21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

A Gamma-Gaussian Mixture Model for Detection of Mitotic Cells in Breast Cancer Histopathology Images

Adnan M. Khan[†], Hesham El-Daly[⋆], Nasir M. Rajpoot[†]
[†]*Department of Computer Science, University of Warwick, UK*
[⋆]*University Hospitals Coventry & Warwickshire, UK*
Corresponding Authors: {amkhan,nasir}@dcs.warwick.ac.uk

Abstract

In this paper, we propose a statistical approach for mitosis detection in breast cancer histological images. The proposed algorithm models the pixel intensities in mitotic and non-mitotic regions by a Gamma-Gaussian mixture model and employs a context-aware post-processing in order to reduce false positives. Experimental results demonstrate the ability of this simple, yet effective method to detect mitotic cells in standard H&E stained breast cancer histology images.

1 Introduction

Detection of Mitotic Cells (MCs) in breast histopathology images is one of three components (the other two being tubule formation, nuclear pleomorphism) required for developing computer assisted grading of breast cancer tissue slides [2]. This is very challenging since the biological variability of the MCs makes their detection extremely difficult (see Figure 1). Additionally, if standard H&E staining is used (which stains chromatin rich structures, such as nucleus, apoptotic and MCs dark blue), it becomes extremely difficult to detect the later given the fact that former two are densely localized in the tissue sections. As a consequence, two categories of relevant works have been reported in literature. One that use an additional stain (e.g. PHH3) to stain MCs exclusively, and detect exclusively stained MCs in the images [7]. Other that use a video sequence to detect mitotic events over time by incorporating spatial and temporal information [3]. Since the exclusive stain costs additionally and videos are not at all used in standard histopathological practices, therefore, a gap exists in literature.

In this paper, a robust MCs detection technique is developed and tested on 35 breast histopathology images, belonging to 5 different tissue slides. To the best of our knowledge, there is not existing method in the literature for detection of MCs in standard H&E stained breast histology images. The proposed method mimics a pathologist's approach to MCs detection under microscope. The main idea is to isolate tumor region from non-tumor areas (lymphoid/inflammatory/apoptotic cells), and search for MCs in the reduced space by statistically modeling the pixel intensities from mitotic and non-mitotic regions. In order to further enhance the positive predictive value (PPV), Context Aware Post-Processing (CAPP) has been introduced. The experimental results show that the proposed system achieves a high sensitivity of 0.82 with PPV of 0.29. Employing CAPP on these results produce 241% increase in PPV at the cost of less than 15% decrease in sensitivity.

2 The Proposed Algorithm

2.1. Stain Normalization

Tissue staining is commonly used to highlight distinct structures in histology images. Among many different stains, Hematoxylin & Eosin (H&E) is one of the most commonly used. It selectively stains nuclei structures *blue* and cytoplasm *pink*. Although staining enables better visualization of tissue structures, however due to non-standardization in histopathological work flow, stained images vary a lot in terms of color and intensity. Stain normalization is used to achieve a consistent color and intensity appearance. Among several approaches reported in literature, we used [5] to normalize the color and intensity of breast histology images.

2.2. Tumor Segmentation

Breast Cancer histology images can be divided into two regions: tumor and non-tumor. Mitosis events are more likely to exist in tumor regions. Therefore, an intelligent mitosis detection system must first remove non-tumor areas from the tissue slide in order to minimize the search space for MCs. We have developed

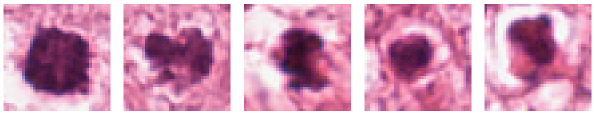

Figure 1. How hard is it to identify MCs in breast histology images? First 3 images (from left) are MCs and last 2 images are non-mitotic images.

