COMPUTER-AIDED GLEASON GRADING OF PROSTATE CANCER HISTOPATHOLOGICAL IMAGES USING TEXTON FORESTS

Parmeshwar Khurd, Claus Bahlmann, Peter Maday, Ali Kamen
Siemens Corporate Research,
Princeton, NJ, USA

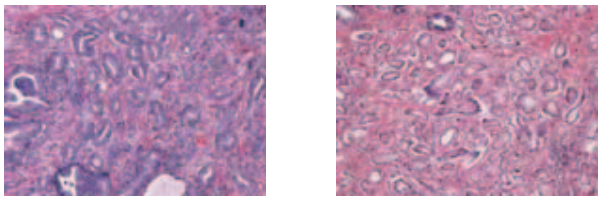
Summer Gibbs-Strauss, Elizabeth M. Genega, John V. Frangioni[1]
Beth Israel Deaconess Medical Center,
Boston, MA, USA

ABSTRACT

The Gleason score is the single most important prognostic indicator for prostate cancer candidates and plays a significant role in treatment planning. Histopathological imaging of prostate tissue samples provides the gold standard for obtaining the Gleason score, but the manual assignment of Gleason grades is a labor-intensive and error-prone process. We have developed a texture classification system for automatic and reproducible Gleason grading. Our system characterizes the texture in images belonging to a tumor grade by clustering extracted filter responses at each pixel into textons (basic texture elements). We have used random forests to cluster the filter responses into textons followed by the spatial pyramid match kernel in conjunction with an SVM classifier. We have demonstrated the efficacy of our system in distinguishing between Gleason grades 3 and 4.

Index Terms— Gleason grading, prostate cancer, texture classification

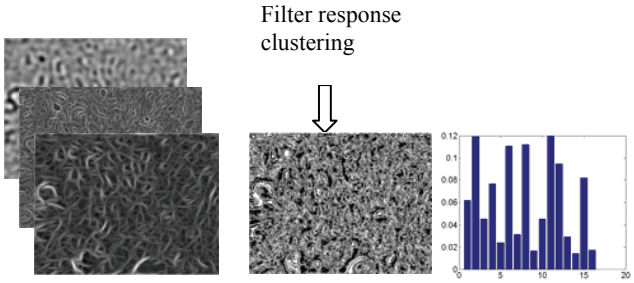
1. INTRODUCTION



(a) Gleason 3 (b) Gleason 4

Fig. 1: H&E images of cancerous prostate tissue

Prostate cancer is the second leading cause of death in American men, after lung cancer. Candidates suspected to have prostate cancer commonly undergo tissue biopsy in order to assess the presence and aggressiveness of cancer. The biopsied tissue samples are imaged with a microscope after hematoxylin and eosin (H&E) staining and assigned tumor grades according to the Gleason grading system (grades 1-5)[12]. In Fig. 1, we have shown H&E images of prostate cancer tissue samples corresponding to Gleason grades 3 and 4. The Gleason grade characterizes tumor differentiation, i.e., the degree to which the tumor resembles healthy tissue. The sum of the primary and the secondary Gleason grades yields the Gleason score, the single most important prognostic indicator for prostate cancer patients. The Gleason score plays an important role in deciding the future course of treatment. However, the assignment of Gleason scores is a time-consuming, error-prone process that depends upon the samples obtained during core biopsy as well as on the expertise of the pathologist. One way to validate the Gleason score obtained during core biopsy is to re-calculate this score in patients who undergo radical prostatectomy, thereby eliminating the sampling error. This form of validation is also required in order to study the correlation of prostate cancer biomarkers observed in other macroscopic imaging modalities with the Gleason score [1]. Calculation of Gleason scores on the entire prostate specimen can also help in the design of a procedure for selecting optimal biopsy locations [2]. Computer-aided Gleason grading becomes essential when we need to assign tumor grades to the entire prostate specimen. Therefore, we have developed a computer-aided system to assign Gleason grades in an automatic and reproducible manner.



