Prostate Histopathology: Learning Tissue Component Histograms for Cancer Detection and Classification

Lena Gorelick, Olga Veksler, Mena Gaed, José A. Gómez, Madeleine Moussa, Glenn Bauman, Aaron Fenster, and Aaron D. Ward*

***Abstract*—Radical prostatectomy is performed on approximately 40% of men with organ-confined prostate cancer. Pathologic information obtained from the prostatectomy specimen provides important prognostic information and guides recommendations for adjuvant treatment. The current pathology protocol in most centers involves primarily qualitative assessment. In this paper, we describe and evaluate our system for automatic prostate cancer detection and grading on hematoxylin & eosin-stained tissue images. Our approach is intended to address the dual challenges of large data size and the need for high-level tissue information about the locations and grades of tumors. Our system uses two stages of AdaBoost-based classification. The first provides high-level tissue component labeling of a superpixel image partitioning. The second uses the tissue component labeling to provide a classification of cancer versus noncancer, and low-grade versus high-grade cancer. We evaluated our system using 991 sub-images extracted from digital pathology images of 50 whole-mount tissue sections from 15 prostatectomy patients. We measured accuracies of 90% and 85% for the cancer versus noncancer and high-grade versus low-grade classification tasks, respectively. This system represents a first step toward automated cancer quantification on prostate digital histopathology imaging, which could pave the way for more accurately informed postprostatectomy patient care.**

***Index Terms*—Automated prostate cancer detection, cancer grading, digital pathology image analysis, machine learning, quantitative pathology, superpixels.**

I. Introduction

PROSTATE cancer (PCa) is the most common noncutaneous cancer among men and radical prostatectomy, in which the entire prostate is surgically removed, is performed on approximately 40% of men with organ-confined PCa [1].

The postprostatectomy pathologic assessment of the resected specimen by the pathologist yields crucial prognostic information that predicts surgical success and guides recommendations for adjuvant therapy [2]. Each tumor is assessed for its location within the prostate, volume, degree of differentiation (using the Gleason grading system [3]), extension into the seminal vesicles [termed as seminal vesicle invasion (SVI)] or beyond the prostate [termed as extra-prostatic extension (EPE)], and existence at the surgical resection margin [termed as a positive surgical margin (PSM)]. The prognosis of a patient is known to be related to the volumes and Gleason grades of the tumors observed in the resected specimen, as well as on SVI, EPE, and PSM status [2]. Adjuvant therapies such as radiation or hormone therapy may be considered for individuals with adverse pathology features.

Tumor volume and EPE status, in particular, are challenging to report quantitatively, and substantial inter-observer variability has been reported in the methods used clinically to make these assessments [4], [5]. The advent of high-resolution (e.g., 0.25 μm/pixel) whole-slide scanners is fostering a transition to a digital pathology workflow, similar to the transition from light-box viewing of film images to digital imaging in radiology. This transition opens the possibility for the integration of computational tools into the digital pathology workflow in order to enable quantitative assessments and reporting in a clinically feasible fashion. In the postprostatectomy prostate cancer setting, the quantitative reporting of tumor grade, location, volume, SVI, EPE, and PSM status depends on an accurate classification of each local tissue region as being cancerous or benign.

Prostate cancer detection on H&E-stained prostate tissue images and assessment of cancer grade are challenging due to the complexity of appearance of normal tissue and cancerous tissue of different grades, as illustrated in Fig. 1(a)–(c). To perform this task accurately, experts employ high-level knowledge regarding the expected morphologic, geometric, and color attributes of

Manuscript received March 06, 2013; accepted May 21, 2013. Date of publication May 31, 2013; date of current version October 02, 2013. This work was supported in part by the Canadian Institutes of Health Research (funding reference number CTP 87515) and in part by the Natural Sciences and Engineering Research Council of Canada. A. Fenster holds a Canada Research Chair in Biomedical Engineering, and acknowledges the support of the Canada Research Chairs Program. A. D. Ward holds a Cancer Care Ontario Research Chair in Cancer Imaging. *Asterisk indicates corresponding author.*

L. Gorelick is with the Departments of Computer Science and Medical Biophysics, and the Robarts Research Institute, The University of Western Ontario, London, ON, N6A 5K8 Canada.

O. Veksler is with the Department of Computer Science, The University of Western Ontario, London, ON N6A 5B7 Canada.

M. Gaed is with the Robarts Research Institute, The University of Western Ontario, London, ON, N6A 5K8 Canada, and also with the Lawson Health Research Institute, London, ON, N6C 2R5 Canada.

J. A. Gómez and M. Moussa are with the Department of Pathology, The University of Western Ontario, London, ON, N6A 3K7 Canada.

G. Bauman is with the Department of Oncology, The University of Western Ontario, London, ON, N6A 4L6 Canada, and also with the Lawson Health Research Institute, London, ON, N6C 2R5 Canada.

A. Fenster is with the Robarts Research Institute, The University of Western Ontario, London, ON, N6A 5K8 Canada, and with the Lawson Health Research Institute, London, ON, N6C 2R5 Canada, and also with the Department of Medical Imaging, The University of Western Ontario, London, ON, N6A 4L6 Canada.

*A. D. Ward is with the Lawson Health Research Institute, London, ON, N6C 2R5 Canada, and also with the Departments of Medical Biophysics, Biomedical Engineering, and Oncology, The University of Western Ontario, London, ON, N6A 4L6 Canada.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2013.2265334

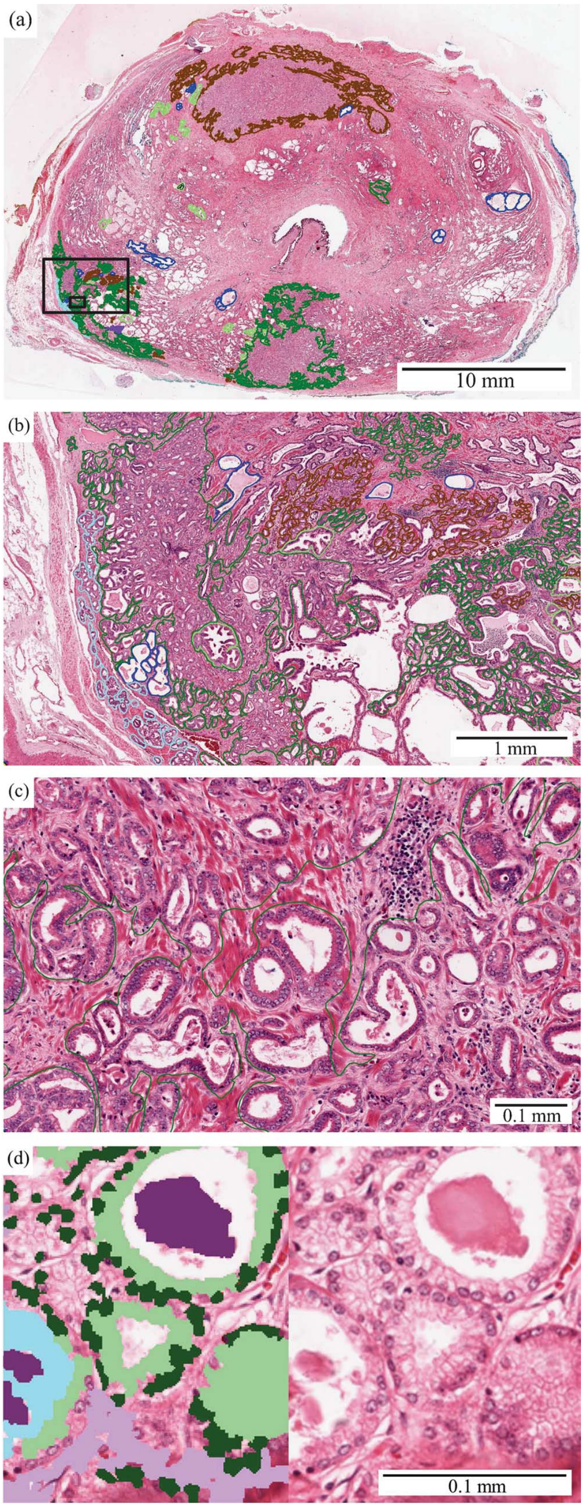

Fig. 1. (a) Sample contoured prostate histopathology image (brown: G3; gray: G4; dark green: G3+4; purple: G4+3; blue: Atrophy; cyan: EPE; light green: PIN). (b) Zoomed according to the large rectangle in (a). (c) Zoomed region according to the small rectangle in (a). (d) Sample of manually labeled superpixels (dark green: epithelial nuclei; light green: epithelial cytoplasm; light purple: stroma; dark purple: secretions; cyan: lumen).

normal and cancerous tissues. Another challenge to this assessment pertains to the sizes of the tissue structures that need to be examined, relative to the usual size of a human prostate. Pathologists frequently assess prostate tissue at a magnification corresponding to a pixel size of 0.5 μm/pixel in a scanned image. To illustrate a typical data size encountered, each prostate section can be approximately 4 cm × 3 cm in size, yielding an image of 80 000 × 60 000 pixels. With approximately 2–6 such images obtained from each prostate, tens of billions of RGB values are usually obtained in total. Thus, this classification problem poses the dual challenges of the need for high-level tissue information and the processing of very large data sets.

To address these challenges, in this paper, we describe and quantitatively evaluate a software system for automatic prostate cancer detection on H&E-stained prostate tissue images. We preprocess all H&E-stained prostate tissue images by automatically partitioning each image into a set of nonoverlapping superpixel regions using the method proposed in [6]. Pixels within each superpixel are similar in color and texture, and superpixel boundaries are aligned with intensity edges in the image. The superpixel partitioning provides a decrease in data size of several orders of magnitude while encoding potentially useful morphometric, geometric, and appearance information within the superpixels themselves. After superpixel partitioning, our system employs a two-stage classification. The first classification stage assigns to each superpixel a tissue component label (e.g., stroma, lumen, epithelial nuclei) having semantic meaning to an expert pathologist based on morphometric, geometric, and appearance information contained within and around each superpixel. Thus, we compute higher-level tissue information from the low-level image pixels. The second stage classifies image regions (each containing multiple labeled superpixels) as either low-grade cancer, high-grade cancer, or noncancer based on local histograms of tissue component labels within each sub-region. To the best of our knowledge, our manuscript describes the first application of a superpixel algorithm to the problem domain of prostate histopathology image analysis. Our results demonstrate that the superpixel image partitioning and superpixel features are inherently suited to the problem at hand.

Specifically, we hypothesize that 1) an AdaBoost classifier trained on morphometric, geometric, and appearance attributes of superpixels will classify superpixels into tissue components with an accuracy of $\geq$80%; 2) an AdaBoost classifier trained on tissue component histograms of superpixel labels will classify 0.3 mm × 0.3 mm image sub-regions as cancer or noncancer with an accuracy of $\geq$90% and a false negative rate of $\leq$15%; and 3) an AdaBoost classifier trained on tissue component histograms of superpixel labels will classify 0.3 mm × 0.3 mm cancerous image sub-regions as high-grade (Gleason 4) or low-grade (Gleason 3) cancer with an accuracy of $\geq$85% and a false negative rate of $\leq$10% for high-grade cancer detection.

