Context-Based Classification of Cell Nuclei and Tissue Regions in Liver Histopathology

Phattthanaphong Chomphuwiset[1]
scpc@leeds.ac.uk
Derek Magee[1]
d.r.magee@leeds.ac.uk
Roger Boyle[1]
r.d.boyle@leeds.ac.uk
Darren Treanor[2]
darrentreanor@nhs.net

[1] School of Computing
University of Leeds
Leeds, UK
http://www.comp.leeds.ac.uk

[2] Department of Histopathology
St James University Hospital
Leeds, UK
http://www.virtualpathology.leeds.ac.uk

Abstract

This paper presents a novel technique for classifying both cell nuclei and tissue regions in liver specimens by incorporating context information, linking cell nuclei and tissue regions using Bayesian networks. The method works in two stages: (i) initial classification of cell nuclei and tissue regions; and (ii) integrating the initial classifications using a Bayesian network to enforce consistancy (thus including context). Results demonstrate that our method of incorporating context information is superior to the classification that uses only object based features for both nucleus and region classification.

1 Introduction

One of the key challenges for automated analysis of histopathology images is cellular segmentation and classification. For segmentation, a number of algorithms have been proposed to cope with the complexity of tissue structure and colour variation of tissue images, e.g. active contour-based, watershed, classification-based and edge-based techniques [4]. In classification, cell nuclei are characterised by distinct appearances using object-based features (OBFs), such as morphology, intensity and texture, to classify them into different classes [5].

Using OBFs to represent cellular structures (e.g. cell nuclei) for classification can produce poor results when morphology, texture and appearance of nuclei are only marginally different between classes. This is often the case in interpretation of histopathology images, and in practice histopathologists make extensive use of other information to identify and classify tissue. To improve classification accuracy, domain-specific information, such as the tissue type of the region containing the cell and spatial relationships between neighbouring cell nuclei, can be used. Figure 1(c) illustrates such domain-specific context information for liver cell nuclei that are found in certain tissue regions, e.g. hepatocytes (blue bounding boxes) are usually found in parenchyma [1]. In addition, there are expected spatial relationships among nuclei types such as epithelia nuclei (represented in yellow bounding boxes).

In this paper, we present a novel technique to perform cell and tissue region classification simultaneously in liver tissues, by incorporating context information linking cell nuclei type

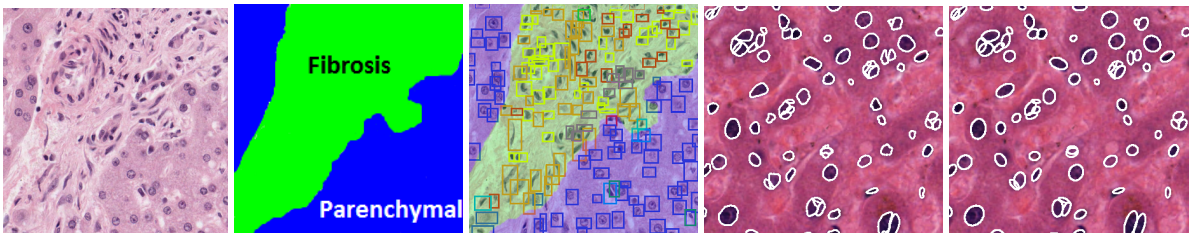
[1] the parenchyma are the functional parts of an organ in the body

and tissue region type to improve performance of classification. For nuclear classification, we first apply a Hough Transform-based technique to detect cell nuclei. Initial cell type classification is performed using area-based and shape-based descriptors. In tissue region classification, we classify images into different region types using texture features. The final stage of classification is implemented by combining initial classifications and context information to improve results using a Bayesian network.