Filter responses Texton map Texton histogram

Fig. 2: Texton histogram for an input H&E image

Our system relies on accurate texture characterization and classification in order to automatically compute the Gleason grade of an imaged prostate specimen. We use filtering followed by clustering in order to characterize

[1] This work was supported by grant 1R01CA134493-01A1 from the NIH.

textures via basic texture elements or textons. The distribution of these textons provides a discriminative signature for each tumor grade and is used as the input to a support vector machine (SVM) classifier. For improved accuracy and speed, we have used random forests for clustering and the spatial pyramid match kernel in the SVM classifier. In Sec. 2, we describe our texture classifier in detail and use it to distinguish between Gleason grades 3 and 4 in Sec. 3, where we also provide comparisons with alternative clustering and classification methods. We conclude with some final remarks in Sec. 4.

2. METHODS

2.1. Texture Classification Framework

We follow the texton-based texture classification framework introduced in [3] as the basis for our work and briefly review this approach below: For every image belonging to a specific texture, an appropriate rotationally invariant filter bank was first used to extract responses at each pixel. The high-dimensional feature space of filter responses provides an accurate description of the texture characteristics of each image, but for accurate classification, we need a sparser compact representation that preserves the information content. Therefore, clustering using the K-means algorithm was performed in this filter response space in order to identify basic texture elements or textons for each texture class. The cluster centers from the different texture classes were then concatenated and texton maps, i.e., cluster (texton) assignments at each pixel, for each textural image were obtained. It was shown in [3] that the spatial histogram of the texton map provides a rich discriminative signature for each textural image. An example of a few filter response images, a texton map and a texton histogram corresponding to the input H&E image in Fig. 1(a) is shown in Fig. 2. Several model histograms for each texture class were stored during training and the texture class for an image during the test phase was assigned using a K-NN classifier by comparing the texton histogram of the test image with the model histograms using the χ^2 distance.

Building upon the approach in [3], we replace the clustering and classification algorithms with potentially more powerful alternatives described below. We continue to use the MR-8 filter bank introduced in [3].

2.2. Texton Identification using Random Forests

Since clustering using the K-means framework is slow during the training phase, tree-structured alternatives are used for speed-up. During the training phase, a binary clustering tree uses suitably determined binary splits in order to recursively divide the entire training set and each training sample ends up being assigned to a tree leaf. The same leaves can then be used for cluster assignment during the testing phase by using the recursive binary splits determined during training. Instead of relying upon a single space-partitioning tree to characterize the data, random forests (RF) [5,6] train an ensemble of trees to capture a richer description of the input feature space. We shall use two types of random trees, namely, random projection trees (RPT) [4] and extremely randomized decision trees (ERT) [5,6], as components within our random forests.

RPTs use two forms of binary-splits, viz., projection splits and distance splits. A projection split picks a random direction for projecting the input features and splits them about the median projection, the underlying intuition being that a randomly chosen direction can yield almost as effective a projection direction as the optimal one determined via principal component analysis. A distance split is only used when the distance between the farthest two points is significantly larger than the average distance between any two points, suggesting that no single direction can provide a good split. In the distance split, the median distance from the mean of all input features is used to perform a spherical split of the feature space. It is shown in [4] that RPTs can adapt to the local covariance dimension of the data and can hence yield a more accurate description of the features in case they lie in a non-linear subspace.

