Proceedings of the 8th INFORMS Workshop on Data Mining and Health Informatics (DM-HI 2013)
O. Seref, N. Serban, D. Zeng, eds.

Discovering Anomalous Patterns in Large Digital Pathology Images

Sriram Somanchi, Daniel B. Neill
Event and Pattern Detection Laboratory
Carnegie Mellon University
Pittsburgh, PA
somanchi@cmu.edu, neill@cs.cmu.edu

Abstract

Advances in medical imaging technology have created opportunities for computer-aided diagnostic tools to assist human practitioners in identifying relevant patterns in massive, multi-scale digital pathology slides. This work presents Hierarchical Linear Time Subset Scanning (HLTSS), a novel pattern detection method which exploits the hierarchical structure inherent in data produced through virtual microscopy in order to accurately and quickly identify regions of interest for pathologists to review. We take a digital image at various resolution levels, identify the most anomalous regions at a coarse level, and continue to analyze the data at increasingly granular resolutions until we accurately identify its most anomalous sub-regions. We demonstrate the performance of our novel method in identifying cancerous locations on digital bslides of prostate biopsy samples, and show that our methods detect regions of cancer in a few minutes with high accuracy both as measured by the ROC curve (measuring ability to distinguish between benign and cancerous slides) and by the spatial precision-recall curve (measuring ability to pick out the malignant areas on a slide which contains cancer).

Keywords: Anomalous Pattern Detection, Pathology Informatics, Digital Pathology

1 Introduction

Anatomic pathology is a medical speciality which includes diagnosing a disease from biopsy samples of an organ. For decades, the pathology work flows have been highly manual, where thin slices of biopsy specimens are histochemically stained on to a glass slide and are analyzed by a pathologist, under an optical microscope, for cell composition. More recently, advances in computer aided medical diagnostics have introduced a digital work flow for pathology. Digital pathology is designed to achieve similar goals as traditional pathology using computerized software and has grown dramatically in the last ten years [2]. Many pathology laboratories are on the path towards modernizing and updating their work flows using these advanced techniques. Digital pathology offers many advantages over traditional anatomic pathology: remote pathology, secure and easy distribution of images among experts, and the ability use informatics tools to analyze the digital images. It is typically believed that a pathologist is under time pressure to analyze these images and to deliver accurate and timely diagnosis [3]. In this paper, we introduce novel detection methods that can aid a pathologist by quickly identifying potential regions of interest in a very high resolution digital image. These regions are automatically detected and highlighted in an image viewer for further examination by a pathologist. This process helps in timely diagnosis by a pathologist and eventually provides input for secondary opinions on regions of interest that might have been originally missed by a pathologist.

1.1 Approach based on Subset Scanning

Typically, the regions of interest for a pathologist in a tissue biopsy contain patterns that are abnormal as compared to the rest of the tissue structure. Hence, our approach for aiding medical diagnostics is to detect regions that contain anomalous patterns in these large-scale multiresolution images. Specifically, we approach pattern detection as a "subset scann" problem [5], where we search over subsets of data with the goal of finding the most anomalous subsets.

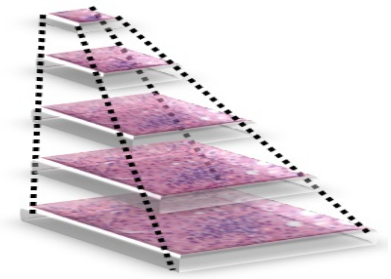

Figure 1: An example of H & E stained Whole Slide Image (WSI) where images at multiple resolutions are stored as pyramidal structure. This helps a pathologist to view in a virtual microscopy.

There are typically two main challenges in subset scan based approaches. Statistically quantifying the anomalousness of a subset and efficiently finding the most anomalousness subset. Anomalousness of a subset can be quantified by a loglikelihood ratio statistic like the Expectation Based Poisson [4] or Expectation Based Gaussian [5] statistics. A major computational hurdle of "subset scan" approach to pattern detection is that we need to optimize the score function over an exponential number of subsets. Linear Time Subset Scanning (LTSS) [5] is a novel approach to anomalous pattern detection that addresses this issue by identifying the most anomalous subset without exhaustive search.