978-4-9906441-1-6 ©2012 IAPR

a feature based texture segmentation framework (RanPEC: Random Projections with Ensemble Clustering [4]) to segment tumor regions. Broadly, the algorithm follows the following pipeline: (1) A library of texture features is computed over a range of scales and orientations, (2) low dimensional embedding (using Random Projections) is performed to avoid over fitting and curse of dimensionality, and finally (3) tumor segmentation is performed in low dimensional space. This produces an accurate and totally unsupervised tumor segmentation. On our dataset, we achieved hight sensitivity of 93% (i.e. 215 out of 231 mitotic cells were retained in the tumor areas obtained as a result of tumor segmentation).

2.3. Statistical Modeling of Mitotic Cells

MCs appear as relatively dark, jagged and irregularly textured structures (see Figure 1). Due to sectioning artifacts, some appear too dim to notice with a naked eye. In terms of *shape*, *color* and *textural* characteristics, lymphoid/inflammatory cells and apoptotic cells that are densely present in tissue slides possess almost similar characteristics, thus could easily be confused with MCs.

In this paper, we propose Gamma-Gaussian Mixture Model (GGMM) for detecting MCs in breast histology images. Image intensities (L channel of La*b* color space) are modeled as random variables sampled from one of the two distributions; Gamma and Gaussian. Intensities from MCs are modeled by a Gamma distribution and those from non-mitotic regions are modeled by a Gaussian distribution. The choice of Gamma and Gaussian distribution is mainly due to the observation that the characteristics of the distribution match well with the data it models (see Figure 2).

2.3.1 Gamma-Gaussian Mixture Model

Figure 2 shows two marginal distributions (solid lines) and their fitted models (dotted lines). The left and right marginal distributions show the probability distributions of pixels belonging to mitotic and non-mitotic regions respectively. Close fit to the marginal distributions was achieved by GGMM. The GGMM is a parametric technique for estimating probability density function. In our context, it can be formulated as follows.

For pixel intensities x, the proposed mixture model is given by:

$$f\left(x;\theta\right) = \rho_1\Gamma(x;\alpha,\beta) + \rho_2 G(x;\mu,\sigma) \tag{1}$$

where ρ_1 and ρ_2 represent the mixing proportions (priors) of intensities belonging to mitotic and non-mitotic regions, and $\rho_1 + \rho_2 = 1$. $\Gamma(x;\alpha,\beta)$ represents the Gamma density function parameterized by α (the shape parameter) and β (the scale parameter). $G(x;\mu,\sigma)$ represents Gaussian density function parameterized by μ (mean) and σ (standard deviation). $\theta = [\alpha,\beta,\mu,\sigma,\rho_1,\rho_2]$ represents the vector of all unknown parameters in the model.

2.3.2 Parameter Estimation

In order to estimate unknown parameters(θ), we employ maximum likelihood estimation (MLE). Given image intensities $x_i, i = 1, 2, ..., n$ where n is number of pixels, log-likelihood function (ℓ) of parameter vector θ is given by

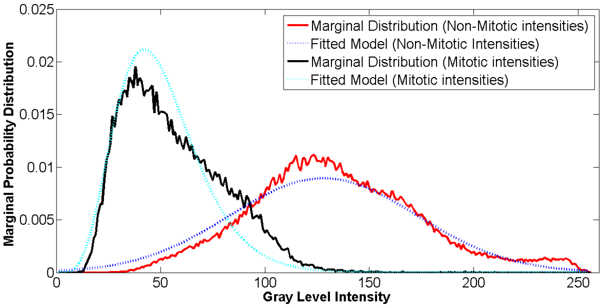

Figure 2. Marginal distributions (solid line) and fitted models (doted lines) by the two-component Gamma-Gaussian Mixture Model

$$\ell\left(\theta\right) = \sum_{i=1}^{n} \log f\left(x_i;\theta\right) \tag{2}$$

where $f\left(x_i;\theta\right)$ is the mixture density function in equation (1). The MLE of θ can be represented by

$$\hat{\theta} = \underset{\theta}{argmax} \quad \ell(\theta) \tag{3}$$

A convenient approach to obtain a numerical solution to the above maximization problem is provided by the Expectation Maximization (EM) algorithm [1]. In our context, the EM algorithm can be set up as follows.

