Title
Prostate Cancer Detector for Pathology Images

Authors
Kabra, Mayank
Baird, Stephen
Mahooti, Sepi
et al.

Publication Date
2012-09-06

Peer reviewed

Prostate Cancer Detector for Pathology Images

Mayank Kabra[a], Stephen Baird[b], Sepi Mahooti[b], Jihoon Kim[c,d], Lucila Ohno-Machado[c,d], Yoav Freund[e,*]

[a]*Electrical and Computer Engineering, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093*
[b]*Dept. of Pathology, School of Medicine, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093*
[c]*Division of Biomedical Informatics, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093*
[d]*San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093*
[e]*Computer Science and Engineering, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093*

Abstract

Availability of slide scanners and electronic medical record systems have led to increased digitization of pathology images. With digitization, Computer-Aided Diagnosis (CAD) tools can be built to reduce pathologist's fatigue and improve diagnosis workflow. As a first step towards building such tools, we developed a cancer detector for prostate needle core biopsy images. The detector is trained using boosting. We use color and texture features. We also built new structural features that express visual cues not captured by color and texture features. These structural features inform the classifier whether cancer has deformed the glands, the main functional unit of prostate. On the test images we scored, we got an Area Under the ROC Curve (AUC) of 0.94.

Keywords: prostate, histopathology, adaboost, morphological features, Computer-Aided Diagnosis

*Corresponding author
Email addresses: mkabra@ucsd.edu (Mayank Kabra), sbaird@ucsd.edu (Stephen Baird), smahooti@ucsd.com (Sepi Mahooti), j5kim@ucsd.edu (Jihoon Kim), lohnomachado@ucsd.edu (Lucila Ohno-Machado), yfreund@cs.ucsd.edu (Yoav Freund)

1. Introduction

Prostate cancer is the second leading cause of cancer death of males in the USA. The American Cancer Society predicts that, during 2010, over 217,730 new cases will be diagnosed and over 32,050 men will die of the disease. As early cancer detection is critical for survival, aging men are regularly screened for the disease. Among the methods to diagnose prostate cancer, pathologic examination of prostate tissues is the only definitive method (Matlaga et al., 2003). As a result, around a million biopsies are performed annually in the US making prostate one of the most commonly examined pathological tissue.

The recent availability of affordable digital scanners is moving the practice of detecting cancer forward. The digital scanning microscopes can generate digital images of whole slides at 400x magnification necessary for diagnosis. These scanners are increasingly being adopted by hospitals and slide images are becoming a part of the patient's electronic medical records. Digital images provide an immediate benefit to pathologists who can now examine slides on monitors instead of bending over microscopes. But in the long term digitization can improve pathologist's efficiency and accuracy by developing Computer-Aided Diagnostic (CAD) tools.

Tissue slides are complex which makes it difficult to build completely automated CAD tools. Even for humans, diagnosing tissue slides demands skill and years of practice. Typical tissue samples contain a mixture of benign and malignant structures that sometimes appear indistinguishable to a lay person. Even if completely automated CAD tools cannot be built right now, based on current technology CAD tools can be built that can improve current diagnostic practice. For example, when a biopsy is performed on a patient, multiple tissue cores are extracted and mounted on different slides. The pathologist examines all the slides to find the ones that can be used to reliably diagnose cancer. An automated cancer detector could reduce the number of slides examined by the pathologist by sorting the slides according to the likelihood of cancer. The sorting would increase the odds of pathologists encountering the cancer in the first few slides they examine.

Another possible CAD tool could reduce the cost of doing remote diagnosis using tele-pathology. Setting up the infrastructure for tele-pathology requires considerable investment. One of the main expenses is the large bandwidth required for transmitting large-sized digital pathology images. Pathology images are larger than other medical images as the entire slide has to be scanned at the high magnification of 400x. The bandwidth required to diagnose cancer can be reduced by using a cancer detector. Such a detector can be used to summarize the images and send only those parts of the image that were detected as cancer or suspicious for cancer. If the pathologist is not satisfied, a backup option could be provided where the whole high-

power image is transmitted. If the detector can detect cancer on 90% of the images, then the summarization would reduce the bandwidth by one order of magnitude.

As a step towards building a platform for such pathology CAD tools, we undertook the task of building a cancer detector for prostate tissue images. We chose prostate because its cancer is one of the most common causes of cancer fatality in the USA. Furthermore, prostate tissues are similar to breast and lung biopsy tissues, so it is likely that techniques developed for prostate cancer detectors would be useful when developing breast and lung cancers detectors.

At the core of our cancer detector are machine learning techniques that have been successfully applied in analyzing other non-medical images. We use boosting, a machine learning technique, as it provides the capability to learn an accurate detector based on multiple visual cues.

One of the important visual cues that pathologists look for are changes in the structure of glands which undergo mild to severe deformations when malignant. Capturing these structural changes is difficult using color and texture features. We developed a learning based technique that detects three main types of gland structures: healthy glands, small glands typical in low grade cancer and completely unstructured mass of epithelial cells found in high grade cancer.

