IDENTIFYING THE HISTOMORPHOMETRIC BASIS OF PREDICTIVE RADIOMIC MARKERS FOR CHARACTERIZATION OF PROSTATE CANCER

by

GREGORY PENZIAS

Submitted in partial fulfillment of the requirements

For the degree of Master of Science

Thesis Advisor: Dr. Anant Madabhushi

Department of Biomedical Engineering

CASE WESTERN RESERVE UNIVERSITY

January, 2017
Identifying the Histomorphometric Basis of Predictive Radiomic Markers for Characterization of Prostate Cancer

Case Western Reserve University

We hereby approve the thesis of

GREGORY PENZIAS

for the degree of

Master of Science

Dr. Anant Madabhushi 09/08/2016
Committee Chair, Advisor
Department of Biomedical Engineering

Dr. Satish Viswanath 09/08/2016
Committee Member
Department of Biomedical Engineering

Dr. David Wilson 09/08/2016
Committee Member
Department of Biomedical Engineering

We certify that written approval has been obtained for any proprietary material contained therein.
# Table of Contents

List of Tables .............................................................................. v

List of Figures ........................................................................... vi

Acknowledgements .................................................................... xii

Abstract ....................................................................................... xiv

Chapter 1. Introduction ................................................................. 1

Problem Background ...................................................................... 1

Scope of this Work ........................................................................ 2

Thesis Overview ........................................................................... 4

Chapter 2. AutoStitcher: An Automated Program for Reconstructing Digitized Histology .......................................................... 5

Introduction .................................................................................. 5

Methods ....................................................................................... 10

Experimental Design ..................................................................... 18

Results and Discussion .............................................................. 24

Chapter 3. Identifying the Morphologic Basis for Radiomic Features in Prostate Cancer ........................................................... 34

Introduction .................................................................................. 34

Methods ....................................................................................... 39

Experimental Design ..................................................................... 47

Results ........................................................................................... 49

Discussion ...................................................................................... 52
## List of Tables

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Notation</td>
<td>11</td>
</tr>
<tr>
<td>2.2</td>
<td>Parameter Selection</td>
<td>23</td>
</tr>
<tr>
<td>2.3</td>
<td>Multi-Site Evaluation</td>
<td>31</td>
</tr>
<tr>
<td>2.4</td>
<td>Time to stitch a single section using AutoStitcher vs. HistoStitcher</td>
<td>33</td>
</tr>
<tr>
<td>3.1</td>
<td>MR Parameters</td>
<td>40</td>
</tr>
<tr>
<td>3.2</td>
<td>List of Extracted QH Features</td>
<td>44</td>
</tr>
<tr>
<td>3.3</td>
<td>List of Extracted Radiomic Features</td>
<td>45</td>
</tr>
<tr>
<td>3.4</td>
<td>Selected Radiomic and QH Features</td>
<td>50</td>
</tr>
</tbody>
</table>
List of Figures

2.1 Illustration of assumptions made in related problem domains in image assembly and document reconstruction. (a) A pair of overlapping images of a natural scene (green box highlights region of overlap). (b) Photograph of a natural scene with no missing pieces (whole image highlighted in green, stitched subimages boxed in dashed lines). (c) Interlocking archetypal jigsaw puzzle pieces (green arrows highlight interlocking edges). (d) Two corresponding prostate histology quadrants, which do not overlap (red box highlights region of non-overlap). (e) A stitched PWMHS which lacks completeness (whole image outlined in green, missing pieces outlined in red). (f) Non-interlocking prostate histology quadrants (red X’s highlight lack of interlocking edges). ................. 9

2.2 Illustration of AutoStitcher’s preprocessing and low resolution initialization workflow. (a) Initial unprocessed quadrant image. (b) Resizing, flipping, and conversion to grayscale. (c) Segmentation of tissue mask. (d) Bounding-box fitting. (e) Identification of edges. (f) Identification of best-fit lines, followed by automatic rotation. (g) Initial unprocessed quadrant images for a single slice. (h) Automatically rotated and flipped quadrant images. (i) Low resolution initialization. ........................................... 13

2.3 Illustration of the computation of the cost function for AutoStitcher used for stitching at high resolution. (a) Stitched image, with two
corner points (stars), which are used to compute misalignment $M$, plotted on the ends of each quadrant’s best-fit lines. (b) A pair of corresponding patches on quadrants $Q^{ur}$ and $Q^{lr}$. (c) Intensity histograms of the two patches. Since these patches appear visually similar, their histograms are also similar, thus this pair will have low dissimilarity $D$. (d) Another pair of corresponding patches on quadrants $Q^{ur}$ and $Q^{lr}$, and (e) their intensity histograms, which are more dissimilar, reflecting the apparent visual dissimilarity of their patches.

2.4 Final reconstruction results of (a) AutoStitcher vs. (b) HistoStitcher, with ASF evaluation fiducials plotted (via red, green, blue, and yellow asterisks). The enlarged boxed regions in (c,d) display sample ASF fiducials for each result. Note that since fiducials were selected on the HistoStitcher reconstruction, the panel in (d) shows both points being superposed onto the same yellow asterisk, compared to distinct points when mapped onto the AutoStitcher reconstruction in (c). Normalized ASF error was computed to be 2.61% in this example.

2.5 Final reconstruction results of (a) AutoStitcher vs. (b) HistoStitcher, with MSF evaluation over the fiducials plotted (in red, green, blue, and yellow asterisks). The enlarged boxed regions in (c) and (d) display sample MSF fiducials for the AutoStitcher and HistoStitcher reconstructions, respectively. Normalized MSF error was computed to be 1.89% in this example.
2.6 Hausdorff distance evaluation computed on the convex hull of (a,e) a PWMHS reconstructed using AutoStitcher next to (b,f) a PWMHS reconstructed using HistoStitcher. The convex hulls are first transformed to a common coordinate system as shown in (c,g), then aligned such that the Hausdorff distance is minimized in (d,h). The location of the maximum Hausdorff distance is highlighted in the black boxes, resulting in normalized Hausdorff errors of 3.81% (d) and 2.43% (h).

2.7 Visual inspection of PWMHSs stitched by (a,c) AutoStitcher and (b,d) HistoStitcher, to reveal common causes of reconstruction error: (i) misalignment of quadrants, (ii) under-compensation for missing tissue gaps, (iii) lack of scaling, and (iv) excessive overlap. AutoStitcher reconstruction in panel (a) has MSF, ASF, and Hausdorff errors of 6.56%, 5.85%, and 5.46% compared to corresponding HistoStitcher reconstruction in Figure 2.7b. AutoStitcher reconstruction in Figure 2.7c has MSF, ASF, and Hausdorff errors of 5.51%, 3.27%, and 5.70% compared to HistoStitcher reconstruction in Figure 2.7d.

2.8 Box and whisker plots of evaluation results between the training cohort, the testing cohort, and the validation cohort, with boxes marking the median, lower-quartile, and upper-quartile, dashed-lines connecting the extremes; and crosses marking outliers. Boxes are colored green, cyan, and purple to distinguish between the three cohorts.
3.1  (a) An H&E-stained histology tumor patch. (b) Gland lumen are segmented using an automatic region-growing algorithm within the tumor region. (c) Segmentation boundaries enclosing the gland lumen provide measures of shape, such as Fourier descriptors, which are decompositions that capture various levels of shape complexity, as illustrated here with Fourier descriptor 9 (blue = low, red = high). Gland lumen architectural features are extracted from (d) voronoi diagram, (e) delaunay triangulation, and (f) minimum spanning tree. (g) Identification of directional tensors for each segmented gland lumen allows extraction of features such as co-occurring gland angularity disorder (arrows and colors depict tensor orientations from 0 to $\pi$). (h) Haralick texture features are computed within local rectangular windows within the tumor region, resulting in a texture value for each pixel, with Haralick correlation shown here (low = blue, high = yellow). (i) Tissue component densities are computed from gland lumen (green), epithelium (blue), nuclei (yellow), and stroma (uncolored)............. 43

3.2  (a) non-standardized intensity distribution (red) vs. standardization template (blue). (b) Standardized intensity distribution (red) vs. standardization template (blue).................................... 45

3.3  (a) A T2w MRI of the prostate (tumor outlined in red). (b) Zoomed-in view of tumor. Sample radiomics feature maps from each class of radiomic features are plotted (blue = low, yellow = high). (c) Standard deviation within 9x9 windows. (d) Entropy
within 5x5 windows. (e) Sobel yx filter. (f) Kirsch 1 filter. (g) Gabor cosine filter at an angle of $\frac{\pi}{8}$ and wavelength 4. (h) Haralick correlation within 5x5 windows. (i) Laws feature 15.

3.4 Receiver operating characteristic (ROC) curves depicting random forest classification performance using radiomic features selected in experiment 1 (a, b), and QH features selected in experiment 3 (c, d) on both the training (a, c) and validation (b, d) cohorts.

3.5 The maximum correlation coefficients for each QH feature group (columns) and each discriminative radiomic feature (rows) selected in experiment 1 are plotted (red = negative correlation, blue = positive correlation), with gray X’s denoting lack of statistical significance.

3.6 Differential expression for a pair of strongly correlated and discriminative radiomic and QH features between tumors with Gleason score 6 (a-d) and Gleason score 9 (e-h). The radiomic feature values are plotted within the tumor (blue = low, yellow = high) shown within the whole prostate (a,e) as well as zoomed-in (b,f). The corresponding QH feature values are plotted within the tumor region on the segmented gland lumen boundaries (blue = low, red = high), shown within the whole tumor (c,g) as well as zoomed-in (d,h). Boxplots summarizing the differential expression of these features over all tumors are plotted (q, r). Note the low number of colors in the tumor with high (f) relative to low (b) Gleason score, indicating lower variance of the depicted Gabor
feature within the high Gleason score tumor. Additionally, note the greater number of colors in the tumor with high (h) relative to low (d) Gleason score, indicating lower kurtosis of the depicted shape feature within the high Gleason score tumor.
I would like to begin by thanking my research supervisor, Satish Viswanath, and my research advisor, Anant Madabhushi, for their constant support and guidance. Their insightful feedback played a crucial role in helping me develop my ability to think critically about designing experiments, interpreting results, and planning next steps in my scientific investigations. Since pushing the boundaries of human knowledge is not always easy, the path leading me to this point contained many obstacles. I am grateful to Satish and Anant for their patience as I made mistakes and learned from them, as well as their impatience and encouragement, which pushed me to work harder and persist longer than I knew I was capable of. I have no doubt that the skills I have learned will not only serve me in my scientific career, but also in my future endeavors in all other domains of life. I feel privileged to have had the opportunity to contribute to what is undoubtedly some of the most important work of our time, and could not have done this without the help of my mentors.

I would also like to thank my colleague and collaborator Andrew Janowczyk for playing a similarly important role on my first project. Andrew’s wit, wisdom, and enthusiasm proved invaluable in helping me achieve success on my first paper as a lead author.