ERTs split the data at each node by picking the best feature component (filter response) and threshold from a set of randomly chosen feature components and thresholds so that a measure of information gain is maximized. We use the following definition [6] of the information gain of a feature component f split at s:

$$IG_s(f) = -\frac{n_{left}}{n} Entropy_{left} - \frac{n_{right}}{n} Entropy_{right},$$

where n denotes the total number of samples at a node, n_{left} denotes the number of samples with $f < s$, n_{right} denotes the number of samples with $f \geq s$ and $Entropy_A$ denotes the Shannon entropy of the class labels in set A. As mentioned in [5,6], randomized decision trees have an advantage over RPTs and K-means clustering because they use the class labels to find discriminative space partitions during training. Although an ERT is only used for space-partitioning and not for classification during the test phase, its corresponding texton map should be more useful during the subsequent SVM classification stage described in Sec. 2.3. However, ERTs only use a single feature component for splitting the data, whereas RPTs utilize the information from all feature components during the projection or distance splits. Therefore, we shall use both RPTs and ERTs in our experiments in Sec. 3.

2.3. SVMs and the Spatial Pyramid Match Kernel

The 2-class SVM [14] is designed to find a max-margin linear classifier separating the classes (e.g., grade 3 vs. grade 4) in a higher-dimensional feature space. An appropriately selected kernel K, designed to measure the similarity between any two input features (texton maps or

texton histograms), controls the mapping from the input features into the higher-dimensional space.

In order to accurately represent the differences between the multi-scale spatial content present in two different visual word (texton) maps, the positive-definite spatial pyramid match kernel (SPM) was introduced in [7]. The SPM kernel computed over L+1 levels for any two texton maps P and Q is given by:

$$K(P,Q) = \mathrm{I}_L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}} (\mathrm{I}_l - \mathrm{I}_{l+1}),$$

where the histogram intersection $\mathrm{I}_l = \sum_j \min(P_l[j], Q_l[j])$, and P_l and Q_l are portions of histograms at level l with j indexing all nodes at level l. Since each tree in a random forest yields its own texton map, as in [6], we use the mean SPM kernel value over all trees.

3. RESULTS AND DISCUSSION

Our prostate cancer dataset consisted of 25 H&E images of Gleason grade 3 and 50 images of Gleason grade 4. Each image was acquired at 10X resolution with 0.625 micron pixel size and was of size 1392x1040 pixels. (We note that a few of our grade 4 images contained small regions belonging to other tumor grades, but we still assigned the grade 4 label to these images.)

For training our clustering algorithms, we selected 30 images (15 of each grade). As in [3] [2], using the K-means algorithm, filter responses from 50000 randomly chosen pixels in each of the 15 images belonging to one class were clustered to yield 8 centers, resulting in a total of 16 concatenated cluster centers from both classes. A texton map and a 16-bin texton histogram were then computed for each of the 30 images. As in [3], we have found the final classification performance of K-means clustering to be relatively insensitive to the number of clusters selected. Using a similar scheme, we trained two forests, one for each texture (grade) class and each with T trees, consisting of depth-3 RPTs. The resulting 8 leaves from each tree yielded class-specific texton maps, whose histograms were then concatenated to yield a single histogram for each training image. For T=1, we would thus get a 16-bin histogram for each training image. Note that the tumor grade labels of the images themselves get used during this clustering phase since the textons corresponding to each texture class are separately identified. However, for our forest of ERTs, we train a single forest of depth-4 ERTs by using the filter responses from both classes. This step would have been detrimental to the classification performance of K-means or RPTs, but it is the key to the success of ERTs. We have used the values T=1 and 4 for RF-RPTs and RF-ERTs to check the benefit of using a forest with many trees. After the training phase for each clustering algorithm was over, texton maps (and histograms) were then obtained for the remaining 45 test images as well.

In the classification stage, we used the SVM with a standard radial basis function kernel (γ=1/histogram-bins, C=1) operating on the base-level histogram (after Z-score normalization) and with a 3-level SPM kernel operating on the texton map(s). To provide a comparison with the baseline approach in [3], we have also provided a comparison with the K-NN (K nearest neighbor) classifier (K=4) using the χ^2 distance. To train and validate our classifier, we used the same 30 training images used during clustering and then validated the resulting classifier on the remaining 45 test images. Thus, both the clustering and the classifier training were oblivious to the final test set used for validation.

