Ensemble Prostate Tumor Classification in H&E Whole Slide Imaging via Stain Normalization and Cell Density Estimation

Michaela Weingant[1], Hayley M. Reynolds[2,3], Annette Haworth[2,3], Catherine Mitchell[4], Scott Williams[5], and Matthew D. DiFranco[6(✉)]

[1] Vienna University of Technology, Vienna, Austria
[2] Department of Physical Sciences, Peter MacCallum Cancer Centre, Melbourne, Australia
[3] Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Australia
[4] Department of Pathology, Peter MacCallum Cancer Centre, Melbourne, Australia
[5] Division of Radiation Oncology and Cancer Imaging, Peter MacCallum Cancer Centre, Melbourne, Australia
[6] Center for Medical Physics and Biomedical Engineering, Medical University Vienna, Vienna, Austria
matthew.difranco@meduniwien.ac.at

Abstract. Classification of prostate tumor regions in digital histology images requires comparable features across datasets. Here we introduce adaptive cell density estimation and apply H&E stain normalization into a supervised classification framework to improve inter-cohort classifier robustness. The framework uses Random Forest feature selection, class-balanced training example subsampling and support vector machine (SVM) classification to predict the presence of high- and low-grade prostate cancer (HG-PCa and LG-PCa) on image tiles. Using annotated whole-slide prostate digital pathology images to train and test on two separate patient cohorts, classification performance, as measured with area under the ROC curve (AUC), was 0.703 for HG-PCa and 0.705 for LG-PCa. These results improve upon previous work and demonstrate the effectiveness of cell-density and stain normalization on classification of prostate digital slides across cohorts.

Keywords: Machine learning · Digital pathology · Tumor prediction · Prostate · Cell counting

1 Introduction

A variety of clinical and research applications would benefit from computer aided methods for accurate and reliable tumor classification in whole mount

H.M. Reynolds—Funded by a Movember Young Investigator Grant awarded through Prostate Cancer Foundation of Australia Research Program.

A. Haworth—Supported by PdCCRS grant 628592 with funding partners: Prostate Cancer Foundation of Australia; Radiation Oncology Section of the Australian Government of Health and Ageing; Cancer Australia.

L. Zhou et al. (Eds.): MLMI 2015, LNCS 9352, pp. 280–287, 2015.
DOI: 10.1007/978-3-319-24888-2_34

histology images. Current fields of research include computer aided methods to classify tumor location [3] and correlation studies assessing histology ground-truth data with multi-parametric MRI (mpMRI) [1,5].

Computer-aided interpretation of prostate H&E histopathology images relies heavily on color information. However, the chromatic appearance of digital whole-slide imagery is subject to many sources of variability, such as manufacturer-dependent staining agents for dying the tissue, institute-dependent staining protocols and hardware-dependent scanning conditions [8]. Stain normalization using color deconvolution, in which RGB pixel data is transformed into stain-specific color channels, helps overcome this obstacle. Fixed stain vector values have been determined for a range of histology stains, including H&E [10], more recent publications proposed image-specific stain vectors [2,6,7].

The use of color and texture-based features for predicting LG-PCa and HG-PCa on whole slide images is desired due to the efficiency and simplicity of feature extraction on such large images. However, a recent study [4] reported unsatisfactory results (AUC=0.632 for HG-PCa, 0.486 for LG-PCa) when training a classifier on cohort and testing on another, suggesting the need for stain-normalized color and texture features, as well as more descriptive measures such as cell density [9].

In this work we aimed to improve on the cross-cohort classifier performance from [4] by (1) including a cell density feature into the classification framework and (2) by applying stain normalization as proposed in [7] prior to color and texture feature extraction in order to account for variations in H&E staining (Fig. 1).

2 Methods

Datasets. Three H&E-stained datasets were used in this work. The first, available as a supplement to [11], has nuclei annotations and served as the ground-truth for developing and quantifying the optimized cell density estimation. It consisted of 36 H&E image-tiles of 600x600 pixels from multiple human tissue sites with a resolution of 0.23 m per Pixel (MPP).