Special thanks are due to my professional and clinical collaborators, who took time out of their busy schedules to provide data, perform annotations, and offer their clinical expertise: Francisco Aguila, Ann Tillet, Andrew Janowczyk, Asha Singanamalli, Mirabela Rusu, Jay Gollamudi, Robin Elliot, Natalie Shih, Michael Feldman, Phillip D. Stricker, Warick Delprado, Sarita Tiwari, Maret Böhm, Anne-Maree Haynes, Lee Ponsky. Their efforts were essential to these research projects.
Additionally, I would like to thank my fellow labmates at the Center for Computational Imaging and Personalized Diagnostics (CCIPD) for providing helpful feedback on my progress reports during our weekly lab meetings, as well as their camaraderie and support: Asha Singanamalli, Mirabela Rusu, Prateek Prasanna, Pallavi Tiwari, George Lee, Patrick Leo, Shoshana Ginsburg, Jacob Antunes, Xiangxue Wang, Lin Li, Jay Patel, Mahdi Orooji, Rakesh Shiradkar, Ahmad Alghary, Yu Zhou, Sahir Ali, Cheng Lu, Jon Whitney, Soumya Ghose, Jhimli Mitra, Niha Beig, Nathaniel Braman, Ania Gawlik, German Corredor-Prada, David Romo-Bucheli, Jun Xu, Ajay Basavanhally.

Many thanks to all my friends at Case Western Reserve University for their camaraderie and support, especially: Benjamin Cowen, Madhukar Kumar, Michael Fishman, Jeremy Kinkade, Jonathan Marc, Sukwoo “Xavier” Yoon, and Andrew Latham.

Last, but not least, I would like to thank my family for their unconditional love and support: Mom, Dad, Josh, Rebecca, my Grandparents, Steve, Jeff, Rich, Louise, Karen, and Kiki.
Abstract

Identifying the Histomorphometric Basis of Predictive Radiomic Markers for Characterization of Prostate Cancer

by

GREGORY PENZIAS

Radiomics has shown promise for *in vivo* prediction of cancer risk, thus providing a potential avenue for reducing over-treatment and unnecessarily invasive biopsy-based diagnosis. Radiomics could be particularly beneficial for stratifying patients into different risk groups in the context of prostate cancer (PCa), for which limitations of current *in vivo* risk assessment result in over-diagnosis and over-treatment. Despite its promise, successful translation of radiomics into the clinic may require a more comprehensive understanding of the underlying morphologic tissue characteristics they reflect. Few studies, however, have attempted to establish the biological or histomorphometric basis for the performance of radiomics. Accomplishing this requires fusing the information obtained from the imaging modalities of radiology and histopathology, since the gold standard definition of PCa comes from histopathologic analysis of whole-mount specimens.

The first step in performing this radiology-pathology fusion in PCa entails achieving spatial correspondence between preoperative *in vivo* magnetic resonance imaging
(MRI) and *ex vivo* hematoxylin & eosin (H&E)-stained whole-mount radical prostatectomy specimens via deformable co-registration. Co-registration, however, requires whole-mount histology sections (WMHSs), which are not always feasible to obtain. In such cases, large specimens are cut into multiple smaller tissue fragments. This thesis presents work on two related modules of radiology-pathology fusion in PCa: First, a novel automated program called AutoStitcher, which reconstructs pseudo whole-mount histology sections (PWMHSs) by digitally stitching together multiple smaller tissue fragments, thus enabling co-registration with *in vivo* radiographic imagery. AutoStitcher reconstructed PWMHSs with less than 3% error relative to manually stitched PWMHSs. Second, comprehensive sets of radiomic features extracted from MRI and quantitative histomorphometric features from H&E were extracted and then spatially co-localized to characterize each tumor region. Correlative analysis revealed a set of promising predictive radiomic markers that could accurately distinguish low-from intermediate-/high-risk PCa and a set of QH features that may form their histomorphometric basis. Results were validated on an independent dataset from a different institution.
1 Introduction

1.1 Problem Background

Histopathological analysis via biopsy and post-surgery analysis provide the current clinical gold standard for diagnosing and characterizing prostate cancer (PCa). However, such procedures are highly invasive and can have significant undesirable side effects and complications. Furthermore, since biopsies are prone to spatial sampling errors, they may not accurately reflect the true aggressiveness of a patient’s disease.\(^1\) Once diagnosed positive on a biopsy, PCa patients often receive a radical prostatectomy to surgically remove the entire prostate. Although PCa is one of the most common cancers among men in the United States, with 1 in 7 being diagnosed with the disease, only 1 in 39 die of it.\(^2\) Thus, many patients undergo unnecessarily invasive diagnosis and therapy.

To combat these problems, there has been considerable recent interest in using radiographic imaging modalities such as magnetic resonance imaging (MRI) for identification and characterization of PCa. Radiography, in contrast to histopathology is non-invasive, thus poses little risk to the patient. Additionally, radiography can be used to analyze the entire tumor volume, thus mitigating sampling errors. The recent introduction of the Prostate Imaging Reporting and Data System (PIRADS)
presented a set of standardized guidelines to aid clinicians in assessing disease risk on in vivo multi-parametric MRI, however, it can be subject to inter-observer variability.\textsuperscript{3} This variability could potentially be mitigated by leveraging recently explored methods of high-throughput computerized analysis of radiographic images, termed radiomics. Radiomics leverages advanced computer vision and machine learning techniques to identify subtle and complex radiographic imaging features for predicting disease using features that may not be visually discernible, thus may provide a useful adjunct to expert radiologist assessment for predicting the risk of disease metastasis in vivo.\textsuperscript{4–6}

1.2 Scope of this Work

In order to develop and validate radiomic predictors, significant challenges must be overcome. Despite the promising potential of radiomics, few studies have attempted to establish the biological or histomorphometric basis for the performance of these features in PCa.\textsuperscript{7–10} Since interpretability of features may be important for translation to the clinic, there is therefore a clear need for further such studies. Elucidating the histomorphometric basis of requires fusing the information obtained from the imaging modalities of radiology and histopathology, since the gold standard definition of PCa comes from histopathologic analysis of whole-mount specimens. The first main step in performing this radiology-pathology fusion entails achieving spatial correspondence between the multimodal images via deformable co-registration. Once spatial correspondence is established, a careful correlative analysis can then be
Introduction

performed to explore the morphological basis for the discriminability of the identified radiomic features. This morphological basis can then be explained in context of existing knowledge of the biological and morphological manifestation of the disease, providing a useful foundation for clinical adoption. Thus, radiology-pathology fusion plays an essential role in the validation of these radiomic predictors of disease.

For PCa patients who opt for radical prostatectomy, the entire prostate gland is surgically excised. Once removed, the prostate specimen is sliced into sections and then stained with hematoxylin and eosin (H&E) for microscopic analysis. This presents the opportunity for post-surgery histopathologic analysis of the prostate specimen to obtain (a) detailed annotation of disease extent, and (b) assessment of disease risk via Gleason grading. Prior to mapping these disease annotations onto in vivo imaging, it is necessary to first obtain whole-mount histology sections (WMHSs) that correspond to the same spatial plane as the imaging sections. However, since preparation of WMHSs is not always feasible in routine clinical practice, many centers have opted to cut large specimens into multiple smaller tissue fragments. These smaller tissue fragments must then be digitally reconstructed into pseudo whole mount histological sections (PWMHSs) prior to being mapped onto their corresponding imaging sections. Since manual reconstruction of PWMHSs can be laborious and time-consuming, as well as prone to inter- and intra-user variability, an automated method is desirable.

This work thus presents (1) the development of an automated method for performing reconstruction of PWMHSs, and (2) a framework for co-registration, feature extraction, and correlative analysis to identify the morphologic basis for radiomic features in discriminating low- from intermediate- and high-risk PCa.
1.3 Thesis Overview

This thesis presents work on a novel automated method for performing reconstruction of PWMHSs based on the published in Penzias et. al. (2016), and radiology-pathology fusion in the context of PCa. The remainder of this thesis is organized as follows:

Chapter 2: Automatic Reconstruction of Digitized Whole Histological Sections from Tissue Fragments presents an automated algorithm called “AutoStitcher” for efficiently and robustly reconstructing PWMHSs, with extensive evaluation in the context of prostate cancer.

Chapter 3: Identifying the histomorphometric basis for radiomic features in distinguishing different Gleason scores of prostate cancer on MRI presents a correlative analysis of radiomic features with quantitative histomorphometric (QH) features that aims to improve the interpretability of radiomic features of prostate cancer.

Chapter 4: Conclusions and Future Directions concludes the thesis by summarizing the most important findings of our work and proposing potential avenues for future investigation.
2 AutoStitcher: An Automated Program for Reconstructing Digitized Histology

2.1 Introduction

Whole-mount histological sections (WMHSs) allow for visual and spatial co-registration with pre-operative in vivo imaging. This approach can thus allow for spatially mapping disease extent annotated on the ex vivo pathology images onto the in vivo imaging\textsuperscript{11}. However, preparation of whole-mounts is not always feasible in routine clinical practice, as large histological specimens can be difficult to slice thin enough to obtain sections without compromising tissue integrity. In addition to requiring significant amounts of technical expertise and preparation time as well as specialized equipment, whole-mount specimens can be too large to fit on a standard glass microscopy slide. As a result, several clinical centers have adopted the solution of cutting large specimens into multiple smaller fragments, resulting in their examining or annotating multiple slides per section\textsuperscript{12,13}.

Unfortunately, utilizing tissue fragments spread across multiple slides presents a significant challenge when the pathologic region-of-interest crosses the boundary of the fragments. An example of this is when in vivo radiological imagery is cognitively
evaluated against corresponding ex vivo pathology, it is considerably more demanding to correlate fragmented pathology with whole section in vivo MR images. In addition to the visual differences between the two modalities, having to switch between multiple slides to cognitively combine disjoint visual cues make it difficult to spatially localize tissue fragments directly on imaging. Similarly, when co-registering ex vivo pathology and in vivo imaging\textsuperscript{14}, whole-mount sections are required on both modalities; which are easier to spatially correlate to one another than tissue fragments. Under these circumstances, reconstructing pseudo whole-mount histological sections (PWMHSs) from component tissue fragments fulfills a clear need in pathology annotation and radiology-pathology correlation workflows.

Chappelow et. al.\textsuperscript{12} presented a semi-interactive computerized tool called HistoStitcher which enabled PWMHS reconstruction from component digitized histological image fragments\textsuperscript{12,13}. The tool takes as input several user-selected fiducials along the edges of a pair of fragments. A transform to “stitch” the edges together is then computed based on calculating an affine transform (encoding rotation, translation, and scaling) of the image fragments such that the fiducials are brought into spatial alignment. HistoStitcher was shown to be easier to use and more memory-efficient than photo-editing tools such as Photoshop. However, a significant limitation with HistoStitcher lies in its requirement of manual identification of corresponding fiducial points prior to reconstruction. In addition to being time-consuming, this makes HistoStitcher subject to inter- and intra-user variability. In this paper, we present AutoStitcher, an fully automated algorithm for reconstructing whole-mount histology images from image fragments which is able to overcome most current limitations associated with HistoStitcher.
A clear parallel to solving the problem of PWMHS reconstruction may lie in the variety of automated approaches that have been developed for jigsaw puzzle assembly, panoramic photo stitching, and shredded document reconstruction. However, these approaches cannot be directly translated over to our specific problem domain of histological image reconstruction as they make at least one of the following assumptions about the input images:

(a) Overlap: Most photo stitching algorithms require using images that depict overlapping regions of a common scene, as shown in Figure 2.1a, as they operate by aligning matching key-points within these regions\textsuperscript{15–17}. Only Poleg and Peleg\textsuperscript{18} have previously presented an algorithm for stitching non-overlapping images, however, it is specifically designed for rectangular photographic images that have no missing pieces.

