The Contribution of Morphological Features in the Classification of Prostate Carcinoma in Digital Pathology Images

Nicholas McCarthy
University College Dublin
nicholas.mc-carthy@ucdconnect.ie

Gillian O'Hurley
University College Dublin
gillian.ohurley@ucd.ie

Pádraig Cunningham
University College Dublin
padraig.cunningham@ucd.ie

Abstract

In this paper we present work on the development of a system for automated classification of digitized H&E histopathology images of prostate carcinoma (PCa). In our system, images are transformed into a tiled grid from which various texture and morphological features are extracted. We evaluate the contribution of high-level morphological features such as those derived from tissue segmentation algorithms as they relate to the accuracy of our classifier models. We also present work on an algorithm for tissue segmentation in image tiles, and introduce a novel feature vector representation of tissue classes in same. Finally, we present the classification accuracy, sensitity and specificity results of our system when performing three tasks: distinguishing between cancer and non-cancer tiles, between low and high-grade cancer and between Gleason grades 3, 4 and 5. Our results show that the novel tissue representation outperforms the morphological features derived from tissue segmentation by a significant margin, but that neither feature sets improve on the accuracy gained by features from low-level texture methods.

1. Introduction

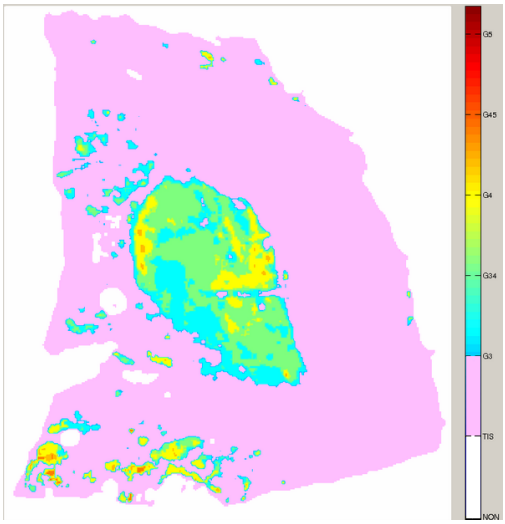
Prostate cancer is the second most common form of cancer affecting men in the western world [3]. The discovery of prostate-specific antigen (PSA) in 1980 and the implementation of prostate cancer screening methods has greatly improved detection rates, so the focus of much research has changed from detection of prostate cancer to the staging of the disease and selection of an appropriate clinical response.

Currently, the staging of prostate cancer is performed by pathologists using microscopic analysis. This process is difficult, time-consuming and can be affected by a number of factors which cause not only inter-pathologist differences, but also intra-pathologist differences (differences in subsequent analyses of the same tissue sample by a single pathologist) [4].

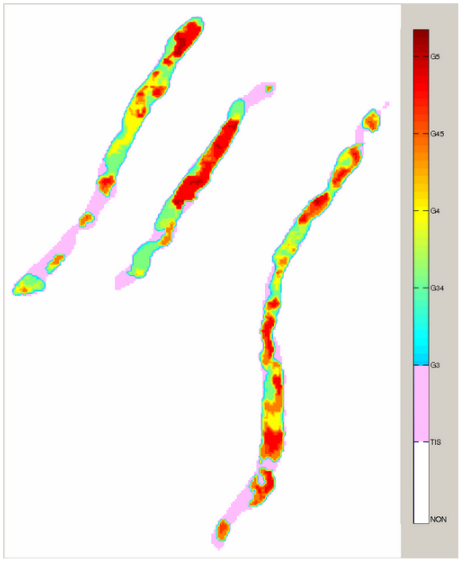
The Gleason Grading scale is the most common system used to stage the disease [12, 10]. The system operates on a scale of 1 to 5, where pattern 1 describes tissue that is most like normal prostate tissue and pattern 5 is indicative of aggressive prostate cancer. A Gleason Score is assigned to a histological sample by summing the grades of the two most prevalent patterns. Patterns 1 and 2 are rarely seen in a clinical setting, so the distinction between patterns 3, 4 and 5 are the most important in determining whether the disease is benign, indolent or aggressive.

