33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Staging Tissues with Conditional Random Fields

Jagath C. Rajapakse and Song Liu

***Abstract*—We present a framework for identifying disease states by classifying cells in the pathological regions of tissues into different categories. We use conditional random fields (CRF) to incorporate characteristics of cells and their spatial distributions. The efficacy of CRF to model cell-cell feature interactions is demonstrated by using a lung tissue dataset and a synthesized cancer tissue dataset. Comparisons with an independent cell model and a contextual model based on a Markov random field indicate that CRF effectively incorporates features of both cells and their spatial distributions for identification of pathological cells.**

I. INTRODUCTION

Accurate diagnosis of cancer and many other diseases usually requires histopathological examination of samples. Traditionally, pathologists examine histopathological images of biopsy samples extracted from patients to assess deviations of cell structures and changes of the spatial distribution of cells across the tissue. They make judgments based on their personal experiences, which are often subjective and time consuming and lead to inter- and intra-rater variations.

To circumvent the drawbacks of manual examination of pathological tissues, it is important to develop automated computational techniques to render quantitative measures of tissue status for disease diagnosis. Several image analysis techniques have been recently proposed for identification of regions of tissues obtained in neuroblastoma [1], breast [2], and prostate [3] cancers. Typically, the first step of cell identification in tissues is to segment individual cells or separate cytological components from tissue images. After segmentation, different categories of features including wavelets [2], texture [3], and color [4] are exploited to classify cells and thereby characterize the tissue segments.

Using these features, classifiers are built to automatically distinguish healthy and pathological regions of tissues. Artificial neural networks [5], support vector machines (SVM) [6] have been used for cancer diagnosis. by identifying pathological cells.

In this paper, we introduce conditional random fields (CRF) to model the distribution of cells in the regions of the tissues. After predicting cell types by the classifier, CRF incorporates interactions of cells in the neighborhood. Initial structure of the CRF is obtained by computing the connection ratio, that is, neighborhood based on the ratio of the number of edges to all possible edges in the graph. Then we introduce an algorithm to find the optimal structure of the graph. Two sets of tissue data including synthesized tissue datasets were used in the experiments.

II. SINGLE CELL CLASSIFICATION

Consider 2-D tissue image $f : \Omega \rightarrow \mathbb{R}$ where $f(z)$ denotes the image intensity at pixel $z \in \Omega$ and $\Omega \subset \mathbb{N}^2$ denotes the image domain. Single cell classification consists of three steps: (i) cell segmentation with level sets, (ii) cell feature extraction with wavelet packets, and (iii) cell classification using support vector machines.

The initial segmentation of cells was performed by using multi-phase level sets [7], followed by a marker-controlled watershed algorithm. After segmentation, 30 Daubechies wavelet features [8] were obtained to represent spatial frequency information of the original image. For classification of cells, support vector machines (SVM) was used as they provide classifiers with optimal margin of separation [9]. Features and the class labels of the cells are the inputs and the outputs to SVM, respectively.

III. TISSUE CELL CLASSIFICATION

A tissue consists of a population of cells distributed in 2D space. Not only the cell features such as wavelet features, but also tissue features such as the topological distribution of cells is important to identify and thereby determine the pathological state. Our aim is to find the cell type y_i based on the features $x = \{x_i\}_{i \in \Omega}$ and the labels $y = \{y_i\}_{i \in \Omega}$ of all the cells. The type of the cell is obtained by the maximum a posteriori (MAP) estimation y_i^*:

$$y_i^* = \arg\max_{l \in \mathcal{L}} p(y_i = l | x, y) \quad (1)$$

where $\mathcal{L}$ denotes the set of cell labels.

A. Conditional Random Fields

For tissue cell classification, features and distributions of the neighboring cells are important. Suppose cells are represented by the vertices of a graph connecting one another and we assume that random variables (x, y) are represented in a conditional random field. Then for every cell i, using the fundamental theory of random fields [10], the right hand side of (1) can be written as

$$p(y_i | x, y_j, j \neq i) = p(y_i | x, y_{\mathcal{N}_i}) \quad (2)$$

where $\mathcal{N}_i$ represents the neighborhood of connecting nodes/cells to the node/cell i of the graph.