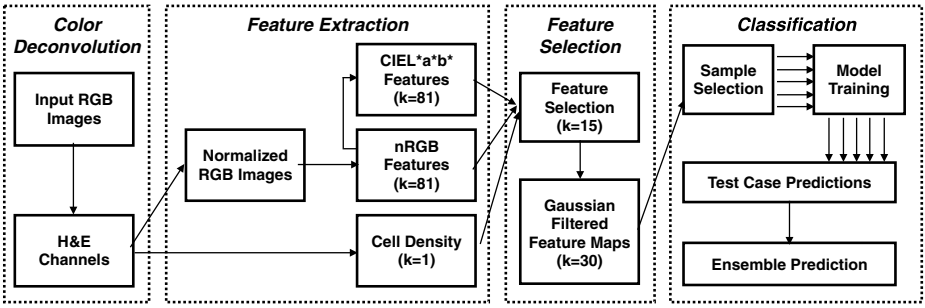


Fig. 1. Flowchart of normalization and cell density within classification framework

Table 1. Number of tiles from annotated regions for each image, with number of annotated detailed ROIs in parentheses.

Sample	Non-Cancer	LG-PCa	HG-PCa	Sample	Non-Cancer	LG-PCa	HG-PCa
A01	14939 (17)	924 (5)	1444 (10)	B01	855 (6)	646 (9)	11 (4)
A02	14770 (14)	63 (3)	668 (18)	B02	2073 (49)	1 (1)	36 (11)
A03	25673 (8)	0 (0)	319 (3)	B03	0 (0)	4516 (2)	0 (0)
A04	6499 (7)	900 (4)	866 (16)	B05	6337 (15)	95 (12)	80 (10)
A05	9668 (6)	445 (4)	221 (4)	B06	16569 (18)	159 (18)	458 (20)
A06	1898 (7)	951 (4)	290 (3)	B07	1026 (15)	411 (15)	14 (3)
A07	0 (0)	775 (9)	453 (3)	B08	14167 (20)	114 (24)	0 (0)
A08	2462 (13)	845 (7)	719 (3)	B09	2335 (13)	174 (10)	34 (5)
A09	8969 (9)	800 (3)	0 (0)	B10	12525 (23)	286 (15)	106 (13)
A10	11159 (7)	39 (1)	914 (4)				
A11	10684 (5)	450 (1)	2190 (12)				
A12	5622 (16)	1273 (2)	1087 (2)				
A13	1310 (1)	800 (8)	94 (3)				
A14	3613 (7)	608 (2)	0 (0)				
Cohort A	117266 (117)	8873 (53)	9265 (81)	Cohort B	55887 (159)	6402 (106)	739 (66)

The second and third dataset, called *Cohort A* and *Cohort B* and previously described in [4], consisted of post-resection prostate H&E digital slides. Cohort A consisted of 14 H&E prostate histopathology whole slide images from one pathology center, taken at 400x magnification and stored with 0.238 MPP. Cohort B was composed of 9 H&E prostate histopathology whole slide images taken at 200x magnification and stored with 0.504 MPP from a different pathology center. Each cohort was annotated by a different expert pathologist.

These two datasets included annotated ROIs, refered to as *detailed* in the results, of homogeneous high-grade, low-grade and benign tissue patterns used here to define training and testing data for various supervised classification scenarios. Tile and region counts for these detailed ROIs are shown in Table 1. For Cohort B, we also have clinical ROIs from the original glass slides which delineate tumor and non-tumor regions more generally, and we refer to these in the results as *clinical* ROIs.

H&E Color Deconvolution and Stain Normalization. In order to mitigate the effects of H&E stain variation between our datasets we included a color normalization method based on [7]. The normalization process involves identification of image-specific stain-vectors, stain deconvolution, and finally normalization to a target appearance in RGB space.

For the stain vector identification, the RGB values of the image in question were first converted to optical density (OD) [10]. The transformation is calculated as $OD = -log_{10}(I)$ where I holds the RGB intensity values of each pixel, normalized to [0 1]. After identifying the image-specific stain vectors V according to [7], the deconvolution was realized using the equation $C = V^{-1}OD$ on