II. Related Work

There has been a substantial research focus on the problem of automatically detecting and grading prostate cancer on digital histopathology imaging; these research efforts have yielded valuable insights into the nature of this problem. The common approach is to compute feature information from the images, and then train a classifier to distinguish cancer from noncancer, or to distinguish cancers of different grades, based on the computed features.

Within the context of the prior work, we distinguish between low-level and high-level features. Low-level features contain information about local distributions of pixel intensities, color and texture, intensity edges and their local orientations. Such features can be directly computed from images using standard image processing techniques, but have little or no semantic meaning to a pathologist. High-level features contain structural information about the content present in the image, such as tissue components (e.g., glands, lumina, stroma, nuclei), their shape, color, size, and geometric arrangement. These features have semantic meaning to a pathologist and are used in practice to guide the assessment [7]. In the following summary, we classify previously developed methods according to this distinction between low- and high-level features.

The majority of methods relying on low-level features utilize a multi-scale/multi-resolution approach and resort to machine learning techniques to train a classifier [8]–[12]. Doyle *et al.* [8] employed decision trees and trained a multi-scale Adaboost classifier to classify image pixels as either cancer or noncancer based on 600 texture features. When tested on 22 images from 22 subjects, their method obtained 88% overall accuracy. Similarly, Doyle *et al.* [9] trained a multi-resolution Bayesian classifier with AdaBoost and tested on 100 biopsy cores from 58 subjects, obtaining 74% pixel accuracy in the cancer versus noncancer classification task. Diamond *et al.* [10] calculated morphological and texture characteristics on samples from 12 subjects to distinguish cancer from noncancer images with 79% accuracy; they noted that long processing times were problematic. Huang *et al.* [11] found that multi-scale fractal dimension features were useful for prostate cancer grading, with 90%–94% classification accuracy; the number of subjects in their study, the origin of the tissue (biopsy or surgical specimen), and the false positive/negative rates were not specified. Khurd *et al.* [12] demonstrated that multi-scale texton characterization of the tissue images can be helpful to prostate cancer grading, obtaining 94% accuracy; the number of subjects in their study was not specified.

Several recent methods aimed to extract high-level structural information from the histopathology images [13]–[15]. Ideally, such high-level features should correspond to the information assessed by pathologists. In a paper focused on automated cancer grading, Doyle *et al.* [13] computed a set of graph-based features relying on manually labeled gland and nucleus centroid locations. Their automated grading experiment used 11 grade 3 and 7 grade 4 images and yielded an accuracy of 76.9% in distinguishing the two grades; the number of subjects in their study was not specified. This same research group reported in [14] improved results of 80% accuracy on a data set consisting of 16 grade 3 and 11 grade 4 images using a method intended to automate the selection of gland and nucleus centroids. The authors indicated that evaluation of their method on a larger data set is the subject of future work; the number of subjects in this study was not specified, and it is not clear if this was a superset of the data set used in [13]. Arif *et al.* [15] have also developed a method intended for automated nucleus extraction, with the ultimate aim of potentially supporting computer-aided diagnosis on prostate histopathology in the future. It is not clear how to generalize the methods in [14], [15] to other prostate tissue components (e.g., stroma, epithelium cytoplasm, intraluminal secretions, etc.), which are often used by pathologists for assessment.

Several other methods explored the utility of high-level features for automatic prostate histopathology assessment. Wittke *et al.* [16] manually segmented input images into epithelium, lumen, and stroma, and used derived features to distinguish low- from high-grade cancer; they reported 67% accuracy on images from 78 subjects. Xu *et al.* [17] used thresholding in the hue-saturation-intensity color space to segment the lumen and nuclei prior to support vector machine-based classification of high-grade versus low-grade cancer, and reported 75% accuracy. The number of subjects used was not specified. Nguyen *et al.* [18] perform K-means in the RGB space to distinguish between nuclei, stroma, lumen, and cytoplasm prior to segmentation and two-/three-way classification based on structural and contextual information. They experiment on 48 900×1500 pixel images from 20 patients with manual labeling of 525 artefacts, 931 normal glands and 1375 cancer glands, and report $93 \pm 0.04\%$ accuracy on gland versus artefact classification, $79 \pm 0.08\%$ accuracy on normal gland versus cancer gland classification, and $77 \pm 0.07\%$ accuracy on a three-way artefact versus normal gland versus cancer gland classification. Farjam *et al.* [19] demonstrated that explicit incorporation of knowledge specific to the domain of prostate histopathology assessment into a procedure for the segmentation of the images into stroma, lumen, and nuclei and subsequent computation of domain-specific features can yield more than 95% accurate classification of cancer grades. The number of subjects used was not specified.

Most relevant to our work is the method of Tabesh *et al.* [20] developed by Aureon Laboratories Inc. The authors took a hybrid low-level/high-level approach and investigated the utility of color, texture, and morphometric features for distinguishing cancer from noncancer and low-grade from high-grade cancer on tissue microarray cores. This method achieved 97% accuracy for the cancer versus noncancer classification task when tested on a set of 367 images and 81% accuracy for the high-grade versus low-grade classification when tested on a set 218 images, respectively. The method was applied to relatively large 1600×1200 images, while some of the images were reported to have as little as 5% of the area covered by cancer, making direct tumor volume quantification difficult. Furthermore, the approach required approximately 30 min of processing time for each 1600×1200 image; pathologists wishing to use the (now closed) company's analysis services were to physically ship biopsy samples for offline analysis.

Methods using low-level features along with multi-scale/multi-resolution approaches may result in the extraction of features which, although not directly meaningful to a pathologist, may correlate to high-level structures of pathological interest. The results of the described previous work, however, point to the idea that designing methods that directly extract high-level information from the image data may be beneficial to the task of automated prostate cancer detection and classification.

Although previous work has made important discoveries and strides toward a practically useful solution to this problem, a clinically-adopted solution remains elusive. Also, the lack of a standardized data set for this problem challenges the direct per-

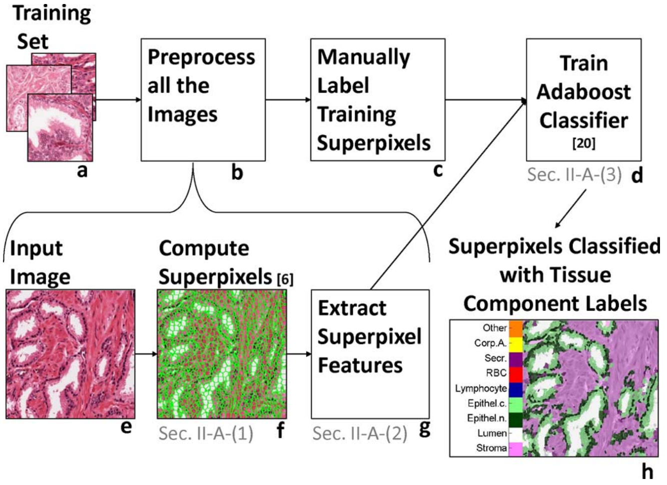
Training Set
a
Preprocess all the Images
b
Manually Label Training Superpixels
c
Train Adaboost Classifier [20]
Sec. II-A-(3)
d
Input Image
e
Compute Superpixels [6]
Sec. II-A-(1)
f
Extract Superpixel Features
Sec. II-A-(2)
g
Superpixels Classified with Tissue Component Labels
Other
Corp.A.
Secr.
RBC
Lymphocyte
Epithel.c.
Epithel.n.
Lumen
Stroma
h

Fig. 2. Overview of the tissue classification method. Given a set of training images (a), each image is preprocessed (b) by automatically computing a superpixel partition (f) and the superpixel features (g). (For a description of the features see Table I and Fig. 4.) Training superpixels are manually labeled with the tissue component labels (c) and used to train an Adaboost classifier (d). After training, given a new input image (e), the same preprocessing (f, g) is performed. The trained adaboost classifier (d) is then used to classify image superpixels and assign each superpixel a tissue component label (h).

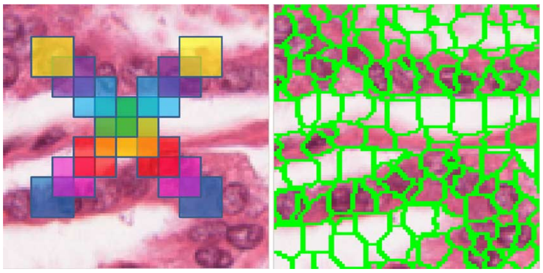

Fig. 3. Illustration of the superpixel algorithm [6]. Each pixel is assigned a label corresponding to one of the patches that overlap with it (left). The labels are stitched so that the boundaries between labels align with the intensity edges in the image and so that the intensities within each label are homogeneous (right). The optimal stitching is found automatically using an energy minimization framework.

formance comparison of methods. For a system to be clinically adopted, it may be beneficial: 1) to classify tissue based on attributes that are used by and therefore intuitive to pathologists, improving their confidence in the software system; 2) to be robust and general via the avoidance of direct incorporation of potentially brittle domain-specific knowledge into the design of the system (as opposed to using expert labeling for training); and 3) to be evaluated on a data set with well-known subject characteristics and a fully described reference standard. Our proposed approach meets all three criteria. Without incorporating any domain-specific knowledge, our approach provides a fully automatic partitioning of the image into intermediate-size superpixel regions and assigns to each superpixel one of nine high-level tissue component labels having semantic meaning to pathologists. Our system for prostate cancer detection and grading based on these higher-level labels was evaluated against a well-described reference standard data set, and with appropriate statistical inferences to account for the size of the data set and variability of performance of our system.

III. Methods

Our approach consists of two main components. The first assigns each image pixel a high-level tissue component label based on a superpixel partitioning and low-level superpixel features. The second classifies image sub-regions as cancer or noncancer, and as high-grade or low-grade cancer, based on high-level tissue component information. These two components are described in the following subsections.

A. *Tissue Component Classification*

The objective of tissue component classification is to assign to each image pixel a tissue component label having semantic meaning to an expert pathologist. We used nine tissue component labels: stroma, lumen, epithelial nucleus, epithelial cytoplasm, lymphocyte, red blood cell (RBC), intraluminal secretions, corpus amylaceum, and other. Fig. 2 shows a general overview of the tissue component classification method. It is based on three main steps: 1) automatic tessellation of the image into superpixels using a graph-cut based method [6], 2) extraction of superpixel appearance, morphometric and geometric features, and 3) classification of superpixels into nine tissue component classes based on the extracted features using Modest AdaBoost [21]. Each of these components is described in the following subsections.

1) Superpixels: There exist several methods to compute superpixels, such as mean shift [22], graph-based [23], and normalized cuts [24]. We chose the efficient graph-cut based method in [6] since it proposed a principled approach to compute superpixels with *regular shapes and sizes* in an energy minimization framework. Superpixels with regular shapes are less likely to straddle object (tissue component) boundaries, since such boundaries are mostly smooth. Moreover, by controlling the maximum size of each superpixel, we can influence the overall error in superpixel tessellation, where error is defined to be a superpixel split between two objects (tissue components).

The basic superpixel algorithm is illustrated in Fig. 3. An image is covered with overlapping square patches of fixed size (Fig. 3 left). Each pixel is contained in several patches, and we need to assign each pixel to one of the patches, thereby producing a superpixel tessellation. If two nearby pixels are assigned to the same patch, there is no penalty. If they belong to different patches, then there is a discontinuity penalty to pay. We set this penalty to be inversely proportional to the image gradient between these pixels. This encourages smoother, regularized boundaries that are well aligned with the intensity edges present in the image.