(b) Completeness: Most automatic jigsaw puzzle solving algorithms assume that all available puzzle fragment images together depict the entire jigsaw image, implying that none of the pieces are missing (illustrated in Figure 2.1c\textsuperscript{19,20}). Liu et. al.\textsuperscript{21} have presented an algorithm for assembling hand-shredded photos that does account for missing shreds, however, it relies on matching the boundary contours of adjacent pieces.

(c) Interlock: As seen in archetypal jigsaw shapes which fit into one another (illustrated by Figure 2.1e); this factor is typically accounted for as contours that can be matched based on their curvature. Most automatic jigsaw puzzle solving algorithms make the assumption that the jigsaw fragments “interlock”\textsuperscript{19,22}. Shredded document reconstruction algorithms focus on reconstructing documents that have been processed by traditional strip- or
cross-cut paper shredders\textsuperscript{23}, and thus tend to rely on curve matching to identify corresponding shredded fragments\textsuperscript{24}.

As depicted in Figure 2.1, all of the above assumptions are violated when reconstructing whole-mount histological images. Typically, histological fragments (a) are non-overlapping, since they are cut from a single object; (b) lack completeness as significant tissue loss can occur during pathological processing; and (c) do not interlock, since uneven warping and tissue loss during processing renders boundary contours too dissimilar to match reliably. Further, tissue processing of whole-mount prostates involves quartering and sectioning of a tissue block. This could potentially lead to variable slice depths and orientations between the tissue fragments that are supposed to lie on the same plane\textsuperscript{25}. This further exacerbates the lack of interlock, completeness, and overlap between these tissue fragments.

AutoStitcher utilizes a novel cost function that does not make any of the three limiting assumptions of (a) overlapping, (b) completeness, and (c) interlocking images; as previously presented in image reconstruction literature. Our cost function is inspired by two features humans use to perform manual stitching: (i) alignment of similar tissue regions and (ii) contiguity of the prostate boundary. Tissue region similarity is quantified by grayscale intensity histograms, and contiguity of the prostate boundary is quantified by distance between automatically detected points along the boundary of each fragment. AutoStitcher achieves computational efficiency by working hierarchically, utilizing approximate reconstructions performed at lower resolutions to reduce the amount of computation necessary at higher resolutions.

To evaluate AutoStitcher’s performance, 113 sections were stitched both automatically via AutoStitcher and manually via HistoStitcher; which were then qualitatively
and quantitatively compared. Quantitative comparison was facilitated via automatically and manually selected fiducial points, as well as Hausdorff distance between reconstruction contours. To investigate whether AutoStitcher could work on an independent cohort from a different institution, parameters were learned on a sub-cohort, and tested on the remaining sections from both institutions.

The remainder of the paper is organized as follows. In the next section, we describe the methodological details of AutoStitcher. Then, we explain our experimental design
to evaluate AutoStitcher on histological data from 2 different institutions. Finally, we present and discuss our experimental results, and end with concluding remarks in the last section.

2.2 Methods

Ethics Statement

Data analysis was waived review and consent by the IRB board, as all data was being analyzed retrospectively, after de-identification. All experimental protocols were approved under the IRB protocol # 02-13-42C with the University Hospitals of Cleveland Institutional Review Board, and all experiments were carried out in accordance with approved guidelines.

Notation

Notation employed in this paper has been summarized in Table 1. We denote an image $Q$ where each image has dimensions $[X, Y]$, and each pixel in the image has coordinates $(x, y) = \{(x, y) | x \in [1, X] \text{ and } y \in [1, Y]\}$.

Preprocessing and Initialization

The only input to AutoStitcher is a set of 4 tissue fragments comprising a single 2D section (selected by a user), which are then pre-processed as follows (see Figure 2.2):

(i) Down-sampling the unprocessed high-resolution tissue fragment images to lower resolution for computational efficiency.
Table 2.1. Notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q = {Q^{ul}, Q^{ur}, Q^{ll}, Q^{lr}}$</td>
<td>The set of quadrant images corresponding to the upper-left, upper-right, lower-left, and lower-right quadrants, respectively.</td>
</tr>
<tr>
<td>$T = {T^{ul}, T^{ur}, T^{ll}, T^{lr}}$</td>
<td>The set of transformations of all 4 quadrant images, where each quadrant’s transformation stores the x-translation, y-translation, and degrees of rotation.</td>
</tr>
<tr>
<td>$D_{pair}(Q^1, Q^2)$</td>
<td>Pair-wise intensity-based dissimilarity computed on each quadrant pair $(Q^1, Q^2)$.</td>
</tr>
<tr>
<td>$M_{pair}(Q^1, Q^2)$</td>
<td>Pair-wise distance-based misalignment computed on each quadrant pair $(Q^1, Q^2)$.</td>
</tr>
<tr>
<td>${V^1(k), V^2(k)}$</td>
<td>The $k^{th}$ set of corresponding histogram vectors or pixel values $V^1$ and $V^2$ of a pair of quadrants, where $V^1$ belongs to $Q^1$ and $V^2$ belongs to $Q^2$.</td>
</tr>
<tr>
<td>${C^1_{outer}, C^2_{outer}}, {C^1_{inner}, C^2_{inner}}$</td>
<td>The set of corresponding outer and inner corner points $C^1$ and $C^2$ of a pair of quadrants, where $C^1$ belongs to $Q^1$ and $C^2$ belongs to $Q^2$.</td>
</tr>
<tr>
<td>$w_1, w_2$</td>
<td>Empirically determined cost function component weights $(w_1, w_2)$</td>
</tr>
<tr>
<td>$w_3, w_4$</td>
<td>Inner- and outer-point weights $(w_3, w_4)$</td>
</tr>
<tr>
<td>$m, n$</td>
<td>$m =$ Number of overhanging, and $n =$ number of corresponding pixels for a pair of adjacent edges</td>
</tr>
<tr>
<td>$p, b$</td>
<td>$p =$ Size of rectangular patch, and $b =$ number of histogram bins for cost computation</td>
</tr>
<tr>
<td>$\phi$</td>
<td>Dissimilarity for a non-corresponding edge pixel</td>
</tr>
</tbody>
</table>

(ii) Applying user-specified flipping to correct for human errors in microscopy slide digitization, where tissue fragments from the same sectioned plane may be scanned on different sides of the original sectioning plane. After flipping, the images are converted to grayscale (depicted in Figure 2.2b) to improve computational efficiency. The grayscale pixel values are rescaled from the conventional 8-bit intensity range $[0,255]$ to decimals within the range $[0,1]$.

(iii) Segmenting the tissue foreground mask from background, as shown in Figure 2.2c, to identify the fragment boundaries.
An approximate initial reconstruction is first computed at low resolution, which is used as an initialization for all subsequent algorithmic processing. Initialization is performed by:

(i) Computing the minimum-area bounding box, as highlighted in blue in Figure 2.2d, and identifying the edges (cyan and green lines in Figure 2.2e) as pixels on the tissue boundary contour that fall between “corner” points (green stars in Figure 2.2d) closest to the three relevant corners of the bounding box. The fourth corner, highlighted in red, is external to the prostate, so is not used to define an edge.

(ii) Computing best-fit lines to the edges using Theil-Sen linear regression\(^{26}\), which achieves robustness to outliers by choosing the median slope of all possible lines through sample points, and rotating the quadrant such that the horizontal best-fit line, as plotted in green in Figure 2.2f, is parallel to the x-axis. Figure 2.2g displays the initially non-rotated quadrants, which is followed by the rotated quadrants in Figure 2.2h.

(iii) Finally, the rotated fragments are translated together such that the boundaries of the masks of adjacent quadrants are joined, shown in Figure 2.2i.

**Domain-Inspired Cost Function**

AutoStitcher uses a two-component, domain-inspired cost function to drive stitching, based on quantifying the dissimilarity of adjacent quadrant regions across the quadrant stitch boundaries, as illustrated in Figures 2.3b-g, while the misalignment component quantifies how well the stitched result maintains the continuity of the prostate shape (as illustrated in Figure 2.3a).
Figure 2.2. Illustration of AutoStitcher’s preprocessing and low resolution initialization workflow. (a) Initial unprocessed quadrant image. (b) Resizing, flipping, and conversion to grayscale. (c) Segmentation of tissue mask. (d) Bounding-box fitting. (e) Identification of edges. (f) Identification of best-fit lines, followed by automatic rotation. (g) Initial unprocessed quadrant images for a single slice. (h) Automatically rotated and flipped quadrant images. (i) Low resolution initialization.

Dissimilarity

Dissimilarity is defined as the degree to which corresponding tissue regions of adjacent quadrants are dissimilar. This is quantified via the following steps:

(i) Identification of corresponding edge-pixels: Corresponding edge pixels are pairs of pixels on adjacent sections that share a common edge, as illustrated by the two pairs of corresponding patches in Figure 2.3b,e. They are defined as follows:
$(x_1, y_1) \in Q^1$ and $(x_2, y_2) \in Q^2$, where:

\[ x_1 = x_2 \text{ if } (Q^1, Q^2) \in \{(Q^u, Q^d), (Q^u, Q^r)\} \]

\[ y_1 = y_2 \text{ if } (Q^1, Q^2) \in \{(Q^l, Q^u), (Q^l, Q^r)\} \]

Pixels are defined to be non-corresponding if they do not have any corresponding pixel on the adjacent quadrant.

(ii) Patch extraction and histogram (or intensity value) computation: For each corresponding edge-pixel identified in step (i) for any pair of quadrants $Q_1$ and $Q_2$, square patches of size $p$ are extracted. Intensity histograms with $b$ bins are computed from the patches, excluding non-tissue background pixels, which are then normalized by computing the discrete probability density function of the intensity bins. Corresponding histograms (or pixel intensity values) $V^1(k)$ and $V^2(k)$ are then computed for all $k \in \{1, \ldots, n\}$, where $n$ is the number of corresponding pixels or patches, such that:

$V^1(k)$ is the $k^{th}$ pixel value or histogram vector centered on $(x_1, y_1)$ in $Q^1$, and $V^2(k)$ is the $k^{th}$ pixel value or histogram vector centered on $(x_2, y_2)$ in $Q^2$, such that:

$V^{1/2}(k) = \{V | V = \langle V_1, V_2, \ldots, V_b \rangle, V_b \in [0, 1], |V|_1 = 1\}$, if $V$ is a histogram

$V^{1/2}(k) = \{V | V = V \in [0, 1]\}$, if $V$ is a pixel
Since the patch size is kept at a constant real-world size, patches encompass only a single pixel at the lowest resolution. Therefore, at the lowest resolution, pixel intensity values are used instead of histograms.