Let $z_{ik}, k = 1, 2$, be indicator variables showing the component membership of each pixel x_i in the mixture model (1). Note that these indicator variables are hidden (unobserved). The log-likelihood (2) can be extended as follows:

$$\begin{aligned} \ell^c\left(\theta\right) &= \textstyle\sum_{i=1}^{n}\sum_{k=1}^{2} z_{ik}\log\rho_k \\ &+ \textstyle\sum_{i=1}^{n}\left\{z_{i1}\log\left[\Gamma(x_i;\alpha,\beta)\right]\right. \\ &+ \left. z_{i2}\log\left[G(x_i;\mu,\sigma)\right]\right\} \end{aligned} \tag{4}$$

The EM algorithm finds $\hat{\theta}$ iteratively, as outlined in Algorithm 1. Let $\theta^{(m)}$ be the estimate of θ after m iterations of the algorithm. The EM algorithm seeks to find the MLE of the marginal likelihood by iteratively applying Expectation and Maximization steps.

2.4. Classification

The posterior probabilities of a pixel x_i belonging to class 1 (Mitotic) or 2 (Non-Mitotic) are calculated as follows,

$$\begin{aligned} p_{i1} &= \frac{\rho_1\Gamma(x;\alpha,\beta)}{\rho_1\Gamma(x_i;\alpha,\beta)+\rho_2 G(x_i;\mu,\sigma)} \\ p_{i2} &= 1 - p_{i1} \end{aligned} \tag{9}$$

Given the pixel-wise posterior probability maps, Otsu thresholding is then used to classify mitotic and non-mitotic pixels. It was found empirically that the area of mitotic cell was between 60 and 1,000 pixels. There-

```
Algorithm 1 Expectation Maximization (EM)
1: Expectation Step (E step): Calculate the expected value of the log-likelihood function $\ell^c(\theta)$, with respect to $P\left(z|x,\theta^{(m)}\right)$, where $z = \{z_{ik},\quad i=1,2,...,n, k=1,2\}$. The conditional expectation can be given as:
   $Q\left(\theta;\theta^{(m)}\right) = \sum_{i=1}^{n}\sum_{k=1}^{2} w_{ik}^{(m)} \log \rho_k + \sum_{i=1}^{n}\left\{w_{i1}^{(m)} \log\left[\Gamma(x_i;\alpha,\beta)\right] + w_{i2}^{(m)} \log\left[G(x_i;\mu,\sigma)\right]\right\}$   (5)
   where
   $w_{i1}^{(m)} = \frac{\rho_1^{(m)}\Gamma\left(x_i;\alpha^{(m)},\beta^{(m)}\right)}{f\left(x_i;\theta^{(m)}\right)},$   (6)
   and
   $w_{i2}^{(m)} = \frac{\rho_2^{(m)}G\left(x_i;\mu^{(m)},\sigma^{(m)}\right)}{f\left(x_i;\theta^{(m)}\right)}$   (7)
   are the conditional expectations of $z_{ik}$.
2: Maximization Step (M step): The M-step maximizes the function $Q\left(\theta;\theta^{(m)}\right)$ with respect to $\theta$ using a numerical optimization.
   $\theta^{(m+1)} = \underset{\theta}{argmax} \quad Q(\theta,\theta^{(m)})$   (8)
3: Convergence Criteria: The above two steps are repeated until $\left\|\theta^{(m+1)} - \theta^{(m)}\right\| < \epsilon$ for a pre-specified value of tolerance $\epsilon$.
```

fore, area thresholding is performed to remove all potentially mitotic regions having area out of this range.

