Classification of breast tissue-microarray spots using texton histograms

Telmo Amaral[a,*], Stephen McKenna[a], Katherine Robertson[b], and Alastair Thompson[c]

[a]School of Computing, [b]Pathology and Neuroscience, [c]Surgery and Molecular Oncology
University of Dundee, DD1 4HN, Dundee, UK

Abstract. Breast-tissue microarrays facilitate the survey of very large numbers of tumours but their scoring by pathologists is time consuming, typically highly quantised and not without error. Automated segmentation of cells and intra-cellular compartments in such data can be problematic for reasons that include cell overlapping, complex tissue structure, debris, and variable appearance. This paper proposes a computationally efficient approach that approximates the density of colour and local invariant features by clusters in the feature space, and characterises each spot by a frequency histogram of nearest cluster centres. Spots are classified into four main types based on their histograms. This approach does not rely on accurate segmentation of individual cells. Classification performance was assessed using 344 spots from the Adjuvant Breast Cancer (ABC) Chemotherapy Trial. A two-layer neural network yielded better classification results than a nearest-neighbour classifier or a single-layer network. Some reasons for classification results in disagreement with pathologist-provided labels are discussed and include the existence of spots containing large proportions of different tissue types.

1 Introduction

Tissue microarrays (TMAs) are an array-based high-throughput technique proposed by Kononen et al. [1], to facilitate gene expression and the survey of very large numbers of tumours. Hundreds of cylindrical biopsies (named *cores*) from individual tumours can be distributed in a single microarray block. Sections of the block provide targets for parallel in situ detection of DNA, RNA, and protein targets in each specimen on the array, and consecutive sections allow the rapid analysis of hundreds of molecular markers in the same set of specimens. Camp et al. [2] have concluded that two cores per patient are sufficient to adequately represent the expression of three common antigens in invasive breast carcinoma.

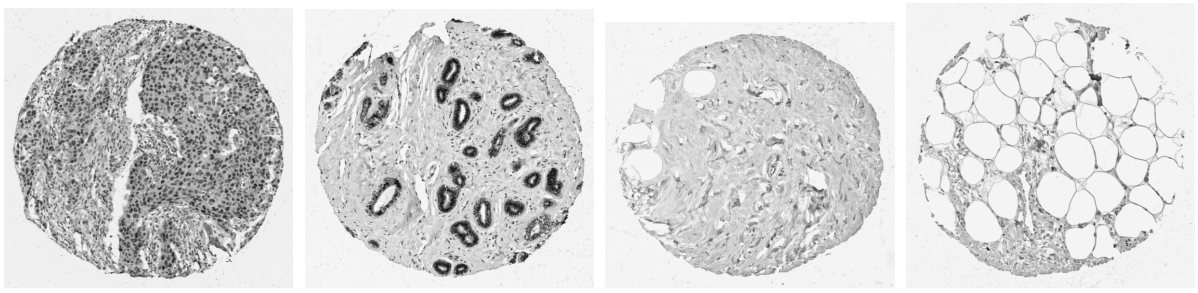
The assessment by pathologists of breast-TMA sections starts with the classification of each spot into one of several types, namely: tumour, normal, stroma, fat, blood, and invalid. Each tumour or normal spot subjected to nuclear staining is then assigned a Quickscore [3] that reflects the estimated proportion of epithelial nuclei that are immuno-positive, and the estimated strength of staining of those nuclei. Applying this procedure to breast-TMA sections from large numbers of individuals is time consuming and suffers from inter- and intra-observer variability, perceptual errors, and severe quantisation that leads to the loss of potentially valuable information. Thus, there is strong motivation for the development of automated methods for quantitative analysis and grading of breast-TMA image data.

Accurate segmentation of cells and intra-cellular compartments in such data can be problematic for reasons that include cell overlapping, complex tissue structure, debris and variable appearance. This paper explores the hypothesis that automated classification of entire tissue spots need not rely on the detection and segmentation of individual cells and intra-cellular compartments. Rather we propose a computationally efficient system that approximates the joint probability distribution of local features by clusters in the feature space, and then characterises the appearance of each spot by a frequency histogram of nearest cluster centres. Spots are classified into four main types based on their histograms. Thus, our work follows an approach similar to that developed by Varma et al. [4] for statistical texture classification, although we use rotation-invariant features of a different type and a different classifier.