(iii) Dissimilarity computation: While patches of tissue on opposite sides of a cut can appear to be visually different, corresponding patches on either side of the cut can be expected to have more similar intensity distributions compared to image patches that do not correspond. For example, given a cut through a gland-dense region, patches on either side of the cut are likely to have intensity distributions skewed towards higher frequencies of high intensities, since gland lumen are mostly white or very light-colored. This would be reflected in the similarity in their intensity distributions. By contrast, distributions of patches in gland-dense and gland-sparse regions would be very different. Dissimilarity between pairs of histograms (or pixels) is computed using the $L^2$-norm of the histogram vector (or intensity value) differences. Given that pixel values are in [0,1] and histogram vectors are discrete probability density functions that sum to 1, the maximum possible dissimilarity, denoted $\phi$, between histograms is equal to 2 and between pixels is equal to 1. Non-corresponding pixels are treated consistently by setting their dissimilarity to $\phi$. The pair-dissimilarity $D_{\text{pair}}$ for the pair of adjacent images is thus defined as:

$$D_{\text{pair}}(Q^1, Q^2) = \frac{1}{n + m} \left( m \times \phi + \sum_{k=1}^{n} \|V^1(k) - V^2(k)\|_2^2 \right)$$

(2.1)

where $n$ is the number of corresponding pixels, and $m$ is the number of non-corresponding pixels.
Misalignment

Misalignment is defined as the degree to which adjacent quadrants are incorrectly localized and oriented relative to one another. To compute misalignment, the endpoints of the best-fit lines of the first and second quadrants in the quadrant-pair are identified, as shown in Figure 2.3a. The pair of corner points nearest to the outer-boundary of the prostate is denoted \((C_{\text{outer}}^1, C_{\text{outer}}^2)\), while the point pair closest to the center of the prostate is denoted \((C_{\text{inner}}^1, C_{\text{inner}}^2)\). Misalignment \(M_{\text{pair}}\) is computed as:

\[
M_{\text{pair}}(Q^1, Q^2) = w_3 \times \|C_{\text{outer}}^1 - C_{\text{outer}}^2\|^2 + w_4 \times \|C_{\text{inner}}^1 - C_{\text{inner}}^2\|^2 \tag{2.2}
\]

where \(w_3\) and \(w_4\) \((w_4 = 1 - w_3)\) weight the relative contributions of the two pairs of points.

Total computation for all pairs of quadrants

AutoStitcher evaluates dissimilarity and misalignment on each of the four pairs of vertically or horizontally adjacent quadrant images \((Q^{ul}, Q^{ur}), (Q^{ur}, Q^{lr}), (Q^{lr}, Q^{ll}), (Q^{ll}, Q^{ul})\); as follows:

\[
D_{\text{tot}} = \frac{1}{4} \times [D_{\text{pair}}(Q^{ul}, Q^{ur}) + D_{\text{pair}}(Q^{ur}, Q^{lr}) + D_{\text{pair}}(Q^{lr}, Q^{ll}) + D_{\text{pair}}(Q^{ll}, Q^{ul})] \tag{2.3}
\]
Cost Function

The optimal reconstruction is found by determining the set of quadrant transformations $T$ of quadrants $Q$ that minimizes the cost function below. This cost function involves combining the dissimilarity $D_{tot}$ and misalignment $M_{tot}$ in a weighted sum (equation 5), where $w_2 = 1 - w_1$. The set of quadrant transformations $T$ specifies the rigid-body transformations of three moving quadrants relative to the fourth fixed quadrant. Each quadrant’s rigid-body transformation is parameterized by three degrees of freedom (two for translation, one for rotation).

\[
\text{argmin}_T \cdot (T, Q) = w_1 \times D_{tot} + w_2 \times M_{tot}
\]  

Optimization of equation (5) is done via genetic algorithms for each resolution in the hierarchy. Genetic algorithms were utilized as they are well-suited to such problem domains, that are highly nonlinear and have many local optima\textsuperscript{27}. In our implementation, each “generation” of the genetic algorithm comprises a set of 20 solutions, where each solution is a nine-element vector specifying a combination of transformations of three quadrants relative to the fourth quadrant. Optimization concludes after twenty-five consecutive generations that have not demonstrated a significant improvement in the cost function. To ensure computational efficiency, the

\[
M_{tot} = \sqrt{\frac{1}{4} \times [M_{pair}(Q^{ul}, Q^{ur}) + M_{pair}(Q^{ur}, Q^{lr}) + M_{pair}(Q^{lr}, Q^{ll}) + M_{pair}(Q^{ll}, Q^{ul})]}
\]  

(2.4)
maximum and minimum possible translations and rotations are restricted to within a local search window of the next lower image resolution.

2.3 Experimental Design

Data Collection and Processing

Data from prostate cancer patients who underwent radical prostatectomy were acquired from two institutions: (1) 19 patients from University of Pennsylvania, and (2) 17 patients from St. Vincent’s Hospital. Each surgically resected prostate gland
was fixed in formalin, embedded in paraffin, and sectioned axially in a plane perpendicular to the urethral axis from apex to base in 3-4 mm sections. Each slice was then sectioned into four quadrants (see Figure 2.2g for sample data), stained with hematoxylin & eosin (H+E), and digitized via an Aperio® whole slide scanner at 20X magnification and 0.5 /pixel resolution. Digitized slides were de-identified and labeled with the anatomic location of the slide (left- or right-anterior or posterior of the prostate gland).

Our final curated dataset contains a total of 113 sections (452 quadrants) from 36 patients. Note that 143 (out of a total of 256) sections were excluded from our cohort due to (a) large quantities of missing or extra-prostatic tissue (100 out of 143), (b) not having been sectioned into quadrants (e.g., left-to-right-sliced apex or base) (43 out of 143).

All 113 sections were reconstructed into PWMHS via each of AutoStitcher and HistoStitcher. AutoStitcher parameters are described in parameter selection in the next subsection. Manual reconstruction via HistoStitcher was performed by a user with 2 years of previous experience in utilizing the software.

**Evaluation**

A total of 113 sections were utilized for evaluation, of which 20 were used for parameter selection, and 44 from University of Pennsylvania plus 49 from St. Vincent’s Hospital were used for independent testing.

As there is no “perfect” ground truth whole-mount section for the clinical data utilized in this study, we quantitatively evaluated the accuracy of reconstruction for AutoStitcher with respect to HistoStitcher using three different error measures.
Two of the measures are based on fiducial points, which provide an intuitive error measure as well as being computationally efficient for tracking image transformations. Fiducial error is also commonly used in similar applications where automated methods are evaluated with respect to manual gold-standard results, such as medical image registration\textsuperscript{28–30}.

Note that in addition to computing these measures in micrometers (\(\mu m\)), we additionally normalized these evaluation measures by the average of length and width of the HistoStitcher reconstruction, to yield a percentage error. In such a case, an error of 100\% would result from an AutoStitcher reconstruction containing fiducial points extremely far from where they were positioned on the corresponding HistoStitcher reconstruction. This stitched result would also appear to be overwhelmingly and obviously inaccurate. Similarly, an error of 0\% would result from an AutoStitcher reconstruction that is completely identical to the corresponding HistoStitcher reconstruction. It should be noted that an error of 0\% does not necessarily represent a perfect result, but simply that the AutoStitcher and HistoStitcher yielded an identical PWMHS reconstruction. Error measures utilized in this study were:

(i) Automatically selected fiducials (ASF) error: A total of ten pairs of corresponding fiducial points are automatically identified on the endpoints and midpoints of the edges of the HistoStitcher reconstruction. Although there are a total of twenty fiducial points (ten pairs), only ten are visible since there is zero distance between each pair as shown in the green, red, and yellow stars in Figure 2.4b. These points are then mapped onto the AutoStitcher reconstruction, revealing twenty visible points as shown in Figure 2.4a. The ASF
error is computed for each AutoStitcher reconstruction as the mean distance between all 10 pairs of points (in $\mu m$).

(ii) Manually selected fiducials (MSF) error: A total of nine or more pairs of fiducial points were identified by an expert on each HistoStitcher reconstruction. These were selected based on visually identified corresponding landmarks or regions of apparent similarity on the reconstructed PWMHS. Utilizing MSF points thus provides an error measure complementary to the ASF by providing a domain knowledge-based measure, which more accurately reflects an expert’s judgment of the reconstruction quality than the automatically identified ASF points. These points are mapped onto the AutoStitcher reconstruction as depicted in Figure 2.5a. The MSF error is computed as the mean distance between all identified point pairs, for each AutoStitcher reconstruction.

(iii) Hausdorff distance (H): A stitched result that maintains the expected outline shape is particularly desirable because this information forms a crucial reference to guide manual selection of anatomic fiducials during image registration. We measure the similarity between shapes of the AutoStitcher reconstruction and HistoStitcher reconstruction outline contours via the Hausdorff distance$^{31}$. The steps involved in the computation are as follows:

1. Binary tissue masks of both the manually and automatically stitched images are obtained from the segmentations performed during pre-processing.
2. The convex hull is computed for each of these masks, as shown in Figure 2.6a,b, in order to eliminate concavities in the reconstruction outlines.
Note that using the original reconstruction outlines would produce large Hausdorff distances despite visually similar reconstructions.

(3) The convex hulls are mapped to a common coordinate system (Figure 2.6c).

(4) The convex hulls are rotated and translated such that they are maximally aligned, such that the Hausdorff distance between them is minimized (Figure 2.6d).

(5) The Hausdorff distance is calculated as the distance between points at which the two aligned convex hulls are farthest apart. An ideal Hausdorff distance of zero would occur when the outlines are identical, and thus, the more dissimilar the outlines, the larger the Hausdorff distance.

Parameter Selection

Optimal parameter values for AutoStitcher were experimentally determined on a training subset of data consisting of 20 sections from 9 patients from the University of Pennsylvania (see Table 2). Based on the likely ranges for each parameter, optimal parameters were identified as those that produced the minimum error of ASF on the training set, over a total of 72 experiments comprising every possible combination of the parameters $w_1, w_2, w_3, w_4,$ and $p$. Parameter $b$ was determined empirically and fixed prior to selection of the other parameters in order to restrict the number of degrees of freedom of the parameter space and make the problem computationally feasible. In synthetic testing (not shown), $b$ was not found to significantly affect the reconstruction accuracy. The value for this parameter was picked such that different-looking patches were distinguishable via their histograms while ensuring that the histograms were not sparse.
Table 2.2. Parameter Selection

<table>
<thead>
<tr>
<th></th>
<th>$w_1$</th>
<th>$w_2 = 1 - w_1$</th>
<th>$w_3$</th>
<th>$w_4 = 1 - w_3$</th>
<th>$p$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Optimal Value</strong></td>
<td>0.989</td>
<td>0.011</td>
<td>0.4</td>
<td>0.6</td>
<td>81x1</td>
<td>16</td>
</tr>
<tr>
<td><strong>Range</strong></td>
<td>0.98-0.995</td>
<td>0.005-0.02</td>
<td>0.3-0.5</td>
<td>0.5-0.7</td>
<td>41x41</td>
<td>-6-20</td>
</tr>
</tbody>
</table>

Multi-site evaluation

Learned parameters from the “training” subset of 20 sections were kept fixed when stitching the remaining 93 sections. Inter-institutional variability of AutoStitcher could then be evaluated by comparing ASF error, MSF error, and Hausdorff distance in terms of (a) training error on 20 sections (9 patients) from UPenn, (b) testing error on 44 sections (17 patients) from UPenn, and (c) validation error on 49 sections (17 patients) from St. Vincent’s Hospital. Note that data used to compute testing and validation error was not used in any way to optimize the algorithmic parameters.

