whose elements are the respective $P(\omega_k|f(c))$ values. The implementation of the EM algorithm is summarized below [12]:

Parameters initialization: The initial parameters $\{\mu_k^0, \Sigma_k^0, p_k^0\}$ are randomly selected.

E-step: Calculate the posterior probabilities using the current parameters γ_k^i,

$$P(\omega_k|f(c)) = \frac{p_k^i \mathcal{N}(f(c)|\mu_k^i, \Sigma_k^i)}{\sum_{j=1}^{K} p_j^i \mathcal{N}(f(c)|\mu_j^i, \Sigma_j^i)},$$

where

$$\mathcal{N}(f(c)|\mu_k^i, \Sigma_k^i) =$$
$$(2\pi)^{-\frac{D}{2}}|\Sigma_k^{-\frac{1}{2}}| \exp\{-\frac{1}{2}(f(c) - \mu_k^i)^T \Sigma_k^{-1}(f(c) - \mu_k^i)\},$$

is a D dimensional Gaussian distribution. For intensities from an RGB image, D is set to 3.

M-step: The mean, covariance and the priori probability of each class are updated by the posterior probabilities obtained in E-step and are computed as follows,

$$\mu_k^{i+1} = \frac{1}{n_k} \sum_{c \in C}^{|C|} P(\omega_k|f(c)) f(c),$$

$$\Sigma_k^{i+1} = \frac{1}{n_k} \sum_{c \in C}^{|C|} P(\omega_k|f(c))(f(c) - \mu_k^{i+1})(f(c) - \mu_k^{i+1})^T,$$

$$p_k^{i+1} = \frac{n_k}{|C|},$$

where $n_k = \sum_{c \in C}^{|C|} P(\omega_k|f(c))$ and $|C|$ is the cardinality of C.

Convergence Evaluation: Convergence is evaluated by calculating the Euclidean distance of log likelihood between current and preceding iterations. Assuming that the Gaussian distribution for every pixel c within the image is independent of one another, based on Gaussian mixture model, the log likelihood function of the model with parameters obtained in *M-step* can be computed as follows,

$$\begin{aligned} L^{i+1}(\mathcal{C}|\mu, \sigma, p) &= \ln\{\sum_{k=1}^{K} p_k^{i+1} \mathcal{N}(f(C)|\mu_k^{i+1}, \Sigma_k^{i+1})\}, \\ &= \ln\{\sum_{k=1}^{K} p_k^{i+1} [\prod_{c \in C}^{|C|} \mathcal{N}(f(c)|\mu_k^{i+1}, \Sigma_k^{i+1})]\}, \\ &= \sum_{c \in C}^{|C|} \sum_{k=1}^{K} \ln\{p_k^{i+1} \mathcal{N}(f(c)|\mu_k^{i+1}, \Sigma_k^{i+1})\}, \end{aligned}$$

where L^{i+1} is the log likelihood estimation of $\mathcal{C}$ with Gaussian mixture model. The convergence criterion can be expressed as the following inequality:

$$\left\| \frac{L^{i+1} - L^i}{L^i} \right\| \leq \epsilon$$

where ϵ is an empirically determined threshold and $||\cdot||$ is the L2 norm. The termination of iterations is decided by the convergence criterion. In our experiments ϵ was set to 10^{-5}. If the convergence criterion is not reached, the algorithm returns to the E-step. Otherwise, the algorithm will return a probability matrix obtained from latest E-step which is used to group every pixel $c \in C$ in to K different classes.

Given the convergence of the EM algorithm $\forall c \in C$ as $P(\omega_k|f(c))$, construct K likelihood scenes $\mathcal{L}_j = (C, \ell_j)$, $j \in \{1, \ldots, K\}$, where $\ell_j(c) = P(\omega_k|f(c))$. For each $\mathcal{L}_j$, derive corresponding binarized scenes $\mathcal{L}_j^B = (C, \ell_j^B)$ where

$$\ell_j^B(c) = \begin{cases} 1, & \text{if } \ell_j = \max_k [P(\omega_k|f(c))] \\ 0, & \text{otherwise.} \end{cases} \qquad (1)$$

The binarized scene $\mathcal{L}_j^B$ represents the EM based segmentation for objects belonging to class $\omega_k \in$ {lymphocyte nuclei, stroma, cancer nuclei}. The appropriate scene for lymphocyte nuclei is identified and used to initialize the active contour model. A sample image with its corresponding binarized lymphocyte nuclei scene obtained via EM is shown in Figure 2.