By decomposing the right hand side of (2) to nodal and neighborhood interactions [11]:

$$p(y_i | x, y) \propto p(y_i | x_i) \prod_{j \in \mathcal{N}_i \backslash i} p(y_i, y_j | x_i, x_j) \quad (3)$$

where $p(y_i | x_i)$ and $\{p(y_i, y_j | x_i, x_j)\}_{j \in \mathcal{N}_i \backslash i}$ give the likelihood of cell type given its features and the features of its

Song Liu and Jagath C. Rajapakse are with BioInformatics Research Centre, School of Computer Engineering, Nanyang Technological university, Singapore (email: {y060101}@e.ntu.edu.sg, {asjagath}@ntu.edu.sg).

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

neighbors, respectively. By using the Hammersley-Clifford theorem [12], the distributions in (3) can be written as a Gibbs distribution:

$$p(y|x) \propto \exp\left\{-\sum_i \left(V_i(y_i|x_i) + \sum_{j\in\mathcal{N}_i\backslash i} V_{i,j}(y_i,y_j|x)\right)\right\} \tag{4}$$

where $V_i(y_i|x_i)$ represents the single-cell potential and $V_{i,j}(y_i,y_j|x)$ pairwise cell interaction potentials.

The single-voxel potential in (4) can then be written as:

$$V_1(y_i = l|x_i) = \ln\left\{\frac{1}{1+\exp(1+Ag_l(x_i))}\right\} \tag{5}$$

where $g_l(x_i)$ is the output discriminant value of SVM, corresponding to lth class and A is a constant.

The pairwise potential term in (4) can be written as V_2:

$$V_2(y_i,y_j|x) = V_2(y_i,y_j) + V_2(y_i,y_j|x_{\mathcal{N}_i},x_{\mathcal{N}_j}) \tag{6}$$

where the pairwise potential $V_2(y_i,y_j)$ models the interactions of two neighbors, which can be given as a function of pairwise cliques (y_i,y_j):

$$V_2(y_i,y_j) = \begin{cases} \alpha & \text{if } y_i = y_j, \\ 1 & \text{otherwise} \end{cases} \tag{7}$$

$V_2(y_i,y_j|x_{\mathcal{N}_i},x_{\mathcal{N}_j})$ models the interactions between the nodes depending on the observed features and can be written as

$$V_2(y_i,y_j|x_{\mathcal{N}_i},x_{\mathcal{N}_j}) = \\ \beta\frac{\|x_i-x_j\|\delta(y_i-y_j)}{\frac{1}{|\mathcal{N}_i|}\sum_{q\in\mathcal{N}_i}\|x_q-x_i\|+\frac{1}{|\mathcal{N}_j|}\sum_{q\in\mathcal{N}_j}\|x_q-x_j\|} \tag{8}$$

Once the CRF is built, belief propagation is used to calculate the marginal probabilities of every node in the CRF [13]. The label of each node is then determined by (1).

IV. Parameter Estimation

A. Initialization of CRF

Let 2D spatial coordinates of cell i be denoted by z_i and the distance between cells i and j be $z_{ij} = |z_i - z_j|$. A locally connected graph is built by using the spatial distance between the cells and its neighborhood $\mathcal{N}_i = \{j : \|z_i - z_j\| < d_T\}$ where d_T denotes a distance threshold. The threshold distance should reflect the distribution of cells in the tissues and is selected to optimize the classification. The CRF is characterized by a connectivity matrix $\mathcal{C} = \{c_{ij}\}_{n\times n}$ where $c_{ij} = 1$ indicates the presence of a connection and c_{ij} indicates an absence of the connection. Since the local connectivity among the nodes varies due to different characteristics of the regions of the image, we use the average pairwise distance as this threshold distance to obtain an initial configuration of the CRF.