Our classification results on the test dataset of 45 images are displayed in Table 1. For each combination of clustering and classification algorithm, we have tabulated the classifier accuracy in correctly identifying each tumor grade separately, the overall classifier accuracy as well as the area under the ROC curve (AUC). Note that the AUC $\in [0,1]$ is immune to the classifier trade-off between grade 3 and grade 4 accuracy. Since all our methods use some form of randomization, we have reported the average performance on 10 runs along with error bars. Among the three classification techniques, the RBF kernel and the K-NN classifier yield similar performance. The relatively poor performance of the SPM kernel could be because of over-fitting problems due to the small non-grade-4 regions present in our grade 4 images. Between the five clustering methods compared, highlighted AUC values for the RBF kernel show no significant difference in performance, barring the T=1 RF-ERT result. A forest of 4 ERTs (rows 13-15) has better mean performance with lower error bars than a single ERT (rows 7-9), but an RF of 8 RPTs (rows 10-12) does not outperform 2 class-specific RPTs (rows 4-6) as significantly.

We note that using our un-optimized implementations, for a test image, filtering requires about 3 sec., classification time is negligible and textonization using K-means, RF-RPT-T=4 and RF-ERT-T=4 requires 0.3 sec., 2 sec. and 3 sec., respectively. Training/testing times for the results in rows 1-3, rows 10-12 and rows 13-15 were about 200 min., 120 min. and 140 min., respectively.

Related Work: Our approach to histopathology texture analysis is most similar to the work in [8], although they have used the standard K-means algorithm for clustering and have used a boosting algorithm for classification. Moreover, this work is not concerned with Gleason grading. Automated Gleason grading was the focus in [9-12]. However, the work in [9] only used global texture features and the work in [10] used architectural features. Moreover,

[2] We thank Roberto Tran and Rene Vidal for providing an implementation of [3].

both these papers used fewer grade 3 and grade 4 samples for cross-validation. The work in [11] used global fractal dimensions for texture characterization and used a larger sample population of all Gleason grades, but they do not provide results on the individual accuracies for distinguishing between grades 3 and 4, whereas the work in [12] only attempted to discriminate between low-grade and high-grade tumors.

Table 1: Classification Results

	Gr. 3 %Acc.	Gr. 4 %Acc.	Net %Acc.	AUC
1. K-means, χ^2 K-NN	90.0 ± 0.00	92.9 ± 1.50	92.2 ± 1.17	NA
2. K-means, RBF-SVM	90.0 ± 0.00	95.1 ± 1.38	94.0 ± 1.07	**0.976**± 0.005
3. K-means, SPM-SVM	100 ± 0.00	85.1 ± 2.63	88.4 ± 2.04	0.974± 0.005
4. RF-RPT,T=1, χ^2 K-NN	96.0 ± 5.16	90.3 ± 2.76	91.6 ± 1.75	NA
5. RF-RPT,T=1, RBF-SVM	95.0 ± 5.27	92.3 ± 5.05	92.9 ± 3.60	**0.981**± 0.010
6. RF-RPT,T=1, SPM-SVM	94.0 ± 8.43	87.4 ± 7.99	88.8 ± 5.44	0.968± 0.016
7. RF-ERT,T=1, χ^2 K-NN	89.0 ± 7.38	83.7 ± 5.56	84.9 ± 3.11	NA
8. RF-ERT,T=1, RBF-SVM	86.0 ± 8.43	85.7 ± 5.55	85.8 ± 4.22	**0.956**± 0.033
9. RF-ERT,T=1, SPM-SVM	90.0 ± 4.71	64.6 ± 10.9	70.2 ± 7.71	0.895± 0.037
10. RF-RPT,T=4, χ^2 K-NN	93.0 ± 4.83	94.0 ± 0.90	93.8 ± 1.41	NA
11. RF-RPT,T=4, RBF-SVM	94.0 ± 5.16	93.4 ± 1.92	93.6 ± 1.26	**0.984**± 0.006
12. RF-RPT,T=4, SPM-SVM	99.0 ± 3.16	79.7 ± 8.24	84.0 ± 6.27	0.961± 0.020
13. RF-ERT,T=4, χ^2 K-NN	98.0 ± 4.21	89.1 ± 3.51	91.1 ± 2.34	NA
14. RF-ERT,T=4, RBF-SVM	96.0 ± 6.99	86.8 ± 4.08	88.8 ± 3.47	**0.978**± 0.011
15. RF-ERT,T=4, SPM-SVM	69.0 ± 1.97	92.3 ± 6.03	87.1 ± 3.44	0.950± 0.029