This system described in this paper falls under the category of computer-aided diagnosis (CAD) systems. Its role in a clinical setting is not to replace the pathologist but to provide decision support, and it achieves this in a number of ways. Firstly, a heatmap of benign versus cancerous regions can be used to draw the attention of the pathologist to that region for further inspection. Other heatmap-like representations of the image classification results can show regions corresponding to different Gleason grades, while also enabling Gleason scoring of a sample.

The additional advantages of employing an automated analysis system are that it can find small foci of PCa that may have gone unnoticed by a pathologist. This also allows much greater accuracy when estimating the percentage of cancer involvement in a sample which is an important predictor in clinical outcome [1]. While not providing an authoritative answer to the



(a) Gleason-classified prostatectomy



(b) Gleason-classified biopsy

selection of a clinical response, it gives pathologists a common reference in the staging of the disease, which would help mitigate inter and intra-pathologist grading differences, and as an overall benefit reduces the amount of time spent performing microscopic analysis.

In section 3 of this paper we describe the image features that are implemented in our system and considered in this evaluation. We present some tissue morphology features that should be distinctive for different tumour stages. Surprisingly, when these features are included in the classification process they do not improve on the performance of lower level texture features taken on their own. This is perhaps an example of a situation in which a large volume of data trumps in-depth morphological analysis.

2. Related Work

Previous work in computer-aided diagnosis of prostate cancer has used a variety of methods and features to characterize and classify the disease. These can be grouped into approaches that evaluate gland morphology [13, 11] and those that employ a tiling procedure [5, 8, 7].

Diamond et al. [5] used co-occurrence texture features [9] and morphological features such as lumen area in a binary classification problem to assign image subregions into two classes corresponding to benign stroma or cancerous tissue patterns. They reported an accuracy of 79.3% evaluated on $100px^2$ tiles taken from 8 digitized tissue slides at magnification 40x.

Doyle et al. [8] used co-occurrence texture features, wavelets and first-order statistics in combination with Bayes classifiers as weak learners for an Adaboost algorithm. They reported an accuracy of 88% in discriminating cancer versus non-cancerous tissue patterns on a dataset of 22 images at magnification 40x.

Naik et al. [13] used morphological shape features in order to classify cancer grade. A Bayes classifier was used to assign each image pixel by its colour into one of three classes corresponding to cell types; lumen, nuclei and cytoplasm. Connectivity algorithms and gland-size constraints were used to detect lumen objects, and a level-set method was employed to identify the inner boundary area of the gland which consists of nuclei and cytoplasm. Eight shape descriptors were calculated for each lumen object that were then used to discriminate between the classes benign tissue and Gleason grades 3 and 4 in three separate binary classification problems. They reported accuracy rates of 86.35% in grade 3 versus benign tissue, 92.9% in grade 4 versus benign tissue and 95.19% in grade 3 versus grade 4.

DiFranco et al. [6] used co-occurrence texture features, first-order colour-channel statistics and a convolution of those features in a tile-based classification scheme. Using random-forest feature selection and an SVM classifier, they reported a highest accuracy of 95.2% when discriminating between cancer and non-cancer tiles.

Nguyen et al. [11] implemented a segmentation-based approach that made use of low-level domain information to derive structural features relating to glandular structure in benign, grade 3 and grade 4 carcinoma. Using a set of 26 images of prostate tissue sam-

ples at 20x magnification, they reported an accuracy of 85.5% when discriminating between ROIs belonging to grade 3 and grade 4 using ROI-based cross-validation and an accuracy of 70.7% using specimen-based cross-validation with an SVM classifier.

3. Methodology

Our system is based on performing a tiling operation on each digitized slide, allowing us to extract a feature vector from each tile for use in subsequent classification tasks. The advantages of this approach are its adaptability to current state-of-the-art supervised-learning methods and the ease with which image pre-processing, feature extraction and classification tasks can be parallellized.