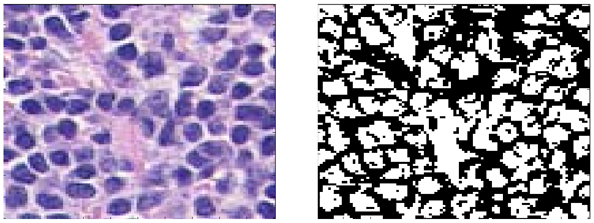
(a) (b)

Figure 2: (a) Original BC histopathology image with corresponding (b) binarized lymphocyte nuclei scene. The EM algorithm provides the probability for each pixel c belonging to class ω_k given f. The maximum posterior class conditional probability is used to assign c to class ω_k that generates the corresponding binarized scene $\mathcal{L}_j^B, j \in \{1, \ldots, K\}$. The appropriate $\mathcal{L}_j^B$ corresponding to lymphocytes is visually identified.

4.3. EM driven Geodesic Active Contour Model (EMaGAC)

An active contour model is used in our boundary segmentation problem where a set of contours S is evolved using level set method [13] to find target object boundaries. S is composed of numerous closed sub-contours $s \in S$, the number of which depends upon the initialization and the final evolved result. Each closed sub-contour s is composed of an ordered set of M points such that $s = \{c_w | w \in \{1, \ldots, M\}\}$ where each point c_w is connected to only to two of its adjacent points c_{w-1} and c_{w+1} with $c_{M+1} = c_1$ and $c_0 = c_M$ to form a closed loop. The entire set of contours S is evolved in time t over a 2D image $\mathcal{C}$. The contour is represented as a zero level set $s = \{c|\phi(t, c) = 0\}$ of a level set function ϕ. In general, the evolution of the level set function ϕ is accomplished by an iterative solution to the following PDE:

$$\frac{\partial \phi}{\partial t} + \mathbf{f}|\nabla \phi| = \mathbf{0}$$

where $\mathbf{f}$ is the speed function for evolution that drives the contour s towards the object boundary. $\mathbf{f}$ is unique to an active contour model and its design depends on how the information in the image is used. In this paper, a specific initialization is provided to the magnetostatic active contour (MAC) model described in [7].

4.3.1. Magnetostatic active contour (MAC) model. The MAC model implements a bidirectional force field F generated from a hypothetical magnetostatic interaction between the contour s and the object boundary. The force field F drives the contour towards the boundary. Both the boundary and the contour are treated as current carrying loops and the magnetic field from the boundary computed using the well known Biôt-Savart law determines the force F acting on the contour. F is defined over the entire image $\mathcal{C}$. The level set implementation of the contour as proposed in [7] takes the form:

$$\frac{\partial \phi}{\partial t} = \alpha g(\mathcal{C}) \nabla \cdot \left(\frac{\nabla \phi}{|\nabla \phi|} \right) |\nabla \phi| - (1 - \alpha) F(\mathcal{C}) \cdot \nabla \phi$$

where α is a real constant, $g(\mathcal{C}) = 1/(1 + |\nabla \mathcal{C}|)$ and $\nabla(\cdot)$ represents the 2D gradient $\left(\frac{\partial(\cdot)}{\partial X}, \frac{\partial(\cdot)}{\partial Y} \right)$ along the X and Y axes.