If a scoring function satisfies the LTSS property then we can find the subset that maximizes the scoring function by evaluating only a linear number of subsets (i.e., the number of subsets evaluated is proportional to N rather than 2^N where N is the number of data elements). Many scoring functions in the literature satisfy the LTSS property [5] which helps in efficiently finding the best subsets. However, when the data is huge (i.e., number of data elements is in billions or trillions) even evaluating a linear number of subsets might not be feasible. Also, finding the best subset without constraints on the subsets might result in including a lot of noise in the detected subsets as there could be many elements which are individually anomalous by chance. In this paper, we introduce a new framework called Hierarchical Linear Time Subset Scanning (HLTSS) to address these two issues. The HLTSS method assumes that the data is stored in hierarchical fashion, where elements are aggregated at multiple levels of hierarchy. For example, in the case of digital pathology, the image of a biopsy sample is stored at multiple resolutions in order to help a pathologist analyze it through virtual microscopy. HLTSS takes advantage of this hierarchical structure in the data to improve the efficiency and accuracy in detecting anomalous subsets. We show that any scoring function that satisfies the LTSS property can be incorporated into our new HLTSS framework.

1.2 Application to Digital Pathology

We apply the HLTSS framework to the problem of identifying regions of interest in digital pathology images. As explained earlier, the digital pathology images are stored at multiple resolutions, with the lowest (coarsest) resolution image providing the overall "big picture" and the highest (finest) resolution image providing very detailed structure of the biopsy sample. Figure 1 provides an example image with multiple resolution images shown in the form of a pyramid. At the highest resolution, these images are very large, typically in the range of $100K \times 100K$ pixels. A pathologist typically uses all of the various resolutions of an image, starting from very low resolution and "zooming in" to high resolution image, to pinpoint abnormal locations on a slide. Our method is motivated by this procedure: we take a digital image at various resolution levels, identify the most anomalous regions at a coarse level, and continue to analyze the data into more granular resolutions until we accurately identify its most anomalous sub-regions.

Each pixel in an image is typically represented in ARGB format, a 32 bit representation with 8 bits each for Alpha, Red, Green and Blue. Alpha is the amount of transparency, which is typically opaque in digital pathology (and therefore ignored for the purposes of this paper). The rest of the components (Red, Green and Blue) are used to represent the color of the pixel. Pyramidal images are typically representative of the hierarchical aggregation of each component [6]. That is, if a given pixel at a higher (coarser) level is representative of a grid of 4x4 pixels at a lower level in the hierarchy, then each component of the higher level pixel is approximately the aggregation of the corresponding component of the 16 pixels below it. Also we note that each pixel at a lower level is typically aggregated to one and only one pixel at a higher level.

Digital pathology images are typically Hematoxylin and Eosin (HE) stained images. The staining is done in such a way that the resulting image is typically composed of hematoxylin (violet color), eosin (pink color) and white (background color). Hematoxylin is typically indicative of the nuclei of the cells and pink is indicative of other tissue regions like cytoplasm. In a pathology image a region is interesting for pathologist due to various abnormalities. In this work we concentrate on one specific feature of a region based on anomalous coloration. More precisely we focus on the identification of regions that are interesting because they contain a higher than expected concentration of violet pixels (hematoxylin dye). There are various applications where identifying this pattern is useful: identifying regions of inflammation in gastrointestinal tracts for Crohn.s disease; finding regions of inflammation (gastritis) in the stomach, which may be indicative of colonization by *Helicobacter* pylori; and diagnosis of prostatic intraepithelial neoplasia, which may lead to prostate cancer. Specifically, we apply this methodology for identifying cancerous regions in a prostate biopsy sample.