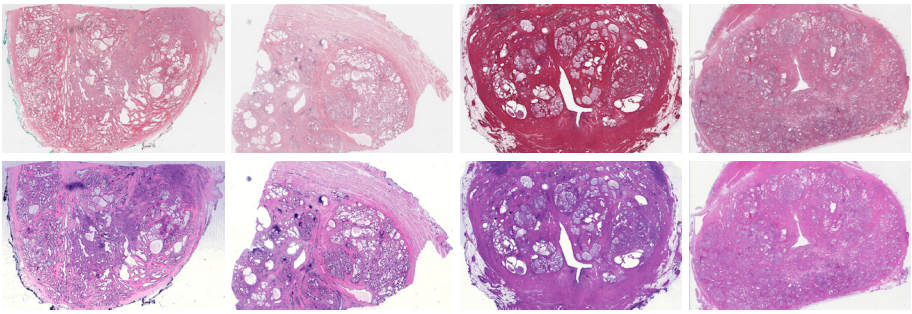


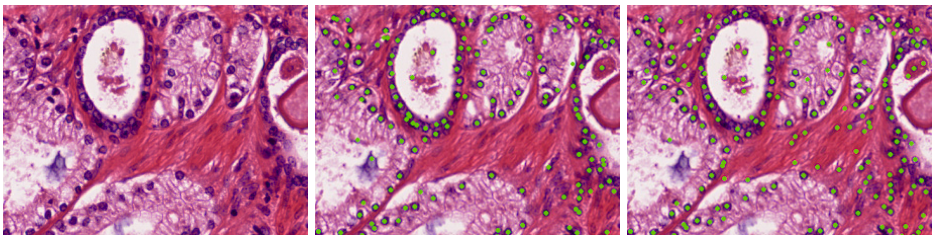
Fig. 2. Two images from Cohort A (left, middle left) and two images from Cohort B (middle right, right). Top row shows original staining, bottom row shows normalized staining.

all OD-tuples, where the three-channel matrix C contains the intensity in every pixel for each of the stains (i.e. H, E and a vector orthogonal to H and E). The results of color deconvolution were used in the cell density algorithm.

Finally, the normalized images I_{norm} were computed using $I_{norm} = e^{-V_{target}*C}$, where V_{target} denotes the stain-matrix containing the desired appearance. The effect of the normalization can be seen in Fig. 2. The previously heterogeneous appearances of different images are transformed to a commonly used uniform appearance, applying a standard H&E vector V_{target} from [10]. Normalized images were used for subsequent color and texture feature extraction (Fig. 1).

Cell Density Estimation. The parameters used in [9] had been empirically tuned to images from Cohort B using a subset of representative benign and tumor tiles. Here, we aimed to determine the parameters more objectively to ensure they were suitable for application across data from multiple centers rather than biased towards one particular center. The optimization was conducted via a grid-search of the parameter space, using the multiple-site dataset from [11] with per-nucleus annotations to quantify the respective outcome. The five parameters were: (1) the radius r_1 of an object to be accepted as a nucleus candidate in the radial symmetry transform (2) the roundness constraint α, which lies between 0, allowing an arbitrary shape and 4, allowing only strictly round shapes (3) the minimum intensity gradient threshold γ for pixels contributing to the symmetry measurement (4) the region radius r_2 of the non-maxima suppression (NMS) and (5) a threshold t below which local maxima are ignored as a maximum candidate for the NMS.

The final parameters were chosen considering two metrics: (1) an adjusted recall measure ($Recall + \frac{FalsePositives}{TruePositives+FalseNegatives}$) of close to 1.5, and (2) a high F_2-scores (similar to the F-score, but weighting recall twice as important as precision).



A B C

Fig. 3. Nuclei identification on a sample image from Cohort B: (A) original ROI (B) using fixed value deconvolution and empirically determined parameters according to [9] (C) using the proposed adaptive deconvolution and optimized parameters.

Classification Framework. Our approach to whole-slide classification comes from [3], and our training scenarios are the same as [4]. Here we use stain normalization prior to calculating color channel histogram and gray level co-occurrence texture features for each tile, and we also include cell density in the feature vector. Images were divided into non-overlapping tiles of approximately 127 m, and for each tile cell density ($k_{cd} = 1$) co-occurence texture features ($k_{co} = 21$) as well as histogram-based features ($k_{hist} = 6$) for each of the six channels of the RGB and CIEL*a*b* colorspaces were calculated, for a total of ($k = 163$) features per tile. Expert annotations were used to assign a class label to each training tile based on the underlying tissue class. For each classification scenario, a Random Forest based feature selection was performed, and the 15 highest ranking features based on Gini-importance were chosen to build a training model. After feature selection, 2-D Gaussian filters ($\sigma = 2.4, 2.8$) were applied to generate smoothed feature maps for inclusion into training models.

For each classification model, up to 3000 tiles were randomly sampled, with the constraint that each ROI from the training images was represented in the training set. Classifiers were trained using radial basis function support vector machines (RBF-SVMs) with $C = 2^{-2}$ and $\gamma = 2^{-9}$. For each training scenario, multiple models were generated and used as an ensemble for pooling predictions on the test sets. For each query image, predictions always based on training models that did not include that image. In total, we ran 5 training scenarios: 01 and 02 used only data from Cohort A, 04 and 05 used only data from Cohort B, and 03 used data from both cohorts. Classifier performance is reported as area under the ROC curve (AUC) for each cohort of images.