2 Materials and Methods

2.1 Image Data

Tissue sections were processed in a routine way by formalin fixing and staining with Heamatoxylin and Eosin (H&E). These were digitised to produce virtual slides at a resolution of 103000 dpi using an Aperio XT virtual slide scanner (Aperio, San Diego) at 40× magnification. Images were selected from five sample tissues of patients who have cirrhosis. Cell nuclei ground truth were provided by a consultant pathologist specialising in liver disease producing bounding boxes around the nuclei of different types, example is shown in Figure 1(c)), and tissue regions were manually delineated (figure 1(b)).



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

Figure 1: example of image data. (a) original image; (b) region map; (c) nuclei and region maps in the tissue specimen; (d) detected nuclei (NMBR); and (e) detected nuclei (MBR).

2.2 Tissue Region Classification

Images are classified on a per-pixel basis into 3 classes (parenchyma, fibrosis and background) using Grey Scale Co-occurence Matrices (GLCMs) [6]. GLCMs are generated based on co-occurring grey values of pixels with particular spatial relationships with respect to the displacement (d) and direction (θ). We set $\theta = \{0°, 45°, 90°, 135°\}$, d=2 and use a sliding window of size 85×85 pixels (determined by cross validation). Thus, 4 occoccurrence matrices are constructed. After normalisation, texture features are extracted by calculating statistical measurements from the matrices (entropy, correlation, contrast, inverse difference moment and angular second moment).

Classification is performed using a Random forest classifier [1]. This combines multiple decision trees (T) such that the structure of each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. Prediction is made based on the majority decisions from the trees in forest.

2.3 Nuclei Detection and Classification

To detect cell nuclei we use a Hough transform-based technique to detect cell nuclei. This is based on the assumption that cell nuclei are rounded with an approximately constant size (dim) [7]. After nuclei detection, each cell nuclei is represented by a set of boundary points, which is considered as non model-based representation (NMBR). We additionally use a model-based representation (MBR) to represent cell nuclei by fitting a ellipse $E_{\mathbf{m}}$, where

$\mathbf{m}=\{c_x, c_y, a, b, \theta\}$, to the boundary points using the direct least squares technique of Fitzgibbon [2]. Where; c_x, c_y is the centre of the ellipse, a and b are minor and major axis, θ is the angle of the major axis. Example representations are shown in figures 1(d) and 1(e).

To classify cell nuclei, we extract features from the two nuclei representations(NMBR and MBR). In the MBR, we first classify pixels into different classes (purple, pink and white) by applying an unsupervised learning method. This involves fitting Gaussian Mixture model using Variational Bayesian approach. Colour priors for the mixture model are based the mean colour of pink, purple and white pixels in training data. Features are then extracted as follows: (i) histograms of purple and white pixels at different distances from the nuclei boundary (generated using a signed distance transform), (ii) the ratio of purple pixels inside and outside the ellipse. We define the area "outside" of an ellipse by scaling major/minor axis with a constant c ($c > 1$) and (iii) length of major and minor axis of the ellipses. In the NMBR, we extract shape descriptors using Hu moments [3].

We assume that there is degree of dependancy between nuclei types and tissue regions. Therefore, classification is performed under the dependency assumption ($P(N|S,A)$ where N denotes nuclei classes, S denotes tissue region types and A denotes OBFs), using a Random forest classifier.

2.4 Combining and Improving Nuclei and Region Classification

In previous sections, we presented techniques for classifying cell nuclei and tissue region type. In this section we present a method for improving classification accuracy by combining the classification of nuclei and tissue regions with context information from nearby nuclei and regions. Images are divided into non-overlapping tiles, or “super pixels”, (S). Each super pixel contains a set of nuclei (N). For each S, there is associated class prior probabilities, $P(C_s|S)$ (where C_s is the tissue region type), which are derived from the initial per-pixel classification by averaging. Similarly for cell nuclei (N), prior probabilities on classes are defined by the initial classification, $P(C_n|N)$ where C_n is the set of nuclei classes. We combine the two different types of information from cellular structures and regions using a Bayesian network (figures 2(a) and 2(b)) - a graph that is constructed from 2 layers (super pixels and nuclei). In the Bayesian Network framework, information from nuclei can be used to derive posterior probabilities of region classes and vice versa. We use loopy belief propagation techniques to approximate inference by organising message passing across graphs. There are multiple possible factorisations of pairwise joint probability distributions (learned