2.5. Context-Aware Post-processing

The results produced as a result of the algorithmic steps stated so far achieve 86% sensitivity, however given a large no of similar looking objects (apoptotic cells, lymphoid/inflammatory cells etc), a number of false positives (FPs) are also obtained. In order to reduce the FPs without significantly reducing sensitivity, CAPP is performed on the classification results. A small context window (see Figure 3) is defined around the bounding box of each potentially mitotic cell. In each context window, four representative features are computed over a set of textural features. The representative features are used to train a Support Vector Machine (SVM) classifier using a Gaussian kernel. The trained classifier is then used to predict unseen candidate contexts of MCs.

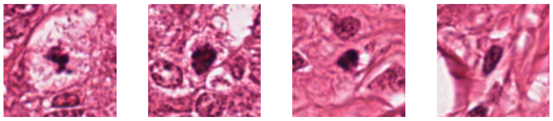

Figure 3. Four examples of 50×50 context patches, cropped around the bounding box of candidate MCs (detected using the proposed algorithm). First 2 (from left) are mitotic, last 2 are false positives.

3 Experimental Results and Discussion

Our experimental dataset consisted of 35 digitized images of breast cancer biopsy slides with paraffin embedded sections stained with Hematoxylin and Eosin (H&E) and scanned at 40× using an Aperio ScanScope slide scanner. After stain normalization, background removal and unsupervised tumor segmentation over all 35 images, seven images were selected to extract mitotic and non-mitotic pixel intensities (L channel of La*b* color space) for model fitting using GGMM. We chose 500 iterations and tolerance ($\epsilon = 0.01$) for the EM algorithm. Although EM provides estimates of priors (ρ_1 and ρ_2), a more accurate estimate of priors ($\rho_1 = 0.0014$ and $\rho_2 = 0.9986$) was used based on the ratio of mitotic and non-mitotic data used for model fitting.

The set of textural features extracted from a window of size 30×30 pixels around the bounding box of each candidate mitosis are as follows: 32 PG features (16 orientations, 2 scales) [6], 1 roughness feature, 1 entropy feature. From each of these 34 features, 4 representative features were computed: (1) mean, (2) standard deviation, (3) skewness, (4) kurtosis. This gave a $136-$dimensional features vector for each pixel inside the context window. The resulting 136 dimensional vector was used in training and testing of SVM.

Since the data consisting of candidate potential mitotic cells, identified before CAPP was applied, was unbalanced (mitotic-29.1%, non-mitotic-70.9%), therefore a balanced mix of mitotic and non-mitotic examples were randomly selected as training data. A total of 69.90% of data was used for training and remaining 30.10% for testing. Grid search was used to find optimal parameters for the Gaussian kernel of the in SVM. A higher penalty for misclassification in the SVM was set for mitotic class, since the original data was unbalanced. Table 1 provides details of the quantitative results obtained with a five-fold cross-validation. According to these results, more than 200% of Positive Predictive Value (PPV) was enhanced at the cost of less than 15% reduction in sensitivity.

4 Conclusions

In this paper, we presented a Gamma-Gaussian Mixture Model (GGMM) for detection of mitotic cells in breast cancer histopathological images. In addition, we introduced Context-Aware Post Processing (CAPP) as a tool to increase the Positive Predictive Value (PPV) with