Let us state the superpixel tesselation problem formally as a labeling problem, which is solved with the graph-cut optimization approach [25]. Given a set of pixels $\mathcal{P}$ and a finite set of labels $\mathcal{L}$, the task in a labeling problem is to assign a label $l \in \mathcal{L}$ to each $p \in \mathcal{P}$. In our case, let the patches be numbered with consecutive integers $1, \ldots, k$, where k is the total number of patches. The ith patch is identified with an integer label i, and therefore the set of all possible labels is $\mathcal{L} = \{1, 2, \ldots, k\}$.

Let f_p denote the label assigned to pixel p, and let f be the collection of all label assignments. There are two types of constraints. Unary constraints $D_p(l)$ express how likely is a label l for pixel p. Let $S(l)$ be the set of the pixels in patch $l \in \mathcal{L}$. Label l can be assigned only to pixels in $S(l)$. Therefore, the data term is defined as

$$D_p(l) = \begin{cases} 1, & \text{if } p \in S(l) \\ \infty, & \text{otherwise} \end{cases}. \tag{1}$$

Binary constraints $V_{pq}(l_1, l_2)$ express how likely labels l_1 and l_2 are for neighboring pixels p and q. Based on [26], we use Potts model

$$V_{pq}(f_p, f_q) = w_{pq} \cdot \min(1, |f_p - f_q|) \tag{2}$$

with w_{pq} from [27]

$$w_{pq} = \exp\left(-\frac{\|I_p - I_q\|^2}{2\sigma^2}\right) \cdot \frac{1}{\operatorname{dist}(p, q)}.$$

Here, I_p is the color of pixel p, and $\operatorname{dist}(p, q)$ is the Euclidean distance between p and q. The coefficients w_{pq} are inversely proportional to the gradient magnitude between p and q, encouraging discontinuities to coincide with intensity edges.

An energy function that combines the unary and binary constraints is

$$E(f) = \sum_{p \in \mathcal{P}} D_p(f_p) + \lambda \sum_{\{p,q\} \in \mathcal{N}} w_{pq} \cdot V_{pq}(f_p, f_q). \tag{3}$$

In (3), the second sum is over eight-connected neighborhood system $\mathcal{N}$. This energy is NP-hard to optimize, and we use the expansion algorithm from [25], which has factor of 2 approximation. To compute max-flow/min-cut, we use [28].

The initial labeling is set to a random labeling with finite energy. Even though the number of labels is quite large, the expansion algorithm is very efficient because expansion on label l is performed only for pixels in $S(l)$.

With D_p as defined in (1), the data term in (3) is equal for all finite energy labelings, and we set $\lambda = 1$ in (3) since it has no effect on optimization.

The maximum superpixel size is equal to the patch size. Small superpixels are discouraged because they contribute a higher discontinuity penalty, since longer boundaries are required. Thus, the sizes of the superpixels are also regularized. A sample partitioning result is shown in Fig. 3, right. This basic algorithm can be extended to other formulations, which allow a trade-off between a less regular spatial tessellation but more accurate boundaries or better efficiency. A complete description of this superpixel algorithm can be found in [6].

2) Superpixel Features: Scale invariant feature transform (SIFT) is a commonly used image texture descriptor [29] that summarizes the distribution of local intensity gradients around each point of interest. Our use of SIFT features was based on the observation that different tissue components appear to have different spatial organizations of image gradients. For example, nuclei are consistently bright in the middle, with a circumferential gradient transitioning to a darker perimeter, whereas stroma has weaker gradients but more variable gradient orientation depending on the predominant orientation of the tissue in the local area.

We, therefore, computed a spatially dense set of SIFT descriptors for each of the images in our training set using the code provided by [30]. This resulted in a 128-feature vector $\mathbf{v}_p$ per image pixel $p \in \Omega$, where $\Omega \subset \mathbb{R}^2$ is the image domain. We reduce the space of all possible local textures by clustering all SIFT descriptors from all training images into a set of SIFT representatives $\{\mathbf{V}_k\}_1^K$ using K-means, similarly to [31]. Based on the results of our preliminary experiments we selected $\mathrm{K} = 100$ (varying K between 80 to 120 did not affect the results significantly). We then labeled each image pixel $p \in \Omega$ with the index i_p of the closest SIFT representative, namely

$$i_p = \operatorname{argmin}_{k \in \{1 \ldots K\}} \|\mathbf{v}_p - \mathbf{V}_k\|_2.$$

Next, we applied superpixel partitioning [6] to all the images in our training set. Each superpixel was then represented as a set of appearance, morphometry and geometry features. The appearance features describe the distribution of the color/intensity within each superpixel, including five-bin per-channel intensity histograms. The morphometry features describe the shape, size, and relative aspect ratio of each superpixel. Finally, the geometry features describe the distribution of the local gradients within superpixels and the immediate neighborhood. To that end, we used K-bin histograms over SIFT representative labels i_p within each superpixel. Using a histogram over SIFT labels further reduces the dimensionality (as opposed to using all SIFT descriptors within a superpixel), summarizing the distribution of local gradients over all points in a superpixel and its neighborhood. The number K determines the size of the histogram that is used as a part of the superpixel feature vector. See Table I for the full description of the features.

It is commonly assumed that spatial relationships between local appearances play an important role in classification of underlying structure in histopathology. Therefore, for each superpixel we also encoded its contextual spatial information. In addition to characterizing the appearance and shape of each superpixel, we also characterized the appearance of the neighborhood around each superpixel. Let $S \subset \Omega$ be a superpixel. Let p_S be the centroid of the superpixel and r be a radius in pixels, $r \in \{10, 20, 30\}$ corresponding to $\{10, 20, 30\}$ μm, respectively. We defined three rings of neighborhood

$$N_r(p_S) = \{p \in \Omega | r - 10 \leq \|p - p_S\| \leq r, p \notin S\}.$$

Because our superpixels are approximately 10 pixels in diameter, we selected the numbers 10, 20, and 30 to roughly correspond to one, two, and three layers of superpixel neighbors, respectively. We then computed color histograms (the last row in Table I) within these neighborhood rings and appended those histograms to the feature vector of each superpixel (adding total of $15 \times 3 = 45$ spatial appearance features for each superpixel). Fig. 4 shows an illustration of the neighborhoods N_r for a superpixel and the construction of its feature vector.

3) Using Adaboost for Classifying Superpixels: We used an Adaptive Boosting (AdaBoost) machine learning framework [32] to learn the superpixel descriptors characterizing

TABLE I
SUMMARY OF THE FEATURES USED TO DESCRIBE EACH SUPERPIXEL

Morphometry 7 features	Area	Number of pixels in the region
	Convex area	Number of pixels in the convex hull of the region
	Eccentricity	The eccentricity of the ellipse with the same second moments as the region: the ratio of the distance between the foci of the ellipse and its major axis length
	Equiv. diameter	The diameter of a circle with the same area as the region
	Extent	The ratio of pixels in the region to pixels in the bounding box
	Major/minor axis length	The length of the major/minor axis of the ellipse with the same second moments as the region
Geometry 100 features	SIFT histogram	K-bin histogram over SIFT representative labels i_p within each superpixel
Appearance Min/Max/Mean 3×4=12 features	Min/max/mean intensity	For the gray and RGB channels
Appearance Histogram 5×3=15 features	Intensity histogram	5-bin histogram of intensities for each of the RGB channels

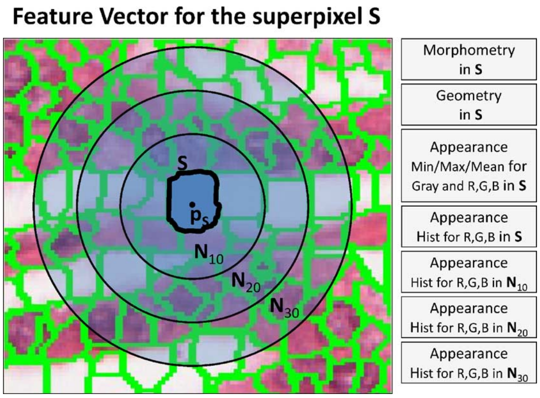
Feature Vector for the superpixel S
S
p_S
N_10
N_20
N_30
Morphometry in S
Geometry in S
Appearance Min/Max/Mean for Gray and R,G,B in S
Appearance Hist for R,G,B in S
Appearance Hist for R,G,B in N_10
Appearance Hist for R,G,B in N_20
Appearance Hist for R,G,B in N_30

Fig. 4. Superpixel features: illustration of the ring neighborhoods.

each tissue component type based on the features described above. AdaBoost is a meta-algorithm; it can be used in conjunction with many different weak classifiers to improve their performance. Specifically, we use Real AdaBoost [21] that utilizes simple decision trees as weak learners. We trained nine binary classifiers, one per tissue component type. Each component-specific binary classifier assigns each superpixel either the “class” or “nonclass” label for that component type (e.g., “stroma” or “not stroma”), along with a confidence value. The outcome of the nine binary classifiers was combined into one multi-label classifier by choosing the component label with the highest confidence.

B. Prostate Cancer Detection and Classification

Below, we describe our method for prostate tissue classification. This method relies on the tissue component classification described in Section III-A as a preprocessing step.

We assume that for each image, a superpixel partition is computed, features are extracted and each superpixel is classified as a specific tissue component type. We then counted the number of pixels in the image labeled with each tissue component and compute a nine-bin *tissue component histogram* per image. We used tissue component histograms as feature vectors for cancer detection. Namely, we considered the task of binary “cancer” versus “noncancer” classification using boosted decision trees (Modest AdaBoost) [21].

We further focused on cancer tissue classification problem and considered the task of binary “high-grade” versus “low-grade” classification, again using tissue component histograms as feature vectors and Modest AdaBoost [21].

IV. EXPERIMENTS

A. Materials

The image data for our experiments was acquired as part of a study that was approved by the research ethics board of our institution; written informed consent was obtained from all subjects. All subjects were suitable for and consented to radical prostatectomy, and had histologically confirmed prostate cancer (clinical stage T1 or T2). For each of 15 subjects, following radical prostatectomy, the resected prostate was fixed in 10% buffered formalin for 48 h. Each specimen was then transversely sliced into 4.4-mm-thick sections. The sections were processed using standard paraffin embedding, yielding whole-mount H&E-stained microscope slides, each containing a single 4-μm-thick section of tissue taken from each paraffin block face. The slides were digitized using a ScanScope GL (Aperio Technologies, Vista, CA, USA) bright field slide scanner. The acquired images were 24-bit color with isotropic 0.5 μm pixels. From each subject, between 2 and 5 (median 3) whole-mount sections were obtained; 50 such sections were obtained in total.