4.3.2. Initialization scheme. The EMaGAC and the EMaGACOR model use EM derived segmentation to provide a specific initialization to the GAC contour. For a given image $\mathcal{C}$, the corresponding lymphocyte nuclei segmented scene $\mathcal{L}_j^B$ is manually selected over all k. For all objects O_ϱ, where $\varrho \in \{1, \ldots, \Omega\}$ detected in $\mathcal{L}_j^B$ via connected component labeling, centroids $q_\varrho = \frac{1}{|O_\varrho|} \sum\limits_{c \in O_\varrho} c$ are computed that serve as seeds points for initializing the GAC (Figure 6(a)). The initial contour $\phi_0 = \phi(0, c)$, is defined as a circle of radius r centered at each q_ϱ (Figure 6(b)). r is chosen to be approximately half the radius of the target object (empirically estimated). The contour is then evolved till the differences in the contours of the current iteration (ϕ^t) to the next (ϕ^{t+1}) are below an empirically determined threshold ρ. The algorithm for initializing the contour model is illustrated in Figure 3. The active contour provides a segmentation that focuses on lymphocyte nuclei and cancer nuclei (Figure 6 (c)). Note that various contours contain two or more overlapping/touching objects. Contours enclosing multiple objects are processed in the next step where we explain a contour splitting scheme based on concavity detection, the *edge-path* algorithm and lymphocyte nuclei size heuristic.

```
Input : Image scene $\mathcal{C}$, number of classes $K$, radius $r$,
    thresholds $\epsilon$ and $\rho$, time step $\Delta t$ and speed function $\mathbf{f}$.
Output : Final evolved contour.
  BEGIN
    Randomly initialize $\gamma^0 = \{\mu_k^0, \Sigma_k^0, p_k^0\}$, compute $L^0$,
      while $\left\| \frac{L_{i+1} - L_i}{L_i} \right\| > \epsilon$,
        calculate $P(\omega_k | f(c))$,
        update $\gamma_{i+1}$,
        compute $L_{i+1}$,
      end
    Obtain $K$ class likelihood scenes $\mathcal{L}_k = (C, \ell_k)$,
      where $\ell_k(c) = P(\omega_k | f(c))$,
    Binarize $\mathcal{L}_k$ to obtain $\mathcal{L}_j^B$ (Equation 1),
    Select target class scene $\mathcal{L}_j^B$,
    Determine all objects $O_\varrho$ by connected component labeling,
    Obtain corresponding centroids $q_\varrho = \frac{1}{|O_\varrho|} \sum_{c \in O_\varrho} c$,
    Define initial contour $\phi_0 = \phi(0, c)$ as circle of radius $r$,
      centered at each $q_\varrho$,
      while $||\phi^{t+1} - \phi^t|| > \rho$,
        Evolve contour, $\phi^{t+1} = \phi^t + [\mathbf{f}(\nabla \phi^t)](\Delta t)$,
      end
  END
```

Figure 3: EMaGAC algorithm

```
Input : Sub contour $s$, threshold $\theta_{\max}$, object size $\tau_A$,
Output : Concavity points $V_s$.
  BEGIN
    for all $s \in S$,
      compute $\mathcal{A}(s)$,
        if $\mathcal{A}(s) > \tau_A$,
          for all points $c_w \in s$,
            compute vectors $(c_w - c_{w-1})$ and $(c_{w+1} - c_w)$,
            compute $\theta(c_w)$ between vectors,
            if $\theta(c_w) \leq \theta_{\max}$ and cross product $\geq 0$, save $c_w \rightarrow V_s$,
          end
        end
    end
  END
```

Figure 4: Concavity detection algorithm

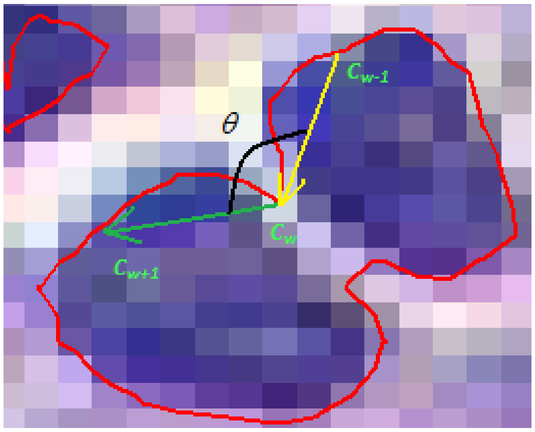

Figure 5: Concavity detection: Three consecutive points on s (c_{w-1}, c_w and c_{w+1}) are used to define two vectors (shown with arrows). The angle $\theta(c_w)$ between them is a measure of concavity/convexity of the point c_w. Concavity points can be distinguished from convex points by computing the cross product between the vectors where a positive cross product would indicate a concavity point if moving in a counter-clockwise direction on s.

4.4. Resolving overlap - EMaGACOR model

In addition to providing a specific initialization to GAC via EM, the EMaGACOR model aims at improving segmentation by providing overlap resolution where contours enclosing multiple objects are split using a size heuristic. A concavity detection scheme [10], [11] is employed to obtain high concavity points on contours that serve as an input to our *edge-path* algorithm to define a split. High concavity points are characteristic of contours that enclose multiple objects and represent junctions where object intersection occurs (Figure 5).