We measured the performance of our detector by measuring the area under the ROC curve for classifying pixels. The AUC for our detector was 0.94 which compares favorably to previous published best of 0.84 by Doyle et al. (2010). The images, training data and code are publicly available (Appendix A).

The paper is organized as follows. In Section 2 we review related work. In Section 3 we briefly review the prostate anatomy. We give a brief overview of our system in Section 4. We review boosting in Section 5, give details of the feature engineering in Section 6 and 7 and give the details of the data in Section 8. In Section 9 we give the results we obtained and finally conclude in Section 11.

2. Related Work

The work most related to ours is the recent work by Doyle et al. (2010). They also developed a cancer detector in prostate needle core images. The features used were first order pixel value statistics and texture features based on co-occurence and Gabor filters. For each feature, probability densities were estimated for the cancer and non-cancerous pixels. Based on these density estimates the optimal threshold values were calculated. These weak rules were then combined using boosting to learn a pixel classifier. To speed up the processing, the cancerous regions were found by classifying the image at multiple resolutions, starting with the coarsest resolution

and discarding obvious negatives and then moving to finer resolutions. The area under the ROC curve for their detector was 0.84.

Recently, Xu et al. (2010) developed a method to detect glands and study the morphological structure of the glands for needle core prostate biopsies. They combined normalized cuts with active contours to segment the glands. The segmented structures are then classified as cancerous or non-canceorous based on the shape features developed by Sparks and Madabhushi (2010). Similarly, Monaco et al. (2010) also developed a gland segmentation and classification technique for whole prostate sections. They used a region growing technique to segment the glands. Glands with similar sizes are clustered using markov random fields to mark the extent of similar regions.

Prior to these studies most of the work (Tabesh et al., 2007; Diamond et al., 2004; Farjam et al., 2007) was concentrated on separating small homogenous patches of tissue mainly because whole slide scanners were unavailable. Lack of whole slide scanners also meant that the data sets used were small. Diamond et al. (2004) developed a system to distinguish homogenous patches of stroma, benign epithelium and prostate cancer with an accuracy of 79.3%. Tabesh et al. (2007) developed a system to distinguish small spots on a tissue microaccray. Their system could separate the cancerous and non-cancerous images with an accuracy of 96.7% and low-grade cancer and high-grade cancer with an accuracy of 81.0%. They detect tissue parts such as nuclei, cytoplasm, red blood cells etc. using hand-coded parameters. Properties of these tissue components are used as features. In our work, we also detect structure but we detect different type of structures and also we use learning to detect the structures. Farjam et al. (2007) achieved an accuracy of 96.5% in grading homogenous image patches.

3. Prostate Anatomy

The visual appearance of prostate tissue is determined by its anatomy and hence a good understanding of the anatomy is crucial for building the cancer detector. In this section we briefly describe the anatomy (Ackerman and Rosai (1974)).

The prostate's function is to produce seminal fluid. The fluid is secreted by specialized epithelial cells. The epithelial cells are arranged in glands and ducts (Figure 1). The lumen form a network that collects the seminal fluid and delivers it out of the prostate. The network of glands and ducts is surrounded by connective tissue (stroma). Between the connective tissue and the epithelial cells, there is a layer of basement membrane that keeps the epithelial cells attached to the stroma and confined within the glands. The prostate mainly consists of stroma and glands.

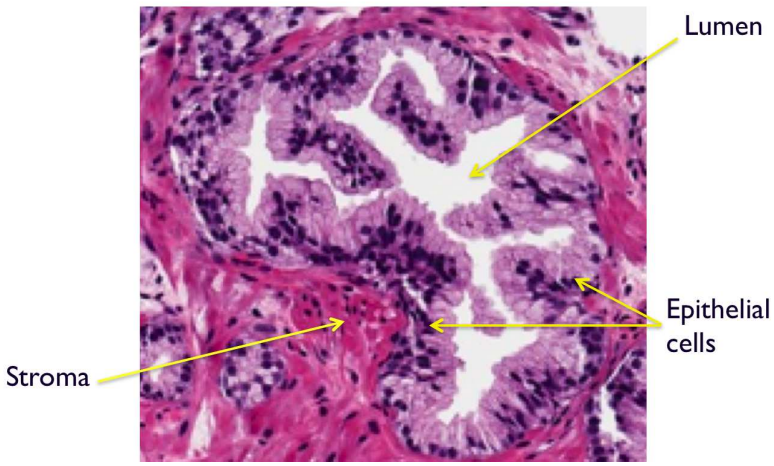

Figure 1: An example of prostate gland. The glands are the main functional unit of the prostate. The epithelial cells that line the lumen secrete seminal fluid. The network of ducts and glands that collects the fluid is supported by stroma. In prostate cancer, the epithelial cells mutate and alter the gland structure and invade the stroma. These cells can also metastasize. Details of the glands are best visible at 40x magnification.