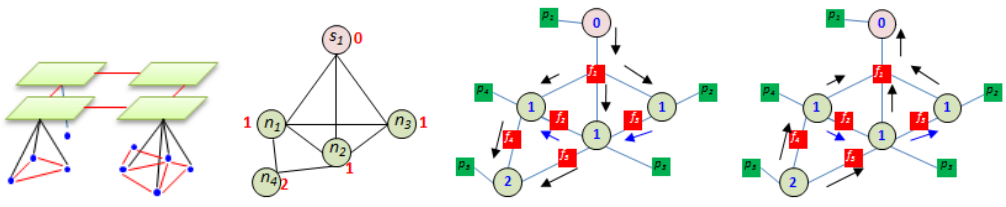


Figure 2: Example of Bayesian Network; (a) super pixels and nuclei in tissue; (b) a simple graph; (c) directed graph for message passing (the blue arrows are the direction that is selected arbitrarily); and (d) the reverted direction.

from training data) over the class variables that may be represented as different Bayesian networks. There is no inherent directionality in the conditional probabilities (i.e. no reason for factorising $P(A,B)$ as $P(A|B)P(B)$ rather than $P(B|A)P(A)$), so to build our network we use the following algorithm:

1. Generate factor graphs (figure 2(c)).

2. Arbitrarily define one vertex as the root node
3. Calculate the graph distance between each other node and the root node (this can be achieved efficiently by propagating distances out from the root. All edges to have a distance of 1.)
4. Build the Bayesian network by defining conditional probabilities between connected vertex with the highest distance (H) to the lowest (L), i.e. $P(L|H)$. Where the distances are equal an arbitrary choice is made (figures 2(c) and 2(d)).
5. Leaf nodes are then defined as vertices that are not conditioned on any other node.

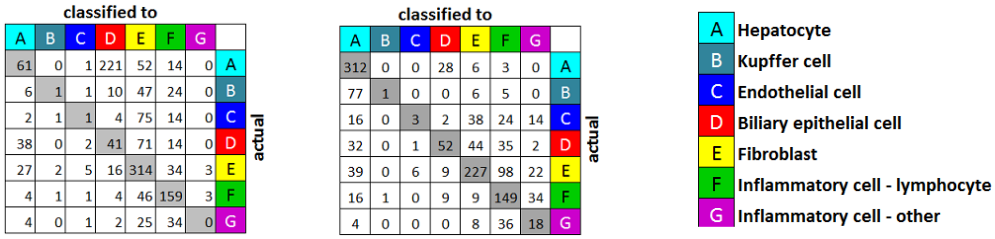
Reasoning is performed using loopy belief propagation and multiple conditional probabilities are combined using an independence assumption.

3 Experiments and Results

Our approach was evaluated using 18 liver tissue images comprising in total of 1386 cell nuclei of 7 types (figure 3(c)). Nuclei detection was performed using prior knowledge of the diameter of cell nuclei (at 40× magnification) *dim*=30 pixels. Cell type classification was performed using various combinations of features from the NMBR and MBR. We conduct 10-fold cross validation on the feature subsets. Results are shown in Table 1.

Table 1: Nuclei classification results using the feature subsets.

Feature	Accuracy
Histogram of puple/white vs. distance transform	30.13%
Major/minor+ raito of purple pixels	**41.68%**
Hu Moments	24.14%
MBR+NMBR	32.03%
Random classifier (the baseline)	12.50%



(a) (b) (c)

Figure 3: The confusion matrices; (a) initial classification and (b) context-based classification. (c) nuclei classes.

MBR features and the ratio of purple pixels give the highest accuracy (41.68%). Texture features from both MRB and NMRB do not provide good classification. This is because there are marginal differences of textural appearance between cell nuclei subtypes, e.g. hepatocytes and epithelia cell nuclei.