4.4.1. Concavity detection. The area $\mathcal{A}(s)$ of the closed sub-contour s is compared to predetermined area of an ideal lymphocyte nuclei τ_A. For our experiments τ_A was set to 35. Hence a sub-contour is eligible for split if $\mathcal{A}(s) > \tau_A$. Since $c = (x, y)$, the difference between any two points c_w

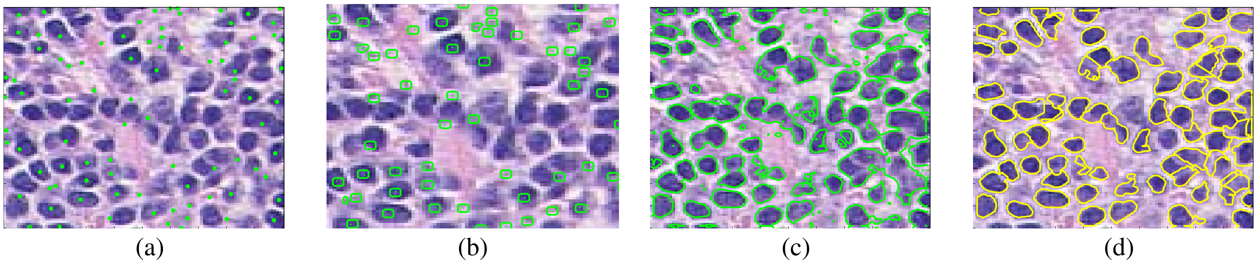

Figure 6: (a) Centroids q_ϱ (green) obtained from EM derived binarized scene $\mathcal{L}_k^B$ that serve as seed points for initializing the active contour S (b) initialized S (GAC). Initialization is defined as circles centered at each q_ϱ (c) Contour result after evolution (d) Improved segmentation after splitting contours enclosing multiple objects by using the *edge-path* algorithm.

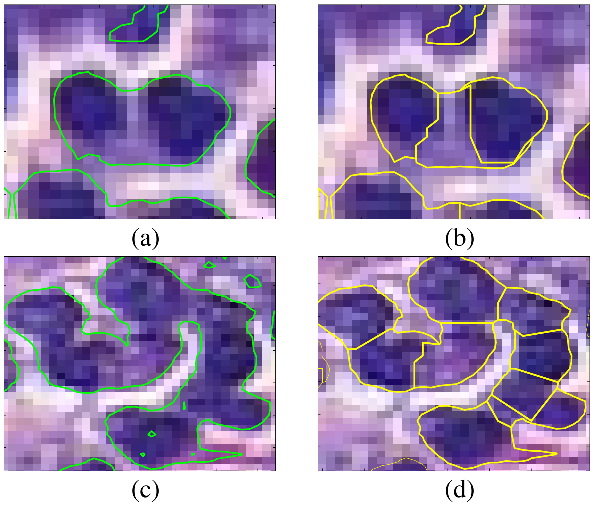

Figure 7: (a), (c) Examples of a contour enclosing overlapping lymphocytes. Lack of edges/weak edges prevent the contour from providing accurate object segmentations (b), (d) Contour split by *edge-path* algorithm using size heuristic.

and c_{w-1} will represent a vector in 2D. Concavity points are detected by computing the angle between vectors defined by three consecutive points $(c_{w-1}, c_w, c_{w+1}) \in s$. The degree of concavity/convexity is proportional to the angle $\theta(c_w)$ as shown in Figure 5. $\theta(c_w)$ can be computed from the dot product relation as shown below:

$$\theta(c_w) = \pi - \arccos\left(\frac{(c_w - c_{w-1}) \cdot \ (c_{w+1} - c_w)}{||(c_w - c_{w-1})|| \ \ ||(c_{w+1} - c_w)||}\right).$$

Concavity points can be distinguished from convexity points by computing the cross product of the vectors $(c_w - c_{w-1})$ and $(c_{w+1} - c_w)$ where a concavity point would yield a positive cross product if moving in a counter-clockwise direction on s (Figure 5). The value of $\theta(c_w)$ for an eligible concavity point c_w is constrained to be less than an empirically determined value $\theta_{\max}$. The value of $\theta_{\max}$ serves as a threshold for detecting meaningful concavity points and in our case it was found that $\theta_{\max} = \frac{8}{9}\pi$ yielded optimal results. The concavity detection algorithm is summarized in Figure 4.