Similar tile-based systems have used tiles of size $100px^2$ [5, 8] and $512px^2$ [7] at the same magnification level of 40x. In initial evaluations we performed tiling at $512px^2$, $256px^2$ and $128px^2$. However, the $128px^2$ tiles showed reduced accuracy in our classifier models so we settled on $256px^2$ tiles as it increases the number of 'pixels' in the subsequent tile-classified image (in which each pixel corresponds to a single input tile), allowing increased fidelity with the input image and more accurate tumour-region consolidation in post-processing steps.

3.1. Dataset Acquisition and Description

The dataset used in this work is based on digitized whole-mount prostate and biopsy histological samples supplied by Beaumont Hospital, Dublin, Ireland in conjunction with the Prostate Cancer Research Consortium (PCRC). The slides had been prepared from parafin-embedded prostates removed by radical prostatectomy and from biopsy samples, which were then stained with hemotoxylin-and-eosin (H&E). The slides were scanned on a Leica SCN400 Digital Slide Scanner at 40x magnification. Identified regions of interest (ROI) are areas containing Gleason pattern 3, 4, 5, and intermediary patterns G3/4 and G4/5. Annotation of these regions of interest was performed by a panel of pathologists (E.W.K, S.F, C.B.), who reached consensus on a label for each region.

3.2. Feature Sets

The features implemented in this system fall into three groups: colour, texture, and domain-specific (i.e. morphological).

The colour features are first-order statistics computed over colour-channel histograms in selected colour-spaces.

Texture features are the set of features that can be extracted from gray-level co-occurrence matrices (GLCM).

The domain-specific features implemented are common nuclear and tissue morphology features, derived from tissue-segmented image tiles. We also introduce a novel feature vector based on the adjacency of tissue types in an image sub-region.

3.3. Colour & Texture Features

Texture and colour features are extracted from images tile in two colour-spaces: RGB and CIELab, and using pixel quantization values of 16, 32 and 64.

In order to represent texture in an image, symmetric GLCMs are constructed for each colour-channel and quantization level at $0°$, $45°$, $90°$ and $135°$ angles from the reference pixel at a distance of 1, 2 and 4 pixels.

Thirteen texture features proposed by Haralick [9] (excepting Maximal Correlation Coefficient due to computational ineffeciency) and two additional texture features proposed by Conners [2] are extracted from each GLCM.

Nine colour-channel histogram statistics are also calculated in each colour-space and quantization level; min, mean, max, standard deviation, variance, kurtosis and skewness.

3.4. Morphological Features

This section describes the domain and morphological features implemented in this system. Morphologic features have been used to classify PCa in previous work [5, 11, 13], and as the Gleason patterns are characterized by a breakdown and fusing of gland boundaries, and infiltration of glands into stromal tissue [1, 10], inclusion of features based on this expert knowledge should increase the ability of a classifier to differentiate Gleason grades.

Tissue and nuclear morphological features are calculated from tissue-segmented images. Our method for identifying and segmenting tissue regions involves pixel-classification followed by a sequence of morphological operators, described in the algorithm below. We also describe a novel feature vector for representing discrete classes in an image.

Pixel Classification Our scheme involves assigning pixels into one of 4 classes: lumen, stroma, cytoplasm and nuclei. Representative pixels from each tissue class