A physician (Gaed) trained by two genitourinary pathologists (Moussa and Gómez) assessed each image using the ScanScope ImageScope v11.0.2.725 software (Aperio Technologies, Vista, CA, USA) and contoured every tumor focus, as well as any identified areas of benign prostatic hyperplasia (BPH), atrophy (ATR), and prostatic intraepithelial neoplasia (PIN) using a Cintiq 12WX pen-enabled display (Wacom Co. Ltd., Saitama, Japan). Tumor areas were classified as either Gleason grade $3\ (\mathrm{G3}), 3+4\ (\mathrm{G3}+4), 4+3\ (\mathrm{G4}+3)$ or $4\ (\mathrm{G4})$. $\mathrm{G3}+4$ referred to a region that contained a mixture of Gleason grade 3 and Gleason grade 4 cancer, with more than 50% of the region comprising Gleason grade 3. $\mathrm{G4}+3$ referred to a region that contains a mixture of Gleason grade 3 and Gleason grade 4 cancer, with more than 50% of the region comprising Gleason grade 4. No Gleason grade 5 was observed in our data set. An illustrative example of this contouring is provided in Fig. 1. Contouring was performed at the highest image resolution, with the objective of generating contours enclosing regions consisting purely of the designated label (e.g., a “G4” contour is intended to contain only tissue that is cancerous with Gleason

TABLE II
NUMBER OF SUB-IMAGES PER PROSTATE TISSUE TYPE IN THE DATA SET USED FOR PROSTATE CANCER DETECTION AND CLASSIFICATION

Tissue type	Number of sub-images
Atrophy (ATR)	493
Benign prostatic hyperplasia (BPH)	104
Prostatic intraepithelial neoplasia (PIN)	216
Gleason 3 (G3)	99
Mixed Gleason 3 and 4, majority 3 (G3+4)	46
Mixed Gleason 3 and 4, majority 4 (G4+3)	12
Gleason 4 (G4)	21

TABLE III
NUMBER OF SUB-IMAGES PER PROSTATE TISSUE TYPE IN THE DATA SET USED FOR TISSUE COMPONENT CLASSIFICATION EXPERIMENT

Tissue type	Number of sub-images
Atrophy	21
Benign prostatic hyperplasia (BPH)	17
Prostatic intraepithelial neoplasia (PIN)	17
Gleason 3 (G3)	20
Mixed Gleason 3 and 4, majority 3 (G3+4)	12
Mixed Gleason 3 and 4, majority 4 (G4+3)	5
Gleason 4 (G4)	5

grade 4, devoid of normal tissue or cancer of other grades). Contouring and assessment in this fashion required approximately 70 h of operator time per subject; Fig. 1(b) and (c) provides an illustration of the attention to detail applied to this contouring task. All contours and assessments were performed by the trained physician (Gaed) and verified by a genitourinary pathologist (Moussa or Gómez).

From these images, we extracted 991 0.3 mm × 0.3 mm sub-images sampled onto a 301 pixel × 301 pixel grid (i.e., the pixel size was approximately 1 μm × 1 μm). Each sub-image resided entirely within one of the contoured regions of interest and thus inherited a corresponding label. Table II summarizes the distribution of sub-image labels in this set. This data set was used for the prostate cancer detection and classification experiment described in Section IV-C.

Our decision to classify small (0.3 mm × 0.3 mm) sub-images rather than large regions is motivated by the pathologist's objective of characterizing Gleason grade heterogeneity within larger regions such as tumors. By classifying each sub-image as cancer or noncancer, and each cancerous sub-image as low-grade or high-grade cancer, one can then examine the proportion of high-grade and low-grade sub-image classifications within tumors to characterize the overall aggressiveness of each tumor.

From the data set described in Table II, we extracted a subset described in Table III for use in the tissue component classification experiment described in Section IV-B. For each of these images, we performed a superpixel partitioning as described in Section III-A1. A custom user interface developed in-house was used by a physician (Gaed) trained by two genitourinary pathologists (Moussa and Gómez) to label a portion of superpixels in each image in the set. Each labeled superpixel was given one of the following nine tissue component labels: stroma, lumen, epithelial nucleus, epithelial cytoplasm, lymphocyte, red blood cell, secretion, corpus amylaceum, other. For superpixels containing tissue falling into more than one of the preceding categories, the physician applied the label representing the majority of the superpixel's contents. A total of 63 926 superpixels were manually labeled in this fashion. A sample of such a manual labeling is provided in Fig. 1(d).

B. Tissue Component Classification

Because classification of tissue into components (stroma, lumen, epithelial nucleus, epithelial cytoplasm, lymphocyte, red blood cell, secretion, corpus amylaceum, other) is an integral part of our overall method for prostate cancer detection and classification, our first experiment evaluates tissue component classification in isolation from the other components of our system. We performed 10 repeated subsampling cross-validation trials using the 63 926 manually-labeled superpixels from the data set described in Table III. In each trial, we randomly split the set of labeled superpixels into a training set comprising 80% of the labeled superpixels, and a test set comprising the remaining 20%. We performed a superpixel partitioning, as described in Section III-A1. For all superpixels, we calculated superpixel features, as described in Section III-A2. We then performed superpixel classification using AdaBoost, as described in Section III-A3. AdaBoost requires two parameters: the maximal depth d for tree learners and the maximal number I of iterations. In this experiment, we used the default $d = 3$ and $I = 179$ (the number of features per superpixel). For each of the 10 trials, nine binary AdaBoost classifiers (one per each tissue component type) were trained and tested, resulting in nine confidence values (one per label) for each superpixel in the test set. Each superpixel then was assigned the label having the largest confidence value.

We measured the mean ± standard deviation of the multiclass classification error (the number of incorrectly labeled superpixels in the test set, divided by the total number of superpixels in the test set) across the 10 trials as well as the confusion matrix for each trial. As a qualitative assessment of the tissue component labeling, we calculated a set of histograms showing the distribution of tissue components (stroma, lumen, epithelial nucleus, epithelial cytoplasm, lymphocyte, red blood cell, secretion, corpus amylaceum, other) within each prostate tissue type (Atrophy, PIN, BPH, G3, G3+4, G4+3, G4). This was done in order to evaluate the agreement of these observed distributions against pathologists' knowledge of the expected distributions. As a further qualitative assessment of the tissue component labeling, we also rendered a set of images with classifier outputs color-coded, to compare the resulting superpixel labeling to the visible tissue components in the H&E-stained test images. Finally, *to test hypothesis (1)* (Section I), that our method will classify superpixels into tissue components with accuracy of $\geq$80%, we performed a one-tailed t-test of the null hypothesis that the mean multiclass classification error ≥ 0.2.

C. Prostate Cancer Detection and Classification

This section describes two experiments. The first is on prostate cancer detection; the classification of sub-images as containing cancerous tissue, or noncancerous tissue (Section IV-C1). The second is on prostate cancer classification; the classification of cancer-containing sub-images as containing high-grade cancer, or low-grade cancer (Section IV-C2). Both experiments were conducted using the 991-sub-image data set described in Table II. A superpixel partitioning was computed for each sub-image, as described in Section III-A1. The number of superpixels per sub-image was recorded in order to measure the reduction in data size achieved by this partitioning. Each superpixel was then represented by a set of features, as described in Section III-A2, and one of the nine tissue component labels was applied to each superpixel, as described in Section III-A3. A tissue component histogram was computed for each sub-image and used for training and testing of a Modest AdaBoost classifier, as described in Section III-B. To evaluate classifier performance, we computed the false positive (FP), false negative (FN), true positive (TP), true negative (TN), and accuracy (TP + TN) rates.

1) Prostate Cancer Detection: This experiment evaluated the ability of a Modest AdaBoost classifier to classify sub-images as containing cancer or noncancer, based on tissue component histograms. We performed 10 repeated subsampling cross-validation trials using the 991 sub-images in the data set described in Table II. In each trial, we randomly split the set of labeled superpixels into a training set comprising 80% (792) of the sub-images, and a test set comprising the remaining 20% (199). For this experiment, each sub-image in the data set was labeled as cancer ("positive") or noncancer ("negative"). Any sub-image having a label of Atrophy, BPH, or PIN was designated as noncancer. Any sub-image having a label of G3, $\mathrm{G3} + 4$, $\mathrm{G4} + 3$, or G4 was designated as cancer.

Since, in the clinical workflow, a failure to detect cancer may result in the denial of necessary postprostatectomy treatment, it may be beneficial to tune the classifier to achieve a decreased FN rate, even if compromising in terms of an increased FP rate. Depending on the classifier at hand, this can be done by either skewing the error cost matrix, or by using unbalanced class probability priors, or both. In these experiments, due to the relatively lower number of positive data points, we implicitly skewed the error cost matrix by artificially augmenting the training set with an additional number s of duplicates for each positive sub-image in the set. This way, during classifier training, each FN error carried heavier weight and was counted s times compared to FP errors. Therefore we explicitly biased the classifier toward obtaining a lower FN rate by preferring FP errors over FN errors. We varied the number of duplicates used in training in order to observe how the classification results changed as a function of number of duplicates. To this end, each of the cross-validation experiments was performed five times, one for each $s \in \{0, 5, 10, 15, 20\}$.

We performed the following measurements for all values of s. We measured the mean ± standard deviation (where applicable) of the classification error across all cross-validation trials. Classification error was calculated as the number of incorrectly labeled sub-images in the test set, divided by the total number of sub-images in the test set. In addition to the FN and FP rates, we measured the TP and TN rates. We also plotted a precision-recall curve showing the mean TN versus mean TP (averaged across all trials) as a function of s in order to visualize the trade-off resulting from increasing the number of positive duplicates in the data set. *To test hypothesis (2)* (Section I), that our method will classify sub-images as cancer or noncancer with accuracy of $\geq$90% and a false negative rate of $\leq$15%, we performed one-tailed t-tests of the null hypotheses that the mean classification error ≥ 0.1 and that the FN rate ≥ 0.15.

2) Prostate Cancer Classification: This experiment evaluated the ability of a Modest AdaBoost classifier to classify cancer-containing sub-images as containing high-grade cancer or low-grade cancer, based on tissue component histograms. It was conducted identically to the prostate cancer detection experiment described in Section IV-C1, with the exception of the following details. We performed a leave-one-out cross-validation experiment using the 120 G3 and G4 sub-images in the data set described in Table II. Leave-one-out cross-validation was performed due to the relatively small size of the data set for this experiment. For this experiment, each sub-image in the data set was labeled as low-grade cancer or high-grade cancer. Sub-images having a label of G3 were designated as low-grade cancer. Sub-images having a label of G4 were designated as high-grade cancer. In this experiment, for purposes of computing TP, TN, FP, and FN, high-grade cancer was considered as "positive" and low-grade cancer as "negative." *To test hypothesis (3)* (Section I), that our method will classify sub-images as high-grade or low-grade cancer with accuracy of $\geq$85% and a false negative rate of $\leq$10% for high-grade cancer detection, we performed one-tailed t-tests of the null hypotheses that the mean classification error ≥ 0.15 and that the FN rate ≥ 0.1.

V. Results

A. Tissue Component Classification

This section reports the results of the tissue component classification experiment described in Section IV-B. The mean ± standard deviation of the multiclass error rate over 10 trials was 0.16 ± 0.009. The data passed a one-sample Kolmogorov–Smirnoff test of normality ($p = 0.89, \alpha = 0.05$). A one-tailed t-test of the null hypothesis that the mean multiclass classification error ≥ 0.2 yielded a p-value of 1.4×10^{-10}. As this test involved 10 samples generated via cross-validation, a Bonferroni-corrected α of $0.05/10 = 0.005$ was used. This null hypothesis was therefore rejected and thus we assert that *hypothesis (1)* (Section I) is true. The Bonferroni-adjusted 95% confidence interval on the mean classification error was found to be (0.1598, 0.1687). This confidence interval suggests that a mean classification error of $\sim$0.16, corresponding to an accuracy of $\sim$84%, can be expected from the tested method.