4. CONCLUSION

We have demonstrated the efficacy of our texture classification system in distinguishing between Gleason grades 3 and 4. In future work, we plan to train our system on additional Gleason grades, stroma and benign epithelium and to then use our automatic Gleason grading system on whole-mount histopathology slides. In addition, we also plan to use our texture classifiers for distinguishing between PIN (Prostatic Intraepithelial Neoplasia) and BPH (Benign Prostatic Hyperplasia). In order to increase the accuracy of our system, we plan to investigate the use of projection or distance splits within our random decision trees and the use of unified texton generation and classification [13].

5. REFERENCES

[1] T. Franiel, L. Ludemann, B. Rudolph, H. Rehbein, C. Stephan, M. Taupitz and D. Beyersdorff, "Prostate MR imaging: Tissue characterization with pharmacokinetic volume and blood flow parameters and correlation with histologic parameters," *Radiology*, vol. 252, no. 1, pp. 101–108, 2009.

[2] Y. Ou, D. Shen, J. Zeng, L. Sun, J. Moul and C. Davatzikos, "Sampling the spatial patterns of cancer: Optimized biopsy procedures for estimating prostate cancer volume and Gleason score," *Medical Image Analysis*, vol. 28, no. 7, pp. 1037–1050, 2009.

[3] M. Varma and A. Zisserman, "Classifying images of materials: Achieving viewpoint and illumination independence," in *ECCV (3)*, 2002, pp. 255–271.

[4] Y. Freund, S. Dasgupta, M. Kabra, and N. Verma, "Learning the structure of manifolds using random projections," in *NIPS*, pp. 255–271, 2007.

[5] F. Moosmann, B. Triggs, and F. Jurie, "Fast discriminative visual codebooks using randomized clustering forests," in *NIPS*, 2006, pp. 985–992.

[6] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *CVPR*, 2008.

[7] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR (2)*, 2006, pp. 2169–2178.

[8] L. Yang, W. Chen, P. Meer, G. Salaru, L.A. Goodell, V. Berstis, and D.J. Foran, "Virtual microscopy and grid-enabled decision support for large-scale analysis of imaged pathology specimens," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 13, no. 4, pp. 636–644, 2009.

[9] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, M. Feldman and J. Tomaszeweski, "Automated grading of prostate cancer using architectural and textural image features," in *ISBI*, 2007, pp. 1284–1287.

[10] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman and J. Tomaszewski, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology," in *ISBI*, 2008, pp. 284–287.

[11] P-W. Huang and C-H. Lee, "Automatic classification for pathological prostate images based on fractal analysis," *Medical Imaging, IEEE Transactions on*, vol. 28, no. 7, pp. 1037–1050, 2009.

[12] A. Tabesh, M. Teverovskiy, H-Y. Pang, V.P. Kumar, D. Verbel, A. Kotsianti, and O. Saidi, "Multifeature prostate cancer diagnosis and Gleason grading of histological images," *Medical Imaging, IEEE Transactions on*, vol. 26, no. 10, pp. 1366 – 1378, 2007.

[13] L. Yang, R. Jin and R. Sukthankar, "Unifying discriminative visual codebook generation with classifier training for object category recognition," in *CVPR,* 2008.

[14] C-C. Chang and C-J. Lin, *LIBSVM: a library for support vector machines*, 2001.