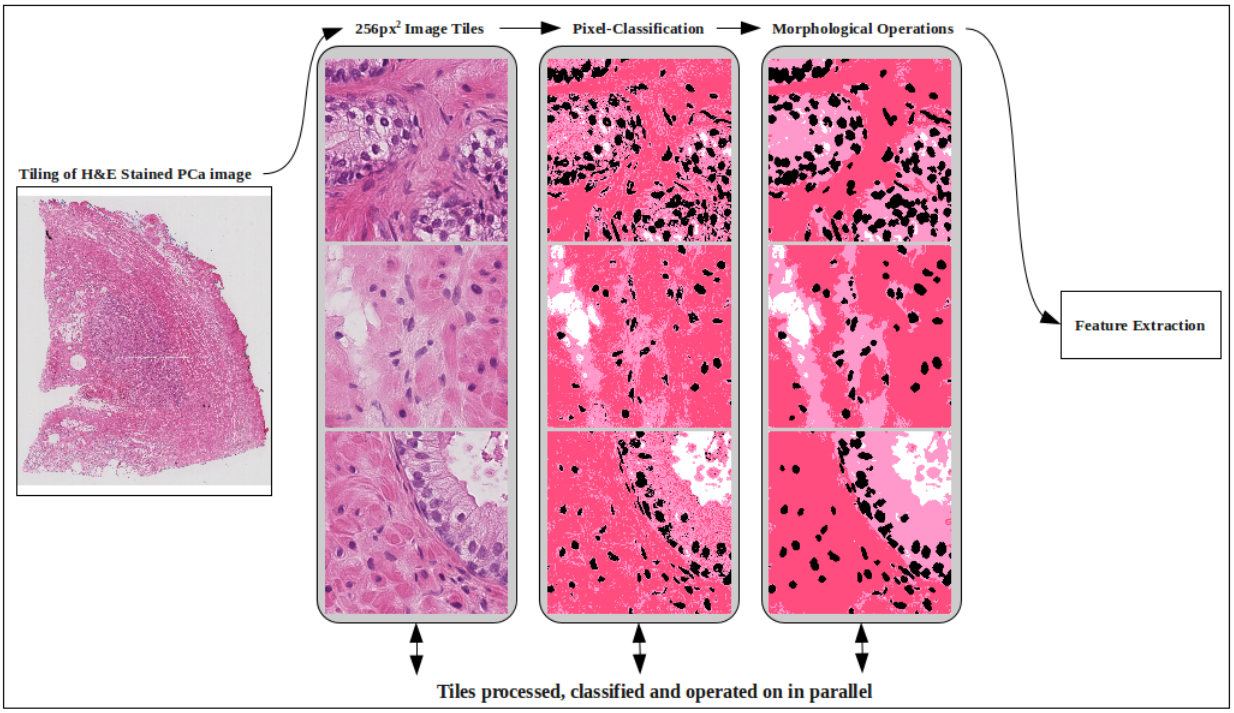


Figure 1. Pixel classification and processing pipeline

were extracted by manual ROI-selection in each image in our dataset. Outlier pixels in the stroma, cytoplasm and nuclei classes - the majority of which are white pixels incorporated by the ROI selection process - are excluded by manual thresholding. We also applied a mean-shift procedure on each pixel class that shifted each pixel towards the mean of its class.

Pixel values were sampled in the RGB colour-space, and were also converted to the CIELab, HSV, YCbCr and decorrelated H&E [14] colour-spaces.

We chose a naive Bayes model to perform pixel-classification, using the pixel intensity values of each colour channel as the input vector. Evaluation of our pixel-classifier model is performed by 10-fold cross-validation on the initial pixel dataset and by visual analysis of a set of representative tiles selected from our image dataset.

The RGB trained model achieved an accuracy of 80.1% and was generally poorer at classifying pixels in stroma and cytoplasm areas than the CIELab model on visual inspection of the tiles. The H&E colourspace was the worst performing of the colour-spaces, achieving only 73.8% accuracy. We found that the best performing input vector was the CIELab colour-space by itself, which achieved 94% accuracy, narrowly outperforming the YCbCr which achieved 93.9%, and any other concatentation of input vectors.

(a) Segmented Nuclei in Stroma and Cytoplasm Regions

In order to reduce the number of colour-space transformations needed to perform pixel classification we selected just the CIELab colour-space for our final model.

Nuclear and Tissue Morphology features When an image tile has been tissue-segmented, nuclear and tissue morphology features are extracted. These include the total area of lumen, stroma and cytoplasm tissue types, and the ratios of each tissue type to another within an image tile. Nuclear features include the total number, area and disorder of nuclei in an image tile, and the same features considering only the nuclei in stroma or cytoplasm regions within an image tile.