Equivalence testing\(^{32}\) was utilized to compare each of the three evaluation measures described in the previous subsection over all reconstructions, with the null hypothesis being that evaluation measures were not statistically significantly equivalent between the two institutions. To perform this test, the error distributions were confirmed to be normal using the Kolmogorov-Smirnov test. The mean and standard deviation of the errors were computed, and then a 90% confidence interval relating to the difference between the means of the errors on the testing and validation cohorts was computed. The equivalence margin $\delta$ was set to 1%, which was estimated to be the smallest difference in mean errors between the two institutions that would render them meaningfully different. If the 90% confidence interval fell within the equivalence margin bounds $[-\delta, \delta]$, the error measure was considered to statistically equivalent between the testing and validation cohort at a significance level of $\alpha =0.05$. 
Speed Comparison

We estimated the required time of stitching for novice and expert users by surveying three users of HistoStitcher for the typical length of time required to perform manual reconstruction of a PWMHS. Additionally, we measured the length of time required for AutoStitcher to perform automated reconstruction of each PWMHS. Based on these times, we estimated an approximate range of times required for each of (a) the novice users and (b) expert users to perform stitching using (i) AutoStitcher and (ii) HistoStitcher, respectively.

2.4 Results and Discussion

Experiment 1: Qualitative and Quantitative Evaluation of Reconstruction Accuracy in terms of fiducial error

Figures 2.4 and 2.5 illustrate a high degree of apparent visual similarity between the AutoStitcher and HistoStitcher reconstructions (depicted for 2 different sections from 2 different patients). This qualitative similarity is supported by ASF and MSF errors of under 3% (see Table 3), over all 113 sections. In the absence of a true “gold standard” PWMHS reconstruction for the real world clinical data used by us in this study, a fiducial-based error of less than 3% indicates that the AutoStitcher reconstruction is highly similar to the HistoStitcher result even when considering tissue regions that are in correspondence across a stitched edge. It should be noted that our HistoStitcher reconstructions were performed by only a single operator. This
was a study limitation and in future work we will look to perform multi-user studies to evaluate the extent of inter-user variability in the use of the AutoStitcher tool.

Figure 2.4. Final reconstruction results of (a) AutoStitcher vs. (b) HistoStitcher, with ASF evaluation fiducials plotted (via red, green, blue, and yellow asterisks). The enlarged boxed regions in (c,d) display sample ASF fiducials for each result. Note that since fiducials were selected on the HistoStitcher reconstruction, the panel in (d) shows both points being superposed onto the same yellow asterisk, compared to distinct points when mapped onto the AutoStitcher reconstruction in (c). Normalized ASF error was computed to be 2.61% in this example.

Visual inspection of some of the less similar-appearing reconstructions reveals common patterns among sections where AutoStitcher’s performance deviates from that of HistoStitcher (illustrated in Figure 2.7). These include:
Figure 2.5. Final reconstruction results of (a) AutoStitcher vs. (b) HistoStitcher, with MSF evaluation over the fiducials plotted (in red, green, blue, and yellow asterisks). The enlarged boxed regions in (c) and (d) display sample MSF fiducials for the AutoStitcher and HistoStitcher reconstructions, respectively. Normalized MSF error was computed to be 1.89% in this example.

(i) Misalignment, which occurs when regions on adjacent quadrants with similar appearance do not correctly line-up in the reconstructed image. This is depicted in Figure 2.7a, in which the lower-left quadrant is misaligned relative to the lower-right and upper-left quadrants. The correctly aligned quadrants are shown in the corresponding HistoStitcher reconstruction in Figure 2.7b, based on the operator correctly identifying where the edges should line up.
between quadrants. Note that misalignment could also be caused by variation in the depth of tissue sectioning and the orientations of quadrants within a section\textsuperscript{25}. This is because regions on adjacent quadrants that come from significantly different depths are less likely to have a similar appearance.

(ii) Over-compensation for missing tissue gaps which occurs when there are large portions of tissue missing from the fragments, and AutoStitcher stitches the quadrants together too tightly. In Figure 2.7a, the result of over-compensation by the algorithm is apparent in the marked closeness of the lower-left and lower-right quadrants, which contrasts with the significant gap left between these two quadrants in the HistoStitcher reconstruction shown in Figure 2.7b. While missing tissue is accounted for in HistoStitcher reconstruction by the operator visualizing how the gaps would appear on a PWMHS, this remains one of the main sources of errors in the case of AutoStitcher.

(iii) Lack of scaling of quadrants by AutoStitcher, as this has not been implemented in the current version of the algorithm. This is evident in the differences in sizes of the upper-right and lower-right quadrants in the AutoStitcher reconstruction shown in Figure 2.7c and the HistoStitcher reconstruction shown in Figure 2.7d.

(iv) Excessive overlap of adjacent quadrants, as both HistoStitcher and AutoStitcher allow for some overlap to ensure optimal alignment of quadrants. Despite the fact that there can theoretically be no overlap between quadrant images, both HistoStitcher and AutoStitcher allow for overlap as they utilize only rigid-body transformations when reconstructing a PWMHS. An example of excessive overlap can be seen in the upper-left and lower-left
quadrants in the AutoStitcher reconstruction, shown in Figure 2.7c. In the HistoStitcher reconstruction shown in Figure 2.7d, there is considerably less overlap between these quadrants, based on the operator manually selecting fiducials that would ensure this.

Figure 2.6. Hausdorff distance evaluation computed on the convex hull of (a,e) a PWMHS reconstructed using AutoStitcher next to (b,f) a PWMHS reconstructed using HistoStitcher. The convex hulls are first transformed to a common coordinate system as shown in (c,g), then aligned such that the Hausdorff distance is minimized in (d,h). The location of the maximum Hausdorff distance is highlighted in the black boxes, resulting in normalized Hausdorff errors of 3.81% (d) and 2.43% (h).

**Experiment 2: Qualitative and Quantitative Evaluation of Reconstruction Accuracy in terms of Hausdorff Distance**

Figure 2.6 illustrates the high degree of visual similarity between the reconstruction outlines of AutoStitcher (blue) compared to HistoStitcher (red). These visualizations are supported by a median normalized Hausdorff Distance of 3%, over all 113 sections.
(see Table 3), indicating that the reconstruction outlines are highly similar between AutoStitcher and HistoStitcher.

Further inspection of the HistoStitcher reconstructions shown in Figure 2.6b, d indicates that they appear slightly wider than the automated reconstructions in Figures 2.6a, c. This may be because the HistoStitcher reconstructions allow for more space near the center of the PWMHS to account for missing tissue. As discussed in section the previous subsection and illustrated in Figure 2.7, misalignment of tissue fragments and differences in the handling of missing tissue between AutoStitcher and HistoStitcher account for the majority of AutoStitcher’s error.

**Experiment 3: Multi-Site Evaluation**

AutoStitcher’s performance was not found to be significantly different between the two institutions for all measures (Table 3). The errors were seen to be marginally higher across all evaluation measures on the validation dataset from St. Vincent’s hospital. As depicted in Figure 2.8 via box-and-whisker plots comparing the training, testing, and validation cohorts, ASF and MSF errors have similar ranges between all 3 cohorts. However, the Hausdorff distance errors have noticeably larger ranges and are slightly higher on average on the testing and validation cohorts, compared to the training cohort. This can be explained by the fact that the Hausdorff measure is computed based on a single pair of points, and uses the distance between points at which the aligned outlines are furthest apart, whereas the ASF and MSF are computed based on averages of several pairs of points.
Figure 2.7. Visual inspection of PWMHSs stitched by (a,c) AutoStitcher and (b,d) HistoStitcher, to reveal common causes of reconstruction error: (i) misalignment of quadrants, (ii) under-compensation for missing tissue gaps, (iii) lack of scaling, and (iv) excessive overlap. AutoStitcher reconstruction in panel (a) has MSF, ASF, and Hausdorff errors of 6.56%, 5.85%, and 5.46% compared to corresponding HistoStitcher reconstruction in Figure 2.7b. AutoStitcher reconstruction in Figure 2.7c has MSF, ASF, and Hausdorff errors of 5.51%, 3.27%, and 5.70% compared to HistoStitcher reconstruction in Figure 2.7d.

Experiment 4: Comparison of required time for AutoStitcher vs. HistoStitcher

As shown in Table 4, a novice user of HistoStitcher can take as little as 40 or as long as 80 minutes to complete a stitch (depending on how many times they repeat the process
Table 2.3. Multi-Site Evaluation

<table>
<thead>
<tr>
<th>Institution</th>
<th>University of Pennsylvania (Training)</th>
<th>University of Pennsylvania (Testing)</th>
<th>St. Vincent’s Hospital (Validation)</th>
<th>Cumulative Institutional Difference 90% Confidence Interval (University of Pennsylvania (test) vs. St. Vincent’s Hospital).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Patients</td>
<td>9</td>
<td>17</td>
<td>17</td>
<td>36</td>
</tr>
<tr>
<td>Number of Sections</td>
<td>20</td>
<td>44</td>
<td>49</td>
<td>113</td>
</tr>
<tr>
<td>Median PWMHS Size (µm) Width x Height</td>
<td>39,382 x 31,216</td>
<td>39,768 x 31,796</td>
<td>41,609 x 34,043</td>
<td>40,750 x 33,080</td>
</tr>
<tr>
<td>Median Absolute Error (ASF)</td>
<td>945.45 µm</td>
<td>974.64 µm</td>
<td>1164.4 µm</td>
<td>1061.6 µm</td>
</tr>
<tr>
<td>Median Normalized Error (ASF)</td>
<td>2.55%</td>
<td>2.75%</td>
<td>3.03%</td>
<td>2.89%</td>
</tr>
<tr>
<td>Median Absolute Error (MSF)</td>
<td>764.41 µm</td>
<td>988.71 µm</td>
<td>1172.3 µm</td>
<td>976.24 µm</td>
</tr>
<tr>
<td>Median Normalized Error (MSF)</td>
<td>2.29%</td>
<td>2.67%</td>
<td>2.75%</td>
<td>2.70%</td>
</tr>
<tr>
<td>Median Absolute Error (H)</td>
<td>1014.3 µm</td>
<td>1153.7 µm</td>
<td>1348.3 µm</td>
<td>1141.3 µm</td>
</tr>
<tr>
<td>Median Normalized Error (H)</td>
<td>2.85%</td>
<td>3.36%</td>
<td>3.44%</td>
<td>3.20%</td>
</tr>
</tbody>
</table>

to get a reasonably appearing reconstruction), while an experienced user can perform a stitch in as little 10-20 minutes since they require fewer repetitions. On the other
Figure 2.8. Box and whisker plots of evaluation results between the training cohort, the testing cohort, and the validation cohort, with boxes marking the median, lower-quartile, and upper-quartile, dashed-lines connecting the extremes; and crosses marking outliers. Boxes are colored green, cyan, and purple to distinguish between the three cohorts.

By comparison, once the set of fragments comprising a single section are identified for AutoStitcher reconstruction, the process of stitching them together is fully...
automated. Therefore, while a stitching time of 4-10 minutes per section for AutoStitcher vs. 10-20 minutes per section for HistoStitcher may seem like a moderate improvement, the user-interaction time for AutoStitcher is negligible relative to that of HistoStitcher. To provide some context, while a pathologist or researcher seeking to reconstruct a dataset consisting of one-hundred sections using HistoStitcher might spend 2 hours per day at 15 minutes per section and thus take a total of 2.5 weeks to stitch the entire dataset, the same dataset could be automatically reconstructed by AutoStitcher in under 12 hours at 7 minutes per section, with a majority of that time being spent on computation.