Prostate cancer is the most prevalent form of cancer and the second most common cause of cancer deaths among men in the U.S. [8]. About one in every six men will be diagnosed with prostate cancer during their lifetime [8]. Pathologists rely on examination of biopsy samples under a microscope for detecting the cancerous cases. It is extremely important for a pathologist to quickly identify cases of cancer so that there can be early intervention to improve the prognosis. The most important features that are used by a pathologist in order to differentiate cancerous locations from benign are based on nucleic, cytoplasmic, luminal and architectural characteristics of the cells. The first two features are based on color characteristics and the last two are based on shape characteristics. The nucleic features that are different include prominent and enlarged nucleoli. The cytoplasmic features include darkness in the shade. In both of these nuclear and cytoplasmic features, the locations that are indicative of cancer are typically more violet in color than the typical benign location [7]. Also the presence of a large nucleoli is the single most important criteria for diagnosis of prostate cancer [1]. Hence, we hypothesize that our anomaly detection algorithm which finds regions which have higher than expected concentrations of violet can be used for potentially identifying locations indicative of cancer; this hypothesis is confirmed by our evaluation results presented below.

We quantify the anomalousness of a region using the Expectation Based Binomial (EBB) scoring function, where we map every pixel to a continuum of white, pink and violet and pinpoint regions containing a higher than expected proportion of violet pixels. We show that EBB satisfies the LTSS property, and hence can be easily incorporated into our HLTSS framework. More detailed information of our algorithm and framework is available in our detailed technical report [9] We apply our methodology to digital images of prostate biopsy samples from the Department of Pathology at University of Pittsburgh Medical Center (UPMC) to identify cancerous locations in these images. We show that our methodology helps in differentiating between cancerous and benign images. This is very important for a pathologist to prioritize his work as it helps him to concentrate on cancerous images. Further, we show that our methods have good accuracy in picking out the malignant areas on a slide which contains cancer.

2 Empirical Evaluation and Results

In order to demonstrate the performance of our method we pose the problem of identifying regions of interest on digital pathology image as the subset scanning problem of identifying the most anomalous groups of pixels. We work with pathology images of prostate biopsy samples of patients who are suspected to have prostate cancer. This method of identifying cancerous regions can be used to first differentiate the cancerous images from benign images. Once we have identified the cancerous images, we can find regions of interest within each large-scale image, which aids a pathologist to make further decisions.

2.1 Data

We have digital pathology images of prostate biopsy samples from the Department of Pathology, University of Pittsburgh Medical Center. There are 14 images which contained cancerous locations and 14 images which were examples of benign prostate biopsy samples. The former 14 cancerous images were also annotated indicating the locations of cancer. These annotations were used for evaluating the performance of our method in detecting regions of interest within a digital image. Each image contained about 5 billion pixels at the most granular level and take up around $10GB$ of space in uncompressed form. We are in the process of acquiring more images with annotations to show the effectiveness of our method on a large number of examples.

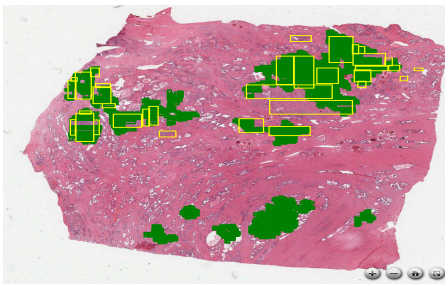

Figure 2: An example of our detected regions as compared to ground truth. The green region is detected by our HLTSS method and yellow are ground truth boxes drawn by a pathologist.

2.2 Methods and Metrics

We apply our Hierarchical Linear Time Subset Scanning (HLTSS) algorithm for all the images obtained from UPMC. We compare the performance of our method with the original Linear Time Subset Scanning (LTSS) algorithm [5]. In order to show the effectiveness in using hierarchically constrained subsets, we run the LTSS algorithm at the most coarse level (LTSS-coarse) and at the most granular level (LTSS-granular) considered by HLTSS. We evaluate the performance of our methods to distinguish between benign and cancerous slides by showing the Receiver Operator Characteristics (ROC) curve. Further we evaluate the ability to pick out the malignant areas on a slide which contains cancer by showing Precision-Recall curves. Precision provides the percentage of detected locations that are actually cancerous, while Recall gives the percentage of cancerous locations that are detected.