4.4.2. Edge-path algorithm. The algorithm defines a path between a pair of concavity points through edge points enclosed within s. From all the possible paths between various concavity points, the shortest path that satisfies a split yielding a sub-contour whose size is close to that of a lymphocyte nuclei is favored. Let V_s be the set of N concavity points detected on s such that $V_s = \{c_m | m \in \{1, \ldots, N\}\}$ and $N \leq M$. Let E represent the set of H edge points enclosed by s such that $E = \{c_u | u \in \{1, \ldots, H\}\}$. For a given pair of concavity points $\{c_a, c_b\} \in V_s$ and $a \neq b$, $a, b \in \{1, \ldots, N\}$, the path Q_{ab} between them is defined through specific h number of ordered edge points in E, such that $Q_{ab} = \{c_a, c_1, \ldots, c_h, c_b\}$ is an ordered set with each of its points connected to only to two of its adjacent points and satisfies the following condition:

$$||c_a - c_b|| \geq ||c_1 - c_b|| \geq \ldots \geq ||c_h - c_b|| \geq 0$$

and there does not exist c_{h+1} such that $||c_{h+1} - c_b|| \leq ||c_h - c_b||$. Hence the path Q_{ab} is defined in an incremental

```
Algorithm : Edge-path
Input : Sub-contour $s$, object size $\tau_A$,
        Edge points $c_u \in E$, Concavity points $V_s$,
Output : Final path $Q_{ab}$ to split $s$ into $s_1$ and $s_2$.
BEGIN
  if $|V_s| > 2$,
    set $\mathcal{D} = \infty$,
    compute $\psi = \frac{\mathcal{A}(s) - \tau_A}{\tau_A}$,
      for each pair of concavity points $\{c_a, c_b\} \in V_s, a \neq b$,
        set $c_A = c_a$, set $e = E$, mark $c_b$ as edge point,
        set $Q_{ab} = \{c_a\}$,
        while $c_A \neq c_b$,
          find closest edge point $c_u \in e$ to $c_A$,
            if $||c_A - c_b|| \geq ||c_u - c_b||$,
              save $c_u \rightarrow Q_{ab}$,
              set $c_A = c_u$,
            end
            remove $c_u$ from $e$,
        end
        Split $s$ in to $s_1$ and $s_2$ using $Q_{ab}$,
        compute $\Gamma = \frac{\mathcal{A}(s_1)}{\mathcal{A}(s_2)}$,
          if $\psi^{-1} \leq \Gamma \leq \psi$ and $||Q_{ab}|| < \mathcal{D}$,
            accept split and reject any previous split,
            set $\mathcal{D} = ||Q_{ab}||$,
          end
      end
  end
END
```

Figure 8: *Edge-path* algorithm

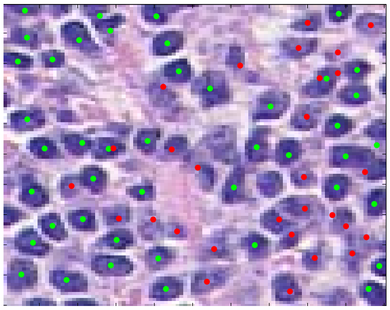

Figure 9: Final detection result (green dots are lymphocyte nuclei). The result is obtained by clustering textural features (average and standard deviation in intensity) extracted from final contour result into 2 groups. Red dots refer to objects identified as non-lymphocytes

manner to ensures that relevant edges between any two given concavity points $\{c_a, c_b\}$ are included while simultaneously guaranteeing a path between them. Once a path Q_{ab} is defined, it splits s in to two sub-contours - s_1 and s_2. The area of the split contours $\mathcal{A}(s_1)$ and $\mathcal{A}(s_2)$ are computed and compared to the predetermined size of an ideal lymphocyte τ_A. The ratio of the areas of the split contours $\Gamma = \frac{\mathcal{A}(s_1)}{\mathcal{A}(s_2)}$ are constrained to a threshold ψ described below:

$$\psi = \frac{\mathcal{A}(s) - \tau_A}{\tau_A}$$

The path Q_{ab} that splits s is accepted if the area ratio Γ satisfies the condition:

$$\psi^{-1} \leq \Gamma \leq \psi \qquad (2)$$

The condition in Eqaution 2 favors the size heuristic. Of all the paths between various pairs of concavity points that satisfy this condition (Equation 2), the shortest path is selected to define the split. The process of splitting contours is repeated till all contours have an area comparable to τ_A. The final contour segmentation is shown in Figure 6 (d) and is used to extract textural information to distinguish lymphocyte nuclei from other object classes. Note that majority of contours enclosing multiple objects have been split via paths representing weak edges or potential overlap boundaries. Detailed splitting of such contours is illustrated in Figure 7. Note that lack of edge information does not impede the algorithm in segregating overlapping objects. The algorithm is summarized in Figure 8.