Fig. 5 shows a classification confusion matrix for one of the cross-validation trials. The rows represent the actual tissue component labels and the columns represent the labels predicted by the Real AdaBoost classifier (Section IV-B). A cell (i, j) represents the frequency of an event whereby a superpixel with the actual label i was classified as label j. The diagonal elements of the confusion matrix depict the classification accuracy for each label. Each off-diagonal matrix element (i, j) shows the rate of error whereby the label i was misclassified and assigned the label j. The rows of the matrix are normalized to sum to one. The numbers on the left of the figure show frequencies of each label in the test set. As can be seen in the matrix, the proportions of correct classifications are above 0.8 for most of the

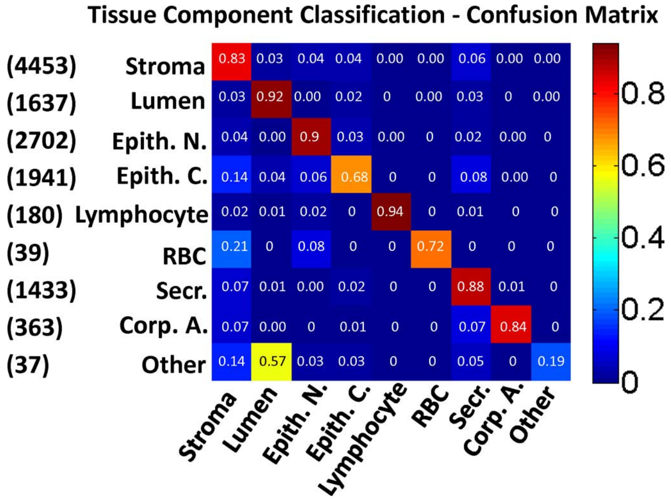

Fig. 5. Confusion matrix for the tissue component classification: Actual labels are the rows, predicted labels are the columns. A cell (i, j) represents the frequency of an event whereby a superpixel with the actual label i was classified as label j. The rows of the matrix are normalized separately for each label (row) to sum to one. The numbers on the left of the figure represent the frequency of each label in the test set.

tissue component labels, with the "other" label having a confounding effect.

Fig. 6 shows the frequency of occurrence of each tissue component in each prostate tissue type, obtained by running our classifier on the 991 subimages in the data set described in Table II. For example, the amount of lumen is lower in cancer, compared to noncancer. The amount of lumen decreases and the amount of epithelial nuclei increases as a function of Gleason grade. Corpora amylacea are mostly present in Atrophy, BPH, and PIN, but not as much in the cancerous tissue.

Fig. 7 show examples of test images with color-coded tissue component labels superimposed on each original image. There are seven rows, one for each prostate tissue type. We show examples of four test images for each tissue type.

To mitigate the potential bias involved in training and testing on neighbouring superpixels in our cross validation experiment, we repeated the experiment with leave-one-patient-out cross validation. Due to the uneven distribution of tissue component labels among patients in the training set, this experiment was especially challenging to the classifier as it was trained on an incomplete set of labels in some folds, artificially degrading its performance. As this is a nine-way classification, the expected error rate arising from chance performance is not 0.5 but rather is 0.89 (i.e., 8/9). We recorded a mean ± std error rate of 0.28 ± 0.1, illustrating that the performance of the classifier is substantially better than chance even with incomplete training data. Note that for the subsequent cancer detection and classification experiments (Section V-B) we use tissue component classifier that was trained on all available labeled images described in Table III. Therefore, the kind of cross validation performed for the evaluation of the tissue component classification experiment has no effect on the performance of the cancer detection and classification experiments described next (Section V-B).

B. Prostate Cancer Detection and Classification

The experiments described in Section IV-C were conducted on a Microsoft Windows (Redmond, WA, USA) platform based on the Intel Xeon E5645 processor (Intel Corporation, Inc., Santa Clara, CA, USA). The software implementation comprises a combination of C++ and Matlab 7.12.0 (The Mathworks, Inc., Natick, MA, USA) code. The implementation was developed as a platform to support research and is not optimized for speed or parallel processing; we observe substantial room for improvement of our implementation in these areas. Nevertheless, for the experiments described in Section IV-C, each test sub-image was classified by this platform in under 2 min, including all necessary steps (superpixel partitioning, calculation of features, and AdaBoost classification). Since each sub-image is independently classified, this system is inherently highly parallelizable. The mean±std number of superpixels per sub-image for the data set described in Table II was 1396 ± 221.

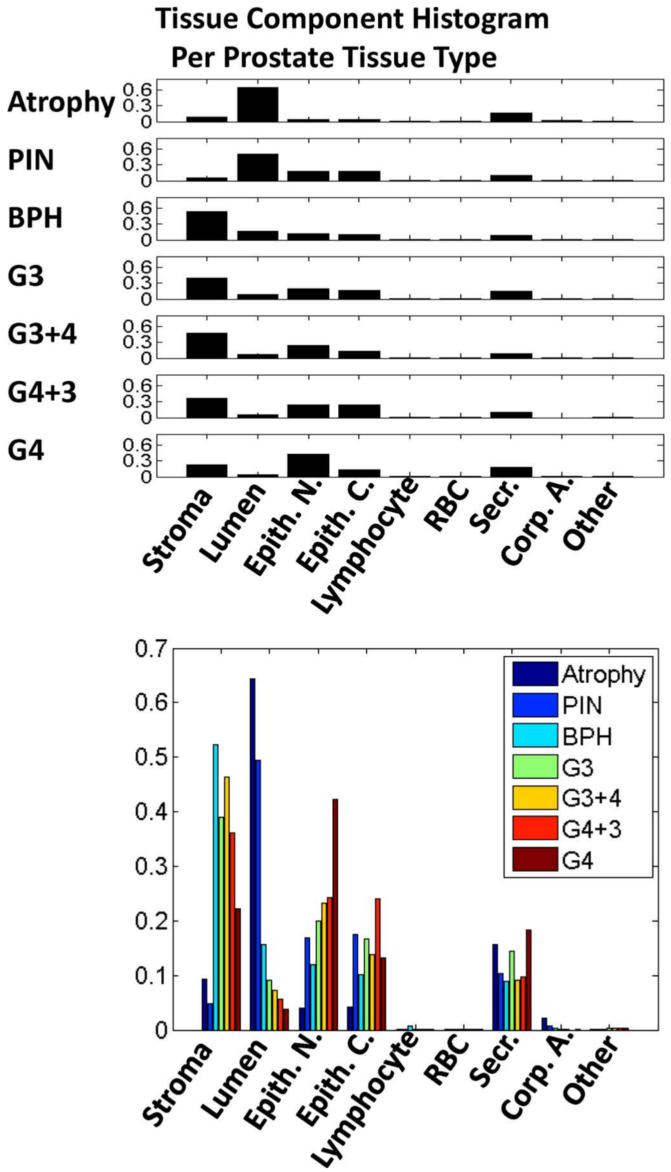

Fig. 6. Top: Distribution of tissue component labels per prostate tissue type. Bottom: The histograms are reorganized and for each tissue component we show how its relative proportion varies across different prostate tissue types.

1) Prostate Cancer Detection: This section reports the results of the prostate cancer detection experiment described in Section IV-C1. See Table IV, first row for the mean ± standard

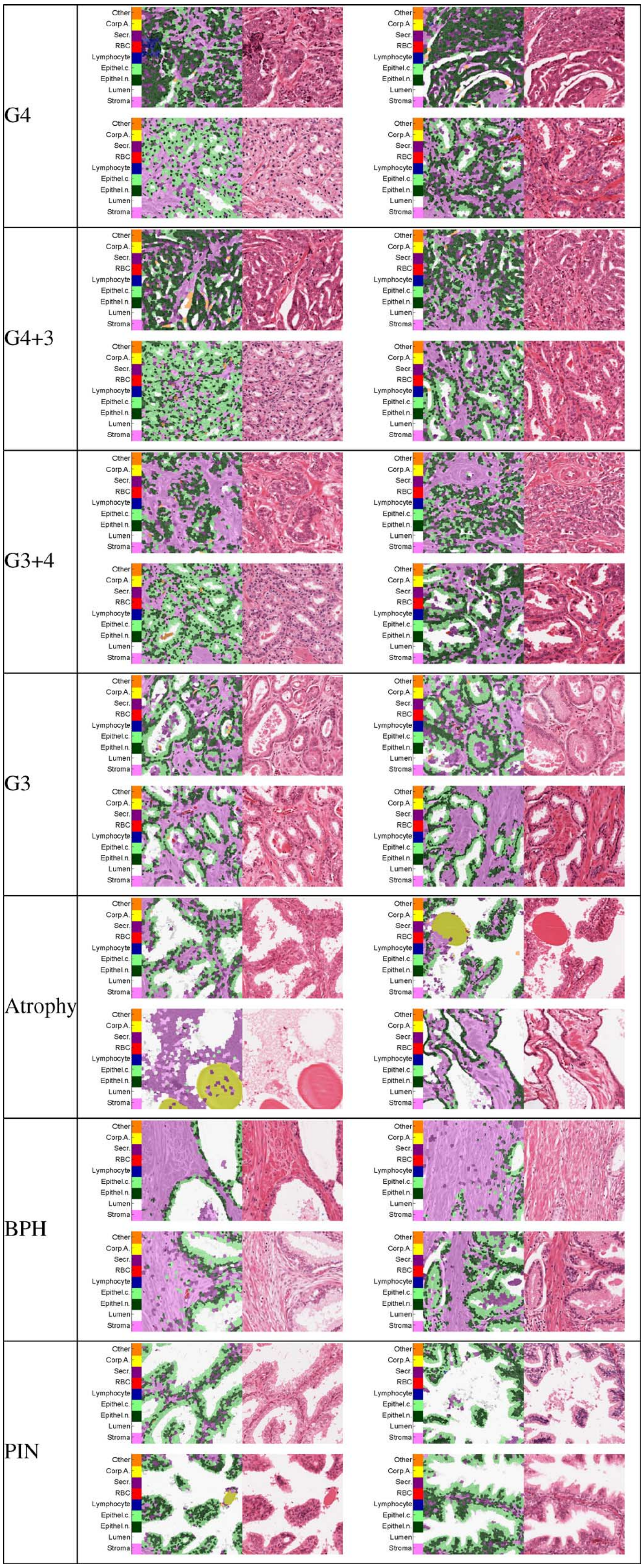

Fig. 7. Tissue component classification superimposed over several example test images.