3 Results

The optimized parameters and the updated metrics for the proposed cell density method in comparison to [9] are summarized in Table 2. An illustration of the nuclei detection on a sample image before and after optimization is shown in Fig. 3.

Classification results are shown in Table 3. Scenario 03 includes data from both cohorts, and we see comparable AUC values for LG-PCa classification between detailed annotations from cohorts A and B (0.854 and 0.870, respectively). For HG-PCa, we note that scenario 03 performs better on Cohort B than Cohort A (0.952 vs. 0.875). For classification of Cohort B using Cohort A for training (Runs 01 and 02), an AUC of 0.705 and 0.657 for LG-PCa and 0.686 and 0.703 for HG-PCa is achieved. Classification of Cohort A using Cohort B as training data (Runs 04 and 05) produces AUC values of 0.563 and 0.568 for LG-PCa and 0.484 and 0.487 for HG-PCa.

Fig. 4(A) shows an example heat map using our ensemble results for image B05 which was clinically annotated with a large, heterogeneous lesion (blue outline) which was given a score of Gleason 3+4. Fig.s 4(B) and (D) show results for LG-PCa (yellow) and HG-PCa (blue) from scenario 02, while figures 4(C) and (E) are from scenario 03. Scenario 02 detected regions of low-grade tumor within the larger tumor, but completely failed to detect high-grade tumor. Scenario 03 was able to detect both low-grade and high-grade tumor regions.

4 Discussion

In this study, we have adapted the cell density algorithm of [9] by introducing adaptive color deconvolution and by optimizing the parameters for cell detection (see Table 2) based on a validation dataset from [11]. Our method improved the F2-score and moved our adjusted recall metric to a desired median detection rate of 1.5. Fig. 3 illustrates that the parameters in [9] led to a detection-bias for nuclei near ducts, compared to the presented parameter setting, which detects nuclei more homogeneously across multiple tissue sites. The latter promises a more stable and reliable cell density estimation across whole slides, seeing as the presence of ducts will not distort the local density in a ROI.

Although the cell density measure overestimates the true cell count per tile, it presents a meaningful estimation for comparison between different tiles of a whole-slide image. A true quantitative verification of the nuclei detection on

Table 2. Update of the parameters and quantifying metrics (MAD = Median absolute deviation) based on the dataset from [7] with 0.23 m per pixel.

	Reynolds et al. [9]	proposed
Parameters		
r_1, radius of the object in question in px	5:8	5:8
α, roundness constraint	4	3
γ, minimum intensity gradient threshold	15	15
r_2, region radius for NMS	1	2
t, minimum threshold for being a NMS candidate	1.5	1.6
Metric		
Adj. Recall (Median / MAD)	1.13 / 0.20	1.55 / 0.19
F2-score (Median / MAD)	0.63 / 0.11	0.68 / 0.06

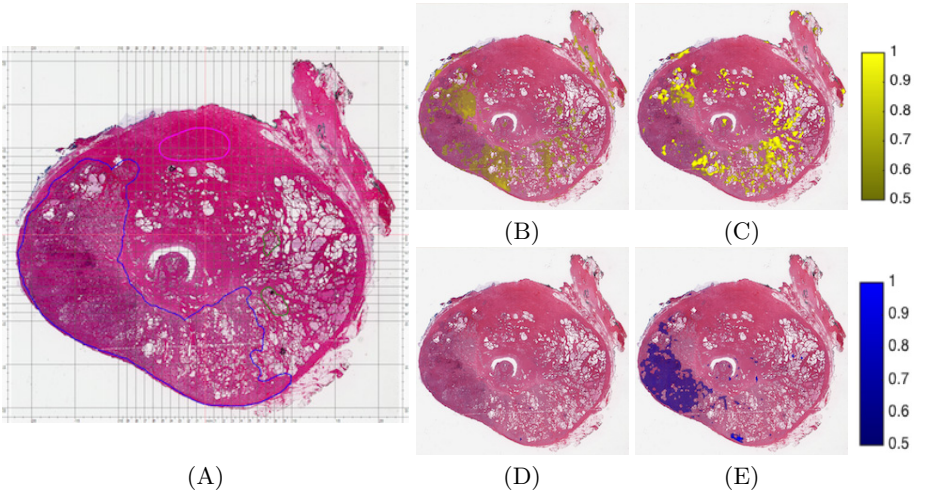


Fig. 4. Prediction maps of LG-PCa (B and C) and HG-PCa (D and E) for imageB05 in Cohort B, along with original annotations (A). Note that for the large tumor region (blue ROI in A), the classification has separated the tumor into regions of HG tumor (E, bottom left) and LG tumor (B,C). The failure to detect tumor in D indicates the inability of the model built using Cohort A data to detect HG-tumor in Cohort B.

prostate images such as those from Cohort A and B would be beneficial, but would require tedious manual cell counting.