4.5. Texture based clustering to detect lymphocytes

The final contour result is used to distinguish lymphocytes from other objects in the image via clustering of simple textural features using the k-means algorithm. The standard deviation σ and average κ of intensity from three channels (R, G and B) is computed for the region enclosed by a contour. Thus each candidate object O_ζ, $\zeta \in \{1, \ldots, T\}$, is described by a 6 dimensional attribute vector $\mathbf{F}(O_\zeta)$, comprised of σ and κ values within O_ζ for each of the three color channels. The popular k-means algorithm is then used to distinguish all O_ζ in to one of 2 object classes based on $\mathbf{F}(O_\zeta)$. The motivation behind this final step is to distinguish lymphocytes from other objects based on their textural attributes. Figure 9 illustrates the final result showing the detected lymphocytes by obtaining the centroids of the closed contours s (green dots).

5. Experiments

A total of 62 images were analyzed using the proposed EMaGACOR model. We compare segmentation results on these images from randomly initialized active contour (GAC) against EMaGAC and EMaGACOR. Quantitative and qualitative comparison of detection results from these three models were performed.

6. Results and Discussion

6.1. Qualitative results

Qualitative results for 3 of the 62 different studies are illustrated in Figure 10 and indicate the superiority of EMaGACOR over the EMaGAC and GAC models respectively in segmenting lymphocytes. Note how initialization from the EM algorithm allows the contour to focus on objects of interest and prevents the contour from enclosing numerous objects (Figures 10(g)-(i)). In addition, the concavity detection scheme and the *edge-path* algorithm provide overlap resolution by splitting contours enclosing multiple overlapping/touching objects and improve segmentation results (Figures 10 (j)-(l)). The experimental trials aim at portraying the difficulty in detecting lymphocytes in BC images by an active contour alone and how the EMaGACOR model is successful in tackling this problem.

6.2. Quantitative results

Quantitative results are summarized in Table 1 which again are indicative of the improved performance of EMaGACOR over GAC and EMaGAC. The active contour model is limited in its ability to segment objects of interest (Figures 10(d)-(f)). The initialization of the GAC contour using EM results proves to be useful in targeting lymphocytes and constraining the contour from erroneously evolving in to large contours enclosing multiple objects. While a Bayesian framework may be used for initialization, the EM algorithm proves to be a better option as it improves automation by replacing the need for annotated training. As discussed earlier, representative data sets for training are difficult to define owing to variability across images. The contour results obtained from EMaGAC (Figures 10(g)-(i)) are clearly sub-optimal in that several overlapping/touching objects are identified as single objects. The concavity detection scheme followed by splitting of contours by the *edge-path* algorithm greatly improves segmentation accuracy in relation to overlapping objects (EMaGACOR). Quantitative evaluation of the three models was done via (a) statistical measurements and by computing the (b) overlap detection ratio (OR).

6.2.1. Statistical measurements. The statistical measurements used to evaluate the experimental results include the sensitivity (SN) and the positive predictive value (PPV) that are computed for each of the three models in our experiment (Table 1) in terms of lymphocyte detection. These values are computed from the true positive (TP), false positive (FP) and false negative (FN) values. TP refers to the number