The spatial orientation of cells, sub-cellular compartments, and arrangements of such structures in breast-tissue histological sections varies greatly. This suggests that the use of local features invariant to rotation could lead to a better use of the training data, in that features computed for a given training pixel based on its neighbourhood would help teach the system to deal with any test pixels with a similar neighbourhood, regardless of the orientation. In our work, the luminance signal is characterised through differential invariants computed as combinations of Gaussian partial derivatives, as discussed by Schmid et al. [5] and theoretically studied by Koenderink et al [6].

Perhaps the most closely related work to that described here concerns automated Gleason grading of prostate tissue images. Specifically, Tabesh et al. [7] reported the use of colour-channel histograms with texture features transformed to

*tamaral@computing.dundee.ac.uk



(a) Tumour (b) Normal (c) Stroma (d) Fat

Figure 1. A spot characteristic of each type.

achieve rotational invariance, and Doyle et al. [8] used morphological and texture features for this purpose. Differential invariants were used by Ranzato et al. [9] for recognition of several categories of biological particles (but not for tissue analysis). In recent work on breast-tissue analysis reported by Kostopoulos et al. [10], colour textural features are used in the assessment of the oestrogen receptor's status of breast carcinomas, through an approach that involves the pre-segmentation of nuclei.

Section 1.1 presents a summary of our previous related work. Section 2 describes the available data and the methods involved. Section 3 details the experiments carried out and presents the results. In section 4 some specific results are discussed. Section 5 provides some conclusions and recommendations for future work.

1.1 Previous work

Our previous work included the classification of breast-TMA spots into two classes, as to the presence or absence of immunopositive epithelial nuclei (regardless of the type of spot) [11]. The analysed data consisted of 110 spots (2 for each of 55 participants) subjected to progesterone-receptor (PR) nuclear staining and whose immunostates (positive or negative) were assigned by a pathologist. In addition, the contours of several hundred epithelial nuclei were marked within randomly selected sub-regions and labelled as immunopositive or negative. In a first stage, the pixels within annotated sub-regions were used to estimate the likelihoods of RGB and differential invariant features (computed for two scales up to the 2nd order) for three classes, namely: epithelial positive, epithelial negative, and background. Assuming these features to be independent, their likelihoods were then used to classify the pixels of whole spots into the three considered classes, using Bayes' rule. In a second stage, the previously classified pixels were used to compute features for each spot that aimed to formalise the two Quickscore values assigned by pathologists. A generalised linear model (GLM) was then trained to classify spots as to their immunostate, based on the two computed features. A leave-2-out experiment was carried out, in order to assess the ability of the system to deal with data from new participants. Different combinations of features were tested, leading to the conclusion that the use of differential invariants in addition to colour yielded a small improvement in accuracy. The most favourable combination of features resulted in a correct-classification rate of 84%.

2 Data and methods

The data used in this work consist of colour images of breast-TMA spots originating from the National Cancer Research Institute's Adjuvant Breast Cancer (ABC) Chemotherapy Trial [12]. A total of 344 spots subjected to PR nuclear staining were used, 86 for each of 4 spot types. The considered types were tumour (T), normal (N), stroma (S), and fat (F), and the assignment by a pathologist of each spot to one of these types was known. Figure 1 shows a spot characteristic of each type, although they are highly variable in appearance. The images have a resolution of 0.23 μm / pixel and a typical spot has a diameter of 700 μm (i.e. about 3000 pixels).

The original image is first converted to grey-scale and down-sampled to both a 4th and a 16th of its size through REDUCE operations [13], in order to build a 3-level Gaussian pyramid. (A REDUCE operation corresponds to the 2D-convolution of the input image with a [1 4 6 4 1]/16 pattern of weights, followed by down-sampling to half the input size on both dimensions.) A set of Gaussian derivative kernels with a standard deviation $\sigma = 8$ pixels are convolved with the whole pyramid. In the case of spots containing epithelial cells (i.e. spots of types T and N), these kernels are expected to focus on parts of cells at the pyramid's base level, on whole cells at the 2nd level, and on cells and their immediate surroundings at the 3rd level, as the average cell radius is about 16 pixels. The filter responses are then used

to compute the 1st- and 2nd-order differential invariants defined in (1), where G_i and G_{ii} are the 1st- and 2nd-order derivatives of a Gaussian kernel along direction i, respectively, I is the luminance (grey-scale) function of the image, and * and • are the convolution and scalar product operators, respectively [5].