B. Structure Learning

In this section, we present an algorithm to determine the best graph structure based on the local connectivity. Since majority of cells (or nuclei) do not have a regular distribution over the tissue, the best structure is learned by optimizing the local structure. The algorithm is designed to find a suitable local structure that is aimed to bring improvement to the classification.

Suppose that the neighborhood of cell i has a configuration s. In order to optimize the performance of classification, a goodness measure $\varepsilon(y_i|s,\theta)$ is introduced to characterize the local graph structure that could bring the best overall classification where θ denotes the parameters of the local structure s. The parameter $\theta = \{n,\mu,\sigma\}$ consists of three variables: the relative size n_s of the neighborhood, and the mean μ and variance σ of distances from the node to its neighbors. The neighborhood configuration $s = (s_l)_{l=1}^L$ is an L dimensional vector where $s_l = \lceil N_s * \frac{|R_l|}{\sum_{l'=1}^L |R_{l'}|}\rceil$ is the proportion of cells of the lth class in the neighborhood where R_l stands for the set of neighbors of class l. $n_s = \lceil N_n * \frac{|\mathcal{N}_i|}{\max_{i'\in\Gamma}\{|\mathcal{N}_{i'}|\}}\rceil$ is the relative size the neighborhood size of node i compared to the maximum neighborhood size. N_n and N_s are the total number of neighborhoods and the neighborhood configurations over the image, respectively.

Furthermore, we define

$$\begin{aligned}\varepsilon(y_i = l|s,\theta) &= p(y_i,s|\mu,\sigma) - p(y_i \neq l,s|\mu,\sigma) \\ &= (p(s,y_i = l,\tilde{y}_i) - p(s,y_i \neq l,\tilde{y}_i))p(\mu)p(\sigma)\end{aligned} \tag{9}$$

where function $p(s,y_i = l,\tilde{y}_i)$ expresses the increment of likelihood, which for the classification by node i taking the label l.

On the contrary, the function $p(s,y_i \neq y_i^*,\tilde{y}_i,)$ denotes the probability that node i with configuration s brings impair to the performance of CRF.

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_\mu^2}}exp[-\frac{(\mu-\mu_\mu)^2}{\sigma_\mu^2}] \tag{10}$$

and

$$p(\sigma) = \frac{1}{\sqrt{2\pi\sigma_\sigma^2}}exp[-\frac{(\sigma-\mu_\sigma)^2}{\sigma_\sigma^2}] \tag{11}$$

represent the distribution of $\mu = \frac{1}{\sum_{j\in\mathcal{N}_i} d_{ij}}|\mathcal{N}_i|$ and $\sigma^2 = \frac{1}{|\mathcal{N}_i|-1}\sum_{j\in\mathcal{N}_i}(d_{ij}-\mu)^2$ under the configuration s. If $|\mathcal{N}_i| = 0$, then $p(\mu) = 1$ and $p(\sigma) = 1$. Once the structure of the CRF is learned, all the three variables s, μ and σ are obtained by using Algorithm 1.

Algorithm 1 Classification using CRF

```
Initialize CRF $\mathcal{C} = \{c_{ij}\}$,
repeat
  flips=$\emptyset$;
  for each edge with $c_{ij} = 1$ do
    Set $c_{ij} = 0$ and $\epsilon = 0$
    for each node/cell $i$ do
      Find configuration $s$ of cell $i$, and parameters $\mu$ and $\sigma$
      Using (10),
      $\epsilon = \epsilon + \varepsilon(y_i|s, \theta)$
      $$\varepsilon(y_i|s, \theta) = \sum_{l=1}^{L} \varepsilon(y_i = l|s, \mu, \sigma)$$
    end for
    if $\epsilon > \epsilon^{\text{old}}$ then
      $\epsilon^{\text{old}} = \epsilon$;
    else
      reset $c_{ij} = 1$; flips = flips++
    end if
  end for
until flips = 0
return $\mathcal{C}$
```

V. Experiment and Results

Cells in the images were first segmented into individual cells by using the multi-phase level set framework. Thereafter, Daubechie wavelet features were extracted from individual cell images. SVM was used for classification of the segmented cells. The likelihoods were then used for forming CRFs over the image. The optimal threshold for forming the CRF was obtained by using the Algorithm (1).