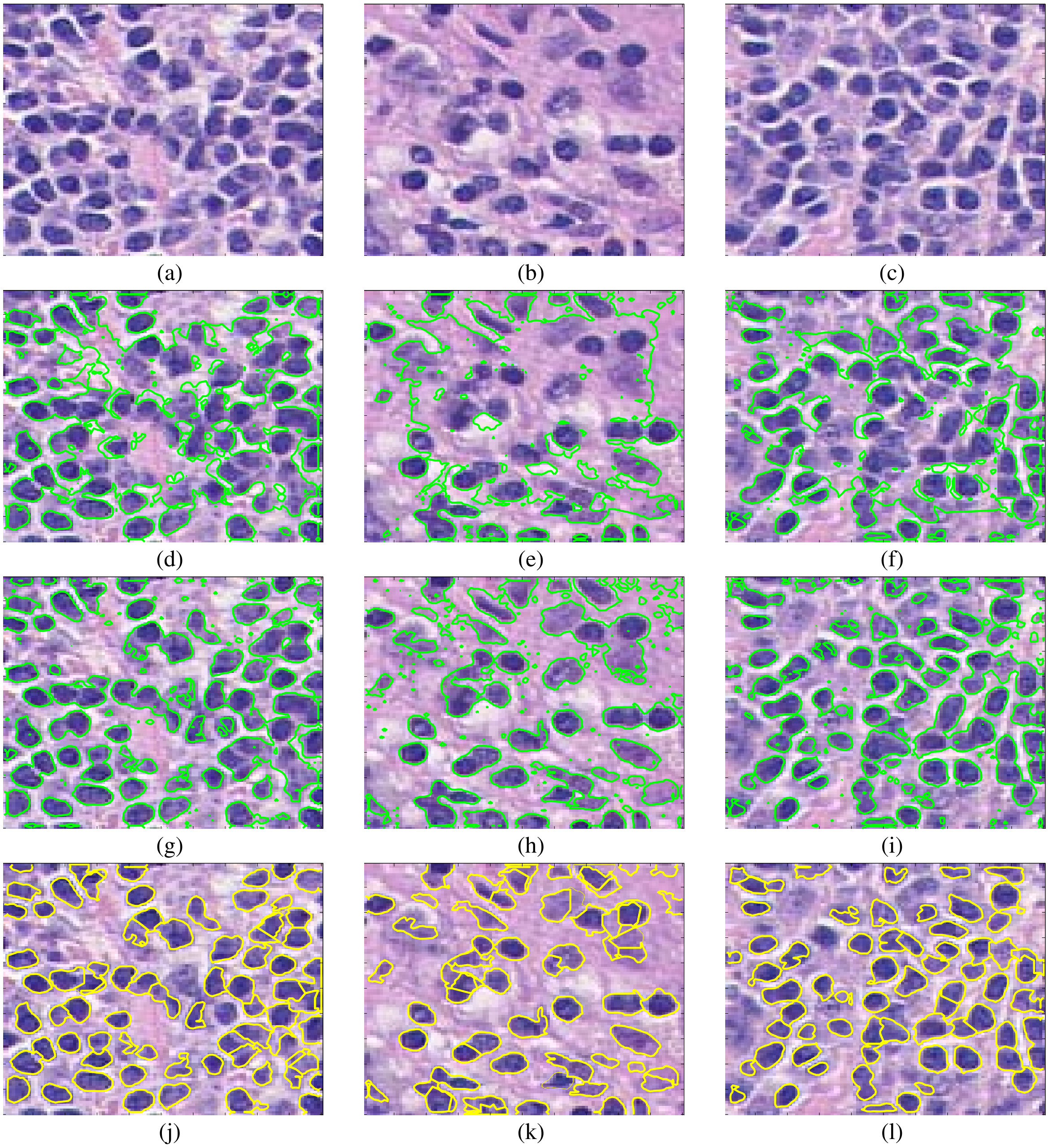

Figure 10: Qualitative results: (a)-(c) Original BC histopathology image Lymphocyte segmentation result from (d)-(f) randomly initialized GAC (g)-(i) EMaGAC, and (j)-(l) after contour splitting using concavity detection and *edge-path* algorithm (EMaGACOR).

of lymphocytes correctly identified while FP refers to the number of objects incorrectly identified as lymphocytes and FN refers to the number of lymphocytes missed by each of the EMaGACOR, EMaGAC and GAC models respectively. The SN and PPV values listed in Table 1 reflect the efficacy of the EMaGACOR model in detecting lymphocytes in BC images as compared to GAC and EMaGAC models.

6.2.2. Overlap detection ratio (OR). The overlap detection ratio (OR) (Table 1) is computed as follows:

$$OR = \frac{Number\ of\ Overlaps\ resolved}{Total\ number\ of\ overlaps}$$

An overlap is characterized by the existence of a common boundary between two objects and in our case may be between two or more lymphocyte nuclei, cancer nuclei or both. A total of 704 cases of overlapping objects were manually identified in 62 images and our model was able to resolve 651 (92.5%) overlaps. The improvement in OR value is indicative of the enhanced overlap resolution provided by the EMaGACOR model using the *edge-path* algorithm.

Table 1: Results of quantitative evaluation between EMaGACOR, EMaGAC and GAC models over 62 images in terms of detection SN, PPV and OR.

	SN	PPV	OR
GAC	26.7	70.8	1.7
EMaGAC	57.4	76.3	10.8
EMaGACOR	90.6	78.1	92.5

6.2.3. Test of statistical significance between models. SN, PPV and OR values were compared for every pair of models using the paired t-test under the null hypothesis that there is no significant difference in these values between the EMaGACOR, EMaGAC and GAC models respectively. Each of the t-tests above returned a p-value $\leq$ 0.05 (Table 2) indicating that improvements in detection results due to the EMaGaCOR model when compared to EMaGAC and GAC models are statistically significant.