$$\begin{aligned} d_A &= (G_x * I) \bullet (G_x * I) + (G_y * I) \bullet (G_y * I) \\ d_B &= (G_{xx} * I) + (G_{yy} * I) \end{aligned} \tag{1}$$

The 2nd and 3rd levels of the resulting differential invariant pyramids are subjected to the necessary EXPAND operations. (The effect of an EXPAND operation is to expand an $(M+1)$-by-$(N+1)$ array into a $(2M+1)$-by-$(2N+1)$ array by interpolating new node values between the given values, making use the same pattern of weights as in REDUCE.) Thus, 9 features become available for each position in the image, namely the R, G, and B colour values, and d_A and d_B for each of the 3 considered scales. We call this vector of 9 features a *jet*. Each feature is then normalised to zero mean and unit variance, the normalisation constants being computed based on a training set of images.

K-means clustering is applied to the jets of a randomly selected subset of image positions from the training set, so as to determine the centres of a number of clusters in the 9-dimensional feature space. The obtained set of centres constitutes essentially a *texton dictionary*, given the nature of the extracted features. Nearest-neighbour classification is then applied to the jets of a randomly selected subset of image positions (much larger than that used at the clustering stage) from the complete data set, in order to assign each jet to the nearest texton. For each image in the data set, a normalised histogram is then computed to reflect the distribution of jets over the clusters.

In the final stage, a two-layer neural network [14] (i.e. a multi-layer perceptron, or MLP) is trained with the training spots' histograms and used to classify the test spots into each of the four spot-type classes. The learning algorithm used was scaled conjugate-gradients (SCG) optimisation, and softmax was chosen as the activation function. For each spot, this function outputs four values that can be interpreted as the posterior probabilities of the spot belonging to each class. As discussed later in sections 3 and 4, the choice of a classifier with this type of output proved to be useful. In the learning stage, *regularisation* was used to penalise weights becoming too large. This means that a term with weight-decay penalty α is added to the entropy error function $E = -\sum_{n,k} t_k^n \ln y_k^n$ (where t_k^n and y_k^n are respectively the target binary code and the softmax output value, for the *k*th class and *n*th input vector), which becomes $E_r = E + \alpha \sum w_i^2$ (where w_i is the *i*th weight). Each test spot was classified as belonging to the class with the highest softmax value. As an alternative to the two-layer network, a nearest-neighbour classifier and a single-layer network (a GLM) were also used to classify spots based on the normalised texton histograms.

3 Experiments

The 344 available spots were divided into two subsets of 172 spots, suitable for running leave-172-out experiments. The size of the texton dictionary was set at 160 centres, so that the implementation of k-means that was used could still process a relatively large number of data points from the training set, specifically 610,000 points, representing 0.06% of the total. In turn, the nearest-neighbour algorithm used to compute the texton histograms was based on 3.95% of each spot's data points. The number of hidden units was fixed at 3, and different values for α were tested. All the code was implemented in Matlab, making use of the Sussex convolution function [15] and the Netlab [16] implementations of k-means, k-nearest-neighbour, and neural networks.

For each of the two passes of a leave-172-out experiment, 10 runs of training and testing with the two-layer network were executed. Table 1(b) shows the resulting minimum, average, and maximum correct-classification rates over the 20 runs, for four different choices of α. Table 1(c) shows the confusion matrix for the best case ($\alpha = 0.1$). For comparison, Table 1(a) shows the average classification rates obtained using not the two-layer network, but a nearest-neighbour classifier (with χ^2 and Euclidean distance metrics) and a single-layer network trained through SCG optimisation with $\alpha = 0.1$.

The spots shown previously in Figure 1 are typical examples of correctly classified spots. In turn, Figure 2 shows four examples of misclassified spots. Below each image, the values of the softmax activation function (i.e., the output posterior probabilities) are presented for each class. The entropy of the posteriors distribution for each spot can be used as a simple measure of classification confidence (the lower the entropy, the higher the confidence). Table 1(d) shows the fractions of test spots that can be classified below different entropy thresholds, averaged over the experiment's 20 runs. Also shown are the mean classification rates and the mean rates of misclassified T (tumour) spots.