Table 2.4. Time to stitch a single section using AutoStitcher vs. HistoStitcher

<table>
<thead>
<tr>
<th></th>
<th>Novice</th>
<th>Expert</th>
</tr>
</thead>
<tbody>
<tr>
<td>HistoStitcher</td>
<td>40-80 minutes</td>
<td>10-20 minutes</td>
</tr>
<tr>
<td>AutoStitcher</td>
<td>4-10 minutes</td>
<td>4-10 minutes</td>
</tr>
</tbody>
</table>
3 Identifying the Morphologic Basis for Radiomic Features in Prostate Cancer

3.1 Introduction

Radiomics involves extracting quantitative features from medical images for disease characterization\textsuperscript{4–6}. Specifically radiomic features attempt to capture sub-visual image attributes of the disease which may not be visually discernible. Radiomic approaches have recently demonstrated promise for assessing disease risk \textit{in vivo} in a number of different disease domains including prostate\textsuperscript{33–36}, breast\textsuperscript{37–39}, lung\textsuperscript{40–43}, brain\textsuperscript{44,45}, colorectal\textsuperscript{46,47}, renal cell carcinoma\textsuperscript{48}, and head and neck cancer\textsuperscript{49,50}. Radiomic features can capture a variety of different types of information. For example, Haralick features capture textural appearance based on statistical relationships between gray levels of neighboring pixels\textsuperscript{51}; Gabor\textsuperscript{52}, Sobel, Kirsch, Gauss, and gradient features capture edge orientation information; Hess features capture second order derivatives; and standard deviation can capture heterogeneity.

Radiomics could be particularly beneficial for stratifying patients into different risk groups in the context of prostate cancer (PCa), for which limitations of current \textit{in vivo} risk assessment result in over-diagnosis and over-treatment.\textsuperscript{53} Viswanath et
al found that Gabor and Haralick features extracted from 3 Tesla (T) endorectal, in vivo T2-weighted (T2w) magnetic resonance imaging (MRI) were able to accurately detect prostate cancer (PCa).\textsuperscript{54} Ginsberg et al. also found Gabor and Haralick features to be useful for detecting PCa on T2w MRI.\textsuperscript{55} Litjens et al. found that Hess, Gabor, and Gaussian derivative features extracted from T2w MRI were able to discriminate different benign confounders from PCa.\textsuperscript{56} Fehr et al. found that a combination of T2w and ADC MRI Haralick texture features were able to distinguish low from intermediate and high Gleason scores with 92% accuracy on a cohort of 147 patients.\textsuperscript{34} Additionally, Tiwari et al. found that an embedding derived from magnetic resonance spectroscopy (MRS) and directional (Sobel, Kirsch, and gradient) and Haralick texture features of T2w MRI demonstrated promise for distinguishing low from high Gleason grades (a well established surrogate for PCa risk).\textsuperscript{57} Since the current clinical gold standard for PCa diagnosis relies on biopsies, which are highly invasive and prone to inaccuracies due to spatial sampling errors, a radiomic approach could potentially provide a safer and more accurate alternative for \textit{in vivo} assessment of PCa risk.

While there is substantial interest in imaging and radiomic markers, successful translation of these markers into the clinic may require a more comprehensive understanding of the underlying morphologic and histomorphometric underpinning of the most predictive and discriminating features. Towards this end, recent correlative studies have attempted to elucidate the biological and morphologic basis of radiomic features for different disease domains, including non-small cell lung cancer (NSCLC), colorectal cancer, Crohns disease, and prostate cancer (PCa). Ganeshan et al.
assessed correlations between contrast-enhanced and unenhanced computed tomography (CT)-derived radiomic texture and histopathologic markers of tumor hypoxia and angiogenesis, and found that radiomic features capturing heterogeneity (such as standard deviation) on contrast-enhanced CT were associated with these histologic markers. The authors note that an important component of intratumoral heterogeneity is blood vessel irregularity, which can result in a heterogeneous vascular supply that can cause areas of hypoxia. The authors suggest that since the radiomic features extracted are all measures of image heterogeneity, the radiomic features might be reflecting hypoxia and angiogenesis, which are histopathologic consequences of intratumoral vascular heterogeneity. Several other studies in other disease domains have identified correlations between similar radiomic features and histopathologic measures of vascularity and hypoxia. The mutually supportive findings of these studies indicate the potential for radiomic features to reflect histopathologic attributes.

Apart from exploring the correlation of imaging features with IHC attributes of vascularity and hypoxia, some researchers have been looking at the association of pathologist-based histologic features of PCa with imaging measurements. Langer et. al., in looking at ex vivo whole mount radical prostatectomy histology images with pre-operative in vivo MRI, found that ADC and T2w MRI were associated with gross measurements of nuclear and lumen space areas; however, none of the parameters or features they evaluated were discriminative of Gleason scores within their cohort. Chatterjee et. al. found that gland component volumes of epithelium, stroma, and lumen space in WMH were more strongly associated with Gleason pattern and ADC than the conventionally cited cellularity metrics of nuclear count and nuclear area. Kobus et. al. found that in vivo 1.5-T ADC was correlated with lumen
space area and inversely associated with nuclear area on WMH, and lumen space area was correlated with Gleason pattern. The associations these studies identified indicate strong potential for identifying morphologic correlates of imaging appearance. However none of these studies specifically looked at radiomic measurements from the MRI scans.

While radiomic features can represent subtle and complex image properties such as texture and orientation, quantitative histomorphometry (QH) features can quantify morphologic attributes of texture, shape, and architecture that may not be visible to a physician. Indeed, there has been an extensive amount of work demonstrating the effectiveness of QH features for detecting and assessing the aggressiveness of PCa. For example, Lee et. al. showed that QH features of gland lumen orientations were more disordered in more versus less aggressive prostate cancer, and could more accurately predict biochemical recurrence in PCa than post-operative nomograms. QH features may therefore provide more detailed characterization of histopathology than gross measurements of tissue component areas. This could then potentially provide a more informative set of measurements for correlating with in vivo imaging features. For instance, since radiomic features such as Sobel, Kirsch, Gabor, and gradient filters can capture edge orientation information within the image, and Haralick texture features can capture image texture and heterogeneity, this then raises the question of whether these directional and Haralick features that were able to distinguish low- from high-risk PCa are actually capturing fundamental disorder in gland lumen orientation architecture. This careful quantitative correlative analysis of radiomic features with morphologic features has not, to the best of our
knowledge, been previously done, and would go a long way in explaining the basis for why specific radiomic features work.

In this study, we utilize radiomic features extracted from T2w MRI and QH features extracted from hematoxylin and eosin (H&E)-stained histology. We utilize deformable registration to map cancer annotation regions from digitized ex vivo H&E-stained histopathology sections onto corresponding preoperative in vivo T2w MRI slices. Then, correlative analysis is performed using three main experiments. First, the top seven radiomic features most discriminative of Gleason scores are identified on a per-region basis using feature selection over 1000 repetitions of cross-validation. Second, the QH features that potentially form the morphologic basis for these radiomic features are identified by correlating each QH feature with each of the top seven radiomic features, using multiple hypothesis correction to determine which correlations are statistically significant. If a QH feature is predictive of Gleason score, and it is correlated with a radiomic feature that is also predictive of Gleason score, then it is likely that the QH feature is part of the morphologic basis for the discriminability of that radiomic feature. Therefore, the third experiment involves identifying the subset of the most predictive QH features found in the second experiment to be correlated with predictive radiomic features. The results of each experiment are then validated on an independent dataset from a different institution to test the generalizability of these findings. The remainder of the paper is organized as follows. The next section describes the details of patient selection, data preprocessing, and feature extraction. Then, the experimental design is described. Finally, we present and discuss the findings of these experiments in the context of previous work, and end with concluding remarks in the last section.
3.2 Methods

3.2.1 Ethics Statement

Ethics Statement

Data analysis was waived review and consent by the IRB board, as all data was being analyzed retrospectively, after de-identification. All experimental protocols were approved under the IRB protocol # 02-13-42C with the University Hospitals of Cleveland Institutional Review Board, and all experiments were carried out in accordance with approved guidelines.

3.2.2 Data Collection and Processing

Subjects who underwent 1.5 or 3 Tesla (T) T2w MRI prior to radical prostatectomy at two different institutions were eligible for this study: (1) 54 patients from University of Pennsylvania (UPenn) imaged between 2009 and 2011, and (2) 17 patients from St. Vincent’s Hospital (SV) imaged between 2012 and 2014. 23 out of the 54 patients from UPenn were selected to include a spectrum of cases with different Gleason scores, and 13 out of the 17 patients from SV were selected by excluding 4 patients whose tumor regions were too small to be accurately co-registered with MRI. Data from UPenn served as our training cohort, and data from SV served as our independent validation cohort. Details regarding MR scanning and acquisition are summarized in Table 1.

Each surgically resected prostate gland was fixed in formalin, embedded in paraffin, and sectioned axially in a plane perpendicular to the urethral axis from apex to base in 3-4 mm sections. Each slice was then sectioned into four quadrants, stained
Table 3.1. MR Parameters

<table>
<thead>
<tr>
<th>Cohort (Training)</th>
<th>N</th>
<th>Sequence</th>
<th>Scanner</th>
<th>Echo Time (ms)</th>
<th>Repetition Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
<td>T2w MRI</td>
<td>3T Verio®, Siemens</td>
<td>107-127</td>
<td>3690-7090</td>
</tr>
<tr>
<td>2 (Validation)</td>
<td>11</td>
<td>T2w MRI</td>
<td>3T, Philips Medical Systems</td>
<td>67-100</td>
<td>2525-3567</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>T2w MRI</td>
<td>1.5T Siemens</td>
<td>119</td>
<td>3760</td>
</tr>
</tbody>
</table>

with hematoxylin and eosin (H&E), and digitized via an Aperio whole slide scanner at 20X magnification and 0.5 um/pixel resolution. Quadrants were then digitally reconstructed into PWMHSs using previously presented tools. A pathologist (R.E.) and radiologist (J.G) then worked together to identify corresponding T2w MRI and PWMHS slices, which were then registered with corresponding T2w MRI slices using thin plate splines (TPS), a landmark based registration method. Manually selected landmarks were used to align prostate boundaries and visible internal structures (such as urethra) between the moving histology image and the target T2w MRI image.

Tumor regions were annotated by a pathologist (N.S., 8.5 years of experience; R.E., XX years of experience) on digitized H&E stained slides using ImageScope (Aperio) software. Tumor regions smaller than 18mm² were excluded from our study, as these regions were deemed too small to accurately co-register. Pathologists determined Gleason scores for each tumor region (M.F for St Vincents; R.E. for UPenn). Based on these Gleason scores, regions were categorized as either low- (score of 3+3), intermediate- (score of 3+4 or 4+3), or high- (score of 4+4 or higher) risk. Based on this categorization, the training cohort consisted of 21 low-, 30 intermediate-, and 14 high-risk prostate cancer regions, while the validation cohort consisted of 26 low-, 8 intermediate-, and 6 high-risk regions.
3.2.3 Quantitative Histomorphometric Feature Extraction

Our choice of QH features was based on previous work demonstrating their capability for detecting malignancy, discriminating between low and high Gleason grades, and predicting outcomes.\textsuperscript{62-65} These features quantify the following different characteristics from tissue specimens: gland lumen architecture, shape, and angularity; relative areas of different tissue components (lumen, nuclei, epithelium, stroma); and texture.