Classifier performance using mixed cohorts (scenario 03) was comparable to that from [4]. However, training with Cohort A and testing with Cohort B shows an improvement of 21% for LG-PCa and 8% for HG-PCa, demonstrating the impact of including stain normalization prior to texture and histogram feature extraction.

Results for training with Cohort B and testing with Cohort A only achieved 56% and 48% AUC, respectively. This discrepancy likely arises from the differences in annotation detail levels between the two cohorts, since ROIs were drawn by two different pathologists. Furthermore, smaller, more specific ROIs

Table 3. Area under the ROC curve (AUC) for classification of both LG-PCa and HG-PCa for images from Cohort A and B using both clinical and detailed ground truth annotations.

	Low-grade (LG-PCa)			High-grade (HG-PCa)		
Run	Detailed A	Detailed B	Clinical B	Detailed A	Detailed B	Clinical B
01	0.910	0.705	0.642	0.834	0.686	0.760
02	0.893	0.657	0.682	0.882	0.703	0.674
03	0.854	0.870	0.711	0.875	0.952	0.847
04	0.563	0.968	0.756	0.484	0.978	0.842
05	0.568	0.961	0.748	0.487	0.961	0.808

were drawn from cohort B in comparison to cohort A. The difference in magnification may also play a role here.

5 Conclusions

Improving the reliability of whole-slide image classification is of critical importance for the adoption of this technology into pathology workflows. The results presented here show a promising improvement in classification of LG- and HG-PCa in whole slide H&E images when training and testing on separate cohorts. This improvement can be attributed to the inclusion of stain normalization for texture-based features, as well as the introduction of a cell-density feature. Future work to perform more rigorous parameter tuning, feature extraction and feature selection should lead to further improvements in performance.

References

1. Borren, A., Groenendaal, G., Moman, M.R., Boeken Kruger, A.E., van Diest, P.J., van Vulpen, M., Philippens, M.E.P., van der Heide, U.A.: Accurate prostate tumour detection with multiparametric magnetic resonance imaging: Dependence on histological properties. Acta Oncol. **53**(1), 88–95 (2014)
2. Cosatto, E., Mille, M., Grad, H.P., Meyer, J.S.: Grading nuclear pleomorphism on histological micrographs. In: 19th Int. Conf. Pattern Recogn. (2008)
3. DiFranco, M.D., O'Hurley, G., Kay, E.W., Watson, W.G., Cunningham, P.: Ensemble based system for whole-slide prostate cancer probability mapping using color texture features. Comput. Med. Imaging Graph. **35**, 629–645 (2011)
4. DiFranco, M.D., Reynolds, H.M., Mitchell, C., Williams, S., Allan, P., Haworth, A.: Performance assessment of automated tissue characterization for prostate H and E stained histopathology. In: SPIE Medical Imaging, p. 94200M (2015)
5. Gibbs, P., Liney, G.P., Pickles, M.D.: Correlation of ADC and T2 measurements with cell density in prostate cancer at 3.0 Tesla. Invest. Radiol. **44**(9), 572–576 (2009)
6. Khan, A.M., Rajpoot, N., Treanor, D., Magee, D.: A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. IEEE Trans. Biomed. Eng. **61**(6), 1729–1738 (2014)
7. Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, C.: A method for normalizing histology slides for quantitative analysis. In: Proceedings of IEEE ISBI 2009, pp. 1107–1110 (2009)
8. McCann, M.T., Ozolek, J.A., Castro, C.A., Parvin, B., Kovacevic, J.: Automated histology analysis: Opportunities for signal processing. IEEE Signal Process. Mag. **32**(1), 78–87 (2015)
9. Reynolds, H.M., Williams, S., Zhang, A.M., Ong, C.S., Rawlinson, D., Chakravorty, R., Mitchell, C., Haworth, A.: Cell density in prostate histopathology images as a measure of tumor distribution. In: SPIE Medical Imaging, p. 90410S (2014)
10. Ruifrok, A.C., Johnston, D.A.: Quantification of histochemical staining by color deconvolution. Analyt. Quant. Cytol. Histol. **23**, 291–299 (2001)
11. Wienert, S., Heim, D., Saeger, K., Stenzinger, A., Beil, M., Hufnagl, P., Dietel, M., Denkert, C., Klauschen, F.: Detection and segmentation of cell nuclei in virtual microscopy images: A minimum-model approach. Scientific Reports **2**, 503 (2012)