Table 1. Correct-classification rates using (a) nearest-neighbour classifiers and GLM, and (b) two-layer network with different choices of α. (c) Confusion matrix for best results with two-layer network ($\alpha = 0.1$), and (d) fractions of spots classified below different entropy thresholds, with corresponding rates of correct classification and of missed tumour spots.

(a)

Method	Avg. rate (%)
Nearest-neigh. (χ^2)	60
Nearest-neigh. (Eucl.)	59
GLM	67

(b)

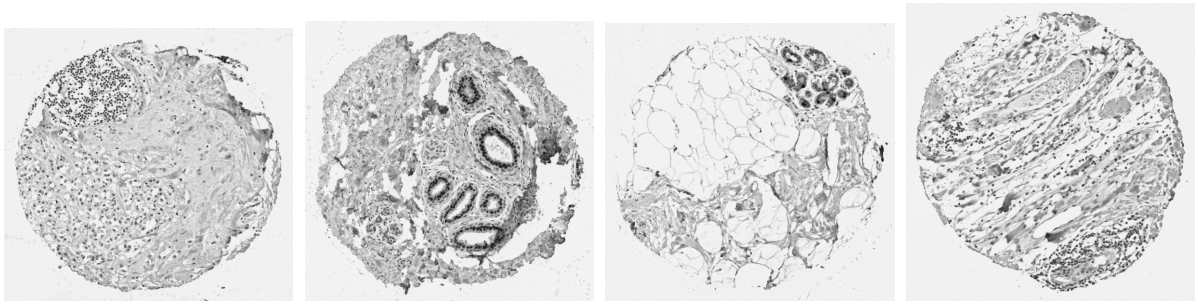
α	Min.	Avg.	Max.
0.00	63	70	74
0.01	70	73	74
0.10	73	75	76
0.20	70	70	70

(c)

Truth	Predicted			
	T	N	S	F
T (%)	74	14	12	0
N (%)	21	56	23	0
S (%)	7	11	73	8
F (%)	0	0	5	95

(d)

Entropy threshold	Avg. classified fraction (%)	Avg. rate (%)	Avg. missed-T rate (%)
1.89	100	75	26
1.69	97	75	25
1.49	85	80	15
1.29	72	84	9
1.09	56	90	5
0.89	41	92	2
0.69	32	95	0
0.49	26	96	0
0.29	17	97	0
0.09	4	100	0



(a) T predicted as N

T	N	S	F
14	**49**	37	0

(b) N predicted as T

T	N	S	F
68	28	04	0

(c) N predicted as S

T	N	S	F
05	11	**45**	**40**

(d) S predicted as N

T	N	S	F
34	**49**	17	0

Figure 2. Examples of misclassified spots and corresponding softmax values for the four classes.

4 Discussion

The epithelial cells in spot 2(a) are unusually far apart, which may explain the low posterior probability of true class T and suggest that a model solely based on local features is incapable of capturing some of the relevant information contained in the spot. Spot 2(b), when compared with the correctly classified spot 1(b), shows similar ring-like arrangements of epithelial cells, but a much larger quantity of scattered epithelials, which seems to have caused the posterior probability of class T to be much higher than that of true class N. In spot 2(c), large regions of stroma and fat boosted the posterior probabilities of classes S and F, when what truly counts for the pathologist is the small portion of normal tissue in the top-right region of the spot. This indicates a difficulty in dealing with heterogeneous spots that contain large proportions of different types of tissue. The scattered but non-epithelial cells in spot 2(d) seem to have been perceived as epithelials, leading to a very low posterior for the true class S. This may indicate that the filters used do not provide enough detail.

As Table 1(d) shows, the system is capable of classifying reasonable fractions of spots with high levels of confidence, achieving both high correct-classification rates and very low or zero misclassified-tumour rates (as is desirable in this application). This suggests that the system could be used to automatically classify the more unequivocal spots, while pointing out to the pathologist the more difficult spots in need of manual assessment.

5 Conclusions and recommendations

This paper presented a system based on texton histograms that classifies breast-TMA spots into four distinct types, achieving a correct-classification rate of 75%. Higher rates are attainable for fractions of the data, by setting classification-confidence thresholds. Some instances of misclassification were discussed.