For disease diagnosis thin slices of tissue samples are stained with hematoxylin and eosin (H&E stain). The thin slices mounted on glass slides are first examined at low magnification to study the gross appearance and then at a higher magnification to study the finer details.

At low magnification, the glands are distinctive with a peculiar structure. While the lumen is irregular, the epithelial cells are constrained to be in a single layer. Also the nuclei of the epithelial cells are programmed to stay away from lumen. Since nuclei are stained dark blue by hematoxylin, lumens are always surrounded by first a pale outline of the epithelial cell's cytoplasm which in turn is surrounded by a dark line of epithelial cell nuclei. Stromal cells on other hand have smaller nucleus and larger cell size. Hence it appears smooth without any activity and is colored pink by eosin.

The underlying cause of prostate cancer is mutation of the genome of the epithelial cells. Prostate cancer is diagnosed by studying the changes brought about by these mutations. The stroma does not usually undergo important changes and is not usually studied for diagnosis. It sometimes does undergo increased cell division when stimulated by invading epithelial cells in which case it is studied for diagnosis.

Two important changes exhibited in H&E examination are the change in nuclear to cytoplasmic ratio, and the change in the cytoplasmic color of the epithelial cells

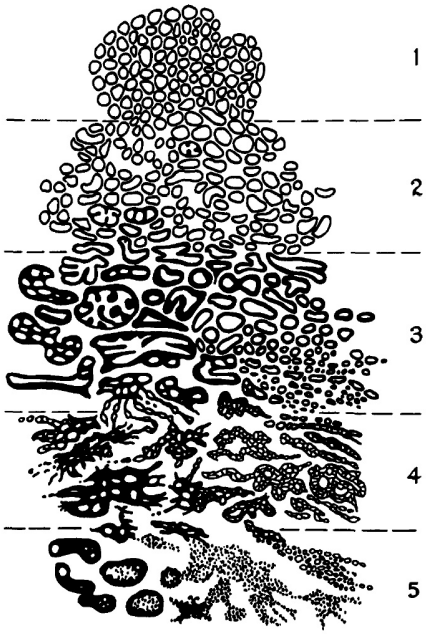

Figure 2: Deformation of the gland structure according to the cancer grade. The glands become more and more unstructured with increasing gleason grade (Gleason, 1966) denoted on the right.

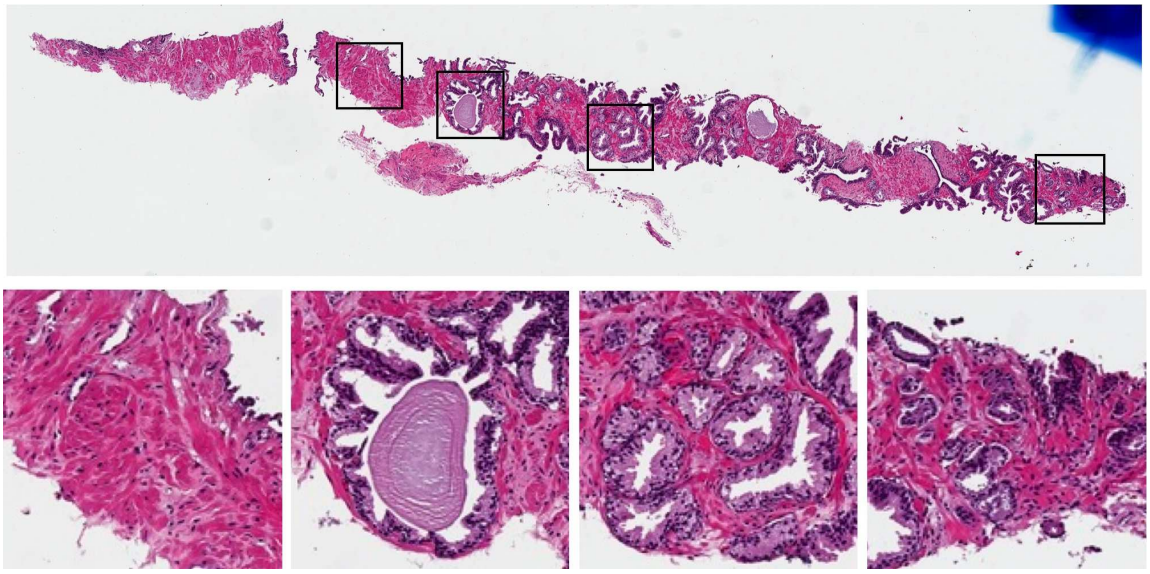

Figure 3: An example of needle core biopsy. Most of the tissue is composed of stroma which is featureless and not difficult to separate. Most of the diagnostic information is in the structure of glands that can be distorted and broken in needle core biopsies. A tissue sample has large number of structures such as concretions in the second patch that make diagnosing cancer a difficult task.