As explained earlier, the cancerous images were annotated by indicating cancerous locations on these images. We use this for evaluating our precision and recall. However the ground truth results were provided as rectangles at a very coarse level as shown in Figure 2 (the yellow rectangles are the regions marked by a pathologist which contain cancerous cells). The ground truth information is provided at a coarse level as it would take a lot of time for a pathologist to mark exactly mark every small region as cancerous. As our HLTSS algorithm finds the locations in a more granular level than the ground truth data, we draw a bounding box each subset detected by HLTSS and report the bounding box as our detected cluster. Note that the bounding box is only used for visualization and evaluation purposes, the anomalousness and score of a bounding box is still based on the most granular level pixels detected within the bounding box. We repeat the same procedure for our LTSS-coarse and LTSS-granular algorithms as well.

2.3 Experimental Results

Figure 2 shows an example of our detected regions as compared to the ground truth results. The green regions are the pixels that we have identified and yellow are the coarse level cancerous locations marked by a pathologist. We can see that we have identified most of the regions with high accuracy though there were a few false positives and false negatives. We currently show only the top 10 clusters in the image. Note that the regions we could not identify as cancerous (false negatives) were because either they were too subtle locations to be included in top 10 or they were cancerous because of shape rather than color characteristics. Figure 2 is an example and we provide overall accuracy results in the following sections.

2.3.1 Accuracy in identifying benign slides from cancerous slides

Figure 3 provides the Receiver Operating Characteristic curve for differentiating cancerous images from benign images. This is very important for a pathologist as they have a huge number of slides to analyze per day and effectively identifying potentially cancerous slides helps them in prioritizing their work. For each image (both cancerous and non-cancerous), we run our algorithms to score an image (based on the highest region score found for that image) and sort the scores in decreasing order. We go through the sorted list and to determine the true and false positive rate. Essentially we are examining each method's tradeoff between its false positive rate (fraction of non-cancerous images

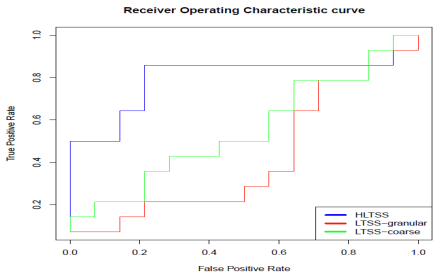

Figure 3: ROC curve for differentiating cancerous images from benign images

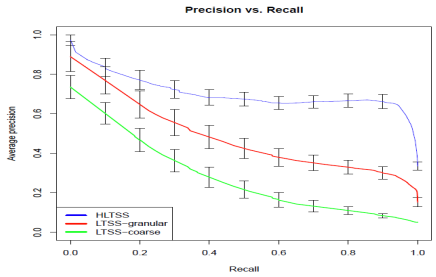

Figure 4: Precision and Recall curve for finding cancerous regions within a cancerous image.

labeled as cancerous) and true positive rate (fraction of cancerous images labeled as cancerous).

From Figure 3, we can see that the HLTSS method has higher true positive rate for most false positive rates. Also it has high true positive rate (over 80%) for a low (20%) false positive rate, while the LTSS-based methods over less than 40% true positive rate. The LTSS-based methods are either similar to or worse than random chance. This is because they are picking up just scattered violet pixels rather than potentially relevant clusters of pixels, and hence are unable to differentiate cancerous images from benign images.