The underlying model of tissue section may need to incorporate not only local features, but also morphological information that can reflect arrangements of nuclei (or simply their sparse scattering). A possibility is to model tissue structures as graphs whose nodes are the centres of pre-detected nuclei [17, 18]. The system could also benefit from the

use of other 2nd-order (or even 3rd-order) differential invariant features, so that more nuclear detail could be perceived and non-epithelial nuclei would not be mistaken for epithelial. Alternatively, it would be worth testing the system with the "Gabor-like" filters proposed by Schmid [19], which were shown to perform better than rotation-invariant combinations of derivatives, in the context of content-based image retrieval. In addition, it would be worth testing the system with a more perceptually uniform colour space than RGB, given that colour in histological staining is designed for human evaluation.

The approach described in this paper should be incorporated into a wider system, possibly along with the previous work described in section 1.1. The goal of this wider system should be to rule out S and F spots, classify T and N spots into immunopositive and immunonegative, and assign Quickscores to those spots identified as immunopositive.

Acknowledgements

This research was funded by the Breast Cancer Research Trust[1].

References

1. J. Kononen, L. Bubendorf, A. Kallionimeni et al. "Tissue microarrays for high-throughput molecular profiling of tumor specimens." *Nature Medicine* **4(7)**, pp. 844–847, 1998.
2. R. Camp, L. Charette & D. Rimm. "Validation of tissue microarray technology in breast carcinoma." *Laboratory Investigation* **80(12)**, pp. 1943–1949, 2000.
3. S. Detre, G. Saccani Jotti & M. Dowsett. "A "quickscore" method for immunohistochemical semiquantitation: validation for oestrogen receptor in breast carcinomas." *Journal of Clinical Pathology* **48(9)**, pp. 876–878, September 1995.
4. M. Varma & A. Zisserman. "A statistical approach to texture classification from single images." *International Journal of Computer Vision* **62(1-2)**, pp. 61–81, April 2005.
5. C. Schmid & R. Mohr. "Matching by local invariants." Technical Report RR-2644, INRIA, 1995.
6. J. Koenderink & A. van Doom. "Representation of local geometry in the visual system." *Biological Cybernetics* **55(6)**, pp. 367–375, 1987.
7. A. Tabesh & M. Teverovskiy. "Tumor classification in histological images of prostate using color texture." In *Asilomar Conference on Signals, Systems and Computers*, pp. 841–845. IEEE, 2006.
8. S. Doyle, M. Hwang, K. Shah et al. "Automated grading of prostate cancer using architectural and textural image features." In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1284–1287. IEEE, 2007.
9. M. Ranzato, P. Taylor, J. House et al. "Automatic recognition of biological particles in microscopic images." *Pattern Recognition Letters* **28(1)**, pp. 31–39, 2007.
10. S. Kostopoulos, D. Cavouras, A. Daskalakis et al. "Colour-Texture based image analysis method for assessing the Hormone Receptors status in Breast tissue sections." In *International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4985–4988. IEEE, 2007.
11. T. Amaral, S. McKenna, K. Robertson et al. "Classification of breast tissue microarray spots using colour and local invariants." In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 999–1002. IEEE, Paris, France, May 2008.
12. Adjuvant Breast Cancer Trials Collaborative Group. "Polychemotherapy for early breast cancer: Results from the international adjuvant breast cancer chemotherapy randomized trial." *Journal of the National Cancer Institute* **99(7)**, pp. 506–515, 2007.
13. P. J. Burt. "The Laplacian pyramid as a compact image code." *IEEE Transactions on Communications* **31**, pp. 532–540, April 1983.
14. C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
15. D. Young. "Convolutions in matlab." `http://www.cogs.susx.ac.uk/courses/compvis/matlab_demos/convolution_demo.html`. Checked on March 24, 2008.
16. I. Nabney. *NETLAB: algorithms for pattern recognition*. Springer-Verlag, New York, 2002.
17. K. Rodenacker & P. Bischoff. "Quantification of tissue sections." *Pattern Recognition Letters* **11(4)**, pp. 275–284, 1990.
18. J.-M. Geusebroek, A. W. Smeulders, F. Cornelissen et al. "Segmentation of tissue architecture by distance graph matching." *Cytometry Part A* **35(1)**, pp. 11–22, 1999.
19. C. Schmid. "Constructing models for content-based image retrieval." In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 39–45. IEEE, 2001.

[1] http://www.breastcancerresearchtrust.org.uk/