CICM features We have implemented a novel feature vector, dubbed class-index co-occurrence matrix (CICM) features - to distinguish it from the gray-level

Data: I, an image tile
Output: G, an indexed image such that each pixel is assigned an index in the set C, where $C = \{LUMEN, STROMA, CYTOPLASM, NUCLEI\}$

Classify pixels in $I \rightarrow$ class-indexed image CI ;
From CI, obtain binary images M_c for each $c \epsilon$ C ;

foreach M_c **do**
  if *c is* $STROMA$ **then**
    1. Morphological opening with disk structuring element of size 3.
    2. Remove regions with area $< 20px^2$, defined using 8-connectivity.
    3. Region filling using 8-connectivity.
  end
  if *c is* $CYTOPLASM$ **then**
    1. Remove regions with area $< 20px^2$, defined using 8-connectivity.
    2. Remove areas with area $< 20px^2$, defined using 8-connectivity.
    3. Region filling using 8-connectivity.
  end
  if *c is* $NUCLEI$ **then**
    1. Remove nuclei regions with area $< 10px^2$.
    2. Morphological opening with disk structuring element of size 3.
    3. Dilation with disk structuring element of size 2.
    4. Region filling using 8-connectivity.
  end
end
Apply each image M_c as indexed layers in the output image G in the following order: ;

1. $STROMA$
2. $CYTOPLASM$
3. $LUMEN$
4. $NUCLEI$

Label any unmarked pixels with index from CI ;
Apply 3x3 median-filter ;

Algorithm 1: Tissue-Segmentation

co-occurrence matrix (GLCM). These are created by constructing an symmetric co-occurrence matrix of distance 1 on an image where each pixel has been assigned a class index in a set. In our scheme, pixels are assigned to 4 classes; lumen, stroma, cytoplasm and nuclei, thus the adjacency matrix constructed is a count of how many nuclei pixels are adjacent to stroma pixels, etc. Values in the matrix are normalized to unit range, and the upper triangle matrix values are extracted to form a feature vector.

As this feature vector then constitutes a form of second-order statistics based on the adjacency of discrete tissue types in an image, and thus the 'pattern' of tissue types in the image it should offer some discriminatory power for distinguishing between the different tissue patterns found in Gleason grades.

From our pixel classification and tissue segmentation procedures we can draw two images: the pixel-classified image and the tissue-segmented image. This method is applicable to both images, and initial evaluations showed comparable accuracy rates for feature vectors calculated on both images. Adhering to Occam's razor, we chose to continue with the simpler model in our evaluations. Accordingly, the accuracy rates reported in the results section are those based on a simple pixel classification scheme - without further morphological processing - from which the CICM feature vector is drawn.

4. Results

Our dataset is made up of 20 digitized prostatectomy and biopsy images, consisting of a total of 1,760,789 $256px^2$ tiles at 40x magnification. There are in total 108,513 tiles labeled with a Gleason grade by our annotation procedure.

We evaluated each feature set using leave-images-out cross-validation. In this method, we remove all samples of a class belonging to one or more images and use them as the test set, as this ensures no tiles from the same image are used to train and test a classifier. As an image may contain tiles of one class and few or none of the opposing class, minor accuracy rates are calculated for all folds.

Initial evaluations showed that texture and colour features calculated from the CIELab colourspace outperformed those from the RGB colour-space, so the Colour Histogram and Haralick accuracy rates reported below correspond only to those given by the CIELab colour-space.

The number of features in each set is given in square brackets after each feature set in the tables below. The model we used for classification was a Random Forest

model, and no feature selection was performed.

4.1. Cancer vs Noncancer

In this problem, we assign all of our annotated Gleason grade areas to the 'Cancer' class. 'Noncancer' are a random selection of tiles from unannotated image areas until classes are balanced, and are therefore made up of benign stroma and benign gland regions.

Feature set	Accuracy	Sensitivity	Specificity
ALL [242]	77.57%	0.76	0.78
Colour Histogram [81]	76.04%	0.75	0.76
Haralick [135]	77.65%	0.75	0.78
Morphological [16]	72.75%	0.71	0.73
CICM [10]	76.26%	0.74	0.77

4.2. Gleason grades and Low vs High

Here, we present the accuracy rates for binary classification of each Gleason grade. Also included are rates for classifying low-risk versus high-risk areas, in which low-risk constitutes Gleason grades 3 and intermediary pattern 3/4 that have been assigned to Gleason 3, and high-risk constitutes the Gleason 4, 5 and intermediary pattern 4/5.