TABLE IV
CANCER CLASSIFICATION AND DETECTION RESULTS: MEAN FN+FP ± STD AND MEAN FP± STD AS A FUNCTION OF NUMBER s OF DUPLICATES USED IN TRAINING. NOTE THAT CANCER CLASSIFICATION EXPERIMENT IS LEAVE-ONE-OUT CROSS-VALIDATION AND THEREFORE DOES NOT HAVE MEAN AND STD

		s=0	s=5	s=10	s=15	s=20
Cancer Detection	FP+FN	0.07±0.01	0.08±0.02	0.09±0.02	0.10±0.02	0.11±0.02
	FN	0.22±0.08	0.06±0.03	0.05±0.04	0.04±0.03	0.03±0.03
Cancer Classif.	FP+FN	0.08	0.08	0.09	0.13	0.13
	FN	0.05	0.04	0.04	0.05	0.04

deviation of the error rate (FP+FN)/(FP+FP+TN+TP) over 10 trials as a function of number of duplicates s used in training. The data (aggregated over all $s \in \{0, 5, 10, 15, 20\}$) passed a one-sample Kolmogorov–Smirnoff test of normality ($p = 0.71$, $\alpha = 0.05$). A one-tailed t-test of the null hypothesis that the mean classification error ≥ 0.1 yielded a p-value of 0.0003. As this test involved 50 samples generated via 10 cross-validation trials for each value of s, a Bonferroni-corrected α of $0.05/50 = 0.001$ was used. The null hypothesis was therefore rejected. The Bonferroni-adjusted 95% confidence interval on the mean classification error was found to be (0.0780, 0.0994). This confidence interval suggests that a mean classification error of not more than ~ 0.1, corresponding to an accuracy of $\sim$90%, can be expected from the tested method.

Table IV, also reports the mean ± standard deviation of the FN/(FN+TP) rate over 10 trials. The data (aggregated over all s) passed a one-sample Kolmogorov–Smirnoff test of normality ($p = 0.07, \alpha = 0.05$). A one-tailed t-test of the null hypothesis that the mean false negative rate ≥ 0.15 yielded a p-value of <0.0001. As this test involved 50 samples generated via 10 cross-validation trials for each value of s, a Bonferroni-corrected α of $0.05/50 = 0.001$ was used. The null hypothesis was therefore rejected. The Bonferroni-adjusted 95% confidence interval on the mean FN rate was found to be (0.04, 0.12). This confidence interval suggests that a mean FN rate of not more than $\sim$12% can be expected from the tested method. Based on the rejection of both of the above null hypotheses, we assert that *hypothesis (2)* (Section I) is true.

Fig. 8–10 show the mean ± standard deviation over 10 trials of FN, FP, and accuracy (TP+TN) rates, respectively, as a function of the number s of duplicates used for each positive (cancer) example in the training. Fig. 11 shows the recall-precision curve.

Table V shows the distribution of classification errors across prostate tissue types as a function of the number s of duplicates used in the training data set. For each experiment corresponding to $s \in \{0, 5, 10, 15, 20\}$ and prostate tissue type, we show the proportion of the images with this type classified as cancer (C) or noncancer (NC), averaged over 10 cross-validation trials. As the number s of positive (cancer) duplicates used in training increases, the number of cancer images misclassified as noncancer decreases. The higher the grade of cancer, the faster the FN rate decreases. (Standard deviations can be seen in Fig. 8–10 and are omitted from Table V for clarity.)

2) Prostate Cancer Classification: This section reports the results of the prostate cancer classification experiment described in Section IV-C2. See Table IV, second row for the error rate (FP+FN)/(FN+FP+TP+TN) of the leave-one-out cross validation classification as a function of number of duplicates s used in training. The data (aggregated over all s) passed a one-sample Kolmogorov–Smirnoff test of normality ($p = 0.83, \alpha = 0.05$).

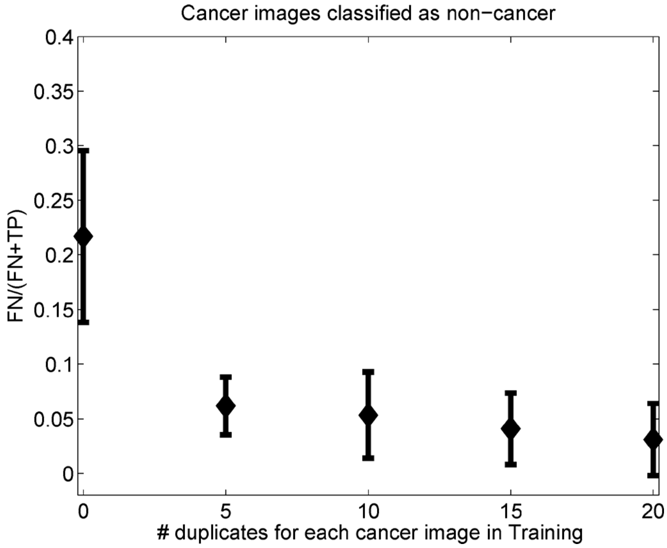

Fig. 8. Cancer detection average false negative rate: FN/(FN+TP) as a function of the number s of duplicates used for each cancer image in training.

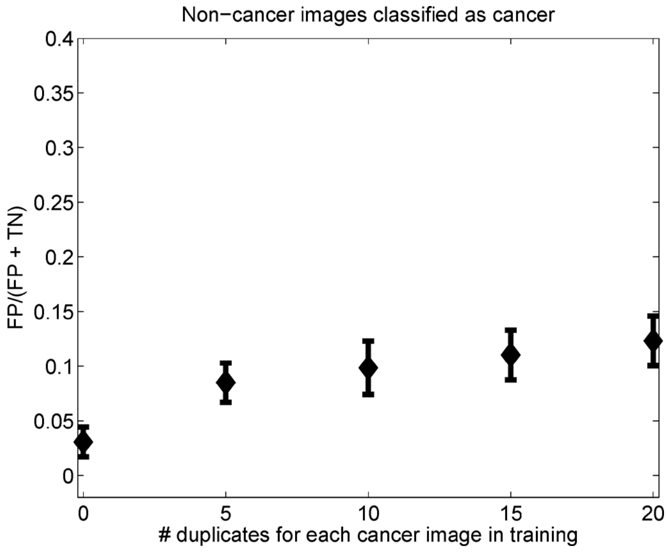

Fig. 9. Cancer detection average false positive rate: FP/(FP+TN) as a function of the number s of duplicates used for each cancer image in training.

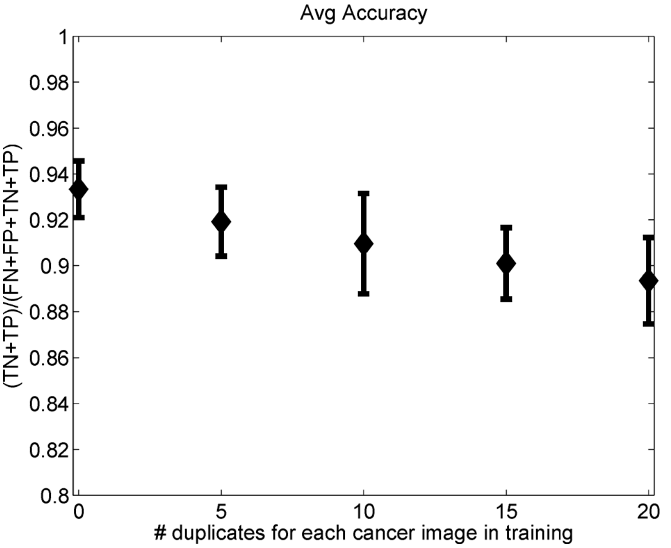

Fig. 10. Cancer detection average accuracy rate: (TP+TN)/(FN+TP+FP+TN) as a function of the number s of duplicates used for each cancer image in training.

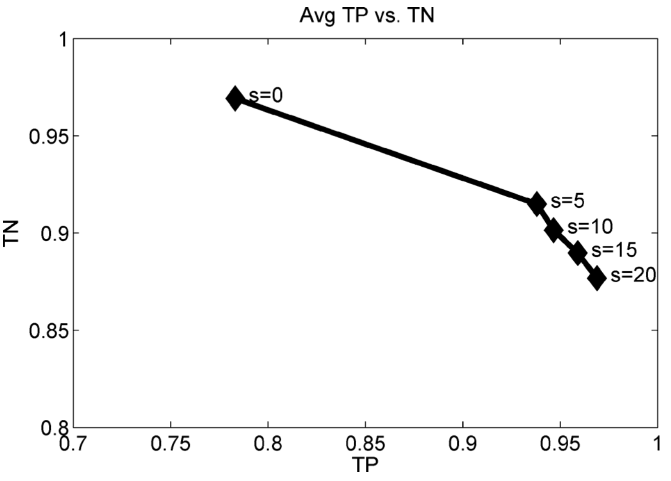

Fig. 11. Cancer detection recall-precision rate curve: average TN versus average TP as a function of the number s of duplicates used for each cancer image in training.

TABLE V

Average Confusion Matrix as a Function of Number of Duplicates Used in Training. For Each Experiment and Prostate Tissue Type, We Show the Proportion of the Images With This Type Classified as Cancer (C) or Noncancer (NC), Averaged Over 10 Trials. Numbers in Bold Show the Proportion of Images Classified Correctly

	Number of duplicates s									
	$s=0$		$s=5$		$s=10$		$s=15$		$s=20$	
	C	**NC**	**C**	**NC**	**C**	**NC**	**C**	**NC**	**C**	**NC**
ATR	0.01	**0.99**	0.04	**0.96**	0.04	**0.96**	0.05	**0.95**	0.06	**0.94**
BPH	0.00	**1.00**	0.02	**0.98**	0.03	**0.97**	0.04	**0.96**	0.05	**0.95**
PIN	0.20	**0.80**	0.42	**0.58**	0.48	**0.52**	0.54	**0.46**	0.55	**0.45**
G3	**0.75**	0.25	**0.95**	0.05	**0.96**	0.04	**0.96**	0.04	**0.97**	0.03
G3+4	**0.71**	0.29	**0.87**	0.13	**0.87**	0.13	**0.90**	0.10	**0.95**	0.05
G4+3	**0.93**	0.07	**1.00**	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**	0.00
G4	**1.00**	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**	0.00

A one-tailed t-test of the null hypothesis that the mean classification error ≥ 0.15 yielded a p-value of 0.005. As this test involved five samples generated via one leave-one-out cross-validation for each values of s, a Bonferroni-corrected α of $0.05/5 = 0.01$ was used. The null hypothesis was therefore rejected. The Bonferroni-adjusted 95% confidence interval on the mean classification error was found to be (0.0514, 0.1486). This confidence interval suggests that a mean classification error of not more than ~ 0.15, corresponding to an accuracy of $\sim 85\%$, can be expected from the tested method.