A. Lung Tissue Images:

Immunofluorescent 2D images of lung tissue sections were obtained with double stained of SPC and CC10 antibody with immunofluorescent markers. The ground truth was obtained through manual segmentation by expert neurobiologists. There are totally 9551 cells for the classification.

B. Synthesized Cancer Cell Images:

Simulated cells consist of two main compartments: cytoplasm and nuclei. The morphology of the two compartments are approximated by translated polygons. Furthermore, textures were added to each compartment. The cancer tissue invasion model used is a reaction-diffusion model which provides the cancer and normal cells distribution in the tissue [14]. There are totally 2773 cells for the classification.

C. Results

A sample of original lung image and the cell graph built with formulating a CRF are shown in Fig. 1. As seen, the proposed method produced a graph where only those clustered nodes are highly connected to each other. The accuracies of performance of cells into healthy and infected with influenza are given in the Table I. CRF showed significantly improved performance over the single cell classification or MRF classification.

TABLE I
Classification of lung tissue images by SVM, MRF, and CRF

Method	Accuracy	Sensitivity	Specificity
SVM	82.59 ± 2.93	32.91 ± 6.50	96.51 ± 0.69
MRF	82.79 ± 2.91	33.21 ± 6.42	96.52 ± 0.69
CRF	90.26 ± 1.69	48.30 ± 14.95	96.52 ± 0.64

A sample dataset and the CRF obtained were given in Fig. 1. As seen from Table II, the classification of cells obtain significant improvement by forming the cell graph with a conditional random field.

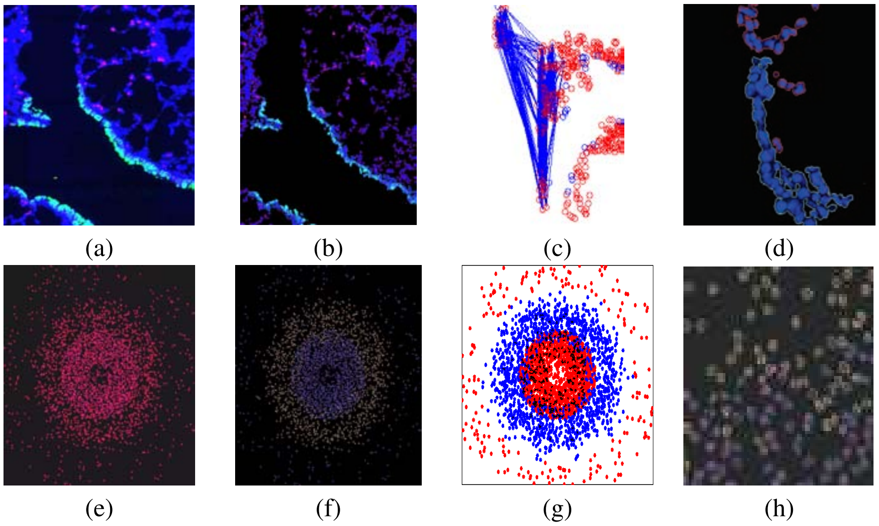
TABLE II
Classification of synthesized cancer cells by SVM, CRF and MRF

Method	Accuracy	Sensitivity	Specificity
SVM	79.87 ± 0.65	75.90 ± 1.55	85.07 ± 2.73
MRF	87.02 ± 1.45	97.74 ± 3.12	80.38 ± 2.89
CRF	88.05 ± 0.98	90.66 ± 1.32	86.09 ± 3.92

VI. CONCLUSIONS

The conditional random field was able to combine both features of cells as well as the cells' spatial location information. A cell graphs was built with a conditional random field (CRF) of the features and types of cells on the tissue. The CRF enables the transmission of messages among neighboring nodes. Those misclassified cells during the execution of SVM receive information from its nearby cells which are correctly classified. Therefore, these misclassified cells are able to re-categorized to its true cell type.