Feature set	G3-G4	G4-G5	LOW-HIGH
ALL [242]	82.10%	91.81%	79.26%
Colour Histogram [81]	78.12%	87.71%	77.28%
Haralick [135]	80.99%	90.68%	78.26%
Morphological [16]	62.38%	79.69%	69.85%
CICM [10]	75.25%	83.94%	74.11%

5. Conclusion

As CAD systems such as the one described in this paper move towards implementation in clinical practice, it is worth considering the implementation cost with regards to data transfer and feature extraction, the most likely bottlenecks in a high-data, high-throughput system. Based on the figures reported here, it appears that morphological features are probably not worth the extra computation required to obtain them. It is also worth noting that the accuracy rates reported are done without feature selection or classifier parameter tuning, and including these steps would almost certainly improve accuracy rates.

5.1. Acknowledgements

The authors wish to acknowledge and thank Prof. Elaine Kay, Dr. Stephen Finn and Dr. Ciara Barrett for their expert knowledge and assistance in annotating the dataset used in this work. We would also like to thank the Prostate Cancer Research Consortium and Prof. William Watson, without whom this work would not have been possible.

References

[1] M. Che, W. Sakr, and D. Grignon. Pathologic features the urologist should expect on a prostate biopsy. *Urologic oncology*, 21(2):153–61, 2003.

[2] R. W. Conners, M. Trivedi, and C. Harlow. Object detection based on gray level cooccurrence. *Computer Vision, Graphics and Image Processing*, pages 25:273–310, 1984.

[3] J.-E. Damber and G. Aus. Prostate cancer. *Lancet*, 371(9625):1710–21, May 2008.

[4] a. De La Taille. Evaluation of the interobserver reproducibility of gleason grading of prostatic adenocarcinoma using tissue microarrays. *Human Pathology*, 34(5):444–449, May 2003.

[5] J. Diamond, N. H. Anderson, P. H. Bartels, R. Montironi, and P. W. Hamilton. The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia. *Human Pathology*, 35(9):1121–1131, Sept. 2004.

[6] M. D. Difranco, G. O. Hurley, E. W. Kay, and R. G. William. Generating a Heatmap of Suspected Prostatic Carcinoma Using Global , Illumination-invariant Texture Features. Workshop proceedings., 2009.

[7] M. D. DiFranco, G. O'Hurley, E. W. Kay, R. W. G. Watson, and P. Cunningham. Ensemble based system for whole-slide prostate cancer probability mapping using color texture features. *Computerized medical imaging and graphics*, 35(7-8):629–45, 2011.

[8] S. Doyle and M. Hwang. Automated grading of prostate cancer using architectural and textural image features. *IEEE International Symposium on Biomedical Imaging*, pages 1284–1287, 2007.

[9] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(6):610–621, Nov. 1973.

[10] P. Humphrey. Gleason grading and prognostic factors in carcinoma of the prostate. *Modern pathology*, 17(3):292—-306, Mar 2004.

[11] N. K, S. B, and J. A. K. Prostate cancer grading: Gland segmentation and structural features. *Pattern Recognition Letters*, 33(7):951–961, May 2012.

[12] A. Lopez-Beltran, G. Mikuz, R. J. Luque, R. Mazzucchelli, and R. Montironi. Current practice of Gleason grading of prostate carcinoma. *Virchows Archiv*, 448(2), Feb. 2006.

[13] S. Naik, A. Madabhushi, J. Tomaszeweski, and M. D. Feldman. A quantitative exploration of efficacy of gland morphology in prostate cancer grading. *2007 IEEE 33rd Annual Northeast Bioengineering Conference*, pages 58–59, Mar. 2007.

[14] C. Ruifrok and D. Johnston. Quantification of histochemical staining by color deconvolution., Aug. 2001.