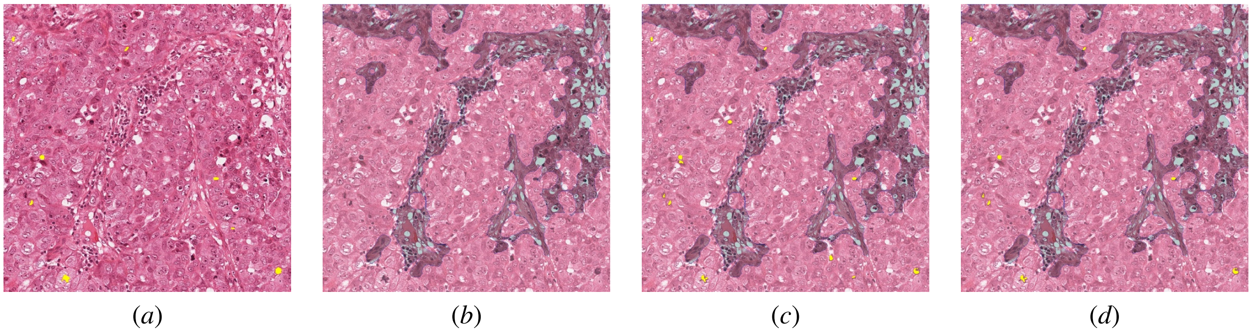

Figure 5. Visual results of MC detection in a sample image: (*a*) Original image with ground truth marked MCs shown in yellow color; (*b*) Results of Tumor segmentation (as outlined in Section 2.2) where non-tumor areas are shown in a slightly darker contrast with blue boundaries; (*c*) Results of MC detection (in yellow color) without CAPP (Sensitivity= **0.87**, PPV= **0.54**) and (*d*) Results of MC detection (in yellow color) with CAPP (Sensitivity= **0.87**, PPV= **0.87**).

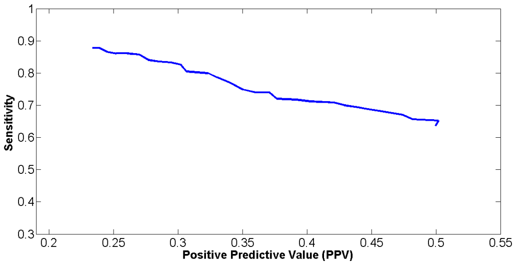

Figure 4. Plot of sensitivity vs. PPV when area-threshold is varied on the candidate MCs. High sensitivity and low PPV is obtained when small values of area-threshold were used. Table 1 shows how introduction of CAPP appreciates PPV without significantly degrading sensitivity.

Table 1. Quantitative Comparison of sensitivity and PPV with and without using CAPP for a fixed value of area threshold = 120 over 35 breast histology images containing 231 mitotic cells. By employing CAPP, PPV is doubled on unseen data, without drastically reducing the sensitivity (i.e. less than 15% only).

	Without CAPP	**With CAPP**
Sensitivity	**0.82**	0.72
PPV	0.29	**0.70**

a minimal loss in sensitivity. We evaluated the performance of the proposed detection algorithm in terms of sensitivity and PPV over a set of 35 breast histology images selected from 5 different tissue slides and showed that a reasonably high value of sensitivity can be retained while increasing the PPV. Our future work will aim at increasing the PPV further by modeling the spatial appearance of regions surrounding mitotic events.

5 Acknowledgments

The authors would like to thank the organizers of ICPR 2012 contest for mitosis detection in breast cancer. The images used in this paper are part of MITOS dataset, a dataset setup for ANR French project MICO.

References

[1] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[2] C. Elston and I. Ellis. 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, 1991.

[3] S. Huh, D. Ker, R. Bise, M. Chen, and T. Kanade. Automated mitosis detection of stem cell populations in phase-contrast microscopy images. *Medical Imaging, IEEE Transactions on*, 30(3):586–596, 2011.

[4] A. Khan, H. Daly, and N. Rajpoot. RanPEC: Random Projections with Ensemble Clustering for segmentation of tumor areas in breast histology images. In *Medical Image Understanding and Analysis*, pages 17–23, 2012.

[5] D. Magee, D. Treanor, P. Chomphuwiset, and P. Quirke. Context aware colour classification in digital microscopy. In *Proc. Medical Image Understanding and Analysis*, pages 1–5, 2010.

[6] K. Murtaza, S. Khan, and N. Rajpoot. Villagefinder: Segmentation of nucleated villages in satellite imagery. *British Mission Vision Conference*, 2009.

[7] V. Roullier, O. Lézoray, V. Ta, and A. Elmoataz. Multi-resolution graph-based analysis of histopathological whole slide images: Application to mitotic cell extraction and visualization. *Computerized Medical Imaging and Graphics*, 35(7-8):603–615, 2011.