Through inspecting the lung datasets it is found that the distribution of the cells has no principled rule, for example, there are more clustered cells images than the remaining ones and the location of the cluster is also not fixed. Since there is no fixed structure of the tissue images, it brings difficulty to build a commonplace model through structure learning. For


(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 1. Lung images: (a) original immunofluorescent image, (b) segmented immunofluorescent image, (c) cell-graph obtained by building the conditional random field, blue circles denote cancer cells, red circles denote benign cells, and (d)sub-region with enlarged resolution, green contours are cancer cells, red contours are benign cells. Synthesized images: (e) synthesized immunofluorescent image, (f) segmented immunofluorescent image, and (g) cell-graph obtained by building the conditional random field, red dots denote cancer cells, blue dots denote cancer cells, and (h)sub-region with enlarged resolution, green cells are benign cells, red or violet cells are cancer cells.

the synthetic cancer tissue image, though the concentration of the cells is provided by the model, the cells distribute randomly, which brings difficulty to find a concise distribution. In order to solve these problems, we propose an algorithm which construct the CRF by adding or deleting the edges among the cells after evaluating the resulting graphical model. The measure used to evaluate the graphical model calculates the probability that one node obtain classification improvement under certain neighborhood configuration.

References

[1] M. Gurcan, T. Pan, H. Shimada, and J. Saltz, “Image analyis for neuroblastoma classification: Segmentation of cell nuclei,” in *28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society Conference*, 2006.

[2] S. Doyle and et al., “Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features,” in *International Symposium on Biomedical Imaging: From Nano to Macro*, 2008.

[3] ——, “Automated grading of prostate cancer using architectural and textural image features,” in *International Symposium on Biomedical Imaging: From Nano to Macro*, 2007.

[4] D. Altunbay and et al., “Color graphs for automated cancer diagnosis and grading,” *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 3, pp. 665–674, 2010.

[5] D. Tasoulis, P. Spyridonos, and N. Pavlidis, “Urinary bladder tumor grade diagnosis using on-line trained neural networks,” in *Lecture Notes in Artificial Intelligence Subseries of Lecture Notes in Computer Science 2773 PART 1*, 2003.

[6] D. Glotsos, P. Spyridonos, D. Cavouras, P. Ravazoula, P. Dadioti, and G. Nikiforidis, “An image-analysis system based on support vector machines for automatic grade diagnosis of brain-tumour astrocytomas in clinical routine,” *Medical Informatics and the Internet in Medicine*, vol. 30, no. 3, pp. 179–193, 2005.

[7] L. Vese and T. Chan, “A multiphase level set framework for image segmentation using the mumford and shah model,” *International Journal of Computer Vision*, vol. 3, no. 50, pp. 271–293, 2002.

[8] A. Laine and J. Fan, “Texture classification by wavelet packet signatures,” *IEEE Trans PAMI*, vol. 15, pp. 1186–1191, 1993.

[9] N. Cristianini and B. Scholkopf, “Support vector machines and kernel methods: The new generation of learning machines,” *AI Magazine*, vol. 23, no. 3, pp. 31–41, 2002.

[10] J. Hammersley and P. Clifford, “Markov fields on finite graphs and lattices,” *Unpublished manuscript*, 1971.

[11] C. Huang and A. Darwiche, “Inference in belief networks: A procedural guide,” *International Journal of Approximate Reasoning*, vol. 15, no. 3, pp. 225–263, 2006.

[12] J. Besag, “On the statistical analysis of dirty pictures,” *J.Roy.Statist.Soc.B*, vol. 48, pp. 259–302, 1986.

[13] J. Yedidia and W. Freeman, “Understanding belief propagation and its generalizations.”

[14] R. Gatenby and E. Gawlinski, “A reaction-diffusion model of cancer invasion,” *Cancer Research*, vol. 56, pp. 5745–5753, 1996.