Table IV also reports the FN/(FN+TP) error rate in the leave-one-out cross validation classification. The data (aggregated over all $s \in \{0, 5, 10, 15, 20\}$) passed a one-sample Kolmogorov–Smirnoff test of normality ($p = 0.41, \alpha = 0.05$). A one-tailed t-test of the null hypothesis that the mean FN rate ≥ 0.1 yielded a p-value of <0.0001. As this test involved five samples generated via one leave-one-out cross-validation for each value of s, a Bonferroni-corrected α of $0.05/5 = 0.01$ was used. The null hypothesis was therefore rejected. The Bonferroni-adjusted 95% confidence interval on the mean FN

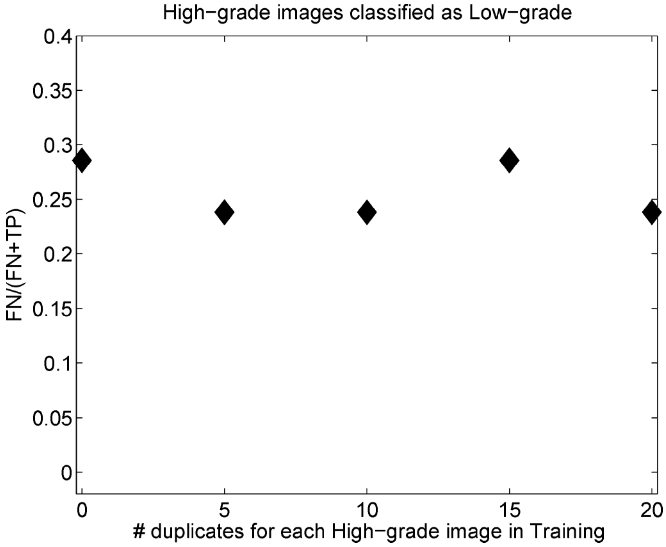

Fig. 12. Cancer classification: The rate of high grade cancer images classified as low grade FN/(FN+TP), as a function of the number s of duplicates used for each high-grade cancer image in training.

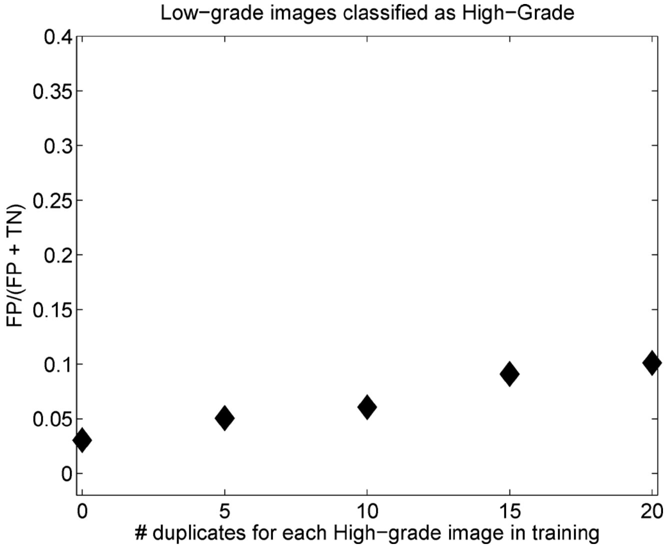

Fig. 13. Cancer classification: The rate of low grade cancer images classified as high grade (FP/(FP+TN), as a function of the number s of duplicates used for each high-grade cancer image in training.

rate was found to be (0.04, 0.05). This confidence interval suggests that a mean FN rate of not more than ~5% can be expected from the tested method. Based on the rejection of both of the above null hypotheses, we assert that *hypothesis (3)* (Section I) is true.

Fig. 12–14 show the FN, FP, and accuracy (TP+TN) rates, respectively, as a function of the number of duplicates used for each positive (high grade) image in the training computed with leave-one-out cross-validation. Fig. 15 shows the recall-precision curve.

VI. Discussion

The focus of this work was to evaluate the accuracy of the proposed system for cancer detection and classification.

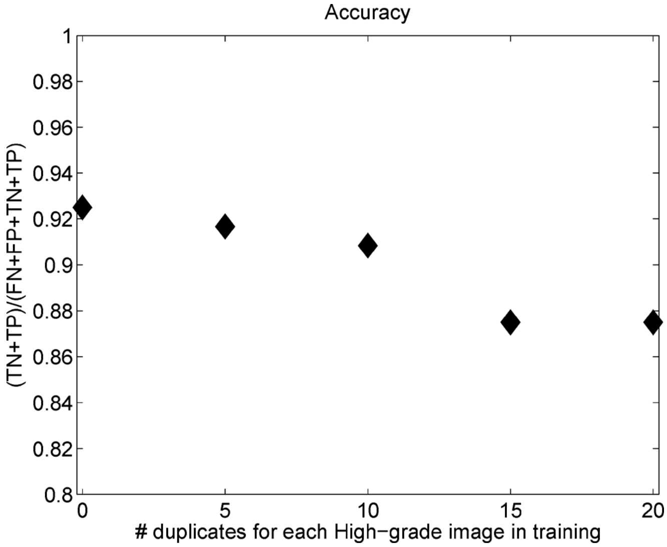

Fig. 14. Cancer classification accuracy rate: (TP+TN)/(FN+TP+FP+TN) as a function of the number s of duplicates used for each high-grade cancer image in training.

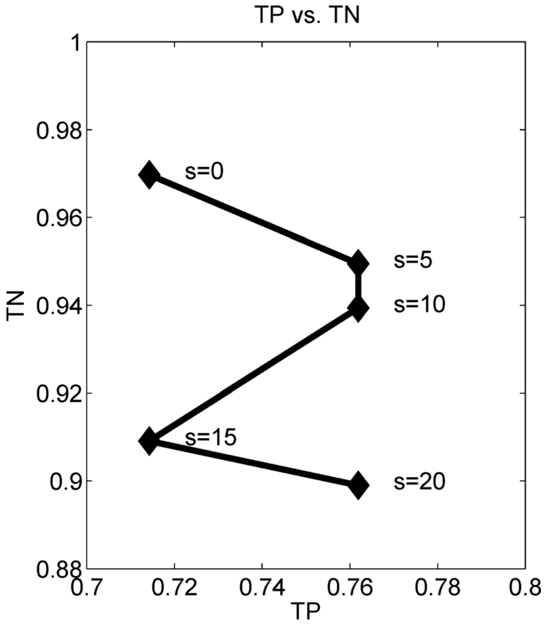

Fig. 15. Cancer classification recall precision rate curve: TN versus TP as a function of the number s of duplicates used for each high grade cancer image in training.

Although speed was measured, a high-speed algorithm/implementation was out of the scope of this work and there exists substantial room for speed improvement in both the tuning of the algorithm and the optimization of the implementation. A high-speed implementation will leverage a key strength of our approach, which is the independent classification of each sub-image; in principle, with sufficient parallel processors, an entire slide could be processed in the same amount of time as a single sub-image. With the advent processing-dedicated GPUs with ever-growing numbers of processors (currently >2500 on a single adapter), a parallel implementation of this algorithm at the pathologist's desk becomes ever more feasible.

A. Tissue Component Classification

Many computer vision applications (e.g., [33]–[40]) have benefited from representing an image as a collection of superpixels. A superpixel is considered to be a perceptually meaningful, atomic, arbitrarily-shaped image subregion whose borders are intended to be better aligned with intensity edges than those of a rectangular region. Usually, a superpixel contains pixels that are similar in color and texture. Such pixels are likely to belong to the same physical world object or, in our case, the same tissue component. Superpixel partitioning of an image reduces data dimensionality and can naturally be used for computing features that need spatial support [6]. To the best of the authors' knowledge, this work represents the first application of a superpixel partitioning to the problem of prostate pathology image classification. In this work, the superpixel partitioning provided more than an order of magnitude reduction in the number of elements to be processed within each sub-image, from $301 \times 301 = 90,601$ pixels to a mean of 1396 superpixels per sub-image (>60-fold reduction in data size). Morphometric, geometric, and appearance features computed based on the superpixel partitioning provided our tested classifier with the necessary data to accurately assign to each superpixel a higher-level tissue component label having semantic meaning to an expert pathologist. This result points to the suitability of the superpixel partitioning to this problem domain.

For classification, we chose to use AdaBoost as it is commonly used in computer vision, data mining, pattern recognition, and medical imaging applications (e.g., [41]–[46]) for its relative simplicity and speed. We observed that this classifier provided useful accuracy in tissue component labeling of superpixels, with the exception of an observed confounding effect of the "other" label (Fig. 5). Further inspection revealed that the "other" label accounted for several additional, infrequently occurring, tissue subtypes of heterogeneous appearance. For example, it includes superpixels lying in the empty space surrounding glands, covering cristalloids or blood vessel walls. The main reason for the relatively high error rate for the "other" label is the low frequency of these sub-types in the training and test data, high variability of appearance and the lack of higher level information that allows for distinguishing e.g., the space surrounding gland from lumen (both uniformly white), or vessel walls from stroma (both red). We observed (Fig. 6) that the proportions of tissue component labels for each prostate tissue type correspond to expected distributions based on expert knowledge. For example, it is known that the ratio of stroma to epithelial nuclei decreases with increasing Gleason grade, and this is reflected in Fig. 6. The decrease in lumen with increasing Gleason grade shown in Fig. 6 is also anticipated as a consequence of more advanced cancer.

B. Prostate Cancer Detection and Classification

The results of our prostate cancer detection and classification experiments can be interpreted in the context of our reference standard data set, which contained meticulously drawn manual contours [e.g., Fig. 1(a)–(c)] verified by a minimum of two experts. The quality of this data set contributed to the effectiveness of classifier training and validity of testing.

The results of our work demonstrate the ability of the trained classifiers to imitate the judgements of the specific observer regarding cancer versus noncancer and low-grade versus high-grade classification, based on the annotations provided by the observer for the training data. From the standpoint of creating a tool that may ultimately be useful to a pathologist in clinical practice, it remains to be seen whether an ideal system should be separately trained to mimic the idiosyncrasies of each individual pathologist, or whether the performance of an ideal system should somehow reflect the performance of pathologists in general. A separate study involving multiple observers in the training and test sets could elucidate the performance of our features and classifiers and compare the results to measured intra- and inter-observer variability in human experts.

1) Prostate Cancer Detection: In our cancer detection experiment, we noted a general improvement in FN rate with increasing s, without a deleterious concomitant increase in FP rate (Table V). However, this is not true for PIN; with increasing s, the accuracy for PIN images decreases rapidly with a high FP rate for PIN images. In fact, at $s = 20$, PIN images are classified as cancer 55% of the time and as noncancer 45% of the time. Upon reflection, this result is unsurprising on account of the fact that PIN is widely considered to be tissue in a *precancerous* state; not quite cancer, but certainly abnormal and with a predisposition to become cancerous. This leads to the observation that PIN images may not in fact be correctly classifiable as either cancer or noncancer, and our overall FP rate for cancer may be improved by placing PIN within a distinct third category.

2) Prostate Cancer Classification: The results of our cancer classification experiment should be interpreted in the context of the relatively small number of images in the data set (99 low-grade and 21 high-grade). We restricted the data set for this experiment to only the sub-images having pure G3 or G4 tissue type, excluding the $\mathrm{G}3+4$ and $\mathrm{G}4+3$ sub-images. Of the total of 991 sub-images in our data set (Table II), the 178 sub-images containing cancer could potentially be used for this experiment, with the remaining 813 noncancer images being relevant only to the cancer versus noncancer experiment (Section V-B1). Of these 178 sub-images, 120 $(99 + 21)$ came from within homogeneous tumors consisting purely of Gleason 3 and Gleason 4 cancer. Since these small (0.3 mm × 0.3 mm) sub-images were drawn from homogeneous tumors, it is reasonable to assign to each sub-image a gold standard label of Gleason 3 (low grade) or Gleason 4 (high grade); the homogeneity of the tumors supports the correctness of this gold standard labeling. However, 58 sub-images $(46 + 12)$ came from within heterogeneous tumors consisting of a mixture of Gleason 3 and Gleason 4 cancer. Due to the small size (0.3 mm × 0.3 mm) of our sub-images, one sub-image drawn from such a heterogeneous tumor is reasonably expected to consist purely of either Gleason 3 (low grade) or Gleason 4 (high-grade) cancer, but due to heterogeneity of the tumor from which the sub-image was extracted, there is no way to determine the correct golden standard label against which to compare classifier output. This is the reason that these 58 sub-images were not used for this experiment and the remaining 120 images were retained. This issue does not disrupt the prostate cancer detection experiment, since G3+4 and G4+3 sub-images uniformly contain cancer.