As a pre-processing step, four different tissue components were segmented as follows. Gland lumen were segmented using an automatic region-growing algorithm\textsuperscript{64} at 1.25x equivalent microscope zoom. Nuclei vs. cytoplasm, and epithelium vs. stroma were segmented using convolutional neural networks trained at 20x microscope zoom. Epithelium segmentations were combined with nuclei and cytoplasm segmentations to obtain epithelial nuclei and epithelial cytoplasm segmentations.

Using these tissue component segmentations, a total of 1024 QH features were extracted as summarized in Table 2 and illustrated in Figure 1. A variety of different classes of features were extracted:

(1) Tissue component density (TCD): Each segmented tissue component, we computed the area of each tissue component divided by the area of the parent tumor region as well as relative ratios of TCD between different tissue components.

(2) The classes of features described in the bullets below were summarized for each tumor region using 15 different descriptive statistics: mean, standard deviation, range, minimum, maximum, mode, median, variance, kurtosis, harmonic mean, skewness, mean, absolute deviation, interquartile range, disorder, min/max.
(3) Gland lumen shape: a total of 25 statistical measures of shape (e.g., smoothness, fourier descriptors, long/short distance)\textsuperscript{65} ratio were computed for each gland lumen segmentation perimeter, resulting in a total of 375 unique shape features.

(4) Graph-based architectural: These features were computed using statistical measures of graph characteristics extracted from voronoi diagrams, Delaunay triangulations, and minimum-spanning-tree graphs.\textsuperscript{71,72} The vertices of the graphs represent segmented gland lumen.

(5) Nearest-neighbor (NN) architectural: These features quantify the number of gland lumen within specified radii, as well as the minimum radii needed to enclose different numbers of gland lumen.

(6) Co-occuring gland angularity (CGA)\textsuperscript{66} features: The dominant orientation of each gland lumen is identified, then co-occurrence features capturing the statistical relationships between neighboring lumens are computed.

(7) Haralick texture features.\textsuperscript{51}

### 3.2.4 Radiomic Feature Extraction

Our choice of radiomic features was based on previous work demonstrating their promise for detecting malignancy, and discriminating between less and more aggressive cancer in a number of different disease domains. These features are based on quantifying the following different image characteristics: signal intensity, texture, orientation, and co-occuring gradient orientations.

Prior to feature extraction, T2w MRI signal intensity was corrected for intensity drift, which is a well-known issue in MRI that causes voxel intensity values to not
Figure 3.1. (a) An H&E-stained histology tumor patch. (b) Gland lumen are segmented using an automatic region-growing algorithm within the tumor region. (c) Segmentation boundaries enclosing the gland lumen provide measures of shape, such as Fourier descriptors, which are decompositions that capture various levels of shape complexity, as illustrated here with Fourier descriptor 9 (blue = low, red = high). Gland lumen architectural features are extracted from (d) voronoi diagram, (e) delaunay triangulation, and (f) minimum spanning tree. (g) Identification of directional tensors for each segmented gland lumen allows extraction of features such as co-occurring gland angularity disorder (arrows and colors depict tensor orientations from 0 to $\pi$). (h) Haralick texture features are computed within local rectangular windows within the tumor region, resulting in a texture value for each pixel, with Haralick correlation shown here (low = blue, high = yellow). (i) Tissue component densities are computed from gland lumen (green), epithelium (blue), nuclei (yellow), and stroma (uncolored).

have a fixed tissue-specific meaning.\textsuperscript{73} To correct for intensity drift, T2w MRI images from the training and validation cohorts were standardized to a template distribution, defined as the per-patient median of intra-prostatic pixel intensities from the training cohort, using the algorithm presented by Nyul et. al.,\textsuperscript{74} as illustrated in Figure 2.
Table 3.2. List of Extracted QH Features

<table>
<thead>
<tr>
<th>Modality</th>
<th>Type</th>
<th>Features</th>
<th>No. of Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>H&amp;E stained histology (TCD)</td>
<td>Tissue Component Density</td>
<td>Lumen, epithelium, stroma, nuclei, epithelial nuclei, and epithelial cytoplasm fractions. Epithelium/stroma, epithelium/lumen, lumen/stroma ratios.</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Shape</td>
<td>Area ratio, distance ratio, standard deviation of distance, variance of distance, long/short distance, perimeter ratio, smoothness, invariant moments 1-7, fractal dimension, fourier descriptors 1-10.</td>
<td>375</td>
</tr>
<tr>
<td></td>
<td>Architectural: voronoi diagram</td>
<td>Polygon area, perimeter length, and chord length.</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Architectural: delaunay triangulation</td>
<td>Triangle area and side length.</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Architectural: minimum spanning tree (MST)</td>
<td>Edge length.</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Architectural: nearest neighbors (NN)</td>
<td>(a) Distance to 3, 5 and 7 NN and (b) number of NN in 10, 20, 30, 40, 50 pixel radius (PR).</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>Architectural</td>
<td>Area, number, and density of polygons.</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Texture</td>
<td>Contrast energy, contrast inverse moment, contrast average, contrast variance, contrast entropy, average, variance, entropy, energy, correlation, information measure 1, information measure 2.</td>
<td>195</td>
</tr>
<tr>
<td></td>
<td>Co-occurring Gland Angularity (CGA)</td>
<td>Contrast energy, contrast inverse moment, contrast average, contrast variance, contrast entropy, average, variance, entropy, energy, correlation, information measure 1, information measure 2.</td>
<td>195</td>
</tr>
<tr>
<td></td>
<td>Cell Cluster Graph (CCG)</td>
<td>Number of Nodes, Number of Edges, Average Degree, Average Eccentricity, Diameter, Radius, Average Eccentricity 90%, Diameter 90%, Radius 90% Average Path Length, Clustering Coefficient C, Clustering Coefficient D Clustering Coefficient E, Number of connected components, giant connected component ratio average connected component size, number isolated nodes, percentage isolated nodes number end nodes, percentage end nodes, number central nodes, percentage central nodes mean edge length, standard deviation edge length, skewness edge length, kurtosis edge length.</td>
<td>37</td>
</tr>
</tbody>
</table>

Then, radiomic features were extracted using a large set of co-occurrence methods and convolutional filter kernels, including Gabor, Haralick, and Laws features, as illustrated in Figure 3. A total of 157 2D radiomic feature responses were computed, and then summarized for each tumor region using 13 different descriptive statistics (mean, standard deviation, range, minimum, maximum, mode, median, variance,
kurtosis, harmonic mean, skewness, mean, absolute deviation, interquartile range), resulting in a total of 2001 radiomic features, as summarized in Table 3.

Figure 3.2. (a) non-standardized intensity distribution (red) vs. standardization template (blue). (b) Standardized intensity distribution (red) vs. standardization template (blue).

<table>
<thead>
<tr>
<th>Table 3: List of Extracted Radiomic Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Modality</strong></td>
</tr>
<tr>
<td>T2w MRI</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Entropy</td>
</tr>
<tr>
<td>Gradient features and gradient-like kernel operations</td>
</tr>
<tr>
<td>Gabor</td>
</tr>
<tr>
<td>Haralick texture</td>
</tr>
<tr>
<td>Laws texture energy</td>
</tr>
</tbody>
</table>
Figure 3.3. (a) A T2w MRI of the prostate (tumor outlined in red). (b) Zoomed-in view of tumor. Sample radiomics feature maps from each class of radiomic features are plotted (blue = low, yellow = high). (c) Standard deviation within 9x9 windows. (d) Entropy within 5x5 windows. (e) Sobel yx filter. (f) Kirsch 1 filter. (g) Gabor cosine filter at an angle of \( \frac{\pi}{8} \) and wavelength 4. (h) Haralick correlation within 5x5 windows. (i) Laws feature 15.
3.3 Experimental Design

The morphologic basis for the discriminability of radiomic features on MRI was identified via a sequence of three experiments:

(a) **Identifying radiomic features discriminable of Gleason score:** The set of radiomic features most discriminable of Gleason score was identified by performing minimum redundancy maximum relevance (mRMR) feature selection\(^7^5\) on the training cohort. To reduce the likelihood of over-fitting the predictive model to the training data, the top seven (roughly one tenth of the number of samples) most frequently selected features from one thousand repetitions of three-fold cross-validation (CV) were identified. The predictive performance was averaged over all runs of CV to obtain an initial estimate of the model’s predictive potential. These seven features were then used to train a random forest classifier, which has been shown to be most prognostic and stable classifiers for radiomic features\(^7^6\) on the entire training cohort. The classifier’s generalizability was then evaluated by testing on the independent validation cohort.

(b) **Identifying QH features correlated with discriminable radiomic features:** Identifying QH features that are correlated with the discriminative radiomic features identified in step 1 is necessary because any QH feature that is part of the morphologic basis of those radiomic features must also be associated with those radiomic features. Although demonstrating that a set
of QH features is correlated with a set of radiomic features is not necessarily demonstrative of a causal relationship by itself, it can provide promising hypotheses for further validation and explanation.

To identify the set of correlated QH features, the pairwise correlation between each of the 7 radiomic features selected in step 1 and each of the 828 informative QH features was computed, resulting in a total of 5796 pairs of features being tested for correlation. The correlations identified as statistically significant were then evaluated on the independent validation dataset to assess their generalizability.

(c) **Identifying the most discriminative QH features that are correlated with the discriminative radiomic features:** If a QH feature is discriminable of Gleason score, and it is correlated with a radiomic feature that is also discriminable of Gleason score, then it is likely that the QH feature is part of the morphologic basis for the discriminability of that radiomic feature. Therefore, this third and final experimental step involves identifying the subset of QH features identified in step 2 that are most discriminative. As in step 1, we select the top seven features from one thousand repetitions of CV, and then evaluate discriminability on the independent validation cohort.

### 3.3.1 Statistical Evaluation

**Selection frequency to identify top QH or radiomics features:** Selection frequency for each feature was computed as the total number of folds in which the feature was selected among the top 7 features for that training fold, divided by the
total number of training folds. The top 7 radiomic and QH features were selected by choosing those with the highest selection frequencies over all repetitions of CV.

**AUC to evaluated classifier performance:** Receiver operating characteristic (ROC) curves were used to evaluate classification performance. The overall ROC and upper- and lower-bound curves were computed as the mean of curves over all repetitions of CV. Upper and lower bounds were computed using threshold averaging within each repetition of CV.

**Multiple hypothesis correction to evaluate correlations:** A correlation coefficient and associated p-value were computed for each pair of features using Spearman’s rank correlation test, which is robust to outliers and quantifies monotonicity. Statistically significantly correlated features were identified using False Discovery Rate (FDR) for multiple hypothesis testing. On the training set and validation set, statistically significant correlations were identified as those with FDR $\leq 0.05$ and FDR $\leq 0.2$, respectively.