Table 2: p-values of t-test between EMaGACOR, EMaGAC and GAC models for SN, PPV and OR over 62 images.

	SN	PPV	OR
GAC/EMaGAC	1.1×10^{-26}	2.3×10^{-4}	6.0×10^{-7}
GAC/EMaGACOR	6.8×10^{-43}	1.4×10^{-6}	1.6×10^{-58}
EMaGAC/EMaGACOR	2.9×10^{-27}	3.0×10^{-2}	9.9×10^{-44}

7. Conclusion

In this paper, we have presented a new segmentation model for detection of lymphocytes in HER2+ BC histopathology images. Our segmentation scheme overcomes a number of issues that plague popular segmentation schemes. Specifically, our model is able to overcome detection sensitivity associated with random initialization and allows resolving object overlap. In addition, the scheme differs from supervised classifier detection methods that are encumbered by the need for a large number of annotated training samples. The initialization of the contour from EM results (1) enhances contour performance, (2) eliminates the need for training and (3) improves automation. The concavity detection scheme in conjunction with the *edge-path* algorithm effectively resolves the issue of segmenting overlapping objects. Experimental results have shown that our new EMaGACOR model performs significantly better compared to EMaGAC and GAC models over a total of 62 images.

The ability to accurately and automatically segment lymphocytes in BC histopathology images may serve to be useful in studying LI and its relation to BC prognosis. Future work will focus on quantifying appropriate LI features to explore the potential of LI as a prognostic tool.

Acknowledgment

This work was made possible due to grants from The Coulter Foundation, Society for Imaging Informatics in Medicine (SIIM) award, the Life Sciences Commercialization Award, New Jersey Commission on Cancer Research, and The National Cancer Institute (R21-CA127186-01, R03-CA128081-01) and the Aresty Research Grant.

References

[1] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray, and M. J. Thun, "Cancer statistics, 2008." *CA Cancer J Clin*, vol. 58, pp. 71–96, 2008.

[2] G. Alexe, G. S. Dalgin, D. Scanfeld, P. Tamayo, J. P. Mesirov, C. DeLisi, L. Harris, N. Barnard, M. Martel, A. J. Levine, S. Ganesan, and G. Bhanot, "High expression of lymphocyte-associated genes in node-negative her2+ breast cancers correlates with lower recurrence rates." *Cancer Res*, vol. 67, pp. 10 669–10 676, 2007.

[3] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," in *Proc. 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro ISBI 2008*, 14–17 May 2008, pp. 496–499.

[4] L. Latson, B. Sebek, and K. A. Powell, "Automated cell nuclear segmentation in color images of hematoxylin and eosin-stained breast biopsy." *Anal Quant Cytol Histol*, vol. 25, pp. 321–331, 2003.

[5] S. Petushi, F. U. Garcia, M. M. Haber, C. Katsinis, and A. Tozeren, "Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer." *BMC Med Imaging*, vol. 6, p. 14, 2006.

[6] S. Essafi, R. Doughri, S. M'hiri, K. B. Romdhane, and F. Ghorbel, "Segmentation and classification of breast cancer cells in histological images." in *Information and Communication Technologies*, 2006.

[7] X. Xie and M. Mirmehdi, "Mac: magnetostatic active contour model." *IEEE Trans Pattern Anal Mach Intell*, vol. 30, pp. 632–646, 2008.

[8] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology," in *Proc. 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro ISBI 2008*, 2008, pp. 284–287.

[9] A. Basavanhally, S. Agner, G. Alexe, G. Bhanot, S. Ganesan, and A. Madabhushi, "Manifold learning with graph-based features for identifying extent of lymphocytic infiltration from high grade, her2+ breast cancer histology," in *MIAAB*, 2008.

[10] D. Chetverikov and Z. Szabo, "A simple and efficient algorithm for detection of high curvature points in planar curves." in *The 23rd Workshop of the Austrian Pattern Recognition Group*, 1999.

[11] L. Yang, O. Tuzel, P. Meer, and D. J. Foran, "Automatic image analysis of histopathology specimens using concave vertex graph." *Med Image Comput Comput Assist Interv Int Conf Med Image Comput Assist Interv*, vol. 11, pp. 833–841, 2008.

[12] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer–Verlag, 2006.

[13] J. Sethian, *Level Set Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science.* Cambridge Univ. Press, 1996.