We also observed oscillation in the precision recall curve (Fig. 15) for this experiment (as compared with the precision recall curve for the prostate cancer detection experiment shown in Fig. 11). This fluctuation effect is possibly attributable to the relatively small size of our data set. Nevertheless, we did observe a reduction in the FN rate for high-grade cancer without a material concomitant rise in FP rate at $s = 5$ (Figs. 12 and 13). Testing on a larger sample size would provide further evidence against which to compare these results. However, with the advent of prostate-specific antigen testing and 18-gauge needle core biopsy, prostate cancer is being diagnosed earlier, making the presence of high-grade cancer at prostatectomy increasingly rare, challenging the acquisition of a large data set for high-grade versus low-grade classification evaluation.

VII. Conclusion

We have designed and evaluated a software system for prostate cancer detection and classification on digitized, hematoxylin & eosin-stained digital histopathology images. Our system uses two stages of AdaBoost-based classification. The first provides high-level tissue component labeling of a superpixel partitioning of the images. The second uses the tissue component labeling to provide a classification of cancer versus noncancer, and low-grade versus high-grade cancer. The superpixel partitioning provided a more than 60-fold reduction in data size, increasing processing efficiency. Using our database of 991 sub-images, our statistical testing measured accuracies of 90% and 85% for the cancer versus noncancer and high-grade versus low-grade classification tasks, respectively. We also measured a false-negative (FN) rate for cancer detection of 12% and for high-grade cancer detection (in our high-grade versus low-grade classification experiment) of 5%.

Our system determines the high-level tissue component labeling of superpixels without the use of any explicitly encoded domain knowledge, automatically learning the labeling from a training set. It therefore can potentially be trained to classify tissue components used by pathologists in analysis of other organs (e.g., breast, brain, etc.). To the best of the authors' knowledge, this work represents the first application of a superpixel image partitioning to the problem of digital histopathology image processing. Tissue component labels were applied based on morphometric, geometric, and appearance information derived from the superpixel partitioning. These labels encode high-level information, similar to that used by pathologists for the task of cancer detection and classification, and were found to support robust automation of these tasks in this study. This system represents a first step toward automated cancer quantification on prostate digital histopathology imaging, which could pave the way for better informed postprostatectomy patient care.

References

[1] J. H. Burkhardt, M. S. Litwin, C. M. Rose, R. J. Correa, J. H. Sunshine, C. Hogan, and J. A. Hayman, "Comparing the costs of radiation therapy and radical prostatectomy for the initial treatment of early-stage prostate cancer," *J. Clin. Oncol.*, vol. 20, no. 12, pp. 2869–2875, 2002.

[2] L. Egevad, J. R. Srigley, and B. Delahunt, "International society of urological pathology (ISUP) consensus conference on handling and staging of radical prostatectomy specimens: Rationale and organization," *Modern Pathol.*, vol. 24, no. 1, pp. 1–5, 2011.

[3] D. F. Gleason, *The Veteran's Administration Cooperative Urologic Research Group: Histologic Grading and Clinical Staging of Prostatic Carcinoma*. Philadelphia, PA: Lea and Febiger, 1977, pp. 171–198.

[4] T. H. van der Kwast, M. B. Amin, A. Billis, J. I. Epstein, D. Griffiths, P. A. Humphrey, R. Montironi, T. M. Wheeler, J. R. Srigley, L. Egevad, and B. Delahunt, "International society of urological pathology (ISUP) consensus conference on handling and staging of radical prostatectomy specimens. working group 2: T2 substaging and prostate cancer volume," *Modern Pathol.*, vol. 24, no. 1, pp. 16–25, 2011.

[5] C. Magi-Galluzzi, A. J. Evans, B. Delahunt, J. I. Epstein, D. F. Griffiths, T. H. van der Kwast, R. Montironi, T. M. Wheeler, J. R. Srigley, L. L. Egevad, and P. A. Humphrey, "International society of urological pathology (ISUP) consensus conference on handling and staging of radical prostatectomy specimens. working group 3: Extraprostatic extension, lymphovascular invasion and locally advanced disease," *Modern Pathol.*, vol. 24, no. 1, pp. 26–38, 2011.

[6] O. Veksler, Y. Boykov, and P. Mehrani, "Superpixels and supervoxels in an energy optimization framework," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 211–224.

[7] J. I. Epstein and G. J. Netto, *Biopsy Interpretation of the Prostate*. Philadelphia, PA: Lippincott Williams Wilkins, 2007.

[8] S. Doyle, C. Rodriguez, A. Madabhushi, M. Feldman, and J. Tomaszeweski, "A boosting cascade for automated detection of prostate cancer from digitized histology," in *Proc. Med. Image Comput. Comput. Assist. Intervent.*, 2006, pp. 504–511.

[9] S. Doyle, M. Feldman, J. Tomaszeweski, and A. Madabhushi, "A boosted Bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1205–1218, May 2012.

[10] 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 the tissue composition in the prostatic neoplasia," *J. Human Pathol.*, vol. 35, no. 9, pp. 1121–1131, 2004.

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

[12] P. Khurd, C. Bahlmann, P. Maday, A. Kamen, S. Gibbs-Strauss, E. M. Genega, and J. V. Frangioni, "Computer-aided Gleason grading of prostate cancer histopathological images using texton forests," in *Proc. Int. Symp. Biomed. Imag.*, 2010, pp. 636–639.

[13] S. Doyle, M. Hwang, S. Kinsuk, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated grading of prostate cancer using architectural and textural image features," in *Proc. Int. Symp. Biomed. Imag.*, 2007, pp. 1284–1287.

[14] 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 *Proc. Int. Symp. Biomed. Imag.*, 2008, pp. 284–287.

[15] M. Arif and N. Rajpoot, "Classification of potential nuclei in prostate histology images using shape manifold learning," in *Proc. Int. Conf. Mach. Vis.*, 2007, pp. 113–118.

[16] C. Wittke, J. Mayer, and F. Schweiggert, "On the classification of prostate carcinoma with methods from spatial statistics," *IEEE Trans. Inf. Technol. Biomed.*, vol. 11, no. 4, pp. 406–414, Jul. 2007.

[17] X. Xu, X. Xing, Y. Huang, and Z. Wang, "On the classification of prostate pathological images based on Gleason score," in *Proc. Int. Symp. Intell. Inf. Technol. Appl. Workshops*, 2008, pp. 605–608.

[18] K. Nguyen, A. Sarkar, and A. l. Jain, "Structure and context in prostatic gland segmentation and classification," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2012*. Berlin: Springer, 2012, vol. 7510, Lecture Notes in Computer Science, pp. 115–123.

[19] R. Farjam, H. Soltanian-Zadeh, K. Jafari-Khouzani, and R. A. Zoroofi, "An image analysis approach for automatic malignancy determination of prostate pathological images," *Cytometry B Clin. Cytom.*, vol. 72, no. 4, pp. 227–240, 2007.

[20] 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," *IEEE Trans. Med. Imag.*, vol. 26, no. 10, pp. 1366–1378, Oct. 2007.

[21] A. Vezhnevets and V. Vezhnevets, "Modest adaboost-teaching adaboost to generalize better," presented at the Graphicon, Novosibirsk, Russia, 2005.

[22] D. Comaniciu, P. Meer, and S. Member, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[23] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, 2004.

[24] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 1997.

[25] Y. Boykov, O. Veksler, and R. Zabih, "Efficient approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 12, pp. 1222–1239, Dec. 2001.

[26] Y. Boykov and V. Kolmogorov, "Computing geodesics and minimal surfaces via graph cuts," in *Proc. Int. Conf. Comput. Vis.*, 2003, pp. 26–33.

[27] Y. Boykov and G. F. Lea, "Graph cuts and efficient n-d image segmentation," *Int. J. Comput. Vis.*, vol. 70, no. 2, pp. 109–131, 2006.

[28] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 137–148, 2004.

[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints.," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[30] A. Vedaldi and B. Fulkerson, VLFeat: An open and portable library of computer vision algorithms 2008 [Online]. Available: http://www.vlfeat.org

[31] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering objects and their location in images," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, Oct. 2005, vol. 1, pp. 370–377, vol 1.

[32] Y. Freund and R. E. Schapire, "A short introduction to boosting," *J. Jpn. Soc. Artif. Intell.*, vol. 14, no. 5, pp. 771–780, 1999.

[33] X. Ren and J. Malik, "Learning a classification model for segmentation," in *Proc. Int. Conf. Comput. Vis.*, 2003, vol. 1, pp. 10–17.

[34] G. Mori, X. Ren, A. A. Efros, and J. Malik, "Recovering human body configurations: Combining segmentation and recognition," in *Proc. Comput. Vis. Pattern Recognit.*, 2004, vol. 2, pp. 326–333.

[35] D. Hoiem, A. Efros, and M. Hebert, "Geometric context from a single image," in *Proc. Int. Conf. Comput. Vis.*, 2005, pp. 654–661.

[36] G. Mori, "Guiding model search using segmentation," in *Proc. Int. Conf. Comput. Vis.*, 2005, pp. 1417–1423.

[37] X. He, R. S. Zemel, and D. Ray, "Learning and incorporating top-down cues in image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 338–351.

[38] T. Malisiewicz and A. A. Efros, "Improving spatial support for objects via multiple segmentations," in *Proc. Br. Mach. Vis. Conf.*, Sep. 2007, pp. 55.1–55.10.

[39] C. Pantofaru, C. Schmid, and M. Hebert, "Object recognition by integrating multiple image segmentations," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 481–494.

[40] B. Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 670–677.

[41] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. 2001 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2001, vol. 1, pp. 511–518.

[42] X. S. Zhou and T. S. Huang, "Relevance feedback in image retrieval: A comprehensive review," *Multimedia Syst.*, vol. 8, pp. 536–544, 2003.

[43] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, pp. 1–37, 2008.

[44] O. Pujol, M. Rosales, P. Radeva, and E. Nofrerias-Fernndez, "Intravascular ultrasound images vessel characterization using adaboost," in *Functional Imaging and Modeling of the Heart*, I. Magnin, J. Montagnat, P. Clarysse, J. Nenonen, and T. Katila, Eds. Berlin, Germany: Springer, 2003, vol. 2674, Lecture Notes in Computer, p. 1006.

[45] R. A. Ochs, J. G. Goldin, F. Abtin, H. J. Kim, K. Brown, P. Batra, D. Roback, M. F. McNitt-Gray, and M. S. Brown, "Automated classification of lung bronchovascular anatomy in CT using adaboost," *Med. Image Anal.*, vol. 11, no. 3, pp. 315–324, 2007.

[46] A. Quddus, P. Fieguth, and O. Basir, "Adaboost and support vector machines for white matter lesion segmentation in MR images," in *Proc. 27th Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2005, pp. 463–466.