### 3.4 Results

**Experiment 1:** The top 7 selected radiomic features and their selection frequencies are listed in Table 3, and ROC curves of their discriminability on the training and validation cohorts are shown in Figure 4a,b. This feature subset, as listed in Table 4, comprises 3 Gabor, 2 Laws, 1 Haralick texture, and 1 descriptive statistic feature, which were able to distinguish low- from intermediate- and high- risk prostate cancer tumors with an AUC of 0.69 on the training cohort and 0.71 on the validation cohort, as shown in Figure 4.
Table 3.4. Selected Radiomic and QH Features

<table>
<thead>
<tr>
<th>Radiomic Feature</th>
<th>Selection Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gabor: sin: theta = 1.1781: lambda = 2: Mean</td>
<td>40.1</td>
</tr>
<tr>
<td>Laws15: Mean</td>
<td>38.5</td>
</tr>
<tr>
<td>Haralick: correlation: ws5: Maximum</td>
<td>37.3</td>
</tr>
<tr>
<td>Laws14: Kurtosis*</td>
<td>21.0</td>
</tr>
<tr>
<td>Median: Window Size 7: Maximum</td>
<td>20.3</td>
</tr>
<tr>
<td>Gabor: cos: theta = 0.3927: lambda = 4: Variance*</td>
<td>19.7</td>
</tr>
<tr>
<td>Gabor: sin: theta = 0: lambda = 2: Median</td>
<td>19.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>QH Feature</th>
<th>Selection Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shape: Fourier Descriptor 9 Median</td>
<td>59.9</td>
</tr>
<tr>
<td>Shape: Fourier Descriptor 3 Minimum</td>
<td>59.7</td>
</tr>
<tr>
<td>Shape: Fourier Descriptor 5 Disorder</td>
<td>43.3</td>
</tr>
<tr>
<td>Shape: Fractal Dimension Mean</td>
<td>40.8</td>
</tr>
<tr>
<td>Shape: Fourier Descriptor 9 Kurtosis*</td>
<td>36.1</td>
</tr>
<tr>
<td>Shape: Invariant Moment 6 Interquartile Range*</td>
<td>34.7</td>
</tr>
<tr>
<td>Shape: Fractal Dimension Maximum*</td>
<td>33.6</td>
</tr>
</tbody>
</table>

*Features significantly correlated on the independent validation dataset.

**Experiment 2:** Of the 5796 pairs of discriminative radiomic features and QH features, 401 pairs consisting of 234 unique QH features were statistically significantly correlated on the training cohort (FDR $< 0.05$). These correlations are visualized in Figure 5, which plots the maximum correlation coefficient for each QH feature group. These 401 correlations were then tested on the independent validation cohort, resulting in 56 out of 401 statistically significant (FDR $< 0.2$) correlations. Three of the 56 pairs of features contained radiomic and QH features that were both selected among the top discriminative features in experiments 1 and 3, as highlighted in Table 4. The differential expression between low- and intermediate-/high-risk tumors of one of these pairs of features is illustrated in Figure 6.

**Experiment 3:** The top 7 QH features selected from the set of 401 unique correlated QH features identified in experiment 2 are listed in Table 4. This feature subset
comprises 7 shape features, which were able to distinguish low- from intermediate- and high-risk prostate cancer tumors with an AUC of 0.75 on the training cohort and 0.66 on the validation cohort, as shown in Figure 4. Since these features are both discriminative of Gleason score and correlated with discriminative radiomic features, they may therefore form the morphologic basis to help explain the discriminability of the radiomic features identified in experiment 1.

Figure 3.4. Receiver operating characteristic (ROC) curves depicting random forest classification performance using radiomic features selected in experiment 1 (a, b), and QH features selected in experiment 3 (c, d) on both the training (a, c) and validation (b, d) cohorts.
3.5 Discussion

Radiomics can potentially enable less invasive stratification of prostate cancer risk on \textit{in vivo} imaging, however the morphologic basis for the discriminative radiomic features is currently unknown. In this study, we sought to identify the QH features that form this morphologic basis by performing (1) feature selection of discriminative radiomic features, (2) identification of QH features correlated with these discriminative radiomic features, and (3) feature selection of the most discriminative QH features correlated with radiomic features on a per-region basis. To accomplish this, we validated each experiment on an independent dataset from a different institution. Ex vivo radical prostatectomy histopathology images were carefully co-registered with pre-operative MRI to map the gold-standard disease annotations onto MRI. Comprehensive sets of radiomic and quantitative histomorphometric features were spatially co-localized to characterize each tumor region. This framework enabled us to identify robust correlations among subtle, complex, and sub-visual features that can accurately predict cancer risk.
Figure 3.6. Differential expression for a pair of strongly correlated and discriminative radiomic and QH features between tumors with Gleason score 6 (a-d) and Gleason score 9 (e-h). The radiomic feature values are plotted within the tumor (blue = low, yellow = high) shown within the whole prostate (a,e) as well as zoomed-in (b,f). The corresponding QH feature values are plotted within the tumor region on the segmented gland lumen boundaries (blue = low, red = high), shown within the whole tumor (c,g) as well as zoomed-in (d,h). Boxplots summarizing the differential expression of these features over all tumors are plotted (q, r). Note the low number of colors in the tumor with high (f) relative to low (b) Gleason score, indicating lower variance of the depicted Gabor feature within the high Gleason score tumor. Additionally, note the greater number of colors in the tumor with high (h) relative to low (d) Gleason score, indicating lower kurtosis of the depicted shape feature within the high Gleason score tumor.
This study identified a promising set of T2w MRI radiomic features as well as a set of QH features correlated with these radiomic features, which were both discriminative of Gleason score. Since the QH features are both discriminative and correlated with the discriminative radiomic features, they may partially explain the morphologic basis for the discriminability of these radiomic features. While it is difficult to demonstrate perfectly that there is a causal relationship between these features, we designed our experiments in such a way that we are reasonably confident that there is some form of causal relationship between these features. For instance, the relationships between some of the features identified in this study appear to have plausible biophysical explanations. As Gleason grade increases, the gland lumen shrinks and the orientations become more disorderly. The set of QH features discriminative of Gleason score that were identified in experiment 3 appear to be capable of reflecting these changes in gland lumen shape, as all 7 selected QH features are shape features. Since the identified radiomic features capture measures of orientation, it is plausible that they might be reflecting gland lumen shape characteristics on a larger scale. Although gland lumen are typically substantially smaller than the size of a voxel on MRI, it may be possible that the cumulative contributions of many smaller gland lumen shapes can produce concomitant changes in orientation characteristics of the T2w MRI signal as measured by Gabor filters. Since gross measurements of lumen area have been previously found to be correlated with T2w MRI signal intensity, it is not surprising that other characteristics of gland lumen and T2w MRI are correlated. Notably, one of the identified Gabor features was significantly correlated with 2 of the gland lumen shape features on both the training and the validation cohort, which indicates a robust association, especially when taking into account differences between
the radiology and histology data that could hinder comparability across institutions (e.g., varied voxel size, histological staining, etc.).

There are a limited number of prior reports in the literature evaluating the ability of T2w MRI radiomic features to predict Gleason score.\textsuperscript{34,35,57,79} However, none of them have evaluated this in a multi-site setting. In this study, we achieved AUCs of 0.68 and 0.7 on the training and validation cohorts, respectively, using T2w MRI radiomic features to predict Gleason score, indicating that these features are capable of generalizing to independent data from different institutions.
4 Conclusions and Future Directions

In this work, we presented two modules of a radiology-pathology fusion framework for improving the characterization of PCa:

First, we presented an automated method called AutoStitcher, which robustly and efficiently digitally stitches pseudo whole mount histological sections from multiple smaller tissue fragments with an accuracy and quality consistent with previous semi-automated methods\textsuperscript{12}. Reassembly of whole histological sections from smaller tissue fragments allows for visual and spatial co-registration with pre-operative in \textit{vivo} imaging. This approach can thus allow for spatially mapping disease extent annotated on the \textit{ex vivo} pathology images onto the \textit{in vivo} imaging. AutoStitcher is the first attempt at an algorithm for automated stitching of histology images, and is fundamentally different from algorithms used in similar problem domains in that it makes none of the assumptions of (a) overlap, (b) completeness, and (c) interlock. The algorithm utilizes a novel cost function that performs alignment using a combination of (i) a quantitative measure of image similarity and (ii) automatically detected fiducial points, and achieves computational efficiency on high-resolution histology images by optimizing this function hierarchically at multiple resolution levels.
AutoStitcher resulted in PWMHS reconstructions that were evaluated as quantitatively and qualitatively similar to reconstructions performed manually by humans (via HistoStitcher), as measured by human-selected and automatically detected fiducial points as well as the difference between reconstruction contours (difference error of 3% for all measures considered, over 113 PWMHS reconstructions). Our experimental design included separate training and testing cohorts (from 1 institution) and independent validation of the method and parameters on data from a different institution. The differences in the performance of AutoStitcher on data from across the 2 institutions was statistically equivalent to within 1%. AutoStitcher requires approximately 4-10 minutes to reconstruct a single section with a negligible amount of user-interaction, while an expert performing manual stitching using HistoStitcher requires 10-20 minutes of user-interaction. When considering large data cohorts that require reconstruction of large numbers of sections, AutoStitcher could thus save substantial amounts of time and substantially reduce user error.

Second, this thesis identified promising associations between \textit{in vivo} T2w MRI radiomic features capable of predicting prostate cancer risk and quantitative histomorphometric features. Radiomics can potentially enable less invasive stratification of prostate cancer risk on in vivo imaging, however the morphologic basis for the discriminative radiomic features is currently unknown. In this study, we sought to identify the QH features that form this morphologic basis by performing (1) feature selection of discriminative radiomic features, (2) identification of QH features correlated with these discriminative radiomic features, and (3) feature selection of the most discriminative QH features correlated with radiomic features on a per-region basis. To accomplish this, we validated each experiment on an independent dataset
from a different institution. Ex vivo radical prostatectomy histopathology images were carefully co-registered with pre-operative MRI to map the gold-standard disease annotations onto MRI. Comprehensive sets of radiomic and quantitative histomorphometric features were spatially co-localized to characterize each tumor region. This framework enabled us to identify robust correlations among subtle, complex, and sub-visual features that can accurately predict cancer risk, thus providing potential explanations for the morphologic basis of some of these radiomic features.

This work, however was subject to some limitations. Due to the many degrees of freedom in determining an optimal PWMHS reconstruction, the current version of AutoStitcher does not take into account scaling or non-linear warping. Additionally, since gold standard references, such as high resolution \textit{ex vivo} MR acquired prior to sectioning, were not available for comparison, AutoStitcher was evaluated with respect to manually reconstructed PWMHSs. Finally, the correlative analysis described in Chapter 3 was performed on relatively small datasets, and was limited to analysis of MRI and H&E histopathology.

Envisioned future directions of this work are as follows:

(1) Incorporating compensation for scaling and non-linear warping into AutoStitcher, and exploring its utility in disease domains other than prostate cancer, such as Renal Cell Carcinoma.

(2) Evaluating AutoStitcher against gold standard references such as high resolution \textit{ex vivo} MR acquired prior to sectioning.

(3) Performing correlative analysis on larger datasets from multiple institutions to further validate the identified associations for radiomic, QH features. Also
Conclusions and Future Directions

possibly extending the framework to accommodate imaging correlations with proteomic and genomic features.
References


[47] Cuishan Liang, Yanqi Huang, Lan He, Xin Chen, Zelan Ma, Di Dong, Jie Tian, Changhong Liang, and Zaiyi Liu. The development and validation of a CT-based radiomics signature for the preoperative discrimination of stage I-II and stage III-IV colorectal cancer. Oncotarget, April 2016.