2.3.2 Accuracy in identifying cancerous regions

We use the precision vs. recall curve as a metric to evaluate and compare the accuracy of detected regions within a cancerous image. In order to calculate this metric we record the list of scores returned for each pixel in an image. The score of each pixel is the score of the detected group of which it is a member. Any pixels that are not detected in any group are scored individually and ranked below all "detected" pixels. We sort the list of scores and we use each unique score value as a threshold for classifying locations, and calculate the precision (fraction of correctly identified cancerous pixels to total number of pixels in the detected region) and the recall (fraction of correctly identified cancerous pixels to total number of cancerous pixels). These precision and recall values are plotted on the curve for each image and are vertically averaged to obtain a single curve for each method with standard errors shown in Figure 4. From the figure, we can see that our HLTSS method maintains almost constant precision while capturing most of the cancerous locations in an image. However, non-hierarchical LTSS-based methods suffer large drops in precision for high recall values. One reason for this is that LTSS-based methods identify even small violet locations in an image as cancerous and hence have low precision. They have to essentially pick out all "violet" locations in whole image to capture the cancerous locations, whereas HLTSS is accurate in providing only the locations which might be interesting to a pathologist. Hence our results show that the HLTSS algorithm can not only differentiate cancerous images from benign images, but also accurately identifies regions of interest within a cancerous image.

3 Conclusions and Future Work

In this work, we developed and evaluated a novel Hierarchical Linear Time Subset Scanning (HLTSS) framework for detecting regions of interest in massive, multi-scale digital pathology slides. Such images typically consist of $10B$ pixels or more at the finest resolution, but are stored as multiple "layers" each representing a hierarchical aggregation of data from the previous layer. HLTSS exploits this hierarchical structure inherent in data produced through virtual microscopy in order to accurately and quickly identify regions of interest for pathologists to review. We proposed an Expectation Based Binomial (EBB) scoring function that can be used in our HLTSS framework to quantify the anomalousness of a group of pixels and identify regions which contain anomalously high concentrations of violet color. We demonstrate the performance of our novel methods in identifying cancerous locations on digital slides of prostate biopsy samples, and show that our methods detect regions of cancer in a few minutes with high accuracy both as measured by the ROC curve (measuring ability to distinguish between benign and cancerous slides) and by the spatial precision-recall curve (measuring ability to pick out the malignant areas on a slide which contains cancer).

Acknowledgments

This work was partially supported by National Science Foundation grants IIS-0916345, IIS-0911032, and IIS-0953330. We thank Dr. Anil Parwani and University of Pittsburgh Medical Center for their support in providing us the annotations and digital pathology images.

References

[1] Mettlin C, Murphy GP, Lee F, Littrup PJ, Chesley A, Babaian R, Badalament R, Kane RA, and Mostofi FK. Characteristics of prostate cancers detected in a multimodality early detection program. the investigators of the american cancer society-national prostate cancer detection project. *Cancer*, 2013.

[2] Romero Lauro G, Cable W, Lesniak A, Tseytlin E, McHugh J, Parwani A, and Pantanowitz. Digital pathology consultations-a new era in digital imaging, challenges and practical applications. *Journal of Digital Imaging*, 2013.

[3] Parwani AV Ho J, Aridor O. Use of contextual inquiry to understand anatomic pathology workflow: Implications for digital pathology adoption. *Journal of Pathology Informatics*, 2012.

[4] D. B. Neill. Expectation-based scan statistics for monitoring spatial time series data. *International Journal of Forecasting*, 25:498–517, 2009.

[5] D. B. Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society (Series B: Statistical Methodology)*, 74(2):337–360, 2012.

[6] Parwani AV Park S, Pantanowitz L. Digital imaging in pathology. *Clinics in laboratory Medicine*, 2012.

[7] Anil Parwani, 2013. http://teleconference.upmc.edu/2013/0418b2013/0418b2013.html.

[8] Amer. Cancer Soc., 2013. Cancer Facts and Figures.

[9] Sriram Somanchi. Discovering anomalous patterns in large digital pathology images. Technical report, Carnegie Mellon University, 2013.