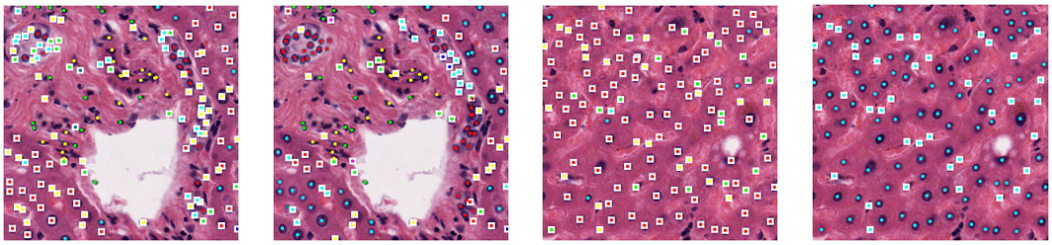
We classifed tissue regions in the same 18 images using GCLM described in section 2.2 (sub-sampling images to 8× resolution). Evaluation is performed by 10-fold cross validation and an accuracy of 78.20± 12.24% is achieved.

To perform the final classification using contextual information, we combine information from initial nucle and tissue regions classifications and generate a Bayesian network (section 2.4). The results, in Figure 3(a) and 3(b), show that context-based techniques significantly improves both nuclei classification and region classification accuracy. Accuracy of the nuclei classification is improved to 54.97%, while the tissue region classification accuracy is 82.35%±10.42. Example result images are illustrated in Figure 4.

4 Discussion and Conclusion

This paper presents techniques for classifying nuclei and tissue region type in liver histopathology images. Cell nuclei are detected using a Hough transform-based approach before nuclei classification is performed using morphological and area-based features. For tissue region classification, we use a texture descriptor (GLCM) and perform pixel-based classification using a Random Forest. We apply contextual information linking cell nuclei and tissue regions using Bayesian network and loopy belief propagation.

The results show that integration of context information into classification improves classification performance (figure 3(a) and 3(b)). Nonetheless, there is still a significant difficulty in distinguishing between fibroblast and lymphocyte nuclei types. In general, fibroblast and lymphocyte nuclei are mixed in tissue regions, therefore the context information is not able to effectively separate these two types of cell nuclei. This leads us to consider the cell nuclei detection technique which assumes that cell nuclei are rounded in shape. Therefore, elongated shape cell nuclei are, e.g. lymphocyte, are not correctly detected, instead they are segmented as rounded objects resulting mis-classification. Ongoing work is attempting to improve the cell nuclei detection technique so that elongated cell nuclei will be properly segmented.



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

Figure 4: Results of nuclear classification; (a) and (c) initial classification; (b) and (d) context-based classification (white boxes: incorrect classified nuclei)

References

[1] L. Breiman. Random forests. *Mach. Learn.*, 45:5–32, October 2001. ISSN 0885-6125.

[2] A. Fitzgibbon, M. Pilu, and R.B. Fisher. Direct least square fitting of ellipses. pages 476 – 480. IEEE Transactions Pattern Analysis and Machine Intelligence, May 1999.

[3] M. K. Hu. Visual pattern recognition by moment invariants. pages 179 – 187. IRE Transactions on Information Theory, February 1962.

[4] G. Lehmann. The watershed transform in itk - discussion and new developments. *The Insight Journal*, January-June 2006.

[5] R. Nasir, A. Munammad, and B. Abhir. Unsupervised learning of shape manifolds. In *Proceedings British Machine Vision Conference*, September 2007.

[6] J. K. Shuttleworth, A. G. Todman, R. N. G. Naguib, B. M. Newman, and M. K. Bennett. Colour texture analysis using co-occurrence matrices for classification of colon cancer images. In *Proceeding of the 2002 IEEE Canadian Conference on Electrical & Computer Engineering*, pages 1134–1139, 2002.

[7] Y. Zhu. Technical report: Computerised analysis and quantification of tma. Computing University of Leeds, September 2008.